frontiers in Genetics

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fgene.2021.712170

Predicting LncRNA–Disease Association by a Random Walk With Restart on Multiplex and Heterogeneous Networks

*Yuhua Yao[1,2,3], Binbin Ji[4], Yaping Lv[1], Ling Li[5], Ju Xiang[6,7,8], Bo Liao[1] and Wei Gao[9]**

[1] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2] Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, Haikou, China, [3] Key Laboratory of Computational Science and Application of Hainan Province, Hainan Normal University, Haikou, China, [4] Geneis Beijing Co., Ltd., Beijing, China, [5] Basic Courses Department, Zhejiang Shuren University, Hangzhou, China, [6] School of Computer Science and Engineering, Central South University, Changsha, China, [7] Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [8] Department of Computer Science, Changsha Medical University, Changsha, China, [9] Departments of Internal Medicine-Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fuzhou, China

Studies have found that long non-coding RNAs (lncRNAs) play important roles in many human biological processes, and it is critical to explore potential lncRNA–disease associations, especially cancer-associated lncRNAs. However, traditional biological experiments are costly and time-consuming, so it is of great significance to develop effective computational models. We developed a random walk algorithm with restart on multiplex and heterogeneous networks of lncRNAs and diseases to predict lncRNA–disease associations (MHRWRLDA). First, multiple disease similarity networks are constructed by using different approaches to calculate similarity scores between diseases, and multiple lncRNA similarity networks are also constructed by using different approaches to calculate similarity scores between lncRNAs. Then, a multiplex and heterogeneous network was constructed by integrating multiple disease similarity networks and multiple lncRNA similarity networks with the lncRNA–disease associations, and a random walk with restart on the multiplex and heterogeneous network was performed to predict lncRNA–disease associations. The results of Leave-One-Out cross-validation (LOOCV) showed that the value of Area under the curve (AUC) was 0.68736, which was improved compared with the classical algorithm in recent years. Finally, we confirmed a few novel predicted lncRNAs associated with specific diseases like colon cancer by literature mining. In summary, MHRWRLDA contributes to predict lncRNA–disease associations.

Keywords: lncRNA, disease, association, networks, random walk, predict

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology, China

Reviewed by:
Wei Peng,
Kunming University of Science and Technology, China
Kebo Lv,
Ocean University of China, China

****Correspondence:***
Wei Gao
13960986882@163.com

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

Received: *20 May 2021*
Accepted: *23 July 2021*
Published: *19 August 2021*

Citation:
Yao Y, Ji B, Lv Y, Li L, Xiang J, Liao B and Gao W (2021) Predicting LncRNA–Disease Association by a Random Walk With Restart on Multiplex and Heterogeneous Networks. Front. Genet. 12:712170. doi: 10.3389/fgene.2021.712170

INTRODUCTION

Numerous studies have indicated that protein-coding genes accounted for less than 2% of the human genome (Crick et al., 1961; Yanofsky, 2007). There are many non-translatable RNAs called non-coding RNAs (ncRNAs), which have been considered as transcriptional noise for a long time (Zhang et al., 2017; Xu et al., 2020). Long non-coding RNAs (lncRNAs) whose length are greater than 200 nucleotides are a class of important ncRNAs (Mercer et al., 2009). There are increasing evidence that lncRNAs play key roles in many important biological processes and

diseases (Akerman et al., 2017; Wang et al., 2019; Peng et al., 2020). For example, HOTAIR was considered as a potential biomarker for liver cancer (Yang et al., 2011; Li et al., 2019), lung cancer (Li G. et al., 2014a), and colorectal cancer (Kogo et al., 2011; Maass et al., 2014), and UCA1 was a potential biomarker for bladder cancer diagnosis (Zhang et al., 2012). Li J. et al. (2014b) summarized the important role of lncRNA such as MALAT1, HOTAIR, and other specific lncRNAs for hepatocellular carcinoma. LncRNAs associated with tumor immune invasion in non-small cell lung cancer (NSCLC) have important value in improving clinical efficacy and immunotherapy, compared with normal controls, and the expression of gabpb1-it1 was significantly downregulated in NSCLC. In addition, overexpression of gabpb1-it1 in cancer samples is associated with increased survival in NSCLC patients (Sun et al., 2020). Inferring the association between lncRNA and diseases can better study human diseases and help the diagnosis and treatment of diseases, and accelerate the identification of potential drug response predictors (Liu et al., 2016, 2020). Therefore, the exploration of lncRNA–disease association has attracted more and more attention from biologists. The establishment of an effective computational model to predict the association between lncRNAs and diseases can save time and money spent in biological experiments (Yao et al., 2019; Yan et al., 2020).

At present, many machine learning methods have been proposed to predict the lncRNA–disease association, for example, Laplacian regulated least square method (LRLSLDA; Chen and Yan, 2013), propagation algorithm (Yang et al., 2014), a method based on Bayesian classifier (Zhao et al., 2015), and a method based on induction matrix (Lu C. et al., 2018a). However, these machine learning methods need negative samples, which are difficult to obtain. In order to solve this problem, network-based methods emerge as the times require. With the increasing importance of revealing the molecular basis of human diseases, network-based methods have been widely used in exploring disease-related genes (Yan et al., 2015; Hu et al., 2018; Lu M. et al., 2018b; Yang et al., 2020). For example, Xiang et al. (2021) proposed a multibiological network (NIDM) network pulse dynamics framework and a fast network embedding (Xiang et al., 2020) to predict disease-related genes. Network-based algorithms have also been widely studied in predicting lncRNA–disease association. Bellucci et al. (2011) combined the expression similarity of lncRNA with the Gaussian nuclear interaction spectrum similarity of lncRNA, and proposed a potential protein determination method based on sequence information to predict the function of lncRNA. In the study of Xiao et al. (2015), the function of lncRNA was predicted by constructing the regulatory network between lncRNA and protein coding genes. In the BPLLDA study, the authors estimated the potential relationship between disease and lncRNAs by connecting the length of the disease and lncRNA pathway (Xiao et al., 2018). KATZLDA was a computing method to predict lncRNA–disease association based on the similarity between heterogeneous network nodes (Chen, 2015a). The random walk model is also widely used in the field of data mining and Internet, and many researches use this method to predict potential association (Xing et al., 2012; Yang et al., 2016, 2017; Gu et al., 2017). Zhou et al. (2015) proposed a new method by integrating the related lncRNA–lncRNA network, disease–disease similarity network, and the heterogeneous lncRNA–disease association network, and then realized random walk on the heterogeneous network. Sun et al. (2014) proposed a method for constructing lncRNA–lncRNA functional similarity network and then developed a calculation method based on global network (RWRlncD). Recently, Lei and Bian (2020) used random walk to weight the structural features of circRNA–disease pairs and combined it with k-nearest neighbor algorithm to get the prediction score of each circRNA–disease pair. Although these methods have been proposed to predict lncRNA–disease association successfully, it is still a challenge to make full use of multi-source biological data.

In this study, a random walk algorithm with restart on multiplex and heterogeneous networks was developed. The downloaded known lncRNA–disease association data were used to calculate lncRNA functional similarity, lncRNA Gaussian interaction kernel similarity, disease semantic similarity, and disease Gaussian interaction kernel similarity, respectively. Then, these similarity networks and lncRNA–disease association network were constructed into multiplex and heterogeneous networks. A random walk with restart was carried out on the multiplex and heterogeneous networks, and the potential lncRNA–disease association was predicted using the final stable probability.

MATERIALS AND METHODS

LncRNA–Disease Association

LncRNADisease (Chen, 2015b), Lnc2Cancer (Ning et al., 2016), MNDR (Wang et al., 2013), and other databases stored the known lncRNA–disease association data, which have been of great help in predicting novel association. In this study, 285 lncRNA–disease association was downloaded from lncRNADisease database, including 117 lncRNAs and 159 diseases. We used LD to represent the lncRNA–disease association adjacency matrix. If lncRNA(i) is related to disease(j), then $LD(i, j) = 1$; otherwise, $LD(i, j) = 0$, that is:

$$LD(i,j) = \begin{cases} 1, & if\ lncRNA\ (i)\ is\ associated\ with\ disease\ (j) \\ 0 & otherwise \end{cases} \quad (1)$$

Disease Similarity

Disease Semantic Similarity

Directed acyclic graphs ($DAGs$) were used to calculate disease–disease similarity, for disease d_k, let $DAG(d_k, T(d_k), E(d_k))$ be its directed acyclic graph, where $T(d_k)$ are ancestor nodes of d_k, and $E(d_k)$ represents the corresponding set of edges from parent node to child nodes. Semantic similarity of diseases was calculated by R package called DOSim (Li et al., 2011); for any disease k in $DAG(d_k, T(d_k), E(d_k))$, the semantic contribution of k to d_k was

defined as:

$$D_{d_k}(k) = \begin{cases} 1, & if\ k = d_k \\ max\{0.5 * D_{d_k}(k') | k' \in children\ of\ k\}, & if\ k \neq d_k \end{cases} \quad (2)$$

The above formula indicates that the contribution of the disease to its semantic value is 1. Semantic contribution decreased with the increase of the distance between disease k and other diseases. Then, the semantic similarity between d_i and d_j was defined as:

$$DSS(d_i, d_j) = \frac{\sum_{k \in T_{d_i} \cap T_{d_j}} (D_{d_i}(k) + D_{d_j}(k))}{\sum_{k \in T_{d_i}} D_{d_i}(k) + \sum_{k \in T_{d_j}} D_{d_j}(k)} \quad (3)$$

Gaussian Interaction Profile Kernel Similarity for Diseases

In order to obtain the similarity information between diseases, the Gaussian Interaction Profile kernel similarity between disease was constructed based on the lncRNA–disease association network. First, the Interaction Profile (IP) of each disease represents a binary code in the known lncRNA–disease association network. For example, for given disease d_i, its $IP(d_i)$ represents the ith column of LD. Next, the Gaussian Interaction Profile kernel similarity between d_i and $d_j$was calculated as:

$$DS_{GIP}(d_i, d_j) = exp(-\gamma_d \left|\left|IP(d_i) - IP(d_j)\right|\right|^2 \quad (4)$$

Where γ_d represents the bandwidth that controls the Gaussian Interaction Profile kernel similarity, $\gamma_d = \frac{\gamma'_d}{(\frac{1}{nd}\sum_{i=1}^{nd} \left|\left|IP(d_i)\right|\right|^2)}$; in this study, according to van Laarhoven et al. (2011), we set $\gamma'_d = 1$, and nd represents the number of diseases.

LncRNA Similarity

LncRNA Functional Similarity

Studies have shown that similar lncRNAs are usually associated with similar diseases. Therefore, lncRNA functional similarity can be roughly estimated by their similarity in related diseases (Sun et al., 2014). For any two lncRNAs l_i and l_j, $D_i = \{d_{i_k} | 1 \leq k \leq m\}$ and $D_j = \{d_{j_l} | 1 \leq l \leq n\}$ were disease sets associated with l_i and l_j, respectively. The semantic similarity between disease d and disease set D was firstly defined as:

$$SS(d, D) = \max_{d_l \in D} DSS(d, d_l) \quad (5)$$

Then, the functional similarity between l_i and l_j was defined as:

$$NFS(l_i, l_j) = \frac{\sum_{i=1}^{m} SS(d_{i_a}, D_j) + \sum_{j=1}^{n} SS(d_{j_b}, D_i)}{m + n} \quad (6)$$

Gaussian Interaction Profile Kernel Similarity for LncRNAs

Similar to the disease Gaussian interaction profile kernel similarity. The formula for calculating the Gaussian interaction profile kernel similarity between l_i and l_j was:

$$LS_{GIP}(l_i, l_j) = exp(-\gamma_l \left|\left|IP(l_i) - IP(l_j)\right|\right|^2) \quad (7)$$

Where γ_l represents the bandwidth that controls the property similarity of Gaussian interaction kernel, $\gamma_l = \frac{\gamma'_l}{(\frac{1}{nl}\sum_{i=1}^{nl} \left|\left|IP(l_i)\right|\right|^2)}$; in this study, $\gamma'_l = 1$, nl represents the number of lncRNAs, $IP(l_i)$ and $IP(l_j)$ represent the ith and jth row of the LD, respectively.

A Random Walk With Restart on Multiplex and Heterogeneous Networks

An overview of MHRWRLDA is shown in **Figure 1**. Specifically, we first downloaded the data of known lncRNA–disease association from the LncRNADisease database and got diseased' DO ID from the DO database to calculate disease similarity. After compute disease similarity and lncRNA similarity, a multiplex and heterogeneous network was set up based on these similarity networks and known lncRNA–disease association network. Finally, a random walk algorithm with restart was implemented on networks, and the final stability probability was used to conduct the predictions.

Multiplex and Heterogeneous Network

Based on disease semantic similarity network, disease Gaussian similarity network, lncRNA similarity network, and lncRNA Gaussian similarity network, we constructed a multiplex and heterogeneous network by using lncRNA–disease association. In these networks, the set of lncRNA nodes was defined as: $R_M = \{v_i^\alpha, i = 1, 2, \cdots, n; \alpha = 1, 2, \cdots, L\}$, where v_i^α represents the ith node on the α layer. The set of disease nodes was defined as: $D_M = \{v_j^\beta, j = 1, 2, \cdots, m; \beta = 1, 2, \cdots, K\}$, where v_i^β represents the jth node on the β layer. The adjacency matrix on each layer is:

$$A^{[\alpha]} = A^{[\alpha]}(i, j) = \begin{cases} 1, & if\ the\ i^{th}\ node\ is\ associated\ with \\ & the\ jth\ node\ on\ layer\ \alpha \\ 0, & otherwise \end{cases} \quad (8)$$

A particle can either travel from the previous node v_i^α term to any neighbor node on the same layer, or it can also jump to the same node on a different layer. The matrix A contains different types of jumps that the particle can follow at each step:

$$A = \begin{pmatrix} (1-\delta)A^{[1]} & \frac{\delta}{(L-1)}I & \cdots & \frac{\delta}{(L-1)}I \\ \frac{\delta}{(L-1)}I & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}A \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}I & \frac{\delta}{(L-1)}I & \cdots & (1-\delta)A^{[L]} \end{pmatrix} \quad (9)$$

Where I is the $n \times n$ identity matrix, the diagonal element of A represents the particle walking on same layer, the off-diagonal element represents the particle jumping between different layers, and the parameter $\delta \in (0, 1)$ represents the probability of the particle walking on the same layer or jumping between different layers. If $\delta = 0$, the particles will always walk on the same layer.

$A_{RM(nL \times nL)}$, $A_{DM(mK \times mK)}$ is the matrix of lncRNA similarity and disease similarity on multiplex and heterogeneous networks, respectively. n, L, m, and K are the number of

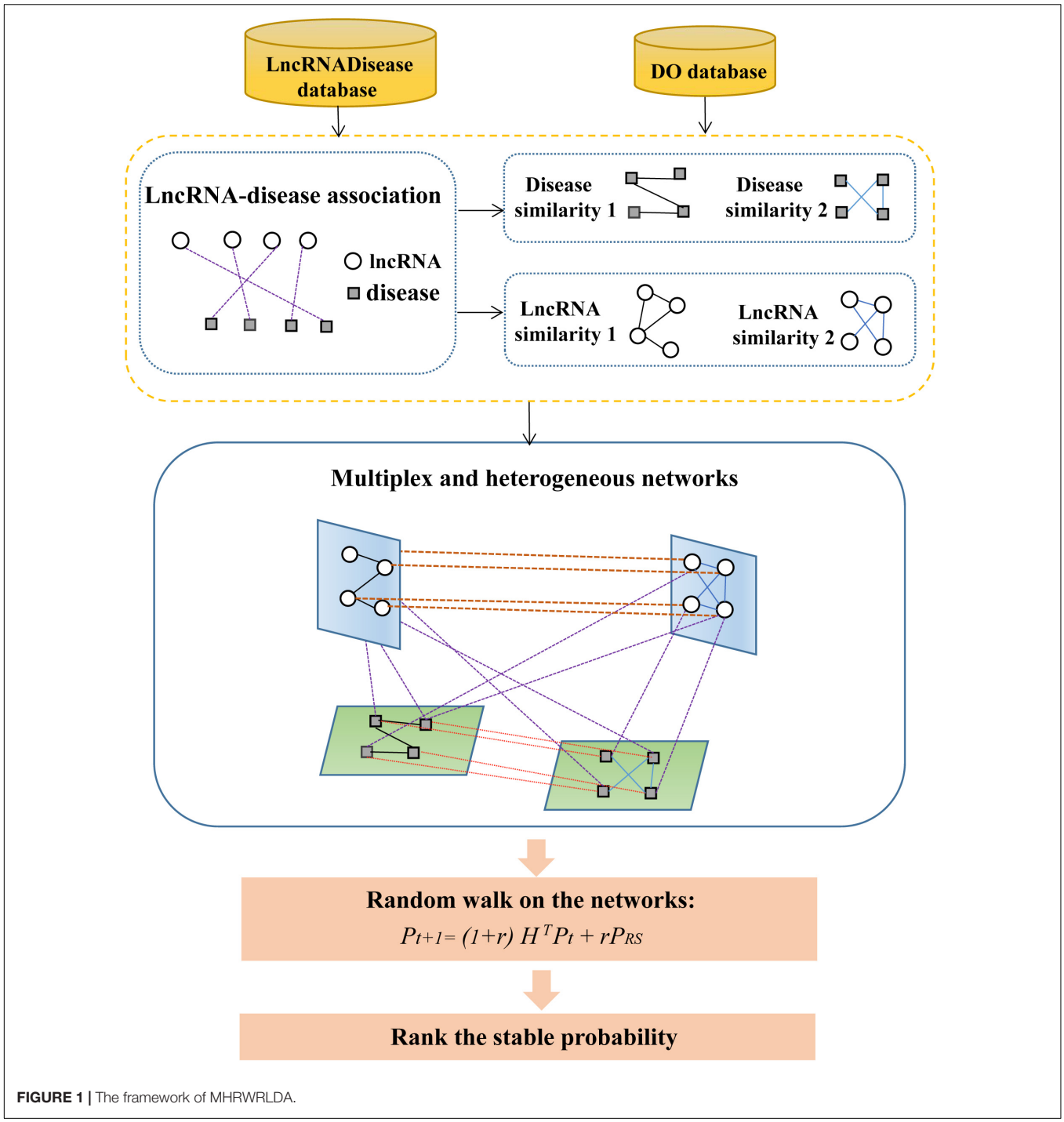

FIGURE 1 | The framework of MHRWRLDA.

lncRNAs, lncRNA similarity networks, diseases, and disease similarity networks, respectively, the adjacency matrix is: $B_{MH} = (B_{n \times m}, B_{n \times m}, \cdots, B_{n \times m})^T$.

The dimension of B_{MH} is $nL \times mK$, which is equivalent to replicating the adjacency matrix $B_{n \times m}$ L^*K times, where $B = LD$. Then, the adjacency matrix of the whole multiplex and heterogeneous networks is: $A = \begin{bmatrix} A_{RM} & B_{MH} \\ B_{MH}^T & A_{DM} \end{bmatrix}$.

Random Walk With Restart on Multiplex and Heterogeneous Networks

A random walk with a restart means that a particle starts at a node and it is faced with two choices at each walk: move to a randomly selected neighbor node, or jump back to the start node. Considering the time is discrete, $t \in \mathbb{N}$, the particle is at node v_t at the tth step. Then, it walks from v_t to v_{t1}. We defined a restart probability $\gamma \in (0, 1)$, and the random walk with restart can be

TABLE 1 | Confusion matrix definitions.

True prediction	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

defined as:

$$P_{t+1} = (1-\gamma)H^T P_t + \gamma P_{RS} \tag{10}$$

Where the vectors P_{t1} and P_t represent the probability distribution of v_t and v_{t1}, respectively. P_{RS} is the initial probability distribution and $P_{RS} = \begin{bmatrix} (1-\eta)R_0 \\ \eta D_0 \end{bmatrix}$; the importance of each network is adjusted by adjusting P_{RS}, where R_0 and D_0 represent the initial probability distribution of lncRNA similarity network and disease similarity network, respectively, and the dimensions of the vectors P_{t+1}, P_t, and P_{RS} are $nL \times mK$. The parameter $\eta \in (0, 1)$ controls the probability of each network restarting; if $\eta < 0.5$, the particle is more likely to be restarted in lncRNA similarity networks. $H = \begin{bmatrix} H_{RR} & H_{RD} \\ H_{DR} & H_{DD} \end{bmatrix}$ represents the transition probability matrix of multiplex and heterogeneous networks, where H_{RR} andH_{DD} represent the transition probability of nodes upstream in the same layer, H_{RD} and H_{DR} represent the transition probability of node jump between different layers. For a given node, if dichotomous correlation exists, the particle can jump between layers or stay in the current layer with probability $\lambda \in (0, 1)$, and the closer it is to 1, the higher the probability of jumping between different networks.

We suppose a particle was located at the node $r_i \in R$. In the next step, the particle can walk to the node $r_j \in R$. The transfer probability is:

$$H_{RR} = \begin{cases} \frac{A_R(i,j)}{\sum_{k=1}^{n} A_R(i,k),} & \text{if } \sum_{k=1}^{m} B(i,k) = 0 \\ (1-\lambda)\frac{A_R(i,j)}{\sum_{k=1}^{n} A_R(i,k),} & \text{otherwise} \end{cases} \tag{11}$$

It can also jump to the node $d_b \in D$ through binary correlation, and the transfer probability is:

$$H_{RD} = \begin{cases} \frac{\lambda B(i,b)}{\sum_{k=1}^{m} B(i,k),} & \text{if } \sum_{k=1}^{m} B(i,k) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Similarly, if the particle was located at the node $d_a \in D$, then the transfer probability of the particle walking to the node $d_b \in D$ is:

$$H_{DD} = \begin{cases} \frac{A_D(a,b)}{\sum_{k=1}^{m} A_D(a,k),} & \text{if } \sum_{k=1}^{n} B(k,b) = 0 \\ (1-\lambda)\frac{A_D(a,b)}{\sum_{k=1}^{m} A_D(a,k),} & \text{otherwise} \end{cases} \tag{13}$$

If the particle jumps to the node $r_j \in R$ through binary correlation, then the transfer probability is:

$$H_{DR} = \begin{cases} \frac{\lambda B(j,a)}{\sum_{k=1}^{n} B(k,a),} & \text{if } \sum_{k=1}^{n} B(k,a) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

When predicting lncRNAs that are potentially associated with the given disease d_i, the node d_i will be used as the seed node in disease similarity networks. The initial probability D_0 is 1 for the given node d_i and 0 for the remaining nodes. If there are known associations among lncRNAs $r_1, r_2 \cdots$ and disease d_i, then the nodes $r_1, r_2 \cdots$ are the seed nodes in lncRNA similarity networks. The initial probability R_0 was assigned to seed node $r_1, r_2 \cdots$, with a probability of 1, and the remaining nodes were 0. P_t converges after some iteration, that is, $P_t - P_{t+1} < 10^{-10}$, and we denoted the stable probability as: $P_\infty = \begin{bmatrix} (1-\eta)R_\infty \\ \eta D_\infty \end{bmatrix}$.

Based on the stabilized R_∞, those seed nodes $r_1, r_2 \cdots$ were removed, and the remaining lncRNAs were ranked. The higher the ranked lncRNA, the more likely it was to be associated with the given disease d_i. Similarly, a lncRNA can also be designated to predict diseases related to it.

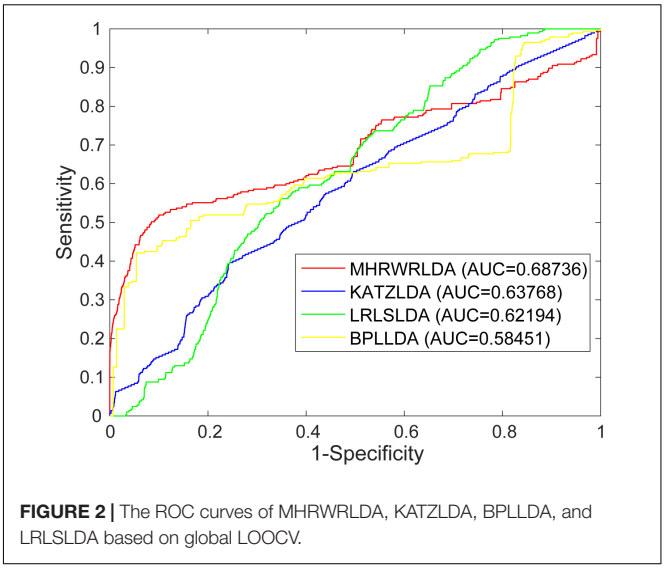

FIGURE 2 | The ROC curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV.

RESULTS

Indicators of Performance Evaluation

For a binary classification problem, the confusion matrix is shown in **Table 1**. Precision, specificity, and sensitivity are evaluation indicators of classification models. They are calculated as:

$$FPR = 1-\text{specificity} = \frac{FP}{TN+FP}$$

$$TPR = \text{sensitivity} = \frac{TP}{TP+FN}$$

To evaluate the performance of MHRWRLDA, the receiver operating characteristic (ROC) curve was drawn by calculating TPR and FPR according to different thresholds. Area under the curve (AUC) is the area under the ROC curve, and this area is less than 1. Since the ROC curve cannot directly indicate which classifier has better effect in many cases, as a value, the larger the AUC is, the better the classifier has an effect.

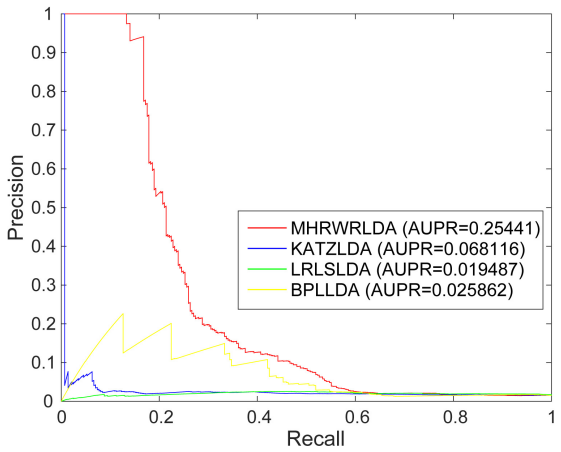

FIGURE 3 | The PR curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV.

TABLE 2 | The predicted top 10 lncRNAs for colon cancer.

Disease	Rank	LncRNA	Evidence
Colon cancer	1	H19	Confirmed
	2	MEG3	Confirmed
	3	CDKN2B-AS1	Confirmed
	4	MALAT1	Confirmed
	5	PVT1	Confirmed
	6	BCYRN1	Unknown
	7	IGF2-AS	Confirmed
	8	Anti-NOS2A	Unknown
	9	WT1-AS	Unknown
	10	UCA1	Confirmed

TABLE 3 | The predicted top 10 lncRNAs for hepatocellular carcinoma.

Disease	Rank	LncRNA	Evidence
Hepatocellular carcinoma	1	H19	Confirmed
	2	MEG3	Confirmed
	3	MALAT1	Confirmed
	4	AIR	Confirmed
	5	HULC	Confirmed
	6	HOTAIR	Confirmed
	7	IGF2-AS	Confirmed
	8	CDKN2B-AS1	Confirmed
	9	PVT1	Confirmed
	10	BCYRN1	Unknown

Performance of MHRWRLDA

In order to evaluate the performance of MHRWRLDA for predicting lncRNA–disease association, we applied the known lncRNA–disease association data to MHRWRLDA, and used Leave-One-Out cross-validation (LOOCV) to verify. For global LOOCV, the scores of all test samples are compared with those of all candidate samples. For local LOOCV, each known lncRNA related to a particular disease is selected as the test sample, and other related lncRNAs are selected as the training samples; the scores of test samples are only compared with those of candidate samples. In this study, there are a total of three parameters, namely, γ, λ, and η, and their range is (0, 1), where γ is the restart probability; λ is the jump probability, reflecting the probability of particles jumping between different networks; and η regulated the probability of each network restarting, When $\eta = \gamma = 0.9$ and $\lambda = 0.9$, the prediction effect is the best; at this point, AUC = 0.68736.

The AUC based on global LOOCV of the KATZLDA (Chen, 2015a), BPLLDA (Xiao et al., 2018), and LRLSLDA (Chen and Yan, 2013) were 0.63768, 0.5845, and 0.6219, respectively. The ROC curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV are shown in **Figure 2**, the PR curves based on global LOOCV are shown in **Figure 3**, and the AUPR values are shown in their legends. Their ROC curves and PR curves based on local LOOCV are shown in **Supplementary Figures 1**, **2**. The results showed that MHRWRLDA performed better than other classical algorithms in predicting lncRNA–disease association.

TABLE 4 | The predicted top 10 lncRNAs for breast cancer.

Disease	Rank	LncRNA	Evidence
Breast cancer	1	H19	Confirmed
	2	CDKN2B-AS1	Confirmed
	3	PVT1	Confirmed
	4	MEG3	Confirmed
	5	BCYRN1	Confirmed
	6	SRA1	Confirmed
	7	XIST	Confirmed
	8	GAS5	Confirmed
	9	HOTAIR	Confirmed
	10	DSCAM-AS1	Confirmed

TABLE 5 | The predicted top five novel disease correlated with MALAT1, PVT1, and MEG3.

LncRNA	Disease	Rank	Evidence
MALAT1	Endometrial stromal sarcoma	1	Confirmed
	Non-small cell lung cancer	2	Confirmed
	Lung adenocarcinoma	3	Confirmed
	Cervical cancer	4	Confirmed
	Osteosarcoma	5	Confirmed
PVT1	Burkitt's lymphomas	1	Confirmed
	Hodgkin's lymphoma	2	Confirmed
	Renal disease	3	Confirmed
	Diabetic nephropathy	4	Confirmed
	Pancreatic cancer	5	Confirmed
MEG3	Pituitary adenomas	1	Confirmed
	Heroin addiction	2	Confirmed
	Nonfunctioning pituitary adenomas	3	Confirmed
	Chronic myeloid leukemia	4	Confirmed
	Myelodysplastic syndrome	5	Confirmed

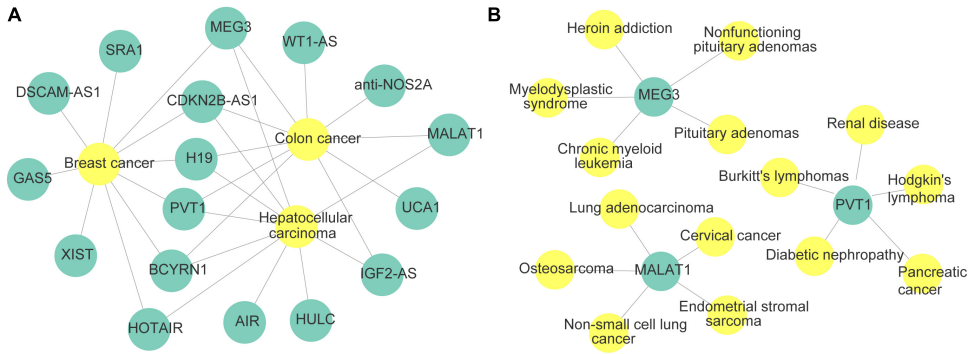

FIGURE 4 | The network of diseases and lncRNAs were made by Cytoscape. **(A)** The network of novel lncRNAs related to colon cancer, hepatocellular carcinoma, and breast cancer. **(B)** The network of novel diseases related to lncRNA MALAT1, PVT1, and MEG3.

Case Study

To further explore the performance of MHRWRLDA in predicting lncRNA–disease association, we selected colon cancer, hepatocellular carcinoma, and breast cancer for the case study. During the experiment, all known associations were considered as the train set, and unknown associations were regarded as the test set. According to LOOCV results, we sorted lncRNAs and selected the top 10 lncRNAs for further verification based on the LncRNADisease database and several recently published studies.

Colon cancer is a malignant tumor, causing nearly 700,000 deaths each year, and has a high incidence rate record in developed countries. We applied MHRWRLDA to colon cancer experiments to predict the top 10 lncRNAs related to colon cancer (**Table 2**). Seven of the top 10 lncRNAs have been confirmed in databases or other literature. Previous studies have found that the third ranked CDKN2B-AS1 up-regulates HCT116, thereby causing cell proliferation (Chiyomaru et al., 2013). In addition, studies have shown that removal of PVT1 (ranked 5) from MCY-driven colon cancer strain HCT116 can reduce carcinogenicity (Tseng et al., 2014).

Hepatocellular carcinoma is one of the most common cancers in the world. Studies have shown that hepatocellular carcinoma is the main component of primary liver cancer. We listed the top 10 lncRNAs related to hepatocellular carcinoma predicted by experiments in **Table 3**. Of the top 10, 9 were all verified in known databases. The overexpression of CDKN2B-AS1, which ranked 8, can inhibit the proliferation and invasion of liver cancer cells (Hua et al., 2015), thereby promoting the apoptosis of liver cancer cells and preventing the occurrence of hepatocellular carcinoma. Ding et al. identified PVT1 (ranked 9) as a novel biomarker for predicting tumor recurrence in patients with hepatocellular carcinoma (Ding et al., 2015).

Breast cancer accounts for 22% of all cancers in women and is the second leading cause of cancer death in women (Donahue and Genetos, 2013; Karagoz et al., 2015). Traditionally, breast cancer has been diagnosed on the basis of histopathological features such as tumor size, grade, and lymph node status. The prediction of breast cancer-related lncRNAs may help diagnose and treat breast cancer (Meng et al., 2014). In order to diagnose and treat breast cancer better, it is necessary to predict lncRNAs associated with breast cancer and identify lncRNA biomarkers (Xu et al., 2015). We implemented MHRWRLDA on breast cancer to predict potentially relevant lncRNAs, and listed the top 10 lncRNAs related to breast cancer in **Table 4**. The downregulation of the top ranked first H19 significantly reduced breast cancer clonal formation and anchored independent growth (Barsytelovejoy et al., 2006). In addition, the incidence of breast cancer is also affected by PVT1 overexpression due to genomic abnormalities (Guan et al., 2007).

Finally, the network of three cases and lncRNAs predicted by MHRWRLDA is shown in **Figure 4A**; it revealed that MEG3, CDKN2B-AS1, H19, PVT1, BCYRN1, HOTAIR, and all three diseases are related. In addition to exploring lncRNAs related to novel diseases, it is also extremely important to predict diseases related to novel lncRNAs. Therefore, taking lncRNA MALAT1, PVT1, and MEG3 as examples, the predicted top five diseases related to them are listed in **Table 5**, and their network is shown in **Figure 4B**. The experimental results proved that MHRWRLDA was useful for predicting the potential lncRNA–disease association.

DISCUSSION

In recent years, the research on the interaction between biomolecules has been growing. Due to the importance of lncRNA, the research on the associations between lncRNAs and diseases has been paid more and more attention. These associations can be characterized by complex networks, so it is urgent to develop network-based computational algorithms to explore functional associations between lncRNAs and diseases. The algorithm of constructing heterogeneous network and implementing random walk on heterogeneous network is widely used in the field of bioinformatics. However, in previous studies, most of them are single heterogeneous networks with a single information source. Therefore, we consider multiple network embedding by integrating different types of edges. Multiplex and heterogeneous networks are the combination of heterogeneous networks connected by multiple interactions; they integrate the framework of multiple information sources, and each layer is

a simplex network with specific types of nodes and edges; when the data set is large, they can produce better results. Multiple heterostructures may provide a richer perspective for the study of the complex relationship between different biological components.

In this study, we extend it to multi-layer heterogeneous networks so as to more effectively predict lncRNA–disease associations. We constitute a multiplex and heterogeneous network by integrating known lncRNA–disease association, lncRNA function similarity, lncRNA Gaussian similarity network, disease semantic similarity network, and disease Gaussian similarity network, and then we generate the final comprehensive predictive scores by the random walk with restart on the multiplex and heterogeneous network, so as to forecast potential lncRNA–disease associations. LOOCV experimental verification results showed that the AUC was 0.68736, which exceeded other algorithms to predict lncRNA–disease association. In novel diseases, the top 10 lncRNAs were verified and predicted by database or literature. In addition, the model can also predict diseases associated with particular lncRNAs.

The network-based approach overcomes the disadvantage of machine learning methods that need to construct negative samples and not only is suitable for predicting lncRNA–disease associations, but also proved to be widely used in exploring disease-related miRNAs, drug repositioning, and prediction of disease–gene associations. Therefore, if the known lncRNA–disease association data are replaced with miRNA-disease association data, MHRWRLDA can be used to predict the potential miRNAs associated with disease; similarly, if it is replaced by drug–disease association data or gene–disease association data, it is possible to make contributions to drug repositioning and the exploration of disease-related genes, respectively. In the future, we will try to apply MHRWRLDA to the above aspects for research.

However, there are some limitations. First, there are only two methods for constructing the similarity network; if the calculation method of the similarity network can be increased, the number of layers in the multi-layer heterogeneous graph can be increased to provide more possibilities for particle migration. Second, the lncRNA–disease association data contain only 117 lncRNAs and 159 diseases, of which there are only 285 pairs of correlations; a small data set may also affect the prediction results. In the future, more association data will be discovered and used to overcome the difficulties caused by the complexity and inconsistency of biological data. In addition, efforts will be made to combine multiple prediction models to achieve more accurate predictions.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/jibinbin171222/MHRWRLDA.

AUTHOR CONTRIBUTIONS

WG conceived, designed, and managed the study. YY and BJ designed the method and wrote the original manuscript. YL and LL revised the original draft. JX wrote the code. BL discussed the proposed method and gave further research. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 61762035), the Hainan Provincial Natural Science Foundation of China (Grant No. 119MS037), the National Natural Science Foundation of China (Grant No. 61702054), the Training Program for Excellent Young Innovators of Changsha (Grant Nos. kq2009093 and kq1905045), and the Joint Funds for the Innovation of Science and Technology, Fujian province (Grant No. 2019Y9038).

ACKNOWLEDGMENTS

The data used to support the findings of this study are available from the corresponding author upon request.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.712170/full#supplementary-material

REFERENCES

Akerman, I., Tu, Z., Beucher, A., Rolando, D. M. Y., Sauty-Colace, C., Benazra, M., et al. (2017). Human pancreatic β Cell lncRNAs control cell-specific regulatory networks. *Cell Metab.* 25, 400–411. doi: 10.1016/j.cmet.2016.11.016

Barsytelovejoy, D., Lau, S. K., Boutros, P. C., Khosravi, F., Jurisica, I., Andrulis, I. L., et al. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* 66, 5330–5337. doi: 10.1158/0008-5472.can-06-0037

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods* 8, 444–445. doi: 10.1038/nmeth.1611

Chen, X. (2015a). KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5:16840.

Chen, X. (2015b). Predicting lncRNA–disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* 5:13186.

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chiyomaru, T., Yamamura, S., Fukuhara, S., Yoshino, H., Kinoshita, T., Majid, S., et al. (2013). Genistein inhibits prostate cancer cell growth by targeting miR-34a and oncogenic HOTAIR. *PLoS One* 8:e70372. doi: 10.1371/journal.pone.0070372

Crick, F., Barnett, L., Brenner, S., and Wattstobin, R. J. (1961). General nature of the genetic code for proteins. *Nature* 192, 1227–1232. doi: 10.1038/1921227a0

Ding, C., Yang, Z., Lv, Z., Du, C., Xiao, H., Peng, C., et al. (2015). Long non-coding RNA PVT1 is associated with tumor progression and predicts recurrence in hepatocellular carcinoma patients. *Oncol. Lett.* 9, 955–963. doi: 10.3892/ol.2014.2730

Donahue, H. J., and Genetos, D. C. (2013). Genomic approaches in breast cancer research. *Brief. Funct. Genomics* 12, 391–396.

Gu, C., Liao, B., Li, X., Cai, L., Li, Z., Li, K., et al. (2017). Global network random walk for predicting potential human lncRNA-disease associations. *Sci. Rep.* 7:12442.

Guan, Y., Kuo, W. L., Stilwell, J. L., Takano, H., Lapuk, A., Fridlyand, J., et al. (2007). Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* 13, 5745–5755. doi: 10.1158/1078-0432.ccr-06-2882

Hu, K., Hu, J. B., Tang, L., Xiang, J., Ma, J. L., Gao, Y. Y., et al. (2018). Predicting disease-related genes by path structure and community structure in protein–protein networks. *J. Statal Mech. Theory Exp.* 2018:100001. doi: 10.1088/1742-5468/aae02b

Hua, L., Wang, C. Y., Yao, K. H., Chen, J. T., Zhang, J. J., and Ma, W. L. (2015). High expression of long non-coding RNA ANRIL is associated with poor prognosis in hepatocellular carcinoma. *Int. J. Clin. Exp. Pathol.* 8, 3076–3082.

Karagoz, K., Sinha, R., and Arga, K. Y. (2015). Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *OMICS J. Integr. Biol.* 19, 115–130. doi: 10.1089/omi.2014.0135

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326. doi: 10.1158/0008-5472.can-11-1021

Lei, X., and Bian, C. (2020). Integrating random walk with restart and k-nearest neighbor to identify novel circRNA-disease association. *Sci. Rep.* 10:1943.

Li, G., Zhang, H., Wan, X., Yang, X., Zhu, C., Wang, A., et al. (2014a). Long noncoding RNA plays a key role in metastasis and prognosis of hepatocellular carcinoma. *Biomed Res. Int.* 2014:780521.

Li, J., Gao, C., Wang, Y., Ma, W., Tu, J., Wang, J., et al. (2014b). A bioinformatics method for predicting long noncoding RNAs associated with vascular disease. *Sci. China Life Sci.* 57, 852–857. doi: 10.1007/s11427-014-4692-4

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., et al. (2011). DOSim: an R package for similarity between diseases based on disease ontology. *BMC Bioinformatics* 12:266. doi: 10.1186/1471-2105-12-266

Li, W., Wang, S., Xu, J., Mao, G., Tian, G., and Yang, J. (2019). Inferring latent disease-lncRNA associations by faster matrix completion on a heterogeneous network. *Front. Genet.* 10:769. doi: 10.3389/fgene.2019.00769

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi: 10.1016/j.omtn.2020.07.003

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.* 6:22811.

Lu, C., Yang, M., Luo, F., Fang-Xiang, W., Li, M., Pan, Y., et al. (2018a). Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* 34, 3357–3364. doi: 10.1093/bioinformatics/bty327

Lu, M., Xu, X., Xi, B., Dai, Q., Li, C., Su, L., et al. (2018b). Molecular network-based identification of competing endogenous RNAs in thyroid carcinoma. *Genes (Basel)* 9:44. doi: 10.3390/genes9010044

Maass, P. G., Luft, F. C., and Bähring, S. (2014). Long non-coding RNA in health and disease. *J. Mol. Med.* 92, 337–346.

Meng, J., Li, P., Zhang, Q., Yang, Z., and Fu, S. (2014). A four-long non-coding RNA signature in predicting breast cancer survival. *J. Exp. Clin. Cancer Res.* 33, 84–84.

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

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., et al. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44, 980–985.

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2020). Probing lncRNA–protein interactions: data repositories, models, and algorithms. *Front. Genet.* 10:1358. doi: 10.3389/fgene.2019.01346

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* 10, 2074–2081. doi: 10.1039/c3mb70608g

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Tseng, Y. Y., Moriarity, B. S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. *Nature* 512, 82–86. doi: 10.1038/nature13311

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Wang, L., Xiao, Y., Li, J., Feng, X., Li, Q., and Yang, J. (2019). IIRWR: internal inclined random walk with restart for LncRNA-disease association prediction. *IEEE Access* 7, 54034–54041. doi: 10.1109/access.2019.2912945

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., et al. (2013). Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis.* 4:e765. doi: 10.1038/cddis.2013.292

Xiang, J., Zhang, J., Zheng, R., Li, X., and Li, M. (2021). NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief. Bioinform.* doi: 10.1093/bib/bbab080

Xiang, J., Zhang, N. R., Zhang, J. S., Lv, X. Y., and Li, M. (2020). PrGeFNE: predicting disease-related genes by fast network embedding. *Methods* 192, 3–12. doi: 10.1016/j.ymeth.2020.06.015

Xiao, X., Wen, Z., Bo, L., Xu, J., and Gu, C. (2018). BPLLDA: predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* 9:411. doi: 10.3389/fgene.2018.00411

Xiao, Y., Lv, Y., Zhao, H., Gong, Y., Hu, J., Li, F., et al. (2015). Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network. *BioMed Res. Int.* 2015:839590.

Xing, C., Liu, M. X., and Yan, G. Y. (2012). RWRMDA: predicting novel human microRNA–disease associations. *Mol. Biosyst.* 8, 2792–2798. doi: 10.1039/c2mb25180a

Xu, J., Zhu, W., Cai, L., Liao, B., Meng, Y., Xiang, J., et al. (2020). LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. *IEEE Access* 8, 80728–80738. doi: 10.1109/access.2020.2990533

Xu, N., Wang, F., Lv, M., and Cheng, L. (2015). Microarray expression profile analysis of long non-coding RNAs in human breast cancer: a study of Chinese women. *Biomed. Pharmacother.* 69, 221–227. doi: 10.1016/j.biopha.2014.12.002

Yan, C., Zhang, Z., Bao, S., Hou, P., and Sun, J. (2020). Computational methods and applications for identifying disease-associated lncRNAs as potential biomarkers and therapeutic targets. *Mol. Ther. Nucleic Acids* 21, 156–171. doi: 10.1016/j.omtn.2020.05.018

Yan, X., Bao, M. H., Luo, H. Q., Xiang, J., and Li, J. M. (2015). A meta-analysis of the association between polymorphisms in MicroRNAs and risk of ischemic stroke. *Genes* 6, 1283–1299. doi: 10.3390/genes6041283

Yang, J., Huang, T., Song, W. M., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016). Discover the network underlying the connections between aging and age-related diseases. *Sci. Rep.* 6:32566.

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42, 353–372. doi: 10.1007/s11357-019-00106-x

Yang, J., Qiu, J., Wang, K., Zhu, L., Fan, J., Zheng, D., et al. (2017). Using molecular functional networks to manifest connections between obesity and obesity-related diseases. *Oncotarget* 8, 85136–85149. doi: 10.18632/oncotarget.19490

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

Yang, Z., Zhou, L., Wu, L., Lai, M., Xie, H., Zhang, F., et al. (2011). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular

carcinoma patients following liver transplantation. *Ann. Surg. Oncol.* 18, 1243–1250. doi: 10.1245/s10434-011-1581-y

Yanofsky, C. (2007). Establishing the triplet nature of the genetic code. *Cell* 128, 815–818. doi: 10.1016/j.cell.2007.02.029

Yao, Y., Ji, B., Shi, S., Xu, J., Xiao, X., Yu, E., et al. (2019). IMDAILM: inferring miRNA-disease association by integrating lncRNA and miRNA data. *IEEE Access* 8, 16517–16527. doi: 10.1109/access.2019.2958055

Zhang, Y., Huang, H., Zhang, D., Qiu, J., Yang, J., Wang, K., et al. (2017). A review on recent computational methods for predicting noncoding RNAs. *Biomed. Res. Int.* 2017:9139504.

Zhang, Z., Hao, H., Zhang, C. J., Yang, X. Y., He, Q., and Lin, J. (2012). Evaluation of novel gene UCA1 as a tumor biomarker for the detection of bladder cancer. *Natl. Med. J. China* 92, 384–387.

Zhao, T., Xu, J., Liu, L., Bai, J., Xu, C., Xiao, Y., et al. (2015). Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Mol. Biosyst.* 11, 126–136. doi: 10.1039/c4mb00478g

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., et al. (2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769. doi: 10.1039/c4mb00511b

Conflict of Interest: BJ was employed by Geneis Beijing Co., Ltd.

The remaining 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.

The handling editor declared a past co-authorship with one of the authors JX.