frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fgene.2022.933009

Finding Lung-Cancer-Related lncRNAs Based on Laplacian Regularized Least Squares With Unbalanced Bi-Random Walk

*Zhifeng Guo, Yan Hui, Fanlong Kong and Xiaoxi Lin**

Department of Oncology, Chifeng Municipal Hospital, Chifeng, China

Lung cancer is one of the leading causes of cancer-related deaths. Thus, it is important to find its biomarkers. Furthermore, there is an increasing number of studies reporting that long noncoding RNAs (lncRNAs) demonstrate dense linkages with multiple human complex diseases. Inferring new lncRNA-disease associations help to identify potential biomarkers for lung cancer and further understand its pathogenesis, design new drugs, and formulate individualized therapeutic options for lung cancer patients. This study developed a computational method (LDA-RLSURW) by integrating Laplacian regularized least squares and unbalanced bi-random walk to discover possible lncRNA biomarkers for lung cancer. First, the lncRNA and disease similarities were computed. Second, unbalanced bi-random walk was, respectively, applied to the lncRNA and disease networks to score associations between diseases and lncRNAs. Third, Laplacian regularized least squares were further used to compute the association probability between each lncRNA-disease pair based on the computed random walk scores. LDA-RLSURW was compared using 10 classical LDA prediction methods, and the best AUC value of 0.9027 on the lncRNADisease database was obtained. We found the top 30 lncRNAs associated with lung cancers and inferred that lncRNAs TUG1, PTENP1, and UCA1 may be biomarkers of lung neoplasms, non-small–cell lung cancer, and LUAD, respectively.

Keywords: lung cancer, lncRNA, biomarker, lncRNA-disease association, laplacian regularized least squares, unbalanced bi-random walk

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Guanghui Li,
East China Jiaotong University, China
JunLin Xu,
Hunan University, China

****Correspondence:***
Xiaoxi Lin
L18047666059@163.com

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

Received: *30 April 2022*
Accepted: *03 June 2022*
Published: *22 July 2022*

Citation:
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

1 INTRODUCTION

Cancers are posing threat for the health of humans (Yang et al., 2013; Liu et al., 2021). Lung cancer is the most common cancer worldwide and one of the leading causes of cancer-relevant deaths, and it has been so for many years. Thus, in 2008, the global statistical analysis demonstrated that approximately 1.6 million new lung cancer cases were diagnosed, and 1.4 million deaths were confirmed globally. In 2012, there were 1.8 million of new lung cancer diagnoses and 1.6 million deaths (de Groot et al., 2018; Howlader et al., 2020). In 2018, the number of new lung cancer cases exceeded 2 million and the number of deaths exceeded 1.7 million (Yuan et al., 2019). In the United States, approximately 234,000 cases of lung cancer were diagnosed the same year. This year, lung cancer diagnosis account for 14 and 13% of new cases in men and women, respectively. Estimation of mortality is 83,550 and 70,500 deaths in men and women, respectively. Lung

carcinoma is one of cancers with the lowest survival rate. It is usually not diagnosed until an advanced stage (de Groot et al., 2018; Howlader et al., 2020).

Despite the fast development of lung cancer therapy, high morbidity and mortality rates still pose a severe challenge for cancer researchers. The majority of patients with advanced-stage lung cancer have been ultimately poorly diagnosed. Thus, designing efficient therapy strategies is extremely important for lung cancer patients. However, existing techniques applied to diagnosis and therapies of lung cancer remain suboptimal. Thus, better strategies supplementing or replacing the existing techniques are urgent. Genome-wide association studies have found numerous genetic variants relevant to various cancers, one-third of which are densely linked to noncoding regions. The noncoding RNAs can be used as biomarkers of lung cancers. Therefore, accurate biomarker identification is urgently required to effectively diagnose lung cancer and boost the survival rate while decreasing its mortality and morbidity (Huang et al., 2017; Roointan et al., 2019; Yang et al., 2020).

Long noncoding RNAs (lncRNAs) are a type of noncoding RNAs that has over 200 nucleotides and post-transcriptional modifications including splicing, capping, and polyadenylation. lncRNAs can be used as a guide for protein-DNA interactions, protein-RNA interactions, and protein–protein interactions (Peng et al., 2020a). With the fast advancement of cancer genomics, many lncRNAs have been demonstrated to be aberrantly expressed in diverse cancers and play key action in the development of tumors through modulation of cancer-related signaling pathways. lncRNAs can regulate survival, metastasis, angiogenesis, and proliferation of tumor cells. Therefore, lncRNAs can be used as potential biomarkers and therapeutic targets in cancers by interacting with proteins (Chandra Gupta and Nandan Tripathi, 2017). For example, Peng et al. and her groups (Peng et al., 2021a; Zhou L. Q. et al., 2021; Peng et al., 2021b; Zhou L. et al., 2021; Tian et al., 2021; Peng et al., 2022) designed a series of state-of-the-art lncRNA-protein interaction prediction methods and significantly improved biomarker identification for various diseases. In addition, lncRNA SNHG14, BCRT1, DSCAM-AS1, MaTAR24, and HOTAIR have been validated to densely link to breast cancer (Niknafs et al., 2016; Dong et al., 2018; Chang et al., 2020; Liang et al., 2020; Yang et al., 2022; Xue et al., 2016). HOTAIR has been reported to be highly expressed in non-small–cell lung cancer (NSCLC) and affect NSCLC tumorigenesis and metastasis. In addition, many biomarkers (for example, CA125, NSE, CEA, VEGF, and EGFR (Khanmohammadi et al., 2020) have been validated to associate with lung cancer.

More importantly, many machine learning methods, especially deep-learning methods, have been applied to identify lncRNA biomarkers of various diseases through lncRNA-disease association prediction. Thus, Fan et al. (2022) designed an LDA prediction method (GCRFLDA) using the graph convolutional matrix completion. Ma Y (Ma, 2022) exploited a deep multi-network embedding-based LDA inference framework. Wu et al. (2021) integrated graph auto-encoder and random forest for LDA prediction. Sheng et al. (2021) developed an attentional multi-level representation encoding method to find new LDAs combining convolutional and variance autoencoders. Zhao et al. (2022) proposed a heterogeneous graph attention network-based LDA identification model. These methods significantly improved the LDA prediction.

With the development of single cell RNA sequencing technologies (Peng et al., 2020b), we can obtain numerous RNA data. These data can improve the analyses of RNA data, for example, SARS-CoV-2 (Xu et al., 2020; Li et al., 2021). By finding new lncRNA biomarkers, we can design corresponding therapeutic strategies for lung cancer based on drug repositioning (Peng et al., 2015; Liu et al., 2020; Meng et al., 2022; Shen et al., 2022).

Although experimental methods found a few biomarkers for lung cancer, they are time-consuming and waste of resources. Therefore, computational techniques have been exploited to infer potential biomarkers for lung cancer. However, the majority of computational approaches need to improve the inference performance. In this study, to analyze the diagnostic, prognostic, and therapeutical potential of lncRNAs in lung cancer patients, we exploit a computational model combining Laplacian regularized least square and unbalanced bi-random walk, LDA-RLSURW, to predict possible lncRNA biomarkers for lung cancer.

2 DATASETS

First, the lncRNA-disease association dataset was collected. The dataset can be obtained from the lncRNADisease database at http://www.cuilab.cn/lncrnadisease (Chen et al., 2012). We obtained 82 lncRNAs, 157 diseases, and 701 associations after excluding lncRNAs without record in the lncRNADisease database and diseases with inappropriate names or without MeSH tree numbers.

3 METHODS

This study developed an lncRNA-disease association prediction method LDA-RLSURW. First, LDA-RLSURW computed disease semantic similarity and lncRNA functional similarity. Second, LDA-RLSURW calculated the initial association probability of each lncRNA-disease pair using unbalanced bi-random walk based on disease similarity matrix and lncRNA similarity, respectively. In conclusion, the computed initial lncRNA-disease association probabilities were further updated Laplacian regularized least squares. The flowchart of LDA-RLSURW is presented in **Figure 1**.

3.1 Disease Semantic Similarity

Semantic similarity between diseases can be computed using the directed acyclic graph (DAGs) based on their MeSH descriptors (Fan et al., 2020). Given a disease A, let its DAG be represented as $DAG_A = \{T_A, E_A\}$, where T_A denotes the ancestor node set of A

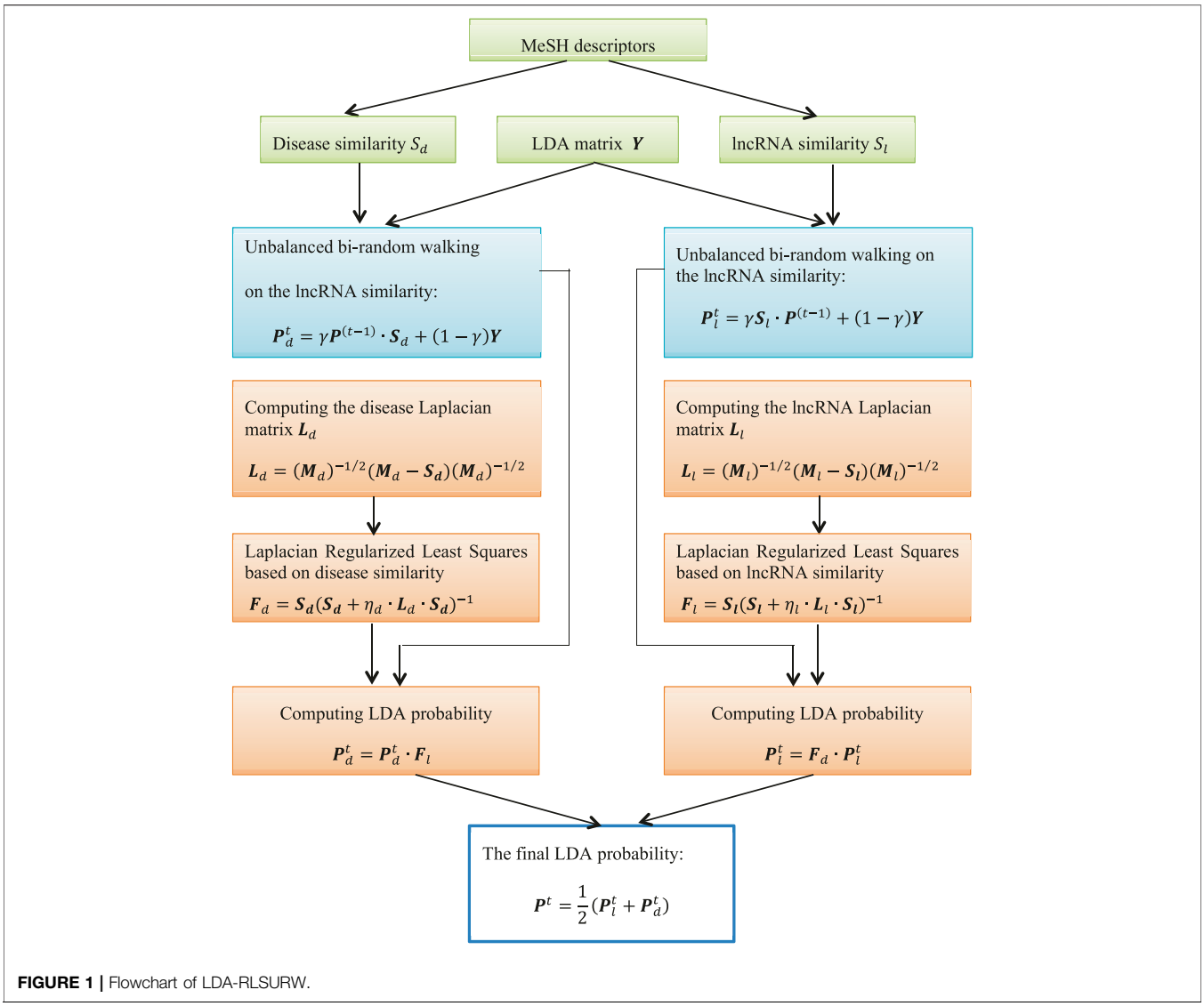

FIGURE 1 | Flowchart of LDA-RLSURW.

including A, and E_A denotes all edge set. For a disease term $t \in T_A$ in DAG_A, its semantic contribution to A can be computed by **Eq. 1** provided by LNCSIM1 (Chen et al., 2015):

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

where $C(t)$ denotes the children of t and α denotes a sematic contribution value of an edge linking t' to t in E_A.

In **Eq. 1**, we assume that terms at one identical layer from DAG_A have identical semantic contribution to A. However, when terms t_1 and t_2 are in the identical layer of DAG_A, and t_1 appears less than t_2 in DAG_A, the results from t_1 may be more specific than t_2. Thus, it could be more reasonable that $SV_A^1(t_1)$ is larger than $SV_A^1(t_2)$.

Considering this situation, we compute another semantic contribution value for disease A by **Eq. 2** provided by LNCSIM1 (Chen et al., 2015):

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

where D denotes the number of all diseases in the MeSH database and $Dags(t)$ denotes the number of DAG s, including the disease term t. In conclusion, the semantic contribution value of disease A in DAG_A can be computed by

$$SV_A^3(t) = \begin{cases} 1 & t = A \\ \max((\alpha + \beta)SV_A^3(t') | t' \in C(t) & t \neq A \end{cases}, \quad (3)$$

where β denotes the information content contribution factor, and

$$\beta = \frac{\max_{k \in K}(Dags(k)) - dags(t)}{D}, \quad (4)$$

where K denotes the disease set from the MeSH database.

Thus, the contribution of all diseases in DAG_A to A can be represented as

$$SV(A) = \sum_{t \in T_A} SV_A^3(t). \quad (5)$$

In summary, the semantic similarity between diseases A and B can be computed by **Eq. 6**:

$$S_d(A,B) = \frac{\sum_{t \in T_A \cap T_B} \left(SV_A^3(t) + SV_B^3(t)\right)}{SV(A) + SV(B)}. \quad (6)$$

3.2 lncRNA Functional Similarity

We calculate the lncRNA similarity using the approach provided by Fan et al. (2020). Assuming that *DG(u)*/*DG(v)* denotes diseases associated with lncRNA u/ v based on the LDA matrix, the lncRNA similarity between u and v was computed through semantic similarity between diseases involved in *DG(u)* and *DG(v)*. First, we construct a disease semantic similarity sub-matrix, where both rows and columns denote all diseases involved in *DG(u)*∪*DG(v)*, and the value of each element can be measured using the semantic similarity between corresponding diseases. Second, let d_u/ d_v denote one disease in *DG(u)*/*DG(v)*; the similarity between d_u/ d_v and *DG(v)*/*DG(u)* can be computed by **Eqs. 7** and **8**:

$$S(d_u, DG(v)) = \max_{d \in DG(v)} (S_d(d_u, d)), \quad (7)$$

$$S(d_v, DG(u)) = \max_{d \in DG(u)} (S_d(d_v, d)). \quad (8)$$

Third, the similarity between *DG(u)* to *DG(v)* and one between *DG(v)* to *DG(u)* can be calculated by **Eqs. 9** and **10**:

$$S_{u \to v} = \sum_{d \in DG(u)} S(d, DG(v)), \quad (9)$$

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

In conclusion, the similarity between two lncRNAs u and v can be computed by **Eq. 11**:

$$S_l(u,v) = \frac{S_{u \to v} + S_{v \to u}}{|DG(u)| + |DG(v)|}, \quad (11)$$

where $|DG(u)|/|DG(v)|$ indicates the number of diseases in $DG(u)/DG(v)$.

3.3 Unbalanced Bi-Random Walk

In this section, inspired by Shen et al. (2022), we consider that the lncRNA similarity network and the disease network and design an unbalance bi-random walk model to score lncRNA-disease pairs. The two networks exhibit different topological structures. Therefore, we use different optimal walking step sizes when randomly walking on these two networks. That is, we propose an unbalanced bi-random walk algorithm. First, we compute lncRNA-disease association scores by randomly walking with the maximal iteration number of n_l on the lncRNA network based on the lncRNA similarity by **Eq. 12**:

$$\boldsymbol{P}_l^t = \gamma \boldsymbol{S}_l \cdot \boldsymbol{P}^{(t-1)} + (1-\gamma)\boldsymbol{Y} \quad for \quad t = n_l. \quad (12)$$

In **Eq. 12**, at each step, the lncRNA similarity is fused with the random walk step by multiplying $\boldsymbol{S}_l$ on the left of the lncRNA-disease association probability matrix. $\gamma \in (0,1)$ is used to decrease the importance of circular bigraphs where the paths are longer during random walk and balance possible and known LDAs.

Second, we compute lncRNA-disease association scores by randomly walking with the maximal iteration number of n_d on the disease network based on the disease similarity by **Eq. 13**:

$$\boldsymbol{P}_d^t = \gamma \boldsymbol{P}^{(t-1)} \cdot \boldsymbol{S}_d + (1-\gamma)\boldsymbol{Y} \; for \; t = n_r. \quad (13)$$

In **Eq. 13**, at each step, disease similarity is fused with the random walk step by multiplying $\boldsymbol{S}_d$ on the right of the lncRNA-disease association probability matrix.

3.4 Laplacian Regularized Least Squares

In the last section, we compute the association probability for each lncRNA and disease using unbalanced bi-random walk method. However, for the algorithm, the jump condition is determined by known LDA data and the two similarity matrices. For a node n_i in an LDA network, if two other nodes n_j and n_k exhibit the same similarity with n_i, n_j and n_k may equally contribute to the jump. However, the node that has lower similarities with other nodes should have more contribution. Thus, we introduce Laplacian regularized least squares to solve the problem. First, the lncRNA Laplacian matrix $\boldsymbol{L}_l$ and the disease Laplacian matrix $\boldsymbol{L}_d$ are normalized to assess the jump probability for each node *via* **Eqs 14**, **15**.

$$\boldsymbol{L}_l = (\boldsymbol{M}_l)^{-1/2}(\boldsymbol{M}_l - \boldsymbol{S}_l)(\boldsymbol{M}_l)^{-1/2}, \quad (14)$$

$$\boldsymbol{L}_d = (\boldsymbol{M}_d)^{-1/2}(\boldsymbol{M}_d - \boldsymbol{S}_d)(\boldsymbol{M}_d)^{-1/2}, \quad (15)$$

where $\boldsymbol{M}_l/\boldsymbol{M}_d$ represent the diagonal matrices of lncRNAs/ diseases whose element $\boldsymbol{M}_l(\boldsymbol{i},\boldsymbol{i})/\boldsymbol{M}_d(\boldsymbol{j},\boldsymbol{j})$ denotes the summation of the i-th/ j-th row of $\boldsymbol{S}_l/\boldsymbol{S}_d$.

Second, to optimize the above minimum problems, the loss functions in the lncRNA and disease spaces are defined based on Laplacian matrices $\boldsymbol{L}_l$ and $\boldsymbol{L}_d$ *via* **Eqs. 11** and **12**, respectively:

$$\min_{F_l}\left[\|\boldsymbol{Y}^T - \boldsymbol{F}_l\|_F^2 + \eta_l \|\boldsymbol{F}_l \cdot \boldsymbol{L}_l \cdot (\boldsymbol{F}_l)^T\|_F^2\right], \quad (16)$$

$$\min_{F_d}\left[\|\boldsymbol{Y} - \boldsymbol{F}_d\|_F^2 + \eta_d \|\boldsymbol{F}_d \cdot \boldsymbol{L}_d \cdot (\boldsymbol{F}_d)^T\|_F^2\right], \quad (17)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $(\cdot)^T$ indicates the transpose, and η_v and η_d represent trade-off parameters. Models (11) and (12) can be solved *via* **Eqs. 13** and **14**, respectively:

$$\boldsymbol{F}_l^* = \boldsymbol{S}_l\left(\boldsymbol{S}_l + \eta_l \cdot \boldsymbol{L}_l \cdot \boldsymbol{S}_l\right)^{-1}\boldsymbol{Y}^T, \quad (18)$$

$$\boldsymbol{F}_d^* = \boldsymbol{S}_d\left(\boldsymbol{S}_d + \eta_d \cdot \boldsymbol{L}_d \cdot \boldsymbol{S}_d\right)^{-1}\boldsymbol{Y}. \quad (19)$$

To comprehensively detect the effect of unbalanced bi-random walk on the inference performance, we replace $\boldsymbol{Y}$ using LDA association probabilities computed by random walks. Assume that **Eqs. 20** and **21** can be defined as follows:

TABLE 1 | AUC values of LDA prediction methods on the lncRNADisease dataset.

	LNCSIM1/LNCSIM2	ILNCSIM	IDSSIM	RWRlncD	IIRWR
5-fold CV	0.8892/0.8881	0.8866	0.8966	0.6976	0.7781
	SIMCLDA	LRLSLDA	LLCPLDA	LDA-LNSUBRW	LDA-RLSURW
	0.7986	0.8174	0.8678	0.8874	0.9027

The LNCSIM1, LNCSIM2, LRLSLDA, and LDA-RLSURW are Laplacian regularized least square-based LDA methods, and the LDA-RLSURW can compute a better AUC. The results demonstrate that integrating unbalanced bi-random random walk can improve the performance. In addition, the IDSSIM and LDA-RLSURW computed the lncRNA similarity and disease similarity using the same method. The IDSSIM used the weighed K nearest known neighbor method to compute the lncRNA-disease association scores. The LDA-RLSURW outperforms IDSSIM, which show that the combination of Laplacian regularized least square and unbalanced bi-random walk can improve the LDA prediction performance compared to weighted K nearest known neighbor method. Both RWRlncD and IIRWR are random walk with restart-based LDA prediction methods. The SIMCLDA is an inductive matrix completion-based method. The LLCPLDA is a locality-constraint linear coding-based method. The LDA-RLSURW computes a better AUC than RWRlncD, IIRWR, SIMCLDA, and LLCPLDA, which further validates the powerful performance of LDA-RLSURW.

TABLE 2 | Inferred top 30 lncRNAs associated with LN.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	MINA	the MNDR database
2	HOTAIR	Known	17	PVT1	the MNDR database
3	MEG3	Known	18	**TUG1**	**Unconfirmed**
4	H19	Known	19	**PANDAR**	**Unconfirmed**
5	GAS5	Known	20	XIST	the MNDR database
6	UCA1	Known	21	**HULC**	**Unconfirmed**
7	CCAT2	Known	22	**HNF1A-AS1**	**Unconfirmed**
8	SPRY4-IT1	Known	23	**PTENP1**	**Unconfirmed**
9	CCAT1	Known	24	**KCNQ1OT1**	**Unconfirmed**
10	CDKN2B-AS1	Known	25	**HIF1A-AS2**	**Unconfirmed**
11	BANCR	Known	26	**DANCR**	**Unconfirmed**
12	BCYRN1	Known	27	**NPTN-IT1**	**Unconfirmed**
13	PCAT1	Known	28	**CRNDE**	**Unconfirmed**
14	SOX2-OT	Known	29	**CBR3-AS1**	**Unconfirmed**
15	CASC2	Known	30	**MIR31HG**	**Unconfirmed**

*The bold values denotes lncRNAs that were predicted to associate with LN and need to further validate in **Table 2**.*

$$\boldsymbol{F}_l = \boldsymbol{S}_l\left(\boldsymbol{S}_l + \eta_l \cdot \boldsymbol{L}_l \cdot \boldsymbol{S}_l\right)^{-1}, \tag{20}$$

$$\boldsymbol{F}_d = \boldsymbol{S}_d\left(\boldsymbol{S}_d + \eta_d \cdot \boldsymbol{L}_d \cdot \boldsymbol{S}_d\right)^{-1}. \tag{21}$$

At the t-th walking, **Eqs. 22** and **23** can be defined as

$$\boldsymbol{P}_l^t = \boldsymbol{F}_d \cdot \boldsymbol{P}_l^t, \tag{22}$$

$$\boldsymbol{P}_d^t = \boldsymbol{P}_d^t \cdot \boldsymbol{F}_l. \tag{23}$$

In conclusion, the LDA-RLSURW calculates the association score for each lncRNA-disease pair by combining association scores from the lncRNA and disease networks using **Eq. 24**:

$$\boldsymbol{P}^t = \frac{1}{2}\left(\boldsymbol{P}_l^t + \boldsymbol{P}_d^t\right). \tag{24}$$

4 EXPERIMENTS

4.1 Experimental Settings and Evaluation

The semantic contribution weight α is set as 0.5, the jump probability γ is set as 0.001, the maximal iteration number on the lncRNA network n_l is set as 31, the maximal iteration number on the disease network n_r is set as 1, and Laplacian regularized least square parameters η_l and η_d are set as 0.01. When the parameters are set as the above values, respectively, the LDA-RLSURW computes the best AUC on the lncRNADisease dataset. Therefore, we choose the parameters as the corresponding values. For other parameters, we set them as defaults provided by corresponding methods. The proposed LDA-RLSURW method and other comparative methods are evaluated using area under the receiver operating characteristic curve (AUC). Larger AUC values denote better performance.

4.2 Performance Comparison With Other Methods

To assess the performance of our proposed LDA-RLSURW method, we compare it with other 10 classical LDA prediction methods, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM (Fan W. et al., 2020). LNCSIM1 and LNCSIM2 measured the disease similarity separately using DAGs and the information content and computed association score for each lncRNA-disease pair by Laplacian regularized least squares. IDSSIM designed novel lncRNA functional similarity and disease semantic similarity computation approaches and computed the lncRNA-disease association scores using the computed similarity matrices and weighed K nearest known neighbor method. **Table 1** shows the AUC

TABLE 3 | Inferred top 30 lncRNAs associated with NSCLC.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	PANDAR	Known
2	HOTAIR	Known	17	HIF1A-AS1	Known
3	MEG3	Known	18	PCAT1	the MNDR database
4	GAS5	Known	19	CASC2	the MNDR database
5	H19	Known	20	SOX2-OT	the MNDR database
6	UCA1	Known	21	HULC	the MNDR database
7	CCAT2	Known	22	**MINA**	Unconfirmed
8	SPRY4-IT1	Known	23	**PTENP1**	Unconfirmed
9	CDKN2B-AS1	Known	24	HIF1A-AS2	the MNDR database
10	PVT1	Known	25	HNF1A-AS1	Known
11	CCAT1	Known	26	KCNQ1OT1	the MNDR database
12	TUG1	Known	27	CRNDE	the MNDR database
13	BANCR	Known	28	DANCR	the MNDR database
14	BCYRN1	Known	29	MIR31HG	the MNDR database
15	XIST	Known	30	NPTN-IT1	the MNDR database

*The bold values denotes lncRNAs that were predicted to associate with NSCLC and need to further validate in **Table 3**.*

values of LDA prediction methods on the lncRNADisease dataset. From **Table 1**, we can see that LDA-RLSURW computes the best AUC, which demonstrates the powerful LDA prediction performance of LDA-RLSURW.

4.3 Case Study

In this section, we conduct case studies to find potential lncRNA biomarkers for lung neoplasms, NSCLC, and adenocarcinoma of lung after confirming the performance of the proposed LDA-RLSURW method.

4.3.1 Finding Potential lncRNA Biomarkers for Lung Neoplasms

Lung neoplasms are one of the leading causes of death associated with malignant tumors in China (Khanmohammadi et al., 2020). Thus, Wang et al. (2020) investigated 14,528 lung cancer patients suffering from multiple primary malignant neoplasms (MPMN) and found 364 MPMN cases. In this section, we inferred the top 30 lncRNA biomarkers associated with lung neoplasms. The results are shown in **Table 2** and **Figure 2**. From **Table 2** and **Figure 2**, we can find that 15 lncRNAs are known to be associated with lung neoplasms in the lncRNADisease database, 3 lncRNAs (MINA, PVT1, and XIST) are unknown to be associated with lung neoplasms in the lncRNADisease database, which can be validated by the MNDR database (Cui et al., 2018). In addition, 12 lncRNAs are predicted to link to lung neoplasms and may be possible biomarkers of lung neoplasms.

More importantly, we predict that lncRNA taurine-upregulated gene 1 (TUG1) may be associated with lung neoplasms. TUG1 is one of lncRNAs that were first identified to associate with human disease. It is linked to diverse physiological processes, for example, gene regulation involved in translation, post-translation, transcription, and post-transcription. In this section, we infer that TUG1 may be the biomarker of lung neoplasms (Guo et al., 2020).

4.3.2 Finding Potential lncRNA Biomarkers for NSCLC

The NSCLC is a subtype of lung cancer. It is one of the leading causes of cancer death in the United States and accounts for 85% of lung cancers among all its subtypes. Although we have achieved important advancements in the NSCLC treatment, our understanding about the biology and mechanisms of NSCLC progression and early detection is still superficial. In this section, we aim to infer new lncRNA biomarkers for NSCLC after confirming the performance of LDA-RLSURW. The predicted top 30 lncRNAs associated with NSCLC are presented in **Table 3** and **Figure 3**. From **Table 3** and **Figure 3**, we can find that 18 lncRNAs associated with NSCLC are known in the lncRNADisease database, 10 lncRNAs associated with NSCLC have been validated in the MNDR database, and 2 lncRNAs (MINA and PTENP1) associated with NSCLC are unknown and require validation. The lncRNA PTENP1 has exerted the tumor-suppressive function through modulating PTEN expression in multiple malignancies. We predict that the

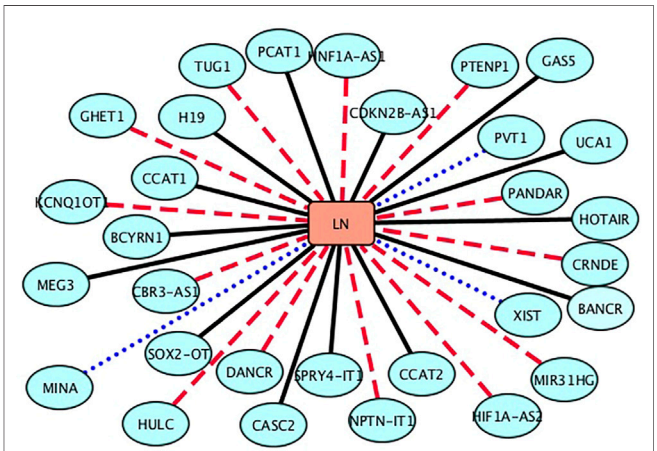

FIGURE 2 | Associations between the inferred top 30 lncRNAs and lung neoplasms (LN). Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of LN.

TABLE 4 | Inferred top 30 lncRNAs associated with LUAD.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	**XIST**	Unconfirmed
2	HOTAIR	Known	17	**PANDAR**	Unconfirmed
3	MEG3	Known	18	**BCYRN1**	Unconfirmed
4	GAS5	Known	19	**PCAT1**	Unconfirmed
5	CCAT1	Known	20	**HULC**	Unconfirmed
6	HNF1A-AS1	the MNDR database	21	**CASC2**	Unconfirmed
7	MIAT	Known	22	**SOX2-OT**	Unconfirmed
8	H19	the MNDR database	23	**PTENP1**	Unconfirmed
9	**UCA1**	Unconfirmed	24	**MINA**	Unconfirmed
10	**CDKN2B-AS1**	Unconfirmed	25	**CRNDE**	Unconfirmed
11	**PVT1**	Unconfirmed	26	**DANCR**	Unconfirmed
12	**TUG1**	Unconfirmed	27	**WT1-AS**	Unconfirmed
13	**CCAT2**	Unconfirmed	28	**KCNQ1OT1**	Unconfirmed
14	**SPRY4-IT1**	Unconfirmed	29	**NPTN-IT1**	Unconfirmed
15	**BANCR**	Unconfirmed	30	**CCDC26**	Unconfirmed

*The bold values denotes lncRNAs that were predicted to associate with LUAD and need to further validate in **Table 4**.*

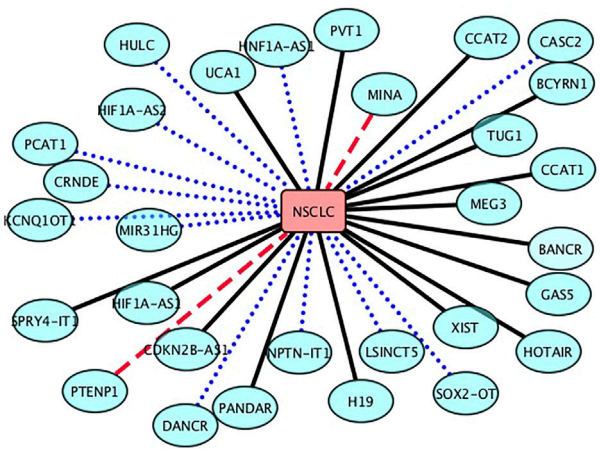

FIGURE 3 | Associations between the inferred top 30 lncRNAs and NSCLC. Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of LN.

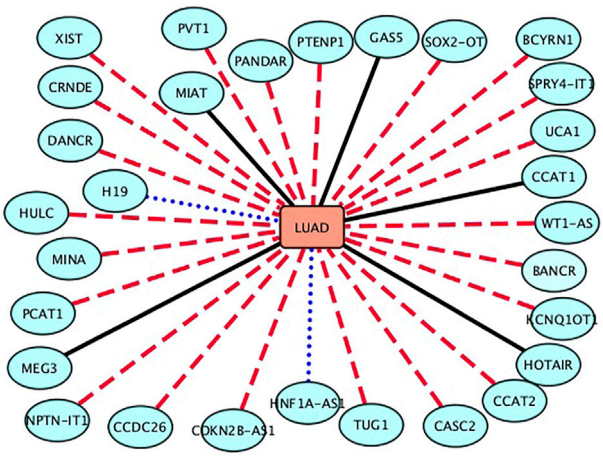

FIGURE 4 | Associations between the inferred top 30 lncRNAs and LUAD. Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of adenocarcinoma of lung.

PTENP1 may be a potential biomarker of NSCLC (Herbst et al., 2018; Arbour and Riely, 2019; Fan et al., 2020; Leighl et al., 2019).

4.3.3 Finding Potential lncRNA Biomarkers for Lung Adenocarcinoma

The NSCLC is divided into three main subtypes: lung squamous cell carcinoma, large-cell lung cancer, and lung adenocarcinoma (LUAD), among which lung squamous cell carcinoma and LUAD are the most prevalent. In this section, we predict possible lncRNAs associated with LUAD. The results are shown in **Table 4** and **Figure 4**. From **Table 4** and **Figure 4**, we can find that 6 lncRNAs are known to associate with LUAD, 2 lncRNAs are not known to associate with LUAD in the lncRNADisease database, although they are known in the MNDR database, and 22 lncRNAs have not been confirmed to associate with LUAD.

Urothelial carcinoma associated 1 (UCA1) is an oncogenic lncRNA. It is highly expressed in many cancers. UCA1 can bind to tumor-suppressive microRNAs, activate a few pivotal signaling pathways, and alter epigenetic and transcriptional regulation. More importantly, its high expression is linked to poor clinicopathological characteristics. In this section, we predict that UCA1 may associate with LUAD and require validation (Yao et al., 2019).

5 DISCUSSION

LNCSIM1 and LNCSIM2 obtained better performance improvements based on cross-validation and case analyses. However, LNCSIM1 cannot effectively distinguish the

semantic contributions of various disease terms from the identical layer. LNCSIM2 computed the IC values only through integrating DAG information. ILNCSIM is an edge-based prediction model. It combined the concept of information content and the hierarchical structure of DAGs to compute disease semantic similarity.

The RWRlncD conducted random walk with restart on the lncRNA similarity network. However, the RWRlncD cannot be used to predict associated information for diseases without any associated lncRNAs. The IRWRLDA improved random walk-based method through setting an initial probability vector to reduce the disadvantages of random walk with restart. The SIMCLDA used an inductive matrix completion model to complement missing LDA information. The LRLSLDA utilized Laplacian regularized least square model to predict LDAs. The LLCLPLDA first applied a locality-constraint linear coding model to project the local-constraint characteristics of lncRNAs and diseases, and then propagated LDAs by the initial LDA. The LDA-LNSUBRW used linear neighborhood similarity measurement and unbalanced bi-random walk algorithm to find possible LDAs.

The LDA-RLSURW obtains better performance for lncRNA-disease association prediction. It has three advantages: First, it utilizes the biological features to compute the lncRNA and disease similarity. Second, it uses unbalanced bi-random walk to compute the lncRNA-disease association probability. In conclusion, it further computes the lncRNA-disease association probability combining Laplacian regularized least squares.

6 CONCLUSION

Lung cancer is one of the most threatening cancer forms worldwide. In this study, we designed a computational method, LDA-RLSURW, to find possible lncRNA biomarkers for lung cancer. LDA-RLSURW effectively combines unbalanced bi-random walk and Laplacian regularized least square. We predict that TUG1, PTENP1, and UCA1 may be the biomarkers of lung neoplasms, NSCLC and LUAD, respectively.

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: ZG, YH, FK, and XL; methodology: ZG, YH, FK, and XL; project administration: XL; software: XL; writing original draft: ZG; writing review and editing: ZG and XL.

REFERENCES

Arbour, K. C., and Riely, G. J. (2019). Systemic Therapy for Locally Advanced and Metastatic Non-small Cell Lung Cancer. *Jama* 322 (8), 764–774. doi:10.1001/jama.2019.11058

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., Russo, S., Bhatia, S., et al. (2020). Matar25 Lncrna Regulates the Tensin1 Gene to Impact Breast Cancer Progression. *Nat. Commun.* 11, 1–19. doi:10.1038/s41467-020-20207-y

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., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA Functional Similarity Network Based on lncRNA-Disease Associations and Disease Semantic Similarity. *Sci. Rep.* 5 (1), 1–12. doi:10.1038/srep11338

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

de Groot, P. M., Wu, C. C., Carter, B. W., and Munden, R. F. (2018). The Epidemiology of Lung Cancer. *Transl. Lung Cancer Res.* 7 (3), 220–233. doi:10.21037/tlcr.2018.05.06

Dong, H., Wang, W., Chen, R., Zhang, Y., Zou, K., Ye, M., et al. (2018). Exosome-mediated Transfer of lncRNA-SNHG14 P-romotes T-rastuzumab C-hemoresistance in B-reast C-ancer. *Int. J. Oncol.* 53, 1013–1026. doi:10.3892/ijo.2018.4467

Fan W, W., Shang, J., Li, F., Sun, Y., Yuan, S., and Liu, J. X. (2020). IDSSIM: an lncRNA Functional Similarity Calculation Model Based on an Improved Disease Semantic Similarity Method. *BMC Bioinforma.* 21 (1), 1–14. doi:10.1186/s12859-020-03699-9

Fan, Y., Chen, M., and Pan, X. (2022). GCRFLDA: Scoring lncRNA-Disease Associations Using Graph Convolution Matrix Completion with Conditional Random Field. *Brief. Bioinform* 23 (1), bbab361. doi:10.1093/bib/bbab361

Guo, C., Qi, Y., Qu, J., Gai, L., Shi, Y., and Yuan, C. (2020). Pathophysiological Functions of the lncRNA TUG1. *Curr. Pharm. Des.* 26 (6), 688–700. doi:10.2174/1381612826666191227154009

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The Biology and Management of Non-small Cell Lung Cancer. *Nature* 553 (7689), 446–454. doi:10.1038/nature25183

Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., et al. (2020). The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N. Engl. J. Med.* 383 (7), 640–649. doi:10.1056/nejmoa1916623

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix Completion with Side Information and its Applications in Predicting the Antigenicity of Influenza Viruses. *Bioinformatics* 33 (20), 3195–3201. doi:10.1093/bioinformatics/btx390

Khanmohammadi, A., Aghaie, A., Vahedi, E., Qazvini, A., Ghanei, M., Afkhami, A., et al. (2020). Electrochemical Biosensors for the Detection of Lung Cancer Biomarkers: A Review. *Talanta* 206, 120251. doi:10.1016/j.talanta.2019.120251

Leighl, N. B., Page, R. D., Raymond, V. M., Daniel, D. B., Divers, S. G., Reckamp, K. L., et al. (2019). Clinical Utility of Comprehensive Cell-free DNA Analysis to Identify Genomic Biomarkers in Patients with Newly Diagnosed Metastatic Non-small Cell Lung Cancer. *Clin. Cancer Res.* 25 (15), 4691–4700. doi:10.1158/1078-0432.ccr-19-0624

Li, T., Huang, T., Guo, C., Wang, A., Shi, X., Mo, X., et al. (2021). Genomic Variation, Origin Tracing, and Vaccine Development of SARS-CoV-2: A Systematic Review. *Innovation* 2 (2), 100116. doi:10.1016/j.xinn.2021.100116

Liang, Y., Song, X., Li, Y., Chen, B., Zhao, W., Wang, L., et al. (2020). Lncrna Bcrt1 Promotes Breast Cancer Progression by Targeting Mir-1303/ptbp3 axis. *Mol. Cancer* 19, 85–20. doi:10.1186/s12943-020-01206-5

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, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor tissue-of-Origin[J]. *Front. Cell Dev. Biol.* 9, 886. doi:10.3389/fcell.2021.619330

Ma, Y. (2022). DeepMNE: Deep Multi-Network Embedding for lncRNA-Disease Association Prediction[J]. *IEEE J. Biomed. Health Inf.* 26 (7), 3539–3549. doi:10.1109/JBHI.2022.3152619

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Jang, J. (2022). A Weighted Bilinear Neural Collaborative Filtering Approach for Drug Repositioning[J]. *Briefings Bioinforma.* 23 (2), bbab581. doi:10.1093/bib/bbab581

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, 1–14. doi:10.1038/ncomms12791

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2015). Predicting Drug-Target Interactions with Multi-Information Fusion. *IEEE J. Biomed. Health Inf.* 21 (2), 561–572. doi:10.1109/JBHI.2015.2513200

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022). EnANNDeep: An Ensemble-Based lncRNA–Protein Interaction Prediction Framework with Adaptive K-Nearest Neighbor Classifier and Deep Models[J]. *Interdiscip. Sci. Comput. Life Sci.* 14, 209–232. doi:10.1007/s12539-021-00483-y

Peng, L. H., Wang, C., Tian, X. F., Zhou, L. Q., and Li, K. Q. (2021b). Finding lncRNA-Protein Interactions Based on Deep Learning with Dual-Net Neural Architecture[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2021, 3116232. doi:10.1109/TCBB.2021.3116232

Peng, L. H., Yuan, R. Y., Shen, L., Gao, P. F., and Zhou, L. Q. (2021a). LPI-EnEDT: an Ensemble Framework with Extra Tree and Decision Tree Classifiers for Imbalanced lncRNA-Protein Interaction Data Classification[J]. *BioData Min.* 14 (1), 1–22. doi:10.1186/s13040-021-00277-4

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2020a). Probing lncRNA-Protein Interactions: Data Repositories, Models, and Algorithms. *Front. Genet.* 10, 1346. doi:10.3389/fgene.2019.01346

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020b). Single-cell RNA-Seq Clustering: Datasets, Models, and Algorithms. *RNA Biol.* 17 (6), 765–783. doi:10.1080/15476286.2020.1728961

Roointan, A., Ahmad Mir, T., Wani, S. I., Mati-ur-Rehman, Hussain, K. K., Ahmed, B., et al. (2019). Early Detection of Lung Cancer Biomarkers through Biosensor Technology: A Review. *J. Pharm. Biomed. analysis* 164, 93–103. doi:10.1016/j.jpba.2018.10.017

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (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

Sheng, N., Cui, H., Zhang, T., and Xuan, P. (2021). Attentional Multi-Level Representation Encoding Based on Convolutional and Variance Autoencoders for lncRNA-Disease Association Prediction. *Brief. Bioinform* 22 (3), bbaa067. doi:10.1093/bib/bbaa067

Tian, X. F., Shen, L., Wang, Z. W., Zhou, L. Q., and Peng, L. H. (2021). A Novel lncRNA–Protein Interaction Prediction Method Based on Deep Forest with Cascade Forest Structure[J]. *Sci. Rep.* 11 (1), 1–15. doi:10.1038/s41598-021-98277-1

Wang, H., Hou, J., Zhang, G., et al. (2019). Clinical Characteristics and Prognostic Analysis of Multiple Primary Malignant neoplasms in patients with lung cancer [J]. *Cancer Gene Therapy* 26 (11), 419–426.

Wu, Q. W., Xia, J. F., Ni, J. C., and Zheng, C. H. (2021). GAERF: Predicting lncRNA-Disease Associations by Graph Auto-Encoder and Random Forest. *Brief. Bioinform* 22 (5), bbaa391. doi:10.1093/bib/bbaa391

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: an Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* 36 (10), 3139–3147. doi:10.1093/bioinformatics/btaa109

Xue, X., Yang, Y. A., Zhang, A., Fong, K.-W., Kim, J., Song, B., et al. (2016). LncRNA HOTAIR Enhances ER Signaling and Confers Tamoxifen Resistance in Breast Cancer. *Oncogene* 35 (21), 2746–2755. doi:10.1038/onc.2015.340

Yang, J., Grünewald, S., and Wan, X.-F. (2013). Quartet-Net: A Quartet-Based Method to Reconstruct Phylogenetic Networks. *Mol. Biol. Evol.* 30 (5), 1206–1217. doi:10.1093/molbev/mst040

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk from Histopathological Images and Clinical Information via Multimodal Deep Learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

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 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yao, F., Wang, Q., and Wu, Q. (2019). The Prognostic Value and Mechanisms of lncRNA UCA1 in Human Cancer. *Cancer Manag. Res.* 11, 7685–7696. doi:10.2147/cmar.s200436

Yuan, M., Huang, L. L., Chen, J. H., Wu, J., and Xu, Q. (2019). The Emerging Treatment Landscape of Targeted Therapy in Non-small-cell Lung Cancer. *Signal Transduct. Target Ther.* 4 (1), 61–14. doi:10.1038/s41392-019-0099-9

Zhao, X., Zhao, X., and Yin, M. (2022). Heterogeneous Graph Attention Network Based on Meta-Paths for lncRNA-Disease Association Prediction. *Brief. Bioinform* 23 (1), bbab407. doi:10.1093/bib/bbab407

Zhou, L. Q., Duan, Q., Tian, X. F., Tang, J. X., 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. doi:10.1186/s12859-021-04485-x

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

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

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