A Machine Learning Framework that Integrates Multi-omics Data Predicts Cancer-related LncRNAs

Lin Yuan
Qilu University of Technology (Shandong Academy of Sciences)
Jing Zhao
Qilu University of Technology (Shandong Academy of Sciences)
Tao Sun
Qilu University of Technology (Shandong Academy of Sciences)
Zhen Shen (wfxueyuan@126.com)
Nanyang Institute of Technology

Research Article

Keywords: LncRNA, multi-omics data, machine learning, neural network, node embedding, cancer

Posted Date: February 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-194877/v1

A Machine Learning Framework that Integrates Multi-omics Data Predicts Cancer-related LncRNAs

Lin Yuan[1], Jing Zhao[1], Tao Sun[1], and Zhen Shen[2]*

Abstract

Background: LncRNAs (Long non-coding RNAs) are a type of non-coding RNA molecule with transcript length longer than 200 nucleotides. LncRNA has been novel candidate biomarkers in cancer diagnosis and prognosis. However, it is difficult to discover the true association mechanism between lncRNAs and complex diseases. The unprecedented enrichment of multi-omics data and the rapid development of machine learning technology provide us with the opportunity to design a machine learning framework to study the relationship between lncRNAs and complex diseases.

Results: In this article, we proposed a new machine learning approach, namely LGDLDA (**L**ncRNA-**G**ene-**D**isease association networks based **L**ncRNA-**D**isease **A**ssociation prediction), for disease-related lncRNAs association prediction based multi-omics data, machine learning methods and neural network neighborhood information aggregation. Firstly, LGDLDA calculates the similarity matrix of lncRNA, gene and disease respectively. LGDLDA calculates the similarity between lncRNAs through the lncRNA expression profile matrix, lncRNA-miRNA interaction matrix and lncRNA-protein interaction matrix. LGDLDA obtains gene similarity matrix by calculating the lncRNA-gene association matrix and the gene-disease association matrix. LGDLDA obtains disease similarity matrix by calculating the disease ontology, the disease-miRNA association matrix, and Gaussian interaction profile kernel similarity. Secondly, LGDLDA integrates the neighborhood information in similarity matrices by using nonlinear feature learning of neural network. Thirdly, LGDLDA uses embedded node representations to approximate the observed matrices. Finally, LGDLDA ranks candidate lncRNA-disease pairs and then selects potential disease-related lncRNAs.

Conclusions: Compared with lncRNA-disease prediction methods, IHI-BMLLR takes into account more critical information and obtains the performance improvement cancer-related lncRNA predictions. Randomly split data experiment results show that the stability of LGDLDA is better than IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA. The results on different simulation data sets show that LGDLDA can accurately and effectively predict the disease-related lncRNAs. Furthermore, we applied LGDLDA to three real cancer data including gastric cancer, colorectal cancer and breast cancer to predict potential cancer-related lncRNAs.

Keywords: LncRNA, multi-omics data, machine learning, neural network, node embedding, cancer

Background

Long non-coding RNAs (lncRNAs) are a type of non-coding RNA molecule with transcript length longer than 200 nucleotides [1, 2]. Many studies have confirmed that the human genome contains massive amounts of lncRNA [3]. Many evidences indicate that lncRNAs regulate the expression level of genes at multiple levels (e.g., epigenetic regulation, genomic splicing, genomic imprinting, chromatin modification, transcriptional activation, transcriptional and post-transcriptional regulation) in the form of RNA [4-7]. The aberrant expression of lncRNA is involved in the proliferation, apoptosis, angiogenesis, and metastasis of tumors [8, 9]. LncRNA is closely related to the diagnosis, prognosis, and prevention and treatment of complex diseases [10]. LncRNA has become a new candidate biomarker for cancer diagnosis and prognosis [11].

The experimentally verified information about disease-related lncRNA is gradually increasing. A large number of databases have been published. The database LncRNADisease contains 3000 lncRNA-disease associations [12]. The database Lnc2Cancer has collected 1500 lncRNA-cancer entries [13]. Moreover, researchers have constructed lncRNA-related databases including NONCODE [14], lncRNAdb [15], LNCipedia [16], lncACTdb [17]. Although the research on lncRNA has progressed rapidly in recent years, the functions of most lncRNAs are still unclear. Bioinformatics calculation methods have been developed to predict the potential lncRNA-disease associations for biological experiment verifications. The calculation methods can greatly reduce the experimental cost and time for finding new disease-related lncRNAs [18, 19].

• Correspondence: wfxueyuan@126.com
[1]School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Daxue Road 3501, Jinan, Shandong 250353, China.
[2]School of Computer and Software, Nanyang Institute of Technology, Changjiang Road 80, Nanyang, Henan 473004, China.
Full list of author information is available at the end of the article.

The disease-related lncRNAs prediction methods can be categorized into network-based approaches and machine learning-based approaches. Biological system is a highly complex heterogenous network involved different molecules. Network-based approaches use multiple features including (but not limited to) lncRNA functional similarity, lncRNA-gene association, gene-gene interaction, gene-disease association, and molecular similarity to construct lncRNA similarity networks, or lncRNA-disease heterogeneous networks, then use network model analysis methods (e.g. propagation algorithms and random walk theory) to predict potential lncRNA-disease associations [20]. RWRlncD constructed a unified network including disease similarity network, lncRNA functional similarity network, and disease-lncRNA association network. The method used the Random Walk with Restart (RWR) method to predict the potential lncRNA-disease association [21]. RWRHLD added miRNA information that interacts with lncRNA, further improving the accuracy of the lncRNA-disease prediction method [22]. LncRDNetFlow used a streaming algorithm to predict lncRNA-disease associations based on multi-omics networks [23]. However, the known lncRNA-disease association data is still insufficient, and those methods cannot be applied to the prediction of related disease without any known lncRNAs information. To avoid the abovementioned problems, researchers attempting to combine known pathogenic gene-miRNA association data, miRNA-lncRNA association data and other data to predict lncRNA-disease association. LncPriCNet used multiple features, including phenotype-gene relations and gene-gene interactions, to construct a multi-level composite network and then used similarity scores to predict lncRNA-disease associations [24]. Ganegoda et al. proposed a model for predicting potential disease-associated lncRNAs by integrating known cancer-associated lncRNAs information and multi-omics data including genomic, regulatory, and transcriptional bios data [25].

Recently, many bioinformatics calculation models based on machine learning algorithms have been proposed to find potential lncRNA-disease associations. Lu et al. used inductive matrix completion and principal component analysis to predict potential lncRNA-disease associations [26]. Based on a review of existing research, Chen et al. proposed a hypothesis that functionally similar lncRNAs tend to be abnormally expressed in similar diseases, and developed a semi-supervised machine learning framework based on laplacian regularized least squares method (named LRLSLDA). Unfortunately, the method suffered from selecting multiple parameters effectively [27]. Wang et al. used lncRNA similarity data and disease similarity data to train a bagging support vector machine (SVM) classifier, and the trained SVM is implemented as a web server to predict potential disease-related lncRNAs [28]. You et al. proposed a method called LDASR to predict latent lncRNA-disease associations by using collaborative filtering and rotating forest [29]. These methods have achieved good results. Although the research on lncRNA has made rapid progress in recent years, unfortunately, these methods often used unmodified traditional machine learning methods, and the omics data used are limited to two or three types. Recently, the accumulation of associated omics data between lncRNAs and diseases and the development of machine learning and deep learning technologies provide researchers with better opportunities to use supervised learning models to predict disease-related lncRNAs.

Meanwhile, modern medical research proves that the alternations of genes may directly or indirectly affect diseases. Earlier studies have shown that RNA-protein interactions regulate gene expression by controlling various post-transcriptional processes. LncRNAs regulate the RNA-protein interactions by recruiting regulatory complexes [30, 31], and the literatures indicate that many lncRNAs also act as regulators to regulate gene expression [32]. Considering the mechanism of lncRNAs regulate genes, and genes regulate diseases provide a better opportunity for obtaining more information about lncRNA-disease associations.

Inspired by currently well-performing neural network technologies [33, 34], we tried to use multiple omics similarity matrices, neural network neighborhood information aggregation and trained supervised learning model to extract association features from lncRNA-gene-disease association network to predict disease-related lncRNAs. In this article, we proposed a new machine learning framework named LGDLDA (LncRNA-Gene-Disease association networks based LncRNA-Disease Association prediction) for disease-related lncRNAs association prediction based multi-omics functional similarity data, machine learning methods and neural network neighborhood information aggregation. Firstly, LGDLDA calculates the similarity between lncRNAs through the lncRNA expression profile matrix, lncRNA-miRNA interaction matrix and lncRNA-protein interaction matrix. The gene similarity matrix is obtained by calculating the lncRNA-gene association matrix and the gene-disease association matrix. The disease similarity matrix is obtained by calculating the disease ontology, the disease-miRNA association matrix, and Gaussian interaction profile kernel similarity. Secondly, LGDLDA integrates neighborhood information by using nonlinear feature learning of neural network. Thirdly, LGDLDA uses embedded node representations to approximate the observed matrices. Finally, LGDLDA ranks candidate lncRNA-disease pairs and then selects potential disease-related lncRNAs. The stability test results show that LGDLDA is more robust and the simulation data experiments show that LGDLDA performs better than four state-of-art methods in predicting lncRNA-disease association. LGDLDA can effectively predict potential cancer-related lncRNAs and provide more candidates for biological experimental verification. Most of predicted cancer-related lncRNAs are supported by

recent literatures.

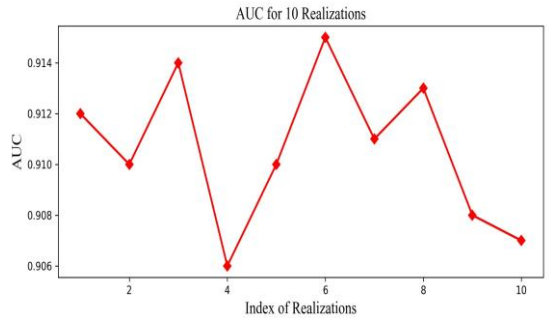

Fig. 1. The AUC values for 10 realizations.

Results

In the results section, the work we do is described as follows: Firstly, we used randomly split samples to observe the robustness of each method. Secondly, we compared LGDLDA with four famous lncRNA-disease association prediction methods on a small lncRNA-disease association simulation network. Four state-of-art methods include NCPLDA [35], IDHI-MIRW [36], LncDisAP [37] and NCPHLDA [38]. Finally, LGDLDA was applied to three real cancer samples to predict potential disease-related lncRNAs.

Comparison of method stability

Before comparing the performance of LGDLDA with four famous lncRNA-disease association prediction methods in small data, we need to evaluate the stability of these methods. We generally randomly divide the data set into two parts: Ω1 and Ω2. In the first step, based on the training set Ω1, we select different parameters and determine the parameter configuration with good performance. In the second step, we expect that the selected parameter configuration can have an accurate prediction in Ω2. There may be two issues to consider: (i) Does the randomness in the randomly divided sample affect the stability of the method?? (ii) Is the stability of LGDLDA better than NCPLDA [48], IDHI-MIRW [36], LncDisAP [37] and NCPHLDA [38] ? To address the two issues, we observed the performance of the method in two experiments. In the first experiment, we performed 10 random splits on a certain comprehensive data set. For each randomly divided data set, we ran LGDLDA on the data set and calculated AUC values. The AUC values for 10 realizations are shown in Fig. 1. The experimental results from Fig. 1 show that random partition strategy has little effect on the method performance. In the second experiment, we performed 50 random splits on a certain comprehensive data set. For each randomly divided data set, we ran ach method on the data set and calculated AUC values. Based on these AUC values, we calculated the minimum, first quartile, median, third quartile and maximum value and draw boxplots. The box plots from Fig. 2 show that the stability of LGDLDA is better than IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA.

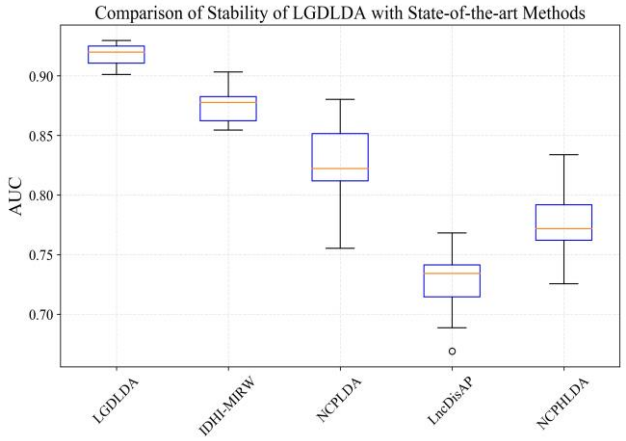

Fig. 2. The box plots of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA.

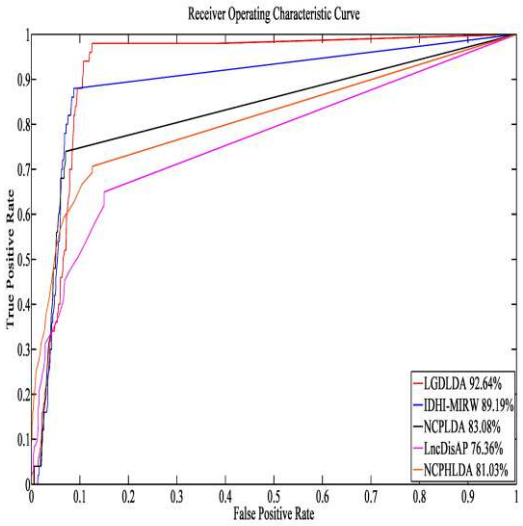

Fig. 3. The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on data that does not contain gene-related information.

Comparison with four state-of-art methods on a small simulation data set

In this section, we compared LGDLDA with four famous methods (i.e., NCPLDA, IDHI-MIRW, LncDisAP and NCPHLDA) on a small lncRNA-disease association simulation network which contains 356 lncRNAs, 354 diseases, 132 genes, 736 known lncRNA-gene associations, 462 gene-disease associations and 2169 known lncRNA-disease associations [36]. LncDisAP [37] and IDHI-MIRW [36] are prediction methods based on multiple biological datasets and RWR algorithm. NCPHLDA [38] and NCPLDA [35] are network-based methods. We performed these experiments on a computer with an Intel i9-10900X CPU and 512RAM.

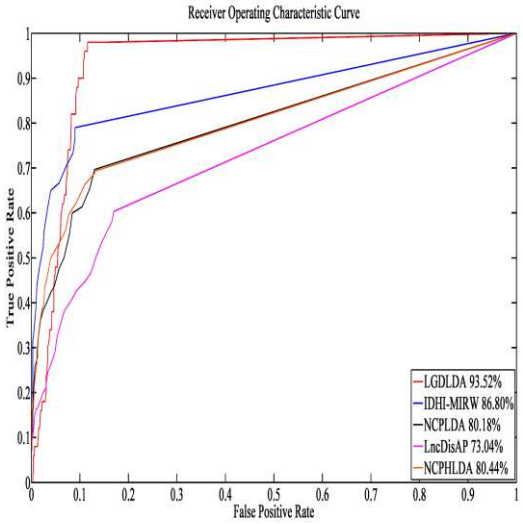

Fig. 4. The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on data containing gene information.

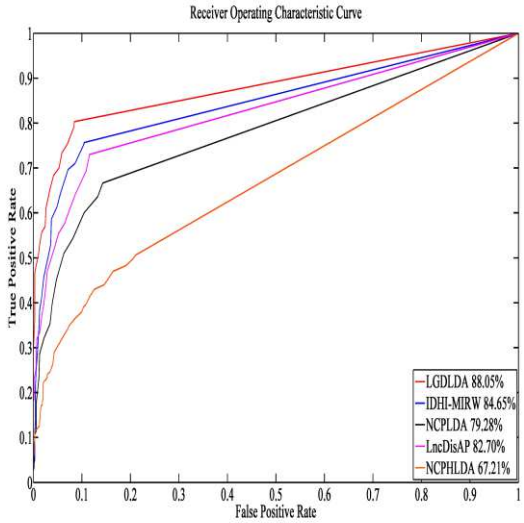

Fig. 5. The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on the data with missing part of the information.

To avoid the small lncRNA-disease association simulation network favoring our own model, we run each method on data that does not contain gene-related information (i.e., data without genes, lncRNA-gene associations, and gene-disease associations). Fig. 3 shows the ROCs and corresponding AUC values of LGDLDA and four competition methods. As shown in Fig. 3, LGDLDA outperformed other four methods in terms of AUC value. The AUC of LGDLDA is 0.9264, which is 0.03345, 0.0956, 0.1628 and 0.1161 higher than that of IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA, respectively. We also run each method on data containing gene information. Fig. 4 shows the ROCs and AUC values of LGDLDA and the four competition methods. As shown in Fig. 4, LGDLDA outperformed other four methods in terms of AUC value. The AUC of LGDLDA is 0.9352, which is 0.0672, 0.1334, 0.2048 and 0.1308 higher than that of IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA, respectively. Considering we often apply method to incomplete data set, we randomly remove 20% of the data and run each method. The ROCs and AUC values of LGDLDA and other four methods are shown in Fig. 5. LGDLDA achieved a better performance than other four methods in terms of AUC. The AUC of LGDLDA is 0.8805, which is 0.0340, 0.0877, 0.0535 and 0.2084 higher than that of IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA, respectively. Although our method LGDLDA is affected by incomplete data, it performs better than other four methods. Compared with the four state-of-art methods, the results on different simulation data sets show that LGDLDA can accurately and effectively predict the disease-related lncRNAs.

Application to cancer data and potential lncRNA-disease associations analysis

In this section, we applied LGDLDA to real cancer data including gastric cancer, colorectal cancer, and breast cancer. For a given disease, all known related lncRNAs are true labels, and other lncRNAs are candidates for disease. Inspired by the work of Guo et al. [29], we used the related information in the LncRNADisease database v2.0, DisGeNet, and LncACTdb to train LGDLDA, and other databases including CRlncRNA [39], MNDR v2.0, LncRNAwiki [40], and Lnc2Cancer, were used to verify the results. We applied the LGDLDA to real cancer data and ranked the lncRNA-disease association scores from large to small, and then identified the top 15 potentially relevant lncRNAs for each cancer.

Gastric cancer is the second most common cancer in the world [41, 42]. Accumulating evidence has demonstrated that many lncRNAs are dysregulated in gastric cancer [43, 44]. It is necessary to use computing methods to predict cancer-related lncRNAs. In the gastric cancer study, we used 1352 associations and gene related associations from databases as positive samples. We randomly selected the same number of samples from the database as negative samples. We constructed the test data set by extracting gastric cancer-related lncRNAs from other databases. Recent literatures supported 12 out of 15 potential gastric cancer-related lncRNAs. The confirmed databases and supporting literature of these 15 cancer-related lncRNAs are shown in Table 1 and Table 2, respectively. For example, Xu et al. [45] found that overexpression of ZFAS1 is significantly related to lymphatic metastasis and TNM staging. The overexpression of ZFAS1 leads to the loss of control of the cell cycle process, which in turn promotes the proliferation and migration of gastric cancer cells. Liu et al. reported that lncRNA H19 is aberrantly highly expressed in gastric cancer cell lines [46]. Zai et al. reported that activated DANCE promotes the proliferation and invasion of gastric cancer cells [47]. LncRNA HOXA11-AS promotes the invasion and

Table 1 The confirmed databases of Top 15 gastric cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Confirmed Database
1	UCA1	CRlncRNA/LncRNAWiki/Lnc2Cancer/LncRNADisease v2.0
2	NEHG1	Unconfirmed
3	TINCR	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
4	HOTAIR	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
5	C1RL-AS1	Unconfirmed
6	SPRY4-IT1	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
7	PVT1	CRlncRNA/Lnc2Cancer/LncRNADisease v2.0
8	NEAT1	LncRNAWiki/LncRNADisease v2.0/CRlncRNA
9	MEG3	LncRNAWiki/Lnc2Cancer/LncRNADisease v2.0
10	MALAT1	LncRNAWiki/Lnc2Cancer/LncRNADisease v2.0
11	DM1-AS	Unconfirmed
12	MIAT	CRlncRNA/LncRNADisease v2.0
13	GHET1	LncRNAWiki/Lnc2Cancer/LncRNADisease v2.0
14	FER1L4	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
15	SUMO1P3	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0

Table 2 The supporting literature of Top 15 gastric cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Verified	Literatures (PMID)
1	UCA1	Yes	28075173
2	NEHG1	No	without evidence
3	ZFAS1	Yes	28285404
4	HOTAIR	Yes	24949306
5	C1RL-AS1	No	without evidence
6	H19	Yes	24810858
7	PVT1	Yes	27756785
8	NEAT1	Yes	29363783
9	MEG3	Yes	26253106
10	MALAT1	Yes	28268166
11	DM1-AS	No	without evidence
12	DANCR	Yes	28951520
13	GHET1	Yes	24397586
14	FER1L4	Yes	24961353
15	HOXA11-AS	Yes	27651312

Table 3 The confirmed databases of Top 15 breast cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Confirmed Database
1	BCRT1	CRlncRNA
2	BORG	CRlncRNA/Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
3	MaTAR25	CRlncRNA
4	SPRY4-IT1	MNDRv2.0/Lnc2Cancer/LncRNADisease v2.0
5	PSORS1C3	Unconfirmed
6	MEG3	LncRNAWiki/Lnc2Cancer/LncRNADisease v2.0
7	PRNCR1	CRlncRNA/Lnc2Cancer
8	UCA1	LncRNADisease v2.0/Lnc2Cancer
9	PTCSC2	Unconfirmed
10	ANAC	Lnc2Cancer/CRlncRNA
11	UCC	Unconfirmed
12	SRA	CRlncRNA/LncRNADisease v2.0/Lnc2Cancer
13	XIST	Lnc2Cancer/LncRNADisease v2.0/CRlncRNA
14	YIYA	Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
15	LUCAT1	CRlncRNA/Lnc2Cancer

Table 4 The supporting literature of Top 15 breast cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Verified	Literatures (PMID)
1	BCRT1	Yes	31300015
2	BORG	Yes	28983112
3	MaTAR25	Yes	33353933
4	SPRY4-IT1	Yes	31736268
5	PSORS1C3	No	without evidence
6	MEG3	Yes	30793226
7	PRNCR1	Yes	31798697
8	UCA1	Yes	31695578
9	PTCSC2	No	without evidence
10	ANAC	Yes	29795261
11	UCC	No	without evidence
12	SRA	Yes	30238005
13	XIST	Yes	30026327
14	YIYA	Yes	29967256
15	LUCAT1	Yes	31300015

Table 5 The confirmed databases of Top 15 prostate cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Confirmed Database
1	PCA3	CRlncRNA/Lnc2Cancer/LncRNAWiki/LncRNADisease v2.0
2	TTTY15	CRlncRNA/Lnc2Cancer
3	HCP5	CRlncRNA/Lnc2Cancer
4	CCAT2	Lnc2Cancer/LncRNADisease v2.0
5	GAS5	CRlncRNA/Lnc2Cancer/LncRNADisease v2.0
6	MSC-AS1	Unconfirmed
7	LINP1	CRlncRNA/Lnc2Cancer
8	TUG1	CRlncRNA/Lnc2Cancer
9	ZFAS1	CRlncRNA/Lnc2Cancer
10	SLINKY	Unconfirmed
11	RNCR3	CRlncRNA/Lnc2Cancer
12	GLIDR	CRlncRNA/Lnc2Cancer
13	PCSEAT	CRlncRNA/Lnc2Cancer/LncRNAWiki
14	IRAIN	Unconfirmed
15	PCOTH	CRlncRNA/Lnc2Cancer/LncRNAWiki

proliferation of gastric cancer by regulating the chromatin modifiers LSD1 and DNMT1 [48]. A large number of studies have shown that LncRNA can be used as a biomarker for the treatment of gastric cancer [49].

Breast cancer is the most common malignant tumor in women and the second leading cause of cancer death [50, 51]. If we can detect cancer-related lncRNA as early as possible and intervene early, it will greatly reduce the incidence of breast cancer. Recent literatures supported 12 out of 15 potential breast cancer-related lncRNAs. The confirmed databases and supporting literature of these 15 cancer-related lncRNAs are shown in Table 3 and Table 4, respectively. For example, Yang et al. found that overexpression of LncRNA BCRT1 can promote the M2 polarization of macrophages, thereby accelerating the development of breast cancer [52]. Schiemann reported that lncRNA BORG regulates the transcriptional repressive activity of TRIM28 to trigger the migration and invasion of potential breast cancer cells [53]. Spector et al. reported that lncRNA MaTAR25 affects the proliferation and metastasis of breast cancer cells by regulating the expression of Tensin1 gene [54].

Prostate cancer is the second most common cancer in men and the fifth leading cause of death worldwide [55, 56]. Recent literatures supported 12 out of 15 potential prostate cancer-related lncRNAs. The confirmed databases and supporting literature of these 15 cancer-related lncRNAs are shown in Table 5 and Table 6, respectively. For example, Zhao et al. [57] reported that overexpression of ANRIL promoted the proliferation and migration of prostate cancer cells. Li et al. reported that lncRNA SNHG1 enhanced the expression of CDK7 and promoted cell proliferation in prostate cancer by negatively regulating miR-199a-3p [58]. Zhang et al. reported that the androgen-reduced transcript of LncRNA GAS5 can promote the proliferation of prostate cancer [59].

Table 6 The supporting literature of Top 15 prostate cancer-associated LncRNAs predicted by LGDLDA.

Rank	Name of LncRNA	Verified	Literatures (PMID)
1	PCA3	Yes	30016891
2	TTTY15	Yes	30527798
3	HCP5	Yes	31746434
4	CCAT2	Yes	32831916
5	GAS5	Yes	31673232
6	MSC-AS1	No	without evidence
7	LINP1	Yes	30058678
8	TUG1	Yes	30915735
9	ZFAS1	Yes	32104094
10	SLINKY	No	without evidence
11	RNCR3	Yes	RNCR3
12	GLIDR	Yes	28211799
13	PCSEAT	Yes	29803673
14	IRAIN	No	without evidence
15	PCOTH	Yes	15930275

Discussion

In case studies, we have found many potential cancer-related lncRNAs. Most of potential association lncRNAs are supported by recent literatures. In future biological experiments, it would be interesting to find the association mechanisms between new potential lncRNAs and diseases.

Finally, the real association mechanism between lncRNAs and disease is much more complicated than what we assumed. For example, the relationship between lncRNAs and complex diseases will change over time. We will try to design a new machine learning framework to analyze association data and time dynamic data simultaneously.

Conclusions

In this article, we proposed a novel machine learning framework, namely LGDLDA, to find cancer-related lncRNAs by integrating analysis of multi-omics data. Firstly, LGDLDA calculates the similarity matrix of lncRNA, gene and disease respectively. LGDLDA calculates the similarity between lncRNAs through the lncRNA expression profile matrix, lncRNA-miRNA interaction matrix and lncRNA-protein interaction matrix. LGDLDA obtains gene similarity matrix by calculating the lncRNA-gene association matrix and the gene-disease association matrix. LGDLDA obtains disease similarity matrix by calculating the disease ontology, the disease-miRNA association matrix, and Gaussian interaction profile kernel similarity. Secondly, LGDLDA integrates the neighborhood information in similarity matrices by using nonlinear feature learning of neural network. Thirdly, LGDLDA uses embedded node representations to approximate the observed matrices. Finally, LGDLDA ranks candidate lncRNA-disease pairs and then selects potential disease-related lncRNAs. LGDLDA incorporates the prior knowledge of biological network topology including lncRNA similarity networks, lncRNA-gene association network, gene-disease association network, disease semantic similarity networks, and lncRNA-disease association network. In this framework, a deep learning model was used to generate feature matrices. In model optimization, the final optimization problem is a popular matrix completion problem, which can be solved using convex optimization methods. In summary, the method takes into account more critical information and obtains the performance improvement cancer-related lncRNA predictions.

Methods and materials

Overview of LGDLDA

In this section, we will introduce the main steps in the LGDLDA method. (1) LGDLDA uses multiple association similarity matrices (including lncRNA functional similarities, gene-disease associations, disease similarities, lncRNA-disease associations, and lncRNA-gene associations matrix) to build lncRNA-gene-disease association network. (2) Based on the matrices generated in the first step, LGDLDA uses the association similarity matrices combined with neural network to calculate the neighborhood information of lncRNAs and diseases, and further embeds it into the low-dimensional spatial node representations. (3) Inspired by the reconstruction matrix algorithm in NNHLDA [33], LGDLDA uses low-dimensional spatial node representations to generate the projection matrices to approximate the observed matrices, and learns as much information in the original matrix as possible in the optimization of the loss function. (4) LGDLDA sorts the elements in the learned association matrix and selects the top values to predict disease-related lncRNAs. Fig. 6 shows the flowchart of LGDLDA method.

Datasets

In this paragraph, we will introduce the mathematical formulas used next. $S \in R^{m \times m}$ is used to represent the lncRNAs functional similarity matrix and $D \in R^{n \times n}$ is used to represent disease similarity matrix, where m and n denote the number of lncRNAs and diseases,

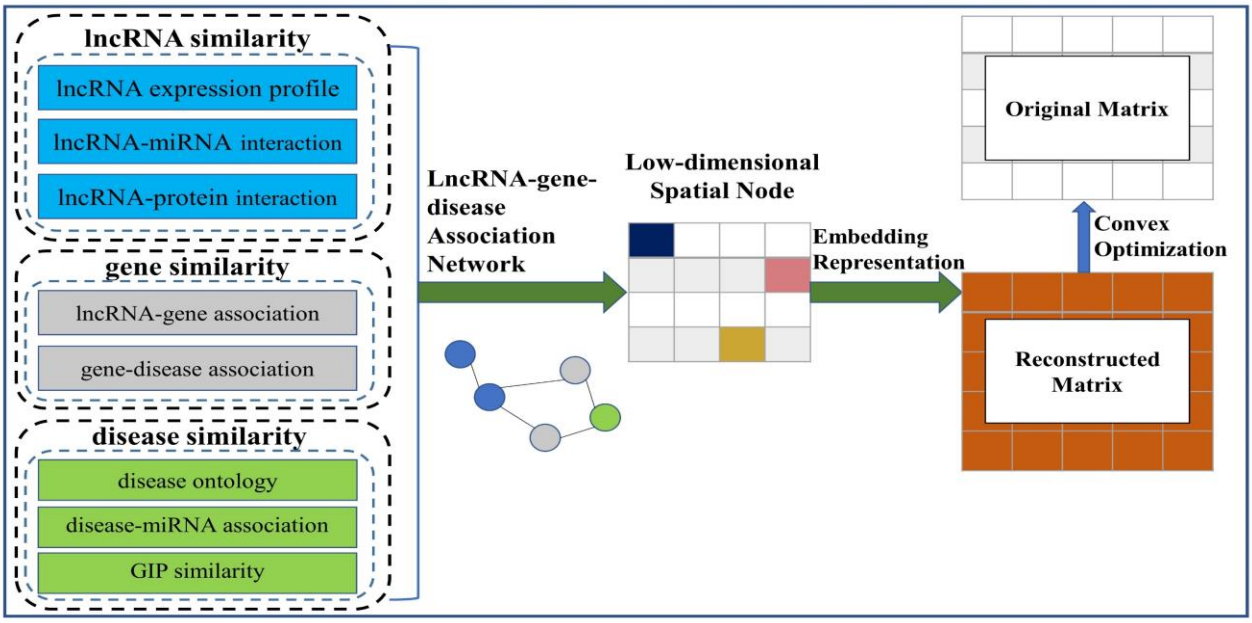

Fig. 6. The flowchart of LGDLDA. (1) LGDLDA uses multiple association similarity matrices to build lncRNA-gene-disease association network. (2) Based on the matrices generated in the first step, LGDLDA uses the association similarity matrices combined with neural network to calculate the neighborhood information of lncRNAs and diseases, and further embeds it into the low-dimensional spatial node representations. (3) LGDLDA uses embedded representations to generate the reconstructed matrix to approximate the original matrix, and learns as much information in the original matrix as possible in the optimization of the loss function. (4) LGDLDA sorts the elements in the learned association matrix and selects the top values to predict cancer-related lncRNAs.

respectively. $A \in R^{m \times n}$ represents lncRNA-disease association matrix, rows represent lncRNAs and columns are used to represent diseases. For each entry a_{ij} in A, the value of a_{ij} is equal to 1 if disease j related to lncRNA I; otherwise, a_{ij} is equal to 0. Let $A_{lg} \in R^{m \times k}$ be the lncRNA-gene association matrix and $A_{gd} \in R^{k \times n}$ represents the gene-disease association matrix, where k represents the number of genes.

For calculating the functional similarity networks of lncRNAs, LGDLDA uses the lncRNA expression profile matrix, lncRNA-protein function association matrix and lncRNA-miRNA association matrix. For calculating the disease similarity network, LGDLDA uses disease information, protein-disease association matrix and miRNA-disease association matrix. All lncRNAs and diseases are annotated with standard corresponding IDs.

Following the work of Zhang et al. on data collection [36], LGDLDA uses the LncRNA expression data from EMBL-EBI. LncRNA-miRNA and lncRNA-protein data come from three databases including starBase v2.0 [60], NPInter v3.0 [61], and RAID v2.0 [62]. Disease-miRNA association data and disease-gene association data come from HMDD v3.0 database [63] and DisGeNet database [64] respectively. LncRNA-disease association data come from LncRNADisease v2.0 [65], Lnc2Cancer [13], and MNDR v2.0 databases [66]. Gene-lncRNA association data are collected from LncACTdb [67].

Constructing lncRNA/disease similarity network

Since the Pearson correlation coefficient is easily affected by outliers, and outliers are inevitably included in the data, we used the biweight midcorrelation coefficient [68, 69]. We computed biweight midcorrelation coefficients between lncRNAs, and constructed the lncRNA similarity weighting network LncSm1.

$$u_i = \frac{x_i - med(\mathbf{x})}{\alpha \bullet mad(\mathbf{x})} \tag{1}$$

$$v_i = \frac{y_i - med(\mathbf{y})}{\alpha \bullet mad(\mathbf{y})} \tag{2}$$

$$mad(\mathbf{x}) = med(|x_i - med(\mathbf{x})|) \tag{3}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent lncRNA x expression vector and lncRNA y expression vector, respectively. x_i is the i-th value in $\mathbf{x}$ vector. $med(\mathbf{x})$ represents the median value of the vector $\mathbf{x}$, $mad(\mathbf{x})$ represents the median absolute deviation value. The weight w_i can be defined as follows:

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \tag{4}$$

If $(1-|u_i|)>0$, its value is equal to 1, otherwise equal to 0. For y_i, we can define a similar weight $w_i^{(y)}$. The parameter α is set to $(mad(\mathbf{x})+med(\mathbf{x}))/2$. The correlation coefficient can be defined as follows:

$$pre(\mathbf{x},\mathbf{y}) = \left(\sqrt{\sum_{i=1}^{m} \left[\left(x_i - med(\mathbf{x}) \right) \cdot w_i^{(\mathbf{x})} \right]^2} \cdot \sqrt{\sum_{i=1}^{m} \left[\left(y_i - med(\mathbf{y}) \right) \cdot w_i^{(\mathbf{y})} \right]^2} \right)^{-1} \quad (5)$$

$$BM(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \left(x_i - med(\mathbf{x}) \right) \cdot w_i^{(\mathbf{x})} \cdot \left(y_i - med(\mathbf{y}) \right) \cdot w_i^{(\mathbf{y})} \cdot pre(\mathbf{x},\mathbf{y}) \quad (6)$$

where m represents the number of elements in the vector. The range of BM value is from -1 to 1. The stronger the correlation, the larger the absolute value of BM.

The radial basis function (RBF) Gaussian kernel function was applied to lncRNA-miRNA interactions to obtain Gaussian interaction profile kernel similarity [70], and constructed the lncRNA similarity weighting network LncSm2. The similarity network can be defined as follows:

$$S_{lm}(i,j) = Exp(-\alpha_{l1} \left\| GIP_{lm}(l_i) - GIP_{lm}(l_j) \right\|^2) \quad (7)$$

$$\alpha_{l1} = \alpha_{l1}^{'} \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \left\| GIP_{lm}(l_i) \right\|^2 \right) \quad (8)$$

Where $GIP_{lm}(l_i)$ represents the lncRNA-miRNA interaction profile, $GIP_{lm}(l_i)$ is a binary vector in which 1 represents presence of interactions between lncRNA l_i and miRNA and 0 represents absence, α_l is the weight factor used to regulate the kernel bandwidth, the parameter $\alpha_l^{'}$ is set to 0.5 empirically and N_l denotes the total number of lncRNAs.

Analogous to lncRNA-miRNA interactions-based Gaussian similarity calculation method, the lncRNA-protein interactions-based Gaussian similarity of lncRNA pairs is calculated by the same method. Formula 7 is replaced by Formula 9:

$$S_{lp}(i,j) = Exp(-\alpha_{l2} \left\| GIP_{lp}(l_i) - GIP_{lp}(l_j) \right\|^2) \quad (9)$$

$$\alpha_{l2} = \alpha_{l2}^{'} \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \left\| GIP_{lp}(l_i) \right\|^2 \right) \quad (10)$$

where $GIP_{lp}(l_i)$ represents the lncRNA-protein interaction profile, $GIP_{lp}(l_i)$ is a binary vector in which 1 represents presence of interactions between lncRNA l_i and protein and 0 represents absence. α_d was used to control the bandwidth of kernel and $\alpha_d^{'}$ is set to 0.5 empirically. Through the above methods we can build a similarity network LncSm3.

We first used the R package "DOSE" to compute the correlation coefficients between diseases [71, 72]. Then, we can build a weighted disease similarity network DisSm1. We used disease-miRNA associations to calculate the kernel similarity of the Gaussian interaction spectrum between disease d_i and d_j, and then construct a weighted disease similarity correlation network DisSm2.

$$S_{dm}(i,j) = Exp(-\alpha_d \left\| GIP_{dm}(d_i) - GIP_{dm}(d_j) \right\|^2) \quad (11)$$

$$\alpha_d = \alpha_d^{'} \left(\frac{1}{N_d} \sum_{i=1}^{N_d} \left\| GIP_{dm}(d_i) \right\|^2 \right) \quad (12)$$

where $GIP_{dm}(d_i)$ denotes the disease-miRNA interaction profile, $GIP_{dm}(d_i)$ is a binary vector in which 1 denotes presence of interactions between disease d_i and miRNA and 0 represents absence. α_d is the weight factor used to regulate the kernel bandwidth, the parameter $\alpha_d^{'}$ is set to 0.5 empirically and N_d denotes the total types of diseases.

Constructing lncRNA/disease topological similarity networks

In order to overcome the loss of information caused by the fusion of similarity networks (i.e., LncSm1, LncSm2, and LncSm3 or DisSm1 and DisSm2), the idea of network diffusion is employed to generate the topological similarity networks. Motivated by the work of Zhang et al. [36], the RWR was applied to each similarity network to construct topological similarity network. RWR algorithm is a widely used complex biological network analysis method [36, 73, 74]. The RWR can be defined as follows:

$$M^{t+1} = (1-\alpha) M^t W + \alpha M^0 \quad (13)$$

$$W(i,j) = \frac{B(i,j)}{\sum_j B(i,j)} \quad (14)$$

where M^t represents the distribution probability matrix of one node visiting another node, and M^0 denotes the initial access probability distribution matrix. α represents the probability of restart. For lncRNA or disease, B denotes the edge-weighted adjacency matrix.

$$TS(i,j) = max(0, \log_2 \frac{M(i,j) \sum_i \sum_j M(i,j)}{\sum_i M(i,j) \sum_j M(i,j)}) \quad (15)$$

where matrix TS is an asymmetric matrix, in which we use the value $(TS(i,j)+TS(j,i))/2$ to represent the similarity value of the network topology between node i and node j. The values in the integration lncRNA network topological similarity matrix LTS can be obtained by calculating the average values of the corresponding position elements in the three lncRNA topological similarity matrices LTS_1, LTS_2, LTS_3 of LncSm1, LncSm2 and LncSm3. The values in the disease network topological similarity matrix DTS can be obtained by calculating the average values of the corresponding position elements in the two disease network topological similarity matrices DTS_1, DTS_2 of DisSm1, DisSm2. LTS represents the lncRNA similarity network LncTSN, and DTS represents the disease similarity network DisTSN.

Node embedding

For nodes representing lncRNA or disease in the heterogeneous network, its characteristic information can be summarized from the neighbor information related to it. For example, lncRNA's features can be aggregated from related lncRNAs, genes and diseases. The aggregation can be defined as follows:

$$lnce_i^{'} = concat(lnce_i, \sum_{j=1}^{m} LTS^{'}\{i,j\} \cdot \sigma_{ll}^{j} + \sum_{j=1}^{n} A^{'}\{i,j\} \cdot \sigma_{ld}^{j} + \sum_{j=1}^{k} A_{lg}^{'}\{i,j\} \cdot \sigma_{lg}^{j}) \quad (16)$$

$$dise_i^{'} = concat(dise_i, \sum_{j=1}^{n} DTS^{'}\{i,j\} \cdot \sigma_{dd}^{j} + \sum_{j=1}^{m} A^{T^{'}}\{i,j\} \cdot \sigma_{dl}^{j} + \sum_{j=1}^{k} A_{gd}^{T^{'}}\{i,j\} \cdot \sigma_{gd}^{j}) \quad (17)$$

$$gee_i^{'} = concat(gee_i, +\sum_{j=1}^{m} A_{lg}{}^{T^{'}}\{i,j\} \cdot \sigma_{lg}^{j} + \sum_{j=1}^{n} A_{gd}^{'}\{i,j\} \cdot \sigma_{gd}^{j}) \quad (18)$$

where $lnce_i^{'} \in R^{2d}$, $dise_i^{'} \in R^{2d}$ and $gee_i^{'} \in R^{2d}$ are the embeddings of $lncRNA_i$, $disease_i$ and gee_i, respectively. The initial representations of lncRNA, disease and gene nodes ($lnce_i \in R^d$, $dise_i \in R^d$ and $gee_i \in R^d$) are randomly set. By considering both node's neighbor information and its own features, we can obtain the network topology feature information of each node, and then calculate the feature vector of this node.

Motivated by the work of Zeng et al. [33], the activation function $\sigma[\cdot]$ (ReLU(x)=max(x,0)) can be defined as follows:

$$\sigma_{xy}^{j} = \sigma\left(\overline{ye_j} \cdot W_{xy} + b\right) \quad (19)$$

where W and b denotes the parameters in the neural networks. The nodes are embedded in low-dimensional vectors and normalized:

$$e_i^{''} = \frac{\sigma\left(e_i^{'} \cdot W_0 + b_0\right)}{\left\|\sigma\left(e_i^{'} \cdot W_0 + b_0\right)\right\|_2} \quad (20)$$

where $e_i^{''}$ stands for either $lnce_i^{''}$, $dise_i^{''}$ or $gee_i^{''}$. Thus, we used a single-layer neural network to non-linearly transform the nodes' representation and obtained a new embedding representation.

Training and evaluation

The information loss function between the reconstructed matrix and the original information matrix can be defined as follows:

$$\begin{aligned} \min_{W,b,E} & \sum\left(A\{i,j\} - lnce_i^{''} E_{ld1}^{i} E_{ld2}^{j\,T} dise_j^{''T}\right)^2 \\ & + \sum(LTS\{i,j\} - lnce_i^{''} E_{ll}^{i} E_{ll}^{jT} lnce_j^{''T})^2 \\ & + \sum(DTS\{i,j\} - dise_i^{''} E_{dd}^{i} E_{dd}^{j\,T} dise_j^{''T})^2 \\ & + \sum\left(A_{lg}\{i,j\} - lnce_i^{''} E_{lg1}^{i} E_{lg2}^{j\,T} gee_j^{''T}\right)^2 \\ & + \sum\left(A_{gd}\{i,j\} - gee_i^{''} E_{gd1}^{i} E_{gd2}^{j\,T} dise_j^{''T}\right)^2 \end{aligned} \quad (21)$$

where $E \in R^{p \times q}$ are the information mapping matrices, which can extract the main features of the nodes from the embedded node information representations. The matrix EE^T is used to enforce symmetry of the recovery.

Since the functions in the method are all differentiable, LGDLDA uses the gradient descent method to train the model parameters. After training, elements in the reconstruction matrix can predict each associations score. The higher a score is, the larger probability we suggest the potential association exists:

$$A\{i,j\}_{recovered} = lnce_i^{''} E_{ld1}^{i} E_{ld2}^{j\,T} dise_j^{''T} \quad (22)$$

In this sense, the final optimization problem is a popular matrix completion problem, which can be solved using convex optimization methods.

Evaluation method and metrics

To be able to fairly evaluate the performance of the methods, we performed LOOCV on the verified lncRNA-disease association data. Given a disease d_i, each known disease-related lncRNA is left out as test sample, meanwhile other disease-related lncRNAs are used as training samples. All irrelevant lncRNAs constitute candidate samples. The test samples are positive samples, and other samples are negative samples. In the predicted association matrix, LGALDA regards elements larger than the threshold as effective associations between lncRNAs and diseases. We used true positive rate (TPR) and false positive rate (FPR) to calculate area under the curve (AUC):

$$TPR = \frac{TP}{TP+FN} \quad (23)$$

$$FPR = \frac{FP}{FP+TN} \quad (24)$$

The change of threshold brings multiple TPRs and FPRs. Based on the values of TPR and FPR, we can calculate the AUC values and draw the receiver operating curve (ROC).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Availability of data and materials

The software of LGDLDA is available at https://github.com/nathanyl/LGDLDA, and the datasets

used are available from the corresponding references.

Competing interests
The authors declare no competing interests.

Funding
This work was supported by the National Key R&D Program of China (Grant nos. 2019YFB1404700, 2018AAA0100100), supported by the grant of National Natural Science Foundation of China (No. 62002189), supported by the grant of Natural Science Foundation of Shandong Province, China (No. ZR2020QF038), and partly supported by National Natural Science Foundation of China (Grant nos. 61861146002, 61732012, 61932008).

Authors' contributions
L.Y. conceived the method. L.Y. and Z.S. designed the method. L.Y. conducted the experiments and wrote the main manuscript text. J.Z. and T.S. prepared figures 1-3. All authors reviewed the manuscript.

Acknowledgements
Not applicable.

Authors' information
[1]School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Daxue Road 3501, Jinan, Shandong 250353, China. [2]School of Computer and Software, Nanyang Institute of Technology, Changjiang Road 80, Nanyang, Henan 473004, China.

References

[1] J. J. Quinn, and H. Y. Chang, "Unique features of long non-coding RNA biogenesis and function," *Nature Reviews Genetics,* vol. 17, no. 1, pp. 47, 2016.

[2] J. Jarroux, A. Morillon, and M. Pinskaya, "History, discovery, and classification of lncRNAs," *Long Non Coding RNA Biology*, pp. 1-46: Springer, 2017.

[3] F. Kopp, and J. T. Mendell, "Functional classification and experimental dissection of long noncoding RNAs," *Cell,* vol. 172, no. 3, pp. 393-407, 2018.

[4] B. Neve, N. Jonckheere, A. Vincent, and I. Van Seuningen, "Epigenetic regulation by lncRNAs: an overview focused on UCA1 in colorectal cancer," *Cancers,* vol. 10, no. 11, pp. 440, 2018.

[5] Y. Long, X. Wang, D. T. Youmans, and T. R. Cech, "How do lncRNAs regulate transcription?," *Science advances,* vol. 3, no. 9, pp. eaao2110, 2017.

[6] R.-Z. He, D.-X. Luo, and Y.-Y. Mo, "Emerging roles of lncRNAs in the post-transcriptional regulation in cancer," *Genes & diseases,* vol. 6, no. 1, pp. 6, 2019.

[7] C.-H. Zheng, L. Yuan, W. Sha, and Z.-L. Sun, "Gene differential coexpression analysis based on biweight correlation and maximum clique." p. S3.

[8] G. Botti, F. Collina, G. Scognamiglio, G. Aquino, M. Cerrone, G. Liguori, V. Gigantino, M. G. Malzone, and M. Cantile, "LncRNA HOTAIR polymorphisms association with cancer susceptibility in different tumor types," *Current drug targets,* vol. 19, no. 10, pp. 1220-1226, 2018.

[9] W.-X. Peng, P. Koirala, and Y.-Y. Mo, "LncRNA-mediated regulation of cell signaling in cancer," *Oncogene,* vol. 36, no. 41, pp. 5661-5667, 2017.

[10] V. Simion, S. Haemmig, and M. W. Feinberg, "LncRNAs in vascular biology and disease," *Vascular pharmacology,* vol. 114, pp. 145-156, 2019.

[11] L. Li, L. Wang, H. Li, X. Han, S. Chen, B. Yang, Z. Hu, H. Zhu, C. Cai, and J. Chen, "Characterization of LncRNA expression profile and identification of novel LncRNA biomarkers to diagnose coronary artery disease," *Atherosclerosis,* vol. 275, pp. 359-367, 2018.

[12] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic acids research,* vol. 41, no. D1, pp. D983-D986, 2012.

[13] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, and L. Wang, "Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers," *Nucleic acids research,* vol. 44, no. D1, pp. D980-D985, 2016.

[14] Y. Zhao, H. Li, S. Fang, Y. Kang, W. Wu, Y. Hao, Z. Li, D. Bu, N. Sun, and M. Q. Zhang, "NONCODE 2016: an informative and valuable data source of long non-coding RNAs," *Nucleic acids research,* vol. 44, no. D1, pp. D203-D208, 2016.

[15] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick, "lncRNAdb: a reference database for long noncoding RNAs," *Nucleic acids research,* vol. 39, no. suppl_1, pp. D146-D151, 2011.

[16] P.-J. Volders, K. Helsens, X. Wang, B. Menten, L. Martens, K. Gevaert, J. Vandesompele, and P.

Mestdagh, “LNCipedia: a database for annotated human lncRNA transcript sequences and structures,” *Nucleic acids research,* vol. 41, no. D1, pp. D246-D251, 2013.

[17] P. Wang, S. Ning, Y. Zhang, R. Li, J. Ye, Z. Zhao, H. Zhi, T. Wang, Z. Guo, and X. Li, “Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer,” *Nucleic acids research,* vol. 43, no. 7, pp. 3478-3489, 2015.

[18] B. Signal, B. S. Gloss, and M. E. Dinger, “Computational approaches for functional prediction and characterisation of long noncoding RNAs,” *Trends in Genetics,* vol. 32, no. 10, pp. 620-637, 2016.

[19] P.-J. Wei, D. Zhang, J. Xia, and C.-H. Zheng, “LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network,” *BMC bioinformatics,* vol. 17, no. 17, pp. 467, 2016.

[20] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, “MicroRNAs and complex diseases: from experimental results to computational models,” *Briefings in bioinformatics,* vol. 20, no. 2, pp. 515-539, 2019.

[21] J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu, and M. Zhou, “Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network,” *Molecular BioSystems,* vol. 10, no. 8, pp. 2074-2081, 2014.

[22] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou, and J. Sun, “Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network,” *Molecular BioSystems,* vol. 11, no. 3, pp. 760-769, 2015.

[23] J. Zhang, Z. Zhang, Z. Chen, and L. Deng, “Integrating multiple heterogeneous networks for novel lncRNA-disease association inference,” *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 16, no. 2, pp. 396-406, 2017.

[24] Q. Yao, L. Wu, J. Li, L. guang Yang, Y. Sun, Z. Li, S. He, F. Feng, H. Li, and Y. Li, “Global prioritizing disease candidate lncRNAs via a multi-level composite network,” *Scientific reports,* vol. 7, pp. 39516, 2017.

[25] G. U. Ganegoda, M. Li, W. Wang, and Q. Feng, “Heterogeneous network model to infer human disease-long intergenic non-coding RNA associations,” *IEEE transactions on nanobioscience,* vol. 14, no. 2, pp. 175-183, 2015.

[26] C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li, and J. Wang, “Prediction of lncRNA–disease associations based on inductive matrix completion,” *Bioinformatics,* vol. 34, no. 19, pp. 3357-3364, 2018.

[27] X. Chen, and G.-Y. Yan, “Novel human lncRNA–disease association inference based on lncRNA expression profiles,” *Bioinformatics,* vol. 29, no. 20, pp. 2617-2624, 2013.

[28] W. Lan, M. Li, K. Zhao, J. Liu, F.-X. Wu, Y. Pan, and J. Wang, “LDAP: a web server for lncRNA-disease association prediction,” *Bioinformatics,* vol. 33, no. 3, pp. 458-460, 2017.

[29] Z.-H. Guo, Z.-H. You, Y.-B. Wang, H.-C. Yi, and Z.-H. Chen, “A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest,” *iScience,* vol. 19, pp. 786-795, 2019.

[30] J. M. Engreitz, J. E. Haines, E. M. Perez, G. Munson, J. Chen, M. Kane, P. E. McDonel, M. Guttman, and E. S. Lander, “Local regulation of gene expression by lncRNA promoters, transcription and splicing,” *Nature,* vol. 539, no. 7629, pp. 452-455, 2016.

[31] K. C. Wang, Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, and R. A. Gupta, “A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression,” *Nature,* vol. 472, no. 7341, pp. 120-124, 2011.

[32] U. A. Ørom, T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, and Q. Huang, “Long noncoding RNAs with enhancer-like function in human cells,” *Cell,* vol. 143, no. 1, pp. 46-58, 2010.

[33] H. Chen, X. Wang, X. Zhang, X. Zeng, T. Song, and A. Rodríguez-Patón, "LncRNA-disease association prediction based on neighborhood information aggregation in neural network." pp. 175-178.

[34] L. Yuan, and D.-S. Huang, “A Network-guided Association Mapping Approach from DNA Methylation to Disease,” *Scientific reports,* vol. 9, no. 1, pp. 1-16, 2019.

[35] G. Li, J. Luo, C. Liang, Q. Xiao, P. Ding, and Y. Zhang, “Prediction of LncRNA-disease associations based on

network consistency projection," *IEEE Access,* vol. 7, pp. 58849-58856, 2019.

[36] X.-N. Fan, S.-W. Zhang, S.-Y. Zhang, K. Zhu, and S. Lu, "Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information," *BMC bioinformatics,* vol. 20, no. 1, pp. 87, 2019.

[37] Y. Wang, L. Juan, J. Peng, T. Zang, and Y. Wang, "LncDisAP: a computation model for LncRNA-disease association prediction based on multiple biological datasets," *BMC bioinformatics,* vol. 20, no. 16, pp. 1-11, 2019.

[38] H. Zhang, Y. Liang, C. Peng, S. Han, W. Du, and Y. Li, "Predicting lncRNA-disease associations using network topological similarity based on deep mining heterogeneous networks," *Mathematical biosciences,* vol. 315, pp. 108229, 2019.

[39] J. Wang, X. Zhang, W. Chen, J. Li, and C. Liu, "CRlncRNA: a manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features," *BMC medical genomics,* vol. 11, no. 6, pp. 114, 2018.

[40] L. Ma, A. Li, D. Zou, X. Xu, L. Xia, J. Yu, V. B. Bajic, and Z. Zhang, "LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs," *Nucleic acids research,* vol. 43, no. D1, pp. D187-D192, 2015.

[41] C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, no. 7418, pp. 61, 2012.

[42] L. Yuan, C.-H. Zheng, J.-F. Xia, and D.-S. Huang, "Module based differential coexpression analysis method for type 2 diabetes," *BioMed research international,* vol. 2015, 2015.

[43] X.-y. Fang, H.-f. Pan, R.-x. Leng, and D.-q. Ye, "Long noncoding RNAs: novel insights into gastric cancer," *Cancer letters,* vol. 356, no. 2, pp. 357-366, 2015.

[44] L. Yuan, L. Zhu, W.-L. Guo, X. Zhou, Y. Zhang, Z. Huang, and D.-S. Huang, "Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 14, no. 5, pp. 1154-1164, 2016.

[45] L. Pan, W. Liang, M. Fu, Z.-h. Huang, X. Li, W. Zhang, P. Zhang, H. Qian, P.-c. Jiang, and W.-r. Xu, "Exosomes-mediated transfer of long noncoding RNA ZFAS1 promotes gastric cancer progression," *Journal of cancer research and clinical oncology,* vol. 143, no. 6, pp. 991-1004, 2017.

[46] H. Li, B. Yu, J. Li, L. Su, M. Yan, Z. Zhu, and B. Liu, "Overexpression of lncRNA H19 enhances carcinogenesis and metastasis of gastric cancer," *Oncotarget,* vol. 5, no. 8, pp. 2318, 2014.

[47] Z. Mao, H. Li, B. Du, K. Cui, Y. Xing, X. Zhao, and S. Zai, "LncRNA DANCR promotes migration and invasion through suppression of lncRNA-LET in gastric cancer cells," *Bioscience reports,* vol. 37, no. 6, 2017.

[48] M. Sun, F. Nie, Y. Wang, Z. Zhang, J. Hou, D. He, M. Xie, L. Xu, W. De, and Z. Wang, "LncRNA HOXA11-AS promotes proliferation and invasion of gastric cancer by scaffolding the chromatin modification factors PRC2, LSD1, and DNMT1," *Cancer research,* vol. 76, no. 21, pp. 6299-6310, 2016.

[49] H. Liu, Z. Zhang, N. Wu, H. Guo, H. Zhang, D. Fan, Y. Nie, and Y. Liu, "Integrative analysis of dysregulated lncRNA-associated ceRNA network reveals functional lncRNAs in gastric cancer," *Genes,* vol. 9, no. 6, pp. 303, 2018.

[50] V. G. Vogel, "Epidemiology of breast cancer," *The breast*, pp. 207-218. e4: Elsevier, 2018.

[51] S.-G. Ge, J. Xia, W. Sha, and C.-H. Zheng, "Cancer subtype discovery based on integrative model of multigenomic data," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 14, no. 5, pp. 1115-1121, 2016.

[52] Y. Liang, X. Song, Y. Li, B. Chen, W. Zhao, L. Wang, H. Zhang, Y. Liu, D. Han, and N. Zhang, "LncRNA BCRT1 promotes breast cancer progression by targeting miR-1303/PTBP3 axis," *Molecular cancer,* vol. 19, pp. 1-20, 2020.

[53] A. J. Gooding, B. Zhang, F. K. Jahanbani, H. L. Gilmore, J. C. Chang, S. Valadkhan, and W. P. Schiemann, "The lncRNA BORG drives breast cancer metastasis and disease recurrence," *Scientific reports,* vol. 7, no. 1, pp. 1-18, 2017.

[54] K.-C. Chang, S. D. Diermeier, T. Y. Allen, L. D. Brine, S. Russo, S. Bhatia, H. Alsudani, K. Kostroff, T. Bhuiya, and E. Brogi, "MaTAR25 lncRNA regulates the Tensin1 gene to impact breast cancer progression," *Nature communications,* vol. 11, no. 1, pp. 1-19,

2020.

[55] P. Rawla, "Epidemiology of prostate cancer," *World journal of oncology,* vol. 10, no. 2, pp. 63, 2019.

[56] L. Yuan, C.-A. Yuan, and D.-S. Huang, "FAACOSE: A fast adaptive ant colony optimization algorithm for detecting SNP epistasis," *Complexity,* vol. 2017, 2017.

[57] B. Zhao, Y.-L. Lu, Y. Yang, L.-B. Hu, Y. Bai, R.-Q. Li, G.-Y. Zhang, J. Li, C.-W. Bi, and L.-B. Yang, "Overexpression of lncRNA ANRIL promoted the proliferation and migration of prostate cancer cells via regulating let-7a/TGF-β1/Smad signaling pathway," *Cancer Biomarkers,* vol. 21, no. 3, pp. 613-620, 2018.

[58] J. Li, Z. Zhang, L. Xiong, C. Guo, T. Jiang, L. Zeng, G. Li, and J. Wang, "SNHG1 lncRNA negatively regulates miR-199a-3p to enhance CDK7 expression and promote cell proliferation in prostate cancer," *Biochemical and biophysical research communications,* vol. 487, no. 1, pp. 146-152, 2017.

[59] Y. Zhang, X. Su, Z. Kong, F. Fu, P. Zhang, D. Wang, H. Wu, X. Wan, and Y. Li, "An androgen reduced transcript of LncRNA GAS5 promoted prostate cancer proliferation," *PLoS One,* vol. 12, no. 8, pp. e0182305, 2017.

[60] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data," *Nucleic acids research,* vol. 42, no. D1, pp. D92-D97, 2014.

[61] Y. Hao, W. Wu, H. Li, J. Yuan, J. Luo, Y. Zhao, and R. Chen, "NPInter v3. 0: an upgraded database of noncoding RNA-associated interactions," *Database,* vol. 2016, 2016.

[62] Y. Yi, Y. Zhao, C. Li, L. Zhang, H. Huang, Y. Li, L. Liu, P. Hou, T. Cui, and P. Tan, "RAID v2. 0: an updated resource of RNA-associated interactions across organisms," *Nucleic acids research,* vol. 45, no. D1, pp. D115-D118, 2017.

[63] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, "HMDD v3. 0: a database for experimentally supported human microRNA–disease associations," *Nucleic acids research,* vol. 47, no. D1, pp. D1013-D1017, 2019.

[64] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, pp. gkw943, 2016.

[65] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, "LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases," *Nucleic acids research,* vol. 47, no. D1, pp. D1034-D1037, 2019.

[66] T. Cui, L. Zhang, Y. Huang, Y. Yi, P. Tan, Y. Zhao, Y. Hu, L. Xu, E. Li, and D. Wang, "MNDR v2. 0: an updated resource of ncRNA–disease associations in mammals," *Nucleic acids research,* vol. 46, no. D1, pp. D371-D374, 2018.

[67] P. Wang, X. Li, Y. Gao, Q. Guo, Y. Wang, Y. Fang, X. Ma, H. Zhi, D. Zhou, and W. Shen, "LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low-and high-throughput experiments," *Nucleic acids research,* vol. 47, no. D1, pp. D121-D127, 2019.

[68] P. Langfelder, and S. Horvath, "Fast R functions for robust correlations and hierarchical clustering," *Journal of statistical software,* vol. 46, no. 11, 2012.

[69] L. Yuan, L.-H. Guo, C.-A. Yuan, Y. Zhang, K. Han, A. K. Nandi, B. Honig, and D.-S. Huang, "Integration of multi-omics data for gene regulatory network inference and application to breast cancer," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 16, no. 3, pp. 782-791, 2018.

[70] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics,* vol. 27, no. 21, pp. 3036-3043, 2011.

[71] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics,* vol. 31, no. 4, pp. 608-609, 2015.

[72] J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, C. Li, X. Li, S. Rao, and X. Li, "DOSim: an R package for similarity between diseases based on disease ontology," *BMC bioinformatics,* vol. 12, no. 1, pp. 266, 2011.

[73] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature*

communications, vol. 8, no. 1, pp. 1-13, 2017.

[74] V. Gligorijević, M. Barot, and R. Bonneau, "deepNF: deep network fusion for protein function prediction," *Bioinformatics,* vol. 34, no. 22, pp. 3873-3881, 2018.

Figures

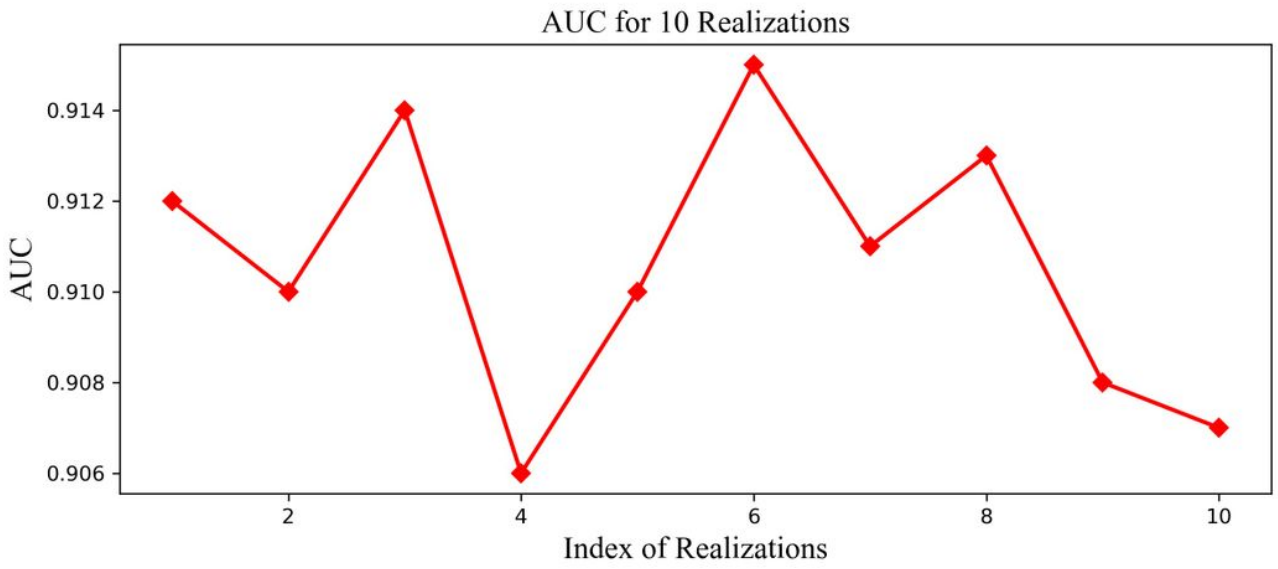

Figure 1

The AUC values for 10 realizations.

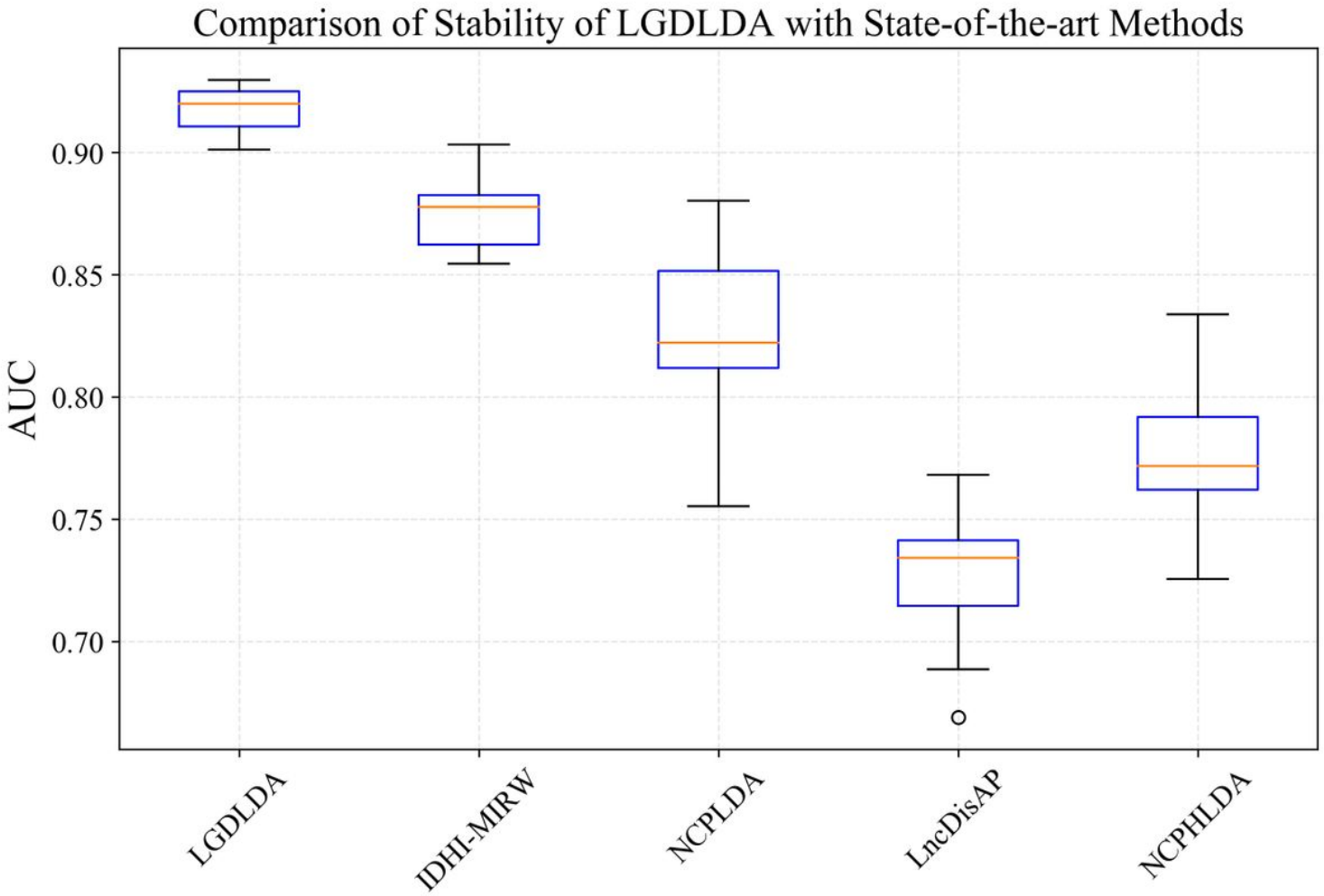

Figure 2

The box plots of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA.

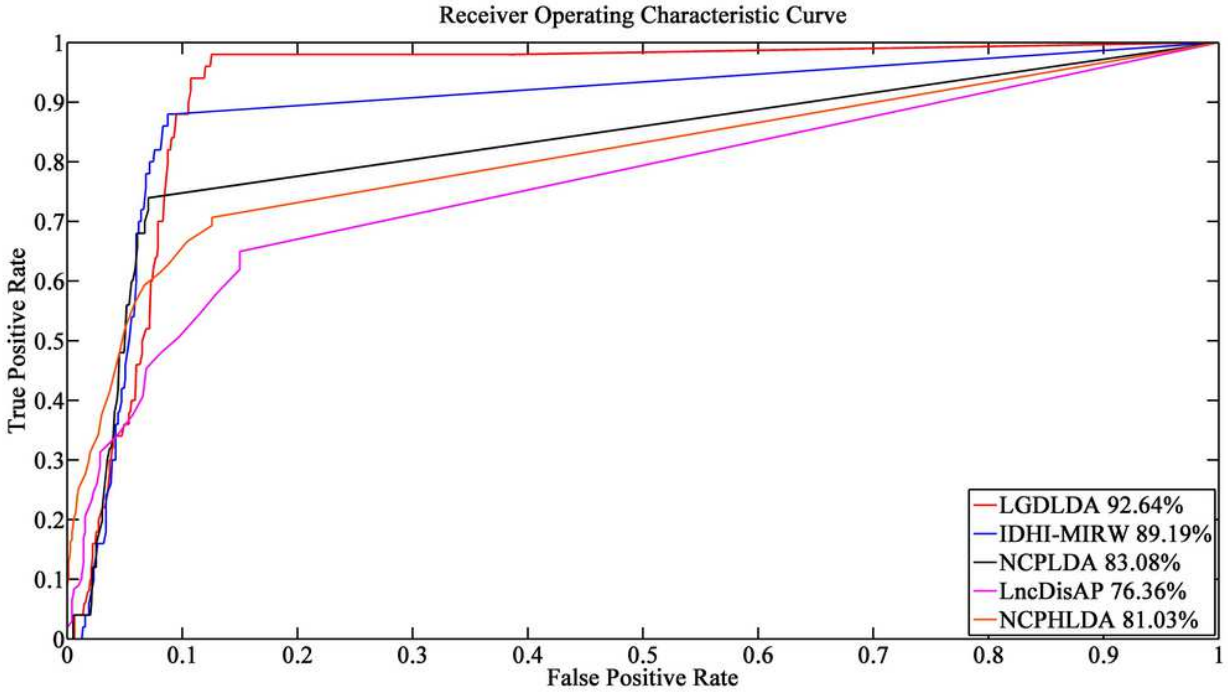

Figure 3

The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on data that does not contain gene-related information.

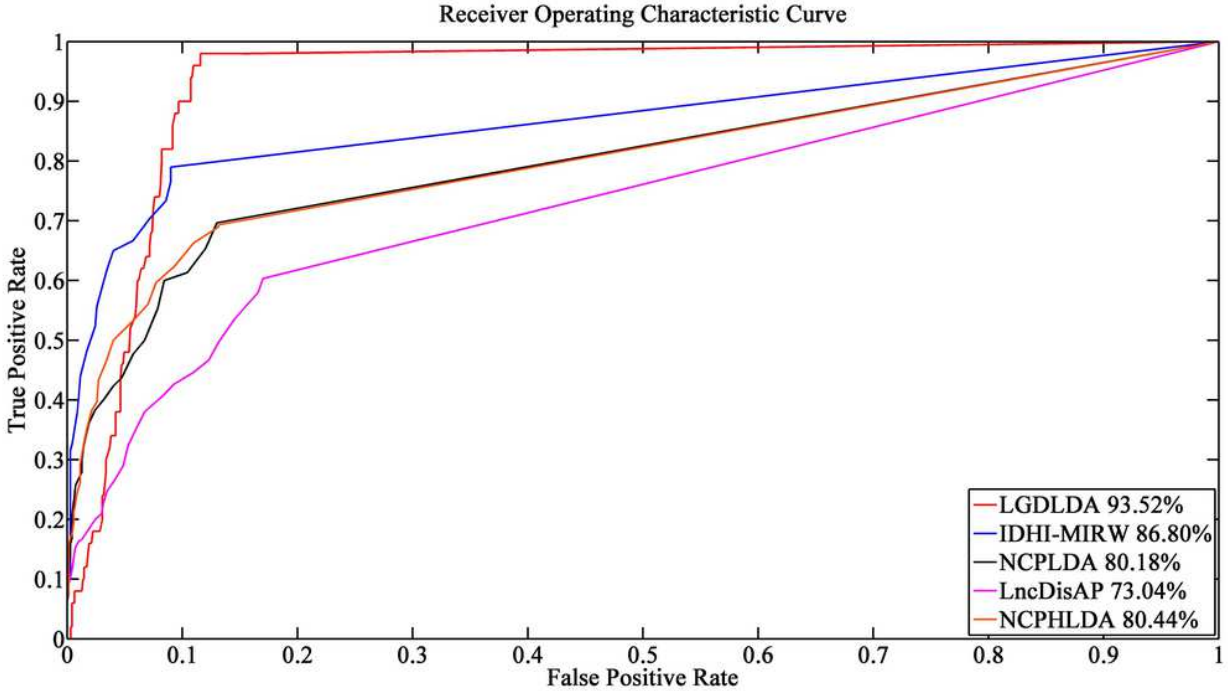

Figure 4

The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on data containing gene information.

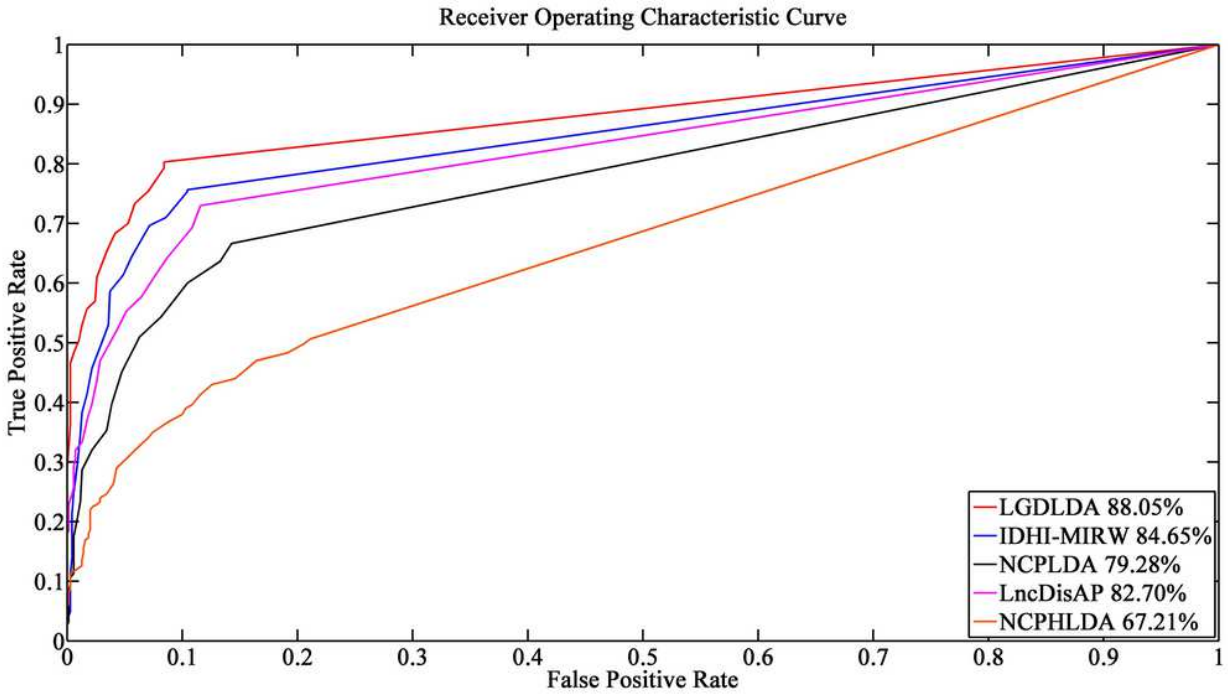

Figure 5

The ROCs and corresponding AUC values of LGDLDA, IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA on the data with missing part of the information.

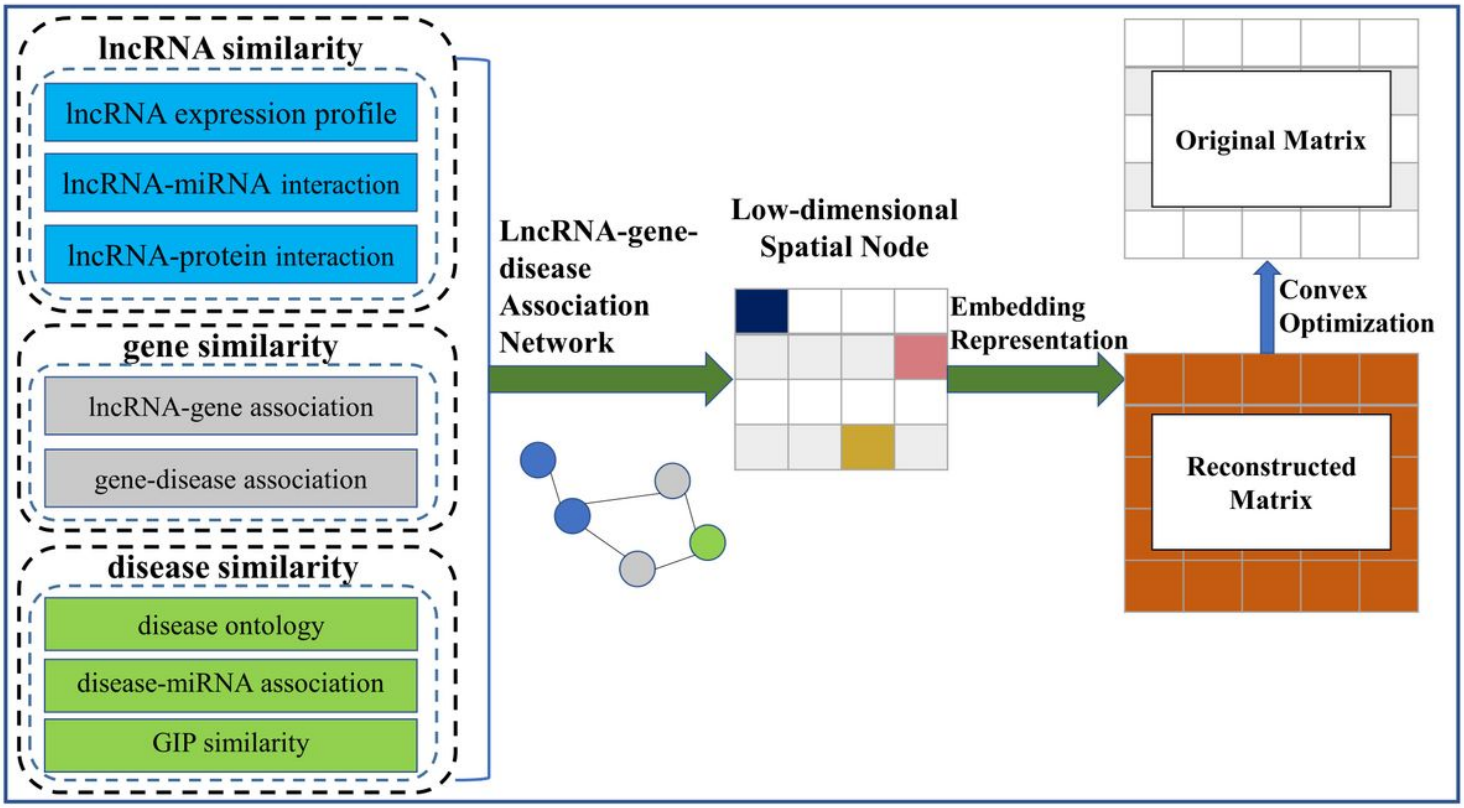

Figure 6

The flowchart of LGDLDA. (1) LGDLDA uses multiple association similarity matrices to build lncRNA-gene-disease association network. (2) Based on the matrices generated in the first step, LGDLDA uses the association similarity matrices combined with neural network to calculate the neighborhood information of lncRNAs and diseases, and further embeds it into the low-dimensional spatial node representations. (3) LGDLDA uses embedded representations to generate the reconstructed matrix to approximate the original matrix, and learns as much information in the original matrix as possible in the optimization of the loss function. (4) LGDLDA sorts the elements in the learned association matrix and selects the top values to predict cancer-related lncRNAs.