A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases

Journal:	*Molecular BioSystems*
Manuscript ID:	MB-ART-06-2014-000339.R1
Article Type:	Paper
Date Submitted by the Author:	15-Sep-2014

Complete List of Authors:	zhou, meng; Jilin University, College of Life Science han, lu; Jilin University, College of Life Science zhang, jiahui; Harbin medical university, College of Bioinformatics Science and Technology hao, dapeng; Harbin medical university, College of Bioinformatics Science and Technology cai, yuanpei; Harbin Engineering University, wang, zhenzhen; Harbin medical university, College of Bioinformatics Science and Technology zhou, hui; Jilin University, College of Life Science sun, jie; Harbin medical university, College of Bioinformatics Science and Technology

A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases

Meng Zhou[a,b,‡], Lu Han[a,‡], Jiahui Zhang[b], Dapeng Hao[b,*], Yuanpei Cai[c], Zhenzhen Wang[b], Hui Zhou[a,*], Jie Sun[b,*]

[*]Corresponding author：

Jie Sun, suncarajie@hotmail.com

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

Hui Zhou, zhouhui@jlu.edu.cn

College of Life Science, Jilin University, Changchun 130012, PR China

Dapeng Hao, haodapeng@ems.hrbmu.edu.cn

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

[‡] These authors contributed equally to this work

Abstract

The complex traits of an organism are associated with complex interplay between genetic factors (GFs) and environmental factors (EFs). However, compared with protein-coding genes and microRNAs, there is a paucity of computational methods and bioinformatic resource platform for understanding the associations between lncRNA and EF. In this study, we developed a novel computational method to identify potential associations between lncRNA and EF, and released LncEnvironmentDB, a user-friendly web-based database aiming to provide a comprehensive resource platform for lncRNA and EF. Topological analysis of EF-related networks revealed the small world, scale-free and modularity structure. We also found that lncRNA and EF significantly enriched interacting miRNAs are functionally more related by analyzing their related diseases, implying that the predicted lncRNA signature of EF can reflect the functional characteristics to some degree. Finally, we developed a random walk with restart-based computational model (RWREFD) to predict potential disease-related EFs by integrating lncRNA-EF associations and EF-disease associations. The performance of RWREFD was evaluated by experimentally verified EF-disease associations based on leave-one-out cross-validation and achieved an AUC value of 0.71, which is higher than randomization test, indicating that RWREFD method has a reliable and high accuracy of prediction. To the best of our knowledge, LncEnvironmentDB is the first attempt to predict and house the experimental and predicted associations between lncRNA and EF. LncEnvironmentDB is freely available on the web at http://bioinfo.hrbmu.edu.cn/lncefdb/.

Introduction

It is well known that complex traits of an organism are associated with genetic factors (GFs) and environmental factors (EFs) as well as the interplay between them. The majority of human complex diseases, such as cancer, heart disease and diabetes, are caused by the complex interplay between GFs and EFs [1-3]. During the past years, a large number of publications have documented plenty of important biological mechanisms and interaction patterns between protein-coding gene/microRNA (miRNA) and EF, and several databases have been released to provide comprehensive collection and available resources for associations between protein-coding gene/miRNA and EF, such as CTD and miREnvironment [4-7]. These databases greatly facilitate further research about the relationship between GF and EF.

Long non-coding RNAs (lncRNAs), mRNA-like transcripts, are a newly discovered class of non-coding RNAs (ncRNAs) and have been identified by large-scale transcriptome analysis and annotation in animals and plants [8-12]. lncRNAs are most commonly defined as > 200 RNA transcripts that lack apparent protein coding capacity [13, 14]. lncRNAs are emerging as important regulators involved in a wide variety of biological progress, including the regulation at the transcriptional and posttranscriptional level, chromatin remodeling and protein transport and trafficking [15, 16]. Moreover, accumulating reports of aberrant lncRNA expression in many complex diseases have suggested that lncRNA are closely associated with human complex diseases [17, 18]. A recent study has reported that lncRNA *IGS*, induced by environmental signals, can function as a molecular switch to regulate the structure and function of nucleolus in mammals [19]. Growing evidence has also demonstrated that the expression patterns of some specific lncRNA can be altered as a response to EFs such as cisplatin, cycloheximide, mercury (II) oxide and doxorubicin [20, 21], implicating the interplay between lncRNA and EF. However, compared with protein-coding gene or miRNA, there are few reports or studies on the association between lncRNA and EF, and there is a paucity of databases linking lncRNA and EF. In this paper, we predicted the putative associations using computational method and released a public

database, LncEnvironmentDB, for housing experimental and predicted associations between lncRNA and EF.

Materials and methods

Data sources

We downloaded lncRNA-miRNA interaction dataset from starBase v2.0 database, which provides the most comprehensive lncRNA-miRNA interaction experimentally supported by large-scale CLIP-Seq data [22]. After getting rid of duplicate associations, 11838 experimentally confirmed lncRNA-miRNA interactions were obtained, including 1114 lncRNAs and 321 miRNAs.

The interactions between EF-miRNA were retrieved from miREnvironment database, which contains a comprehensive collection and curation of experimentally supported EF-miRNA associations [7]. After removing the redundant records, 1815 experimentally confirmed EF-miRNA interactions were obtained, including 268 EFs and 394 miRNAs.

The lncRNA-disease association data was downloaded from LncRNADisease database [23], and EF-disease association data was obtained from miREnvironment database [7].

Prediction of associations between lncRNA and EF

To predict the associations between lncRNA and EF, a computational workflow was developed and shown in Fig.1. First, experimentally supported lncRNA-miRNA and EF-miRNA interactions were combined. Next, the hypergeometric test was executed for each lncRNA-EF pair separately to measure whether the lncRNA and EF significantly shared some miRNA partners which can interact with both of them. The test calculates the *P*-value by using the following function:

$$P=1-\sum_{i=0}^{x-1}\frac{\binom{L}{i}\binom{N-L}{M-i}}{\binom{N}{M}}$$

Where N is the total number of miRNA interacting with lncRNA or EF, M is the number of miRNA which interact with a given lncRNA. L is the number of miRNA

interacting with a given EF and x is the number of miRNA that interact with both of them. All the *P*-values were subject to false discovery rate (FDR) correction [24]. Only the lncRNA-EF pairs with a small *P*-value (FDR < 0.05) were chosen as the predicted associations between lncRNA and EF.

Computational model to predict disease-related EFs

Here, we proposed a novel computational frame, random walk with restart-based method, to infer potential disease-related EFs based on EF-EF association network (EEAN). Random walk method has been successfully applied to identify disease-related genetic factors, such as protein-coding gene, miRNA and lncRNA [25-27]. In this model, for a given disease d, all the EFs with known experimentally validated associations with disease d were taken as seed nodes. Other non-seed EFs in EEAN were considered as candidate EFs. The initial probability p_0 of each seed EF was set as $1/n$ (n is the number of seed EFs), while the initial probability of all non-seed EFs were set as zero. The probability of restarting in random walk at a given time step was denoted by r ($0 < r < 1$). The random walk-based predicting scores can be obtained iteratively as follows:

$$p_{t+1} = rp_0 + (1-r)M_{EFNet}\,p_t$$

Where p_t represents a vector in which the *i*-th element holds the probability of finding the random walker at node *i* at step t. M_{EFNet} is the EF-EF association matrix.

For each candidate EF, the resulting random walk-based predicting score was defined as a stable probability p_∞ which was obtained by performing the iteration until the difference between p_t and p_{t+1} was stable (less than a cutoff measured by L_1 norm). According to previous studies[25, 28], the cutoff was chosen as 10^{-10}. Then

candidate disease-related EF can be prioritized according to p_∞. The high-ranking EFs were expected to have a high probability to be related to the given disease d.

Results and Discussion

Prediction and analysis of associations between lncRNA and EF

In this study, we first focused on the associations between lncRNA and EF. According to the ceRNA hypothesis, lncRNA can function as endogenous decoys for miRNA affecting the interactions between miRNA and their targets [29]. It is very natural to infer that the lncRNA have greater possibilities of associations with this EF if a lncRNA and an EF share significant miRNA partners. Based on above assumption, we proposed a computational method and identified approximately 5649 lncRNA and EF pairs using the common miRNA partners of any pair of lncRNA and EF based on ceRNA hypothesis. Then we constructed two EF-related interaction networks, miRNA-EF interaction network (MEIN) (Supplementary material 1) and lncRNA-EF interaction network (LEIN) (Supplementary material 2), to explore the relationship between EF and other biological molecules in EF networks. MEIN was represented as a bipartite ME graph $ME(M, E, A)$, where $A = \{a_{ij} : m_i \in M, e_j \in E\}$ (Figure 2A). In this bipartite ME graph, one node set corresponded to the miRNA set, the other node set corresponded to the EF set and the edge set corresponded to the association between them. An edge between a miRNA node and an EF node meant that the EF is known to be associated with the miRNA. The MEIN was laid out using Cytoscape 2.8.3 and the global properties were analyzed and shown in Table 1. There are 1815 miRNA-EF interaction associations between 394 miRNAs and 268 EFs. We found that 66.42% of the EFs were associated with at least two miRNAs and 70.81% of the miRNAs were associated with two and more EFs. In order to obtain a global view of the MEIN, we analyzed its characteristics (Table 1). As shown in Figure 2A and Table 1, MEIN revealed short characteristic path length, few connected components, low diameter and density, suggesting a small world and modular architecture like other biological networks. The degree distribution of miRNA and EF in MEIN was investigated and evaluated using four models (Exponential, Power law, Lognormal

and Poisson) (Figure 2B and Table 2). The degree of the miRNA (or EF) node in MEIN is the number of EF (or miRNA) associated with a given miRNA (or EF). On average, each miRNA interacted with ca. 4.6 EFs and each EF was involved with ca. 6.8 miRNAs, implying the complex interactions between miRNA and EF in determining phenotypes and diseases. The results of model fitting demonstrated that the exponential and power law models fit best to the degree distribution of miRNA and EF in MEIN, whereas the poisson model fit worst.

The predicted lncRNA-EF associations were also modeled as a bipartite LE network $LE(L,E,A)$, where $A=\{a_{ij}:l_i\in L,e_j\in E\}$ (Figure 2C). In this bipartite LE graph, there are two distinct sets of vertices corresponding to either lncRNA or EF. Vertices l_i and e_j are linked by an edge in the LEIN if lncRNA i was predicted to be associated with EF j. In total, we observed 5649 lncRNA-EF associations between 808 lncRNAs and 206 EFs in the LEIN. The global properties and degree distribution of the LEIN were analyzed and shown in Table 1, Table 2 and Figure 2D. Notably, approximately 85.4% of lncRNAs were associated with two or more of EFs, and 93.3% of EFs were involved with at least two lncRNAs. Network analysis revealed that LEIN displayed the common characteristics of biological network such as scale-free small word and power-law distribution, suggesting that inferred lncRNA-EF network was not a random network, where the degree followed poisson distribution, and have biological significance.

Database construction of associations between lncRNA and EF

Based on the above data, we constructed the LncEnvironmentDB, a web-based database which developed by JavaSE 6.0 technology and deployed on tomcat 6.0 web server. The database used Hsqldb (version 2.3) as a data storage engine. The application was built upon JSP and Spring MVC technology, which use Spring Framework as a middle business logic layer. The HTML-based web interface has been tested on Google Chrome 6, and Safari 4. LncEnvironmentDB is available at http://bioinfo.hrbmu.edu.cn/lncefdb/, and has user-friendly interface and contains pages for browsing, searching, submitting and downloading.

Analysis of functional relatedness between lncRNA and EF in LEIN

We next investigated whether the inferred lncRNA-EF associations have functional implication and relationships between lncRNA and EF. However, there are few available functional annotation data for lncRNA and EF. Many studies have observed that functionally similar genetic factors (protein-coding genes, miRNAs or lncRNAs) tend to be involved with phenotypically similar diseases [30-32]. Therefore, we investigated the functional relatedness between lncRNA and EF in LEIN by examining whether lncRNA and EF pairs in which they have predicted association in LEIN (LE pairs) were associated with more similar diseases than lncRNA and EF pairs in which they have no predicted association in LEIN (NLE pairs), and whether two lncRNAs associated with common EFs (EL pairs) tend to participate in phenotypically more similar diseases than non-EF-associated lncRNA pairs (NEL pairs). Similarity scores between lncRNA (or EFs)-associated diseases were measured based on the structure of directed acyclic graph (DAG) in Disease Ontology using DOSIM package [33]. The results indicated that there was a significant difference between LE pairs and NLE pairs for disease similarity (p-value = 0.017, Wilcoxon rank sum test) (Fig. 3A). The average disease similarity scores between lncRNA and EF in LE pairs (average 0.221) were significantly higher than those between lncRNA and EF in NLE pairs (average 0.186). In a similar manner, the disease similarity score values between lncRNAs associated with common EFs were significantly higher than those for lncRNAs in NEL pairs (average 0.390 vs. 0.170, p-value < 0.001, Wilcoxon rank sum test) (Figure 3B). Taken together, these results suggested that lncRNAs and EFs are functionally more related in LEIN than those not in LEIN.

Predicting potential EF-disease association based on lncRNA-associated EF-EF network

Previous study has demonstrated that the miRNAs interacting with EF can provide functional information for this EF [34]. So we want to test whether the functional relationship between EFs can be evaluated through their lncRNA signatures. To address this question, we computed the similarity scores between EF-associated diseases, and found that EFs sharing significantly enriched interacting

lncRNAs (LEE pairs) tend to be associated with phenotypically more similar diseases, as opposed to non lncRNA-associated EF pairs (NLEE pairs) (average 0.309 vs. 0.264, p-value < 0.001, Wilcoxon rank sum test) (Figure 3C). This result implied that EFs with similar functions tend to interact with more common lncRNAs in the context of disease. Based on above observation, we constructed an EF-EF functional association network (EEAN) based on their lncRNA signatures (Supplementary material 3). Two EFs were connected in EEAN if they share significant overlap of interacting lncRNAs using a hypergeometric test. As shown in Table 1, EEAN revealed scale-free small word and modularity structure characteristics.

Next, we proposed a random walk with restart-based computational frame (RWREFD) to predict novel EF-disease associations by using an EF-EF association network. In order to evaluate the accuracy and reliability of RWREFD to infer potential EF-disease associations, we implemented leave-one-out cross validation on known experimentally verified EF-disease associations, which is illustrated in Figure 4. For one disease d, one known EF-disease association was held out as the testing case for each cross validation and other known experimentally verified EFs associated with this given disease were considered as training set and taken as seed nodes of RWREFD to prioritize candidate disease-related EFs. The candidate disease-related EF set was composed of held-out EF and all other EFs without known associations with this given disease. Therefore, for each EF-disease association, we obtained a ranking list. If the rank of held-out EF exceeded the given threshold, this method was considered as a successful prediction for this EF-disease association.

RWR algorithm has been reported to be robust to the selection of restart probability r [25, 26]. So, we chose $r = 0.7$ as a weighted choice in leave-one-out cross validation analysis. Then we calculated the true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity) by varying the ranking cutoffs of successful prediction. Sensitivity value refers to the percentage of held-out EF ranked above a given threshold, namely the ratio of the successfully predicted known experimentally verified EF-disease associations to the total known experimentally verified EF-disease associations, and specificity value refers to the percentage of non-held-out EFs ranked

below this threshold. A receiver operating characteristic (ROC) curve and the area under the curve (AUC) were employed as a standard measure to evaluate the performance of RWREFD. The ROC curve can be obtained by plotting true positive rate versus false positive rate at different thresholds. As shown in Figure 5, RWREFD achieved an AUC of 0.71, suggesting that RWREFD method can efficiently recover known EF-disease association in the candidate disease-related EF set.

To further evaluate whether the performance of RWREFD in leave-one-out cross validation analysis was generated by chance and cannot represent biological significance, a randomization test was carried out. In each validation, the seed nodes of RWREFD were chosen randomly from candidate EF set for each disease, and the AUC value was recalculated as above. The result of randomization test showed that the AUC value under random circumstance was 0.56, which is approximate to the uninformative random AUC and is much lower than that in real situations. The result of randomization test demonstrated that the prioritization list obtained by RWREFD method has biological significance.

Conclusion

The complex interactions between genetic factors and environmental factors play important roles in the formation of various phenotypes. The identification and available public resources of associations between genetic factors and environmental factors is crucial not only for understanding the complex mechanism of multifactorial diseases, but also for promoting the identification of novel relationships between EFs and human diseases. In this study, we first proposed a novel method to identify potential associations between lncRNA and EF by considering common interacting miRNA partners based on ‘ceRNA hypothesis’, and released a user-friendly web-based database, LncEnvironmentDB. To the best of our knowledge, LncEnvironmentDB is the first attempt to predict and house the experimental and predicted associations between lncRNA and EF. We will continuously maintain and update the database. With the increasing studies on lncRNA and EF in the future, we will extend this database by incorporating more experimentally supported interactions of lncRNA and EF using literature mining. Then we modeled and analyzed the

associations between lncRNA and EF and their relationship to human disease. We found that LEIN displays the common characteristics of biological networks such as small world, scale-free and power-law distribution, suggesting that inferred lncRNA-EF network is not random network and have strong biological significance. Further studies revealed that lncRNA and EF having interaction association (LE pairs) are associated with more similar diseases than that between lncRNA and EF having no interaction association, implying that the lncRNA signature of EFs can reflect the functional characteristics to some degree. Based on this observation, we constructed an EF-EF association network using lncRNA signatures of EFs, and found EFs with similar functions tend to interact with more common lncRNAs in the context of disease, implying the potential ways to identify disease-related EFs using lncRNA-EF associations. Finally, we tried to introduce a random walk with restart-based computational model to predict potential disease-related EFs by integrating lncRNA-EF associations and EF-disease associations. The result of performance evaluation showed that RWREFD method has a reliable and high accuracy of prediction. Taken together, our study and LncEnvironmentDB will become a useful bioinformatic resource for the analysis of the relationships of lncRNA and EF, and will facilitate further research for understanding the contribution of the environment-lncRNA interaction to environmental human disease.

Competing interests

We have no competing interests

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61403111 and No. 31300185), Natural Science Foundation of Heilongjiang Province of China (Grant No. QC2013C019) and China Postdoctoral Science Foundation (Grant No. 2014M551268）

Notes and references

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

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

[c] Harbin Engineering University, Harbin 150001, PR China

† **Electronic Supplementary Information (ESI) available.**

Authors' contributions. JS, HZ and DPH conceived and designed the experiments. MZ, LH, CJZ and DPH developed the prediction method, implemented the experiments and analyzed the result. YPC developed online database. ZZW and JS analyzed the result. JS and MZ wrote the paper. All authors read and approved the final manuscript

References

1. A. Bhatnagar, *Circ Res*, 2006, 99, 692-705.
2. P. Boffetta and F. Nyberg, *Br Med Bull*, 2003, 68, 71-94.
3. C. Ling and L. Groop, *Diabetes*, 2009, 58, 2718-2725.
4. A. P. Davis, T. C. Wiegers, R. J. Johnson, J. M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, C. G. Murphy and C. J. Mattingly, *PLoS One*, 2013, 8, e58201.
5. Y. C. Lee, C. Q. Lai, J. M. Ordovas and L. D. Parnell, *J Data Mining Genomics Proteomics*, 2011, 2.
6. S. W. Turner, J. G. Ayres, T. V. Macfarlane, A. Mehta, G. Mehta, C. N. Palmer, S. Cunningham, T. Adams, K. Aniruddhan, C. Bell, D. Corrigan, J. Cunningham, A. Duncan, G. Hunt, R. Leece, U. MacFadyen, J. McCormick, S. McLeish, A. Mitra, D. Miller, E. Waxman, A. Webb, S. Wojcik, S. Mukhopadhyay and D. Macgregor, *BMC Med Res Methodol*, 2010, 10, 107.
7. Q. Yang, C. Qiu, J. Yang, Q. Wu and Q. Cui, *Bioinformatics*, 2011, 27, 3329-3330.
8. M. J. Hangauer, I. W. Vaughn and M. T. McManus, *PLoS Genet*, 2013, 9, e1003569.
9. J. Liu, C. Jung, J. Xu, H. Wang, S. Deng, L. Bernad, C. Arenas-Huertero and N. H. Chua, *Plant Cell*, 2012, 24, 4333-4345.
10. 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.
11. M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn and E. S. Lander, *Nature*, 2009, 458, 223-227.
12. Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, I. Yamanaka, H. Kiyosawa, K. Yagi, Y. Tomaru, Y. Hasegawa, A. Nogami, C. Schonbach, T. Gojobori, R. Baldarelli, D. P. Hill, C. Bult, D. A. Hume, J. Quackenbush, L. M. Schriml, A. Kanapin, H. Matsuda, S. Batalov, K. W. Beisel, J. A. Blake, D. Bradt, V. Brusic, C. Chothia, L. E. Corbani, S. Cousins, E. Dalla, T. A. Dragani, C. F. Fletcher, A. Forrest, K. S. Frazer, T. Gaasterland, M. Gariboldi, C. Gissi, A. Godzik, J. Gough, S. Grimmond, S. Gustincich, N. Hirokawa, I. J. Jackson, E. D. Jarvis, A. Kanai, H. Kawaji, Y. Kawasawa, R. M. Kedzierski, B. L. King, A. Konagaya, I. V. Kurochkin, Y. Lee, B. Lenhard, P. A. Lyons, D. R. Maglott, L. Maltais, L. Marchionni, L. McKenzie, H. Miki, T. Nagashima, K. Numata, T. Okido, W. J. Pavan, G. Pertea, G. Pesole, N. Petrovsky, R. Pillai, J. U. Pontius, D. Qi, S. Ramachandran, T. Ravasi, J. C. Reed, D. J. Reed, J. Reid, B. Z. Ring, M. Ringwald, A. Sandelin, C. Schneider, C. A. Semple, M. Setou, K. Shimada, R. Sultana, Y. Takenaka, M. S. Taylor, R. D. Teasdale, M. Tomita, R. Verardo, L. Wagner, C. Wahlestedt, Y. Wang, Y. Watanabe, C. Wells, L. G. Wilming, A. Wynshaw-Boris, M. Yanagisawa, I. Yang, L. Yang, Z. Yuan, M. Zavolan, Y. Zhu, A. Zimmer, P. Carninci, N. Hayatsu, T. Hirozane-Kishikawa, H. Konno, M. Nakamura, N. Sakazume, K. Sato, T. Shiraki, K. Waki, J. Kawai, K. Aizawa, T.

Arakawa, S. Fukuda, A. Hara, W. Hashizume, K. Imotani, Y. Ishii, M. Itoh, I. Kagawa, A. Miyazaki, K. Sakai, D. Sasaki, K. Shibata, A. Shinagawa, A. Yasunishi, M. Yoshino, R. Waterston, E. S. Lander, J. Rogers, E. Birney and Y. Hayashizaki, *Nature*, 2002, 420, 563-573.
13. C. P. Ponting, P. L. Oliver and W. Reik, *Cell*, 2009, 136, 629-641.
14. J. T. Kung, D. Colognori and J. T. Lee, *Genetics*, 2013, 193, 651-669.
15. T. R. Mercer and J. S. Mattick, *Nat Struct Mol Biol*, 2013, 20, 300-307.
16. V. A. Moran, R. J. Perera and A. M. Khalil, *Nucleic Acids Res*, 2012, 40, 6391-6400.
17. E. A. Gibb, C. J. Brown and W. L. Lam, *Mol Cancer*, 2011, 10, 38.
18. J. Sana, P. Faltejskova, M. Svoboda and O. Slaby, *J Transl Med*, 2012, 10, 103.
19. M. D. Jacob, T. E. Audas, J. Uniacke, L. Trinkle-Mulcahy and S. Lee, *Mol Biol Cell*, 2013, 24, 2943-2953.
20. R. Mizutani, A. Wakamatsu, N. Tanaka, H. Yoshida, N. Tochigi, Y. Suzuki, T. Oonishi, H. Tani, K. Tano, K. Ijiri, T. Isogai and N. Akimitsu, *PLoS One*, 2013, 7, e34949.
21. H. Tani and M. Torimura, *Biochem Biophys Res Commun*, 2013, 439, 547-551.
22. J. H. Li, S. Liu, H. Zhou, L. H. Qu and J. H. Yang, *Nucleic Acids Res*, 2013.
23. 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.
24. Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B*, 1995, 289-300.
25. J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu and M. Zhou, *Mol Biosyst*, 2014, 10, 2074-2081.
26. X. Chen, M. X. Liu and G. Y. Yan, *Mol Biosyst*, 2012, 8, 2792-2798.
27. S. Kohler, S. Bauer, D. Horn and P. N. Robinson, *Am J Hum Genet*, 2008, 82, 949-958.
28. H. Chen and Z. Zhang, *The Scientific World Journal*, 2013, 2013.
29. L. Salmena, L. Poliseno, Y. Tay, L. Kats and P. P. Pandolfi, *Cell*, 2011, 146, 353-358.
30. 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.
31. T. Ideker and R. Sharan, *Genome Res*, 2008, 18, 644-652.
32. X. Chen and G. Y. Yan, *Bioinformatics*, 2013, 29, 2617-2624.
33. 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.
34. C. Qiu, G. Chen and Q. Cui, *Sci Rep*, 2012, 2, 318.

Figure Legends

Figure 1. Principle and workflow for predicting the associations between lncRNA and EF. By combining experimentally supported lncRNA-miRNA and EF-miRNA interactions, the hypergeometric test was executed for each lncRNA-EF pair separately to measure whether the lncRNA and EF significantly shared some miRNA partners which can interact with both of them. Finally, all lncRNA-EF pairs with FDR<0.05 were imported into LncEnvironmentDB database and displayed in a web page.

Figure 2. A global view of MEIN and LEIN. (A) In MEIN, black node corresponds to miRNA and orange node corresponds to a distinct EF. An edge is placed between a miRNA node and an EF node if the miRNA has been experimentally validated to be associated with this EF. (B) In LEIN, blue node corresponds to lncRNA and orange node corresponds to a distinct EF. A lncRNA and an EF will be linked by an edge if they shared significantly overlap of miRNAs. (C) Degree distribution and fitted model of nodes in MEIN. (D) Degree distribution and fitted model of nodes in LEIN.

Figure 3. A brief statistic of similarity scores between lncRNAs and EFs in LEIN. (A) The distribution of disease similarity scores based on DAG between LE pairs and NLE pairs. (B) The distribution of disease similarity scores based on DAG between EL pairs and NEL pairs. (C) The distribution of disease similarity scores based on DAG between LEE pairs and NLEE pairs.

Figure 4. The procedure of leave one out cross validation. For a given disease, one disease-related EF and all other EFs without known associations with this given disease formed the test set. The remaining disease-related EFs were taken as seed nodes of RWREFD. The test EFs were ranked by the RWREFD method.

Figure 5. Performance evaluation of the proposed method. The performance comparison between RWREFD and random situation was implemented based on leave-one-out cross-validation tests in terms of ROC curve and AUC values.

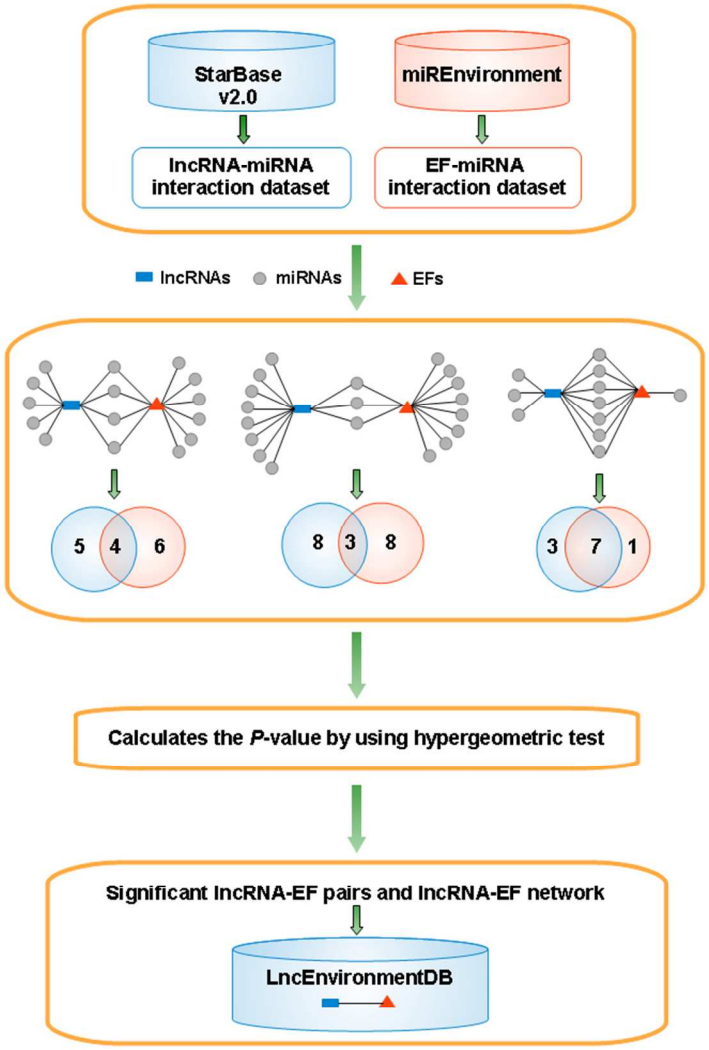
Tables

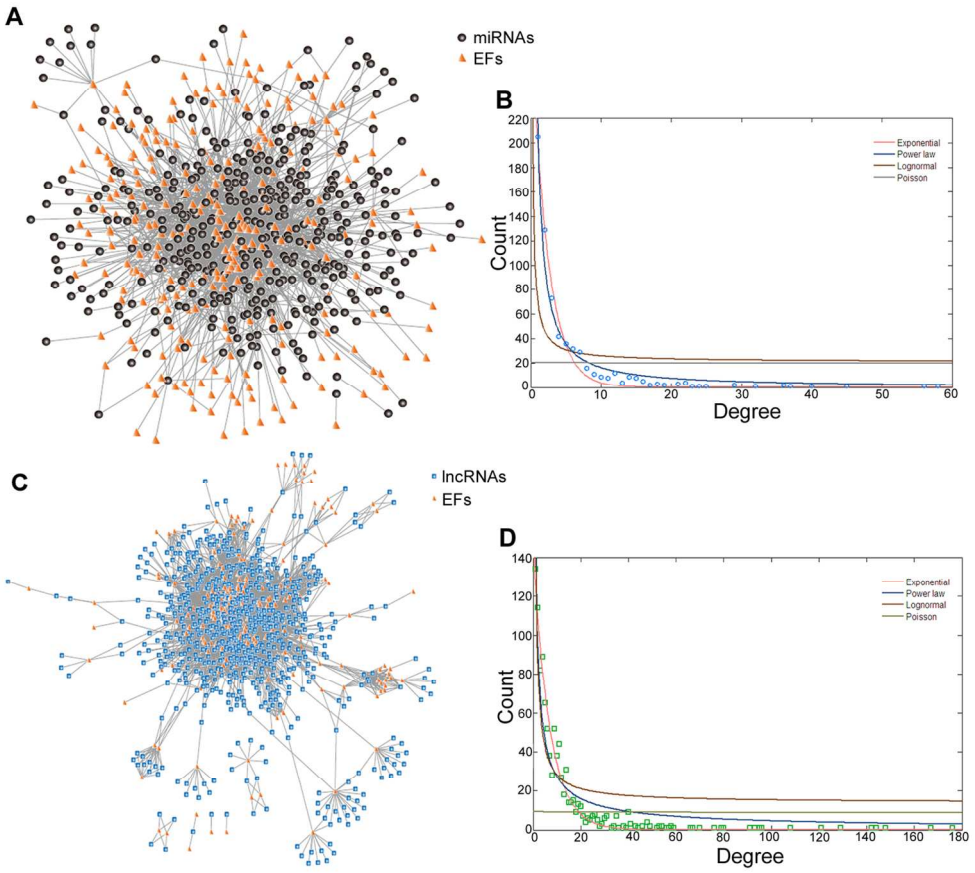
Table 1. The summary properties of three EF-related networks

Property	MEIN	LEIN	EEAN
Number of nodes	662	1014	190
Number of edges	1815	5649	1163
Cluster coefficient	0.0	0.0	0.751
Connected components	3	6	7
Diameter	9	12	9
Radius	1	1	1
Centralization	0.156	0.163	0.170
Shortest paths	431000	975352	28804
Characteristic path length	3.647	4.034	3.275
Average number of neighbors	5.483	11.142	12.253
Density	0.008	0.011	0.065
Network heterogeneity	1.778	1.539	0.784

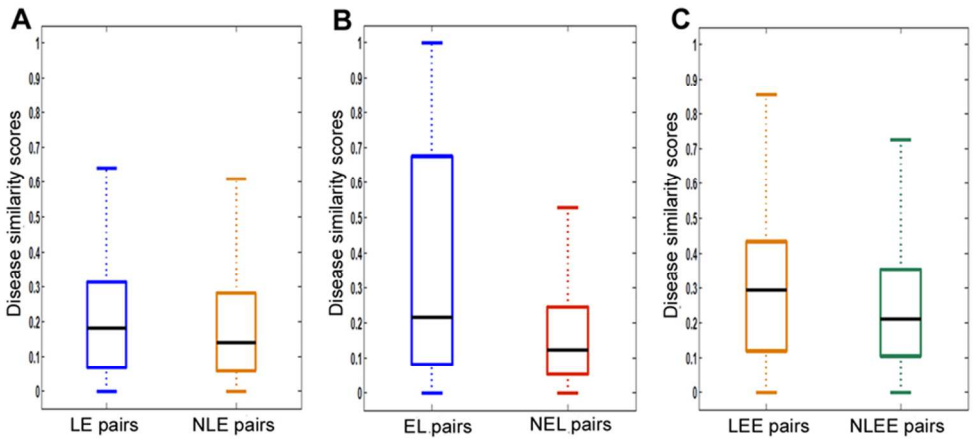
Table 2 Four types of fitted models of the degree distribution for MEIN and LEIN

Model	Parameter	MEIN	LEIN
Exponential $y = Ae^{bx}$	A	309	144.9
	b	-0.435	-0.145
	R^2	0.981	0.964
Power law $y = a \cdot x^b$	a	216.4	161.4
	b	-1.118	-0.773
	R^2	0.970	0.878
Lognormal $y = y_0 + \frac{A}{\sqrt{2\pi}\sigma x} e^{-\frac{\ln(x)-\mu}{2\sigma^2}}$	y_0	21.07	2.827
	μ	1.079	26.417
	σ	378.797	2.548
	A	-2.229	15.46
	R^2	0.366	0.816
Poisson $y = y_0 + \frac{\lambda^x}{x!} e^{-\lambda}$	y_0	20.21	12.65
	λ	0.853	0.945
	R^2	0.004	0.001

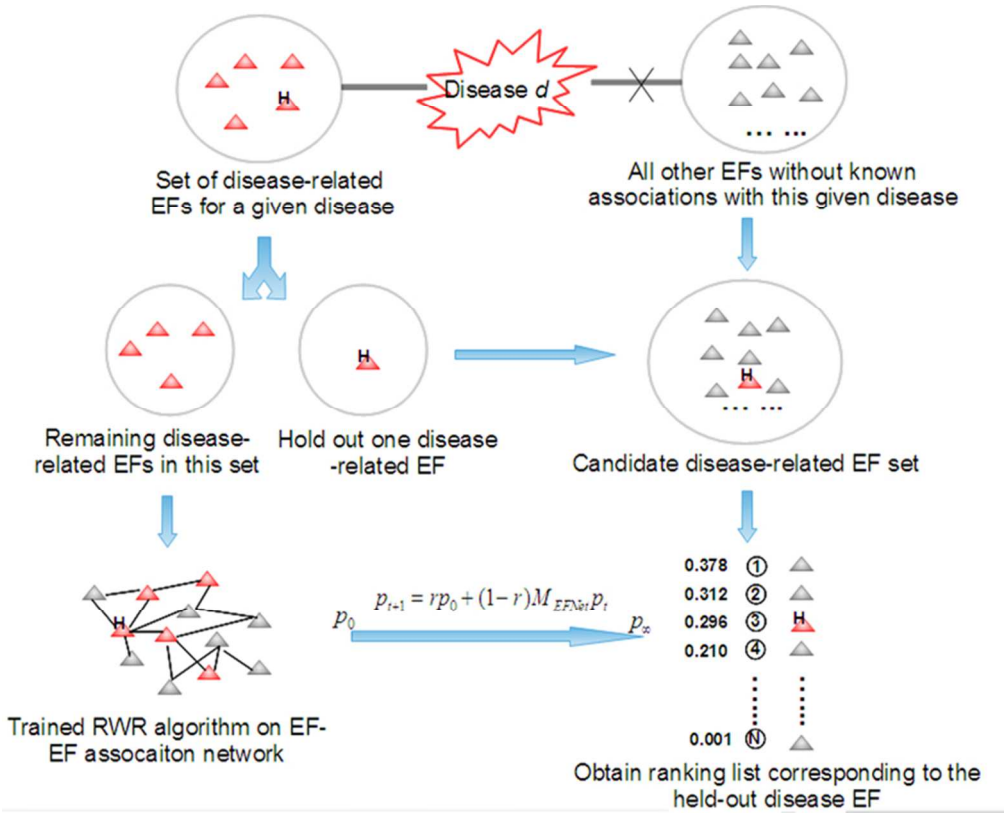

StarBase v2.0
miREnvironment
lncRNA-miRNA interaction dataset
EF-miRNA interaction dataset
lncRNAs
miRNAs
EFs
5 4 6
8 3 8
3 7 1
Calculates the P-value by using hypergeometric test
Significant lncRNA-EF pairs and lncRNA-EF network
LncEnvironmentDB



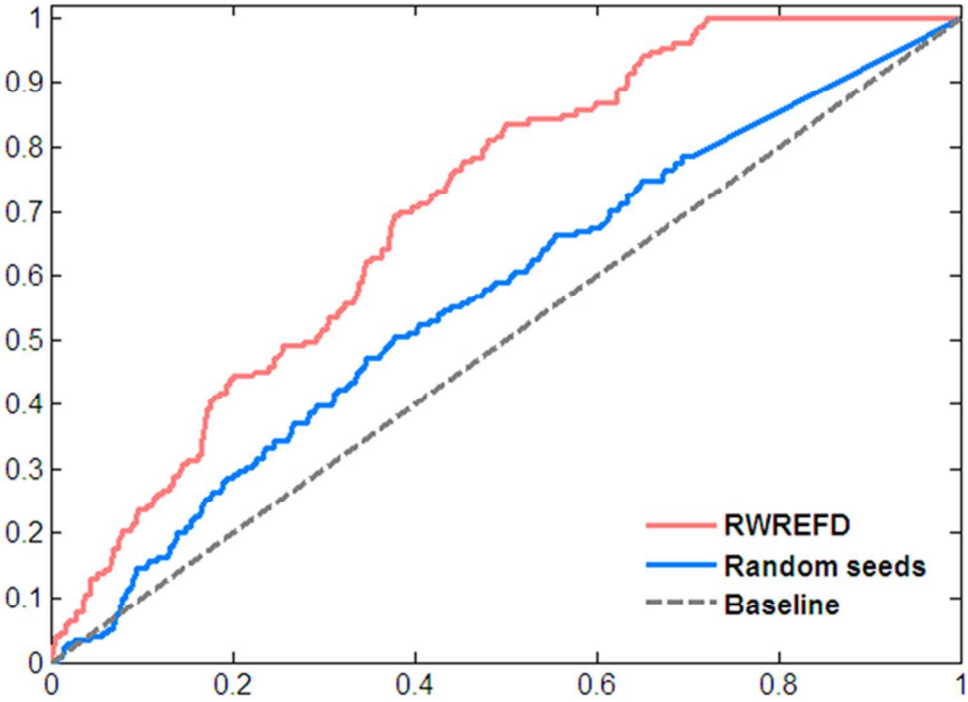
140x123mm (300 x 300 DPI)



80x37mm (300 x 300 DPI)



70x56mm (300 x 300 DPI)



60x43mm (300 x 300 DPI)