A Network Based Method for Analysis of lncRNA-Disease Associations and Prediction of lncRNAs Implicated in Diseases

Xiaofei Yang[1], Lin Gao[1]*, Xingli Guo[1], Xinghua Shi[2], Hao Wu[1], Fei Song[1], Bingbo Wang[1]

[1] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China, [2] Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, North Carolina, Unites States of America

Abstract

Increasing evidence has indicated that long non-coding RNAs (lncRNAs) are implicated in and associated with many complex human diseases. Despite of the accumulation of lncRNA-disease associations, only a few studies had studied the roles of these associations in pathogenesis. In this paper, we investigated lncRNA-disease associations from a network view to understand the contribution of these lncRNAs to complex diseases. Specifically, we studied both the properties of the diseases in which the lncRNAs were implicated, and that of the lncRNAs associated with complex diseases. Regarding the fact that protein coding genes and lncRNAs are involved in human diseases, we constructed a coding-non-coding gene-disease bipartite network based on known associations between diseases and disease-causing genes. We then applied a propagation algorithm to uncover the hidden lncRNA-disease associations in this network. The algorithm was evaluated by leave-one-out cross validation on 103 diseases in which at least two genes were known to be involved, and achieved an AUC of 0.7881. Our algorithm successfully predicted 768 potential lncRNA-disease associations between 66 lncRNAs and 193 diseases. Furthermore, our results for Alzheimer's disease, pancreatic cancer, and gastric cancer were verified by other independent studies.

Citation: Yang X, Gao L, Guo X, Shi X, Wu H, et al. (2014) A Network Based Method for Analysis of lncRNA-Disease Associations and Prediction of lncRNAs Implicated in Diseases. PLoS ONE 9(1): e87797. doi:10.1371/journal.pone.0087797

Editor: Paolo Provero, University of Turin, Italy

Received July 4, 2013; **Accepted** December 31, 2013; **Published** January 31, 2014

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 60933009, 61070137, 91130006, 61303118, & 61303122), the website is http://www.nsfc.gov.cn/publish/portal1/ the Fundamental Research Funds for the Central Universities (XIDIAN UNIVERSITY, No. K50512230005 & K50512230003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: lgao@mail.xidian.edu.cn

Introduction

Long non-coding RNAs (lncRNAs) are similar to mRNAs in gene structure, with length greater than 200 nt [1–3]. LncRNAs play critical roles in many important biological processes such as chromatin modification [4], transcriptional and post-transcriptional regulation [4], and human diseases [2].

More and more studies have reported that mutated and dysfunctional lncRNAs are implicated in a broad range of human diseases. For example, Pasmant et al. [5] performed a GWAS and identified that *ANRIL* was significantly associated with coronary disease, type 2 diabetes, and many types of cancers. *HOTAIR* was increased from 100 to approximately 2,000-fold in breast cancer metastases using quantitative PCR [6]. *MALAT-1* was significantly associated with metastasis in NSCLC patients by quantitative RT-PCR [7]. With regard to Alzheimer's disease, *BCAE1-AS* was shown to have a key role in regulating *BACE1* and in driving pathology [8]. Cui et al. [9] found that the expression of *PlncRNA-1* was significantly higher in prostate cancer cells. Therefore, it is necessary to analyze the available lncRNA-disease associations and predict potential lncRNA-disease associations in human. Such studies will help us understand the molecular mechanisms of human diseases and identify biomarkers for disease diagnosis, treatment, and prevention at lncRNA level [10].

Chen et al. [10] reported a LncRNADisease database that includes approximately 480 entries of experimentally supported associations between 166 diseases and 118 lncRNAs. Moreover, we have manually collected 380 lncRNA-disease associations between 226 lncRNAs and 145 diseases by literature mining. By integrating these two data sets, we obtained 578 lncRNA-disease associations between 295 lncRNAs and 214 diseases. These data were analyzed in a network view and used to predict lncRNA-disease associations.

In this paper, based on the available lncRNA-disease associations, a lncRNA-disease association network was constructed. From the constructed network, two relevant biological networks "lncRNA-implicated disease network" (lncDN) and "disease-associated lncRNA network" (DlncN) were derived, as shown in Figure 1. In lncDN, a node represented a disease, and a link between two nodes indicated that the two corresponding diseases shared at least one lncRNA as their disease-causing lncRNA (Figure 1 and Figure 2-(a)). In DlncN, a node represented a lncRNA, and a link between two nodes represented the fact that the two corresponding lncRNAs were implicated in at least one common disease (Figure 1 and Figure 2-(b)). The known lncRNA-disease associations were represented in a single network framework, and the network topological properties were analyzed to help us investigate all of these associations. Furthermore, a propagation algorithm was applied to predict potential lncRNA-disease associations on the lncRNA-disease association network. In addition, a coding-non-coding gene-disease bipartite network was constructed by integrating coding gene-disease associations obtained from OMIM [11] with lncRNA-disease associations. To achieve better prediction performance, the propagation algorithm was applied to rank the potential gene-disease pairs for all the diseases on the coding-non-coding gene-disease bipartite network. In the Leave-One-Out Cross-Validation (LOOCV) procedure, our method achieved a reliable Area Under Curve (AUC) of 0.7881. We then employed our method to the study of three multi-factorial diseases, Alzheimer's disease, pancreatic cancer and gastric cancer, and provided suggestions of novel disease-causing lncRNAs for further study.

Materials and Methods

Data Sources

The 480 lncRNA-disease associations were downloaded from LncRNADisease database [10], including 118 lncRNAs and 166 diseases. Note that many other lncRNA-disease associations have been reported in the literature, but have not been included in the LncRNADisease database yet. Hence, we retrieved literature from PubMed (http://www.ncbi.nlm.nih.gov/pubmed), employing the key words 'lncRNA and disease', 'lncRNA and cancer', 'long non-coding RNA and disease', 'long non-coding RNA and cancer', 'lincRNA and disease' or 'lincRNA and cancer', and manually extracted 129 articles that reported lncRNA-disease associations. In this way, we collected an additional 380 lncRNA-disease associations between 226 lncRNAs and 145 diseases by literature mining. Integrating these two data sets from both LncRNADisease database and literature search, we finally obtained 578 associations between 295 lncRNAs and 214 diseases. All of these 578 lncRNA-disease associations were then merged into a lncRNA-disease association network.

Of the 214 diseases, 160 diseases and their causative genes could be found using their MIM number in OMIM database [11]. In total, we downloaded 801 disease genes for these 160 diseases from OMIM database. Such data resulted in 980 protein-coding gene-disease associations that were used in our method.

Integrating lncRNA-disease associations and protein-coding gene-disease associations obtained above, we obtained 1558 coding-non-coding gene-disease associations between 1096 genes (295 lncRNAs and 801 protein-coding genes) and 214 diseases. These associations were used to construct the coding-non-coding genes-disease bipartite network.

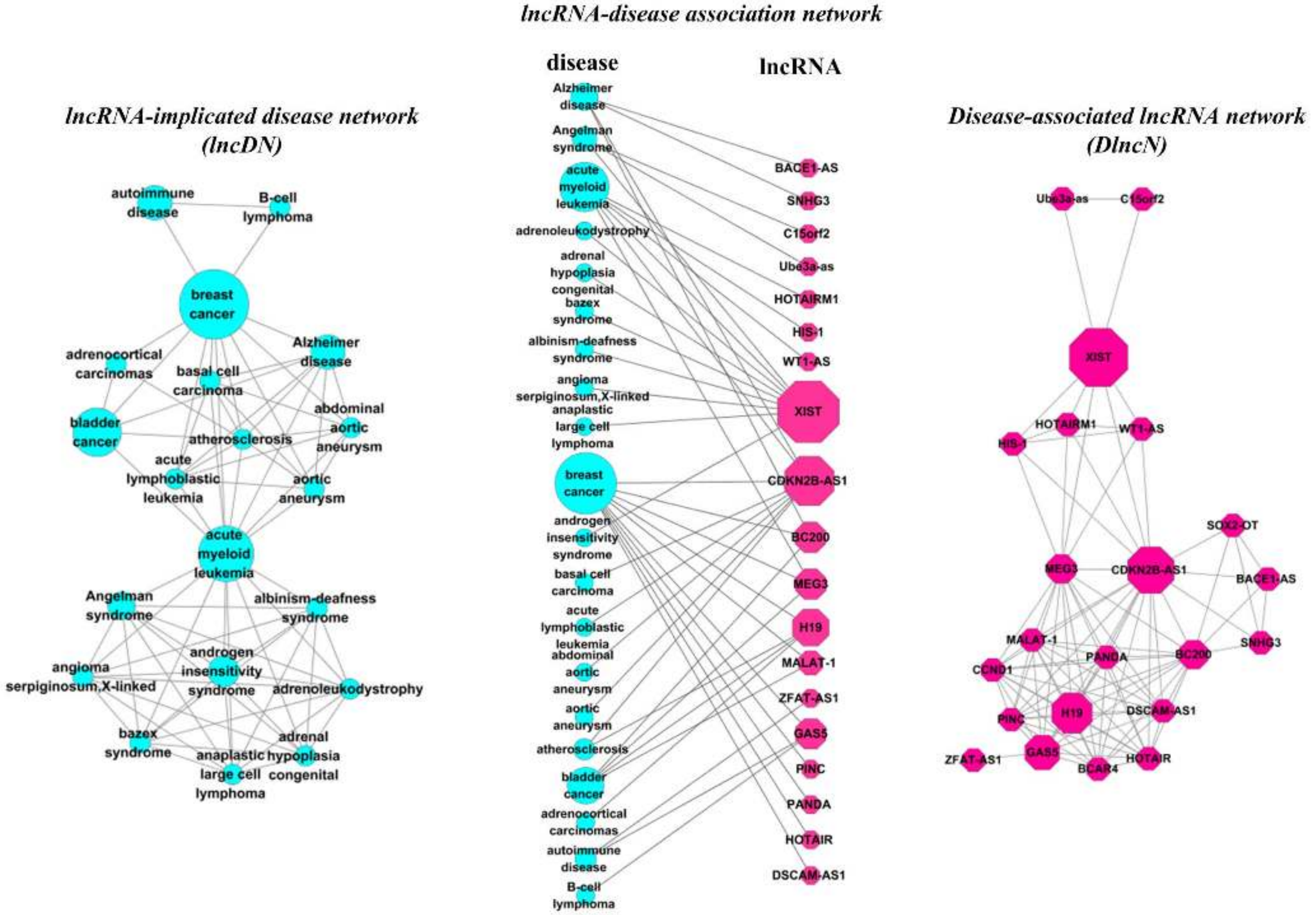

Figure 1. Construction of the lncRNA-disease bipartite network. *(Center)* A subnetwork of the full lncRNA-disease association network (Figure S1), where the blue circles and red hexagons correspond to diseases and lncRNAs, respectively. A link is placed between a disease and a lncRNA if mutations or dysfunctions in that lncRNA lead to the specific disease. The size of a blue circle is proportional to the number of lncRNAs participating in the corresponding disease. The size of a red hexagon is proportional to the number of diseases associated with the corresponding lncRNA. *(Left)* The lncDN projection of the center graph, in which two diseases are connected if there is a lncRNA implicated in both diseases. *(Right)* The DlncN projection of the center graph where two lncRNAs are connected if they are involved in the same disease.
doi:10.1371/journal.pone.0087797.g001

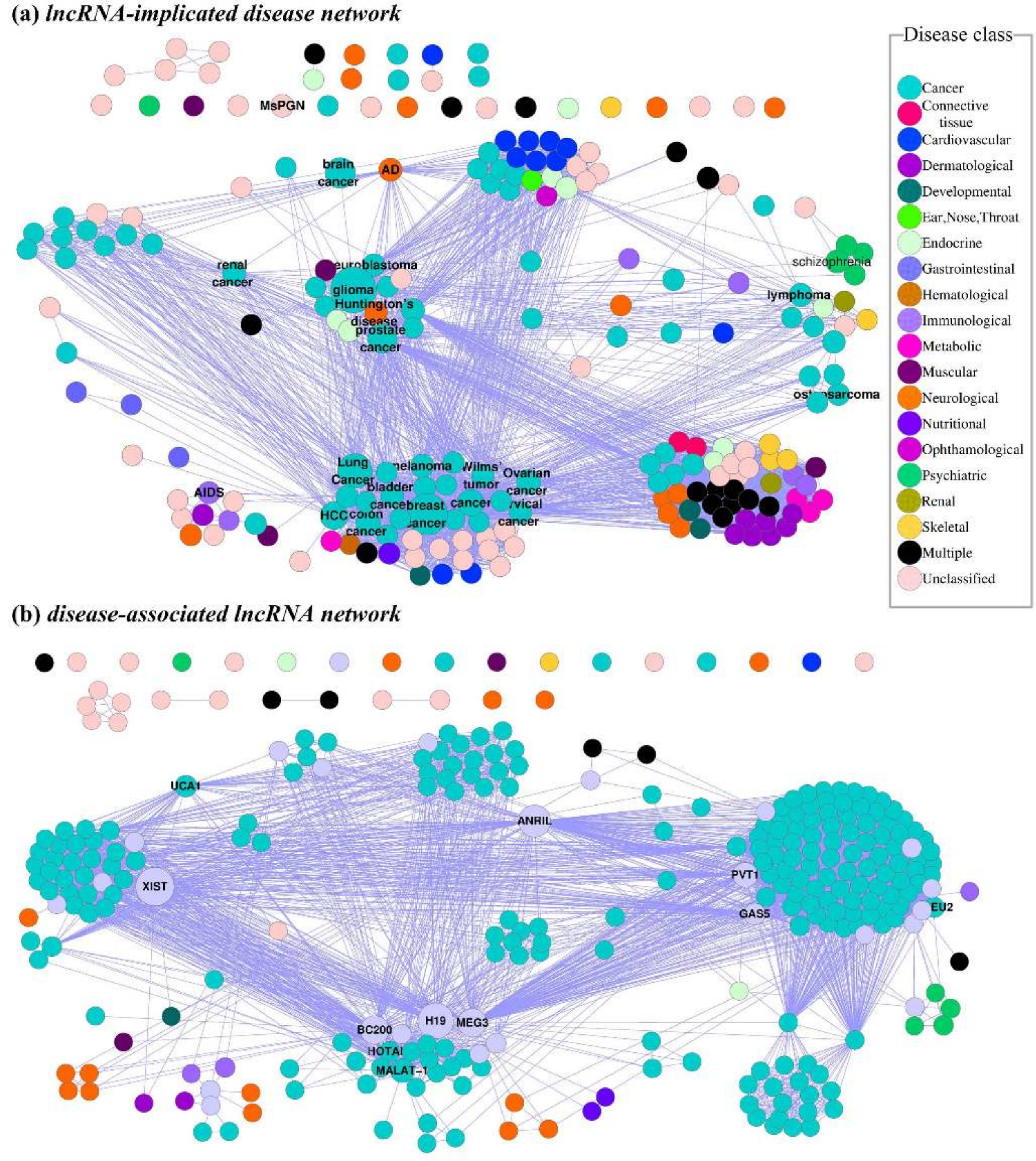

Figure 2. The lncDN and DlncN. (a) In lncDN, each node corresponds to a distinct disease, colored based on the disease class [26] to which it belongs. The names of 20 disease classes are shown on the right panel. A link between two diseases exists if they share at least one implicated lncRNA. The size of the node is proportional to the degree of the node in lncRNA-disease association network. We label the diseases associated with more than five lncRNAs by their names. (b) In DlncN, each node is a lncRNA, with two lncRNAs being connected if they are implicated in the same disease. The size of each node is proportional to the number of diseases in which the lncRNA is implicated. The color of a node is based on the class of diseases in which the corresponding lncRNA implicated. Nodes are light purple if the corresponding lncRNAs are associated with more than one disease class. We label the lncRNAs implicated in more than five diseases by their names.
doi:10.1371/journal.pone.0087797.g002

Methods

Given a bipartite network $G(X,Y,E)$, X and Y were two disjoint node sets, E was the edge set in which the element represented the edge connecting the node from X and the node from Y. The bipartite network could be viewed in one-mode projection onto X and one-mode projection onto Y, called X projection and Y projection respectively. The X projection of G was a network in which nodes were from X, and the edge indicated that the connected nodes were associated with at least one same node from Y. Similarly, the Y projection of G was a network in which nodes were from Y, and the edge indicated that the connected nodes were associated with at least one same node from X. With regard to the lncRNA-disease association network, which could be claimed as a bipartite network, lncDN and DlncN were the disease projection and lncRNA projection of the lncRNA-disease association network. The properties of these two projections were analyzed in the "Results" section. It was found that the lncDN could reflect the relationships between any two diseases at the lncRNA level and that DlncN could reflect the relationships between any two lncRNAs at the disease level. Moreover, we tried to exploit these relationships to predict the hidden lncRNA-disease associations. For better performance, both protein-coding genes and lncRNAs that were implicated in diseases were considered together. As a result, a coding-non-coding gene-disease bipartite network was constructed to reflect the associations between diseases and all the disease-causing genes (i.e. protein-coding genes or lncRNAs). The resource-allocation process [12], as one of the best weighting methods for one-mode projection of a bipartite network, was used to weight the gene projection of the coding-non-coding gene-disease bipartite network. Then a propagation algorithm was applied to compute the association score for each gene that was used to measure how much the gene could be implicated in a disease on the weighted gene projection. For a disease q, every gene had its initial information. Our propagation algorithm could be assumed as a process where genes pumped their initial information to their neighbors, and every gene propagated the information received in the previous iteration to other genes via edges in gene projection.

Next, we illustrated the principle of the resource-allocation process, and then provided the propagation algorithm to compute the score of genes with respect to a specific disease.

Principle of the resource-allocation process

We divided the nodes of a bipartite network into two sets X and Y, and only the connections between two nodes in different sets are allowed. The resource-allocation process is one of the best weighting method for one-mode projection of a bipartite network [12]. This process was illustrated in Figure 3 and included the following two steps. First, we allocated resources from X to Y. Second, we then allocated resources from Y back to X. The initial resource of five nodes was a,b,c,d and e in set X. These two steps of the resource-allocation process were merged into one, and the final resource of X nodes denoted by a',b',c',d' and e', could be written as:

$$\begin{pmatrix} a' \\ b' \\ c' \\ d' \\ e' \end{pmatrix} = \begin{pmatrix} 3/4 & 0 & 1/4 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/4 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1/2 & 1/4 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} \qquad (1)$$

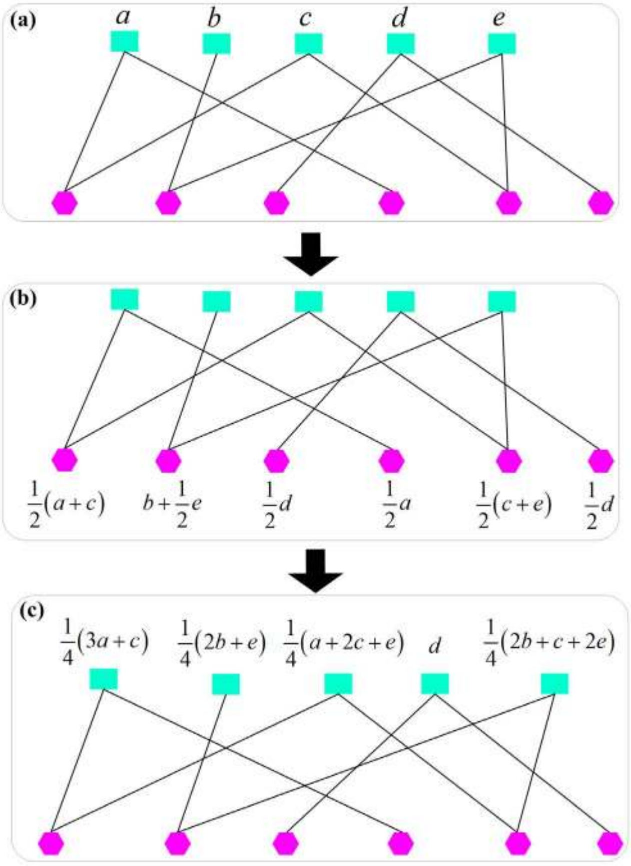

Figure 3. Principle of the resource-allocation process in a bipartite network. The green rectangles represent X nodes and red hexagons represent Y nodes. The whole process consists of two steps: First, the resource flows from X to Y (a→b), and then returns to X (b→c).

doi:10.1371/journal.pone.0087797.g003

The 5×5 matrix, W, represented the weighted X projection. The element w_{ij} represented the fraction of resource that the j-th X node transferred to i-th X node, and could be considered as the importance of node i on node j [12].

For a bipartite network $G(X,Y,E)$, and $|X|=n,|Y|=m$, x_i was the i-th X node and y_l was the l-th Y node. w_{ij} could be calculated as follows [12]:

$$w_{ij} = \frac{1}{k(x_j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(y_l)} \qquad (2)$$

where $1 \le i,j \le n$, a_{il} was the $n \times m$ adjacent matrix of $G(X,Y,E)$, and $k(x_i)$ was the degree of x_i.

The propagation algorithm

The coding-non-coding gene-disease bipartite network was denoted by $A(G,D,E)$, where G was the node set of the genes, D was the node set of the diseases, and E was the edge set. The weighted gene projection of A was denoted by W, where w_{ij} was calculated by Formula (2) and represented the importance of gene i on gene j in terms of their association with disease.

Our propagation algorithm was based on a semi-supervised learning algorithm [13], which had been previously used to prioritize protein-coding genes implicated in human diseases [14] and annotate functions of lncRNAs [15]. The input of the algorithm included $A(G,D,E)$, a query disease q, and W. For disease q, every gene had its own initial information. If a gene was connected with q in the coding-non-coding gene-disease bipartite network, the initial information was one; otherwise the initial information was zero. For a given disease q, the score vector f of genes represented the association scores of genes with q, which was computed by an iterative algorithm. The genes were ranked for q by the final score vector. Of all the genes not associated with disease q, the top 1% ranked genes were considered as the predicted genes. The score vector f was defined as:

$$f = W \times f \tag{3}$$

An iterative process [14] was applied to compute the score vector in Formula (3). Considering the initial information on the genes for the given disease, the score vector f was computed iteratively as follows,

$$f^t = \alpha \times W \times f^{t-1} + (1-\alpha) \times f^0 \tag{4}$$

In Formula (4), the score vector was initialized as f^0 by the initial information on genes. The parameter $\alpha \in (0,1)$ gave the relative importance between the contributed information of other genes and the initial information of itself. The final score vector with respect to disease q was determined by both the information on other genes and its initial information. The iterative computation was controlled by the mean score deviation of the two neighboring score vector. All the testes on the real data and random data had shown that the iterative computation converges eventually (Table S1).

Results

Properties of lncRNA-disease association network

The available lncRNA-disease associations were modeled as a bipartite network, and a subnetwork of this network was shown in Figure 1. In this bipartite network, one node set corresponded to the disease set; the other set corresponded to the lncRNA set. A lncRNA and a disease were connected by a link if the lncRNA was associated with the disease. The constructed bipartite network contained 578 edges between 295 lncRNA nodes and 214 disease nodes.

The degree distribution of the full lncRNA-disease association network (Figure S1) closely followed a power-law distribution (Figure S2-(a)). We also analyzed the degree of disease nodes and that of lncRNA nodes separately. The degree of a disease node, which meant the number of lncRNAs associated with the disease, was denoted by s and had a broad distribution (Figure S2-(b)). These results indicated that most disorders were associated with a small number of lncRNAs, whereas a handful of diseases, such as breast cancer and lung cancer, were related to a large number of lncRNAs. For example, 41 lncRNAs were involved in breast cancer ($s = 41$), 18 lncRNAs were related with prostate cancer ($s = 18$), and 28 lncRNAs were involved in lung cancer ($s = 28$). The degree of a lncRNA node, i.e. the number of diseases associated with the lncRNA, was denoted by d, and had a broad distribution as well (Figure S2-(c)). This indicated that many lncRNAs were related to a few diseases whereas a small number of lncRNAs could be related to dozens of diseases. For example, *XIST* ($d = 50$) was associated with 50 diseases, including 40 skin diseases [16] and certain types of cancers such as testicular cancer [17] and breast cancer [18]. *H19* ($d = 39$) was associated with 39 diseases, including Beckwith-Wiedemann syndrome [19], Silver-Russell syndrome [20,21] and many types of cancer [22]. *MEG3* ($d = 23$) was associated with 23 diseases, including breast cancer [23], bladder cancer [23], glioma [24], and Wilms' tumor [25], etc. These lncRNAs represented major hubs in DlncN (Figure 2-(b)).

Network analysis of lncDN and DlncN

We performed a network analysis of lncDN and DlncN to help us understand the lncRNA-disease associations. Two biologically relevant network projections, lncDN and DlncN, were generated (Figure 2) based on the lncRNA-disease association network. Specifically, lncDN provided a disease centered view of the lncRNA-disease association network (Figure 2-(a)). DlncN was complementary to lncDN and offered a lncRNA centered view of the lncRNA-disease association network (Figure 2-(b)). Especially, the links between two lncRNAs in DlncN signified the disease phenotypic associations, which might be a measure of their functional correlations and could be used in future studies.

Degree distributions of lncDN and DlncN. The degree distribution of the lncDN was investigated (Figure S3-(a)). The results showed that most disorders linked to only a few other diseases, whereas only few disorders represented hubs that were connected to a large number of distinct disorders. Such hub disorders included breast cancer (linked to 150 other disorders, i.e. $n = 150$, here n meant the degree of a node in lncDN), prostate cancer ($n = 144$), and lung cancer ($n = 73$). The degree distribution of the DlncN (Figure S3-(b)) was similar to that of lncDN. We could see that the degrees of most lncRNAs were small, whereas a few lncRNAs linked to a large number of lncRNAs. For example, *MEG3* linked to 196 other lncRNAs, *ANRIL* linked to 166 other lncRNAs, and *PVT1* linked to 162 other lncRNAs. These highly connected lncRNAs represented hubs in DlncN which connected to a large number of diseases in lncRNA-disease association network. We concluded that the degree distributions of both lncDN and DlncN networks closely followed a power-law distribution, despite of the incompleteness and false positive rate of the known lncRNA-disease associations.

Comparison of lncDN and DlncN with random networks. In lncDN, there were 3061 links among 214 individual diseases. Of the 214 diseases, 197 had at least one link to other diseases and 182 diseases formed a giant connected component. In DlncN, there were 6989 links among 295 lncRNAs. Of the 295 lncRNAs, 276 had at least one link to other lncRNAs and 265 lncRNAs formed a giant connected component.

We randomly shuffled the lncRNA-disease association network for 10^4 times, while keeping the degree of each lncRNA and each disease in the bipartite network unchanged [26]. We constructed the corresponding r-lncDN and r-DlncN respectively for the disease and lncRNA centered view of the randomized lncRNA-disease association network. Comparing lncDN and DlncN with r-lncDN and r-DlncN, respectively, we found that the topology property of the two generated networks, lncDN and DlncN, deviated from random. The average size of the giant connected components of 10^4 r-lncDNs was 137 ± 6, which was significantly smaller than 182 ($\text{p-value} < 10^{-8}$, z-test), the actual size of the giant connected component of lncDN. Similarly, the average size of the giant connected components of 10^4 r-DlncNs was 215 ± 7, which was significantly smaller than 265 ($\text{p-value} < 10^{-8}$, z-test), the actual size of the giant connected component of DlncN. Considering disease classes as defined in the Goh et al.'s study [26], we found that diseases (lncRNAs) were more likely to be linked to the diseases in the same class in the actual networks. For example, in the lncDN, there were 806 links between diseases of the same class, a two-fold enrichment with respect to 397 ± 47 links obtained between the same set of nodes in the randomized networks. These differences suggested important pathophysiological clustering of diseases and disease associated lncRNAs.

Clustering coefficients of lncDN and DlncN. To further address the topological properties of lncDN and DlncN, we calculated the average clustering coefficient, a measure of the tendency of nodes in a network to form clusters or groups [27], by NetworkAnalyzer [28], a plugin of cytoscape software [29]. We found that the average clustering coefficients of nodes in both networks approximately diminished when the degree of node increased (Figure S4), indicating that nodes with high degrees tended to be hub nodes in both networks. In addition, we calculated the clustering coefficients of lncDN as the average of the clustering coefficient of all the vertices in lncDN [30], and the clustering coefficients of 550 randomly generated networks with the same degree sequence as lncDN [31]. The average clustering coefficient of the randomized networks was 0.45 ± 0.01, which was significantly smaller than 0.81 ($\text{p-value} < 10^{-10}$, *z*-test), the clustering coefficient of lncDN. Likewise, we generated 550 randomized networks with the same degree sequence as DlncN. The average clustering coefficient of the 550 randomized networks was 0.53 ± 0.01, which was significantly smaller than 0.91 ($\text{p-value} < 10^{-10}$, *z*-test),the clustering coefficient of DlncN. These results indicated that lncDN and DlncN revealed obvious community structure. Therefore, in the following section, we would like to analyze the modules of lncDN and DlncN.

Modules of lncDN and DlncN. We clustered the lncDN and DlncN by MINE (http://apps.cytoscape.org/apps/mine), a plugin of cytoscape software [29]. As a result, we obtained 14 modules of lncDN and 19 modules of DlncN. The size of each module had a board distribution (Figure S5).

Although the lncDN layout was generated without any knowledge on the disease classes, the resulted network was visibly clustered according to major disease classes (Figure 2-(a)). For example, most (seven in 11) diseases that belonged to cardiovascular were associated with *ANRIL* and were clustered together. Most (seven in eight) dermatological diseases were associated with *XIST* and were also clustered into one cluster. However, some lncRNAs might be of special importance as they were implicated in different cancers which were not clustered into a single cluster. For example, *ANRIL* was associated with 14 types of cancer and *MEG3* was associated with 18 types of cancer. These observations suggested the complexity and heterogeneity of different types of cancers.

In DlncN, lncRNA nodes were colored based on the class of diseases in which these lncRNAs were implicated. Nodes were light purple if the corresponding lncRNAs were associated with more than one disease class (Figure 2-(b)). We found that most lncRNAs were only implicated in certain type of cancers, and they were mostly clustered into one module. For example, 17 lncRNAs were only related to brain cancer, 22 lncRNAs were only related to breast cancer, and 84 lncRNAs were only related to glioma. However, the major hubs were related to more than one disease class, such as *XIST* that was related to 12 disease classes, *H19* was related to seven disease classes, *ANRIL* was related to six disease classes, and *MEG3* was related to four disease classes. These results were consistent with the fact that many lncRNAs exhibited tissue-specific expression [32] and that a few lncRNA were expressed across many tissues, such as *MEG3*, *XIST*, and *H19* [33].

Prediction of lncRNAs implicated in diseases

We applied the propagation algorithm to predict the candidate gene-disease associations on the coding-non-coding gene-disease bipartite network. In this algorithm, there were two parameters to be tuned: α and t. The parameter α gave the relative importance between the information that other genes contribute and the initial information. This parameter was tuned by LOOCV tests and "0.618" was chosen as our α based on this procedure. The parameter t represented the number of iterations. The iterative computation would stop if the mean square deviation of the coding-non-coding gene-disease association score matrix between the t-th iteration and the $(t-1)$-th iteration was not greater than 0.00001. With these two parameters, our algorithm ranked 2139 potential gene-disease pairs (768 lncRNA-disease pairs and 1371 coding gene-disease pairs) within top 1% for all the diseases. In the LOOCV procedure, our method achieved an AUC of 0.7881.

Robustness of our bipartite network

We tested the robustness of the coding-non-coding gene-disease bipartite network using the method of Multiple Survival Screening (MSS) [34], which was introduced to test the robustness of cancer causing genes by re-sampling experiments. Here, we performed 1000 times of re-sampling of our coding-non-coding gene-disease associations to predict the potential gene-disease associations. In each re-sampling experiment, we randomly removed 10% edges from the coding-non-coding gene-disease bipartite network, and then applied the propagation algorithm to predict the potential gene-disease associations on the remaining bipartite network with 90% edges.

If a gene g was ranked within top 1% among all the genes according to the score vector for a given disease q, then the gene g was predicted to be associated with the disease q, i.e. the gene-disease pair (g,q) was considered as a predicted association. Applying the propagation algorithm on the coding-non-coding gene-disease bipartite network, we obtained 2139 predicted associations. For a predicted association (g_i,q_i) $(1 \le i \le 2139)$, if the rank of g_i was within top 1% in a re-sampling experiment, then n_i was increased by one. A vector $N = (n_1, n_2, \cdots, n_{2139})$ was obtained, where $n_i \in [0,1000]$, meant the times of the predicted association (g_i,q_i) could be also predicted in 1000 re-sampling experiments. Furthermore, we performed 1000 times of random experiments. In each experiment, we randomly shuffled the coding non-coding gene-disease bipartite network, while keeping the degree of each gene and each disease in the bipartite network unchanged as above, and then applied the propagation algorithm to the randomized network. Similarly, a vector $N^r = \left(n_1^r, n_2^r, \cdots, n_{2139}^r\right)$($r$ represented random) was obtained. We found that N was significantly larger than N^r ($\text{p-value} < 10^{-10}$, *z*-test), with most of n_is larger than 700, and most of n_i^rs smaller than 250 (Figure 4). These findings suggested that even the 10% edges of the coding-non-coding gene-disease bipartite network were deleted, the predictive results were still stable. Therefore, our coding-non-coding gene-disease bipartite network was sufficiently robust to predict potential coding or non-coding gene disease associations.

Leave-one-out cross-validation tests

To evaluate the power of our method, we applied the LOOCV procedure. In each test of LOOCV, a single gene-disease association was removed from the coding-non-coding gene-disease bipartite network, and the method was evaluated by its success in reconstructing the hidden association. If the degree of gene or disease node in the removed gene-disease association was exactly one, then the gene or disease would be an isolated node. An isolated node in the propagation algorithm could not get any information, so we removed the nodes whose degree was one in

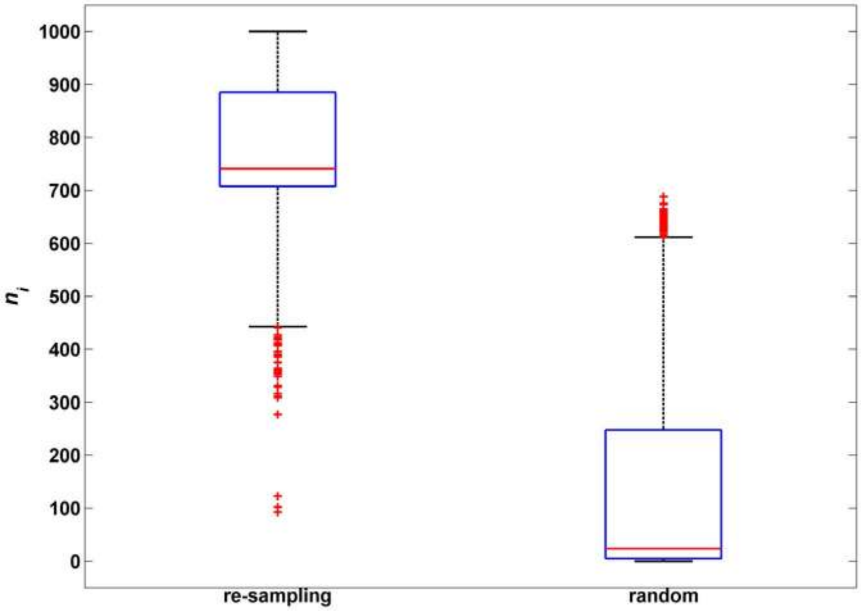

Figure 4. Comparison between re-sampling and random experiments to investigate the robustness of the bipartite network. Two box graphs represent the re-sampling experiments (left) and the random experiments (right), n_i means the time of predicted association (g_i,q_i) can be also predicted in re-sampling experiments and random experiments.
doi:10.1371/journal.pone.0087797.g004

LOOCV. Finally, we kept 532 links between 103 diseases and 163 genes (mapped to 44 lncRNAs and 119 protein-coding genes) that were to be used in LOOCV. To the best of our knowledge, this was the first work of predicting potential lncRNA-disease associations in a network view, therefore, no previous methods could be directly compared with our method. We would compare the predictive performance of the propagation algorithm on different networks.

The receiver operating characteristics (ROC) curve was used to measure the performance of our method, which plotted the true positive rate (TPR) versus the false positive rate (FPR) at different rank thresholds. In LOOCV, for a rank threshold k $(1 \le k \le 100)$, TPR meant the percentage of the leave-out associations obtaining the rank within top k%; FPR meant the percentage of unassociated gene-disease pairs obtaining the rank within top k%. When the rank threshold was varied between 1 and 100, the corresponding TPR and FPR were obtained. In this way, the ROC curve could be plotted, and the AUC could be calculated. Following this procedure, we performed LOOCV over lncRNA-disease association network, and achieved an AUC of 0.6820. The ROC was shown in Figure 5.

Aiming at improving the performance of our method, we integrated the protein coding gene-disease associations with lncRNA-disease associations to construct the coding-non-coding gene-disease bipartite network. Here, we also performed LOOCV procedure over coding-non-coding gene-disease bipartite network and obtained an AUC of 0.7881. The ROC was shown in Figure 5. Clearly, the integration of protein coding gene-disease associations could improve the performance of our method. One reason of the improvement was that the number of edges in the bipartite network was increased by the integration. Therefore, potential genes could get more information from other genes and diseases in propagation and could be better predicted. The better performance might be also attributed to the fact that coding and non-coding genes were cooperated in human diseases. Therefore, the performance of our method would be further improved after obtaining more known lncRNA-disease associations, and more associations between coding genes and non-coding genes.

Moreover, we performed the LOOCV procedure over 50 random networks. The mean FPR and mean TPR were used to plot the ROC curve (Figure 5), and we achieved an AUC of 0.5005, smaller than AUCs of other two cases. This indicated that our coding-non-coding gene-disease bipartite network could reflect some mechanisms of human complex diseases, and our method could discover potential lncRNA-disease associations.

Case study

For each disease, all the genes (including coding and non-coding genes) were ranked according to their association scores with the disease. The genes ranked within top 10 (this was a user-defined threshold, and 10 was used here) were considered as the potential genes involved in the given disease. For all the 214 diseases in the coding-non-coding gene-disease bipartite graph, we uncovered 768 novel lncRNA-disease associations between 66 lncRNAs and 193 diseases.

To further demonstrate the power of our method, we examined the results for three multifactorial diseases (i.e. Alzheimer's disease (MIM: 176807), pancreatic cancer (MIM: 260350) and gastric cancer (MIM: 137215)) as case studies. For each case, the top 10 genes including protein-coding genes and lncRNAs were listed in Table 1.

Results for Alzheimer's disease. Alzheimer's disease (AD) is the most common form of dementia in the elderly [35] and it is characterized by slow progressive loss of memory, cognitive abilities, and intellectual functions [36]. Currently, it has been reported that 23 genes including 6 lncRNAs and 17 protein-coding genes are associated with AD. The association scores of these 23 genes were higher than unassociated genes. In the top-10 ranked genes unassociated with AD, we found that the rank of lncRNA *H19* was two, and the rank of lncRNA *PVT1* was three. *H19* had been associated with glioblastoma [37] and *PVT1* had been associated with glioma [24]. Both glioblastoma and glioma

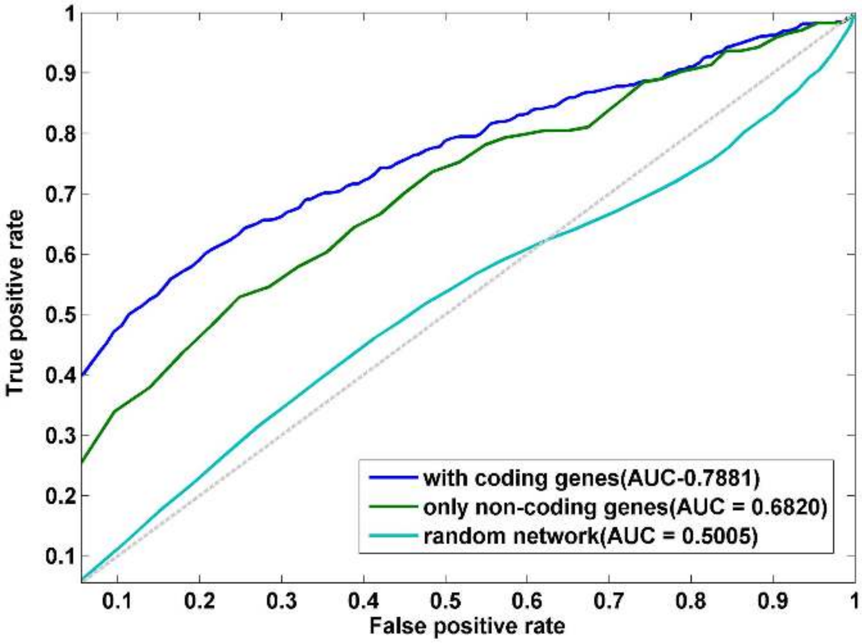

Figure 5. A comparison between the performance of our propagation algorithm on coding-non-coding gene-disease bipartite network and that on lncRNA-disease association network. The blue line represents the ROC curve of taking LOOCV over coding-non-coding gene-disease bipartite network, and an AUC of 0.7881 was obtained. The cyan line represents the ROC curve of taking LOOCV over lncRNA-disease association network, and an AUC of 0.6820 was obtained. The light blue line represents the ROC curve of taking LOOCV over random networks, and an average AUC of 0.5005 was obtained.
doi:10.1371/journal.pone.0087797.g005

Table 1. The top-10 ranked genes for three case studies.

Alzheimer's disease					
gene	**ACC/MIM**	**Rank**	**gene**	**ACC/MIM**	**Rank**
COL4A2	120090	1	IL1RN	147679	6
H19	*NR_002196*	2	EPO	133170	7
PVT1	*NR_003367*	3	SOD2	147460	8
ALOX5AP	603700	4	VEGF	192240	9
PON1	168820	5	F2	176930	10
Pancreatic cancer					
gene	**ACC/MIM**	**Rank**	**gene**	**ACC/MIM**	**Rank**
H19	*NR_002196*	1	FGFR3	134934	6
ANRIL	*NR_003529*	2	*UCA1*	*NR_015379*	7
BC200	*NR_001568*	3	PIK3CA	171834	8
MEG3	*NR_002766*	4	CDH1	192090	9
XIST	*NR_001564*	5	*SRA*	*AF092038*	10
Gastric cancer					
gene	**ACC/MIM**	**Rank**	**gene**	**ACC/MIM**	**Rank**
XIST	*NR_001564*	1	*ANRIL*	*NR_003529*	6
MALAT-1	*NR_002819*	2	TP53	191170	7
MEG3	*NR_002766*	3	FGFR3	134934	8
PVT1	*NR_003367*	4	PTEN	601728	9
BRCA2	600185	5	RAD54L	603615	10

In this table, the susceptibility genes (protein-coding genes and lncRNAs) for three case studies including Alzheimer's disease, Pancreatic cancer and Gastric cancer were listed. The genes in italic were lncRNAs and the others were protein-coding genes.
doi:10.1371/journal.pone.0087797.t001

were brain or neuron related diseases and AD was described as a neurological disease. All these suggested the relationship between these two lncRNAs and AD.

Results for Pancreatic cancer. Pancreatic cancer has a high mortality rate and the 5-year relative survival rate is less than 5% [38]. It has been previously shown that 18 genes including 5 lncRNAs and 13 protein-coding genes are implicated in pancreatic cancer. The association scores of these 18 genes were also higher than unassociated genes. In the top-10 ranked genes unassociated with pancreatic cancer, we found that the rank of *ANRIL* was two. Pasmant et al. [5] confirmed the pivotal role of *ANRIL* in regulation of *CDKN2A/B* expression through a *cis*-acting mechanism in mice and *ANRIL* implicated in proliferation and senescence. Furthermore, the association of *CDKN2A* (MIM:600160) with pancreatic cancer had been curated in OMIM [11]. The rank of *UCA1* was seven, and Kaneko et al. [39] showed that *UCA1* and *BMF* were upregulated in gallbladder epithelia of children with pancreaticobiliary malfunction. Therefore, our results suggested that *UCA1* might be associated with pancreatic disease.

Results for Gastric cancer. Gastric cancer is a high morbidity cancer and has varied morbidities in different populations [40]. It has been presented that 15 genes including 4 lncRNAs and 11 protein-coding genes are implicated in gastric cancer. The association scores of these 15 genes were higher than unassociated genes. In potential genes implicated in gastric cancer, the rank of *XIST* was one. Weakley et al.'s study [41] showed that *XIST* was differentially expressed in preneoplastic cells located in gastric fundus that could lead to gastric cancer. The rank of *MEG3* was three, and *MEG3* was reported to function as a novel lncRNA tumor suppressor [42].

Discussion

The lncRNA-disease association network was constructed, from which two relevant networks, lncDN and DlncN, were generated accordingly. These networks provided a unified framework of all known lncRNA and disease associations and a new network view for the study of the lncRNA-disease associations. The detailed lncRNA-disease association network (Figure S1) showed all the known lncRNA-disease associations. Furthermore, a computational iterative algorithm was applied to mine the hidden lncRNA-disease associations. The results showed that our method could provide insightful suggestions of lncRNA implicated in diseases.

Our method had some limitations that should be acknowledged. First, the analysis of the function of lncRNAs on a whole genome-wide scale was limited due to the diversity, lack of knowledge and specificity of expression of lncRNAs, and the lack of lncRNA functional annotation. Second, the shortage of lncRNA-disease associations limited the analysis of the mechanism of lncRNAs implicated in disease on a larger network. Finally, due to the lack of interactions and similarities between non-coding genes and protein coding genes, it was insufficient in biological meaning to replace the gene similarity matrix in Formula (4) by the weighted gene projection W.

Supporting Information

Table S1 20 tests on random networks to show that the propagation method converges. We did 20 tests. In test i $(1 \leq i \leq 20)$, we applied the propagation method on 100 random networks with the mean square deviation threshold between t-th iteration and $(t-1)$-th iteration being 10E-i. The average iteration times were calculated and listed.
(XLSX)

Figure S1 Bipartite-graph representation of the lncRNA-disease association network. A disease (circle) and a lncRNA (hexagons) are connected if the lncRNA is implicated in the disease. The size of a node is proportional to the degree of the node. The color of a disease node (circle) represents the class which it belongs. The names of 20 disease classes are shown on the right panel. The color of a lncRNA node (hexagons) is based on the class of diseases in which the corresponding lncRNA implicated. LncRNA Nodes are light purple if the corresponding lncRNAs are associated with more than one disease class. We label the diseases (lncRNAs) associated with more than five lncRNAs (diseases) by their names.
(TIF)

Figure S2 Degree distribution of full lncRNA-disease association network. (a) The degree distribution of the full lncRNA-disease association network. It closely follows a power-law distribution. Here, k represents degree, $p(k)$ denotes the fraction of nodes with a given degree k. (b) Degree distribution of disease nodes in lncRNA-disease association network. (c) Degree distribution of lncRNA nodes in lncRNA-disease association network.
(TIF)

Figure S3 Degree distribution of lncDN and DlncN. (a) Degree distribution of lncDN. It closely follows a power-law distribution. Here, k represents degree, $p(k)$ denotes the fraction of nodes with a degree k. (b) Degree distribution of DlncN. It closely follows a power-law distribution. Here, k represents degree, $p(k)$ denotes the fraction of nodes with a degree k.
(TIF)

Figure S4 Degree distributions of average clustering coefficients of nodes in lncDN and DlncN. (a) Degree distribution of average clustering coefficients of nodes in lncDN. (b) Degree distribution of average clustering coefficients of nodes in DlncN. Both distributions are closely following a power-law distribution.
(TIF)

Figure S5 Distribution of module sizes in lncDN and DlncN. (a) The module sizes of 14 modules in lncDN. (b) The module sizes of 19 modules in DlncN.
(TIF)

Acknowledgments

We would like to thank Xiaoke Ma, Xindong Zhang, Yuxuan Hu, Qiangfeng Cliff Zhang, and Shihua Zhang for their helpful advice and discussion. We also thank Brianna Ashlyn Lam for her input on refining the manuscript.

Author Contributions

Conceived and designed the experiments: XFY LG. Performed the experiments: XFY XLG. Analyzed the data: XFY. Contributed reagents/materials/analysis tools: XFY XLG XHS HW FS BBW. Wrote the paper: XFY.

Reference

1. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43: 904–914.
2. Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21: 354–361.

3. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22: 1775–1789.
4. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.
5. Pasmant E, Sabbagh A, Vidaud M, Bieche I (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J 25: 444–448.
6. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464: 1071–1076.
7. Ji P, Diederichs S, Wang W, Boing S, Metzger R, et al. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22: 8031–8041.
8. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, et al. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14: 723–730.
9. Cui Z, Ren S, Lu J, Wang F, Xu W, et al. (2013) The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor. Urol Oncol 31: 1117–1123.
10. Chen G, Wang Z, Wang D, Qiu C, Liu M, et al. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 41: D983–986.
11. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–517.
12. Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. Phys Rev E Stat Nonlin Soft Matter Phys 76: 046115.
13. Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2004) Learning with local and global consistency. Advances in neural information processing systems 16: 321–328.
14. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6: e1000641.
15. Guo X, Gao L, Liao Q, Xiao H, Ma X, et al. (2013) Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res 41: e35.
16. Sun BK, Tsao H (2008) X-chromosome inactivation and skin disease. J Invest Dermatol 128: 2753–2759.
17. Kawakami T, Okamoto K, Sugihara H, Hattori T, Reeve AE, et al. (2003) The roles of supernumerical X chromosomes and XIST expression in testicular germ cell tumors. J Urol 169: 1546–1552.
18. Vincent-Salomon A, Ganem-Elbaz C, Manie E, Raynal V, Sastre-Garau X, et al. (2007) X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors. Cancer Res 67: 5134–5140.
19. Sparago A, Cerrato F, Vernucci M, Ferrero GB, Silengo MC, et al. (2004) Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. Nat Genet 36: 958–960.
20. Eggermann T, Begemann M, Spengler S, Schroder C, Kordass U, et al. (2010) Genetic and epigenetic findings in Silver-Russell syndrome. Pediatr Endocrinol Rev 8: 86–93.
21. Bartholdi D, Krajewska-Walasek M, Ounap K, Gaspar H, Chrzanowska KH, et al. (2009) Epigenetic mutations of the imprinted IGF2-H19 domain in Silver-Russell syndrome (SRS): results from a large cohort of patients with SRS and SRS-like phenotypes. J Med Genet 46: 192–197.
22. Looijenga LH, Verkerk AJ, De Groot N, Hochberg AA, Oosterhuis JW (1997) H19 in normal development and neoplasia. Mol Reprod Dev 46: 419–439.
23. Zhang X, Zhou Y, Mehta KR, Danila DC, Scolavino S, et al. (2003) A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J Clin Endocrinol Metab 88: 5119–5126.
24. Zhang X, Sun S, Pu JK, Tsang AC, Lee D, et al. (2012) Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis 48: 1–8.
25. Bjornsson HT, Brown LJ, Fallin MD, Rongione MA, Bibikova M, et al. (2007) Epigenetic specificity of loss of imprinting of the IGF2 gene in Wilms tumors. J Natl Cancer Inst 99: 1270–1273.
26. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685–8690.
27. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
28. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24: 282–284.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
30. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
31. Chung F, Lu L (2002) Connected components in random graphs with given expected degree sequences. Annals of combinatorics 6: 125–145.
32. Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23: 1494–1504.
33. Gibb EA, Vucic EA, Enfield KS, Stewart GL, Lonergan KM, et al. (2011) Human cancer long non-coding RNA transcriptomes. PLoS One 6: e25915.
34. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, et al. (2010) Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun 1: 34.
35. Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA (2003) Alzheimer disease in the US population: prevalence estimates using the 2000 census. Arch Neurol 60: 1119–1122.
36. Guttman R, Altman RD, Nielsen NH (1999) Alzheimer disease. Report of the Council on Scientific Affairs. Arch Fam Med 8: 347–353.
37. Qureshi IA, Mattick JS, Mehler MF (2010) Long non-coding RNAs in nervous system function and disease. Brain Res 1338: 20–35.
38. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, et al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 41: 986–990.
39. Kaneko K, Ito Y, Ono Y, Tainaka T, Tsuchiya H, et al. (2011) Gene expression profiling reveals upregulated UCA1 and BMF in gallbladder epithelia of children with pancreaticobiliary maljunction. J Pediatr Gastroenterol Nutr 52: 744–750.
40. Bevan S, Houlston R (1999) Genetic predisposition to gastric cancer. Qjm 92: 5–10.
41. Weakley SM, Wang H, Yao Q, Chen C (2011) Expression and function of a large non-coding RNA gene XIST in human cancer. World J Surg 35: 1751–1756.
42. Zhou Y, Zhang X, Klibanski A (2012) MEG3 noncoding RNA: a tumor suppressor. J Mol Endocrinol 48: R45–53.