RESEARCH ARTICLE

Fusion of KATZ measure and space projection to fast probe potential lncRNA-disease associations in bipartite graphs

Yi Zhang[1,2☯], Min Chen[3☯*], Li Huang[4,5], Xiaolan Xie[1], Xin Li[1], Hong Jin[1], Xiaohua Wang[6], Hanyan Wei[6]

1 School of Information Science and Engineering, Guilin University of Technology, Guilin, China, **2** Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, China, **3** School of Computer Science and Technology, Hunan Institute of Technology, Hengyang, China, **4** Academy of Arts and Design, Tsinghua University, Beijing, China, **5** The Future Laboratory, Tsinghua University, Beijing, China, **6** Pharmacy School, Guilin Medical University, Guilin, China

☯ These authors contributed equally to this work.
* chenmin@hnit.edu.cn

Abstract

It is well known that numerous long noncoding RNAs (lncRNAs) closely relate to the physiological and pathological processes of human diseases and can serves as potential biomarkers. Therefore, lncRNA-disease associations that are identified by computational methods as the targeted candidates reduce the cost of biological experiments focusing on deep study furtherly. However, inaccurate construction of similarity networks and inadequate numbers of observed known lncRNA–disease associations, such inherent problems make many mature computational methods that have been developed for many years still exit some limitations. It motivates us to explore a new computational method that was fused with KATZ measure and space projection to fast probing potential lncRNA-disease associations (namely KATZSP). KATZSP is comprised of following key steps: combining all the global information with which to change Boolean network of known lncRNA–disease associations into the weighted networks; changing the similarities calculation into counting the number of walks that connect lncRNA nodes and disease nodes in bipartite graphs; obtaining the space projection scores to refine the primary prediction scores. The process to fuse KATZ measure and space projection was simplified and uncomplicated with needing only one attenuation factor. The leave-one-out cross validation (LOOCV) experimental results showed that, compared with other state-of-the-art methods (NCPLDA, LDAI-ISPS and IIRWR), KATZSP had a higher predictive accuracy shown with area-under-the-curve (AUC) value on the three datasets built, while KATZSP well worked on inferring potential associations related to new lncRNAs (or isolated diseases). The results from real cases study (such as pancreas cancer, lung cancer and colorectal cancer) further confirmed that KATZSP is capable of superior predictive ability to be applied as a guide for traditional biological experiments.

OPEN ACCESS

Citation: Zhang Y, Chen M, Huang L, Xie X, Li X, Jin H, et al. (2021) Fusion of KATZ measure and space projection to fast probe potential lncRNA-disease associations in bipartite graphs. PLoS ONE 16(11): e0260329. https://doi.org/10.1371/journal.pone.0260329

Editor: Qi Zhao, University of Science and Technology Liaoning, CHINA

Received: September 12, 2021

Accepted: November 6, 2021

Published: November 22, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The sources of funding are National Nature Science Foundation of China (Grant No. 62166014) with funder of Yi Zhang, Nature Science Foundation of Guangxi Province (Grant No. 2020GXNSFAA297255) with funder of Yi Zhang, Guangxi key Laboratory Fund of Embedded Technology and Intelligent System (Grant No. 2019-01-06) with funder of Yi Zhang, Scientific

Research Project of Education Department of Hunan Province (Grant No. 19A125) with funder of Min Chen, respectively. Yi Zhang and Min Chen who are regarded as the joint first authors of this manuscript are responsible of conceptualization, original draft, review and editing.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: LOOCV, leave-one-out cross validation; ROC, receiver operating characteristic; AUC, area under the ROC curve; FPR, false positive rate; TPR, true positive rate.

Introduction

Long non-coding RNAs (lncRNAs) whose length are longer than 200 nucleotides (nt) have crucial roles in gene expression control during developmental and differentiational processes [1]. Therefore, there is no surprise that mutation and dysregulation of lncRNAs could contribute to the development of various human complex diseases [2], such as HOTAIR in breast cancer [3] and MALAT1 in early-stage non-small cell lung cancer [4]. LncRNAs can also drive many important cancer phenotypes through their interactions with other cellular macromolecules including DNA, protein, and RNA [5–8]. There is urgent need to discern potential functional roles of lncRNAs to further study the pathology, diagnosis, therapy, prognosis, prevention of human complex diseases, and detect disease biomarkers at lncRNA level [9, 10]. With strong data support from lncRNA related databases (such as LncRNAdb [11], LncRNA-Disease [12], NRED [13], and NONCODE [14]) and similarity calculation based on miRNA information [15–20], the computational prediction models that were built to infer lncRNA–disease associations could supply more accurate targeted candidates [21]: 1) saving cost and time for biological experiments; 2) making bio-experiments focus on deeper study of targets; 3) speeding up understanding the pathogenesis of complex diseases.

The computational models used for inferring lncRNA–disease associations have been divided into three main categories: 1) Machine learning-based inferring models use naive Bayesian classifier model [22, 23], support vector machine (SVM) [24, 25], matrix completion [26, 27], matrix factorization [28–30] to infer potential lncRNA–disease associations. However, the models categorized to this category are not able to achieve high predictive accuracy. 2) Network-based inferring models, based on the biological premise that lncRNAs with similar functions tend to be associated with similar diseases [31, 32], use random walk [33–35], KATZ measure [36, 37], hyper geometric distribution [15], label propagation algorithm [38], propagating information streams [39], lncRNA-miRNA interaction [15, 30] to identify potential lncRNA–disease associations. Nevertheless, the models categorized to this category rely heavily on the information integrated from diverse biological data sources, and it is difficult to integrate heterogeneous data from multiple sources deeply. 3) Convolutional neural network (CNN) based inferring models [40–43], are at the early research stage, with consuming relatively high time complexity and relying on the quality of multiple sources biological data as well. Therefore, those above models still have different limitations, such as, needing negative samples, not being able to infer associations related to isolated diseases and new lncRNAs directly, not high accuracy with singular methodology. Addressing these limitations, we explored a novel prediction method based on the fusion of KATZ Measure and Space Projection to infer potential lncRNA-disease associations in bipartite graphs, namely KATZSP.

KATZ measure such a graph-based computational method could be used to transform the problem of calculating similarities between nodes to link prediction in bipartite graph. In the context of lncRNA-disease association prediction, the heterogeneous networks are represented by matrices (also called bipartite graph). Therefore, calculating similarities between the nodes of lncRNAs and diseases is further transformed into the problem of counting the number of walks that connect the interactive lncRNA-disease pairs in bipartite graph. Furthermore, the number of walks as the lengths decided the potential association probability of this lncRNA-disease pair [36, 44]. Space projection method [45, 46] could improve the lncRNA-disease association predictive ability easily with few regulation parameters, even though the known lncRNA-disease associations exist inherent data sparsity. After simplified and uncomplicated fusion process, KATZ measure and space projection method were fused to form an integrated computational model KATZSP with needing only one attenuation factor, while dropping above limitations.

Experimental evaluation and discussion

Evaluation metrics

Leave One Out Cross Validation (LOOCV) experiments were implemented for evaluating the predictive performance of KATZSP. We divided the dataset of known associations into two parts: the testing subset and the training subset. In the testing subset, each known association was used as a test data in turn, and the remaining known associations formed the training subset. Under the framework of LOOCV, we compared the prediction results on some specific threshold to obtain the following four metrics: true positive (TP), false positive (FP), false negative (FN), true negative (TN). Furthermore, according to some specified thresholds, we calculated the true positive rate ($TPR = \frac{TP}{TP+FN}$) against false positive rate ($FPR = \frac{FP}{TN+FP}$) with which to plot out the receiver operating characteristic curve (ROC). The area under the ROC curve (AUC) was finally calculated to numerically evaluate the overall predictive performance of KATZSP.

Impact with parameter selection

Coefficient β plays as an attenuation factor of weight to control the contribution of lengths coming from walks on calculating the similarities between any two interactive nodes. According to the convergence properties of sequences required by KATZ method, the value of β should be less than the reciprocal of the max-eigenvalue of the adjacency matrix **A**. In order to obtain the optimal value of β, we set $\beta = 1/\max(\text{eig}(\mathbf{A}))^{*}K$ where max(eig(**A**)) denotes the max-eigenvalue of adjacency matrix **A**. Then the value of K was increased from 0.1 to 0.9 with step size of 0.1. With changing the value of K, LOOCV was implemented on all the three datasets built (dataset 1, dataset 2 and dataset 3). The results in Fig 1 showed that AUC could achieve the maximum value on all the three datasets when $K = 0.1$.

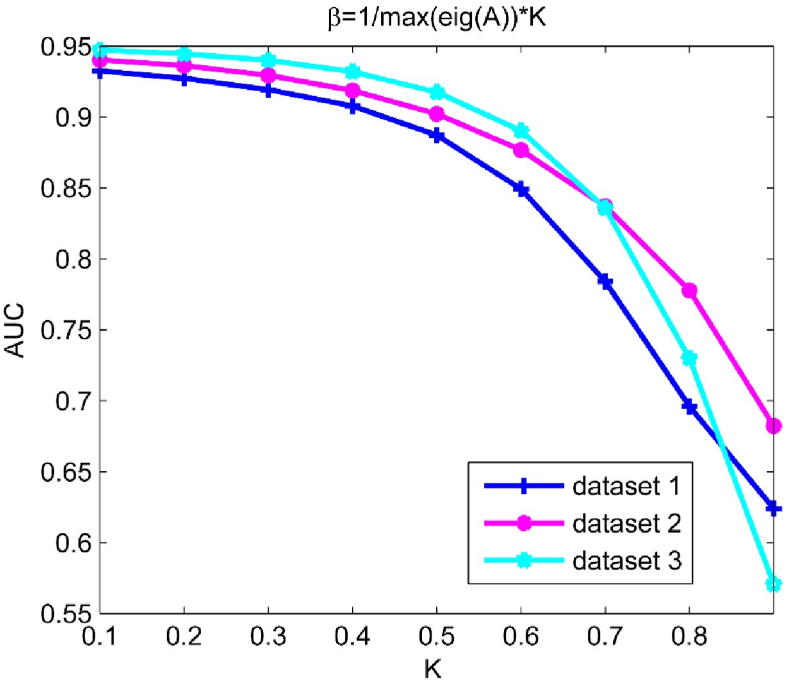


Fig 1. Impact with parameter variation on model prediction accuracy.

https://doi.org/10.1371/journal.pone.0260329.g001

Compare predictive abilities under different solutions

To demonstrate how our technical solution selected performed better than others, LOOCV experiments were implemented under following four technical solutions: only using space projection (SP), only using KATZ (KATZ), using space project first and then KATZ (SPKATZ), using KATZ first and then space projection (KATZSP). The results compared on three datasets (dataset 1, dataset 2 and dataset 3) were shown in Figs 2–4, respectively.

From the comparison results shown in Figs 2–4, we easily found the solution used in our model (KATZSP) achieved AUC values of 0.9324, 0.9403 and 0.9472 on dataset 1, dataset 2 and dataset 3, respectively. Among above four solutions, our KATZSP which performed the best predictive ability on all three datasets with distinct advantage than other three solutions.

Compare performance with other models

To further demonstrate the reliable predictive ability of our model, we chose some the-state-of-art computational models in similar type (NCPLDA [47], LDAI-ISPS [48] and IIRWR [49]) to compare with our model in the framework of LOOCV. To make comparison fairly, we configured the same experimental environment and condition for all models on dataset 1, dataset 2 and dataset 3. From the comparison results shown in Figs 5–7, our KATZSP achieved the highest AUC values on all three datasets with detail analysis shown in Table 1.

Verify predictive ability for new lncRNAs and isolated diseases

To implement the verification in this section, we simulated each lncRNA in the known lncRNA-disease associations dataset to be a new lncRNA by removing all known associations relating to it. Similarly, we simulated each disease in the known lncRNA-disease associations

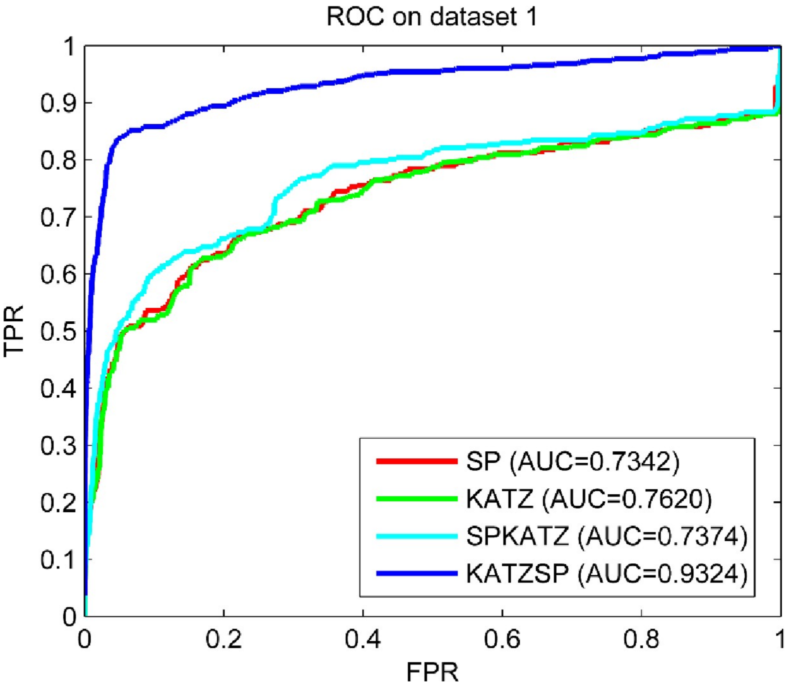


Fig 2. Predictive abilities with different technical solutions on dataset 1.

https://doi.org/10.1371/journal.pone.0260329.g002

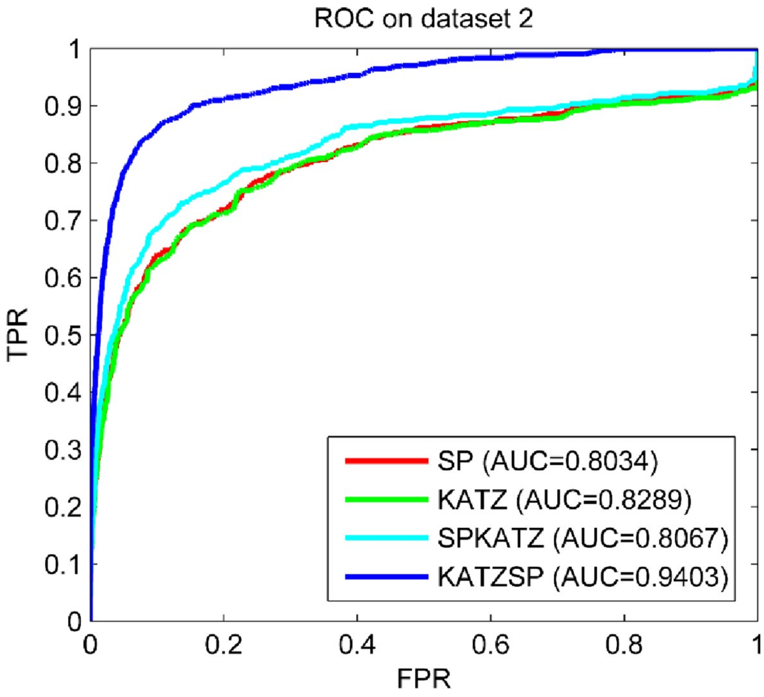


Fig 3. Predictive abilities with different technical solutions on dataset 2.

https://doi.org/10.1371/journal.pone.0260329.g003

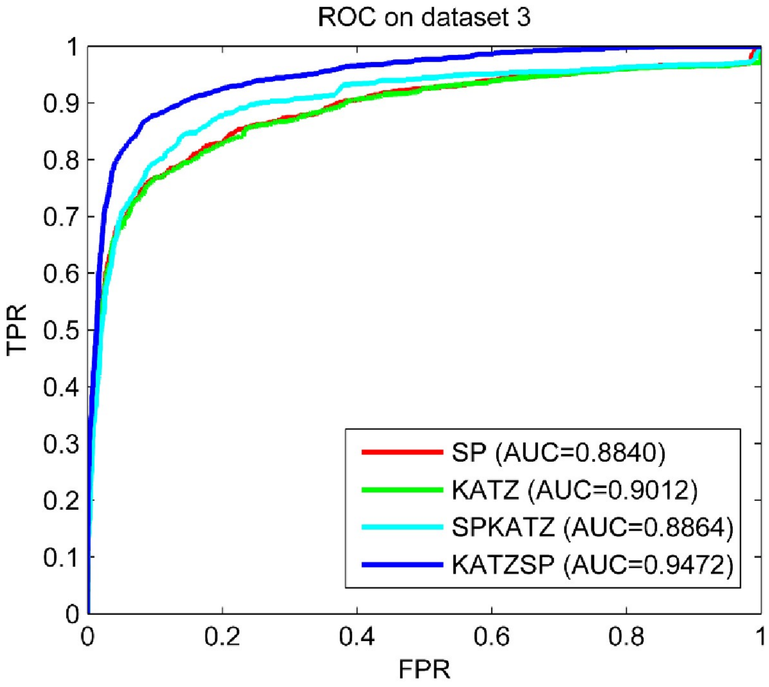
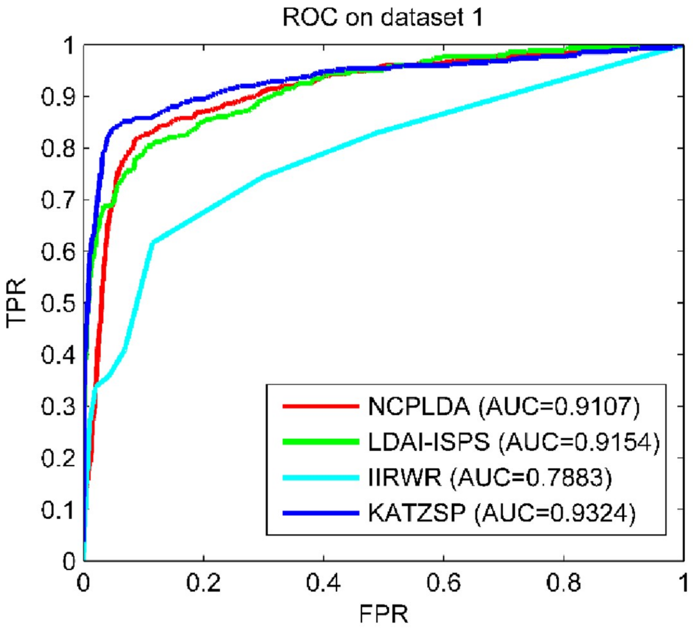


Fig 4. Predictive abilities with different technical solutions on dataset 3.

https://doi.org/10.1371/journal.pone.0260329.g004



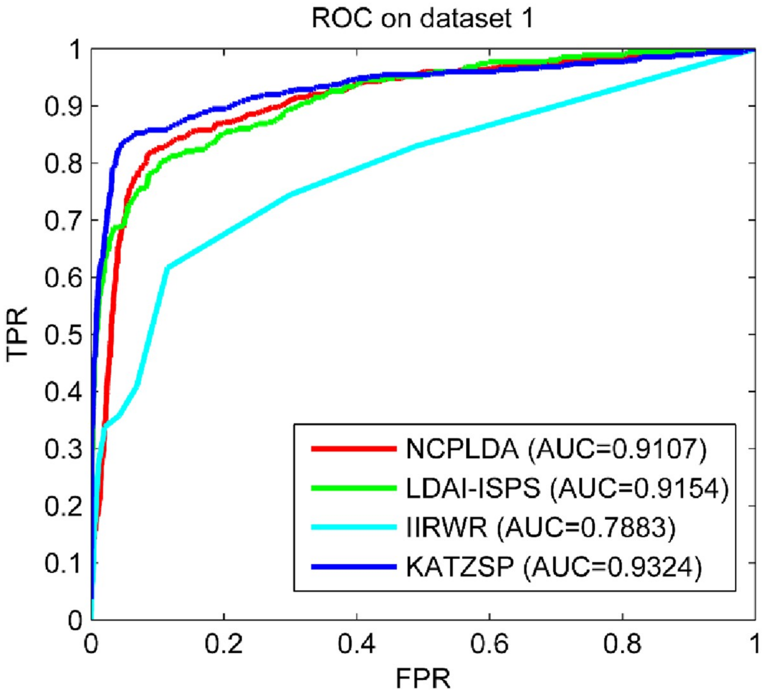


Fig 5. Predictive abilities of KATZSP and other models on dataset 1.

https://doi.org/10.1371/journal.pone.0260329.g005

dataset to be an isolated disease by removing all known associations relating to it. Each new lncRNA (or isolated disease) simulated was specified to be the test sample for model evaluation and the rest lncRNAs (or diseases) in the known lncRNA-disease associations dataset worked as the training samples for model learning. Until the associations between each new

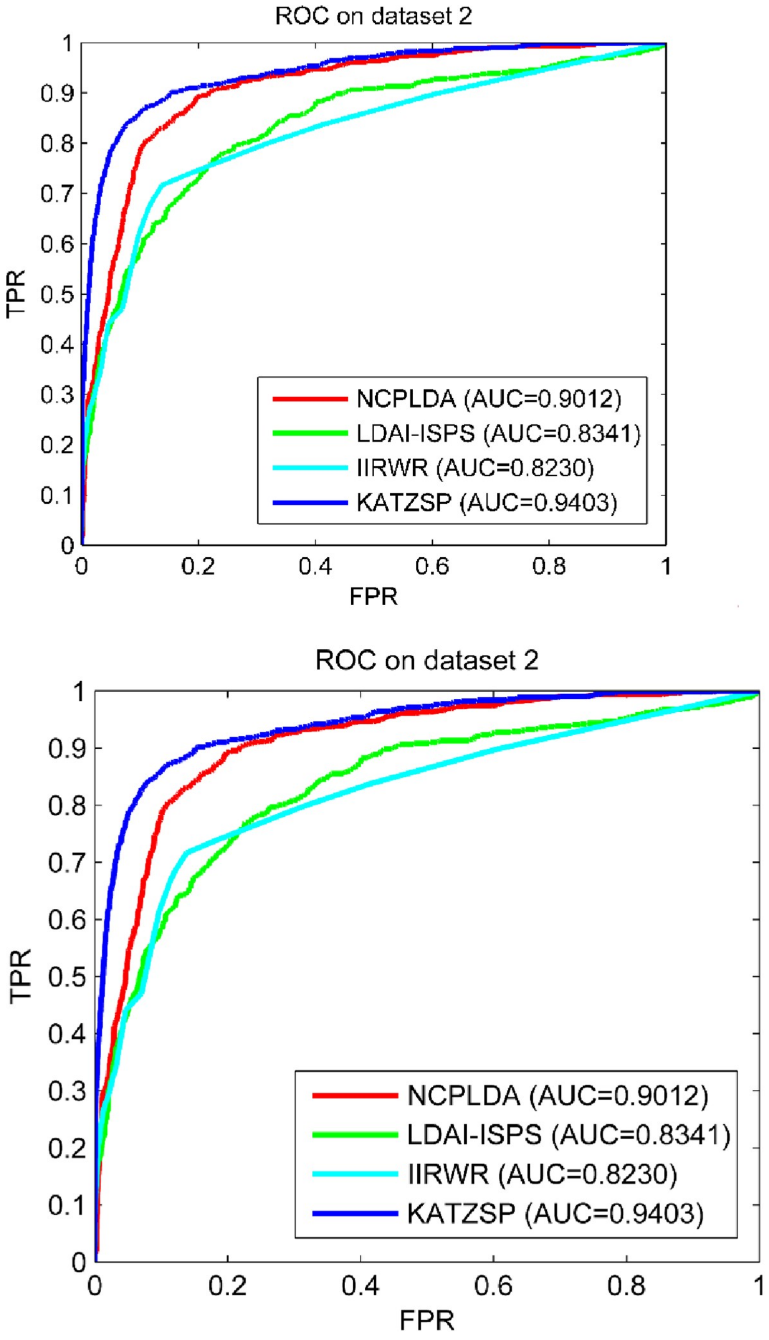


Fig 6. Predictive abilities of KATZSP and other models on dataset 2.

https://doi.org/10.1371/journal.pone.0260329.g006

lncRNA and diseases or the associations between lncRNAs and each isolated disease were inferred by our KATZSP, the inferred results on dataset 1, dataset 2 and dataset 3 were shown in Fig 8.

With the AUC values in Fig 8, it demonstrated that our KATZSP could be effectively applied to infer associations related to new lncRNAs and associations related to isolated diseases.

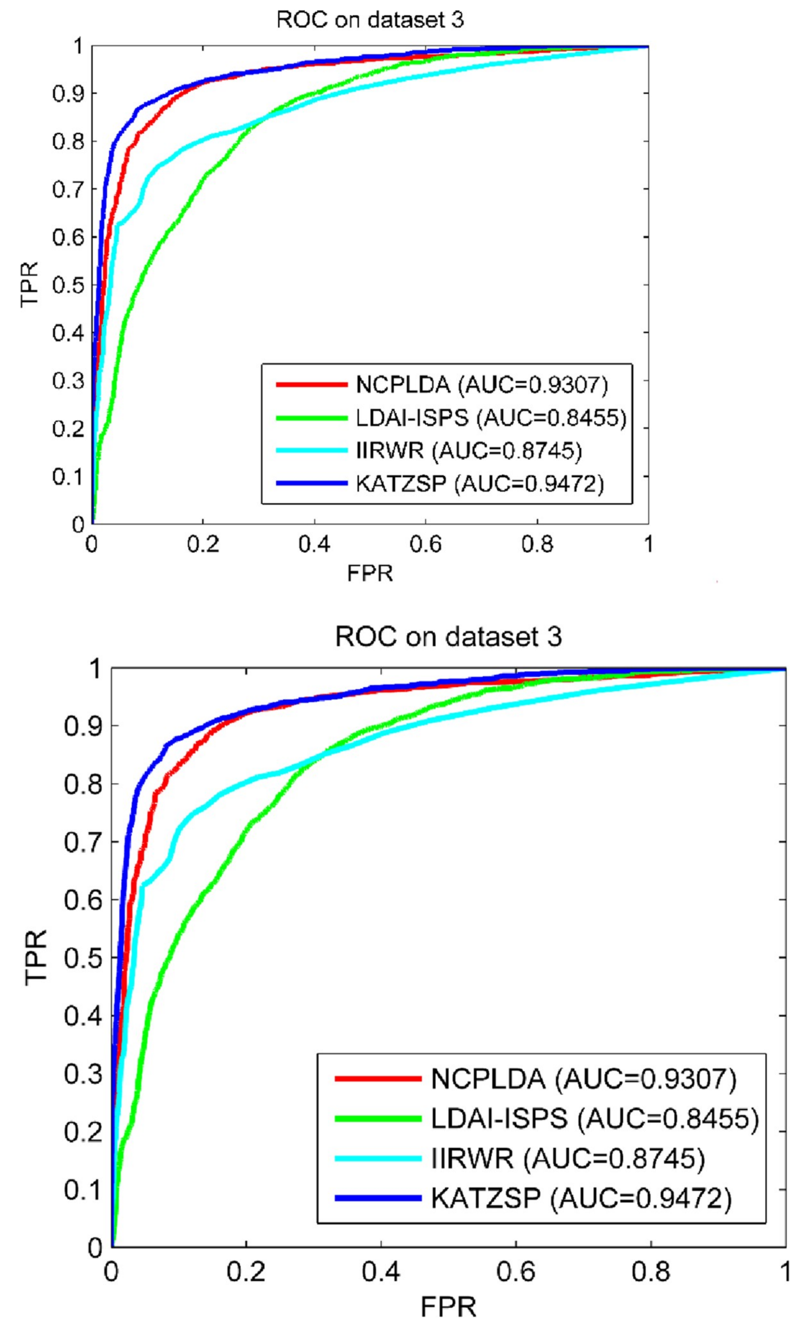


Fig 7. Predictive abilities of KATZSP and other models on dataset 3.

https://doi.org/10.1371/journal.pone.0260329.g007

Cases study

Case study for three specific diseases

To further demonstrate the predictive performance of our KATZSP on real cases study, we selected three specific diseases (pancreas cancer, lung cancer and colorectal cancer) as the cases to examine. With using the training samples composed of the known associations in dataset 2 and the testing samples composed of the unknown associations, our KATZSP

Table 1. AUCs of KATZSP and other models on all three datasets.

Model AUC value	NCPLDA	LDAI-ISPS	IIRWR	KATZSP
AUC on dataset 1	0.9107 (2.3%)	0.9154 (1.9%)	0.7883 (18.3%)	0.9324
AUC on dataset 2	0.9012 (4.3%)	0.8341 (11.8%)	0.8230 (14.3%)	0.9403
AUC on dataset 3	0.9307 (1.7%)	0.8455 (12%)	0.8745 (8.3%)	0.9472

From data of "AUC on dataset 1" in Table 1, our KATZSP was demonstrated with higher AUC values which were 2.3%, 1.9% and 18.3% higher than that of NCPLDA, LDAI-ISPS and IIRWR, respectively. Similarly, the comparison results on dataset 2 demonstrated the AUC values of our KATZSP were 4.3%, 11.8% and 14.3% higher than that of NCPLDA, LDAI-ISPS and IIRWR, respectively. In the last row of Table 1, the 1.7%, 12% and 8.3% higher AUC values of our KATZSP were compared with that of NCPLDA, LDAI-ISPS and IIRWR, respectively. Therefore, our KATZSP was demonstrated with more reliable predictive ability over other previous models on all the three datasets under the evaluation framework of LOOCV.

https://doi.org/10.1371/journal.pone.0260329.t001

focused on inferring the potential lncRNAs relating to above three cases. The lncRNAs with the top five highest prediction scores of each case were listed in Table 2. If the same associations predicted by KATZSP were also found in some literatures or the newest databases, such as LncRNADisease 2.0 (http://www.rnanut.net/lncrnadisease) and Lnc2Cancer 3.0 (http://www.biobigdata.net/lnc2cancer), it could further validate with the supporting evidences that our KATZSP was capable of the reliable predictive ability and practicability.

Case study for isolated diseases

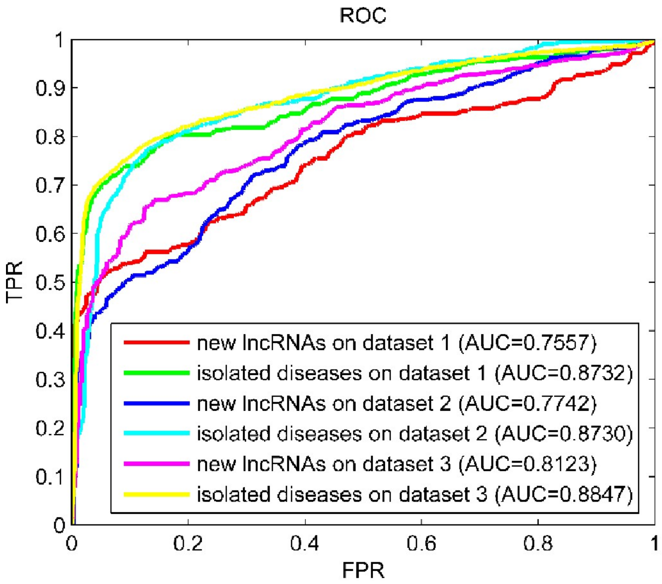
In recent years, many new diseases without any known association r lncRNAs have been gradually discovered, namely isolated diseases. It is important to verify if our KATZSP could be applied to infer the potential lncRNAs associated to such kind of isolated diseases. Above three cases (pancreas cancer, lung cancer and colon cancer) were simulated as the isolated diseases by removing all known associations relating to them in dataset 2. Our KATZSP only used other information to infer the potential lncRNAs associated with these three isolated diseases simulated. The top five lncRNAs with highest prediction scores of each disease were listed in Table 3 where only two prediction results (TC0101441 and KRASP1) couldn't be found supporting evidence from any databases or published literatures.

In Tables 2 and 3, all predicted results except two were confirmed with extra evidences, which validated our KATZSP could be effectively applied in real life with supplying calculated candidates to guide biological experiments.

Materials and methods

Obtain data source

Known lncRNA-disease associations. From a publicly accessible address at http://www.cuilab.cn/lncrnadisease, three versions of the databases which consist of associations between lncRNAs and human diseases were obtained for our work. With processing of the database in version 2013, we built a new dataset (namely dataset 1) with 352 known lncRNA–disease associations involved in 156 lncRNAs and 190 diseases. With processing of the database in version 2016, a new-built dataset (namely dataset 2) consists of 621 known lncRNA–disease associations involved in 285 lncRNAs and 226 diseases. With processing of the database in version 2017, a similar new-built dataset (namely dataset 3) consists of 1695 known lncRNA–disease associations involved in 828 lncRNAs and 314 diseases. The observed lncRNA–disease associations with lncRNA nodes and disease nodes form the bipartite graph denoted by the Boolean



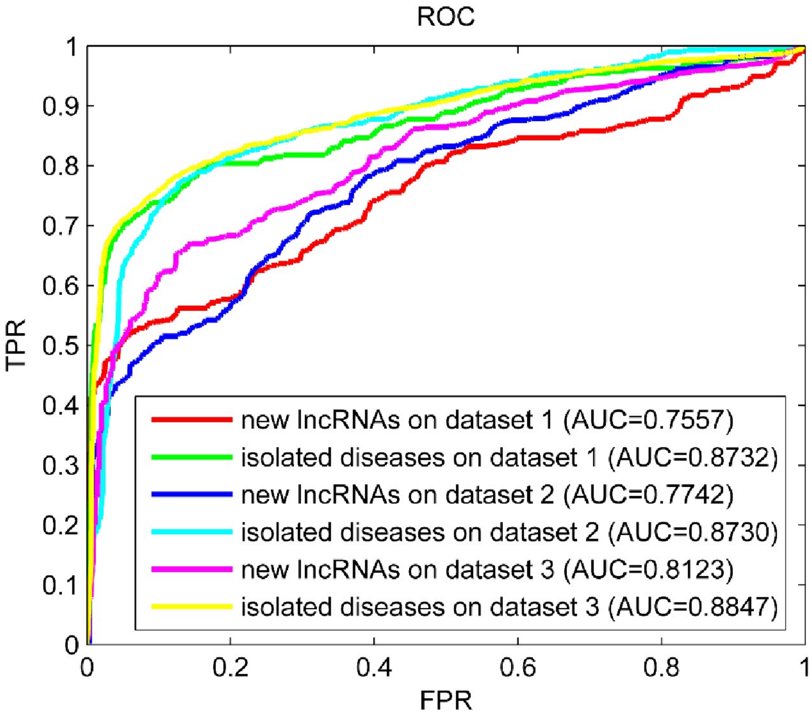


Fig 8. Predictive ability of KATZSP for new lncRNAs and isolated diseases.

https://doi.org/10.1371/journal.pone.0260329.g008

matrix **LD** = $(ld_{ij})_{nl \times nd}$, whose element ld_{ij} is 1 when lncRNA l_i relates to disease d_j. Otherwise, the value of element ld_{ij} is 0. The number of lncRNAs and the number of diseases in matrix **LD** are denoted by nl and nd, respectively.

Disease–disease semantic similarity. Referring to the description by Wang et al. [51], in DAG (Directed Acyclic Graph), the contribution of a disease d_t to the semantics of disease d_i

Table 2. Top 5 specific diseases-related candidate lncRNAs.

Case	LncRNA	Evidences	Rank
Pancreas cancer	H19	LncRNADisease	1
Pancreas cancer	MEG3	LncRNADisease	2
Pancreas cancer	CDKN2B-AS1	LncRNADisease	3
Pancreas cancer	GAS5	LncRNADisease	4
Pancreas cancer	UCA1	LncRNADisease	5
Lung cancer	PVT1	LncRNADisease	1
Lung cancer	GAS5	LncRNADisease	2
Lung cancer	CDKN2B-AS1	LncRNADisease	3
Lung cancer	UCA1	LncRNADisease	4
Lung cancer	NPTN-IT1	Lnc2Cancer	5
Colorectal cancer	PVT1	LncRNADisease	1
Colorectal cancer	CDKN2B-AS1	Lnc2Cancer	2
Colorectal cancer	LSINCT5	Lnc2Cancer	3
Colorectal cancer	GAS5	Lnc2Cancer	4
Colorectal cancer	UCA1	LncRNADisease	5

The data in column "Evidences" of Table 2 showed that all the potential lncRNAs inferred relating to the three specific diseases have been found the evidence in LncRNADisease 2.0 or Lnc2Cancer 3.0. It validated the reliability of the inferred results coming from our KATZSP.

https://doi.org/10.1371/journal.pone.0260329.t002

has following definition with denotation of $D_{d_i}(d_t)$:

$$D_{d_i}(d_t) = \begin{cases} 1, \text{ if } d_t = d_i \\ \max\{\Delta * D_{d_i}(d_{t'}) | d_{t'} \in \text{children of } d_t\}, \text{ if } d_t \neq d_i \end{cases} \tag{1}$$

where Δ was set to be the most suitable value of 0.5.

Based on both the addresses of diseases in DAG graphs and the semantic relations with ancestor diseases, the element dd_{ij} in matrix $\mathbf{DD} = (dd_{ij})_{nd \times nd}$ denotes the semantic similarity

Table 3. Top 5 specific isolated diseases-related candidate lncRNAs.

Disease	lncRNA name	Evidences	Rank
pancreas cancer	HOTAIR	LncRNADisease	1
pancreas cancer	MALAT1	LncRNADisease	2
pancreas cancer	H19	LncRNADisease	3
pancreas cancer	MEG3	LncRNADisease	4
pancreas cancer	TC0101441	No evidence	5
lung cancer	HOTAIR	LncRNADisease	1
lung cancer	MALAT1	LncRNADisease	2
lung cancer	H19	LncRNADisease	3
lung cancer	MEG3	LncRNADisease	4
lung cancer	PVT1	LncRNADisease	5
colon cancer	HOTAIR	LncRNADisease	1
colon cancer	MALAT1	LncRNADisease	2
colon cancer	H19	LncRNADisease	3
colon cancer	EPB41L4A-AS1	Literature [50]	4
colon cancer	KRASP1	No evidence	5

https://doi.org/10.1371/journal.pone.0260329.t003 between diseases d_i and d_j with definition as follows:

$$dd_{ij} = \frac{\sum_{d_t \in T_{d_i} \cap T_{d_j}} (D_{d_i}(d_t) + D_{d_j}(d_t))}{\sum_{d_t \in T_{d_i}} D_{d_i}(d_t) + \sum_{d_t \in T_{d_j}} D_{d_j}(d_t)} \tag{2}$$

where T_{d_i} is the set of all ancestor nodes relating to disease d_i, including node d_i itself in DAG.

LncRNA–lncRNA functional similarity. How to accurately measure the functional similarity between two lncRNAs was detailly descripted in many literatures [47–49, 52]. A group of diseases which have associations with lncRNA l_i were denoted by $D^{(l_i)} = \{d_{i_1}, d_{i_2}, \cdots, d_{i_k}\}$, and the similarity between any disease d_t in $D^{(l_i)}$ and the whole set $D^{(l_i)}$ has following definition:

$$S(d_t, D^{(l_i)}) = \max_{1 \le x \le k} dd_{ti_x} \tag{3}$$

Similarly, set $D^{(l_j)} = \{d_{j_1}, d_{j_2}, \cdots, d_{j_{k'}}\}$ denotes a group of diseases associate with lncRNA l_j. The similarity between any disease d_t in $D^{(l_j)}$ and the whole set $D^{(l_j)}$ has following definition:

$$S(d_t, D^{(l_j)}) = \max_{1 \le x \le k'} dd_{tj_x} \tag{4}$$

Functional similarities between the lncRNAs were denoted by $\mathbf{LL} = (ll_{ij})_{nl \times nl}$ whose element ll_{ij} represents the functional similarity between l_i and l_j with calculation as follows:

$$ll_{ij} = \frac{\sum_{1 \le x \le k} S(d_{i_x}, D^{(l_j)}) + \sum_{1 \le y \le k'} S(d_{j_y}, D^{(l_i)})}{k + k'} \tag{5}$$

Central similarity of the Gaussian interaction profile. Compared to the number of unknown lncRNA–disease associations, the number of known lncRNA–disease associations is very small, which leads the bipartite graph represented by Boolean matrix of known lncRNA–disease associations to have sparsity. In order to reduce the influence from sparsity on prediction precision, the central similarities of Gaussian interaction profile were calculated in accordance with the description in Laarhoven's work [53]. Therefore, the central similarities of Gaussian interaction profile between the diseases were denoted by $\mathbf{DD}^{(g)} = (dd^g_{ij})_{nd \times nd}$ whose element dd^g_{ij} represents the central similarity of Gaussian interaction profile between disease d_i and d_j with following definition:

$$dd^g_{ij} = \exp(-\gamma_d \| \mathbf{LD}(:, i) - \mathbf{LD}(:, j) \|^2) \tag{6}$$

where the ith column of matrix $\mathbf{LD}$ was denoted by $\mathbf{LD}(:,i)$ which represents all the known associations relating to disease d_i; The Gaussian kernel bandwidth here was denoted by γ_d with following definition in accordance to the previous study [54]:

$$\gamma_d = \frac{1}{\frac{1}{nd} \sum_{i=1}^{nd} \| \mathbf{LD}(:, i) \|^2} \tag{7}$$

Similarly, the central similarities of Gaussian interaction profile between the lncRNAs were denoted by $\mathbf{LL}^{(g)} = (ll^g_{ij})_{nl \times nl}$ whose element ll^g_{ij} represents the central similarity of Gaussian interaction profile between lncRNA l_i and l_j with definition as follows:

$$ll^g_{ij} = \exp(-\gamma_l \| \mathbf{LD}(i, :) - \mathbf{LD}(j, :) \|^2) \tag{8}$$

where the ith row of matrix $\mathbf{LD}$ was denoted by $\mathbf{LD}(i,:)$ which represents all the known associations relating to lncRNA l_i; The Gaussian kernel bandwidth here was denoted by γ_l with following definition:

$$\gamma_l = \frac{1}{\frac{1}{nl}\sum_{i=1}^{nl} \|\mathbf{LD}(i,:)\|^2} \tag{9}$$

Integrated similarity of lncRNAs and diseases. The final similarity matrix of diseases denoted by $\mathbf{DD}^{(f)} = (dd_{ij}^{f})_{nd\times nd}$ comes from an integration of **DD** and $\mathbf{DD}^{(g)}$, and the final similarity matrix of lncRNAs denoted by $\mathbf{LL}^{(f)} = (ll_{ij}^{f})_{nl\times nl}$ comes from an similar integration of **LL** and $\mathbf{LL}^{(g)}$. When the original semantic similarity between disease d_i and d_j was 0, the value of element dd_{ij}^{f} in matrix $\mathbf{DD}^{(f)}$ was set as the central similarity of the Gaussian interaction profile, otherwise it was set as the original semantic similarity between disease d_i and d_j. The value of element ll_{ij}^{f} in matrix $\mathbf{LL}^{(f)}$ has a similar setting process as above. For clarity, the formalized acquirement for element values was defined as follows:

$$dd_{ij}^{f} = \begin{cases} dd_{ij}, & \text{if } \ dd_{ij} \neq 0 \\ dd_{ij}^{g}, & \text{otherwise} \end{cases} \tag{10}$$

$$ll_{ij}^{f} = \begin{cases} ll_{ij}, & \text{if } \ ll_{ij} \neq 0 \\ ll_{ij}^{g}, & \text{otherwise} \end{cases} \tag{11}$$

Obtain primary prediction scores

Construct adjacency matrix. Based on KATZ measurement, the number of walks that connect lncRNA nodes and disease nodes in the original bipartite graph were calculated to measure the similarities between these nodes as the potential association probabilities. The different lengths of walks between lncRNA nodes and disease nodes contributed differently to the similarities between these two kinds of nodes. The shorter length of walks contributed more to the similarities than the longer one. To make full use of the heterogeneous network constructed above, matrix $\mathbf{DD}^{(f)}$, $\mathbf{LL}^{(f)}$ and **LD** were integrated into a new heterogeneous network $\mathbf{A}_{(nl+nd)\times(nl+nd)}$ as the adjacency matrix with definition as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{LL}^{(f)} & \mathbf{LD} \\ \mathbf{LD}^{T} & \mathbf{DD}^{(f)} \end{bmatrix} \tag{12}$$

Calculate primary prediction score on KAZT measurement. By applying KATZ measurement, potential association probabilities between node l_i and node d_j could be calculated as follows with denotation of $S^{KATZ}(l_i, d_j)$:

$$S^{KATZ}(l_i, d_j) = \sum_{w=1}^{m} \beta^{w}(\mathbf{A}^{w})_{l_i,d_j} \tag{13}$$

where β is a non-negative coefficient to control the contribution of lengths coming from walks on the similarities between any two nodes, such as l_i and d_j, β^w raised to the power of w, $(\mathbf{A}^{w})_{l_i,d_j}$ denotes the number of paths whose length of walks equals w between corresponding nodes pair, such as l_i and d_j, m denotes the maximum value of the length of walks.

Because bigger value of the length of walks contributes less to the similarities between two nodes, the above formula for similarity calculation could be approximately described in matrix when the value of m tends to be infinity ($m \to \infty$):

$$S^{KATZ} = \sum_{w=1}^{\infty} \beta^w \mathbf{A}^w = \sum_{w \geq 1} \beta^w \mathbf{A}^w = (\mathbf{I} - \beta \mathbf{A})^{-1} - \mathbf{I} \tag{14}$$

where the value of coefficient β was set in range of $(0,\min\{1,1/\|\mathbf{A}\|_2\})$, matrix S^{KATZ} has the same size as adjacency matrix **A**.

Submatrix $S^{KATZ}[1{:}nl,nl+1{:}nl+nd]$ denotes the elements that located at the rows 1 to nl and the columns nl+1 to nl+nd in matrix S^{KATZ}, which has the same location as matrix **LD** in adjacency matrix **A**. In order to express in a consistent way, submatrix $S^{KATZ}[1{:}nl,nl+1{:}nl+nd]$ was denoted by matrix $\mathbf{LD}^{(p)}_{nl \times nd} = (ld^p_{ij})_{nl \times nd}$ to represent the primary prediction results in the first stage.

Refine primary prediction scores

In order to improve the prediction performance of the proposed model, matrix space projection was used to refine the primary prediction scores obtained in the first stage ($\mathbf{LD}^{(p)}_{nl \times nd}$).

Project on lncRNA space. Project the final similarity matrix of lncRNAs ($\mathbf{LL}^{(f)}$) on the matrix of primary prediction scores ($\mathbf{LD}^{(p)}$) to obtain the projection scores on the lncRNA space, which were denoted by $\mathbf{LD}^{(pl)}_{nl \times nd} = (ld^{pl}_{ij})_{nl \times nd}$ with detailed definition as follows:

$$ld^{pl}_{ij} = \frac{\mathbf{LL}^{(f)}(i,:) \times \mathbf{LD}^{(p)}(:,j)}{\|\mathbf{LD}^{(p)}(:,j)\|} \tag{15}$$

where ld^{pl}_{ij} denotes the predicted score of the association between lncRNA l_i and disease d_j with lncRNA space projection, $\|\mathbf{LD}^{(p)}(:,j)\|$ is the 2-norm of vector $\mathbf{LD}^{(p)}(:,j)$.

Project on disease space. Similarly, project the final similarity matrix of diseases ($\mathbf{DD}^{(f)}$) on the matrix of primary prediction scores ($\mathbf{LD}^{(p)}$) to obtain the projection scores on the disease space, which were denoted by $\mathbf{LD}^{(pd)}_{nd \times nl} = (ld^{pd}_{ij})_{nd \times nl}$ with detailed definition as follows:

$$ld^{pd}_{ij} = \frac{\mathbf{DD}^{(f)}(j,:) \times (\mathbf{LD}^{(p)}(i,:))^T}{\|\mathbf{LD}^{(p)}(i,:)\|} \tag{16}$$

where $(\mathbf{LD}^{(p)}(i,:))^T$ denotes the transpose of vector $\mathbf{LD}^{(p)}(i,:)$, and $\|\mathbf{LD}^{(p)}(i,:)\|$ is the 2-norm of vector $\mathbf{LD}^{(p)}(i,:)$.

Integrate space projection scores. In order to fully capture the information of disease similarity, lncRNA similarity, and known lncRNA–disease associations, we integrated the projection scores on lncRNA space ($\mathbf{LD}^{(pl)}_{nl \times nd}$) and the projection scores on disease space ($\mathbf{LD}^{(pd)}_{nd \times nl}$) to obtain the final prediction scores ($\mathbf{LD}^{(f)}_{nl \times nd}$) with detailed definition as follows:

$$\mathbf{LD}^{(f)} = \frac{\mathbf{LD}^{(pl)} + (\mathbf{LD}^{(pd)})^T}{2} \tag{17}$$

Represent workflow model

With the related data preparation, the inferring process with each key step of KATZSP for lncRNA-disease associations was graphically reprensented in Fig 9.

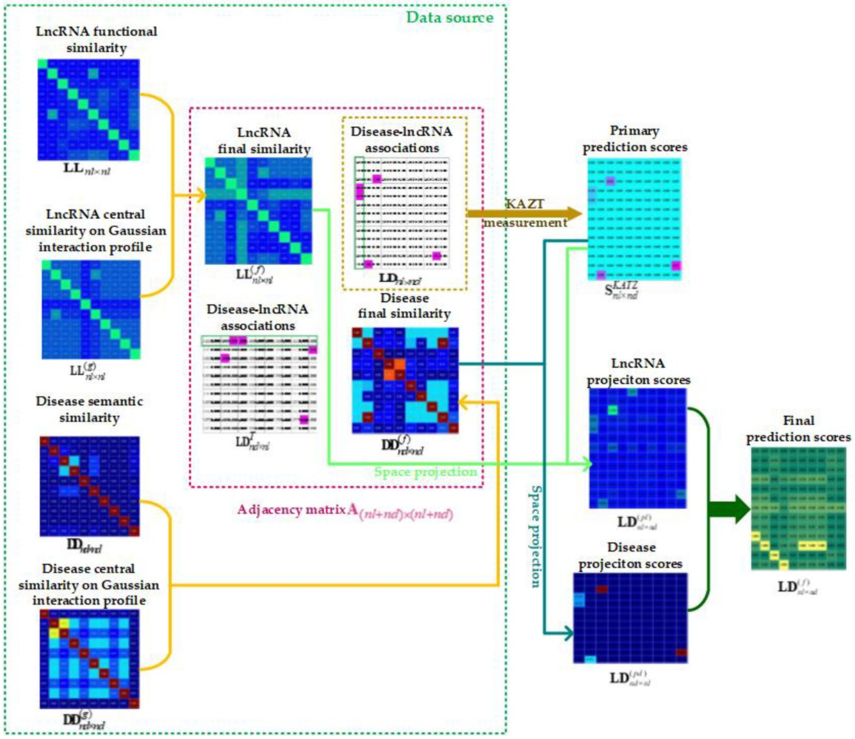

Data source
LncRNA functional similarity
$LL_{nl \times nl}$
LncRNA central similarity on Gaussian interaction profile
$LL^{(g)}_{nl \times nl}$
Disease semantic similarity
$DD_{nd \times nd}$
Disease central similarity on Gaussian interaction profile
$DD^{(g)}_{nd \times nd}$
LncRNA final similarity
$LL^{(f)}_{nl \times nl}$
Disease-lncRNA associations
$LD_{nl \times nd}$
Disease-lncRNA associations
$LD^{T}_{nd \times nl}$
Disease final similarity
$DD^{(f)}_{nd \times nd}$
Space projection
Adjacency matrix $A_{(nl+nd) \times (nl+nd)}$
KAZT measurement
Primary prediction scores
$S^{KATZ}_{nl \times nd}$
Space projection
LncRNA projeciton scores
$LD^{(pl)}_{nl \times nd}$
Disease projeciton scores
$LD^{(pd)}_{nd \times nl}$
Final prediction scores
$LD^{(f)}_{nl \times nd}$

Fig 9. Workflow model of KATZSP.

https://doi.org/10.1371/journal.pone.0260329.g009

Conclusions

In recent years, even though many computational models for inferring lncRNA–disease associations have emerged, those computational methods still have some limitations that motivated us to propose a new model (KATZSP) to infer lncRNA–disease associations. The main contribution of KATZSP is composed of: only needing one attenuation factor β to control the contribution of walk lengths between any two nodes in bipartite graphs; making up the sparsity with simply integrating KATZ measurement and space projection; no needing negative samples; being able to be applied to isolated diseases and new lncRNAs directly. Compared with some state-of-the-art methods in similar type (NCPLDA, LDAI-ISPS and IIRWR), our model KATZSP achieved higher prediction accuracy on all three datasets (dataset 1, dataset 2 and dataset 3). The results from case study further confirmed the stronger predictive performance of KATZSP to be applied for real cases. Our KATZSP still has following limitations that need to be improved in future: further reducing the biases that the predicted results prefer the data with more known associations; the prediction accuracy needing to be enhanced further with fusion of more heterogeneous data.

Supporting information

S1 File. We have released our code publicly at the address of https://github.com/zywait/KATZSP. In the public repository released includes our minimal underlying datasets (data352.mat, data621.mat, data1695.mat).
(ZIP)

Acknowledgments

The authors thank the anonymous reviewers for suggestions that helped improve the paper substantially.

Author Contributions

Conceptualization: Yi Zhang, Min Chen.

Data curation: Yi Zhang.

Formal analysis: Min Chen.

Funding acquisition: Xiaolan Xie.

Methodology: Min Chen.

Resources: Xiaohua Wang, Hanyan Wei.

Software: Xin Li.

Validation: Li Huang, Hong Jin.

Writing – original draft: Yi Zhang.

Writing – review & editing: Yi Zhang.

References

1. Fatica A., Bozzoni I., Long non-coding RNAs: new players in cell differentiation and development, Nature Reviews Genetics, 15 (2014) 7–21. https://doi.org/10.1038/nrg3606 PMID: 24296535
2. Chen X., Yan C.C., Zhang X., You Z.-H., Long non-coding RNAs and complex diseases: from experimental results to computational models, Brief Bioinform, 18 (2017) 558–576. https://doi.org/10.1093/bib/bbw060 PMID: 27345524
3. Xue X., Yang Y.A., Zhang A., Fong K., Kim J., Song B., et al., LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer, Oncogene, 35 (2016) 2746–2755. https://doi.org/10.1038/onc.2015.340 PMID: 26364613
4. Gutschner T., Hämmerle M., Eißmann M., Hsu J., Kim Y., Hung G., et al., The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells, Cancer Research, 73 (2013) 1180–1189. https://doi.org/10.1158/0008-5472.CAN-12-2850 PMID: 23243023
5. Dai X., Zhang S., Zaleta-Rivera K., RNA: interactions drive functionalities, Mol Biol Rep, 47 (2020) 1413–1434. https://doi.org/10.1007/s11033-019-05230-7 PMID: 31838657
6. Hu H., Zhang L., Ai H., Zhang H., Fan Y., Zhao Q., et al., HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy, Rna Biology, 15 (2018) 797–806. https://doi.org/10.1080/15476286.2018.1457935 PMID: 29583068
7. Wang C.-C., Han C.-D., Zhao Q., Chen X., Circular RNAs and complex diseases: from experimental results to computational models, Brief Bioinform, (2021), https://doi.org/10.1093/bib/bbab286 PMID: 34329377
8. Liu W., Jiang Y., Peng L., Sun X., Gan W., Zhao Q., et al., Inferring Gene Regulatory Networks Using the Improved Markov Blanket Discovery Algorithm, Interdisciplinary Sciences: Computational Life Sciences, (2021), https://doi.org/10.1007/s12539-021-00478-9 PMID: 34495484
9. Chen X., Sun Y.-Z., Guan N.-N., Qu J., Huang Z.-A., Zhu Z.-X., et al., Computational models for lncRNA function prediction and functional similarity calculation, Briefings in Functional Genomics, 18 (2019) 58–82. https://doi.org/10.1093/bfgp/ely031 PMID: 30247501
10. Zhang L., Yang P., Feng H., Zhao Q., Liu H., Using network distance analysis to predict lncRNA–miRNA interactions, Interdisciplinary Sciences: Computational Life Sciences, 13 (2021) 535–545. https://doi.org/10.1007/s12539-021-00458-z PMID: 34232474
11. Quek X.C., Thomson D.W., Maag J.L., Bartonicek N., Signal B., Clark M.B., et al., lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs, Nucleic Acids Res, 43 (2014) D168–D173. https://doi.org/10.1093/nar/gku988 PMID: 25332394

12. Geng C., Ziyun W., Dongqing W., Chengxiang Q., Mingxi L., Xing C., et al., LncRNADisease: a database for long-non-coding RNA-associated diseases, Nucleic Acids Res, 41 (2013) D983–D986. https://doi.org/10.1093/nar/gks1099 PMID: 23175614

13. Dinger M.E., Pang K.C., Mercer T.R., Crowe M.L., Grimmond S.M., Mattick J.S., NRED: a database of long noncoding RNA expression, Nucleic Acids Res, 37 (2008) D122–D126. https://doi.org/10.1093/nar/gkn617 PMID: 18829717

14. Bu D., Yu K., Sun S., Xie C., Skogerbø G., Miao R., et al., NONCODE v3. 0: integrative annotation of long noncoding RNAs, Nucleic Acids Res, 40 (2011) D210–D215. https://doi.org/10.1093/nar/gkr1175 PMID: 22135294

15. Chen X., Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA, Sci Rep, 5 (2015) 13186. https://doi.org/10.1038/srep13186 PMID: 26278472

16. Chen X., Xie D., Zhao Q., You Z.-H., MicroRNAs and complex diseases: from experimental results to computational models, Brief Bioinform, 20 (2019) 515–539. https://doi.org/10.1093/bib/bbx130 PMID: 29045685

17. Chen X., Wang L., Qu J., Guan N.-N., Li J.-Q., Predicting miRNA–disease association based on inductive matrix completion, Bioinformatics, 34 (2018) 4256–4265. https://doi.org/10.1093/bioinformatics/bty503 PMID: 29939227

18. Chen X., Zhu C.-C., Yin J., Ensemble of decision tree reveals potential miRNA-disease associations, PLoS Comput Biol, 15 (2019) e1007209. https://doi.org/10.1371/journal.pcbi.1007209 PMID: 31329575

19. Chen X., Huang L., LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction, PLoS Comput Biol, 13 (2017) e1005912. https://doi.org/10.1371/journal.pcbi.1005912 PMID: 29253885

20. Chen X., Yin J., Qu J., Huang L., MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction, PLoS Comput Biol, 14 (2018) e1006418. https://doi.org/10.1371/journal.pcbi.1006418 PMID: 30142158

21. Huang Y.-A., Chen X., You Z.-H., Huang D.-S., Chan K.C., ILNCSIM: improved lncRNA functional similarity calculation model, Oncotarget, 7 (2016) 25902–25914. https://doi.org/10.18632/oncotarget.8296 PMID: 27028993

22. Zhao T., Xu J., Liu L., Bai J., Xu C., Xiao Y., et al., Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features, Mol Biosyst, 11 (2015) 126–136. https://doi.org/10.1039/c4mb00478g PMID: 25354589

23. Yu J., Ping P., Wang L., Kuang L., Li X., Wu Z., A Novel Probability Model for LncRNA–Disease Association Prediction Based on the Naïve Bayesian Classifier, Genes, 9 (2018) 345. https://doi.org/10.3390/genes9070345 PMID: 29986541

24. Chen Q., Lai D., Lan W., Wu X., Chen B., Chen Y.-P.P., et al., ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion, IEEE/ACM Transactions on Computational Biology, 18 (2019) 1106–1112.

25. Lan W., Li M., Zhao K., Liu J., Wu F.-X., Pan Y., et al., LDAP: a web server for lncRNA-disease association prediction, Bioinformatics, 33 (2016) 458–460.

26. Li W., Wang S., Xu J., Mao G., Tian G., Yang J., Inferring latent disease-lncRNA associations by faster matrix completion on a heterogeneous network, Frontiers in genetics, 10 (2019) 769. https://doi.org/10.3389/fgene.2019.00769 PMID: 31572428

27. Lu C., Yang M., Luo F., Wu F.-X., Li M., Pan Y., et al., Prediction of lncRNA–disease associations based on inductive matrix completion, Bioinformatics, 34 (2018) 3357–3364. https://doi.org/10.1093/bioinformatics/bty327 PMID: 29718113

28. Liu J.-X., Cui Z., Gao Y.-L., Kong X.-Z., WGRCMF: A weighted graph regularized collaborative matrix factorization method for predicting novel LncRNA-disease associations, IEEE journal of biomedical, 25 (2020) 257–265.

29. Fu G., Wang J., Yu G., Domeniconi C., Matrix factorization-based data fusion for the prediction of lncRNA–disease associations, Bioinformatics, 34 (2017) 1529–1537.

30. Liu H., Ren G., Chen H., Liu Q., Yang Y., Zhao Q., Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized, Knowledge-Based Systems, 191 (2020) 105261.

31. Zhao Q., Yu H., Ming Z., Hu H., Ren G., Liu H., The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions, Molecular Therapy-Nucleic Acids, 13 (2018) 464–471. https://doi.org/10.1016/j.omtn.2018.09.020 PMID: 30388620

32. Ping P., Wang L., Kuang L., Ye S., Iqbal M.F.B., Pei T., A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 16 (2018) 688–693.

33. Zhou M., Wang X., Li J., Hao D., Wang Z., Shi H., et al., Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network, Mol Biosyst, 11 (2015) 760–769. https://doi.org/10.1039/c4mb00511b PMID: 25502053

34. Yu G., Fu G., Lu C., Ren Y., Wang J., BRWLDA: bi-random walks for predicting lncRNA-disease associations, Oncotarget, 8 (2017) 60429. https://doi.org/10.18632/oncotarget.19588 PMID: 28947982

35. Chen X., You Z.-H., Yan G.-Y., Gong D.-W., IRWRLDA: improved random walk with restart for lncRNA-disease association prediction, Oncotarget, 7 (2016) 57919–57931. https://doi.org/10.18632/oncotarget.11141 PMID: 27517318

36. Chen X., KATZLDA: KATZ measure for the lncRNA-disease association prediction, Sci Rep, 5 (2015) 16840. https://doi.org/10.1038/srep16840 PMID: 26577439

37. Zhang Z., Zhang J., Fan C., Tang Y., Deng L., KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 16 (2017) 407–416. https://doi.org/10.1109/TCBB.2017.2704587 PMID: 28534780

38. Liu Y., Feng X., Zhao H., Xuan Z., Wang L., A novel network-based computational model for prediction of potential LncRNA–disease association, International journal of molecular sciences, 20 (2019) 1549.

39. Zhang J., Zhang Z., Chen Z., Deng L., Integrating multiple heterogeneous networks for novel lncRNA-disease association inference, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 16 (2017) 396–406. https://doi.org/10.1109/TCBB.2017.2701379 PMID: 28489543

40. Xuan P., Sheng N., Zhang T., Liu Y., Guo Y., CNNDLP: a method based on convolutional autoencoder and convolutional neural network with adjacent edge attention for predicting lncRNA–disease associations, International journal of molecular sciences, 20 (2019) 4260. https://doi.org/10.3390/ijms20174260 PMID: 31480319

41. Sheng N., Cui H., Zhang T., Xuan P., Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA–disease association prediction, Brief Bioinform, 22 (2021) bbaa067. https://doi.org/10.1093/bib/bbaa067 PMID: 32444875

42. Xuan P., Jia L., Zhang T., Sheng N., Li X., Li J., LDAPred: a method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs, International journal of molecular sciences, 20 (2019) 4458. https://doi.org/10.3390/ijms20184458 PMID: 31510011

43. Xuan P., Cao Y., Zhang T., Kong R., Zhang Z., Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes, Frontiers in genetics, 10 (2019) 416. https://doi.org/10.3389/fgene.2019.00416 PMID: 31130990

44. Zou Q., Li J., Hong Q., Lin Z., Wu Y., Shi H., et al., Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods, Biomed Res Int, 2015 (2015) 810514. https://doi.org/10.1155/2015/810514 PMID: 26273645

45. Zhang Y., Chen M., Li A., Cheng X., Jin H., Liu Y., LDAI-ISPS: LncRNA–disease associations inference based on integrated space projection scores, International journal of molecular sciences, 21 (2020) 1508. https://doi.org/10.3390/ijms21041508 PMID: 32098405

46. Chen M., Peng Y., Li A., Deng Y., Li Z., A novel lncRNA-disease association prediction model using Laplacian regularized least squares and space projection-federated method, IEEE Access, 8 (2020) 111614–111625.

47. Li G., Luo J., Liang C., Xiao Q., Ding P., Zhang Y., Prediction of LncRNA-Disease Associations Based on Network Consistency Projection, IEEE Access, 7 (2019) 58849–58856.

48. Zhang Y., Chen M., Li A., Cheng X., Jin H., Liu Y., LDAI-ISPS: LncRNA–Disease Associations Inference Based on Integrated Space Projection Scores, International journal of molecular sciences, 21 (2020) 1508. https://doi.org/10.3390/ijms21041508 PMID: 32098405

49. Wang L., Xiao Y., Li J., Feng X., Li Q., Yang J., IIRWR: Internal Inclined Random Walk With Restart for LncRNA-Disease Association Prediction, IEEE Access, 7 (2019) 54034–54041.

50. Bin J., Nie S., Tang Z., Kang A., Fu Z., Hu Y., et al., Long noncoding RNA EPB41L4A-AS1 functions as an oncogene by regulating the Rho/ROCK pathway in colorectal cancer, Journal of cellular physiology, 236 (2021) 523–535. https://doi.org/10.1002/jcp.29880 PMID: 32557646

51. Wang D., Wang J., Lu M., Song F., Cui Q., Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, Bioinformatics, 26 (2010) 1644–1650. https://doi.org/10.1093/bioinformatics/btq241 PMID: 20439255

52. Jie S., Hongbo S., Zhenzhen W., Changjian Z., Lin L., Letian W., et al., Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network, Mol Biosyst, 10 (2014) 2074–2081. https://doi.org/10.1039/c3mb70608g PMID: 24850297

53. van Laarhoven T., Nabuurs S.B., Marchiori E., Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics, 27 (2011) 3036–3043. https://doi.org/10.1093/bioinformatics/btr500 PMID: 21893517

54. Chen X., Huang Y.-A., Wang X.-S., You Z.-H., Chan K.C., FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model, Oncotarget, 7 (2016) 45948–45958. https://doi.org/10.18632/oncotarget.10008 PMID: 27322210