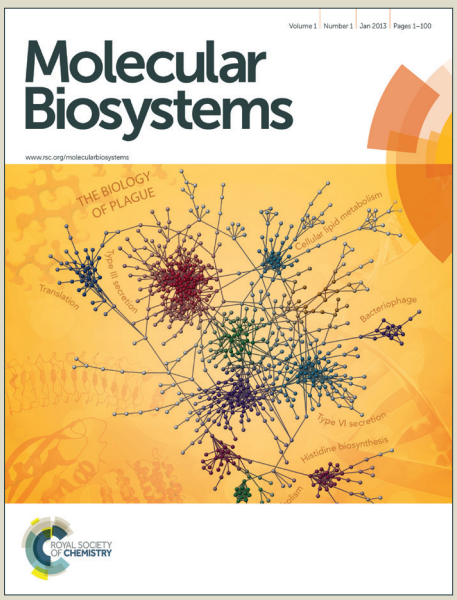
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Inferring novel lncRNA-disease associations based on random walk on lncRNA functional similarity network

Jie Sun[a, †], Hongbo Shi[a, †], Zhenzhen Wang[a], Changjian Zhang[a], Lin Liu[a], Letian Wang[a], Weiwei He[d], Dapeng Hao[a, *], Shulin Liu[b,*], Meng Zhou[a,c, †,*]

*Corresponding author

[†]These authors contributed equally to this work

Abstract

Accumulating evidence demonstrates that long non-coding RNAs (lncRNAs) play important roles in the development and progression of human complex diseases, but predicting novel human lncRNA-disease associations is a challenging and urgently needed work, especially at a time when rich and increasing amounts of lncRNA-related biological data are available. In this study, we proposed a global network-based computational framework, RWRlncD, to infer potential human lncRNA-disease associations by implementing random walk with restart on the lncRNA functional similarity network. The performance of RWRlncD was evaluated by experimentally verified lncRNA-disease associations based on leave-one-out cross-validation. We achieved an area under the ROC curve of 0.822, demonstrating excellent performance of RWRlncD. Of importance, the performance of RWRlncD is robust to different parameter selection. The predicted lncRNA-disease associations with high-ranks in case studies about prostate cancer and Alzheimer's disease were manually confirmed by literature mining, providing evidence to show the good performance and potential value of the RWRlncD method in predicting lncRNA-disease associations.

Introduction

Sequence analysis of the human genome identifies only ~ 20,000 protein-coding genes, consisting of less than 2% of the whole genome [1]. Further studies demonstrate that at least 90% of the genome is likely to be transcribed, yielding tens of thousands of non coding RNAs (ncRNA) [2]. ncRNAs can be further divided into two major categories based on the transcript length: small ncRNA and long ncRNA. Long non-coding RNAs (lncRNAs), which are longer than 200 nucleotides (nt), are commonly defined as RNA polymerase Ⅱ (RNAPⅡ) transcripts [3]. lncRNAs are found to be transcribed within introns of protein-coding genes, in either sense or antisense orientation, or within intergenic regions [4, 5].

A large number of lncRNAs have been identified by experimental as well as bioinformatics approaches, with many being collected into public biological databases, such as lncRNAdb [6], LNCipedia [7], NONCODE [8] and PLncDB [9]. Although the functions of most lncRNAs remain largely unknown, increasing evidence from public studies has demonstrated their critical roles in various biological processes through a variety of mechanisms such as chromatin remodeling, transcriptional co-activation or co-repression, protein inhibition and post-transcriptional modification, or as decoy elements [10]. Differential lncRNA expression has been observed in various diseases, and many disease-associated lncRNAs have been characterized, such as *H19*, *XIST*, etc., (for details, see an excellent recent review [11]). Therefore, identifying and characterizing novel disease-associated lncRNAs will provide novel insights into the molecular mechanisms underlying human complex diseases.

Computational approaches have been successfully applied to the discovery of disease-related protein coding genes or miRNAs in the past decades and considerably accelerated the elucidation of molecular underpinnings of human complex diseases. Meanwhile, many computational approaches have also been developed to predict novel lncRNA genes or their functions by using different biological resources [12-19], but to date few effective computational applications for identifying novel lncRNA-disease associations have been reported, such as LncRNADisease [20] and LRLSLDA [21]. With

the availability of rich and increasing amounts of lncRNA-related biological data, computational tools for the prediction of disease-related lncRNAs is urgently needed.

It is well known that functionally related genes are often associated with phenotypically similar diseases [22, 23]. Based on such findings, we developed a method to measure the functional similarity of lncRNAs and construct an lncRNA functional similarity network. We then proposed a novel computational framework, RWRlncD, to infer potential lncRNA-disease associations by random walk with restart on the lncRNA functional similarity network. The proposed RWRlncD method exploits a global network-based strategy and prioritizes candidate lncRNAs for a disease of interest by integrating lncRNA-disease network, disease similarity network and lncRNA functional similarity network. The RWRlncD was validated by experimentally verified lncRNA-disease associations, which demonstrated excellent performance of our method. Furthermore, the predicted lncRNA-disease associations with high-ranks in case studies about prostate cancer and Alzheimer's disease were confirmed both manually and by literature mining, providing convincing evidence to indicate the good performance and potential value of our proposed RWRlncD method for predicting novel lncRNA-disease associations.

Materials and methods

The human lncRNA-disease association data

The human lncRNA-disease association data were retrieved from a manually curated lncRNA-disease relations database, lncRNADisease (http://cmbi.bjmu.edu.cn/lncrnadisease), which has recorded approximately 600 high quality experimentally verified lncRNA-disease associations from ca. 500 publications [20]. We further verified the names of lncRNAs and diseases. After removing repeating lncRNA-disease entries, we finalized 352 lncRNA-disease associations, including 156 lncRNAs and 190 diseases, to construct a lncRNA-disease association network (LDAN) (Supplementary material 1).

Measuring functional similarity between two lncRNAs

First, we computed the similarity scores between diseases using Wang's measure

in DOSim, which is an R package for calculating the DO-based semantic similarity between diseases by DOID in an ontology sense [24, 25]. Next, we extended the previous method for measuring the functional similarity of miRNA genes or protein-coding genes to calculate functional similarity scores between lncRNAs [26, 27]. We supposed that lncRNA1 was associated with m diseases and lncRNA2 was associated with n diseases. We denoted one disease as d and one disease group as $D = \{d_1, d_2, \cdots\cdots, d_k\}$. The functional similarity scores between the two lncRNAs can be computed as follows:

$$SIM(d, D) = \max_{1 \le i \le k} \left(SIM(d, d_i) \right)$$

$$LncSIM(\ln cRNA1, \ln cRNA2) = \frac{\sum_{1 \le i \le m} SIM(d_{1i}, D_2) + \sum_{1 \le j \le n} SIM(d_{2j}, D_1)}{m + n}$$

Where d_{1i} represents disease i associated with lncRNA1 and d_{2j} represents disease j associated with lncRNA2. D_1 represents a disease group, in which all diseases are associated with lncRNA1, and D_2 represents another disease group, in which all diseases are associated with lncRNA2. *SIM*(d, D) is the maximum similarity score between one disease d and a disease group D, and *LncSIM*(*lncRNA1, lncRNA2*) is the functional similarity scores of two lncRNAs.

Random walk with restart for lncRNA-disease association

The random walks algorithm simulates a random walker that starts on a (or some) given seed node and transits from current nodes randomly to neighbors in the network based on the probabilities of the edges between two nodes. The random walker can also use a given probability to teleport to the start nodes called restart probability [28, 29]. Here, we denote P_0 as the initial probability vector and P_t as a vector in which the i-th element holds the probability of finding the random walker at node i at step t. Let α be the restart probability of the random walk in every time step at source nodes and W be the lncRNA-lncRNA functional similarity matrix. The sketch of the random walk with restart algorithm can be defined as follows:

$$P_{t+1} = (1 - \alpha) W P_t + \alpha \cdot P_0$$

Then the probability of random walk will become stable and can be defined as P_∞ by performing the iteration until the difference between P_t and P_{t+1} measured by the $L1$ norm falls below a given cutoff.

In this study, we proposed a novel computational framework, RWRlncD, to infer potential lncRNA-disease associations by random walk with restart on lncRNA-lncRNA functional similarity network. The schematic representation of RWRlncD method is in Figure 1. For RWRlncD, all the lncRNAs associated with a disease of interest were considered as seed lncRNAs, while other lncRNAs that have no any known relationship with this given disease were regarded as non-seed lncRNAs. These non-seed lncRNAs will be considered as candidate disease-related lncRNAs in the analysis. The initial probability P_0 of each seed lncRNA was set as $1/n$ (n is the number of seed lncRNAs), while the initial probability of all non-seed lncRNAs were set as zero. The stable probability P_∞ of each non-seed lncRNA was obtained by iterative process when the difference between P_t and P_{t+1} is less than 10^{-10}. The stable probability P_∞ can be used as a measure of proximity to seed lncRNAs. If $P_\infty(\ln cRNAi) > P_\infty(\ln cRNAj)$, lncRNAi will be more proximate to seed lncRNAs than lncRNAj in the LFSN. As a result, all candidate lncRNAs can be ranked according to the P_∞, and the top ranked lncRNAs can be expected to have a high probability to be associated with the disease of interest.

Results

Construction and characteristics of lncRNA-disease association network

In our study, there were 352 lncRNA-disease associations between 156 lncRNA and 190 diseases. We denoted the lncRNA set as $L = \{l_1, l_2, \cdots\cdots, l_n\}$ and the disease set as $D = \{d_1, d_2, \cdots\cdots, d_m\}$. The lncRNA-disease association network was

constructed based on the 352 lncRNA-disease associations and was represented as a bipartite LD graph $G(L,D,E)$, where $E=\{e_{ij}: l_i \in L, d_j \in D\}$ (Figure 2A). In this bipartite LD graph, there are two distinct sets of vertices corresponding to either lncRNA or disease. Vertices l_i and d_j are linked by an edge in the LDAN if lncRNA i is associated with disease j.

In order to obtain a global view of the lncRNA-disease association network, we analyzed its characteristics (Table 1). The degree of the lncRNA (or disease) node in LDAN is the number of diseases (or lncRNA) associated with a given lncRNA (or disease). On average, each lncRNA was involved with 2.3 diseases and each disease was associated with 1.9 lncRNAs, implying the regulatory complexity of lncRNAs in diseases. Examinations of the degree distribution of lncRNAs and diseases in LDAN revealed a power-law distribution with R^2 = ~0.9970 for lncRNAs and R^2 = ~0.9985 for diseases (Figure 2B & C), indicating that the LDAN displayed scale-free characteristics like many other biological network.

Construction and characteristics of lncRNA functional similarity network (LFSN)

Based on the reported observations that functionally related genes may be associated with diseases of similar pathogeneses [22, 23], we applied our method in the calculation of the functional similarity scores between any two lncRNAs in LDAN. For this, we first mapped 190 diseases to Disease Ontology (DO) and obtained the corresponding disease DOID. Then, we used the corresponding disease DOID to calculate the DO-based semantic similarity between diseases using DOSIM software. After removing the lncRNA-disease association in which a disease could not be mapped to DO, we finally obtained the pairwise functional similarity scores of 133 lncRNAs (Supplementary material 2). These scores are converted into lncRNA functional similarity matrix A, where the entity $A(i,j)$ in row i column j is the functional similarity score between lncRNA i and j. Based on the above functional

similarity matrix A, we constructed the lncRNA functional similarity network using similarity score cutoff β and found that the number of edges remained relatively stable when the score cutoff β was chosen to be equal or larger than 0.5 (Figure 3A). Therefore we used 0.5 as the cutoff β to construct the lncRNA functional similarity network. In LFSN, if the similarity score is equal or larger than 0.5 between two lncRNAs, these two lncRNAs will be linked by an edge in the LFSN (Figure 3B). In total, we saw 371 lncRNA-lncRNA functional associations between 117 lncRNAs in the LFSN and evaluated the degree distribution of the lncRNAs in the LFSN (Figure 3C). Notably, although approximately 89% lncRNAs were associated with two or more of other lncRNAs, a few of them may interact with multiple functionally similar lncRNAs.

Performance evaluation of the proposed method

In order to further evaluate the performance of our RWRlncD method to infer potential lncRNA-disease associations, we performed leave-one-out across validation analysis on 198 experimentally verified lncRNA-disease associations between 49 diseases associated with more than one lncRNA and 117 lncRNAs. For each disease, each known lncRNA associated with this given disease was left out as the testing case and other known experimentally verified lncRNAs associated with this given disease were taken as seed lncRNAs. All the lncRNAs without known associations with this given disease were placed in the candidate lncRNAs set. We wanted to test how well this testing case might rank relative to all lncRNAs in the candidate lncRNAs set of this given disease. If the ranking of the testing case in the ranking list exceeded a given threshold, this lncRNA-disease association would be deemed to be successfully predicted by the RWRlncD method. For the restart probability α in the RWRlncD method, we chose $\alpha = 0.7$ based on its excellent performance in previous studies [30, 31].

The sensitivity and specificity were calculated for each threshold. Sensitivity measures the percentage of the testing case whose ranking is higher than a given

threshold and specificity measures the percentage of candidate lncRNAs ranked below this given threshold. Finally, a receiver operating characteristics (ROC) curve was plotted by varying the threshold and then the value of area under curve (AUC) was calculated. We used AUC as a standard measure to evaluate performance of the RWRlncD method. The maximum value of AUC is 1, which indicates every testing case is ranked first in the ranking list and AUC = 0.5 indicates random performance. Figure 4 shows the results of performance evaluation of the RWRlncD method using the ROC curves obtained by calculating the sensitivity [$sensitivity = TP/(TP + FN)$] and specificity [$specificity = TN/(TN + FP)$] by varying the threshold, where TP is true positive, TN is true negative, FN is false negative, FP is false positive, sensitivity is the proportion of the testing case ranked higher than a given rank cutoff and specificity is the proportion of the testing case ranked lower than a given rank cutoff. Our RWRlncD method tested on 198 experimentally verified lncRNA-disease associations achieved an AUC of 0.822, demonstrating the excellent performance of the RWRlncD method in recovering the known experimentally verified lncRNA-disease associations. However, taking into account the fact that the lncRNA functional similarity network was constructed relying on known lncRNA-disease associations, and each known lncRNA-disease association was used as the testing case in the leave-one-out across validation analysis, which may over-estimate the performance. Therefore, we re-evaluate the performance of RWRlncD method through re-constructing network from lncRNA-disease associations after removing this relation. As a result, our method achieved an AUC of 0.808 which is slightly decreased. To further determine whether the results of cross validation by the RWRlncD method might have been generated by chance, we performed randomization tests. The seed lncRNAs were generated randomly from candidate lncRNAs for each disease and the AUC value was calculated by performing the leave-one-out cross validation as above (Figure 4). The results showed that the AUC value under randomization tests (0.483) was much lower than that in the real situation, further demonstrating the effective and reliable performance of RWRlncD.

Previous studies have suggested that prediction performance would be enhanced by making use of the global network similarity information [28, 32]. So we used all pairwise functional similarity scores of 133 lncRNAs to construct a weighted lncRNA functional similarity network (WLFSN), in which the edges were assigned different functional similarity scores between lncRNAs. We further improved our RWRlncD method on WLFSN and performed the leave-one-out cross validation as described above; we achieved an AUC of 0.91 when restart probability $\alpha = 0.7$.

Effects of parameters in the proposed method

Our RWRlncD method has two parameters (restart probability α and score cutoff β to construct LFSN), which impact prediction performance. In order to investigate the possible effects of these two parameters on the performance of our RWRlncD method, we assigned different values for these two parameters and performed the above leave-one-out cross validation analysis using LFSN and WLFSN. The AUC values for different combinations of these two parameters were calculated and summarized in Table 2 and Table 3. The AUC for different parameters in RWRlncD method distributed appropriately from 0.75 ~ 0.90, demonstrating that our RWRlncD method could achieve reliable performance for different parameter combinations, and the predictive results are robust to score cutoff β and restart probability α .

Comparisons with other existing similar methods

When our research was in progress, a novel method, Laplacian RegularizedLeast Squares for LncRNA-Disease Association (LRLSLDA) was reported to predict disease-related lncRNA [21]. Unlike RWRlncD, LRLSLDA used the lncRNA expression information. LRLSLDA obtained an AUC of 0.776 in the leave-one-out cross validation on the same known experimentally verified lncRNA-disease associations from lncRNADisease database, but is less than an average AUC of 0.866 obtained by our RWRlncD method tested on the same datasets for different parameters, suggesting that our RWRlncD method can achieve more effective and more reliable performance

for predicting novel lncRNA-disease associations. Meantime, some network-based computational methods have been developed to predict disease-related protein-coding genes. So we used the best outperforming method, ICN[33, 34], to predict lncRNA-disease associations based on lncRNA functional similarity network, and performed performance comparison analysis with RWRlncD method based on the same dataset. The comparison between ICN and RWRlncD was shown in Figure 4. The ICN method achieved an AUC of 0.734, but is less than an average AUC of 0.866 obtained by RWRlncD tested using leave-one-out procedure on the same datasets.

Case studies

To illustrate the application of RWRlncD to infer novel lncRNA-disease associations, we presented case studies for prostate cancer and Alzheimer's disease because of their relatively numerous seed lncRNAs. Here, all known lncRNAs associated with the disease of interest were taken as seed, and all candidate lncRNAs could be ranked by our RWRlncD method according to P_∞. The comprehensive prediction results of potential novel lncRNA-disease associations for the given two diseases were summarized in Supplementary material 3. In these global ranking lists, we manually checked the top 10 predicted lncRNA-disease associations for a disease of interest from the NCBI database. The novel predicted lncRNA-disease associations confirmed by literature mining in the top 10 and supporting evidence were shown in Table 4. Among the top predicted disease-related lncRNAs, 6 lncRNAs-disease associations were validated directly or indirectly by reported biological experiments, and almost all of them were ranked high in the predictive lists. These independent practical applications and high-ranking evidence further demonstrate the reliable performance and potential value of our proposed RWRlncD method for predicting novel lncRNA-disease associations.

Discussion

The identification of potential disease-related lncRNAs is important for understanding their critical roles in the development and progression of human complex diseases and as such may provide new ways for biomarker identification and

drug design in the diagnosis, treatment and prevention of diseases. In this study, we first constructed an lncRNA-disease association network using experimentally verified lncRNA-disease associations and found power-law distribution of degree for lncRNAs and diseases, demonstrating that the associations between lncRNAs and diseases are not random but have biological significance. Also, lncRNA and their target diseases tended to densely cluster. Additionally, we observed some unconnected components, which reiterated the complexity of associations between lncRNAs and diseases. Many lines of evidence have shown that similarity networks between biological concepts, such as miRNAs and phenotypes, may be used to predict potential functions for novel biological molecules, or to infer potential candidate disease-related lncRNAs for guiding further biological experiments [28, 35, 36]. In this study, we improved previous methods to construct an lncRNA-lncRNA functional similarity network based on the assumption that functionally related lncRNAs may be associated with phenotypically similar diseases. Then we proposed a novel computational framework, RWRlncD, to predict novel lncRNA-disease associations by integrating lncRNA-disease network, disease similarity network and lncRNA functional similarity network. The candidate disease-related lncRNAs for a given disease can be prioritized by implementing random walks with restarts on the lncRNA functional similarity network, making full use of global network similarity information.

The results of performance evaluation based on leave-one-out cross validation revealed a high performance for recovering experimentally verified lncRNA-disease associations. The performance of RWRlncD was reliable and stable by assessing the effects of different parameters. These results indicate that the RWRlncD method can be applied to the prediction of novel lncRNA-disease associations. Previous studies have suggested that functionally related genes are associated with diseases of similar pathogeneses [22, 23], which has been successfully applied to detect the associations between protein-coding genes or miRNAs and diseases. Our proposed method added the disease phenotype data and adopted global network information by integrating disease similarity network, lncRNAs functional network and known lncRNA-disease

associations. However, the RWRlncD method is not applicable for lncRNAs, which do not have any known associated diseases. Therefore, with the availability of rich and increasing amounts of lncRNA-related biological data, different kinds of data should be considered and integrated to predict disease-related lncRNAs, such as expression profiles, functional characteristics, regulatory patterns between lncRNA and microRNAs or between lncRNAs and protein. Meanwhile, the RWRlncD method is based on the lncRNA functional similarity network in which lncRNA similarity was calculated using known lncRNA-disease associations. The performance of RWRlncD will be improved greatly by obtaining more lncRNA-disease association data or integrating more bioinformatics data to obtain more accurate functional similarity between lncRNAs.

Acknowledgements

We thank anonymous reviewers for valuable suggestions. This work was supported by Scientific Research Fund of Heilongjiang Provincial Education Department (NO: 12541308).

Notes and references

[a]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China

[b]Genomics Research Center (one of The State-Province Key Laboratories of Biomedicine-Pharmaceutics of China), Harbin Medical University, Harbin 150081, PR China

[c]College of Life Science, Jilin University, Changchun 130012, PR China

[d]Hospital of Harbin institute of technology, Harbin 150001, PR China

† Electronic Supplementary Information (ESI) available.

‡ **Authors' contributions**. MZ and DPH conceived and designed the experiments. JS, HBS and MZ developed the prediction method, implemented the experiments and analyzed the result. ZZW, LTW, CJZ, LL and WWH analyzed the result. JS, MZ and SLL wrote the paper. All authors read and approved the final manuscript

1. *Nature*, 2004, 431, 931-945.
2. P. Kapranov, A. T. Willingham and T. R. Gingeras, *Nat Rev Genet*, 2007, 8, 413-423.
3. C. P. Ponting, P. L. Oliver and W. Reik, *Cell*, 2009, 136, 629-641.
4. D. Rearick, A. Prakash, A. McSweeny, S. S. Shepard, L. Fedorova and A. Fedorov, *Nucleic Acids Res*, 2011, 39, 2357-2366.
5. T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow and R. Guigo, *Genome Res*, 2012, 22, 1775-1789.
6. P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger and J. S. Mattick, *Nucleic Acids Res*, 2011, 39, D146-151.
7. P. J. Volders, K. Helsens, X. Wang, B. Menten, L. Martens, K. Gevaert, J. Vandesompele and P. Mestdagh, *Nucleic Acids Res*, 2012, 41, D246-251.
8. D. Bu, K. Yu, S. Sun, C. Xie, G. Skogerbo, R. Miao, H. Xiao, Q. Liao, H. Luo, G. Zhao, H. Zhao, Z. Liu, C. Liu, R. Chen and Y. Zhao, *Nucleic Acids Res*, 2012, 40, D210-215.
9. J. Jin, J. Liu, H. Wang, L. Wong and N. H. Chua, *Bioinformatics*, 2013, 29, 1068-1071.
10. S. W. Cheetham, F. Gruhl, J. S. Mattick and M. E. Dinger, *Br J Cancer*, 2013, 108, 2419-2425.
11. E. A. Gibb, C. J. Brown and W. L. Lam, *Mol Cancer*, 2011, 10, 38.
12. L. X. Garmire, D. G. Garmire, W. Huang, J. Yao, C. K. Glass and S. Subramaniam, *PLoS One*, 2011, 6, e24051.
13. K. Sun, X. Chen, P. Jiang, X. Song, H. Wang and H. Sun, *BMC Genomics*, 2013, 14 Suppl 2, S7.
14. Q. Liao, C. Liu, X. Yuan, S. Kang, R. Miao, H. Xiao, G. Zhao, H. Luo, D. Bu, H. Zhao, G. Skogerbo, Z. Wu and Y. Zhao, *Nucleic Acids Res*, 2011, 39, 3864-3878.

15. X. Guo, L. Gao, Q. Liao, H. Xiao, X. Ma, X. Yang, H. Luo, G. Zhao, D. Bu, F. Jiao, Q. Shao, R. Chen and Y. Zhao, *Nucleic Acids Res*, 2013, 41, e35.
16. X. Guo, L. Gao, Q. Liao, H. Xiao, X. Ma, X. Yang, H. Luo, G. Zhao, D. Bu, F. Jiao, Q. Shao, R. Chen and Y. Zhao, *Nucleic Acids Res*, 2012, 41, e35.
17. L. Sun, Z. Zhang, T. L. Bailey, A. C. Perkins, M. R. Tallack, Z. Xu and H. Liu, *BMC Bioinformatics*, 2012, 13, 331.
18. K. Liu, Z. Yan, Y. Li and Z. Sun, *Bioinformatics*, 2013, 29, 2221-2222.
19. J. H. Li, S. Liu, H. Zhou, L. H. Qu and J. H. Yang, *Nucleic Acids Res*, 2013.
20. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan and Q. Cui, *Nucleic Acids Res*, 2013, 41, D983-986.
21. X. Chen and G. Y. Yan, *Bioinformatics*, 2013, 29, 2617-2624.
22. T. Ideker and R. Sharan, *Genome Res*, 2008, 18, 644-652.
23. M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, *PLoS One*, 2008, 3, e3420.
24. J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, C. Li, X. Li, S. Rao and X. Li, *BMC Bioinformatics*, 2011, 12, 266.
25. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C. F. Chen, *Bioinformatics*, 2007, 23, 1274-1281.
26. Z. Du, L. Li, C. F. Chen, P. S. Yu and J. Z. Wang, *Nucleic Acids Res*, 2009, 37, W345-349.
27. D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, *Bioinformatics*, 2010, 26, 1644-1650.
28. X. Chen, M. X. Liu and G. Y. Yan, *Mol Biosyst*, 2012, 8, 2792-2798.
29. K. Macropol, T. Can and A. K. Singh, *BMC Bioinformatics*, 2009, 10, 283.
30. Y. Li and J. C. Patra, *Bioinformatics*, 2010, 26, 1219-1224.
31. R. Jiang, M. Gan and P. He, *BMC Syst Biol*, 2011, 5 Suppl 2, S2.
32. X. Yao, H. Hao, Y. Li and S. Li, *BMC Syst Biol*, 2011, 5, 79.
33. S. Navlakha and C. Kingsford, *Bioinformatics*, 2010, 26, 1057-1063.
34. C. L. Hsu, Y. H. Huang, C. T. Hsu and U. C. Yang, *BMC Genomics*, 2011, 12 Suppl 3, S25.
35. H. Chen and Z. Zhang, *BMC Med Genomics*, 2013, 6, 12.
36. Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu and Y. Wang, *BMC Syst Biol*, 2010, 4 Suppl 1, S2.
37. P. C. Lin, Y. L. Chiu, S. Banerjee, K. Park, J. M. Mosquera, E. Giannopoulou, P. Alves, A. K. Tewari, M. B. Gerstein, H. Beltran, A. M. Melnick, O. Elemento, F. Demichelis and M. A. Rubin, *Cancer Res*, 2012, 73, 1232-1244.
38. M. R. Ginger, A. N. Shore, A. Contreras, M. Rijnkels, J. Miller, M. F. Gonzalez-Rimbau and J. M. Rosen, *Proc Natl Acad Sci U S A*, 2006, 103, 5781-5786.
39. T. Chiyomaru, S. Yamamura, S. Fukuhara, H. Yoshino, T. Kinoshita, S. Majid, S. Saini, I. Chang, Y. Tanaka, H. Enokida, N. Seki, M. Nakagawa and R. Dahiya, *PLoS One*, 2013, 8, e70372.
40. K. Sakurai, C. Furukawa, T. Haraguchi, K. Inada, K. Shiogama, T. Tagawa, S. Fujita, Y. Ueno, A. Ogata, M. Ito, Y. Tsutsumi and H. Iba, *Cancer Res*, 2011, 71, 1680-1689.
41. R. Johnson, *Neurobiol Dis*, 2012, 46, 245-254.
42. M. M. Wilhelmus, S. M. van der Pol, Q. Jansen, M. E. Witte, P. van der Valk, A. J. Rozemuller, B. Drukarch, H. E. de Vries and J. Van Horssen, *Free Radic Biol Med*, 2011, 50, 469-476.

Tables

Table 1.Global characteristics of the lncRNA-disease association network

No. of lncRNA	No. of disease	No. of lncRNA-disease associations	Average degree of lncRNA	Average degree of disease
156	190	352	2.3	1.9

Table 2. The effect for different combinations of these two parameters of RWRlncD method on LFSN

α \ β	0.1	0.3	0.5	0.7	0.9
0.1	0.759	0.857	0.813	0.956	0.928
0.3	0.788	0.874	0.819	0.955	0.926
0.5	0.773	0.876	0.821	0.954	0.925
0.7	0.751	0.875	0.822	0.954	0.925
0.9	0.725	0.873	0.823	0.954	0.925

Table 3. The effect for different restart probability value of the RWRlncD method on WLFSN

α	0.1	0.3	0.5	0.7	0.9
AUC	0.857	0.897	0.906	0.91	0.911

Table 4. The newly lncRNA-disease associations confirmed by literature mining in the top 10 predicted results by RWRlncD

LncRNA name	Ranking	References
Prostate cancer		
MIR31HG	2	[37]
PINC	2	[38]
HOTAIR	3	[39]
DNM3OS	6	[40]
Alzheimer's disease		

TUG1	1	[41]
PINK1-AS	2	[42]

Figures

Figure 1. **The schematic representation and overview of the RWRlncD method**. The procedures of RWRlncD can be divided into three steps: calculating the functional similarity scores between lncRNAs based on lncRNA-disease associations; constructing the lncRNA functional similarity network; and predicting the potential lncRNA-disease associations by random walks with restarts on the lncRNA functional similarity network

Figure 2. **Construction and characteristics of lncRNA-disease association network.** (A) lncRNA-disease association network (LDAN), generated by using 352 experimentally verified associations between lncRNA and disease. (B) Degree distribution for lncRNAs in the LDAN. (C) Degree distribution of disease in the LDAN.

Figure 3. **Construction and characteristics of lncRNA functional similarity network.** (A) Cumulative distribution of the edges between lncRNAs using various similarity cutoffs. (B) lncRNA functional similarity network (LFSN). Each node represents one lncRNA and the edge between two nodes indicate the functional similarity score of the two lncRNAs is equal or greater than the score cutoff (here the cutoff is 0.5). (C) Degree distribution for lncRNA in the LFSN.

Figure 4. **ROC curves and AUC values of RWRlncD on LFSN and WLFSN**. The ROC curves were plotted and AUC values were calculated by leave-one-out cross validation on 352 experimentally verified lncRNA-disease associations using RWRlncD on LFSN and WLFSN.

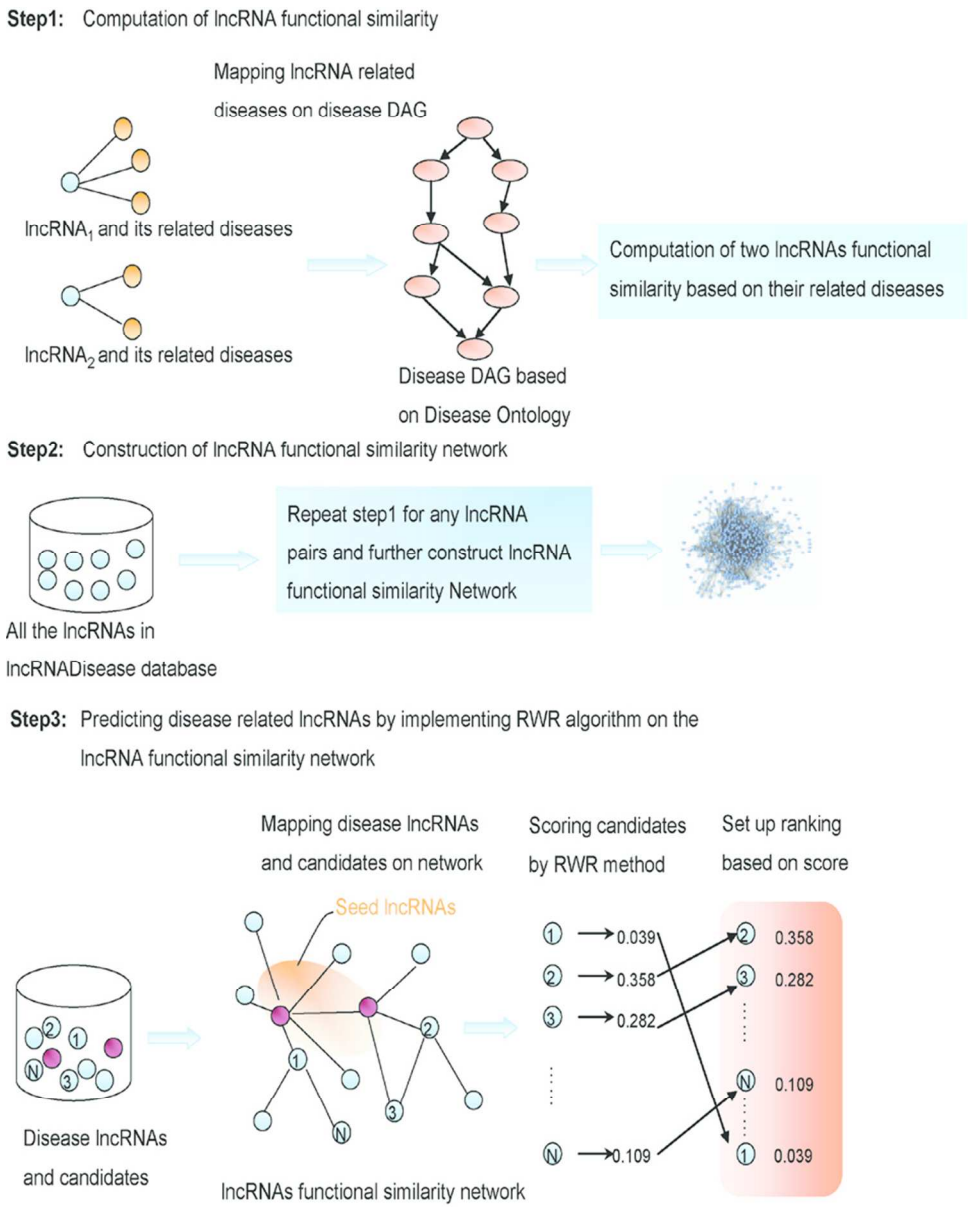


Figure 1. The schematic representation and overview of the RWRlncD method. The procedures of RWRlncD can be divided into three steps: calculating the functional similarity scores between lncRNAs based on lncRNA-disease associations; constructing the lncRNA functional similarity network; and predicting the potential lncRNA-disease associations by random walks with restarts on the lncRNA functional similarity network
80x99mm (300 x 300 DPI)

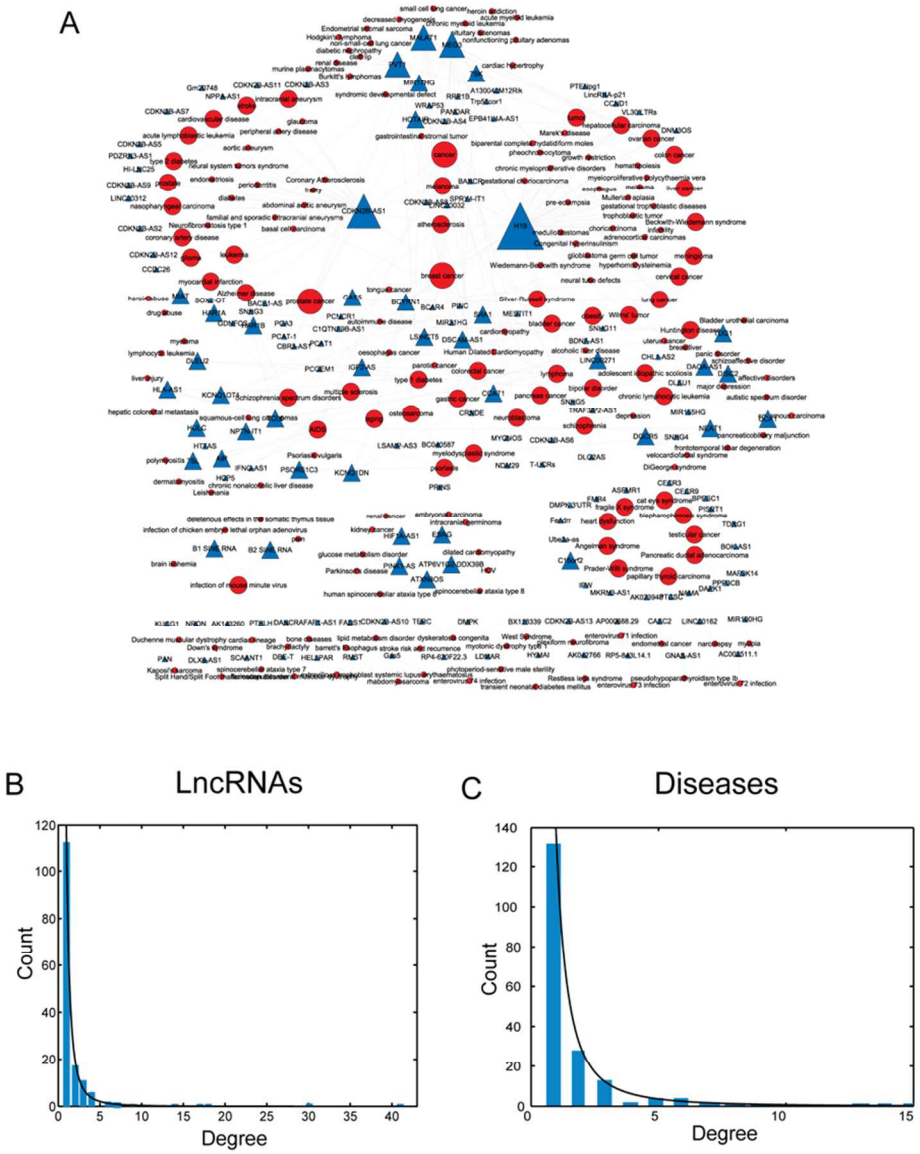


Figure 2. Construction and characteristics of lncRNA-disease association network. (A) lncRNA-disease association network (LDAN), generated by using 352 experimentally verified associations between lncRNA and disease. (B) Degree distribution for lncRNAs in the LDAN. (C) Degree distribution of disease in the LDAN.
80x99mm (300 x 300 DPI)

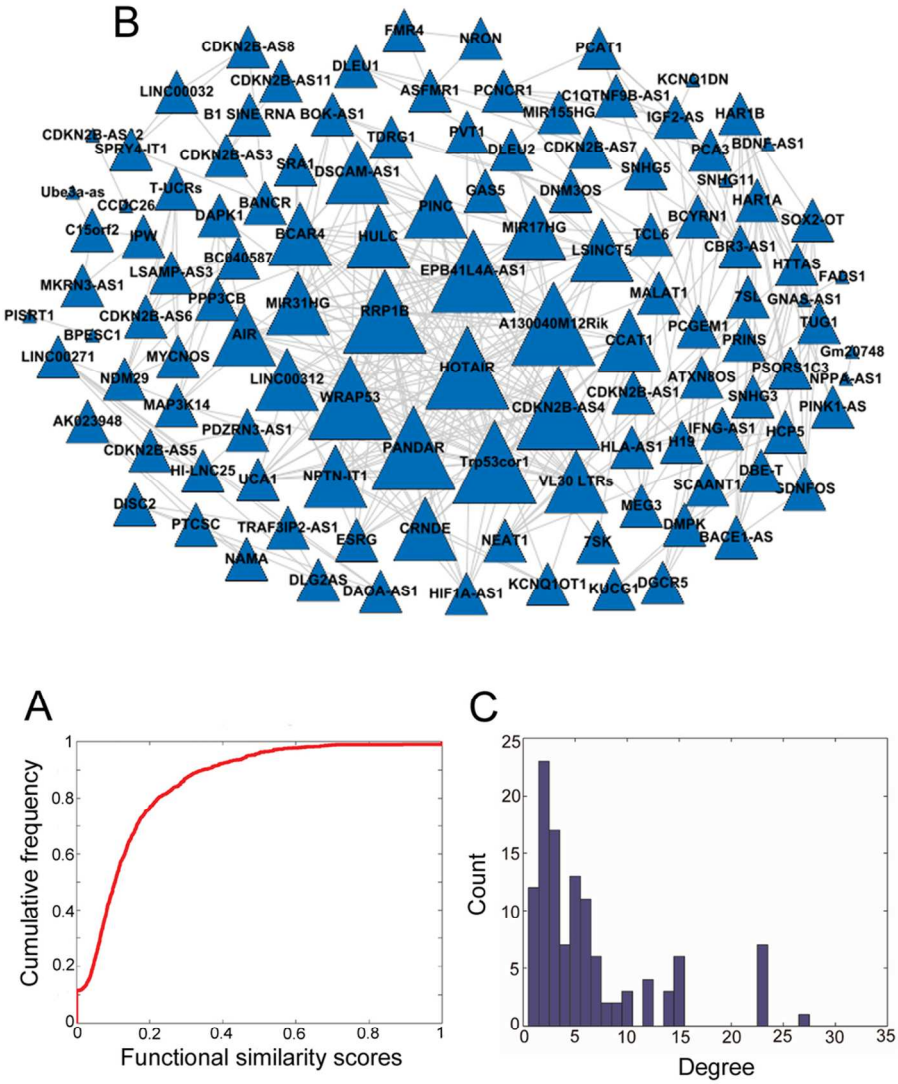


Figure 3. Construction and characteristics of lncRNA functional similarity network. (A) Cumulative distribution of the edges between lncRNAs using various similarity cutoffs. (B) lncRNA functional similarity network (LFSN). Each node represents one lncRNA and the edge between two nodes indicate the functional similarity score of the two lncRNAs is equal or greater than the score cutoff (here the cutoff is 0.5). (C) Degree distribution for lncRNA in the LFSN.
90x115mm (300 x 300 DPI)

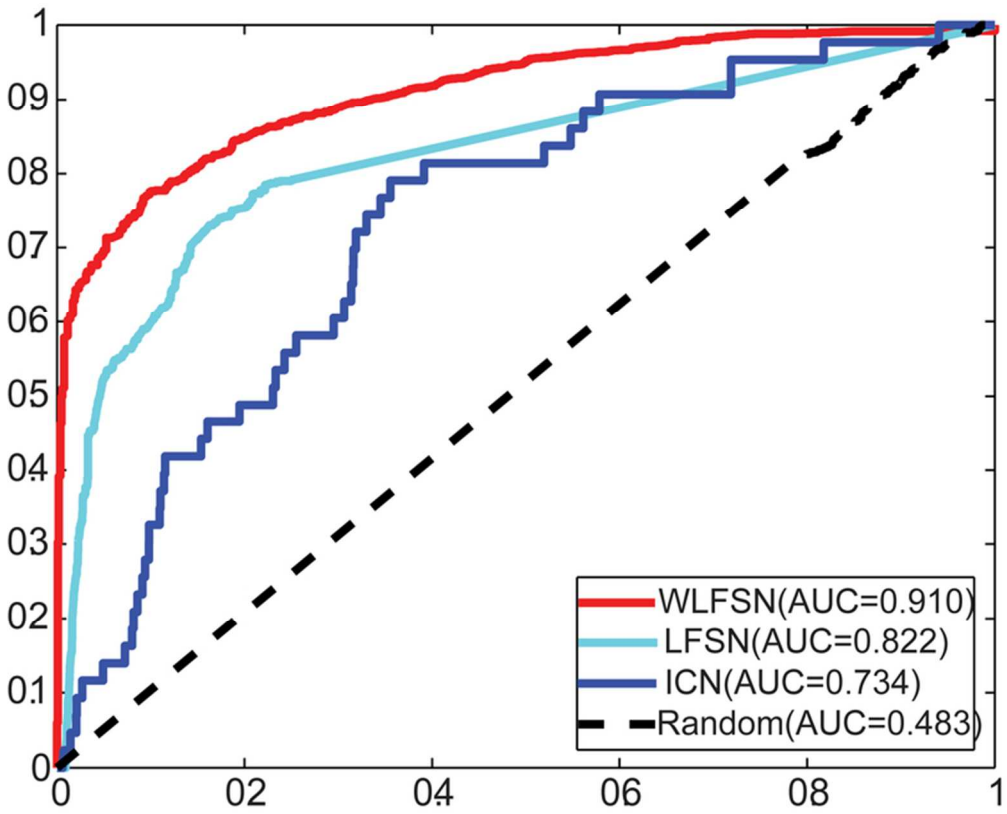


Figure 4. ROC curves and AUC values of RWRlncD on LFSN and WLFSN. The ROC curves were plotted and AUC values were calculated by leave-one-out cross validation on 352 experimentally verified lncRNA-disease associations using RWRlncD on LFSN and WLFSN.
68x56mm (300 x 300 DPI)