A novel target convergence set based random walk with restart for prediction of potential LncRNA-disease associations

Jiechen Li 1,2, Xueyong Li 1, Xiang Feng 1,2, Bing Wang 3, Bihai Zhao 2 and Lei Wang 1,2*

Abstract

Background: In recent years, IncRNAs (long-non-coding RNAs) have been proved to be closely related to the occurrence and development of many serious diseases that are seriously harmful to human health. However, most of the IncRNA-disease associations have not been found yet due to high costs and time complexity of traditional bio-experiments. Hence, it is quite urgent and necessary to establish efficient and reasonable computational models to predict potential associations between IncRNAs and diseases.

Results: In this manuscript, a novel prediction model called TCSRWRLD is proposed to predict potential IncRNA-disease associations based on improved random walk with restart. In TCSRWRLD, a heterogeneous IncRNA-disease network is constructed first by combining the integrated similarity of IncRNAs and the integrated similarity of diseases. And then, for each IncRNA/disease node in the newly constructed heterogeneous IncRNA-disease network, it will establish a node set called TCS (Target Convergence Set) consisting of top 100 disease/IncRNA nodes with minimum average network distances to these disease/IncRNA nodes having known associations with itself. Finally, an improved random walk with restart is implemented on the heterogeneous IncRNA-disease network to infer potential IncRNA-disease associations. The major contribution of this manuscript lies in the introduction of the concept of TCS, based on which, the velocity of convergence of TCSRWRLD can be quicken effectively, since the walker can stop its random walk while the walking probability vectors obtained by it at the nodes in TCS instead of all nodes in the whole network have reached stable state. And Simulation results show that TCSRWRLD can achieve a reliable AUC of 0.8712 in the Leave-One-Out Cross Validation (LOOCV), which outperforms previous state-of-the-art results apparently. Moreover, case studies of lung cancer and leukemia demonstrate the satisfactory prediction performance of TCSRWRLD as well.

Conclusions: Both comparative results and case studies have demonstrated that TCSRWRLD can achieve excellent performances in prediction of potential IncRNA-disease associations, which imply as well that TCSRWRLD may be a good addition to the research of bioinformatics in the future.

Keywords: Potential IncRNA-disease association prediction, Heterogeneous network, Random walk with restart, Target convergence set, Global set
Background

For many years, the genetic information of organism is considered to be stored only in genes used for protein coding, and RNAs have always been thought to be an intermediary in the process of encoding proteins by DNAs [1, 2]. However, recent studies have shown that the genes used to encode proteins only account for a small part (less than 2%) of human genome and more than 98% of human genome are not made up of genes that encode proteins and yield a big mount of ncRNAs (non-coding-RNAs) [3, 4]. In addition, as the complexity of biological organisms increases, so does the importance of ncRNAs in biological processes [5, 6]. Generally, ncRNAs can be divided into two major categories such as small ncRNAs and long ncRNAs (lncRNAs) according to the length of nucleotides during transcription, where small ncRNAs consist of less than 200 nucleotides and include microRNAs and transfer RNAs etc. However, lncRNAs consist of more than 200 nucleotides [7–9]. In 1990, the first two kinds of lncRNAs such as H19 and Xist were discovered by researchers through gene mapping. Since gene mapping approach is extremely time-consuming and labor-intensive, then researches in the field of lncRNAs have been at a relatively slow pace for a long time [10, 11]. In recent years, with the rapid development of high-throughput technologies in gene sequencing, more and more lncRNAs have been found in eukaryotes and other species [12, 13]. Moreover, simulation results have shown as well that lncRNAs play important roles in various physiological processes such as cell differentiation and death, regulation of epigenetic shape and so on [8, 14, 15]. Simultaneously, growing evidences have further illustrated that lncRNAs are closely linked to diseases that pose a serious threat to human health [16–18], which means that lncRNAs can be used as potential biomarkers in the course of disease treatment in the future [19].

With the discovery of a large number of new types of lncRNAs, many databases approach to lncRNAs such as IncRNA-disease [20], IncRNAdb [21], NONCODE [22] and Lnc2Cancer [23] have been established by researchers successively, however, in these databases, the number of known associations between IncRNAs and diseases is still very limited due to high costs and time-consumption of traditional biological experiments. Thus, it is meaningful to develop mathematical models to predict potential lncRNA-disease associations quickly and massively. Based on the assumption that similar diseases tend to be more likely associated with similar lncRNAs [24, 25], up to now, a good deal of computational models for inferring potential lncRNA-disease associations have been proposed. For instance, Chen et al. proposed a computational model called LRLSLDA [26] for prediction of potential lncRNA-disease associations by adopting the method of Laplacian regularized least squares. Ping and Wang et al. constructed a prediction model for extracting feature information from bipartite interactive networks [27]. Zhao and Wang et al. developed a computational model based on Distance Correlation Set to uncover potential lncRNA-disease associations through integrating known associations between three kinds of nodes such as disease nodes, miRNA nodes and lncRNA nodes into a complex network [28]. Chen et al. proposed an lncRNA-disease association prediction model based on a heterogeneous network by considering the influence of path length between nodes on the similarity of nodes in the heterogeneous network [29–31]. However, for some time past, a network traversal method called RWR (Random Walk with Restart) has emerged in the field of computational biology including prediction of potential miRNA-disease associations [32, 33], drug-target associations [34] and lncRNA-disease associations [35–37] etc.

Inspired by the thoughts illustrated in above state-of-the-art literatures, in this paper, a computational model called TCSRWRLD is proposed to discover potential lncRNA-disease associations. In TCSRWRLD, a heterogeneous network is constructed first through combining known lncRNA-disease associations with the lncRNA integrated similarity and the disease integrated similarity, which can overcome a drawback of traditional RWR based approaches that these approaches cannot start walking process while there are no known lncRNA-disease associations. And then, each node in the heterogeneous network will establish its own TCS according to the information of network distance, which can reflect the specificity of different nodes in the walking process and make the prediction more accurate and less time-consuming. Moreover, considering that for a given walker, while its TCS has reached the ultimate convergence state, there may be still some nodes that are not included in its TCS but actually associated with it, then in order to ensure that there is no omission in our prediction results, each node in the heterogeneous network will further establish its own GS as well. Finally, for evaluating the prediction performance of our newly proposed model TCSRWRLD, cross validation are implemented based on known lncRNA-disease associations downloaded from the IncRNA-disease database (2017version), and as a result, TCSRWRLD can achieve reliable AUCs of 0.8323, 0.8597, 0.8665 and 0.8712 under the frameworks of 2-folds CV, 5-folds CV, 10-folds CV and LOOCV respectively. In addition, simulation results in case studies of leukemia and lung cancer show that there are 5 and 7 out of the top 10 predicted lncRNAs having been confirmed to be associated with Leukemia and Lung cancer respectively by recent evidences, which demonstrate as well that our model TCSRWRLD has excellent prediction performance.
Results
In order to verify the performance of TCSRWRDL in predicting potential lncRNA-disease associations, LOOCV, 2-folds CV, 5-folds CV and 10-folds CV were implemented on TCSRWRDL respectively. And then, based on the dataset of 2017-version downloaded from the IncRNA Disease database, we obtained the Precision-Recall curve (P-R curve) of TCSRWRDL. In addition, based on the dataset of 2017-version downloaded from the IncRNA Disease database and the dataset of 2016-version downloaded from the lnc2Cancer database, we compared TCSRWRDL with state-of-the-art prediction models such as KATZLDA, PMFILDA [38] and Ping’s model separately. After that, we further analyzed the influences of key parameters on the prediction performance of TCSRWRDL. Finally, case studies of leukemia and lung cancer were performed to validate the feasibility of TCSRWRDL as well.

Cross validation
In this section, ROC curve (Receiver Operating Characteristic) and the score of AUC (Area Under ROC Curve) will be adopted to measure the performance of TCSRWRDL in different cross validations. Here, let TPR (True Positive Rates or Sensitivity) represent the percentage of candidate lncRNAs-disease associations with scores higher than a given score cutoff, and FPR (False Positive Rates or 1-Specificity) denote the ratio of predicted lncRNA-disease associations with scores below the given threshold, then ROC curves can be obtained by connecting the corresponding pairs of TPR and FPR on the graph. As illustrated in Fig. 1, simulation results show that TCSRWRDL can achieve reliable AUCs of 0.8323, 0.8597, 0.8665 and 0.8712 in the frameworks of 2-folds CV, 5-folds CV, 10-folds and LOOCV respectively, which implies that TCSRWRDL can achieve excellent performance in predicting potential lncRNA-disease associations.

Moreover, in order to further estimate the prediction performance of TCSRWRDL, we will obtain the P-R curve of TCSRWRDL as well. Unlike the AUC, the AUPR (Area Under the Precision-Recall curve) represents the ratio of all true positives to all positive predictions at every given recall rate. As illustrated in Fig. 2, simulation results show that TCSRWRDL can achieve a reliable AUPR of 0.5007.

Comparison with other related methods
From above descriptions, it is easy to know that TCSRWRDL can achieve satisfactory prediction performance. In this section, we will compare TCSRWRDL with some classical prediction models to further demonstrate the performance of TCSRWRDL. Firstly, based on the dataset of 2017-version downloaded from the IncRNA Disease database, we will compare TCSRWRDL with the state-of-the-art models such as KATZLDA, PMFILDA and Ping’s model. As shown in Fig. 3, it is easy to see that TCSRWRDL can achieve a reliable AUC of 0.8712 in LOOCV, which is superior to the AUCs of 0.8257, 0.8702 and 0.8346 achieved by KATZLDA, Ping’s model and PMFILDA in LOOCV respectively.
Moreover, in order to prove that TCSRWLDR can perform well in different data backgrounds, we also adopt the dataset of 2016-version downloaded from the lnc2Cancer database, which consists of 98 human cancers, 668 lncRNAs and 1103 confirmed associations between them, to compare TCSRWLDR with KATZLDA, PMFILA and Ping’s model. As illustrated in Fig. 4, it is easy to see that TCSRWLDR can achieve a reliable AUC of 0.8475 in LOOCV, which is superior to the AUCs of 0.8204 and 0.8374 achieved by KATZLDA and PMFILA respectively, while is inferior to the AUC of 0.8663 achieved by Ping’s model.

Fig. 2 precision-recall curve achieved by TCSRWLDR

Fig. 3 The AUCs achieved by TCSRWLDR, KATZLDA, Ping’s model and PMFILA in LOOCV based on the dataset of 2017-version downloaded from the IncRNA-disease database
Analysis on effects of parameters

In TCSRWRDL, there are some key parameters such as $\gamma_l$, $\gamma_d$, and $\partial$. As for $\gamma_l$ and $\gamma_d$ in the Equation (5) and Equation (11), we have already known that the model can achieve the best performance when the values of $\gamma_l$ and $\gamma_d$ are both set to 1 [39]. Hence, in order to estimate effect of the key parameter $\partial$ on the prediction performance of TCSRWRDL, we will set the value range of $\partial$ from 0.1 to 0.9 and select the value of AUC in LOOCV as the basis of parameter selection in this section. As illustrated in Table 1, It is easy to see that TCSRWRDL can achieve the highest value of AUC in LOOCV while $\partial$ is set to 0.4. Moreover, it is also easy to see that TCSRWRDL can maintain robustness for different values of $\partial$, which means that TCSRWRDL is not sensitive to the values of $\partial$ as well.

Case studies

Up to now, cancer is considered as one of the most dangerous diseases to human health because it is hard to be treated [40]. At present, the incidence of various cancers has a high level not only in the developing countries where medical development is relatively backward, but also in the developed countries where the medical level is already very high. Hence, in order to further evaluate the performance of TCSRWRDL, case study of two kinds of dangerous cancers such as lung cancer and leukemia will be implemented in this section. As for these two kinds of dangerous cancers, the incidence of lung cancer has remained high in recent years, and the number of lung cancer deaths per year is about 1.8 million, which is the highest of any cancer types. However, the survival rate within five years after the diagnosis of lung cancer is only about 15%, which is much lower than that of other cancers [41]. Recently, growing evidences have shown that lncRNAs play crucial roles in the development and occurrence of lung cancer [42]. As illustrated in Table 2, while implementing TCSRWRDL to predict lung cancer related lncRNAs, there are 7 out of the top 10 predicted candidate lung cancer related lncRNAs having been confirmed by the latest experimental evidences. Additionally, as a blood-related cancer [43], Leukemia has also been found to be closely related to a variety of lncRNAs in recent years. As illustrated in Table 2, while implementing TCSRWRDL to predict Leukemia related lncRNAs, there are 5 out of the top 10 predicted candidate Leukemia related lncRNAs having been confirmed by state-of-the-art experiment results as well. Thus, from above simulation results of case studies, we can easily reach an agreement that TCSRWRDL may

![Fig. 4 the AUCs achieved by TCSRWRDL, KATZLDA, Ping’s model and PMFILDA based on the dataset of 2016-version downloaded from the Lnc2Cancer database](image)

Table 1 AUCs achieved by TCSRWRDL in LOOCV while the parameter $\partial$ is set to different values from 0.1 to 0.9

| $\partial$ | 0.1     | 0.2     | 0.3     | 0.4     | 0.5     | 0.6     | 0.7     | 0.8     | 0.9     |
|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| AUC       | 0.8418  | 0.8655  | 0.8707  | 0.8712  | 0.8700  | 0.8680  | 0.8700  | 0.8672  | 0.8660  |
Table 2: Evidences of top 10 potential leukemia-related lncRNAs and lung cancer-related lncRNAs predicted by TCSRWRLD

| Disease name | LncRNA name | RANK | Evidence |
|--------------|-------------|------|----------|
| Lung cancer  | YUG1        | 1    | Lnc2Cancer |
| Lung cancer  | XIST        | 2    | Lnc2Cancer |
| Lung cancer  | PVT1        | 4    | Lnc2Cancer |
| Lung cancer  | PCAT29      | 5    | MDR      |
| Lung cancer  | HOTAIRM1    | 7    | MDR      |
| Lung cancer  | NEAT1       | 9    | Lnc2Cancer |
| Lung cancer  | anti-NOS2A  | 10   | Lnc2Cancer |
| Leukemia     | MALAT1      | 1    | Lnc2Cancer |
| Leukemia     | HOTAIR      | 2    | Lnc2Cancer |
| Leukemia     | H19         | 4    | Lnc2Cancer |
| Leukemia     | MEG3        | 5    | Lnc2Cancer |
| Leukemia     | PVT1        | 7    | Lnc2Cancer |

have great value in predicting potential lncRNA-disease associations.

Discussion
Since it is very time-consuming and labor-intensive to verify associations between lncRNAs and diseases through traditional biological experiments, then it has become a hot topic in bioinformatics to establish computational models to infer potential lncRNA-disease associations, which can help researchers to have a deeper understanding of diseases at the lncRNA level. In this manuscript, a novel prediction model called TCSRWRLD is proposed, in which, a heterogeneous network is constructed first through combining the disease integrated similarity, the lncRNA integrated similarity and known lncRNA-disease associations, which can guarantee that TCSRWRLD is able to overcome the shortcomings of traditional RWR based prediction models that the random walk process cannot be started while there are no known lncRNA-disease associations. And then, based on the newly constructed heterogeneous network, a random walk based prediction model is further designed based on the concepts of TCS and GS. In addition, based on the dataset of 2017-version downloaded from the LncRNADisease database, a variety of simulations have been implemented, and simulation results show that TCSRWRLD can achieve reliable AUCs of 0.8323, 0.8597 0.8665 and 0.8712 under the frameworks of 2-fold CV, 5-fold CV, 10-fold CV and LOOCV respectively. Additionally, simulation results of case studies of lung cancer and leukemia show as well that TCSRWRLD has a reliable diagnostic ability in predicting potential lncRNA-disease associations. Certainly, the current version of TCSRWRLD still has some shortages and deficiencies. For example, the prediction performance of TCSRWRLD can be further improved if more known lncRNA-disease associations have been added into the experimental datasets. In addition, more accurate establishment of Mesh database will help us obtain more accurate disease semantic similarity scores, which is very important for the calculation of lncRNA functional similarity as well. Of course, all these above problems will be the focus of our future researches.

Conclusion
In this paper, the main contributions are as follows: (1) A heterogeneous lncRNA-disease network is constructed by integrating three kinds of networks such as the known lncRNA-disease association network, the disease-disease similarity network and the lncRNA-lncRNA similarity network. (2) Based on the newly constructed heterogeneous lncRNA-disease network, the concept of network distance is introduced to establish the TCS (Target Convergence Set) and GS (Global Set) for each node in the heterogeneous lncRNA-disease network. (3) Based on the concepts of TCS and GS, a novel random walk model is proposed to infer potential lncRNA-disease associations. (4) Through comparison with traditional state-of-the-art prediction models and the simulation results of case studies, TCSRWRLD is demonstrated to be of excellent prediction performance in uncovering potential lncRNA-disease associations.

Methods and materials
Known disease-lncRNA associations
Firstly, we download the 2017-version of known lncRNA-disease associations from the LncRNAdisease database (http://www.cuilab.cn/ lncrnadisease). And then, after removing duplicated associations and picking out the lncRNA-disease associations from the raw data, we finally obtain 1695 known lncRNA-disease associations (see Additional file 1) including 828 different lncRNAs (see Additional file 2) and 314 different diseases (see Additional file 3). Hence, we can construct a 314 × 828 dimensional lncRNA-disease association adjacency matrix A, in which, there is A(i, j) = 1, if and only if there is an known association between the disease d_i and the lncRNA l_j in the LncRNADisease database, otherwise there is A(i, j) = 0. In addition, for convenience of description, let N_L = 828 and N_D = 314, then it is obvious that the dimension of the lncRNA-disease association adjacency matrix A can be represented as N_D × N_L. And the like mentioned above, we can get a cancer-disease associations adjacency matrix which dimension is 98 × 668 (It comes from 2016-version of known lncRNA-disease associations from the Lnc2Cancer database) (see Additional file 4).

Similarity of diseases
Semantic similarity of diseases
In order to estimate the semantic similarity between different diseases, based on the concept of DAGs (Directed

\[ A = (a_{ij})_{314 \times 828} \]
Acyclic Graph) of different diseases proposed by Wang et al. [44, 45], we can calculate the disease semantic similarity through calculating the similarity between compositions of DAGs of different diseases as follows:

**Step 1** For all these 314 diseases newly obtained from the lncRNAdisease database, their corresponding MESH descriptors can be downloaded from the Mesh database in the National Library of Medicine (http://www.nlm.nih.gov/). As illustrated in Fig. 5, based on the information of MESH descriptors, each disease can establish a DAG of its own.

**Step 2** For any given disease $d$, Let its DAG be $\text{DAG}(d) = (d, D(d), E(d))$, where $D(d)$ represents a set of nodes consisting of the disease $d$ itself and its ancestral disease nodes, and $E(d)$ denotes a set of directed edges pointing from ancestral nodes to descendant nodes.

**Step 3** For any given disease $d$ and one of its ancestor nodes $t$ in $\text{DAG}(d)$, the semantic contributions of the ancestor node $t$ to the disease $d$ can be defined as follows:

$$D_d(t) = \begin{cases} \frac{1}{\max\{\Delta \cdot D_d(t')|t' \in \text{children of } t\}} & \text{if } t = d \\ \frac{1}{\max\{\Delta \cdot D_d(t')|t' \in \text{children of } t\}} & \text{if } t \neq d \end{cases}$$

(1)

Where $\Delta$ is the attenuation factor with value between 0 and 1 to calculate the disease semantic contribution, and according to the state-of-the-art experimental results, the most appropriate value for $\Delta$ is 0.5.

**Step 4** For any given disease $d$, let its DAG be $\text{DAG}(d)$, then based on the concept of DAG, the semantic value of $d$ can be defined as follows:

$$D(d) = \sum_{t \in \text{DAG}(d)} D_d(t)$$

(2)

Taking the disease DSN (Digestive Systems Neoplasms) illustrated in Fig. 5 for example, according to the Equation (1), it is easy to know that the semantic contribution of digestive systems neoplasms to itself is 1. Besides, since the neoplasms by site and the digestive system disease located in the second layer of the DAG of DSN, then it is obvious that both of the semantic contributions of these two kinds of diseases to DSN are $0.5^2 = 0.25$. Moreover, since the neoplasms located in the third layer of the DAG of DSN, then its semantic contribution to DSN is $0.5^3 = 0.125$. Hence, according to above formula (2), it is easy to know the semantic value of DSN will be $2.25 (=1 + 0.5 + 0.5 + 0.25)$.

**Step 5** For any two given diseases $d_i$ and $d_j$, based on the assumption that the more similar the structures of their DAGs, the higher the semantic similarity between them will be, the semantic similarity between $d_i$ and $d_j$ can be defined as follows:

$$\text{DisSemSim}(i, j) = \frac{\sum_{t \in \text{DAG}(d_i) \cap \text{DAG}(d_j)} (D_{d_i}(t) + D_{d_j}(t))}{D(d_i) + D(d_j)}$$

(3)

**Gaussian interaction profile kernel similarity of diseases**

Based on the assumption that similar diseases tend to be more likely associated with similar lncRNAs, according to above newly constructed lncRNA-disease association adjacency matrix $A$, for any two given diseases $d_i$ and $d_j$, the Gaussian interaction profile kernel similarity between them can be obtained as follows:

$$\text{GKD}(d_i, d_j) = \exp\left(-\gamma \cdot \|\text{IP}(d_i) - \text{IP}(d_j)\|^2\right)$$

(4)
\[ Y_d = Y_i / \left( \sum_{k=1}^{N_D} ||IP(d_k)||^2 \right) \]  \hspace{1cm} (5)

Here, \( IP(d_i) \) denotes the vector consisting of elements in the \( r \)-th row of the IncRNA-disease adjacency matrix \( A \). \( Y_d \) is the parameter to control the kernel bandwidth based on the new bandwidth parameter \( \gamma_d \) by computing the average number of IncRNAs-disease associations for all the diseases. In addition, inspired by the thoughts of former methods proposed by O. Vanunu et al. [46], we will adopt a logistics function to optimize the Gaussian interaction profile kernel similarity between diseases, and based on above Equation (4), we can further obtain a \( N_D \times N_D \) dimensional adjacency matrix \( FKD \) as follows:

\[ FKD(i, j) = \frac{1}{1 + \exp(-12GK(i,j)+\log(9999))} \]  \hspace{1cm} (6)

**Integrated similarity of diseases**

Based on the disease semantic similarity and disease Gaussian interaction profile kernel similarity obtained above, a \( N_D \times N_D \) dimensional integrated disease similarity adjacency matrix \( KD \) (\( N_D \times N_D \)) can be obtained as follows:

\[ KD(i, j) = \frac{DisSemSim(i, j) + FKD(i, j)}{2} \]  \hspace{1cm} (7)

**Similarity of LncRNAs**

**Functional similarity of LncRNAs**

We can obtain corresponding disease groups of two given IncRNAs \( l_i \) and \( l_j \) from the known associations of IncRNA-disease. Based on the assumption that similar diseases tend to be more likely associated with similar IncRNAs, We define the functional similarity of two given IncRNAs \( l_i \) and \( l_j \) as the semantic similarity between the disease groups corresponding to them. The specific calculation process is as follows:

For any two given IncRNAs \( l_i \) and \( l_j \), let \( DS(i) = \{ d_k \mid A(k, i) = 1 \} \) and \( DS(j) = \{ d_k \mid A(k, j) = 1 \} \), then the functional similarity between \( l_i \) and \( l_j \) can be calculated according to the following steps [31]:

**Step 1** For any given disease group \( DS(k) \) and disease \( d \in DS(k) \), we first calculate the similarity between \( d_i \) and \( DS(k) \) as follows:

\[ S(d_i, DS(k)) = \max_{d \in DS(k)} \{ DisSemSim(d_i, d) \} \]  \hspace{1cm} (8)

**Step 2** Therefore, based on above Equation (8), we define the functional similarity between \( l_i \) and \( l_j \) as \( FuncKL(i, j) \), which can be calculated as follows:

\[ FuncKL(i, j) = \frac{\sum_{d \in DS(i)} S(d_i, DS(j)) + \sum_{d \in DS(j)} S(d_i, DS(i))}{|DS(i)| + |DS(j)|} \]  \hspace{1cm} (9)

Here, \(|D(i)|\) and \(|D(j)|\) represent the number of diseases in \( DS(i) \) and \( DS(j) \) respectively. Thereafter, according to above Equation (9), it is obvious that a \( N_L \times N_L \) dimensional IncRNA functional similarity matrix \( FuncKL \) can be obtained in final.

**Gaussian interaction profile kernel similarity of IncRNAs**

Based on the assumption that similar IncRNAs tend to be more likely associated with similar diseases, according to above newly constructed IncRNA-disease association adjacency matrix \( A \), for any two given IncRNAs \( l_i \) and \( l_j \), the Gaussian interaction profile kernel similarity between them can be obtained as follows:

\[ FKL(l_i, l_j) = \exp \left( -\gamma_i ||IP(l_i)|| - ||IP(l_j)|| \right) \]  \hspace{1cm} (10)

\[ Y_i = Y_i / \left( \sum_{k=1}^{N_L} ||IP(l_i)||^2 \right) \]  \hspace{1cm} (11)

Here, \( IP(l_i) \) denotes the vector consisting of elements in the \( r \)-th column of the IncRNA-disease adjacency matrix \( A \). \( Y_i \) is the parameter to control the kernel bandwidth based on the new bandwidth parameter \( \gamma_i \) by computing the average number of IncRNAs-disease associations for all the IncRNAs. So far, based on above Equation (10), we can obtain a \( N_L \times N_L \) dimensional IncRNA Gaussian interaction profile kernel similarity matrix \( FKL \) as well.

**Integrated similarity of IncRNAs**

Based on the IncRNA functional similarity and IncRNA Gaussian interaction profile kernel similarity obtained above, a \( N_L \times N_L \) dimensional integrated IncRNA similarity adjacency matrix \( KL \) (\( N_L \times N_L \)) can be obtained as follows:

\[ KL(i, j) = \frac{FuncKL(i, j) + FKL(i, j)}{2} \]  \hspace{1cm} (12)

**Construction of computational model TCSRWRLD**

**The establishment of heterogeneous network**

Through combing the \( N_D \times N_D \) dimensional integrated disease similarity adjacency matrix \( KD \) and the \( N_L \times N_L \) dimensional integrated IncRNA similarity adjacency matrix \( KL \) with the \( N_D \times N_L \) dimensional IncRNA-disease association adjacency matrix \( A \), we can construct a new \( (N_L + N_D) \times (N_L + N_D) \) dimensional integrated matrix \( AA \) as follow:
\[ AA(i, j) = \begin{bmatrix} KL(i, j) & A^T(i, j) \\ A(i, j) & KD(i, j) \end{bmatrix} \tag{13} \]

According to above Equation (13), we can construct a corresponding heterogeneous lncRNA-disease network consisting of \( N_D \) different disease nodes and \( N_L \) different lncRNA nodes, in which, for any given pair of nodes \( i \) and \( j \), there is an edge existing between them, if and only if there is \( AA(i, j) > 0 \).

**Establishment of TCS (target convergence set)**

Before the implementation of random walk, for each node in above newly constructed heterogeneous lncRNA-disease network, as illustrated in Fig. 6, it will establish its own TCS first according to the following steps:

**Step 1** For any given lncRNA node \( l_j \), we define its original TCS as the set of all disease nodes that have known associations with it, i.e., the original TCS of \( l_j \) is \( TCS_0(l_j) = \{ d_k \mid A(k, j) = 1, k \in [1, N_D] \} \). Similarly, for a given disease node \( d_i \), we can define its original TCS as \( TCS_0(d_i) = \{ l_k \mid A(i, k) = 1, k \in [1, N_L] \} \).

**Step 2** After the original TCS has been established, for any given lncRNA node \( l_j \), \( \forall d_k \in TCS_0(l_j) \) and \( \forall t \in [1, N_D] \), then we can define the network distance \( ND(k, t) \) between \( d_k \) and \( d_i \) as follows:

\[ ND(k, t) = \frac{1}{KD(k, t)} \tag{14} \]

According to above Equation (14), for any disease nodes \( d_k \in TCS_0(d_i) \) and \( \forall t \in [1, N_D] \), obviously it is reasonable to deduce that the smaller the value of \( ND(k, t) \), the higher the similarity between \( d_k \) and \( d_i \) would be, i.e., the higher the possibility that there is potential association between \( d_k \) and \( d_i \) will be.

**Step 3** According to above Equation (14) and Equation (15), for any given disease node \( d_i \) or any given lncRNA node \( l_j \), we define that the TCS of \( d_i \) as the set of top 100 lncRNA nodes in the heterogeneous lncRNA-disease network that have minimum average network distance to the disease nodes in \( TCS_0(d_i) \), and the TCS of \( l_j \) as the set of top 100 disease nodes in the heterogeneous lncRNA-disease network that have minimum average network distance to the lncRNA nodes in \( TCS_0(l_j) \). Then, it is easy to know that these 100 lncRNA nodes in \( TCS(l_j) \) may belong to \( TCS_0(l_j) \) or may not belong to \( TCS_0(l_j) \), and these 100 disease nodes in \( TCS(l_j) \) may belong to \( TCS_0(l_j) \) or may not belong to \( TCS_0(l_j) \).
Random walk in the heterogeneous LncRNA-disease network

The method of random walk simulates the process of random walker's transition from one starting node to other neighboring nodes in the network with given probability. Based on the assumption that similar diseases tend to be more likely associated with similar lncRNAs, as illustrated in Fig. 7, the process of our prediction model TCSRWRLD can be divided into the following major steps:

**Step 1** For a walker, before it starts its random walk across the heterogeneous lncRNA-disease network, it will first construct a transition probability matrix $W$ as follows:

$$W(i, j) = \frac{AA(i, j)}{\sum_{k=1}^{N_D+N_L} AA(i, k)}$$  

(16)

**Step 2** In addition, for any node $\xi_i$ in the heterogeneous lncRNA-disease network, whether $\xi_i$ is a lncRNA node $l_i$ or a disease node $d_i$, it can obtain an initial probability vector $P_i(0)$ for itself as follows:

$$P_i(0) = \left( p_{i,1}(0), p_{i,2}(0), \ldots, p_{i,j}(0), \ldots, p_{i,N_D+N_L}(0) \right)^T$$  

(17)

$$p_{i,j}(0) = W(i, j) \; j = 1, 2, \ldots, N_D+N_L$$  

(18)

**Step 3** Next, the walker will randomly select a node $\xi_i$ in the heterogeneous lncRNA-disease network as the starting node to initiate its random walk, where $\xi_i$ may be an lncRNA node $l_i$ or a disease node $d_i$. After the initiation of the random walk process, supposing that currently the walker has arrived at the node $\Gamma_i$ from the previous hop node $\Gamma_j$ after $t-1$ hops during its random walk across the heterogeneous lncRNA-disease network, then here and now, whether $\Gamma_i$ is a lncRNA node $l_i$ or a disease node $d_i$, and $\Gamma_j$ is a lncRNA node $l_j$ or a disease node $d_j$, the walker can further obtain a walking probability vector $P_i(t)$ as follows:

$$P_i(t) = (1-\partial) \cdot W^T \cdot P_i(t-1) + \partial \cdot P_i(0)$$  

(19)

Where $\partial$ ($0 < \partial < 1$) is a parameter for the walker to adjust the value of walking probability vector at each hop. Moreover, based on above newly obtained walking probability vector $P_i(t)$, let $P_i(t) = \left( p_{i,1}(t), p_{i,2}(t), \ldots, p_{i,j}(t), \ldots, p_{i,N_D+N_L}(t) \right)^T$, and for con-

---

**Fig. 7** Flow chart of our prediction model TCSRWRLD
venience, supposing that there is \( p_{ik}(k) = \text{maximum}(p_{i1}(t), p_{i2}(t), ... , p_{ik}(t), ... , p_{iN_{D}+N_{L}}(t)) \), then the walker will choose the node \( \psi_k \) as its next hop node, where \( \psi_k \) may be a lncRNA node \( \psi_k \) or a disease node \( \psi_k \). Especially, as for the starting node \( \psi_0 \), since it can be regarded that the walker has arrived at \( \psi_0 \) from \( \psi_0 \) after 0 hops, then it is obvious that at the starting node \( \psi_0 \), the walker will obtain two kinds of probability vectors such as the initial probability vector \( P_i(0) \) and the walking probability vector \( P_i(1) \). However, at each intermediate node \( \Gamma_i \), the walker will obtain two other kinds of probability vectors such as the initial probability vector \( P_i(0) \) and the walking probability vector \( P_i(t) \).

**Step 4** Based on above Equation (19), supposing that currently the walker has arrived at the node \( \Gamma_j \) from the previous hop node \( \Gamma_j \) after \( t-1 \) hops during its random walk across the heterogeneous lncRNA-disease network, let the walking probability vectors obtained by the walker at the node \( \Gamma_i \) and \( \Gamma_j \) be \( P_i(t) \) and \( P_j(t-1) \) respectively, if the L1 norm between \( P_i(t) \) and \( P_j(t-1) \) satisfies \( \| P_i(t) - P_j(t-1) \|_1 \leq 10^{-6} \), then we will regard that the walking probability vector \( P_i(t) \) has reached a stable state at the node \( \Gamma_i \). Thus, after the walking probability vectors obtained by the walker at every disease node and lncRNA node in the heterogeneous lncRNA-disease network have reached stable state, and for convenience, let these stable walking probability vectors be \( P_1(\infty), P_2(\infty), ..., P_{N_D+N_L}(\infty) \), then based on these stable walking probability vectors, we can obtain a stable walking probability matrix \( S(\infty) \) as follows:

\[
S(\infty) = \begin{bmatrix} S_1 & S_2 & \cdots & S_{N_D+N_L} \\
S_1 & S_2 & \cdots & S_{N_D+N_L} \\
S_1 & S_2 & \cdots & S_{N_D+N_L} \end{bmatrix}
= (P_1(\infty), P_2(\infty), ..., P_{N_D+N_L}(\infty))^T
\]  

(20)

Where \( S_1 \) is a \( N_D \times N_D \) dimensional matrix, \( S_2 \) is a \( N_L \times N_L \) dimensional matrix, \( S_3 \) is a \( N_D \times N_L \) dimensional matrix, and \( S_4 \) is a \( N_L \times N_D \) dimensional matrix. And moreover, from above descriptions, it is easy to infer that the matrix \( S_2 \) and the matrix \( S_4 \) are the final result matrices needed by us, and we can predict potential lncRNA-disease associations based on the scores given in these two final result matrices.

According to above described steps of the random walk process based on our prediction model TCSRWRDL, it is obvious that for each node \( \Gamma_i \) in the heterogeneous lncRNA-disease network, the stable walking probability vector obtained by the walker at \( \Gamma_i \) is \( P_i(\infty) = (p_{i1}(\infty), p_{i2}(\infty), ... , p_{iN_D+N_L}(\infty)) \). Moreover, for convenience, we denote a node set consisting of all the \( N_D+N_L \) nodes in the heterogeneous lncRNA-disease network as a Global Set (GS), then it is obvious that we can rewrite the stable walking probability vector \( P_i(\infty) \) as \( P_i^{GS}(\infty) \). Additionally, from observing the stable walking probability vector \( P_i^{GS}(\infty) \), it is easy to know that the walker will not stop its random walk until the \( N_D+N_L \) dimensional walking probability vector at each node in the heterogeneous lncRNA-disease network has reached a stable state, which will obviously be very time-consuming while the value of \( N_D+N_L \) is large to a certain extent. Hence, in order to decrease the execution time and quicken the velocity of convergence of TCSRWRDL, based on the concept of TCS proposed in above section, while constructing the walking probability vector \( P_i(t) = (p_{i1}(t), p_{i2}(t), ... , p_{iN_D+N_L}(t))^T \) at the node \( \Gamma_i \), we will keep the \( p_{i,j}(t) \) unchanged if the \( j \)th node in these \( N_D+N_L \) nodes belongs to the TCS of \( \Gamma_i \), otherwise we will set \( p_{i,j}(t) = 0 \). Thus, the walking probability vector obtained by the walker at \( \Gamma_i \) will turn to be \( P_{i^{TCS}}(t) \) while the stable walking probability vector obtained by the walker at \( \Gamma_i \) will turn to be \( P_{i^{TCS}}(\infty) \).

Obviously, compared with \( P_i^{GS}(\infty) \), the stable state of \( P_{i^{TCS}}(\infty) \) can be reached by the walker much more quickly. However, considering that there may be nodes that are not in the TCS of \( \Gamma_i \) but actually associated with the target node, therefore, in order to avoid omissions, during simulation, we will construct a novel stable walking probability vector \( P_{i^{ANS}}(\infty) \) through combining \( P_i^{GS}(\infty) \) with \( P_i^{TCS}(\infty) \) to predict potential lncRNA-disease associations as follows:

\[
P_{i^{ANS}}(\infty) = \frac{P_i^{GS}(\infty) + P_{i^{TCS}}(\infty)}{2}
\]  

(21)

**Supplementary information**
---

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12859-019-3216-4.

---

**Additional file 1.** The known lncRNA-disease associations for constructing the known lncRNA-disease network. We list 1695 known lncRNA-disease associations which were collected from LncRNADisease dataset is the latest version in the database.

**Additional file 2.** The known 828 lncRNAs name included in the 1695 known lncRNA-disease associations which were collected from LncRNA Disease dataset is the latest version in the database.

**Additional file 3.** The known 314 diseases name included in the 1695 known lncRNA-disease associations which were collected from LncRNA Disease dataset is the latest version in the database.

**Additional file 4.** The known 98 human cancer 668 lncRNAs and 1103 confirmed associations between them from Lnc2Cancer database.

---

**Abbreviations**

10-Fold CV: 10-fold cross-validation; 2-Fold CV: 2-fold cross-validation; 5-Fold CV: 5-fold cross-validation; AUC: Areas under ROC curve; AUPR: Area under precision-recall curve; FPR: False positive rates; GS: Global set; H19: Long non-coding RNA H19; lncRNAs: Long non-coding RNAs; LOOCV: Leave-One Out Cross Validation; ncRNAs: Non-coding RNAs; P-R curve: Precision-recall curve; ROC: Receiver-operating characteristics; RWR: Random walk with...
restarted. TCS: Target Convergence Set; TCSRWRDL: A novel computational model based on improved rand walk with restart is proposed to infer potential lncRNA-disease associations; TPR: True positive rates; Xist: Long non-coding RNA Xist

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

Author’s contributions
JCL conceived the study. JCL, XF, LW improved the study based on the original model. XYL, BW and BHZ implemented the algorithms corresponding to the study. LW supervised the study. JCL and LW wrote the manuscript of the study. All authors reviewed and improved the manuscript.

Funding
This research was partly sponsored by the National Natural Science Foundation of China (No.61873221, No. 61672447) and the Natural Science Foundation of Hunan Province (No.2018J058, No.2019J70010, No.2017J5036). Publication costs were funded by the National Natural Science Foundation of China (No.61873221, No.61672447). The funder of manuscript is Lei Wang(L.W.), whose contribution are stated in the section of Author’s Contributions. The funding body has not played any roles in the design of the study and collection/analysis and interpretation of data in writing the manuscript.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the LncRNADisease repository, http://www.cuilab.cn/.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1 College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, Hunan, People’s Republic of China. 2Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, XiangTan, People’s Republic of China. 3School of Electrical and Information Engineering, Anhui University of Technology, Anhui 243002 Maanshan, People’s Republic of China.

Received: 8 February 2019 Accepted: 12 November 2019

References
1. Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. Nat. 1961;192(4809):1227–32.
2. Yanofsky C. Establishing the triplet nature of the genetic code. Cell. 2007; 128(5):815–8.
3. Jean-Michel F. Fewer genes, more noncoding RNA. Sci. 2005;309(5740):1529–30.
4. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Sci. 2008;322(5909):1845–8.
5. Paul B, Viktor S, Royce TE, Rozowsky JS, Urban AE, Xiaowei Z, Rinn JL, Waraporn T, Manoj S, Sherman W. Global identification of human microRNA and disease associations. PLoS One. 2008;3(10):e3420.
6. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction based on an lncRNA-disease association network. BMC Genomics. 2018;19(1):457–55.
7. Zhao H, Kuang L, Wang L, Pei T. Prediction of microRNA-disease associations based on distance correlation set. BMC Bioinformatics. 2015;16(2):688–93.
8. Zhang J, Zhang W, Wang P, Zhi H, Wang J, Liu Y, Guo M, Yue M, Wang L. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. 2015;44(D1):D980–5.
9. Mitchell G, Pamela R, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–51.
10. Mitchell G, Paul M, Levin JZ, Julie D, Rastan S. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. 1992;71(3):515–26.
11. Mitchell G, Paul M, Levin JZ, Julie D, Rastan S, Xiaowei Z, Rinn JL, Waraporn T, Manoj S, Sherman W. Global identification of human microRNA and disease associations. PLoS One. 2008;3(10):e3420.
33. Chen X. miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. Mol BioSyst. 2016;12(2):624–33.
34. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. Mol BioSyst. 2012;8(7):1970–8.
35. Je S, Hongbo S, Zhenzhen W, Changjian Z, Lin L, Letian W, Weimei H, Dapeng H, Shulin L, Meng Z. Inferring novel IncRNA–disease associations based on a random walk model of a IncRNA functional similarity network. Mol BioSyst. 2014;10(8):2074–81.
36. Chen X, You ZH, Yan GY, Gong DW. RWRLDA: improved random walk with restart for IncRNA–disease association prediction. Oncotarget. 2016;7(36):57919–31.
37. Jie S, Hongbo S, Zhenzhen W, Zhong SY, Zhu L, Su P. Prediction of IncRNA–disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. BMC Bioinformatics. 2019;20(1):87.
38. Xuan Z, Li J, Yu J, Feng X, Zhao B, Wang L. A probabilistic matrix factorization method for identifying IncRNA–disease associations. Genes. 2019;10(2):126.
39. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics. 2011;27(21):3036–43.
40. Spiess PE, Dhillon J, Baumgarten AS, Johnstone PA, Giuliani AR. Pathophysiological basis of human papillomavirus in penile cancer: key to prevention and delivery of more effective therapies. CA Cancer J Clin. 2016;66(6):481–95.
41. Tony G, Monika HM, Moritz E, Jeff H, Youngsoo K, Alexey R, Gavrieli A, Marion S, Matthias G. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 2013;73(3):1180–9.
42. White NM, Cabanski CR, Silva-Fisher JM, Dang HX, Govindan R, Maher CA. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. Genome Biol. 2014;15(8):429.
43. Omer A, Singh P, Yadav NK, Singh RK. microRNAs: role in leukemia and their computational perspective. Wiley Interdiscip Rev. RNA. 2015;6(1):65–78.
44. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinform. 2010;26(13):1644–50.
45. Chen X. Predicting IncRNA–disease associations and constructing IncRNA functional similarity network based on the information of miRNA. Sci Rep. 2015;5:13186.
46. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6(10):e1000641.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.