frontiers in Genetics

METHODS
published: 12 January 2021
doi: 10.3389/fgene.2020.622234

Predicting lincRNA-Disease Association in Heterogeneous Networks Using Co-regularized Non-negative Matrix Factorization

Yong Lin[1]* *and Xiaoke Ma*[2]*

[1] School of Physics and Electronic Information Engineering, Ningxia Normal University, Guyuan, China, [2] School of Computer Science and Technology, Xidian University, Xi'an, China

Long intergenic non-coding ribonucleic acids (lincRNAs) are critical regulators for many complex diseases, and identification of disease-lincRNA association is both costly and time-consuming. Therefore, it is necessary to design computational approaches to predict the disease-lincRNA associations that shed light on the mechanisms of diseases. In this study, we develop a co-regularized non-negative matrix factorization (aka *Cr-NMF*) to identify potential disease-lincRNA associations by integrating the gene expression of lincRNAs, genetic interaction network for mRNA genes, gene-lincRNA associations, and disease-gene associations. The Cr-NMF algorithm factorizes the disease-lincRNA associations, while the other associations/interactions are integrated using regularization. Furthermore, the regularization does not only preserve the topological structure of the lincRNA co-expression network, but also maintains the links "lincRNA → gene → disease." Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art methods in terms of accuracy on predicting the disease-lincRNA associations. The model and algorithm provide an effective way to explore disease-lncRNA associations.

Keywords: disease-lincRNA association, non-negative matrix factorization, heterogeneous network, regularization, network analysis

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Peng Gao,
Children's Hospital of Philadelphia, United States
Zhong-Yuan Zhang,
Central University of Finance and Economics, China
Wanxin Tang,
Sichuan University, China

***Correspondence:**
Yong Lin
linyong@nxu.edu.cn
Xiaoke Ma
xkma@xidian.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *28 October 2020*
Accepted: *03 December 2020*
Published: *12 January 2021*

Citation:
Lin Y and Ma X (2021) Predicting lincRNA-Disease Association in Heterogeneous Networks Using Co-regularized Non-negative Matrix Factorization. Front. Genet. 11:622234. doi: 10.3389/fgene.2020.622234

1. INTRODUCTION

Long intergenic non-coding RNAs (lincRNAs) are transcripts whose lengths are greater than 200 nucleotides with little or no protein coding potential (Kapranov et al., 2007; Mercer et al., 2009; Wang and Chang, 2011). In the traditional view, lncRNAs are considered as "junk RNAs" because they do not code protein sequences. However, it has been proven that many lncRNAs are dysregulated in human cancers and implicated in disease progression through modulating apoptosis, increasing cellular oncogenic potential, or inhibiting tumor growth (Wilusz et al., 2009; Taftet al., 2010).

With the advent of the next generation sequencing (NGS) techniques, a large number of lincRNAs have been identified (Guttman et al., 2009, 2010; Wang et al., 2009; Popadin et al., 2013), providing a great opportunity to investigate the functions of lncRNAs. Unfortunately, very few lincRNAs have been depicted with explicit molecular mechanisms in cancers through biological experiments or computational approaches (Guo et al., 2013; Zhao et al., 2016; Tang et al., 2017).

Thus, discovering lincRNA patterns that are associated with cancers is urgently needed as it sheds light on the underlying mechanism of diseases.

Therefore, great efforts have been devoted to investigating the functions or patterns of lincRNAs by analyzing omics data, such as DNA sequences, expression profiles, and genomic annotations. For instance, Liao et al. (2011) constructed a co-expression network for protein-coding genes and lincRNAs, and predicted the functions of lincRNAs via analyzing the constructed co-expression network. However, it has been criticized because of the fact that the gene expression profile cannot fully characterize the connections between genes and lincRNAs. To overcome this problem, Guo et al. (2013) developed a global prediction algorithm to infer probable functions of lincRNAs at a large scale by integrating gene expression, a protein-protein interaction (PPI) network, and DNA sequences. Ma et al. (2017a) designed a pipeline to discover disease related lincRNA modules across various clinical stages of cancers, rather than predicting the functions of lincRNAs. Ning et al. (2016) extracted the disease associated with SNPs within human lincRNAs.

Despite numerous research contributions to extract various patterns of lincRNAs, few efforts have been devoted to analyzing lincRNA-disease associations, which can be used to predict implicated diseases. The available methods to predict lincRNA-disease associations can be categorized into two classes: biological experiments-based methods and computational based approaches. The biological experiment-based methods have been criticized because they are time-consuming and costly. Computational based approaches are thus an alternative which can provide critical clues for biologists in revealing the mechanisms of diseases.

However, it is non-trivial to design effective and efficient algorithms to predict the lincRNA-disease associations largely due to two reasons. First, to infer the lincRNA-disease associations, large-scale known association data is a prerequisite. Second, diseases, such as cancers, are complex and difficult to characterize. Thus, it is wise to predict the lincRNA-disease associations by integrating omics data with an immediate purpose to improve the accuracy of prediction. Regarding the first concern, as more experimentally validated lincRNA-disease associations accumulate, researchers have summarized these associations as lincRNA-disease database, such as LncRNADisease (Chen et al., 2012) and Lnc2Meth (Zhi et al., 2018). These known associations provide a great opportunity to infer the lincRNA-disease associations.

Regarding the second concern, many algorithms have been developed to address this issue. For example, Yang et al. (2014) predicted the lncRNA-disease associations by constructing two biological networks, such as lncRNA-implicated disease network and disease network. Then, a propagation algorithm is applied to extract similar lncRNAs and diseases from those constructed networks. To integrate the expression profile, Chen et al. (2012) designed the Laplacian regularized least squares for lncRNA-disease associations, where the tissue expression profiles of intergenic lncRNA (lincRNA) from the Human BodyMap LincRNA project (Cabili et al., 2011). Zhang et al. (2017) proposed a label propagation algorithm to predict lncRNA-disease associations by integrating multiple heterogeneous networks. Fu et al. (2018) developed a matrix factorization-based model to predict disease-lncRNA associations, where multiple data matrices from various heterogeneous sources are factorized into low-rank matrices. Lan et al. (2017) designed a web server for the prediction of the lncRNA-disease. These algorithms achieve promising performance in inferring lncRNA-disease associations.

However, all of these studies solely focus on ranking lncRNA-disease associations via integrating the additional features of lncRNA genes and diseases, which cannot make use of the known prior knowledge to further improve the performance of algorithms. The latent features facilitate the identification of biological patterns, such as copy number and driver genes (Xi et al., 2020a,b). Actually, compared to the lincRNAs, knowledge of protein-coding genes is more redundant. How do you effectively incorporate the prior information into algorithms in order to perform a particular function and/or to infer a disease in the biological systems? For instance, Liao et al. (2011) made use of the gene-lncRNA relation to predict the functions of lncRNAs, implying that integration of omic data is promising for improving the performance of algorithms. Recently, Biswas et al. (2015) designed the *iNMF* algorithm by integrating expression profiles of protein-coding and lncRNA genes, lncRNA-disease and gene-disease associations, and gene genetic interaction networks to predict the diseases of lncRNAs. The experimental results demonstrate that it is wise to integrate omics data to infer lncRNA-disease associations a major motivation for this study.

iNMF jointly factorizes expression profiles of lncRNA and protein-coding genes. However, the method ignores the fact that lncRNAs execute their functions via interactions between them. Thus, we develop a novel algorithm, named co-regularized NMF (Cr-NMF), to predict lincRNA-disease associations via the heterogeneous network with multiple types of association, including lincRNA co-expression, lincRNA-disease, gene-disease, gene genetic and lincRNA-gene associations (As shown in **Figure 1**). The Cr-NMF algorithm decomposes the lincRNA-disease associations into the feature and coefficient matrices; the latent features for lincRNAs regularize the topological structure of lincRNA co-expression network. Furthermore, we also expect that the factorization reflects paths from *lincRNA* $\rightarrow$ *gene* $\rightarrow$ *disease*, which is also represented by regularization. Compared to state-of-the-art algorithms, the proposed algorithm is more accurate in the lincRNA-disease prediction. The proposed model and method provide an effective strategy to predict lncRNA-disease associations.

The rest of this study is organized as follows. Section 2 presents the details of the proposed algorithm. Then, in section 3, we set up experiments to validate the performance of Cr-NMF. Finally, conclusions are drawn in section 4.

2. ALGORITHM

The algorithm consists of two major components: the objective function construction and optimization rules, as shown in

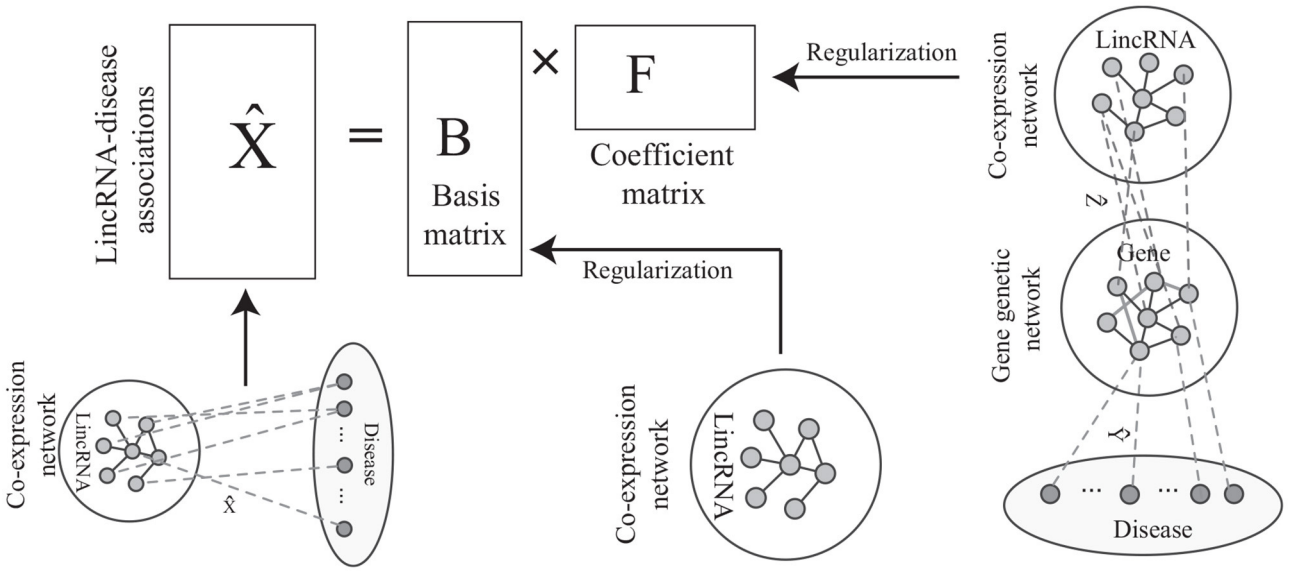

FIGURE 1 | Overview of the Cr-NMF algorithm. It factorizes the known disease-lincRNA associations, regularizing the other associations/interactions. Specifically, the basis matrix is the latent representation of lincRNAs, which preserves the topological structure of lincRNA co-expression networks. The coefficient matrix regularizes the links "lincRNA → gene → disease".

TABLE 1 | Notations.

Notation	Definition and description
n_g, n_d, n_l	Number of genes, diseases, and lincRNAs
$G^{[g]}$	Gene genetic interaction network
$G^{[l]}$	lincRNA co-expression network
$\widehat{X}$	known lincRNA-disease associations
$\widehat{Y}$	known gene-disease associations
$\widehat{Z}$	genes-lincRNAs associations
$W^{[g]}, W^{[l]}$	weighted adjacency matrix for $G^{[g]}$ and $G^{[l]}$
$w_{ij}^{[g]}$	the element at i-th row jth column in matrix $W^{[g]}$
D	the degree diagonal matrix, i.e., $D = diag(d_1, \ldots, d_n)$
$\overline{W}^{[g]}$	normalized $G^{[g]}$, i.e., $\overline{W}^{[g]} = D^{-1/2} W^{[d]} D^{-1/2}$
$W^{'}$	transpose of matrix W
$w_{i.}$	the i-th row of matrix W
$w_{.j}$	the j-th column of matrix W
$\|W\|_F$	Frobenius norm of matrix W
$Tr(W)$	the Tr of matrix W, i.e., $Tr(W) = \sum_i w_{ii}$

Figure 1. The procedure and analysis of the proposed algorithm are addressed in this section.

2.1. Notations

Before presenting the detailed description of the proposed algorithm, let us introduce some terminologies that are widely used in the sections that follow.

The notations for the algorithm are summarized in **Table 1**. Let n_g be the number of genes, n_d be the number of diseases, n_l be the number of lincRNAs. The lincRNA co-expression network is denoted by $G^{[l]} = (V^{[l]}, E^{[l]})$, where $V^{[l]}$ is the set of lincRNAs and $E^{[l]}$ is the interaction sets based on lincRNA co-expression coefficients. The adjacency matrix for $G^{[l]}$ is denoted by matrix $W^{[l]}$, where $w_{ij}^{[l]}$ is the weight on edge (i, j) in $G^{[g]}$. Because $G^{[l]}$ is undirected, $W^{[l]}$ is symmetric. The degree of the i-th lincRNA in $G^{[l]}$ is defined as the sum of weights on the edges connecting to it, i.e., $d_i = \sum_i w_{ij}^{[l]}$. The degree matrix of $G^{[l]}$ is the diagonal one with degree sequence, i.e., $D^{[l]} = diag(d_i^{[l]}, \ldots, d_{n_l}^{[l]})$. Given network $G^{[l]}$, we construct a normalized Laplacian matrix $L^{[l]} = I - (D^{[l]})^{-1/2} W^{[l]} (D^{[l]})^{-1/2}$. Analogously, we construct the normalized Laplacian matrix for $G^{[g]}$ as $L^{[g]} = I - (D^{[g]})^{-1/2} W^{[g]} (D^{[g]})^{-1/2}$.

The known lincRNA-disease associations are represented by $\widehat{X}$, where the row represents a lincRNA and column denotes a disease. The known gene-disease associations are denoted by $\widehat{Y}$, where rows correspond to genes and columns denote diseases. Thegene-lincRNA associations $\widehat{Z}$ are constructed based expression data, where the rows correspond to genes, columns to lincRNAs, and $z_{ij} = 1$ if the i-th gene and j-th lincRNA are associated with at least one disease, 0 otherwise.

2.2. Objective Function

NMF aims at learning the representation parts of the original data (Lee and Seung, 1999) by approximating the target matrix into the product of two low-ranking matrices. Specifically, given matrix W, NMF decomposes W into two non-negative matrices $B_{(m+n)\times k}$ and $F_{(m+n)\times k}$ such that

$$W \approx BF', s.t. B \geq 0, F \geq 0, \quad (1)$$

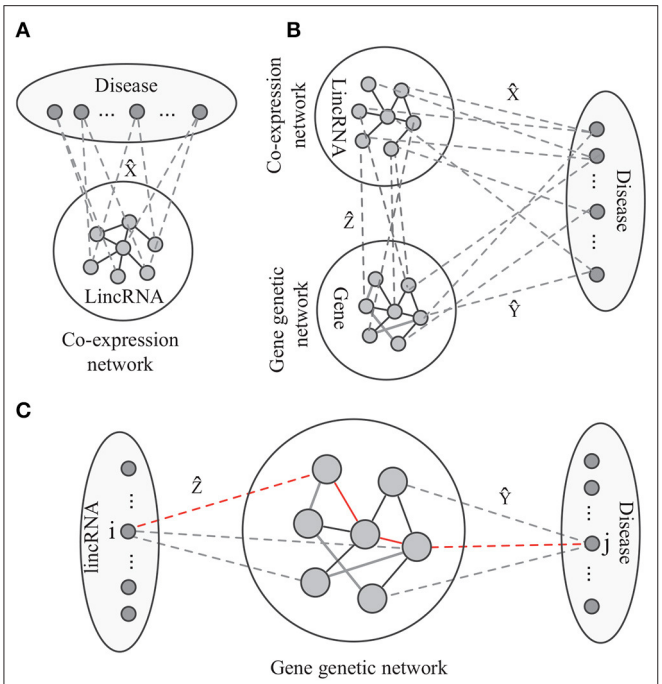

FIGURE 2 | The paradiagrams for disease-lincRNA association prediction. **(A,B)** The label propagation model predicts the disease-lincRNA associations by each single disease based on the lincRNA co-expression network, and **(C)** the proposed algorithm integrates lincRNA co-expression network, gene genetic networks and various associations. The dashed lines show disease-lincRNA association (matrix $\widehat{X}$), disease-gene association (matrix $\widehat{Y}$), and lincRNA-gene associations (matrix $\widehat{Z}$).

where B is the basis matrix and F is the feature matrix. NMF has been widely applied for graph analysis (Ma et al., 2018a), link prediction (Ma et al., 2017b, 2018b), bioinformatics (Chen and Zhang, 2016; Ma et al., 2016, 2018c).

As shown in **Figure 2A**, the label propagation-based model has been widely studied and successfully applied to predict phenotype-gene associations (Hwang and Kuang, 2010; Vanunu et al., 2010; Hwang et al., 2011). The model aims at identifying the disease-lincRNA associations X under some constraints. Thus, the objective function of label propagation model is defined as

$$O_{lp} = \theta Tr(X'L^{[l]}X) + (1-\theta)\|X - \widehat{X}\|_F^2, \tag{2}$$

where $\theta \in (0,1)$ is a parameter to balance the contributions of the two terms, $Tr(\cdot)$ is the Tr function and $\|\cdot\|_F$ is the Frobenius norm. To further improve the performance of label propagation model, Petegrosso et al. (2017) proposed transfer learning-based label propagation model to integrate omics data to predict phenome-genome association.

Given the disease-lincRNA associations $\widehat{X}$, Cr-NMF first factorizes $\widehat{X}$ into the product of matrix B and F, i.e.,

$$\widehat{X} = BF, \quad s.t. \quad B \geq 0, F \geq 0, \tag{3}$$

where $B \in R^{n_l \times r}$ is the basis matrix, $F \in R^{r \times n_d}$ is the feature matrix, r is the number of latent variables (usually, $r \ll \min\{n_l, n_d\}$). By casting Equation (3) as an optimization form, we obtain the following objective function as

$$O_{NMF} = \frac{1}{2}\|\widehat{X} - BF\|_F^2, \quad s.t. \quad B \geq 0, F \geq 0. \tag{4}$$

On the one hand, matrix B is considered to be the representations of lincRNAs in the latent space, where each row $b_{i.}$ is interpreted as latent representation of the i-th lincRNA. We expect the latent representations in matrix B preserve the local topological structure of lincRNAs $G^{[l]}$. Specifically, if a pair of lincRNAs are close in terms of the latent representation, they are well connected in $G^{[l]}$ and vice versa. Cai et al. (2010) demonstrated that

$$\begin{aligned} O_{G^{[l]}} &= \frac{1}{2}\sum_i\sum_j \|b_{i.} - b_{j.}\|^2 w_{ij}^{[l]} \\ &= Tr(B'D^{[l]}B) - Tr(B'W^{[l]}B) \\ &= Tr(B'L^{[l]}B). \end{aligned} \tag{5}$$

On the other hand, the disease-lincRNA associations are also related to the topological structure of the gene interaction network, lincRNA-gene association (**Figure 2B**), and the disease-gene associations. The association between the i-th lincRNA and the j-th disease follows the pattern *lincRNA* $\rightarrow$ *gene* $\rightarrow$ *gene network* $\rightarrow$ *disease*. For example, in **Figure 2C**, the i-th lincRNA and j-th disease are connected by the red path. There is a good biological interpretation for this pattern: the lincRNAs transduce signal to the target genes. The dysfunctional signal possibly leading to an abnormal response via interaction among genes, resulting in diseases. Thus, the disease-lincRNA association w_{ij} can be defined as a product of weights on all the paths connecting the i-th lincRNA and j-th disease, i.e.,

$$x_{ij} = \sum_k \widehat{z}_{ik} w_{ij}^{[g]} \widehat{y}_{kj}. \tag{6}$$

The underlying assumption for Equation (6) is that the more paths connecting a lincRNA and disease, the more likely it is to be a true association. Transforming Equation (6) into matrix form, we obtain

$$X = \widehat{Z}W^{[g]}\widehat{Y}. \tag{7}$$

Transforming Equation (7) into an optimization problem, we obtain

$$O_{G^{[g]}} = \frac{1}{2}\|X - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \tag{8}$$

Because we use NMF to approximate X, Equation (8) is re-written as

$$O_{G^{[g]}} = \frac{1}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \tag{9}$$

Combining Equations (4,5), and (9), the objective function of the proposed algorithm is defined as

$$O = O_{NMF} + \alpha O_{G^{[l]}} + \beta O_{G^{[g]}}, \tag{10}$$

where parameter α, β control the contributions of two terms $O_{G^{[l]}}$ and $O_{G^{[g]}}$. The disease-lincRNA prediction problem is transformed into an optimization problem as

$$\begin{aligned} \min_{B,F} \quad & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2 \\ s.t. \quad & B \geq 0, F \geq 0. \end{aligned} \tag{11}$$

In the next subsection, we address how to optimize the problem in Equation (11).

2.3. Optimization

An iterative two-step strategy is adopted because direct optimization to Equation (11) is difficult, where we optimize matrices B and F by fixing parameters. At each iteration, either matrix B or F is optimized first, whereas the other is fixed. Iteration is repeated until the algorithm converges or the maximum number of iterations is reached.

Let the objective function of Equation (11), i.e.,

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \end{aligned} \tag{12}$$

We handle the non-negative constraints for matrices B and F using the Larange method. Specifically, let ϕ_{ij} and ψ_{ij} be the Larange multiplier for the constraints b_{ij} and f_{ij}, respectively. Considering $\Phi = [\phi_{ij}]$, $\Psi = [\psi_{ij}]$, the Larange $\mathcal{L}$ of Equation (12) can be formulated as

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2 + \Phi B + \Psi F. \end{aligned} \tag{13}$$

The partial derivatives of $\mathcal{L}$ with respect to basis matrix B and feature matrix F are calculated as

$$\frac{\partial \mathcal{L}}{\partial B} = (1+\beta)BFF' - \widehat{X}F' + 2\alpha L^{[l]}B - \widehat{Z}W^{[g]}\widehat{Y}F' + \Phi, \tag{14}$$

and

$$\frac{\partial \mathcal{L}}{\partial F} = B'\widehat{X} - B'BF + \beta B'BF - B'\widehat{Z}W^{[g]}\widehat{Y} + \Psi. \tag{15}$$

According to the Karush-Kuhn-Tucker conditions $\phi_{ij}b_{ij} = 0$ and $\psi_{ij}f_{ij} = 0$, we obtain the updated rules

$$B = \frac{\widehat{X}F' + \widehat{Z}W^{[g]}\widehat{Y}F'}{(1+\beta)BFF' + 2\alpha L^{[l]}B}B, \tag{16}$$

and

$$F = \frac{B'BF + B'\widehat{Z}W^{[g]}\widehat{Y}}{B'\widehat{X} + \beta B'BF}F. \tag{17}$$

The Cr-NMF algorithm is presented in Algorithm 1.

Algorithm 1: The Cr-NMF algorithm

Input:

$G^{[l]}$: Co-expression network for lincRNAs;
$M^{[g]}$: Expression profile for genes;
$M^{[l]}$: Expression profile for lincRNAs;
$\widehat{X}$: Known disease-lincRNA associations;
$\widehat{Y}$: Known disease-gene associations;
α, β: Parameters control relevant importance.

Output:

X: Predicted disease-lincRNA associations.

Step 1: Data Processing

1: Construct co-expression network $G^{[l]}$ for lincRNAs using expression profile $M^{[l]}$;
2: Construct gene-lincRNA associations $\widehat{Z}$ using $M^{[l]}$ and $M^{[g]}$;
3: Construct Laplacian matrix $L^{[g]}$ for $G^{[g]}$;
4: Construct Laplacian matrix $L^{[l]}$ for $G^{[l]}$;

Step 2: Matrix Factorization

5: Make initial matrices B and F;
6: Update matrix B according to Equation (16);
7: Update matrix F according to Equation (17);
8: Goto Step 5 until the algorithm is convergent;

Step 3: Predict disease-lincRNA associations

9: Predict disease-lincRNA association as $X = BF$;
10: **return** X

2.4. Algorithm Analysis

The complexity of algorithm is investigated. On the space complexity of algorithm, the space for the gene genetic network is $O(n_g^2)$. The space for lincRNA co-expression network is $O(n_l^2)$. The space of disease-lincRNA association, disease-gene associations, and gene-lincRNA association is $O(n_d n_l)$, $O(n_d n_g)$, and $O(n_g n_l)$, respectively. The space of basis matrix B and feature matrix F is $O((n_l+n_d)r)$, where r is the number of latent variables. Thus, the total space of Cr-NMF is $O(n_l^2 + n_g^2 + n_d n_l + n_d n_g + n_g n_l + (n_l + n_d)r)$. Because $n_d \ll n_g$ and $n_l \ll n_g$, the total space of the proposed method is $O(n_g^2)$.

The running time of the proposed algorithm depends on the updating rules in Equations (16) and (17). Thus, the time complexity of Cr-NMF is the same as that of NMF, i.e., $O(tkn^2)$, where t is the number of iteration (Lin, 2007). Thus, the overall running time for RNMF-MM is $O(tkn^2) + O(n^2) = O(tkn^2)$, indicating that the proposed algorithm is also efficient in comparison with the NMF algorithm.

3. RESULTS

In this section, we validate the performance of the proposed algorithm. The data, parameter selection as well as the performance of algorithms are addressed in turns.

3.1. Data

The lincRNAs are downloaded from the Human BodyMap project, which provides a catalog of lincRNAs from RNA-seq data across 22 tissues (Cabili et al., 2011). The catalog contains transcript expression profile across the tissues using the Cufflinks (Trapnell et al., 2010).

The association dataset of lincRNAs and diseases are extracted from the LncRNADisease database (Chen et al., 2012) in January 2015. There are 1564 lincRNAs and their associations with 1641 diseases. We employ the OMIM API function call (Hamosh et al., 2005) to retrieve closely matched phenotype IDs, resulting in a set of 684 OMIM phenotypes (mainly disease) associated with lincRNAs. All the diseases without matching any valid OMIM phenotype ID are removed. Finally, we obtain the lincRNA-disease association among 562 lincRNAs and 645 OMIM diseases.

The mRNA-disease associations are downloaded from DisGeNET software (Bauer-Mehren et al., 2010), where 16,666 mRNA genes are associated with 13,135 diseases. Similar to the lincRNA-disease associations, we use the OMIM function call to map disease names to matched phenotype IDs, and only these diseases with at least one lincRNAs are selected. Finally, 180,266 gene-disease associations are obtained among 645 OMIM diseases and 13,425 coding-genes.

The gene genetic interaction network is extracted from Lin et al. (2010), where 4,836,794 interactions among coding-genes. Only these genes associated with at least one disease are retained, resulting 3,264,923 interactions among 13,425 genes.

In this study, we want to make use the connections between lincRNAs and coding-genes. Based on Biswas et al. (2015), we construct the lincRNA-gene association network from diseases. Specifically, if the i-th lincRNA is connected to the j-th coding-gene if and only if both of them are associated with at least a disease. Based on this strategy, there are 1,775,375 edges among 562 lincRNAs and 13,425 coding-genes.

3.2. Settings

To fully validate the performance of the proposed algorithm, we select five well-known algorithms for a comparative comparison: NMF (Lee and Seung, 1999), non-smooth NMF (nsNMF) (Pascual-Marqui et al., 2001), integrated NMF (iNMF) (Biswas et al., 2015), Label Propagation (LP) (Hwang et al., 2011), and Random Walk (RW) (Li and Patra, 2010). All these algorithms can be categorized into two classes: matrix decomposition based and topological structure based methods. The matrix decomposition-based algorithms include NMF, nsNMF, and iNMF, while the topological structure-based methods are LP and RW.

To evaluate the performance of these algorithms, three measures, including mean absolute error (MAE), Accuracy and root mean squared error (RMSE), are employed to quantify the accuracy of algorithms. They are defined as Herlocker et al. (2004):

$$MAE(\widehat{X}, X) = \frac{1}{|\tau|} \sum_{(i,j)\in\tau} |\widehat{x}_{ij} - x_{ij}|, \tag{18}$$

$$Accuracy(\widehat{X}, X) = 1 - MAE(\widehat{X}, X), \tag{19}$$

$$RMSE(\widehat{X}, X) = \sqrt{\frac{1}{|\tau|} \sum_{(i,j)\in\tau} (|\widehat{x}_{ij} - x_{ij}|)^2}, \tag{20}$$

$$RSS(\widehat{X}, X) = \sqrt{\sum_{i,j} (|\widehat{x}_{ij} - x_{ij}|)^2}, \tag{21}$$

where $\widehat{X}$ and X are the observed association matrix and the predicted association matrix, respectively. τ is the set of lincRNA-disease association for prediction, i.e., τ is considered as the test set.

3.3. Parameter Selection

Three parameters are involved in the proposed algorithm, where parameter α determines the relevant importance of lincRNA co-expression networks, parameter β controls the relevant importance of the gene genetic network, and parameter k is the number of features for the basis and coefficient matrices. Similar to Ref., we set $\alpha = \beta$ by assuming that the lincRNA co-expression network and gene genetic network are equally important in discovering the lincRNA-disease associations.

We first investigate how parameter k determines the performance of the proposed algorithm. **Figure 3A** illustrates how RSS changes from 3 to 54 with a gap 3. From **Figure 3A**, we conclude that as k increases from 3 to 33, RSS dramatically decreases, which implies that the accuracy of the proposed algorithm increases. As k increases from 34 to 54, RSS increases. There is a good reason why this occurs. When k is small, the number of the latent features is insufficient to characterize the lincRNA-disease associations. When k is large, the number of the latent features is redundant. $k = 33$ reaches a good balance between them since RSS reaches the minimum. In the experiment, we set $k = 33$.

We then investigate how parameter α and β affect the performance of the Cr-NMF algorithm. **Figure 4** shows that how MAE and RMSE change as $\alpha \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. It is shown that the proposed algorithm achieves the best performance when $\alpha = 1$. Furthermore, the proposed algorithm is robust since the perturbation of performance is subtle if $\alpha \in [10, 100]$, indicating that Cr-NMF is not sensitive to parameter α and β. Even though MAE and RMSE decrease when $\alpha \in [10, 100]$, the change is subtle.

Finally, we check the convergence of the proposed algorithm. **Figure 3B** shows how RSS changes as the number of iterations increases. It is easy to assert that, when the number of iterations reaches 60, the algorithm converges because *RSS* does not change dramatically any more. Thus, the number of iterations is set as 60 in the experiments. The result demonstrates that the proposed algorithm is efficient.

3.4. Performance of Various Algorithms on Predicting lincRNA-Disease Associations

By setting $\alpha(\beta) = 10$, $k = 33$, and the number of iterations as 60, we apply Cr-NMF to the omic data to predict the lincRNA-disease associations. To quantify the performance of various algorithms, the accuracy in Equation (19) is adopted, where it is also used in Biswas et al. (2015). Because all of these compared algorithms have a factor of randomness, we get rid of randomness of algorithms by running each algorithm 50 times, and the mean of accuracy is used to quantify the performance of algorithms.

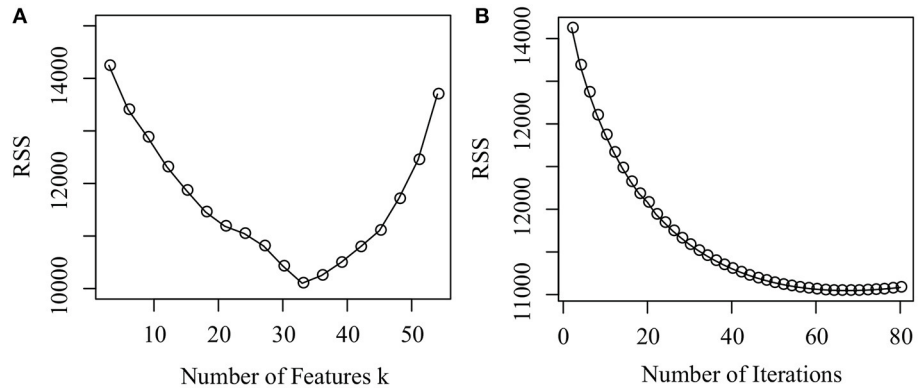

FIGURE 3 | Parameter selection and convergence analysis. **(A)** How the RSS changes as the number of features changes from 3 to 54, and **(B)** How the RSS changes as the number of iterations increases from 1 to 100.

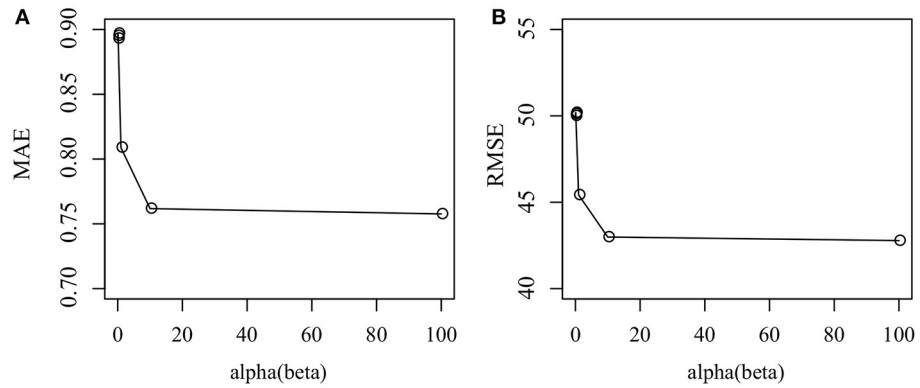

FIGURE 4 | How parameter affects the performance of the proposed algorithm in terms of various measurements: **(A)** MAE, and **(B)** RSME.

The leave-one-out cross validation (LOOCV) is adopted to measure the accuracy of each algorithm. Specifically, for each disease, we remove all the associations between the disease and lincRNA genes. The accuracy of various algorithms is depicted in **Figure 5A**. It is easy to draw conclusions such as: (1) the Cr-NMF algorithm achieves the best performance in LOOCV, followed by the iNMF algorithm. In detail, the accuracy of Cr-NMF is 0.823 ± 0.009, which is 1.9% higher than the iNMF algorithm on predicting disease-lincRNA associations. (2) Both Cr-NMF and iNMF algorithms outperform the rest of the methods, implying the integration of omic data is promising on predicting disease-lincRNA associations. Moreover, (3) The random walk and label propagation algorithms are worst in terms of accuracy. There are two reasons why the proposed algorithm outperforms the other approaches. First, the Cr-NMF algorithm directly factorize associations between diseases and lincRNAs, which captures the latent features to characterize the disease-lincRNA associations. Second, the factorization preserves the paths from "disease $\rightharpoonup$ lincRNA $\rightarrow$ protein-coding gene," which more precisely infers disease-lincRNA associations. The RW and LP algorithms are much worse than the others, implying that the topological structure is insufficient to characterize the relations between diseases and lincRNAs.

In order to further validate the performance of the proposed algorithm, we take the disease-lincRNA associations before 2015 January as training set, and set the data between 2015 and 2017 July as testing set, as shown in **Figure 5B**. It is easy to assert that the proposed algorithm is best, followed by iNMF. Specifically, the accuracy of algorithms is 0.647 (Cr-NMF), 0.594 (iNMF), 0.587 (nsNMF), 0.598 (sNMF), 0.575 (LP), 0.412 (RW). Careful comparison between **Figures 5A,B** indicates that the accuracy of various algorithms in the external validation decreases dramatically. However, the relative performance of these algorithms is similar. The results demonstrate that the proposed algorithm is promising in predicting disease-lincRNA associations.

4. CONCLUSION

LncRNAs are critical regulators in human diseases and disorder pathways. Thus, it is necessary to understand the associations

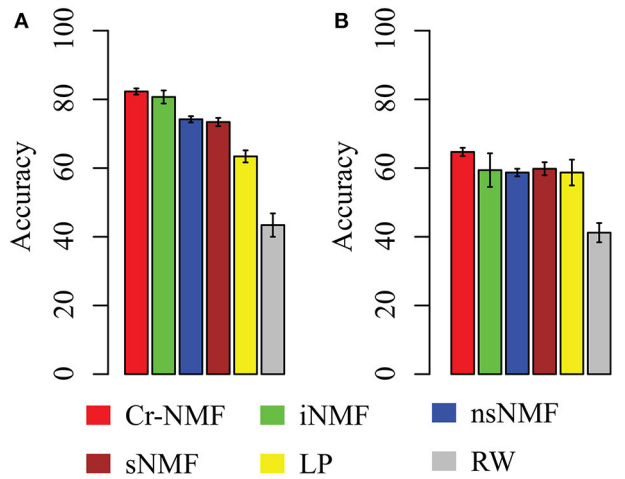

FIGURE 5 | The accuracy of various algorithms on predicting disease-lincRNA associations in terms of various strategies: **(A)** leave-one-out cross validation (LOOCV), and **(B)** external validation, where Y-axis denotes mean accuracy and error bar represents standard deviation.

between lncRNAs and diseases since these relations shed light on revealing the mechanisms of complex diseases. Compared to the protein-coding genes, a very little is known about the associations of lncRNAs and diseases. The next generation of sequencing technique discovers novel lncRNAs at an unprecedent speed. Therefore, there is a critical need to develop sophisticated computational tools to predict the relations between lncRNAs and diseases.

In this study, we proposed an NMF-based algorithm to predict lincRNA-disease associations by integrating multiple types of interaction data, such as co-expression interactions between lincRNAs, disease-lincRNA associations, disease-gene associations, gene genetic interactions, and lincRNA-gene links. There are two advantages of the proposed algorithm. First, it is able to explain each of the associated lincRNA as well as disease in a latent feature space. Second, the proposed algorithm takes the path from lincRNA to disease, i.e., "disease ⇀ lincRNA → protein-coding gene," which improves the accuracy of the prediction. The results demonstrate that the propose method outperforms state-of-the-art algorithms in terms of accuracy.

There are some limits in the proposed algorithm. First, there are two parameters involved in the methods and we solve this issue by a step search strategy in the experiments. A better and faster way to accomplish this needs to be developed. Particularly, how to infer the values of parameters by making use of the biological knowledge in diseases is ideal. Second, even though the proposed algorithm integrates omics data, incorporating additional data, such as disease networks, mutation data in genes would obtain even more meaningful results. In a future study, we will address these issues.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: TCGA.

AUTHOR CONTRIBUTIONS

YL and XM constructed the original idea and designed the experiments. XM wrote the manuscript. YL proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the NSFC (Grant No. 61562070), Scientific Research Projects of Colleges and Universities in Ningxia (NGY2018-136), Major Scientific Research Projects in Ningxia (2019BDE03015), and the Ningxia Science and Technology Leading Talent Project (201601).

REFERENCES

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2010). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Bioinformatics* 26, 2924–2926. doi: 10.1038/msb.2009.47

Biswas, A., King, M., Kim, D.-C., Ding, C. H. Q., Zhang, B., and Wu, X. (2015). Inferring disease associations of the long non-coding RNAs through non-negative matrix factorization. *Netw. Model. Anal. Health Inform. Bioinform.* 4:9. doi: 10.1007/s13721-015-0081-6

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. 25, 1915–1927. doi: 10.1101/gad.17446611

Cai, D., He, X., Han, J., and Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/TPAMI.2010.231

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNAdisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

Chen, J., and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 32, 1724–1732. doi: 10.1093/bioinformatics/btw059

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., et al. (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.* 41:e35. doi: 10.1093/nar/gks967

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D, et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature 07672

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi: 10.1038/nbt.1633

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of

human genes and genetic disorders. *Nucleic Acids Res.* 33(Suppl_1), D514–D517. doi: 10.1093/nar/gki033

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inform. Syst.* 22, 5–53. doi: 10.1145/963770.963772

Hwang, T., and Kuang, R. (2010). "A heterogeneous label propagation algorithm for disease gene discovery," in *Proceedings of the 2010 SIAM International Conference on Data Mining* (Siam, OH: SIAM), 583–594.

Hwang, T., Zhang, W., Xie, M., Liu, J., and Kuang, R. (2011). Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics* 27, 2692–2699. doi: 10.1093/bioinformatics/btr463

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., et al. (2017). Ldap: a web server for lncRNA-disease association prediction. *Bioinformatics* 33, 458–460. doi: 10.1093/bioinformatics/btw639

Lee, D., and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791.

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878. doi: 10.1093/nar/gkq1348

Lin, A., Wang, R. T., Ahn, S., Park, C. C., and Smith, D. J. (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.* 20, 1122–1132. doi: 10.1101/gr.104216.109

Lin, C. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* 19, 2756–2779. doi: 10.1162/neco.2007.19.10.2756

Ma, X., Dong, D., and Wang, Q. (2018a). Community detection in multi-layer networks using joint nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng.* 31, 273–286. doi: 10.1109/TKDE.2018.2832205

Ma, X., Sun, P., and Qin, G. (2017b). Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability. *Pattern Recogn.* 71, 361–374. doi: 10.1016/j.patcog.2017.06.025

Ma, X., Sun, P., and Wang, Y. (2018b). Graph regularized nonnegative matrix factorization for temporal link prediction in dynamic networks. *Phys. A Stat. Mech. Appl.* 496, 121–136. doi: 10.1016/j.physa.2017.12.092

Ma, X., Sun, P., and Zhang, Z. (2018c). An integrative framework for protein interaction network and methylation data to discover epigenetic modules. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 1855–1866. doi: 10.1109/TCBB.2018.2831666

Ma, X., Tang, W., Wang, P., Guo, X., and Gao, L. (2016). Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE ACM Trans. Comput. Biol. Bioinform.* 15, 647–658. doi: 10.1109/TCBB.2016.2625791

Ma, X., Yu, L., Wang, P., and Yang, X. (2017a). Discovering DNA methylation patterns for long non-coding rnas associated with cancer subtypes. *Comput. Biol. Chem.* 69, 164–170. doi: 10.1016/j.compbiolchem.2017.03.014

Mercer, R. T., Dinger, M. E., and Mattick, J. M. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

Ning, S., Yue, M., Wang, P., Liu, Y., Zhi, H., Zhang, Y., et al. (2016). Lincsnp 2.0: an updated database for linking disease-associated snps to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res.* 45, D74–D78. doi: 10.1093/nar/gkw945

Pascual-Marqui, R.D., Pascual-Montano, A.D., Kochi, K.,and Carazo, J.M. (2001). Smoothly distributed fuzzy c-means: a new self-organizing map. *Pattern Recogn.* 34, 2395–2402. doi: 10.1016/S0031-3203(00)00167-9

Petegrosso, R., Park, S., Hwang, T. H., and Kuang, R. (2017). Transfer learning across ontologies for phenome–genome association prediction. *Bioinformatics* 33, 529–536. doi: 10.1093/bioinformatics/btw649

Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E. T., and Antonarakis, S. E. (2013). Genetic and epigenetic regulation of human lincRNA gene expression. *Am. J. Hum. Genet.* 93, 1015–1026. doi: 10.1016/j.ajhg.2013.10.022

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. *J. Pathol.* 220, 126–139. doi: 10.1002/path.2638

Tang, W., Zhang, D. Z. and Ma, X. (2017). RNA-sequencing reveals genome-wide long non-coding RNAs profiling associated with early development of diabetic nephropathy. *Oncotarget* 8:105832. doi: 10.18632/oncotarget.22405

Trapnell, C., William, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6:e1000641. doi: 10.1371/journal.pcbi.1000641

Wang, K., and Chang, H. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909

Xi, J., Li, A., and Wang, M. (2020a). Hetrcna: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE ACM Trans. Comput. Biol. Bioinform.* 17, 422–434. doi: 10.1109/TCBB.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020b). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi: 10.1093/bioinformatics/btz793

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* 9:e87797. doi: 10.1371/journal.pone.0087797

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2017). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 396–406. doi: 10.1109/TCBB.2017.2701379

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., et al. (2016). Noncode 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* 44, D203–D208. doi: 10.1093/nar/gkv1252

Zhi, H., Li, X., Wang, P., Gao, Y., Gao, B., Zhou, D., et al. (2018). Lnc2meth: a manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease. *Nucleic Acids Res.* 46, D133–D138. doi: 10.1093/nar/gkx985

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.