Finding Lung-Cancer-Related IncRNAs 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 (IncRNAs) demonstrate dense linkages with multiple human complex diseases. Inferring new IncRNA-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 IncRNA biomarkers for lung cancer. First, the IncRNA and disease similarities were computed. Second, unbalanced bi-random walk was, respectively, applied to the IncRNA and disease networks to score associations between diseases and IncRNAs. Third, Laplacian regularized least squares were further used to compute the association probability between each IncRNA-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 IncRNA-Disease database was obtained. We found the top 30 IncRNAs associated with lung cancers and inferred that IncRNAs TUG1, PTENP1, and UCA1 may be biomarkers of lung neoplasms, non-small-cell lung cancer, and LUAD, respectively.

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

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
cancer 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; Rooiintan 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.
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 LNC SIM1 (Chen et al., 2015):

$$SV_A^1(t) = \max(\alpha \times SV_A^1(t'), t \in C(t), t \neq A),$$  

where $C(t)$ denotes the children of $t$ and $\alpha$ denotes a semantic 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 LNC SIM1 (Chen et al., 2015):

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

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

$$SV_A^3(t) = \max((\alpha + \beta)SV_A^3(t'), t \in C(t), t \neq A),$$

where $\beta$ denotes the information content contribution factor, and

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

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
\[ SV(A) = \sum_{t\in T_A} SV_A^2(t). \] (5)

In summary, the semantic similarity between diseases \( A \) and \( B \) can be computed by Eq. 6:
\[ S_d(A, B) = \frac{\sum_{t\in T_A\cap T_B} (SV_A^2(t) + SV_B^2(t))}{SV(A) + SV(B)}. \] (6)

### 3.2 IncRNA Functional Similarity

We calculate the IncRNA similarity using the approach provided by Fan et al. (2020). Assuming that \( DG(u)/DG(v) \) denotes diseases associated with IncRNA \( u \)/\( v \) based on the LDA matrix, the IncRNA 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 \( DG(u)\cup DG(v) \), 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:
\[ S(d_u, DG(v)) = \max \{S_d(d_u, d_v)\}, \] (7)
\[ S(d_v, DG(u)) = \max \{S_d(d_v, d_u)\}. \] (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:
\[ S_{D\rightarrow D}(u, v) = \sum_{d\in DG(v)} S(d, DG(v)), \] (9)
\[ S_{D\rightarrow D}(v, u) = \sum_{d\in DG(u)} S(d, DG(u)). \] (10)

In conclusion, the similarity between two IncRNAs \( u \) and \( v \) can be computed by Eq. 11:
\[ S_l(u, v) = \frac{S_{D\rightarrow D}(u, v) + S_{D\rightarrow D}(v, u)}{|DG(u)| + |DG(v)|}, \] (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 IncRNA similarity network and the disease network and design an unbalanced bi-random walk model to score IncRNA-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 IncRNA-disease association scores by randomly walking with the maximal iteration number of \( n_t \) on the IncRNA network based on the IncRNA similarity by Eq. 12:
\[ P_f^t = \gamma S_l \cdot P^{(t-1)} + (1 - \gamma) Y \text{ for } t = n_t, \] (12)

In Eq. 12, at each step, the IncRNA similarity is fused with the random walk step by multiplying \( S_l \) on the left of the IncRNA-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 IncRNA-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:
\[ P_d^t = \gamma P^{(t-1)} \cdot S_d + (1 - \gamma) Y \text{ for } t = n_d. \] (13)

In Eq. 13, at each step, disease similarity is fused with the random walk step by multiplying \( S_d \) on the right of the IncRNA-disease association probability matrix.

### 3.4 Laplacian Regularized Least Squares

In the last section, we compute the association probability for each IncRNA 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 IncRNA Laplacian matrix \( L_i \) and the disease Laplacian matrix \( L_d \) are normalized to assess the jump probability for each node via Eqs. 14, 15.
\[ L_i = (M_i)^{-1/2} (M_i - S_i) (M_i)^{-1/2}, \] (14)
\[ L_d = (M_d)^{-1/2} (M_d - S_d) (M_d)^{-1/2}, \] (15)
where \( M_i/M_d \) represent the diagonal matrices of IncRNAs/diseases whose element \( M_i(i, i)/M_d(j, j) \) denotes the summation of the \( i \)-th/ \( j \)-th row of \( S_i/S_d \).

Second, to optimize the above minimum problems, the loss functions in the IncRNA and disease spaces are defined based on Laplacian matrices \( L_i \) and \( L_d \) via Eqs. 11 and 12, respectively:
\[ \min_{F} \left[ \|Y - F\|_F^2 + \eta_i \|F_i \cdot L_i \cdot (F_i)^T\|_F^2 \right], \] (16)
\[ \min_{F} \left[ \|Y - F\|_F^2 + \eta_d \|F_d \cdot L_d \cdot (F_d)^T\|_F^2 \right], \] (17)
where \( \| \cdot \|_F \) denotes the Frobenius norm, \( (\cdot)^T \) indicates the transpose, and \( \eta_i \) and \( \eta_d \) represent trade-off parameters. Models (11) and (12) can be solved via Eqs. 13 and 14, respectively:
\[ F_i = S_i(S_i + \eta_i \cdot L_i \cdot S_i)^{-1} Y, \] (18)
\[ F_d = S_d(S_d + \eta_d \cdot L_d \cdot S_d)^{-1} Y. \] (19)

To comprehensively detect the effect of unbalanced bi-random walk on the inference performance, we replace \( Y \) using LDA association probabilities computed by random walks. Assume that Eqs. 20 and 21 can be defined as follows:
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 weighted K nearest known neighbor method to compute the lncRNA-disease association scores. The LDA-RLSURW outperforms IDSSIM, which shows that the combination of Laplacian regularized least square and unbalanced bi-random 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 LRLSLDA is a locality-constraint linear coding-based method. The LDA-RLSURW computes a better AUC than RWRlncD, IIRWR, SIMCLDA, and LRLSLDA, which further validates the powerful performance of LDA-RLSURW.

**Table 1** AUC values of LDA prediction methods on the lncRNADisease dataset.

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

**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    | GASS        | 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   | BANCRI      | Known    | 26   | DANCR       | Unconfirmed |
| 12   | BCYRN1      | Known    | 27   | NPTN-IT1    | Unconfirmed |
| 13   | PCAT1       | Known    | 28   | CRNDE       | Unconfirmed |
| 14   | SOX2-OT     | Known    | 29   | CBRR3-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.

The semantic contribution weight \( \gamma \) is set as 0.5, the jump probability \( \alpha \) is set as 0.5, the maximal iteration number on the lncRNA network \( n_l \) is set as 31, the maximal iteration number on the disease network \( n_d \) is set as 1, and Laplacian regularized least square parameters \( \eta_l \) and \( \eta_d \) 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 scores 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 weighted K nearest known neighbor method. Table 1 shows the AUC...
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
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 IncRNA. 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. LNC SIM2 computed the IC values only through integrating DAG information. ILNC SIM 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 RWRRlncD conducted random walk with restart on the lncRNA similarity network. However, the RWRRlncD cannot be used to predict associated diseases for diseases without any associated lncRNAs. The IRWRLDRA 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 LLRLSLDA 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. J ama 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). LncRNA functional Similarity Network Based on lncRNA-Disease Associations and Disease Semantic Similarity. Sci. Rep. 5(1), 1–12. doi:10.1038/srep01338

Fan, Y., Chen, M., and Pan, X. (2022). GCRFDLA: 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:10.2174/138161282666191227154009

Herbst, R. S., Morgenstern, D., and Bosshoff, 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.jinn.2021.100116

Liang, Y., Song, L., 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/s12894-020-01206-5

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
