Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA

Xing Chen

Accumulating experimental studies have indicated that lncRNAs play important roles in various critical biological processes and their alterations and dysregulations have been associated with many important complex diseases. Developing effective computational models to predict potential disease-lncRNA associations could benefit not only the understanding of disease mechanism at lncRNA level, but also the detection of disease biomarkers for disease diagnosis, treatment, prognosis and prevention. However, known experimentally confirmed disease-lncRNA associations are still very limited. In this study, a novel model of HyperGeometric distribution for LncRNA-Disease Association inference (HGLDA) was developed to predict lncRNA-disease associations by integrating miRNA-disease associations and lncRNA-miRNA interactions. Although HGLDA didn’t rely on any known disease-lncRNA associations, it still obtained an AUC of 0.7621 in the leave-one-out cross validation. Furthermore, 19 predicted associations for breast cancer, lung cancer, and colorectal cancer were verified by biological experimental studies. Furthermore, the model of LncRNA Functional Similarity Calculation based on the information of MiRNA (LFSCM) was developed to calculate lncRNA functional similarity on a large scale by integrating disease semantic similarity, miRNA-disease associations, and miRNA-lncRNA interactions. It is anticipated that HGLDA and LFSCM could be effective biological tools for biomedical research.

Based on the assumption of the central dogma of molecular biology, genetic information is stored in protein-coding genes and RNA is just an intermediary between a DNA sequence and its encoded protein. However, sequence analysis shown that there were only ~20,000 protein-coding genes in the human genome and more than 98% of the human genome does not encode protein sequences, yielding tens of thousands of non-coding RNAs (ncRNAs). Based on accumulating experimental evidences, these ncRNAs have played very fundamental and critical roles in various biological processes. Based on whether transcript lengths are larger than 200 nucleotides, ncRNAs can be further divided into small ncRNA (such as miRNA, siRNA, and piRNA) and long ncRNA (lncRNA). Long non-coding RNAs (lncRNAs) are a heterogeneous class of ncRNAs with non-protein-coding transcripts longer than 200 nucleotides. In comparison with protein-coding genes, lncRNAs have the following differences: (1) lncRNAs have less conservation across species; (2) lncRNAs have relatively lower expression level and much more tissue-specific pattern; (3) lncRNAs have longer, but fewer, exons. In the early 1990s, H19 and Xist were first identified based on traditional gene mapping approaches. In the 1990s, H19 and Xist were first identified based on traditional gene mapping approaches. In the 1990s, H19 and Xist were first identified based on traditional gene mapping approaches.

1National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing, 100190, China. 2Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China. Correspondence and requests for materials should be addressed to X.C. (email: xingchen@amss.ac.cn)
In recent years, there has been rapid development in both experimental technology and computational prediction algorithms for IncRNA discovery. Thousands of IncRNAs have been discovered in eukaryotic organisms ranging from nematodes to humans, including cardiovascular diseases, neurological disorders, diabetes, HIV, and various types of cancers, such as breast cancer, hepatocellular carcinoma, prostate cancer, lung cancer, liver cancer, and so on. Therefore, HOTAIR was considered as potential biomarker in various types of cancers. Except for HOTAIR, the dysfunction of IncRNA H19 is also involved in various diseases. For example, H19 could be used as a potential diagnostic marker for the early recurrence of bladder cancer. Furthermore, it has been demonstrated that down-regulation of H19 significantly decreases breast and lung cancer cell clonogenicity and anchorage-independent growth based on a knockdown approach. Several experimental studies have also shown that IncRNA BCAR4 is associated with breast cancer, which is expressed in 27% of primary breast tumors. Specifically, in human ZR-75-1 and MCF7 breast cancer cells, the forced expression of BCAR4 causes cell proliferation in the absence of estrogen and in the presence of various antiestrogens, indicating BCAR4 could considered as a proper target for the treatment of antiestrogen-resistant breast cancer.

Considering the important roles of IncRNAs in various biological processes regulation and complex diseases development and progression, potential disease-IncRNA associations identification could not only benefit the underlying disease mechanism mining at IncRNA level, but also facilitate disease biomarkers detection and drug discovery for disease diagnosis, treatment, prognosis and prevention. Computational models and tools can effectively decrease the time and cost of biological experiments by quantifying the association probability of each IncRNA-disease pair and verifying most promising IncRNA-disease pairs with high scores based on further biological experimental validation. Nowadays, developing effective computational models by integrating various kinds of biological datasets to prioritize disease-related IncRNAs has become one of the most important and attracting topics in the fields of both IncRNAs and complex diseases.

Some computational models have been developed to infer novel disease-IncRNA associations. In the previous study, Chen et al. presented a semi-supervised learning method, LRSLDA, to infer novel human IncRNA-disease associations. LRSLDA was developed based on the assumption that similar diseases tend to interact with functionally similar IncRNAs and the framework of Laplacian Regularized Least Squares. LRSLDA is a reliable tool for IncRNA-disease association prediction. More importantly, it does not need negative samples. However, the parameter selection problem and the problem of combining two different classifiers into the final classifier exist in this method. Based on the same assumption, Sun et al. presented a method to constructed a IncRNA-IncRNA functional similarity network, then they proposed a global network-based computational method named RWIncD by integrating disease similarity network, IncRNAs functional network and known IncRNA-disease associations. However, this method can't be applied to the IncRNAs without any known associated diseases. Li et al. developed a simple genomic location based bioinformatics method for the prediction of novel associations between IncRNAs and vascular disease. However, not all of the IncRNAs are related with their neighbor genes and no statistical tests were used, which resulted in limitations of this method. Yang et al. investigated IncRNA-disease associations by constructing the IncRNA-disease association network and coding-non-coding gene-disease bipartite network based on known associations between diseases and disease genes. Then, a propagation algorithm was applied to infer the underlying IncRNA-disease associations. This method also has some limitations, such as the lack of the information of non-coding genes and protein coding genes interactions and similarities and IncRNA functional annotation. Zhao et al. developed the naive Bayesian classifier to identify cancer-related IncRNAs based on the integration of genome, regulome and transcriptome data. The important limitation of this method is that they
regard the unknown lncRNA-disease associations as negative samples, which would largely influence the predictive accuracy of the method. Recently, based on the findings that lncRNAs that sharing significantly enriched interacting miRNAs tend to be associated with similar diseases, Zhou et al. proposed a novel method named RWRHLD to identify candidate lncRNA-disease associations by integrating miRNA-associated lncRNA-lncRNA crosstalk network, disease-disease similarity network, and known lncRNA-disease association network into a heterogeneous network. Then, a random walk was implemented on this heterogeneous network. This method can only predict associations for the lncRNAs that have lncRNA-miRNA interaction datasets, limiting the wide application of RWRHLD. Aforementioned methods all need the prior information of known experimentally verified lncRNA-disease association. So far, although plenty of biological datasets about lncRNA sequence and expression have been generated and stored in some publicly available databases, such as NRED, IncRNAdb, NONCODE, the number of lncRNAs reported to be associated with diseases is still very limited. Liu et al. developed a method by integrating human IncRNA and gene expression profiles, and human disease-associated gene data. This method didn’t rely on known lncRNA-disease associations and obtained an AUC of 0.7645 for non-tissue-specific lincRNAs. However, too many false positives would be brought based on the ROC curve in that paper.

Nowadays, plenty of experimentally confirmed miRNA-disease associations have been collected in various databases. Therefore, the model of HyperGeometric distribution for LncRNA-Disease Association inference (HGLDA) was developed here to predict potential lncRNA-disease associations by integrating known miRNA-disease associations and lncRNA-miRNA interactions. Although HGLDA didn’t rely on any known disease-related lncRNAs associations, it still obtained a reliable AUC of 0.7621 in the leave-one-out cross validation (LOOCV) based on known experimentally verified lncRNA-disease associations from the LncRNA-Disease database. HGLDA was also applied to predict Breast Cancer, Lung Cancer, and Colorectal Cancer-related IncRNAs. Seven, seven, and five predicted potential associations with false discovery rate (FDR) less than 0.05 have been confirmed by recent biological experiments for these three important human complex diseases, respectively. Above results effectively demonstrated its potential ability of inferring disease-lncRNA associations and detecting biomarkers detection for human disease diagnosis, treatment, prognosis and prevention. Furthermore, the model of LncRNA Functional Similarity Calculation based on the information of MiRNA (LFSCM) was developed to quantitatively calculate lncRNA functional similarity on a large scale by integrating disease semantic similarity, miRNA-disease associations, and miRNA-lncRNA interactions.

**Results**

**Performance evaluation of potential lncRNA-disease association prediction.** HGLDA was applied to the known experimentally verified lncRNA-disease associations in the lncRNADisease database in the framework of LOOCV. Each known disease-lncRNA association was left out in turn as test sample. How well this test sample was ranked relative to the candidate samples (all the disease-lncRNA pairs without the evidence to confirm their association) was evaluated. When the rank of this test sample exceeds the given threshold, this model was considered to provide a successful prediction. When the thresholds were varied, true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity) could be obtained. Here, sensitivity refers to the percentage of the test samples whose ranking is higher than the given threshold. Specificity refers to the percentage of samples that are below the threshold. Receiver-operating characteristics (ROC) curve was drawn by plotting TPR versus FPR at different thresholds. Area under ROC curve (AUC) was further calculated to evaluate the performance of HGLDA. AUC = 1 indicates perfect performance and AUC = 0.5 indicates random performance. As a result, HGLDA achieved an AUC of 0.7621 (see Fig. 1). One important fact must be pointed out is that HGLDA predict potential lncRNA-disease association without relying on the information of known disease-lncRNA associations. Although previous study of predicting potential lncRNA-disease associations by integrating disease-gene associations and gene-lncRNA co-expression relationship obtained a comparable AUC of 0.7645, the ROC curve in that study is much below the ROC curve in this study when FPR is small, which is particularly important for practical biological research. More importantly, available experimentally verified disease-miRNA associations are still comparatively rare relative to the known disease-gene associations. The performance of HGLDA would be further improved when more known miRNA-disease associations could be obtained in the future.

**Case studies of potential lncRNA-disease association prediction.** HGLDA was applied to predict potential disease-lncRNA associations for all the diseases investigated in this article. Potential predictive associations with significant FDR values were publicly released to benefit the biological experimental validation (see Supplementary Table 1). It is anticipated that these potential lncRNA-disease associations which significantly share common miRNAs could be validated by biological experiments and provide important complementary for experimental studies. Especially, plenty of evidences have demonstrated that lncRNAs plays important roles in various kinds of human cancers. Therefore, case studies about three kinds of important cancers were implemented to show the predictive performance of HGLDA. Predictive results were confirmed based on recent experimental literatures.

As the second leading cause of female cancer death, breast cancer comprises 22% of all cancers in women. Breast cancer is caused because of multiple molecular alterations and traditionally diagnosed...
based on histopathological features such as tumor size, grade and lymph node status\textsuperscript{69}. Researches showed that lncRNA plays an important role in many biological processes and is strongly associated with the formation of various cancers including breast cancer\textsuperscript{69,70}. To better diagnose and treat breast cancer, it is necessary to predict breast cancer-related lncRNAs and identify lncRNA biomarkers\textsuperscript{70}. HGLDA was implemented to prioritize candidate lncRNAs for breast cancer. As a result, seven lncRNAs with significant FDR less than 0.05 have been confirmed based on recent experimental literatures (see Table 1).

For example, XIST, KCNQ1OT1 and NEAT1 are there experimentally confirmed breast cancer related lncRNAs, which have been ranked 1st, 8th, and 12th in the predicted list based on the model of HGLDA, respectively. The XIST RNA signal variability in the BRCA1 breast tumor is correlated with chromosomal genetic abnormalities, and BRCA1 breast tumors often contain cells showing multiple XIST RNA domains per nucleus\textsuperscript{71}. KCNQ1OT1 is induced by estrogen in estrogen receptor-alpha (ER\textsubscript{α}) expressing breast cancer cells (see Figure 1).

As a result, HGLDA achieved an AUC of 0.7621, demonstrating its reliable predictive ability even if potential lncRNA-disease associations were predicted without relying on the information of known disease-lncRNA associations in the model of HGLDA.

| Disease            | lncRNA          | Evidence (PMID)                                      |
|--------------------|-----------------|-----------------------------------------------------|
| Breast cancer      | MALAT1          | 24525122;19379481                                   |
| Breast cancer      | H19             | 16707459;14729626;12419837                           |
| Breast cancer      | CDKN2B-AS1      | 17440112;20956613                                   |
| Breast cancer      | NEAT1           | 25417700;23825647                                   |
| Breast cancer      | XIST            | 17545591                                            |
| Breast cancer      | KCNQ1OT1        | 21304052                                            |
| Breast cancer      | HOTAIRM1        | 25296969                                            |
| Lung cancer        | EPB41L4A-AS1    | BCYRN1 1697389;9490301                              |
| Lung cancer        | MALAT1          | 20937273;24757675;24667321                           |
| Lung cancer        | TUG1            | 24853421                                            |
| Lung cancer        | GAS5            | 24357161;23676682                                   |
| Lung cancer        | HOTAIR          | 25491133;24591352;24155936                           |
| Lung cancer        | H19             | 16707459;8838103;7700644                              |
| Lung cancer        | NEAT1           | 25010625                                            |
| Colorectal cancer  | XIST            | 17143621;22879877                                   |
| Colorectal cancer  | HOTAIR          | 24531795;21862635;24667321                           |
| Colorectal cancer  | MALAT1          | 21503572;25446987;25031737                           |
| Colorectal cancer  | KCNQ1OT1        | 16965397;11340379;23660942                           |
| Colorectal cancer  | H19             | 18719115;19926638;22121898                           |

Table 1. HGLDA was applied to three kinds of important cancer (breast cancer, lung cancer, and colorectal cancer). As a result, 19 predicted lncRNA-disease pairs with significant FDR less than 0.05 have been confirmed based on recent experimental literatures.
breast cancer cells and further mediate CDKN1C repression through epigenetic repression. The alternative splicing of NEAT1 may play important role in nicotine induced breast cancer development and breast cancer patients with high level of NEAT1 expression show low survival rate.

Lung cancer, which can be roughly divided into two groups: non-small cell lung cancer (80.4%) and small cell lung cancer (16.8%) considering disease patterns and treatment strategies, is the leading cause of cancer-related death worldwide in both men and women. There are estimated 1.4 million deaths resulting from lung cancer each year. Data show that the risk of lung cancer mortality is even greater than the combination of the next three most common cancers (colon, breast and prostate). Specially, five-year survival rate of lung cancer patients is only approximately 15%, which is much lower than other cancers types. To diagnose and treat lung cancer in a better and more efficiently way, more attentions are focused on the deregulation of protein-coding genes to identify oncogenes and tumor suppressors in the last decades.

Recent researches have shown that IncRNAs play a critical role the development and progression of lung cancers. Potential lung cancer-related IncRNAs were obtained by selecting candidate IncRNAs with FDR less than 0.05. Seven predicted IncRNAs have been confirmed by independent experimental literatures (see Table 1). According to biological experiments in several studies, it has been confirmed that MALAT1 is a non-coding RNA which plays important roles in many different cancers. Specially it has been shown to be highly associated with metastasis of lung cancer and promote lung cancer cell motility by regulating motility related gene expression. Therefore, it could be an important biomarker for metastasis development in lung cancer. TUG is another lung cancer related IncRNA, which can be regulated by P53 to affect non-small cell lung cancer (NSCLC) cell proliferation in part by epigenetically controlling the expression of HOXB8. GASS, which can also be mediated by P53 pathway, is shown to be a tumor suppressor and down-regulated in NSCLC. These three IncRNAs were all ranked in the top five prediction list for lung cancer (11th, 13th, and 41st, respectively).

As the third most common cancer in men and the second in women, colorectal cancer is one of the most common malignancies in the world and an important threat to human health. Data shows that the 5.2% of men and 4.8% of women have the risk of colorectal cancer in the United States and the mortality rate caused by colorectal cancer is nearly 33% in the developed world. Critical mutations underlying the pathogenic mechanism of colorectal cancer have been confirmed. Especially, mutations and dysregulations of some IncRNAs have been linked with the development and progression of colorectal cancer. Five predicted colorectal cancer-related IncRNAs have been confirmed by experimental literature (see Table 1). XIST, MALAT1, H19, and KCNQ1OT1 were ranked in the top four prediction list of colorectal cancer. As a result, recent biological experiments indicated these four IncRNAs all showed high correlation with colorectal cancer. For example, evidences show that expression level change of or DNA amplification of XIST is associated with colorectal carcinoma. Also, MALAT1 plays important role in colorectal cancer development by promoting its invasion and metastasis. Down-regulation of MALAT1 will inhibit colorectal invasion by attenuating Wnt/β-catenin signaling. Moreover, the methylation state of H19 locus is highly related with colorectal cancer and the H19-derived microRNA also regulates colorectal cancer development. Loss of imprinting of KCNQ1OT1 is considered as a useful marker for diagnosis of colorectal cancer because of its frequent occurrences in colorectal cancer samples.

**IncRNA functional similarity.** LFSCM was applied to all the IncRNAs investigated in this study. Therefore, pairwise functional similarity among 1114 IncRNAs has been obtained (See Supplementary Table 2).

**Discussions**

Predicting potential disease-related IncRNAs by integrating various kinds of biological datasets is one of the most important and attracting topics for computational biology research, which is critical for understanding disease mechanism at the IncRNA level and disease biomarkers detection for disease diagnosis, prognosis and prevention. In this study, considering many miRNA-disease associations have been confirmed by recent biological experiments, the model of HGLDA was developed to predict potential disease-IncRNA associations on a large scale by selecting disease-IncRNA pairs which significantly share common miRNA partners. The important difference from previous computational researches about IncRNA-disease inference is that HGLDA doesn't rely on any known IncRNA-disease associations. To validate the performance of HGLDA, LOOCV was implemented on IncRNA-disease association dataset obtained from IncRNASDisease database and case studies were further implemented to three important cancers (Breast cancer, Lung Cancer, and Colorectal Cancer). Reliable performance has been obtained in the above validations. Therefore, to facilitate further biological experiment confirmation, significant IncRNA-disease pairs for all the diseases investigated in this study were publicly released. It is anticipated that HGLDA could further demonstrate its potential value for disease-IncRNA association inference and disease biomarker detection in the future.

Calculating IncRNA functional similarity could benefit IncRNA function inference and disease-related IncRNA prioritization. Therefore, based on the assumption that functionally similar IncRNAs tend to interact with functionally similar miRNAs, the model of LFSCM was further developed to quantitatively calculate IncRNA functional similarity. In this model, disease semantic similarity, miRNA-disease associations, and miRNA-IncRNA interactions were integrated on a large scale.
HGLDA obtained the reliable performance in both LOOCV and case studies about three kinds of important cancers, which could be largely attributed to the following several factors. Firstly, known experimentally verified disease-miRNA associations and lncRNA-miRNA interactions were integrated to infer the potential associations between lncRNAs and diseases. Secondly, both miRNA and lncRNA are ncRNAs, which don’t encode protein sequences. Therefore, predicting lncRNA-disease associations from miRNA-related datasets is more reasonable than previous study of integrating disease genes and gene-lncRNA co-expression relationship. More importantly, HGLDA doesn’t need the prior information of known lncRNA-disease associations, which ensure that this method could be applied to the diseases without any known related lncRNAs. Therefore, HGLDA represents a novel, effective, and important bioinformatics tool for the research of both complex diseases and lncRNAs.

Despite of the reliable performance of HGLDA, there are also some limitations in the model of HGLDA. Although HGLDA doesn’t rely on any known experimentally verified lncRNA-disease associations, its performance was not very satisfactory based on the evaluation of LOOCV and could be further improved by integrating more reliable biological datasets, such as disease semantic similarity, disease phenotypic similarity, lncRNA functional similarity, and lncRNA-related various interactions. Although the model of LFSCM can be applied to the lncRNAs without any known related diseases, it can’t be applied to those lncRNAs without any known miRNA interaction partners. Furthermore, lncRNA functional similarity was calculated based on known miRNA-disease associations and lncRNA-miRNA interactions, hence LFSCM tends to cause bias to lncRNAs with more miRNA interaction partners or/and lncRNAs with miRNA interaction partners which has been associated with more diseases. LFSCM would be further improved when more known datasets could be available and more reliable types of biological datasets could be integrated. More importantly, as what has been pointed out in the literature, it is unwise to use a single disease-related lncRNA to judge cancer risks for all the persons. Therefore, I planned to construct various cancer hallmark networks to effectively evaluate cancer risks based on the lncRNA profiles of each person. Finally, obtaining the tumor recurrence and metastases probability, predicting potential consequences after applying a specific drug to the patients, and identifying molecular signatures to evaluate and predict therapeutic results after cancer treatment in the framework of lncRNAs are three important problems in the personalized medicine, which could be considered in the future.

Methods

Human miRNA-disease associations. The human miRNA-disease association dataset was downloaded from HMDD in January, 2015, which included 10368 high-quality experimentally verified human miRNA-diseases associations from 3511 papers about 572 miRNA and 378 diseases. Then, duplicate associations with the different evidences were discarded and different miRNA copies were merged which produce the same mature miRNA. Finally, 5430 miRNA–disease associations were obtained, including 383 diseases and 495 miRNAs (see Supplementary Table 3).

lncRNA–miRNA interactions. lncRNA–miRNA interaction dataset was downloaded from starBase v2.0 database in January, 2015, which provided the most comprehensive experimentally confirmed lncRNA–miRNA interactions based on large scale CLIP-Seq data. After getting rid of duplicate interactions, 10112 lncRNA-miRNA interactions about 132 miRNAs and 1114 lncRNAs were obtained (see Supplementary Table 4).

Disease-lncRNA associations. To validate the performance of HGLDA, the recent version of lncRNA-disease association dataset in the lncRNADisease database was downloaded and LOOCV was implemented based on this golden-standard dataset. For this dataset, I got rid of duplicate associations with different evidences and the lncRNA-disease associations involved with either diseases or lncRNAs which were not contained in the dataset used in this paper. As a result, 183 lncRNA-disease associations were obtained and LOOCV was implemented based on these experimentally verified high-quality associations (see Supplementary Table 5).

HGLDA. The model of HGLDA was developed to predict potential disease-related lncRNAs (See Fig. 2). The hypergeometric distribution test was implemented for each lncRNA-disease pair by examining whether this lncRNA and disease significantly shared common miRNAs which can interact with both of them. The significance was measured by the P-value defined as follows:

\[
P = 1 - \sum_{i=0}^{x-1} \binom{L}{i} \binom{N-L}{M-i} / \binom{N}{M}
\]

where \(N\) is the total number of miRNAs which are associated with lncRNAs or diseases, \(M\) is the number of miRNAs interacting with this given lncRNA, \(L\) is the number of miRNAs interacting with this given disease, and \(x\) is the number of miRNAs that interact with both of them, respectively. Furthermore, FDR
correction was implemented to all calculated P-values and those lncRNA-disease pairs with FDR less than 0.05 were considered to be potential lncRNA-disease associations.

**LFSCM.** LFSCM is composed of the following three steps (See Fig. 3): calculating disease semantic similarity based on the disease MeSH descriptors and their direct acyclic graphs (DAGs); calculating miRNA functional similarity based on disease semantic similarity and disease-miRNA associations; calculating lncRNA functional similarity based on miRNA functional similarity and lncRNA-miRNA interactions. For the disease semantic similarity calculation, the method in the literature was adopted. The semantic similarity between two diseases was calculated based on the nodes shared by their disease.

**Figure 2.** Flowchart of HGLDA, demonstrating the basic ideas of predicting potential disease-related lncRNAs by integrating miRNA-disease associations and lncRNA-miRNA interactions. Firstly, the hypergeometric distribution test was implemented for each lncRNA-disease pair by calculating the P-value to indicate whether this lncRNA and disease significantly shared common miRNAs which can interact with both of them. Then, FDR correction was implemented to all calculated P-values. Finally, those lncRNA-disease pairs with FDR less than 0.05 were selected to be potential lncRNA-disease associations.
DAGs. The variable $S1$ is denoted as disease semantic similarity matrix, in which the entity $S1(i,j)$ in row $i$ column $j$ represents the semantic similarity between disease $i$ and $j$.

For the miRNA functional similarity, the semantic similarity of their associated disease groups was measured. The similarity calculation between miRNA $u$ and $v$ is taken as an example to demonstrate the procedure, which consisted of three steps: obtaining all the known diseases associated with miRNA $u$ and $v$, which are defined as variable $D(u)$ and $D(v)$, respectively; calculating the similarity between each disease in one disease groups and the other disease groups; calculating the similarity between two disease groups as the functional similarity between miRNA $u$ and $v$. In the second step, taking the similarity calculation between $D(v)$ and disease $D1$ in the groups of $D(u)$ as an example, similarity was defined as follows:

$$S1(D1, D(v)) = \max_{D \in D(v)} (S1(D1, D))$$

(2)

In the third step, the functional similarity between miRNA $u$ and $v$ was defined

$$S2(u, v) = \frac{\sum_{D \in D(u)} S1(D, D(v)) + \sum_{D \in D(v)} S1(D, D(u))}{|D(u)| + |D(v)|}$$

(3)

where $S2$ is the miRNA functional similarity matrix and the entity $S2(i,j)$ in row $i$ column $j$ is the functional similarity between miRNA $i$ and $j$.

For the lncRNA functional similarity calculation, similar method as miRNA functional similarity calculation was adopted. Here, lncRNA $i$ and $j$ is take as an example. Firstly, all the miRNAs interacting with these two lncRNA as miRNA groups are defined as $M(i)$ and $M(j)$, respectively. Then, the similarity between miRNA group $M(j)$ and miRNA $M1$ in the miRNA group $M(i)$ was defined as follows:

$$S2(M1, M(j)) = \max_{M \in M(j)} (S2(M1, M))$$

(4)

Finally, the similarity between two miRNA groups was calculated and regarded as the functional similarity between corresponding two lncRNAs.

$$FS(i, j) = \frac{\sum_{M \in M(j)} S2(M, M(j)) + \sum_{M \in M(i)} S2(M, M(i))}{|M(i)| + |M(j)|}$$

(5)

Figure 3. Flowchart of LFSCM, demonstrating the basic ideas of calculating lncRNA functional similarity based on disease semantic similarity, disease-miRNA associations, and lncRNA-miRNA interactions. Firstly, disease semantic similarity among all the diseases investigated in this paper was calculated based on their disease DAGs. Then, disease set associated with each miRNA was identified and the similarity among these disease sets was calculated and considered to be miRNA functional similarity. Finally, lncRNA functional similarity was calculated based on miRNA functional similarity and lncRNA-miRNA interactions.
where FS is the lncRNA functional similarity matrix and the entity $FS(i,j)$ in row $i$ column $j$ is the functional similarity between lncRNA $i$ and $j$.

References
1. Crick, F., Barnett, L., Brenner, S. & Watts-Tobin, R. General Nature of the Genetic Code for Proteins. Nature 192, 1227–1232 (1961).
2. Yanofsky, C. Establishing the triplet nature of the genetic code. Cell 128, 815–818 (2007).
3. Bertone, P. et al. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242–2246 (2004).
4. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007).
5. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38, 626–635 (2006).
6. Claverie, J. M. Fewer genes, more noncoding RNA. Science 309, 1529–1530 (2005).
7. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–1848 (2008).
8. Kaplanov, P. et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1484–1488 (2007).
9. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
10. Kaplanov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8, 413–423 (2007).
11. Esteller, M. Non-coding RNAs in human disease. Nat Rev Genet 12, 861–874 (2011).
12. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. Nat Rev Genet 10, 155–159 (2009).
13. Guttmann, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154, 240–251 (2013).
14. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927 (2011).
15. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760–1774 (2012).
16. Guttmann, M. et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lncRNAs. Nat Biotechnol 28, 503–510 (2010).
17. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehtler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105, 716–721 (2008).
18. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. Cell 136, 629–641 (2009).
19. Borsani, G. et al. Characterization of a murine gene expressed from the inactive X chromosome. Nature 351, 325–329 (1991).
20. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. Mol. Cell. Biol. 10, 28–36 (1990).
21. Brockdorff, N. et al. 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 71, 515–526 (1992).
22. Khalil, A. M. et al. Many human long intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106, 11667–11672 (2009).
23. Guttmann, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227 (2009).
24. Rinn, J. L. et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311–1323 (2007).
25. Wang, K. C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124 (2011).
26. Wapinski, O. & Chang, H. Y. Long noncoding RNAs and human disease. Trends Cell Biol 21, 354–361 (2011).
27. Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 23, 1494–1504 (2009).
28. Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E. & Mattick, J. S. IncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 39, D146–D151 (2011).
29. Chen, G. et al. LncRNA Disease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 41, D983–D986 (2013).
30. Mohanty, V., Gökmens-Polar, Y., Badve, S. & Janga, S. C. Role of IncRNAs in health and disease—size and shape matter. Brief Funct Genomics 14, 115–129 (2015).
31. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. Mol. Cell 43, 904–914 (2011).
32. Congrains, A. et al. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. Atherosclerosis 220, 449–455 (2012).
33. Johnson, R. Long non-coding RNAs in Huntington’s disease neurodegeneration. Neurobiol Dis 46, 245–254 (2012).
34. Pasman, E., Sabbagh, A., Vidaud, M. & Büche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J 25, 444–448 (2011).
35. Zhang, Q., Chen, C.-Y., Yadavalli, V. S. & Jeang, K.-T. NEAT1 Long Noncoding RNA and Paraspeckle Bodies Modulate HIV-1 Posttranscriptional Expression. MolBio 4, e00596–e00512 (2013).
36. Barysye-Lovejoy, D. et al. The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. Cancer Res 66, 5330–5337 (2006).
37. Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076 (2010).
38. Calin, G. A. et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 12, 215–229 (2007).
39. Panzitt, K. et al. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. Gastroenterology 132, 330–342 (2007).
40. de Kok, J. B. et al. DD3PCA3, a very sensitive and specific marker to detect prostate tumors. Cancer Res 62, 2695–2698 (2002).
41. Szell, M., Bata-Csörgő, Z. & Kemény, L. The enigmatic world of mRNA-like ncRNAs: their role in human evolution and in human diseases. Semin Cancer Biol 18, 141–148 (2008).
42. Zhang, X. et al. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J Clin Endocrinol Metab 89, 5119–5126 (2003).
43. Ji, P. et al. MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031–8041 (2003).
Inferring the human microRNA functional similarity and functional network based on the information of miRNA.

How to cite this article: Chen, X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci. Rep. 5, 13186; doi: 10.1038/srep13186 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/