frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 January 2023
DOI 10.3389/fgene.2022.1023615

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Guanghui Li,
East China Jiaotong University, China
Li Zejun,
Professional Services Review, Australia

*CORRESPONDENCE
Shijun Li,
cflishijun6588@sina.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 20 August 2022
ACCEPTED 10 October 2022
PUBLISHED 20 January 2023

CITATION
Li S, Chang M, Tong L, Wang Y, Wang M and Wang F (2023), Screening potential lncRNA biomarkers for breast cancer and colorectal cancer combining random walk and logistic matrix factorization.
Front. Genet. 13:1023615.
doi: 10.3389/fgene.2022.1023615

Screening potential lncRNA biomarkers for breast cancer and colorectal cancer combining random walk and logistic matrix factorization

Shijun Li*†, Miaomiao Chang†, Ling Tong, Yuehua Wang, Meng Wang and Fang Wang

Department of Pathology, Chifeng Municipal Hospital, Chifeng, China

Breast cancer and colorectal cancer are two of the most common malignant tumors worldwide. They cause the leading causes of cancer mortality. Many researches have demonstrated that long noncoding RNAs (lncRNAs) have close linkages with the occurrence and development of the two cancers. Therefore, it is essential to design an effective way to identify potential lncRNA biomarkers for them. In this study, we developed a computational method (LDA-RWLMF) by integrating random walk with restart and Logistic Matrix Factorization to investigate the roles of lncRNA biomarkers in the prognosis and diagnosis of the two cancers. We first fuse disease semantic and Gaussian association profile similarities and lncRNA functional and Gaussian association profile similarities. Second, we design a negative selection algorithm to extract negative LncRNA-Disease Associations (LDA) based on random walk. Third, we develop a logistic matrix factorization model to predict possible LDAs. We compare our proposed LDA-RWLMF method with four classical LDA prediction methods, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM. The results from 5-fold cross validation on the MNDR dataset show that LDA-RWLMF computes the best AUC value of 0.9312, outperforming the above four LDA prediction methods. Finally, we rank all lncRNA biomarkers for the two cancers after determining the performance of LDA-RWLMF, respectively. We find that 48 and 50 lncRNAs have the highest association scores with breast cancer and colorectal cancer among all lncRNAs known to associate with them on the MNDR dataset, respectively. We predict that lncRNAs HULC and HAR1A could be separately potential biomarkers for breast cancer and colorectal cancer and need to biomedical experimental validation.

KEYWORDS

breast cancer, colorectal cancer, lncRNA, biomarker, lncRNA-disease association, random walk, logistic matrix factorization

1 Introduction

Breast cancer is the second leading cause of cancer-related death in women worldwide and the most common malignant tumor among US woman (Sun et al., 2017; DeSantis et al., 2019; Yang et al., 2013; Waks and Winer, 2019). During the past 25 years, breast cancer mortality rate showed a substantial increase in the world (Garrido-Castro et al., 2019). This increasing rate is one threaten to health for women in the world, in particular women from developing and low-income regions. More than 1.5 million women were diagnosed to breast cancer every year, which accounts for 25% among all women with cancers (Sun et al., 2017). In 2018, breast cancer accounts for approximately 24% of new cancer cases and approximately 15% of cancer deaths in women (Heer et al., 2020). In 2019, it is estimated that about 268,600 new patients suffer from invasive breast cancer and 48,100 patients suffer from ductal carcinoma *in situ* among US women. Moreover, 41,760 women may die from breast cancer in the same year (DeSantis et al., 2019). About 13% of women may suffer from invasive breast cancer in lifetime (DeSantis et al., 2019). The incident rate of breast cancer will increase by more than 46% by 2040 (Heer et al., 2020). Consequently, breast cancer has been one essential problem to be solved around the world.

However, the precise mechanisms of breast cancer remain unclear (Barzaman et al., 2020). Systemic treatment of breast cancer patients mainly consists of chemotherapy, endocrine treatment, and targeted therapy (Campos-Parra et al., 2018). In spite of rapid progress in different treatment strategies, accumulating patients show recurrence of the disease and decreased survival because of therapy resistance, which increases metastasis rates (Sledge et al., 2014). Once the metastasis occurs, the 5-year overall survival rate may be below 25% (Siegel et al., 2013).

Colorectal cancer is the third most frequent cancer and the second most death-caused cancer. It is estimated that there are about 1.9 million new cases and 0.9 million death cases worldwide in 2020 (Xi and Xu, 2021). Of new diagnose cases, 20% of patients have metastases and another 25% with localized disease may later develop metastases (Biller and Schrag, 2021). Its incidence is high in developed countries and is increasing in low- and middle-income countries, which poses a challenge to global public health (Biller and Schrag, 2021; Xi and Xu, 2021).

In this situation, it is essential to discover novel molecular biomarkers that can characterize therapy response for breast cancer and colorectal cancer. We can extend the overall survival rates of patients and delay or prevent the two cancers from metastases based on molecular biomarkers (Campos-Parra et al., 2018). Consequently, screening reliable biomarker is a research hotspot on the diagnosis and treatment of cancer including breast cancer and colorectal cancer (Huang et al., 2019; Yang et al., 2020; Peng et al., 2022a).

A substantial number of evidence suggest that over 80% of the human genome can be transcribed into non-coding RNAs, such as microRNAs (Peng et al., 2017; Peng et al., 2018; Chen et al., 2019; Huang et al., 2021), circle RNAs (Zhao et al., 2019; Lan et al., 2022), and long non-coding RNAs (lncRNAs) (Zhang et al., 2021a; Peng et al., 2021a; Peng et al., 2022b; Zhou et al., 2021a; Zhou et al., 2021b). In particular, lncRNAs obtain emerging interest as diagnostic biomarkers and therapeutic targets (Chandra Gupta and Nandan Tripathi, 2017; Guo et al., 2022). Differential expression of lncRNAs forms specific patterns to various complex diseases including cancer (Wahlestedt C, 2013). Once the regulation effects of lncRNAs are detected, they are promising therapeutic targets.

LncRNAs are closely related to breast cancer and colorectal cancer. For example, lncRNA BCRT1, MaTAR25, DSCAM-AS1, and CDC6 can promote breast cancer progression (Niknafs et al., 2016; Kong et al., 2019a; Chang et al., 2020; Liang et al., 2020), BCRT4 can induce signaling transduction in breast cancer (Xing et al., 2015), LINC00673 can promote cell proliferation of breast cancer (Qiao et al., 2019), and BORG can cause breast cancer metastasis and disease recurrence (Gooding et al., 2017). SNHG11, FEZF1-AS1, RP11, and DLEU1 have been reported to novel biomarkers of colorectal cancer (Bian et al., 2018; Liu et al., 2018; Wu et al., 2019; Xu et al., 2020). Thus, many computational models have been developed to discover lncRNA biomarkers for cancers (Peng et al., 2020a; Shen et al., 2022; Sun et al., 2022), for instance, rotation forest (Guo et al., 2019), KATZ measure (Chen, 2015), collaborative deep learning (Lan et al., 2020), matrix factorization (Fu et al., 2018; Wang et al., 2021a), network consistency projection (Li et al., 2019), and graph autoencoder (Shi et al., 2021).

In this manuscript, inspired by the association prediction method provided by Peng et al. (2020b), we develop a computational method, LDA-RWLMF, to predict LncRNA-Disease Associations (LDAs). LDA-RWLMF integrates random walk and Logistic Matrix Factorization to discover the roles of lncRNA biomarkers in the prognosis and diagnosis for breast cancer and colorectal cancer. First, we compute disease similarity and lncRNA similarity. Second, we first use random walk to extract negative LDAs. Third, we explored a logistic matrix factorization model to predict possible LDAs. The results from 5-fold cross validation show that LDA-RWLMF computes the best AUC value of 0.9312 on the MNDR dataset. Finally, we rank all lncRNA biomarkers for breast cancer and colorectal cancer after determining the performance of LDA-RWLMF.

2 Datasets

2.1 LncRNA-disease associations

Human LDA dataset was collected from the MNDR database (Cui et al., 2018; Fan et al., 2020) (http://www.rna-society.org/mndr/index.html). There are 1,529 LDAs between 89 diseases and 190 lncRNAs after preprocessing. For an LDA matrix between n lncRNAs and m diseases, we use $\boldsymbol{Y} \epsilon \mathrm{R}^{\mathrm{n} \times \mathrm{m}}$ to describe the association information by Eq. 1:

$$\boldsymbol{Y}_{ij} = \begin{cases} 1 \text{ If lncRNA } l_i & \text{associates with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2.2 Disease semantic similarity

We use the method provided by Fan et al. (2020) to compute disease semantic similarity based on the MeSH descriptors. Disease semantic similarity method provided by Fan et al. (2020) was based on LNCSIM1 and LNCSIM2 provided by Chen (2015). For a disease A, suppose that T_A represents its ancestor node set, E_A denotes all edge set, its Directed Acyclic Graph (DAG) is represented as $DAG_A = \{T_A, E_A\}$. For a disease term $t \in T_A$ in DAG_A, its semantic contribution to A is calculated by Eq. 2 (Chen, 2015):

$$SV_A^1(t) = \begin{cases} 1 & t = A \\ \max\ (\Delta \times SV_A^1(t') \big| t' \in C(t) & t \neq A \end{cases} \quad (2)$$

where $C(t)$ indicates the children of t, Δ indicates the sematic contribution factor related to edges that link t' to t, and Δ was usually set as 0.5 (Wang et al., 2010).

The above equation demonstrates that terms at the same layer from DAG_A have the same semantic contribution to A. But if two terms t_1 and t_2 are in the same layer of DAG_A and t_1 appears in less in DAG_A than t_2, the conclusion from t_1 will be more specific than one from t_2, thus, $SV_A^1(t_1)$ is higher than $SV_A^1(t_2)$.

In this case, we compute the second semantic contribution of term $t \in T_A$ to disease A by Eq. 3:

$$SV_A^2(t) = -log\frac{Dags(t)}{D} \quad (3)$$

where D indicates the number of diseases in MeSH, $Dags(t)$ indicates the number of DAGs that contain the disease term t. And the semantic contribution of t in DAG_A can be defined by Eq. 4:

$$SV_A^3(t) = \begin{cases} 1 & t = A \\ \max\ ((\Delta + \nabla) SV_A^3(t') \big| t' \in C(t) & t \neq A \end{cases} \quad (4)$$

where ∇ indicates the contribution factor related to information content, and is computed by Eq. 5:

$$\nabla = \frac{\max_{k \in K}(Dags(k)) - dags(t)}{D} \quad (5)$$

where K indicates the disease set in MeSH.

Furthermore, the contribution of all terms in DAG_A to the disease A is computed by Eq. 6:

$$SV(A) = \sum_{t \in T_A} SV_A^3(t) \quad (6)$$

Finally, the semantic similarity between two diseases (A and B) can be computed by Eq. 7:

$$S_d^s(A, B) = \frac{\sum_{t \in T_A \cap T_B}(SV_A^3(t) + SV_B^3(t))}{SV(A) + SV(B)} \quad (7)$$

2.3 LncRNA functional similarity

We use the method provided by Fan et al. (Fan et al., 2020) to compute lncRNA functional similarity. Let that $DG(u)$ [or $DG(v)$] indicate diseases linking to lncRNA u (or v) on LDA matrix, the similarity between two lncRNAs u and v is obtained through disease semantic similarity in DG (u) and DG (v). A disease semantic similarity sub-matrix is first constructed. In the constructed matrix, rows and columns are diseases in $DG(u) \cup DG(v)$, and each element indicates the semantic similarity between diseases. Suppose that d_u indicate a disease in DG (u), the similarity between d_u and $DG(v)$ is computed by Eq. 8:

$$S(d_u, DG(v)) = \max_{d \in DG(v)}(S_d(d_u, d)) \quad (8)$$

Similarly, the similarity between d_v and DG (u) is computed by Eq. 9:

$$S(d_v, DG(u)) = \max_{d \in DG(u)}(S_d(d_v, d)) \quad (9)$$

And the similarity of $DG(u) \rightarrow DG(v)$ is computed by Eq. 10:

$$S_{u \rightarrow v} = \sum_{d \in DG(u)} S(d, DG(v)) \quad (10)$$

And similarity of $DG(v) \rightarrow DG(u)$ is computed by Eq. 11:

$$S_{v \rightarrow u} = \sum_{d \in DG(v)} S(d, DG(u)) \quad (11)$$

The similarity between lncRNAs u and v is measured based on the disease semantic similarity by Eq. 12:

$$S_l^f(u, v) = \frac{S_{u \rightarrow v} + S_{v \rightarrow u}}{|DG(u)| + |DG(v)|} \quad (12)$$

where $|DG(u)|$ and $|DG(v)|$ are the number of diseases in $DG(u)$ and $DG(v)$.

3 Methods

We want to compute association probability for each lncRNA-disease pair based on disease semantic similarity and lncRNA functional similarity. The pipeline is shown in Figure 1.

3.1 Gaussian association profile similarity and similarity fusion

In this section, we use Gaussian Association Profile (GAP) to compute the GAP similarity of diseases and lncRNAs. For a lncRNA l_i, its GAP $AP(l_i)$ is denoted using the i th row of $\boldsymbol{Y}$. The GAP similarity of lncRNAs l_i and l_j is defined by Eq. 13:

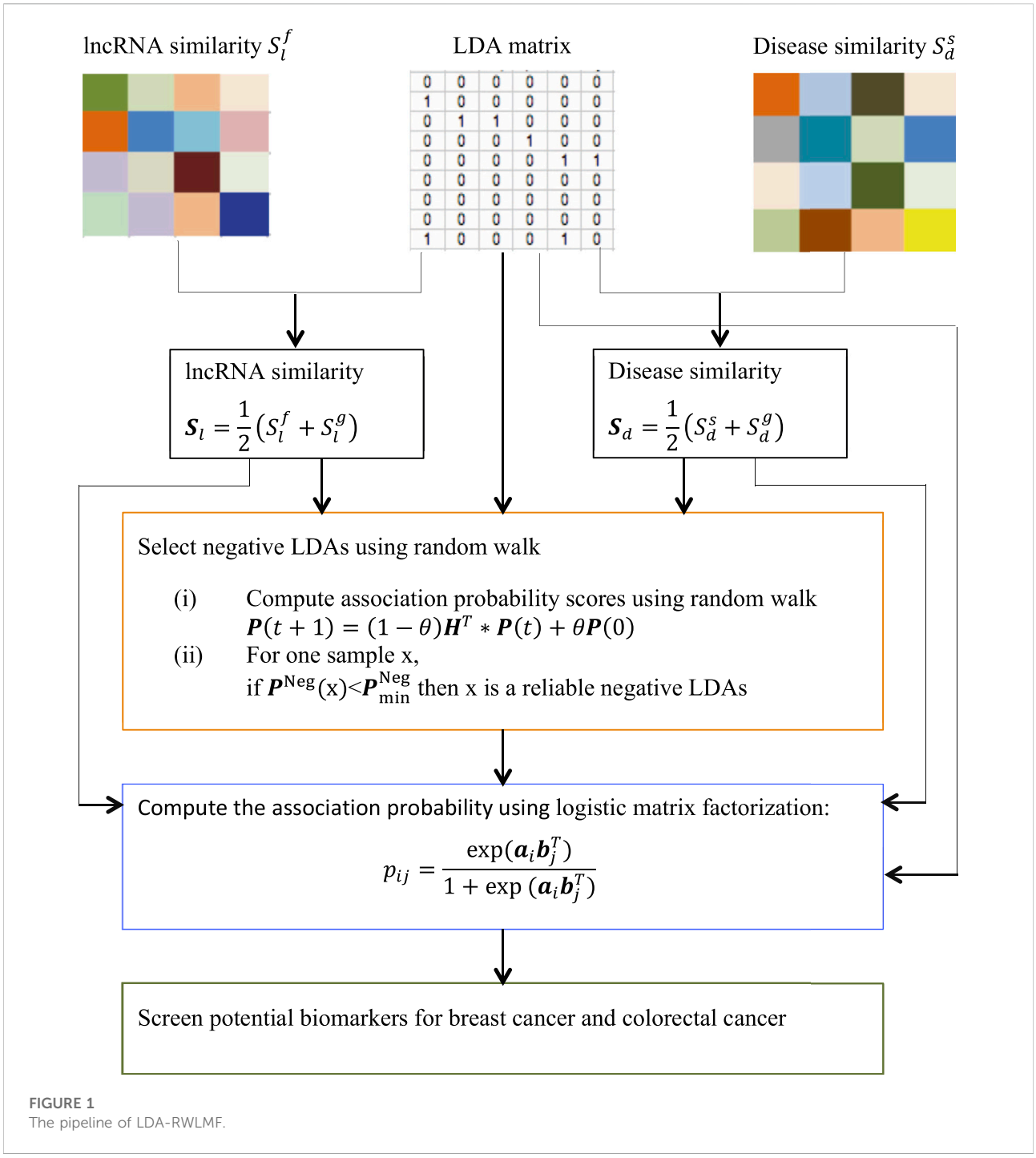

FIGURE 1
The pipeline of LDA-RWLMF.

$$S_l^g(l_i, l_j) = \exp\left(-\gamma_l \left\| AP(l_i) - AP(l_j) \right\|^2\right) \quad (13)$$

where $\gamma_l = \gamma_l' / (\frac{1}{n}\sum_{k=1}^{n} \| AP(l_i) \|^2)$ is the normalized kernel bandwidth with parameter γ_l'. Thus, the lncRNA similarity matrix $\boldsymbol{S}_l$ is computed by Eq. 14:

$$\boldsymbol{S}_l = \frac{1}{2}(S_l^f + S_l^g) \quad (14)$$

Similarly, the disease GAP similarity $\boldsymbol{S}_d$ can be computed.

3.2 Screening negative LDAs

There are not negative LDAs in the MNDR dataset. Credible negative LDAs help improve LDA prediction performance and further more effectively find potential lncRNA biomarkers for

breast cancer and colorectal cancer. Peng et al. (2021b) developed a random walk with restart-based virus-drug association prediction method and obtained better performance. Inspired by the method provided by Peng et al. (2021b), we first compute association probability for each lncRNA-disease pair through random walk with restart and then screen credible negative LDAs.

We first constructed a heterogeneous network composed of lncRNA similarity network, disease similarity network, and LDA network. lncRNA similarity matrix $\boldsymbol{S}_l$, disease similarity matrix $\boldsymbol{S}_d$, and LDA matrix Y are used as the adjacency matrices related to the heterogeneous network. The adjacency matrix related to the heterogeneous network is represented as Eq. 15:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{S}_l & \boldsymbol{Y} \\ \boldsymbol{Y}^T & \boldsymbol{S}_d \end{bmatrix} \tag{15}$$

where $\boldsymbol{Y}^T$ denotes the transpose of $\boldsymbol{Y}$.

We then compute transition probability on the heterogeneous graph. Suppose that $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{ll} & \boldsymbol{H}_{ld} \\ \boldsymbol{H}_{dl} & \boldsymbol{H}_{dd} \end{bmatrix}$ indicate transition probability matrix, where $\boldsymbol{H}_{ll}$ and $\boldsymbol{H}_{dd}$ indicate the walks within lncRNA similarity network and disease similarity network, respectively, $\boldsymbol{H}_{ld}$ and $\boldsymbol{H}_{dl}$ indicate the jumps between networks. For an lncRNA/disease, when there is an association between the lncRNA/disease and diseases/lncRNAs, the node will either continue to walk in the current network based on a transition probability $\lambda \in [0, 1]$ or jump between the above four networks.

The i -th lncRNA will walk to the j -th lncRNA through the transition probability $H_{ll}(i, j)$ by Eq. 16:

$$\boldsymbol{H}_{ll}(i,j) = \begin{cases} \dfrac{S_l(i,j)}{\sum_{k=1}^{n} S_l(i,k)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(i,k) = 0 \\ \dfrac{(1-\lambda) S_l(i,j)}{\sum_{k=1}^{n} S_l(i,k)}, & otherwise \end{cases} \tag{16}$$

or jump to a disease d_j through the transition probability $H_{ld}(i, j)$ by Eq. 17:

$$\boldsymbol{H}_{ld}(i,j) = \begin{cases} \dfrac{\lambda Y(i,j)}{\sum_{k=1}^{m} Y(i,k)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(i,k) \neq 0 \\ 0, & otherwise \end{cases} \tag{17}$$

Similarly, the i -th disease d_i will walk to the j -th disease d_j through the transition probability $\boldsymbol{H}_{dd}(i, j)$ by Eq. 18:

$$\boldsymbol{H}_{dd}(i,j) = \begin{cases} \dfrac{S_d(i,j)}{\sum_{k=1}^{m} S_d(i,k)}, & \text{if } \sum_{k=1}^{n} \boldsymbol{Y}(k,i) = 0 \\ \dfrac{(1-\lambda) S_d(i,j)}{\sum_{k=1}^{m} S_d(i,k)}, & otherwise \end{cases} \tag{18}$$

or jump to an lncRNA l_j through the transition probability $H_{dl}(i, j)$ by Eq. 19:

$$\boldsymbol{H}_{dl}(i,j) = \begin{cases} \dfrac{\lambda Y(i,j)}{\sum_{k=1}^{n} Y(k,i)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(k,i) \neq 0 \\ 0, & otherwise \end{cases} \tag{19}$$

At the $t-$ th step, the association probability matrix between all lncRNA-disease pairs on the heterogeneous network is computed by Eq. 20:

$$\boldsymbol{P}(t+1) = (1-\theta)\boldsymbol{H}^T * \boldsymbol{P}(t) + \theta \boldsymbol{P}(0) \tag{20}$$

where $\boldsymbol{H}^T$ indicates the transpose of $\boldsymbol{H}$, and θ is the restarting probability. $\boldsymbol{P}(0)$ indicates the initial probability with $p_i(0) = \begin{bmatrix} (1-\eta) v_i \\ \eta s_i \end{bmatrix}$, where v_i and s_j indicate the initial probability distributions on disease similarity network and lncRNA similarity network, respectively. And $\eta \in [0, 1]$ is used to control the restarting probability in these two similarity networks. If $\eta < 0.5$, the particle will more tend to restart from one of the seed microbes than from one of the seed diseases.

In the second step, we consider known LDAs as positive sample set P, unknown lncRNA-disease pairs as unlabeled set U and propose a PU learning approach to screen credible negative LDA sample set RN. The method contains the following six steps:

Step 1. Randomly screening positive sample subset D from P

Step 2. Adding D into U;

Step 3. Considering $P - D$ as positive samples, $U + D$ as negative samples;

Step 4. Obtaining LDA score matrix $\boldsymbol{S}^{Neg}$ using random walk with restart;

Step 5. Ranking lncRNA-disease pairs in $\boldsymbol{D}$ based on $\boldsymbol{S}_{min}^{Neg}$ and obtaining the minimum score $\boldsymbol{S}_{min}^{Neg}$ in D;

Step 6. For every lncRNA-disease pair x in U:

If $\boldsymbol{S}^{Neg}(x) < \boldsymbol{S}_{min}^{Neg}$ then $RN = RN \cup x$.

3.3 LDA prediction based on logistic matrix factorization

Logistic matrix factorization has been applied to multiple areas (Liu et al., 2020; Tang et al., 2021; Tian et al., 2022). Inspired by the approaches, we develop a logistic matrix factorization-based LDA prediction method, LDA-RWLMF.

Assume that both lncRNAs and diseases are mapped to r-dimensional shared latent spaces ($r \ll n, m$), thus an lncRNA l_i or disease d_i can be represented as a latent vector $\boldsymbol{a}_i \in \mathfrak{R}^{1\times t}$ or $\boldsymbol{b}_i \in \mathfrak{R}^{1\times t}$. The association probability p_{ij} between l_i and d_i is calculated by Eq. 12:

TABLE 1 AUCs of LDA identification approaches on the MNDR dataset.

Dataset	LNCSIM1	LNCSIM2	ILNCSIM	IDSSIM	LDA-RWLMF
the MNDR dataset	0.9251	0.9280	0.9267	0.9302	0.9312

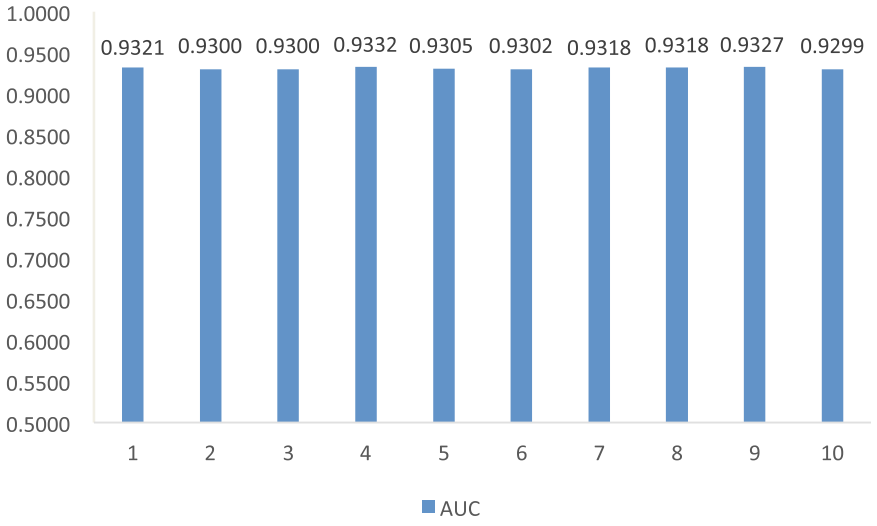

FIGURE 2
The AUC of LDA-RWLMF from 10 time cross validation (t = 1, 2, 3, . . . , 10).

$$p_{ij} = \frac{\exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)}{1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)} \tag{21}$$

The latent vector matrix of all lncRNAs or diseases can be represented as $\boldsymbol{A} \in \Re^{n \times r}$ or $\boldsymbol{B} \in \Re^{m \times r}$ where $\boldsymbol{a}_i$ or $\boldsymbol{b}_i$ indicates the i th or j th row in $\boldsymbol{A}$ or $\boldsymbol{B}$. In addition, known LDAs are more credible than unknown lncRNA-disease pairs. Thus, we assign higher confidence values to known LDAs than unknown lncRNA-disease pairs. Similar to Peng et al. (2020b), we use a constant c to assess the importance of known LDAs and construct a prediction model by Eq. 22:

$$\begin{aligned} p(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{B}) = & \left(\prod_{1 \le i \le n, 1 \le j \le m, y_{ij}=1} \left[p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \right]^c \right) \\ & \times \left(\prod_{1 \le i \le n, 1 \le j \le m, y_{ij}=0} \left[p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \right] \right) \\ = & \prod_{i=1}^{n} \prod_{j=1}^{m} p_{ij}^{cy_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \end{aligned} \tag{22}$$

Model (21) can be optimized based on the Bayesian distribution by Eq. 23:

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{B}} & \sum_{i=1}^{m} \sum_{j=1}^{n} \left(1 + cy_{ij} - y_{ij}\right) \log\left[1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)\right] - cy_{ij} \boldsymbol{a}_i \boldsymbol{b}_j^T + \frac{\lambda_l}{2} \\ & \| \boldsymbol{A} \|_F^2 + \frac{\lambda_d}{2} \| \boldsymbol{B} \|_F^2 \end{aligned} \tag{23}$$

where λ_l and λ_d are two parameters, $\|\boldsymbol{A}\|_F$ indicates the Frobenius norm of $\boldsymbol{A}$. (Zhang et al. 2019a; Zhang et al. 2019b) integrated linear neighborhood information to model (22) to predict various associations. Similarly, we fuse neighborhood information to Eq. 23 by Eq. 24:

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{B}} & \sum_{i=1}^{m} \sum_{j=1}^{n} \left(1 + cy_{ij} - y_{ij}\right) \ln\left[1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)\right] - cy_{ij} \boldsymbol{a}_i \boldsymbol{b}_j^T \\ & + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{A}^T \left(\lambda_l \boldsymbol{I} + \alpha \boldsymbol{L}_l\right) \boldsymbol{A} + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{B}^T \left(\lambda_d \boldsymbol{I} + \alpha \boldsymbol{L}_d\right) \boldsymbol{B}\right.\right. \end{aligned} \tag{24}$$

where $tr\ (\cdot)$ indicates the trace of the matrix. $\boldsymbol{L}_l$ and $\boldsymbol{L}_d$ indicate the corresponding Laplacian matrix of $\boldsymbol{A}$ and $\boldsymbol{B}$. $\boldsymbol{L}_l = (\boldsymbol{D}_l + \tilde{\boldsymbol{D}}_l) - (\boldsymbol{A} + \boldsymbol{A}^T)$ where $\boldsymbol{D}_l$ and $\tilde{\boldsymbol{D}}_l$ are two diagonal matrices and $\boldsymbol{D}_l(i,i) = \sum_{j=1}^{m} a_{ij}$ and $\tilde{\boldsymbol{D}}_l(i,i) = \sum_{i=1}^{m} a_{ij}$. Similarly, $\boldsymbol{L}_d$ can be computed.

We compute $\boldsymbol{A}$ and $\boldsymbol{B}$ by solving Eq. 24 through an alternating gradient ascent approach.

TABLE 2 The rankings of the predicted top 48 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	CASC2	Known	25	PVT1	Known
2	DLEU2	Known	26	RMST	Known
3	MIR17HG	Known	27	TRAF3IP2-AS1	Known
4	DSCAM-AS1	Known	28	HCP5	Known
5	SNHG4	Known	29	LINC00271	Known
6	TCL6	Known	30	GHET1	Known
7	XIST	Known	31	SNHG3	Known
8	CBR3-AS1	Known	32	TDRG1	Known
9	MIAT	Known	33	DAOA-AS1	Known
10	CCAT2	Known	34	BACE1-AS	Known
11	SOX2-OT	Known	35	NAMA	Known
12	GAS5	Known	36	BDNF-AS	Known
13	PCA3	Known	37	SNHG11	Known
14	MALAT1	Known	38	UCA1	Known
15	BANCR	Known	39	SNHG16	Known
16	WT1-AS	Known	40	MIR100HG	Known
17	PANDAR	Known	41	H19	Known
18	HNF1A-AS1	Known	42	TERC	Known
19	HAR1B	Known	43	MEG3	Known
20	CCDC26	Known	44	SPRY4-IT1	Known
21	BCAR4	Known	45	DANCR	Known
22	PDZRN3-AS1	Known	46	KCNQ1OT1	Known
23	HIF1A-AS2	Known	47	IFNG-AS1	Known
24	CRNDE	Known	48	HOTAIR	Known

TABLE 3 The rankings of the remaining 41 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
49	HULC	PMID: 31824174, 33107484, 33745450	70	ZFAT-AS1	Unconfirmed
50	CCAT1	Known	71	PTENP1	PMID: 28731027, 29085464, 29212574, 31196157
51	NPTN-IT1	Unconfirmed	72	HIF1A-AS1	Unconfirmed
52	PCAT1	PMID: 32853955, 28989584, 33850635, 32220602	73	SRA1	Known
53	HAR1A	PMID: 26942882	74	MINA	Unconfirmed
54	LSINCT5	Known	75	DLEU1	Known
55	TUG1	PMID: 28950664, 27848085, 30098551, 33380806	76	PSORS1C3	Unconfirmed
56	MIR155HG	Unconfirmed	77	LINC00032	Unconfirmed
57	DGCR5	PMID: 32521856	78	WRAP53	Unconfirmed
58	IGF2-AS	PMID: 33175607	79	7SK	Unconfirmed
59	BCYRN1	Known	80	RRP1B	Unconfirmed
60	EPB41L4A-AS1	PMID: 35181612	81	MYCNOS	Unconfirmed
61	PINK1-AS	Unconfirmed	82	PRINS	Unconfirmed
62	DNM3OS	Unconfirmed	83	ATP6V1G2-DDX39B	Unconfirmed
63	ADAMTS9-AS2	PMID: 30840279	84	MKRN3-AS1	Unconfirmed
64	MIR31HG	lncRNADisease	85	NRON	Unconfirmed
65	BOK-AS1	Unconfirmed	86	MESTIT1	Unconfirmed
66	ESRG	Unconfirmed	87	LINC00162	Unconfirmed
67	KCNQ1DN	Unconfirmed	88	DISC2	Unconfirmed
68	ATXN8OS	PMID: 31173245, 33385064, 33477683	89	SCAANT1	Unconfirmed
69	CDKN2B-AS1	Known			

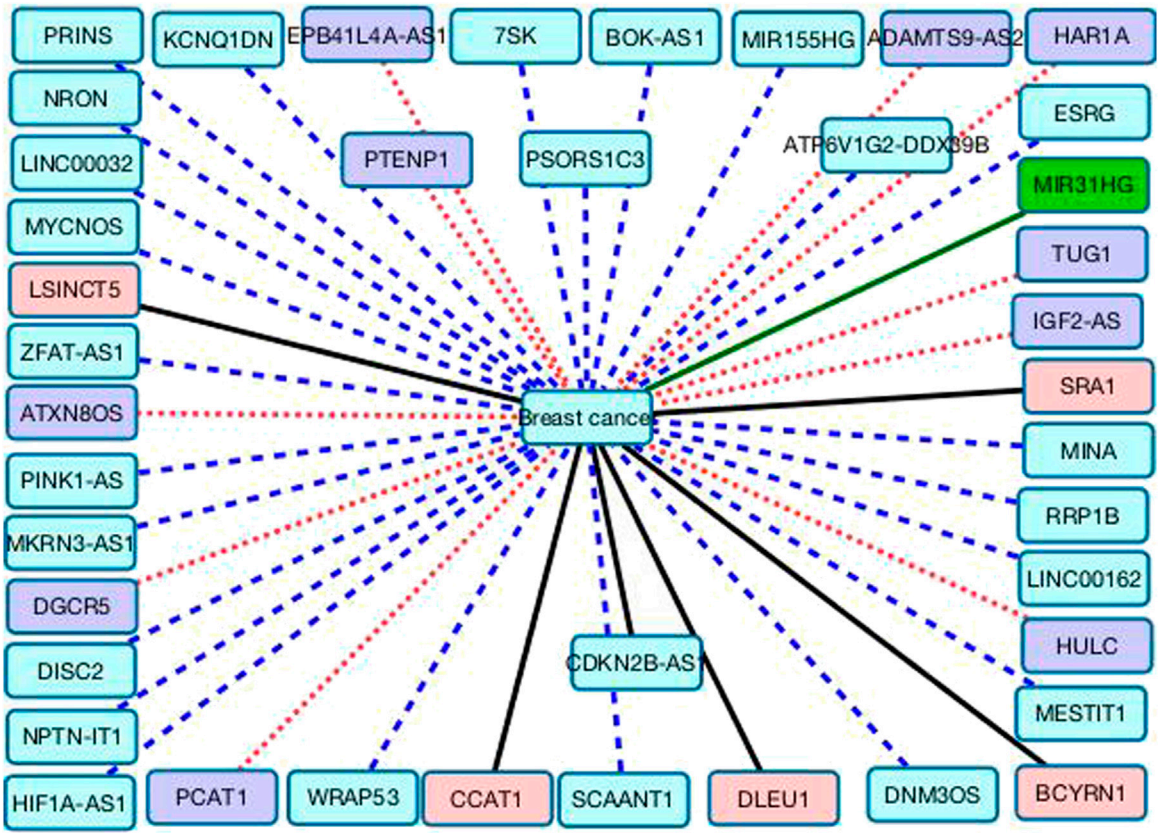

FIGURE 3
The associations between the remaining 41 lncRNAs and breast cancer.

Finally, lncRNA-disease association score $\boldsymbol{Y}_{fin}(i, j)$ for each lncRNA-disease pair can be computed by Eq. 25:

$$\boldsymbol{Y}_{fin} = \boldsymbol{A}\boldsymbol{B}^{T} \tag{25}$$

4 Results

4.1 Experimental settings

We conduct 5-fold cross validation for 10 times to investigate the performance of LDA-RWLMF. AUC is used to evaluate the prediction accuracy of LDA identification models. AUC is the area under the true positive rate (TPR)-false positive rate (FPR) curve, where TPR and FPR are defined by Eqs 26, 27:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{26}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{27}$$

where TP, FP, TN, FN represent the number of true positives, false positives, true negatives, false negatives, respectively. Higher AUC is, better the prediction performance is. In addition, parameters in LDA-RWLMF are set to defaults provided by Peng et al. (2020b). And parameters in the other four comparison LDA prediction methods (LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM) are set to the same values provided by corresponding methods.

4.2 Performance comparison with other methods

To measure the performance of the proposed LDA-RWLMF method, we compare it with four other representative LDA inference approaches on the MNDR dataset. That is, LNCSIM1 (Chen, 2015), LNCSIM2 (Chen, 2015), ILNCSIM (Huang et al., 2016), and IDSSIM (Fan et al., 2020). LNCSIM1 and LNCSIM2 used Laplacian regularized least squares to predict possible LDAs based on disease DAGs and the information content, respectively. ILNCSIM first combined the hierarchical structure of disease DAG and the information content to compute disease similarity and then used Laplacian

TABLE 4 The rankings of the identified top 50 lncRNAs associated with colorectal cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	SOX2-OT	Known	26	NAMA	Known
2	DLEU2	Known	27	WT1-AS	Known
3	CASC2	Known	28	TDRG1	Known
4	TCL6	Known	29	GHET1	Known
5	TRAF3IP2-AS1	Known	30	CRNDE	Known
6	DSCAM-AS1	Known	31	XIST	Known
7	GAS5	Known	32	MALAT1	Known
8	MIR17HG	Known	33	RMST	Known
9	HAR1B	Known	34	SNHG3	Known
10	CCDC26	Known	35	BACE1-AS	Known
11	CBR3-AS1	Known	36	MIR100HG	Known
12	PANDAR	Known	37	IFNG-AS1	Known
13	MIAT	Known	38	DANCR	Known
14	SNHG4	Known	39	SNHG16	Known
15	HIF1A-AS2	Known	40	SNHG11	Known
16	HNF1A-AS1	Known	41	TERC	Known
17	PCA3	Known	42	KCNQ1OT1	Known
18	BANCR	Known	43	MEG3	Known
19	LINC00271	Known	44	HULC	Known
20	PDZRN3-AS1	Known	45	UCA1	Known
21	CCAT2	Known	46	SPRY4-IT1	Known
22	BCAR4	Known	47	PCAT1	Known
23	DAOA-AS1	Known	48	HOTAIR	Known
24	BDNF-AS	Known	49	PVT1	Known
25	HCP5	Known	50	CCAT1	Known

regularized least squares to infer new LDAs. IDSSIM designed a weighted K nearest neighbor approach to identify potential associations between lncRNAs and diseases by integrating disease semantic similarity and lncRNA functional similarity. Table 1 gives the AUC values of the four LDA identification methods and our proposed LDA-RWLMF on the MNDR dataset.

The results from Table 1 demonstrate that LDA-RWLMF computes the highest AUC compared to LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM on the MNDR dataset. Figure 2 gives the results of LDA-RWLMF from 10 time cross validation. From Figure 2, we can find that AUC obtain by LDA-RWLMF is relatively steady during 10 time cross validation.

4.3 Case study

4.3.1 lncRNA biomarker identification for breast cancer

Breast cancer is the commonest life-threatening cancer in women (Key et al., 2001; Sharma, et al., 2010). lncRNAs play important roles in epigenetic regulation, transcriptional regulation and post-transcriptional regulation and have been potential biomarkers of many diseases. Substantial publications have reported that lncRNAs affect proliferation and apoptosis, invasion and metastasis, and cancer stemness of breast cancer. For example, LSINCT5 and Zfas one can promote the proliferation of breast cancer, HOTAIR suppresses invasion and migration of breast cancer, SOX2OT induces SOX2 expression in breast cancer, and SRA is the expression activator of breast cancer (Sun et al., 2017). We want to conduct case analyses to find possible lncRNA biomarkers for breast cancer based on the proposed LDA-RWLMF model.

In the MNDR dataset, there are 89 lncRNAs that may associate with breast cancer, where 54 lncRNAs have been experimentally validated to associate with the cancer and 35 lncRNAs have unknown associations with it. We use the proposed LDA-RWLMF method to rank the 89 lncRNAs for breast cancer. The results are shown in Tables 2, 3. Table 2 demonstrates the ranking results of the predicted top 48 lncRNAs according to the computed association score with breast cancer on the MNDR dataset. These 48 lncRNAs are known to link to breast cancer on the MNDR dataset and are ranked as top 48.

Table 3 gives the rankings of the remaining 41 lncRNAs according to the association scores with breast cancer on the

TABLE 5 The rankings of the remaining 41 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
51	HAR1A	Unconfirmed	71	ZFAT-AS1	Unconfirmed
52	NPTN-IT1	known	72	SRA1	Unconfirmed
53	TUG1	known	73	PSORS1C3	Unconfirmed
54	IGF2-AS	PMID: 32853944, 30581274	74	HIF1A-AS1	Unconfirmed
55	LSINCT5	known	75	MINA	Unconfirmed
56	DGCR5	PMID: 31452812	76	LINC00032	Unconfirmed
57	H19	known	77	WRAP53	Unconfirmed
58	EPB41L4A-AS1	PMID: 32557646	78	DLEU1	Unconfirmed
59	MIR155HG	PMID: 34562123,31228357	79	RRP1B	Unconfirmed
60	CDKN2B-AS1	known	80	7SK	Unconfirmed
61	MIR31HG	PMID: 30447009,35733512,34485123	81	PRINS	Unconfirmed
62	ESRG	PMID: 34896077	82	MYCNOS	Unconfirmed
63	BCYRN1	PMID: 30114690,32944001,31773686	83	ATP6V1G2-DDX39B	Unconfirmed
64	BOK-AS1	Unconfirmed	84	MKRN3-AS1	Unconfirmed
65	PINK1-AS	Unconfirmed	85	NRON	Unconfirmed
66	KCNQ1DN	Unconfirmed	86	SCAANT1	Unconfirmed
67	ATXN8OS	Unconfirmed	87	DISC2	Unconfirmed
68	DNM3OS	Unconfirmed	88	MESTIT1	Unconfirmed
69	PTENP1	Unconfirmed	89	LINC00162	Unconfirmed
70	ADAMTS9-AS2	Unconfirmed			

MNDR dataset. Among all lncRNAs unknown to associate with breast cancer on the MNDR dataset, lncRNA HULC is predicted to link to breast cancer with the highest association scores. Shi et al. (2016) observed that HULC can act as an oncogene biomarker in triple-negative breast cancer and as an independent possible poor prognostic factor in patients suffered from triple-negative breast cancer. Wang et al. (2019) found that HULC can promote the development of breast cancer through regulating the expression of LYPD1. Gavgani et al. (2020) investigated that the HULC knockdown can induce apoptosis and suppress cellular migration in breast cancer cells.

PCAT1 may link to breast cancer with the ranking of three among all lncRNAs unknown to associate with breast cancer on the MNDR dataset. Several studies have reported that PCAT1 can associate with breast cancer although its association with the cancer on the MNDR dataset is unobserved. Abdollahzadeh et al. (2020) reported that the altered regulation of PCAT1 may play crucial roles in the development and pathogenesis of breast cancer. Sarrafzadeh et al. (2017) assessed the expression of PCAT-1 through real-time reverse transcription polymerase chain reaction in breast tumor samples from 47 breast cancer patients and found that PCAT-1 may involve in the pathogenesis of breast cancers. Wang et al. (2021a) observed that PCAT-1 can facilitate breast cancer progression by binding to RACK1 and thus boosting oxygen-independent stability of HIF-1α. Tang et al. (2022) detect that PCAT1 can regulate the expression of PITX2 in breast cancer.

In addition, we predict that nephronectin intronic transcript 1 (NPTN-IT1, also known as lncRNA-LET) may have relationship with breast cancer. NPTN-IT1 has been reported to associate with bladder cancer through attenuating the expression of the target of miR-145 and ILF3 in bladder cancer (Zhang et al., 2021b). It was significantly down-regulated in multiple tumor tissues of colorectal cancer. It also has a regulation role in hypoxia signaling of hepatocellular carcinoma (Sun et al., 2013) and was highly expressed in HepG2 cells (Kong et al., 2019b). We hope that association between three lncRNAs (HULC, NPTN-IT1, and PCAT1) and breast cancer can be validated through wet experiments. Figure 3 shows the associations between the 41 lncRNAs that are ranked as the last 41 and breast cancer. Black solid lines represent known LDAs in the MNDR database. Green solid lines represent LDAs that can be observed in the lncRNA disease database. Red dots lines represent LDAs that are predicted to be potential lncRNA biomarkers of breast cancer and can be confirmed by related publications. Blue equal dash lines represent unknown LDAs.

4.3.2 lncRNA biomarker identification for colorectal cancer

Colorectal cancer is a heterogeneous disease. It has high morbidity and mortality. lncRNAs demonstrate dense associations with colorectal cancer. In this study, we conduct case analyses to identify possible lncRNA

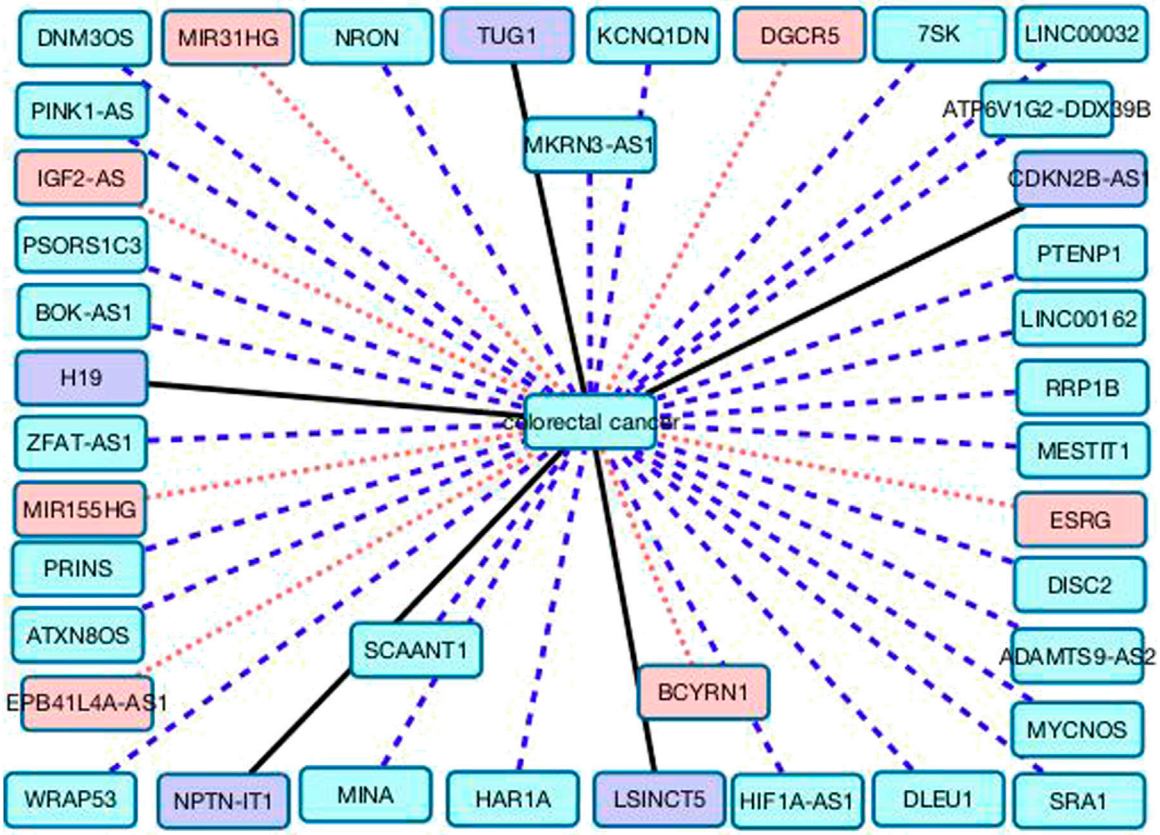

FIGURE 4
The associations between the remaining 39 lncRNAs and colorectal cancer.

biomarkers for colorectal cancer based on LDA-RWLMF. In the MNDR dataset, 89 lncRNAs possibly associate with colorectal cancer, where 55 lncRNAs have been validated to be the biomarkers of the cancer and remaining 34 lncRNAs have not been validated. We use LDA-RWLMF to compute the association scores between all 89 lncRNAs and colorectal cancer and rank the 89 lncRNAs for colorectal cancer. The results are shown in Tables 4, 5. Table 4 shows the rankings of the identified top 50 lncRNAs according to the computed association score with colorectal cancer on the MNDR dataset. The 50 lncRNAs are known to associate with colorectal cancer on the MNDR dataset and are ranked as top 50.

Table 5 gives the rankings of the remaining 39 lncRNAs according to the association scores with colorectal cancer on the MNDR dataset. Among all lncRNAs unknown association with colorectal cancer on the MNDR dataset, lncRNA HAR1A is inferred to link to colorectal cancer with the highest association scores. HAR1A is a favorable prognostic biomarker for patients. Shi et al. (2019) analyzed the expression profiles of HAR1A using RT-qPCR and found its expression level was significantly lower in hepatocullular cancer. Chen et al. (2020) have still reported that the HAR1A expression levels were reduced in hepatocellular carcinoma tissues.

Figure 4 gives the associations between the remaining 39 lncRNAs and colorectal cancer. Black solid lines represent known LDAs in the MNDR database. Red dots lines represent LDAs that are predicted to be potential lncRNA biomarkers of breast cancer and can be confirmed by related publications. Blue equal dash lines represent unknown LDAs.

5 Discussion and conclusion

Breast cancer and colorectal cancer are the most frequent cancers with high mortality rates. They demonstrate very high heterogeneity at molecular and clinical levels. With the fast development of next generation sequencing technologies, we can more accurately characterize the human genome. lncRNAs act mainly as gene expression regulators. The dysregulation of lncRNAs may destroy the normal transcriptional landscape and thus cause malignant transformation. In addition, their highly specific expression

and functional tertiary structure force them to be as promising diagnostic biomarkers and potential targets for various diseases including breast cancer and colorectal cancer.

In this study, we proposed a computational lncRNA-disease association method (LDA-RWLMF) to identify potential biomarkers for breast cancer and colorectal cancer. First, a random walk with restart method was designed to extract negative LDAs. Second, a logistic matrix factorization model was explored to infer possible associations between lncRNAs and diseases. Finally, all lncRNAs are ranked according to association scores with breast cancer and colorectal cancer on the MNDR dataset.

We conduct 5-fold cross validation for 10 times to compare LDA-RWLMF with state-of-the-art LDA prediction models on the MNDR dataset, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM. The results show that LDA-RWLMF computes the best AUC values of 0.9312. We predict that lncRNAs (HULC, NPTN-IT1, and PCAT1) may be possible biomarkers of breast cancer and colorectal cancer.

Our proposed LDA-RWLMF method has two disadvantages. First, it extracted credible negative LDA samples. In the area of association prediction, there are no negative association samples because of the limitation of biomedical experiments, which causes relatively poor performance. Thus, we designed a negative LDA extraction method based on PU learning. Second, the logistic matrix factorization model can effectively discover possible associations between two biological entities. Thus, we used the model to identify new LDAs. In addition, diseases and lncRNAs exhibit abundant biological features. In this study, we failed to consider these diverse features. In the future, we will further integrate more biological information to improve LDA prediction.

In the future, we will further design more effective negative sample screening method based on positive-unlabeled learning. In addition, we will also develop deep learning model for LDA prediction. We anticipate that the proposed LDA-RWLMF method can help design therapeutic regimens for personalized treatment of breast cancer and colorectal cancer and thus opportunely inhibit its recurrence.

Data availability statement

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

Author contributions

Conceptualization: SL, MC, and LT; Methodology: SL, MC, YW, MW, and FW; Project administration: SL; Software: SL and MC; Writing-original draft: SL; Writing-review and editing: SL, MC.

Conflict of interest

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

Publisher's note

References

Abdollahzadeh, R., Mansoori, Y., Azarnezhad, A., Daraei, A., Paknahad, S., Mehrabi, S., et al. (2020). Expression and clinicopathological significance of AOC4P, PRNCR1, and PCAT1 lncRNAs in breast cancer. *Pathol. Res. Pract.* 216 (10), 153131. doi:10.1016/j.prp.2020.153131

Barzaman, K., Karami, J., Zarei, Z., Hosseinzadeh, A., Kazemi, M. H., Moradi-Kalbolandi, S., et al. (2020). Breast cancer: Biology, biomarkers, and treatments. *Int. Immunopharmacol.* 84, 106535. doi:10.1016/j.intimp.2020.106535

Bian, Z., Zhang, J., Min, L., Feng, Y., Xue, W., Jia, Z., et al. (2018). LncRNA-FEZF1-AS1 promotes tumor proliferation and metastasis in colorectal cancer by regulating PKM2 signaling. *Clin. Cancer Res.* 24 (19), 4808–4819. doi:10.1158/1078-0432.CCR-17-2967

Biller, L. H., and Schrag, D. (2021). Diagnosis and treatment of metastatic colorectal cancer: A review. *Jama* 325 (7), 669–685. doi:10.1001/jama.2021.0106

Campos-Parra, A. D., López-Urrutia, E., Orozco Moreno, L. T., Lopez-Camarillo, C., Meza-Menchaca, T., Figueroa Gonzalez, G., et al. (2018). Long non-coding RNAs as new master regulators of resistance to systemic treatments in breast cancer. *Int. J. Mol. Sci.* 19 (9), 2711. doi:10.3390/ijms19092711

Chandra Gupta, S., and Nandan Tripathi, Y. (2017). Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets. *Int. J. Cancer* 140 (9), 1955–1967. doi:10.1002/ijc.30546

Chang, K. C., Diermeier, S. D., Yu, A. T., Brine, L. D., and Spector, D. L. (2020). MaTAR25 lncRNA regulates the Tensin1 gene to impact breast cancer progression [J]. *Nat. Commun.* 11 (1), 1–19.

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 (D1), D983–D986. doi:10.1093/nar/gks1099

Chen, X. (2015). Katzlda: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5 (1), 16840–16911. doi:10.1038/srep16840

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019). MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 20 (2), 515–539. doi:10.1093/bib/bbx130

Chen, Y., Guo, Y., Chen, H., and Ma, F. (2020). Long non-coding RNA expression profiling identifies a four-long non-coding RNA prognostic signature for isocitrate dehydrogenase mutant glioma. *Front. Neurol.* 11, 573264. doi:10.3389/fneur.2020.573264

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: An updated resource of ncRNA-disease associations in mammals *Nucleic Acids Res.* 46 (D1), D371–D374. doi:10.1093/nar/gkx1025

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., et al. (2019). Breast cancer statistics. *Ca. Cancer J. Clin.* 69 (6), 438–451. doi:10.3322/caac.21583

Duffy, M. J., Synnott, N. C., and Crown, J. (2018). Mutant p53 in breast cancer: Potential as a therapeutic target and biomarker. *Breast Cancer Res. Treat.* 170 (2), 213–219. doi:10.1007/s10549-018-4753-7

Fan, W., Shang, J., Li, F., Sun,, Y., and Liu, J. X. (2020). Idssim: An lncRNA functional similarity calculation model based on an improved disease semantic similarity method[J]. *BMC Bioinforma.* 21 (1), 1–14.

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34 (9), 1529–1537. doi:10.1093/bioinformatics/btx794

Garrido-Castro, A. C., Lin, N. U., and Polyak, K. (2019). Insights into molecular classifications of triple-negative breast cancer: Improving patient selection for treatment. *Cancer Discov.* 9 (2), 176–198. doi:10.1158/2159-8290.CD-18-1177

Gavgani, R. R., Babaei, E., Hosseinpourfeizi, M. A., Fakhrjou, A., and Montazeri, V. (2020). Study of long non-coding RNA highly upregulated in liver cancer (HULC) in breast cancer: A clinical & *in vitro* investigation. *Indian J. Med. Res.* 152 (3), 244–253. doi:10.4103/ijmr.IJMR_1823_18

Gooding, A. J., Zhang, B., Jahanbani, F. K., Gilmore, H. L., Chang, J. C., Valadkhan, S., et al. (2017). The lncRNA BORG drives breast cancer metastasis and disease recurrence. *Sci. Rep.* 7 (1), 1–18. doi:10.1038/s41598-017-12716-6

Guo, Z. H., You, Z. H., Wang, Y. B., Yi, H. C., and Chen, Z. H. (2019). A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. *IScience* 19, 786–795. doi:10.1016/j.isci.2019.08.030

Guo, Z., Hui, Y., Kong, F., and Lin, X. (2022). Finding lung-cancer-related lncRNAs based on laplacian regularized least squares with unbalanced Bi-random walk. *Front. Genet.* 13, 933009. doi:10.3389/fgene.2022.933009

Heer, E., Harper, A., Escandor, N., Sung, H., McCormack, V., and Fidler-Benaoudia, M. M. (2020). Global burden and trends in premenopausal and postmenopausal breast cancer: A population-based study. *Lancet. Glob. Health* 8 (8), e1027–e1037. doi:10.1016/S2214-109X(20)30215-1

Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2021). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief. Bioinform.* 22 (3), bbaa140. doi:10.1093/bib/bbaa140

Huang, Y. A., Chen, X., You, Z. H., Huang, D. S., and Chan, K. C. C. (2016). Ilncsim: Improved lncRNA functional similarity calculation model. *Oncotarget* 7 (18), 25902–25914. doi:10.18632/oncotarget.8296

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47 (D1), D1013–D1017. doi:10.1093/nar/gky1010

Key, T. J., Verkasalo, P. K., and Banks, E. (2001). Epidemiology of breast cancer. *Lancet. Oncol.* 2 (3), 133–140. doi:10.1016/S1470-2045(00)00254-0

Kong, J., Qiu, Y., Li, Y., Zhang, H., and Wang, W. (2019b). TGF-β1 elevates P-gp and BCRP in hepatocellular carcinoma through HOTAIR/miR-145 axis. *Biopharm. Drug Dispos.* 40 (2), 70–80. doi:10.1002/bdd.2172

Kong, X., Duan, Y., Sang, Y., Li, Y., Zhang, H., Liang, Y., et al. (2019a). LncRNA-CDC6 promotes breast cancer progression and function as ceRNA to target CDC6 by sponging microRNA-215. *J. Cell. Physiol.* 234 (6), 9105–9117. doi:10.1002/jcp.27587

Lan, W., Dong, Y., Chen, Q., Zheng, R., Liu, J., Pan, Y., et al. (2022). Kgancda: Predicting circRNA-disease associations based on knowledge graph attention network. *Brief. Bioinform.* 23 (1), bbab494. doi:10.1093/bib/bbab494

Lan, W., Lai, D., and Chen, Q. (2020). Ldicdl: LncRNA-disease association identification based on collaborative deep learning[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 1715–1723. doi:10.1109/TCBB.2020.3034910

Li, G., Luo, J., Liang, C., Xiao, Q., Ding, P., and Zhang, Y. (2019). Prediction of LncRNA-disease associations based on network consistency projection. *Ieee Access* 7, 58849–58856. doi:10.1109/access.2019.2914533

Liang, Y., Song, X., Li, Y., Chen, B., Zhao, W., Wang, L., et al. (2020). Retraction note to: LncRNA BCRT1 promotes breast cancer progression by targeting miR-1303/PTBP3 axis. *Mol. Cancer* 19 (1), 131–220. doi:10.1186/s12943-022-01576-y

Liang, Y., Wu, Y., and Zhang, Z. (2022a). Hyb4mC: A hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction[J]. *BMC Bioinforma.* 23 (1), 1–18. doi:10.1186/s12859-022-04789-6

Liang, Y., Zhang, Z. Q., Liu, N. N., Wu, Y. N., Gu, C. L., and Wang, Y. L. (2022b). Magcnse: Predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinforma.* 23 (1), 1–22. doi:10.1186/s12859-022-04715-w

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, T., Han, Z., Li, H., Zhu, Y., Sun, Z., and Zhu, A. (2018). LncRNA DLEU1 contributes to colorectal cancer progression via activation of KPNA3. *Mol. Cancer* 17 (1), 1–13. doi:10.1186/s12943-018-0873-2

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., et al. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nat. Commun.* 7 (1), 12791–12813. doi:10.1038/ncomms12791

Peng, L. H., Chen, Y. Q., Ma, N., and Chen, X. (2017). Narrmda: Negative-aware and rating-based recommendation algorithm for miRNA-disease association prediction. *Mol. Biosyst.* 13 (12), 2650–2659. doi:10.1039/c7mb00499k

Peng, L. H., Sun, C. N., Guan, N. N., Qiang, J., and Chen, X. (2018). Hnmda: Heterogeneous network-based miRNA-disease association prediction. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L. H., Tian, X. F., Shen, L., Kuang, M., Li, T. B., Tian, G., et al. (2020a). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11, 577387. doi:10.3389/fgene.2020.577387

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020b). Rnmfmda: A microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front. Microbiol.* 11, 592430. doi:10.3389/fmicb.2020.592430

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* doi:10.1109/TCBB.2021.3116232

Peng, L. H., Shen, L., Xu, J. L., Tian, X. F., Liu, F. X., Wang, J. J., et al. (2021b). Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures[J]. *Sci. Rep.* 11 (1), 1–11.

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022a). Cell-cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: Data resources and computational strategies. *Brief. Bioinform.* 23 (4), bbac234. doi:10.1093/bib/bbac234

Peng, L. H., Tan, J. W., Tian, X. F., and Zhou, L. Q. (2022b). EnANNDeep: An ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models[J]. *Interdiscip. Sci. Comput. Life Sci.*, 1–24. doi:10.1007/s12539-021-00483-y

Qiao, K., Ning, S., Wan, L., Wu, H., Wang, Q., Zhang, X., et al. (2019). LINC00673 is activated by YY1 and promotes the proliferation of breast cancer cells via the miR-515-5p/MARK4/Hippo signaling pathway. *J. Exp. Clin. Cancer Res.* 38 (1), 418–515. doi:10.1186/s13046-019-1421-7

Sarrafzadeh, S., Geranpayeh, L., and Ghafouri-Fard, S. (2017). Expression analysis of long non-coding PCAT-1in breast cancer. *Int. J. Hematol. Oncol. Stem Cell Res.* 11 (3), 185–191.

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., and Sharma, K. K. (2010). Various types and management of breast cancer: An overview. *J. Adv. Pharm. Technol. Res.* 1 (2), 109–126.

Shen, L., Liu, F. X., Huang, L., Liu, G. Y., Zhou, L. Q., and Peng, L. H. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Shi, F., Xiao, F., Ding, P., Qin, H., and Huang, R. (2016). Long noncoding RNA highly up-regulated in liver cancer predicts unfavorable outcome and regulates metastasis by MMPs in triple-negative breast cancer. *Arch. Med. Res.* 47 (6), 446–453. doi:10.1016/j.arcmed.2016.11.001

Shi, Z., Luo, Y., Zhu, M., Zhou, Y., Zheng, B., Wu, D., et al. (2019). Expression analysis of long non-coding RNA HAR1A and HAR1B in HBV-induced hepatocullular carcinoma in Chinese patients. *Lab. Med.* 50 (2), 150–157. doi:10.1093/labmed/lmy055

Shi, Z., Zhang, H., Jin, C., Quan, X., and Yin, Y. (2021). A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinforma.* 22 (1), 136–220. doi:10.1186/s12859-021-04073-z

Siegel, R., and Naishadhamjemal, D. A. (2013). Cancer statistics, 2013. *Ca. Cancer J. Clin.* 63, 11–30. doi:10.3322/caac.21166

Sledge, G. W., Mamounas, E. P., Hortobagyi, G. N., Burstein, H. J., Goodwin, P. J., and Wolff, A. C. (2014). Past, present, and future challenges in breast cancer treatment. *J. Clin. Oncol.* 32 (19), 1979–1986. doi:10.1200/JCO.2014.55.4139

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief. Bioinform.* 23 (4), bbac266. doi:10.1093/bib/bbac266

Sun, W., Wu, Y., Yu, X., Liu, Y., Song, H., Xia, T., et al. (2013). Decreased expression of long noncoding RNA AC096655.1-002 in gastric cancer and its clinical significance. *Tumour Biol.* 34 (5), 2697–2701. doi:10.1007/s13277-013-0821-0

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., et al. (2017). Risk factors and preventions of breast cancer. *Int. J. Biol. Sci.* 13 (11), 1387–1397. doi:10.7150/ijbs.21635

Tang, W., Lu, G., and Ji, Y. (2022). Long non-coding RNA PCAT1 sponges miR-134-3p to regulate PITX2 expression in breast cancer[J]. *Mol. Med. Rep.* 25 (3), 1–10.

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2021). Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front. Immunol.* 11, 3824. doi:10.3389/fimmu.2020.603615

Tian, X., Shen, L., Gao, P., Huang, L., Liu, G., Zhou, L., et al. (2022). Discovery of potential therapeutic drugs for COVID-19 through logistic matrix factorization with kernel diffusion. *Front. Microbiol.* 13, 13. doi:10.3389/fmicb.2022.740382

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.* 12 (6), 433–446. doi:10.1038/nrd4018

Waks, A. G., and Winer, E. P. (2019). Breast cancer treatment: A review. *Jama* 321 (3), 288–300. doi:10.1001/jama.2018.19323

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, J., Chen, X., Hu, H., Yao, M., Song, Y., Yang, A., et al. (2021b). PCAT-1 facilitates breast cancer progression via binding to RACK1 and enhancing oxygen-independent stability of HIF-1α. *Mol. Ther. - Nucleic Acids* 24, 310–324. doi:10.1016/j.omtn.2021.02.034

Wang, M. N., You, Z. H., Wang, L., Li, L. P., and Zheng, K. (2021a). Ldgrnmf: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* 424, 236–245. doi:10.1016/j.neucom.2020.02.062

Wang, N., Zhong, C., Fu, M., Li, L., Wang, F., Lv, P., et al. (2019). Long non-coding RNA HULC promotes the development of breast cancer through regulating LYPD1 expression by sponging miR-6754-5p. *Onco. Targets. Ther.* 12, 10671–10679. doi:10.2147/OTT.S226040

Wu, Y., Yang, X., and Chen, Z. (2019). m6A-induced lncRNA RP11 triggers the dissemination of colorectal cancer cells via upregulation of Zeb1[J]. *Mol. cancer* 18 (1), 1–16.

Xi, Y., and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol.* 14 (10), 101174. doi:10.1016/j.tranon.2021.101174

Xing, Z., Park, P. K., Lin, C., and Yang, L. (2015). LncRNA BCAR4 wires up signaling transduction in breast cancer. *RNA Biol.* 12 (7), 681–689. doi:10.1080/15476286.2015.1053687

Xu, W., Zhou, G., Wang, H., Liu, Y., Chen, B., Chen, W., et al. (2020). Circulating lncRNA SNHG11 as a novel biomarker for early diagnosis and prognosis of colorectal cancer. *Int. J. Cancer* 146 (10), 2901–2912. doi:10.1002/ijc.32747

Yang, J., Grünewald, S., and Wan, X. F. (2013). Quartet-net: A quartet-based method to reconstruct phylogenetic networks[J]. *Mol. Biol.* 30 (5), 1206–1217.

Yang, J., Peng, S., and Zhang, B. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases[J]. *Geroscience* 42 (1), 353–372.

Zhang, H., Jiang, L., Zhong, S., Li, J., Sun, D., Hou, J., et al. (2021b). The role of long non-coding RNAs in drug resistance of cancer. *Clin. Genet.* 99 (1), 84–92. doi:10.1111/cge.13800

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021a). Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscip. Sci. Comput. Life Sci.* 13 (3), 535–545. doi:10.1007/s12539-021-00458-z

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict MicroRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (2), 405–415. doi:10.1109/tcbb.2019.2931546

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). Sflln: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inf. Sci. (N. Y).* 497, 189–201. doi:10.1016/j.ins.2019.05.017

Zhao, Q., Yang, Y., Ren, G., and Fan, C. (2019). Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans. Nanobioscience* 18 (4), 578–584. doi:10.1109/TNB.2019.2922214

Zhou, L., Duan, Q., Tian, X., Tang, J., and Peng, L. H. (2021a). LPI-HyADBS: A hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification[J]. *BMC Bioinforma.* 22 (1), 1–31.

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021b). LPI-deepGBDT: A multiple-layer deep framework based on gradient boosting decision trees for lncRNA-protein interaction identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8