SCIENTIFIC REPORTS

OPEN

Improved low-rank matrix recovery method for predicting miRNA-disease association

Received: 5 May 2017
Accepted: 23 June 2017
Published online: 20 July 2017

Li Peng[1,2], Manman Peng[1], Bo Liao[1], Guohua Huang[3], Wei Liang[2] & Keqin Li[4]

MicroRNAs (miRNAs) performs crucial roles in various human diseases, but miRNA-related pathogenic mechanisms remain incompletely understood. Revealing the potential relationship between miRNAs and diseases is a critical problem in biomedical research. Considering limitation of existing computational approaches, we develop improved low-rank matrix recovery (ILRMR) for miRNA-disease association prediction. ILRMR is a global method that can simultaneously prioritize potential association for all diseases and does not require negative samples. ILRMR can also identify promising miRNAs for investigating diseases without any known related miRNA. By integrating miRNA-miRNA similarity information, disease-disease similarity information, and miRNA family information to matrix recovery, ILRMR performs better than other methods in cross validation and case studies.

MicroRNAs (miRNAs) comprise a set of 22-nucleotide long, noncoding RNAs, which are widespread in fauna and flora[1]. miRNAs act as crucial regulatory factors of gene expressions that result in post-transcriptional repression or degradation by complementarily binding to specific 3′ untranslated regions of their mRNA[2]. miRNAs participate in various important biological progresses, such as cell survival, apoptosis, differentiation, tumor growth, and metastasis[3]. Therefore, abnormal regulation of miRNA may lead to development and progression of various diseases, including cancer[4,5].

Substantial evidence from over 24,000 peer-reviewed reports revealed that miRNA performs a crucial role in cancer, as mentioned in the paper written by Ganju *et al.*[6], in which they summarized miRNA nanotherapeutics for cancer. Numerous miRNA-disease interactions were revealed over the years. Li *et al.*[7] and Jiang *et al.*[8] collected data from experiments that invalided miRNA-disease interactions and constructed two online databases, namely, human miRNA-disease database (HMDD) and miR2Disease, respectively. Yang *et al.*[9] set up another publicly database named Differentially Expressed MiRNAs in Human Cancers (dbDEMC).

However, known associations between miRNAs and disease are still currently limited. Revelation of potential relationship between diseases and miRNAs is a critical problem not only in uncovering molecular mechanisms of various diseases but also in providing underlying biomarkers for disease diagnosis, treatment, and drug design. Biological experimental methods for finding disease-related miRNAs are expensive and time consuming. With accumulation of available studies and emergence of large amounts of biological data about miRNA, powerful computational approach can be used to mine underlying miRNA-disease associations from these data[10]. Computational approaches sort the most plausible miRNA candidates for further analysis, hence markedly improving efficiency of experiments.

In recent years, numerous approaches were presented to predict miRNA-disease associations from machine-learning-based and network-similarity-based perspective. Jiang *et al.*[11] presented a computational model based on hypergeometric distribution to prioritize microRNAome candidates for predictive diseases to verify potential disease-associated miRNAs. Shi *et al.*[12] utilized functional relationships between miRNA targets and disease genes to mine miRNA-disease associations. Mork *et al.*[13] developed a miRPD method by combining known miRNA-protein associations to identify diseases-related miRNAs and underlying related proteins. However, the above methods strongly depended on miRNA-target associations, and their prediction performances are affected by high false-positive rates resulting from miRNA target prediction. To distinguish positive disease-related miRNAs from negative ones, Xu *et al.*[14] and Jing *et al.*[15] extracted different features and presented

[1]College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China. [2]College of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, 411201, China. [3]College of Information Engineering, Shaoyang University, Shaoyang, Hunan, 422000, China. [4]Department of Computer Science, State University of New York, New Paltz, New York, 12561, USA. Correspondence and requests for materials should be addressed to M.P. (email: pengmanman@hnu.edu.cn)

the support vector machine classification approach. Jiang *et al.*[16] developed a computational framework based on naive Bayes to mine underlying relationships from genomic data. However, negative samples of disease-related miRNAs are difficult even impossible to obtain[17]. These machine-learning-based approaches use unlabeled miRNA-disease associations as negative samples; inevitably, their accuracy of prediction is markedly influenced. Without using negative samples, Chen *et al.*[18] proposed a semi-supervised approach, named Regularized Least Squares for miRNA-Disease Association (RLSMDA), which predicted miRNA-disease association on the framework of regularized least squares.

As summarized in the paper reviewed by Zhou *et al.*[19], network similarity-based methods can be divided into two cases: local network similarity-based methods and global network methods. Xuan *et al.*[20] proposed a locally network-based approach named HMDP based on weighted k of most similar neighbors to detect promising miRNA candidates for investigation of diseases. Computation strategies of miRNA functional similarity were improved in their study by integrating information on disease phenotype similarity, miRNA family, and clusters. Chen *et al.*[21] first applied a global network-based method and advanced a method based on Random Walk with Restart (RWRMDA) for prediction of miRNA-disease associations. Li *et al.*[22] proposed a computational model named MCMDA,which predicts the associations score of each miRNA-disease pair based on matrix completion. Chen *et al.*[23] developed a novel method named miREFRWR based on the framework of random walk with restart to predict potential interactions between disease and miRNA-environmental factor. Chen *et al.*[24] advanced the miREFScan, which is a novel prediction approach based on semi-supervised classifier, to predict underlying disease-related associations between miRNAs and environmental factors(EFs). miREFScan is the first computational approach to predict correlation among miRNAs, EFs, and diseases simultaneously. These approaches performed well in cross validation. However, they cannot be used for diseases without known related miRNAs.

Chen *et al.*[25] proposed Network-Consistency-Based Interface (Net-CBI), another global-based approach, to identify underlying miRNA-disease associations. Net-CBI can isolate disease prediction, but its predictive performance is significantly poorer than that of RWRMDA. By combining multiple data sources, Liu *et al.*[26] constructed a heterogeneous network to predict disease-related miRNAs. Chen *et al.*[27] proposed a method called Restricted Boltzmann machine for multiple types of miRNA-disease association prediction to predict multi-type miRNA-disease relationships.Chen *et al.*[28] proposed a new approach named WBSMDA, which predicts miRNA-disease interactions based on the model of within and between score. By integrating experimentally validated miRNA-disease associations and various similarity information based on miRNA and disease into a heterogeneous graph, Chen *et al.*[29] proposed HGIMDA based on the framework of heterogeneous graph inference to reveal potential associations between miRNA and disease. You *et al.*[30] proposed a path-based prediction model, named PBMDA, to infer underlying miRNA-disease associations. By integrating various reliable biological datasets, PBMDA constructs a heterogeneous graph and applies depth-first search algorithm in the integrated heterogeneous network. Chen *et al.*[31] developed a new computational approach named SDMMDA based on super-disease and super miRNA to predict underlying miRNA-disease interactions. Chen *et al.*[32] proposed a new computational model of ranking-based KNN named RKNNMDA to discover potential relationship between miRNAs and diseases.

As a whole, limitations of previous approaches can be summarized as follows. First, several methods strongly rely on uncertain data, such as miRNA-target associations. Second, several machine-learning-based approaches require negative samples, which are difficult to obtain. Third, several approaches work ineffectively on isolated diseases. Finally, certain approaches, such as Net-CBI, perform poorly in predicting isolated diseases.

To overcome the above deficiency, we developed Improved Low-Rank Matrix Recovery (ILRMR) for prediction of miRNA-disease association. The algorithm of matrix recovery is widely used in recommender systems, shows good prediction performance[33,34], and is successfully applied in other fields, such as movie, commodity, and social tags[35–37]. Our method combines multiple biological data and is based on the hypothesis that similar miRNAs interact with similar diseases. Cross validation and case studies showed that ILRMR performs better compared with other methods.

The main contributions of this study are as follows:

(1) ILRMR is a semi-supervised learning approach that overcomes obstacles in obtaining negative samples in practical problems.
(2) Various biological data are integrated into matrix recovery to precisely capture new underlying associations; these data constitute similarities between miRNA-miRNA and disease-disease, miRNA family information, and known correlations between miRNA and disease.
(3) This study improves computational strategies on miRNA similarities and disease similarities.
(4) ILRMR is a global approach that can predict all disease simultaneously and have the ability to new disease without known link to miRNAs.

Results

Performance evaluation of ILRMR. In this section, we adopt two approaches to evaluate predictive performance of ILRMR: (1) Leave-one out cross validation (LOOCV) was implemented for ILRMR by using a benchmark based on known and experimentally verified miRNA-disease associations. In LOOCV of ILRMR, each known miRNA-disease interaction was excluded as test sample, and remaining interactions were used as training samples to recover predictive matrix. (2) To further prove robustness of ILRMR, we masked portions of interactions according to mask ratios in experiments and evaluated recovery and prediction ability of ILRMR. In comparison between methods, area under the receiver operating curve (ROC) (AUC) value was calculated as performance criterion of evaluation. An AUC value that closely approximates 1 indicates a significantly improved performance. ROC curve[38] plots the true positive rate or sensitivity versus false positive rate or 1-specificity at

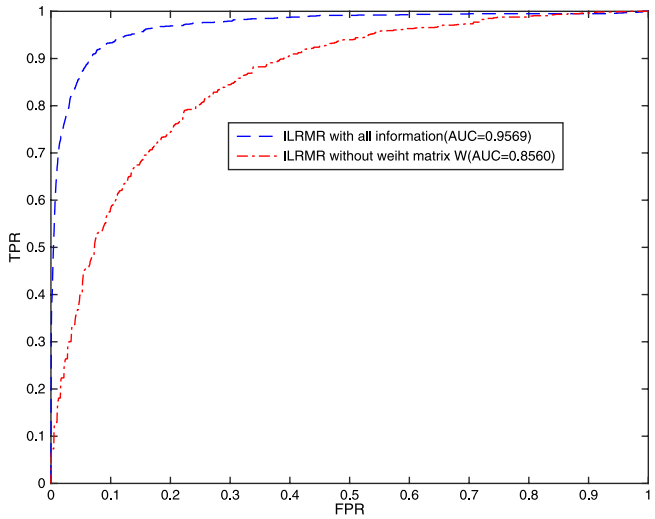

Figure 1. Performance evaluation of ILRMR in two situations based on LOOCV. (1) ILRMR with all information (ILRMR); (2) ILRMR without weight matrix *W*.

Mask Ratio	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
AUC	0.9432	0.9371	0.9280	0.9188	0.9067	0.8846	0.8769	0.8536	0.8301	0.8276	0.8014

Table 1. Comparison of prediction performance of ILRMR under different mask ratios.

different thresholds. Sensitivity refers to ratio of correctly predicted interactions to total experiment-verified miRNA-disease interactions. Specificity refers to proportion of interactions below the given threshold. However, considering a large number of unlabeled associations in the dataset, using only AUC to assess the predictive performance of the model was not insufficient. The area under precision-recall curve (AUPR) is as much as possible to reduce the affect on predictive performance caused by false positive data. Thus, using AUC and AUPR value to evaluate the performance can be more reasonable.

Based on multi-information, including miRNA-functional similarity, miRNA cosine-based similarity, miRNA family information, and disease semantic similarity, ILRMR integrates weight matrix W to recover association matrix. We evaluate predictive performance of ILRMR while considering the following aspects: (1) ILRMR with all information and (2) ILRMR without weight matrix W. Figure 1 plots the ROC curve of the two situations mentioned above.

ILRMR exhibited a commendable performance, and AUC values in the two situations reached 0.8560 and 0.9569, respectively. AUC value increased by 10.09% compared with ILRMR without weight matrix W. Evidently, weight matrix W based on miRNA (disease) similarity benefits improvement of predictive performance of ILRMR.

To further evaluate predictive performance of ILRMR, we assume that known miRNA-disease association matrix is complete and mask part of associations according to its mask ratio. The masked association matrix $X = [x_1, x_2, \ldots, x_n]$, in which only part of associations are kept, was recovered by ILRMR. We varied mask ratios from 0.1 to 0.6 for each sample and with an interval of 0.05. We implemented experiments for 20 times and calculate average performance. Table 1 summarizes performance of ILRMR under different mask ratios in terms of AUC. Results demonstrate that robust ILRMR performs reliably and efficiently mines potential miRNA-disease associations when the numbers of known associations decrease. AUC values markedly declined when mask ratio increased from 0.1 to 0.6. However, the value remains considerable.

Comparison with other methods. To our knowledge, advanced computational approaches in miRNA-disease association prediction include RWRMDA[5], Net-CBI[25], HDMP[20], RLSMDA[18] and the global network method presented by Shi *et al.*[12]. However, RWRMDA and HDMP are local approaches that cannot work on diseases without known related miRNAs. Therefore, these approaches cannot be used for comparisons in this work. Considering that the method presented by Shi *et al.* predicted miRNA-disease association by integrating miRNA-targets association, disease gene associations, and protein interaction, the datasets totally different from the ones used in ILRMR. Moreover, known miRNA-disease associations were not used with their corresponding methods. Hence, this method cannot be reasonably and fairly compared with ILRMR. ILRMR, Net-CBI and RLSMDA all use similar data sets and can predict novel miRNA-disease associations for isolated diseases. In this view, we consider their performances for comparison.

We implement LOOCV on the benchmark to assess predictive performance of ILRMR, Net-CBI, and RLSMDA. Optimal parameters of Net-CBI and RLSMDA were set as described in corresponding literature. Considering the miRNA family information and the similarity of known miRNA-disease association have not been used in the method of NetCBI and RLSMDA, the three approaches were implemented only using miRNA

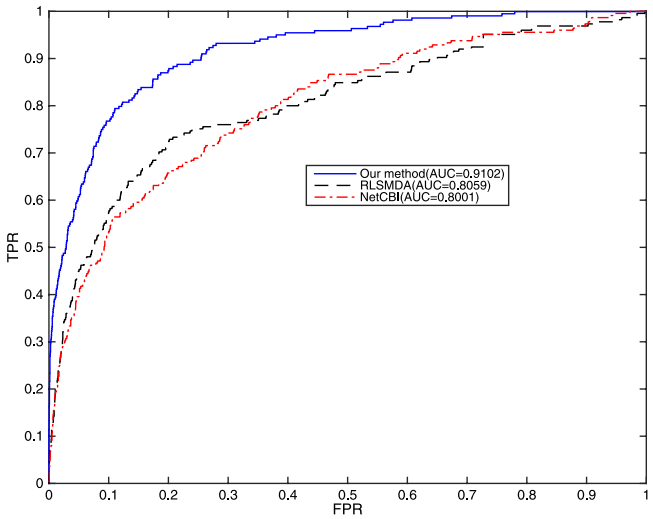

Figure 2. Comparison of methods between ILRMR, Net-CBI, and RLSMDA in terms of ROC curve and AUC on LOOCV. Without miRNA family and similarity of known miRNA-disease association network under consideration, ILRMR outperformed Net-CBI and RLSMDA in LOOCV.

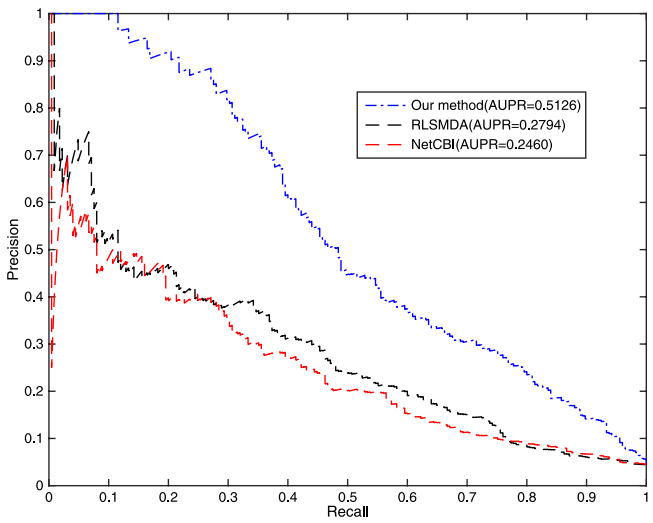

Figure 3. PR curve and AUPR values of ILRMR, RLSMDA and NetCBI.

functional similarities and disease semantic similarities in the comparisons of predicting. Figure 2 shows ROC curve and AUC value of the three methods. Without miRNA family information and cosine-based similarity of miRNA and disease under consideration, ILRMR achieved a reliable AUC of 0.9102. Net-CBI and RLSMDA achieved AUC values of 0.8001 and 0.8059, respectively. Figure 3 shows precision-recall curve and AUPR values of ILRMR, RLSMDA and NetCBI. Evidently, ILRMR outperformed Net-CBI and RLSMDA in LOOCV.

To further prove the strength of algorithms and avoid data dependence, we also implement LOOCV on predictive dataset. An AUC value of 0.9675 was obtained from ILRMR without considering miRNA family information and similarity of known miRNA-disease association network presented. Net-CBI and RLSMDA obtained AUC values of 0.9511 and 0.9560, respectively.

Case studies. To further evaluate the ability of ILRMR to predict underlying disease-related miRNA candidates, we analyze case studies on lung and breast cancers. All known miRNA-disease associations in predictive dataset were used as training set to predict potential disease-related miRNA candidates based on the ILRMR model. Predictive results were verified based on the latest version of HMDD[7]. We also check results on updated miRNA-disease relevant database, miR2Disease[8], and dbDEMC[9]. Table 2 and Supplementary Information S3 respectively list the top 50 lung cancer-related and breast cancer-related miRNAs predicted by ILRMR and confirmations of these associations.

Lung cancer is one of the most common malignant tumors with the highest morbidity and mortality and heavily threatens people's health and life. In the predictive dataset, we discover 72 miRNAs related to lung

rank	miRNA name	evidences	rank	miRNA name	evidences
1	hsa-mir-31	HMDD, dbDEMC, miR2disease	26	hsa-mir-204	miR2disease
2	hsa-mir-542	HMDD	27	hsa-mir-25	HMDD, dbDEMC
3	hsa-mir-222	HMDD, dbDEMC, miR2disease	28	hsa-mir-133a	HMDD, dbDEMC
4	hsa-mir-494	HMDD, dbDEMC	29	hsa-mir-429	dbDEMC, miR2disease
5	hsa-mir-103	HMDD, dbDEMC	30	hsa-mir-339	dbDEMC, miR2disease
6	hsa-mir-7	HMDD, miR2disease	31	hsa-mir-127	HMDD,dbDEMC
7	hsa-mir-10b	HMDD, dbDEMC	32	hsa-mir-215	dbDEMC
8	hsa-mir-93	HMDD, dbDEMC, miR2disease	33	hsa-mir-451	dbDEMC, miR2disease
9	hsa-mir-221	HMDD,dbDEMC	34	hsa-mir-302c	dbDEMC
10	hsa-mir-141	dbDEMC, miR2disease	35	hsa-mir-151	Unconfirmed
11	hsa-mir-99a	dbDEMC, miR2disease	36	hsa-mir-373	dbDEMC
12	hsa-mir-296	dbDEMC	37	hsa-mir-130a	dbDEMC, miR2disease
13	hsa-mir-23a	HMDD, dbDEMC	38	hsa-mir-200c	HMDD, dbDEMC, miR2disease
14	hsa-mir-16	dbDEMC, miR2disease	39	hsa-mir-15b	dbDEMC
15	hsa-mir-98	HMDD, dbDEMC, miR2disease	40	hsa-mir-18b	HMDD
16	hsa-mir-488	dbDEMC	41	hsa-mir-135b	HMDD, dbDEMC
17	hsa-mir-302d	dbDEMC	42	hsa-mir-372	Unconfirmed
18	hsa-mir-200b	HMDD,dbDEMC, miR2disease	43	hsa-mir-27a	HMDD, dbDEMC
19	hsa-mir-200a	HMDD, dbDEMC, miR2disease	44	hsa-mir-423	miR2disease
20	hsa-mir-185	HMDD, dbDEMC	45	hsa-mir-22	HMDD, dbDEMC, miR2disease
21	hsa-mir-320	dbDEMC	46	hsa-mir-107	HMDD, dbDEMC
22	hsa-mir-377	miR2disease	47	hsa-mir-20b	dbDEMC
23	hsa-mir-195	dbDEMC, miR2disease	48	hsa-mir-376b	Unconfirmed
24	hsa-mir-181b	HMDD, dbDEMC, miR2disease	49	hsa-mir-629	HMDD
25	hsa-mir-135a	HMDD, dbDEMC	50	hsa-mir-486	HMDD, dbDEMC

Table 2. The top 50 lung cancer-related miRNAs predicted by ILRMR and the confirmation of these associations. Forty-seven of the top 50 potential lung cancer miRNAs candidates have been confirmed based on the update HMDD, dbDEMC and mir2disease.

cancer. Underlying lung cancer-related miRNA candidates were predicted by ILRMR based on 72 known associations. Table 2 provides the top 50 lung cancer-related miRNAs predicted by ILRMR. One typical example is hsa-miR-31, which ranked first in predictive results. Recent studies[39] demonstrated close connection of miRNAs to clinicopathological parameters in clinical stages of lung cancer. Hsa-mir-31 expression significantly increases in lung cancer patients with poor survival[40]. Among the top 50 prediction list, 47 miRNAs were verified by HMDD, dbDEMC, and miR2Disease; and only hsa-mir-151, hsa-mir-372, and hsa-mir-376b were not confirmed. However, Leidinger P. *et al.*[41] demonstrated that hsa-mir-151 is upregulated in non-small cell lung carcinoma compared with non-tumorous tissues. As described in literature[42], T. Nijjar *et al.* identified that low expression level of hsa-mir-372 can be associated with recurrence case groups of stage I of non-small cell lung cancer. Evidence supported by literature further confirms reliability of ILRMR in predicting new underlying disease-related miRNA candidates.

Breast cancer is a malignant tumor that occurs in glandular epithelium of breasts and is regarded as the first major harm to women's health[43]. In the predictive dataset, 78 miRNAs are related to breast cancer. As shown in Supplementary Table 1, 48 of the top 50 breast cancer-related miRNA candidates predicted by ILRMR were confirmed by the three databases mentioned above. For example, hsa-mir-340[44], which ranked first in the predictive list, inhibits migration and development of breast cancer cell by targeting oncoprotein c-Met. hsa-mir-301a and hsa-mir-301b[45] ranked third and ninth, respectively; they are pivotal oncogenes in human breast cancer and promote nodal or distant relapses through multiple pathways. hsa-mir-7[46,47] family is regarded as tumor suppressor to migration of breast cancer. In our experiment, hsa-mir-7i, hsa-mir-7b, and hsa-mir-7g are ranked first in the top 10 list.

Applicability of ILRMR to predict diseases without any known associated miRNAs. To further verify the ability of ILRMR to predict diseases without any known associated miRNAs, we removed known verified miRNAs-disease associations on predictive diseases mentioned in the predictive dataset. This procedure ensured that prediction only considered similar information and known miRNA-disease association of other diseases. We deployed case studies for lung cancer and breast cancer, and predictive results are listed in Supplementary Table 2 and Table 3, respectively. For lung cancer, we removed 72 known miRNA-lung-cancer-related associations to predict underlying associations by ILRMR. Among the top 50 potential lung cancer miRNA candidates, 48 were based on recently updated HMDD, dbDEMC, and miR2Disease. For breast cancer, 78 known associations related to breast cancer were removed, and 47 of the top 50 predicted miRNAs were verified. The top 30 predictions for lung cancer and breast cancer were all confirmed. Therefore, ILRMR exhibits excellent performance in predicting diseases without known associated miRNAs. Topical subheadings are allowed.

miRNA	disease			
	d_1	d_2	…	d_n
m_1	0	1	…	0
m_2	1	0	…	0
…	0	0	…	1
m_m	1	0	…	0

Table 3. miRNA-disease associations.

Application of ILRMR to predict novel human miRNAs-disease associations. The reliable performance of our algorithm had been thoroughly verified on cross validation and case studies as discussed above. Here, we further demonstrated the application of ILRMR to globally predict new potential miRNA-disease associations. All the known miRNA-disease associations in the predictive dataset were used for prediction. We ranked the unknown associations according to the scores recovered by ILRMR, and manually verified the top 50 associations through three updated HMDD, miR2diseases and dbDEMC. The predictive results and confirmations of these associations are listed in Supplementary Table 4.

Discussion

Revelation of potential relationship between diseases and miRNAs is a critical problem not only in uncovering molecular mechanisms of various diseases but also in providing underlying biomarkers for disease diagnosis, treatment, and drug design. In this paper, we develop ILRMR for miRNA-disease association prediction. Compared with other state-of-the-art computational methods, ILRMR is a global method that can simultaneously prioritize potential associations of all diseases and does not require negative samples. ILRMR can also identify promising miRNAs for investigating diseases without any known related miRNA. By integrating miRNA-miRNA similarity information, disease-disease similarity information, and miRNA family information to matrix recovery, ILRMR performs better compared with other methods in cross validation and case studies.

Reliable performance of ILRMR can be majorly attributed to combination of the following algorithm factors. (1) This algorithm integrates various biological information, specifically on similarities of miRNA and disease, to matrix recovery, thereby significantly improving prediction performance. (2) The algorithm takes full advantages of unlabeled data in the miRNA-disease association matrix. (3) ILRMR solved by augmented Lagrange multipliers (ALMs) shows good convergence to obtain optimal solutions[48].

ILRMR can be a valuable computational tool for predicting miRNA-disease associations. This approach can be further applied to reveal other biological associations, such as lncRNA-disease, gene-disease, and drug-target associations. However, the proposed approach also presents several limitations. First, a more reasonable construction of weight matrix based on miRNA similarity and disease similarity will further improve prediction capabilities. Second, further work can be conducted to extend similarity measures as a regression and to make the model more efficient and general.

Wang *et al.*[49–51] discussed a cancer hallmark network framework and cancer systems biology in the genome sequencing era. It is very interesting and so instructive for our in-depth analysis and understanding of the pathogenesis of cancer. At present, we predicted only whether there is an association between miRNAs and diseases. The specific regulation mechanism has not yet been studied. Whether the miRNAs regulate more cancer hallmark genes deserves a closer look. From this perspective, more research work we may able to carry out in the future work.

Methods

Data Preparation. Data on miRNA-disease associations used in this paper were obtained from HMDD constructed by Li *et al.*[7]. Two versions (September-2009 Version and V2.0 Version) of HMDD associations were used in our study. The first version was used as predictive dataset to predict new miRNA-disease associations. The latest version was used to confirm prediction results. Two other online databases, miR2Disease and dbDEMC, which were constructed by Jiang *et al.*[8] and Yang *et al.*[9], were also used for confirmation of predicted results. To further demonstrate generalization abilities of our methods for certain situations, that is, extremely limited known and experimentally identified miRNA-disease interactions, miRNA-disease association data from ref. 11 were also used as benchmark datasets in the paper. miRNA functional similarity scores were downloaded from http://www.cuilab.cn, which is a reliable website that provides biological data to facilitate research for biologists and medical scientists. Disease semantic similarities were calculated similarly as those in other studies[52], whereas similarity score can be obtained from supplementary material in ref. 18.

Problem Description. We considered m miRNAs and n diseases and supposed that original matrix $A_{m\times n}$ represents adjacency matrix of miRNA-disease association, where $A_{ij}=1$ is the i^{th} miRNA that interacts with the j^{th} disease; otherwise, $A_{ij}=0$. As shown in Table 3, a value of 1 represents corresponding miRNA-disease association verified though biological experiments and exists in databases, including mir2Disease, HMDD, and dbDEMC. A value of 0 represents a missing value (unknown association that will be predicted). The masked association matrix $X_{m\times n}$ is obtained from the original association matrix $A_{m\times n}$, and masked part of interaction according the mask ratio demanded in the cross validation. The work we need to do is to estimate the missing value of the matrix based on the existing association and relevant information.

Model of LRMR for predicting miRNA-disease association. Low-rank matrix recovery (LRMR) is a highly effective algorithm for predicting missing values. This algorithm uses different mathematical or machine learning methods to decompose potential characteristics from an original matrix to explain and to predict missing values. Limited validated numbers are available for known miRNA-disease associations through biological experiments, whereas negative samples are difficult or impossible to obtain. Matrix $A_{m\times n}$ of miRNA-disease association is sparse and imbalanced. Furthermore, a certain degree of potential similarity exists among column (row) vectors in association matrix. Given the characteristics of association matrix mentioned above, we considered recovery of matrix by using robust principal component analysis (rPCA), which is one of the powerful models used in ILRMR.

We predicted unknown miRNA-disease associations based on the robust PCA model by using (1), which minimizes errors between known association matrix X and resuming matrix R_{mir_dd}:

$$\min_{R_{mir\_dd},E} \|R_{mir\_dd}\|_* + \lambda\|E\|_1, \quad subject\ \ to \quad X = R_{mir\_dd} + E \tag{1}$$

where $\|R_{mir\_dd}\|^*$ denotes the nuclear norm[53] of the resuming miRNA-disease association matrix R_{mir_dd}, $\|E\|_1$ reprents the $\ell_1 - norm$ of the discrepancy matrix E, weight parameter λ denotes weight sparse error term in the cost function, and $0 \le \lambda \le 1$. Optimization model can be solved using the exact ALM method from a previous study[48].

Calculating miRNA-based similarity. To improve the accuracy of association matrix recovery and the prediction effects, we combined the weight matrix W with the robust PCA model, which includes miRNA-miRNA similarity and disease-disease similarity. By Comparing with a similarity measure method in a previous study[52,54], we calculate miRNA similarity by integrating multi-information, including miRNA functional similarity, cosine-based similarity, and miRNA family information. Considering each miRNA as a vector of the frequency of the interaction with the diseases, we then computed the cosine value of the angle formed by two miRNA vectors[55]. Assuming that Sim_{mir_cos} represents the miRNA similarity matrix, we calculate cosine-based similarity by (2)

$$Sim_{mir\_cos}(i,j) = \frac{x_i x_j^T}{\|x_i\|\|x_j\|} \tag{2}$$

Then, we integrate the miRNA functional similarity (matrix Sim_{mir_fun}), information on miRNA families (matrix Sim_{FAM}) and cosine-base similarity(matrix Sim_{mir_cos}) into Eq. (3) to construct the final miRNA-miRNA similarity:

$$Sim_{mir}(i,j) = Sim_{mir\_fun}(i,j) \times (1 + Sim_{mir\_cos}(i,j)) \times (1 + FAM(i,j)) \tag{3}$$

miRNA functional similarity score calculation was based on the method proposed by Wang[52]. miRNA family information was obtained from the miRBase database[56]. When miRNAs i and j belong to the same family, value of $FAM(i,j)$ is 1, otherwise the value is 0. $Sim_{mir}(i,j)$ denotes the final similarity score between miRNA i and j. When i^{th} and j^{th} miRNAs are more similar and belong to the same family, $Sim_{mir}(i,j)$ is higher.

Calculating disease-based similarity. Similar to the calculations of miRNA cosine-based similarity, disease cosine-based similarity was computed. We assume that $X = [x_1, x_2, \ldots, x_n]$ represents the miRNA-disease association matrix. Sim_{dd_cos} represents similarity matrix between diseases according to known correlations in the miRNA-disease associaton network. We calculate the matrix by (4) based on the vector cosine-based similarity measure method:

$$Sim_{dd\_cos}(i,j) = \frac{X_i X_j^T}{\|X_i\|\|x_j\|} \tag{4}$$

where X_i denotes i^{th} row of X.

Then, we integrate the disease semantic similarity and cosine-based similarity into Eq. (5) to construct final disease-disease similarity

$$Sim_{dd}(i,j) = Sim_{dd\_phe}(i,j) \times (1 + Sim_{dd\_cos}(i,j)) \tag{5}$$

where $Sim_{dd\_phe}(i,j)$ corresponds to the semantic similarity score of diseases i and j. From the Medical Subject Heading database (a strict system for disease description and classification), diseases were described in a DAG. Disease semantic similarity can be calculated based on the assumption that two diseases sharing more parts of DAGs are more similar[18]. $Sim_{dd}(i,j)$ represents the final similarity between diseases i and j. When two diseases are more similar, score is higher.

Calculating weight matrix *W* based on miRNA(disease) similarity and prediction of novel association. To further improve the prediction accuracy, we integrate weight matrix W based on the miRNA and disease similarity mentioned above to matrix recovery algorithm. Thus, we obtain the following prediction formula (6):

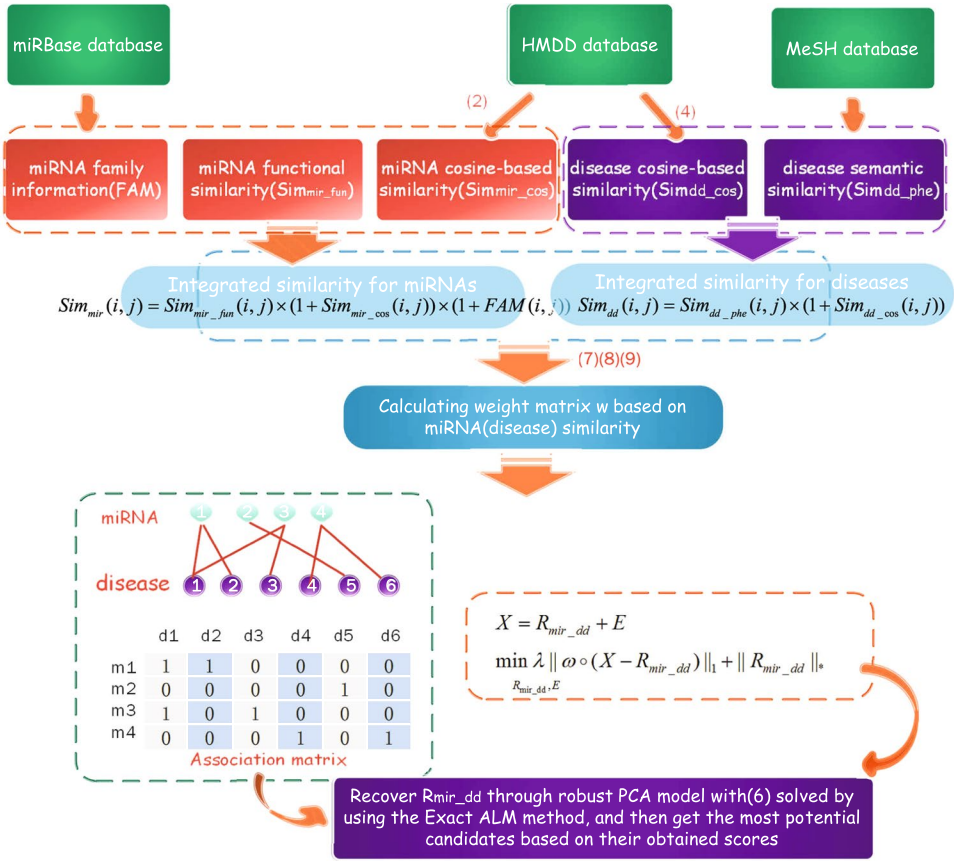

Figure 4. The overall flowchart of ILRMR.

$$X = R_{mir\_dd} + E \min_{R_{mir\_dd},E} \lambda \| W \circ (X - R_{mir\_dd}) \|_1 + \| R_{mir\_dd} \|_* \tag{6}$$

where W denotes weight matrix based on miRNA (disease) similarity. Figure 4 shows the overall flowchart of ILRMR method. The algorithm is summarized as follow.

In our method, association matrix $X_{m\times n}$ is decomposed into sum of low rank matrix R_{mir_dd} and sparse noise matrix E, and low rank matrix R_{mir_dd} is then recovered by solving the nuclear norm optimization problem. $\ell_1 - norm$ is used to suppress noise. We used Hadamard product between weight matrix W and discrepancy matrix to improve accuracy of the recovery. Considering that matrix $X_{m\times n}$ measures $m\times n$ and Hadamard product[57] is a class of matrix operation, in which operation of two matrices matrix must be of the same order, we calculate W by (7) based on calculation of similarity mentioned above with appropriate transformation:

$$W_{(i,j)} = \frac{W_{mir}(i,j) + W_{dd}(i,j)}{2} \tag{7}$$

Algorithm 1: ILRMR for miRNA-disease interaction predictio

Input: $A_{m\times n}, Sim_{mir\_fun}, Sim_{FAM}, Sim_{dd\_phe}, \lambda$
Output: R_{mir_dd}
Obtain the masked miRNA-disease association matrix $X_{m\times n}$;
Compute $Sim_{mir\_cos}(i,j)$ using (2)
Compute $Sim_{mir}(i,j)$ using (3);
Compute $Sim_{dd\_cos}(i,j)$ using (4);
Compute $Sim_{dd}(i,j)$ using (5);
Compute $W_{mir}(i,j)$and $W_{dd}(i,j)$ using (8), (9);
Compute $W(i,j)$ using (7);
Recover R_{mir_dd} through robust PCA model with (6) solved by using the Exact ALM method[48]
Sort the result in R_{mir_dd} in descending order;
Return obtained the ranking list of miRNA-disease interaction

Weight matrix W consists of two parts, namely, $W_{mir}(i, j)$ and $W_{dd}(i, j)$, which weights based on miRNA-miRNA similarity and disease-disease similarity, respectively. $W_{mir}(i, j)$ denotes the miRNA-based similarity weight value, and is calculated by (8)

$$W_{mir}(i, j) = \frac{Sim_{mir_i} \times X_j}{X_j} \tag{8}$$

where Sim_{mir_i} corresponds to i^{th} row of matrix Sim_{mir} and vector comprising the similarities between miRNA i and all other miRNAs. X_j denotes the j^{th} column of association matrix $X_{m\times n}$ and the vector consisting of the interactions between disease j and all miRNAs. $|X_j|$ represents the length of vector X_j (the norm of vector X_j). Evidently, higher value of $W_{mir}(i, j)$, indicates higher possibility that miRNA i is associated with disease j.

Similarly, $W_{dd}(i, j)$ denotes disease-based similarity weight value, and is calculated by (9)

$$W_{dd}(i, j) = \frac{X_i \times Sim_{dd_j}}{|X_i|} \tag{9}$$

where Sim_{dd_j} corresponds to j^{th} column of matrix Sim_{dd} and the vector consisting of similarities between disease j and all other diseases. X_i corresponds to the i^{th} row of matrix $X_{m\times n}$ and the vector consisting of interactions between miRNA i and all diseases. Notebly, higher value of $W_{dd}(i, j)$, indicates a higher probability that miRNA i is associated with disease j.

References

1. Ambros, V. The functions of animal micrornas. *Nature* **431**, 350–355 (2004).
2. Bartel, D. P. Micrornas: target recognition and regulatory functions. *cell* **136**, 215–233 (2009).
3. Bartel, D. P. Micrornas: genomics, biogenesis, mechanism, and function. *cell* **116**, 281–297 (2004).
4. Paul, P. *et al*. Interplay between mirnas and human diseases: A review. *Journal of Cellular Physiology* (2017).
5. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics. *CA: a cancer journal for clinicians* **65**, 5–29 (2015).
6. Ganju, A. *et al*. mirna nanotherapeutics for cancer. *Drug discovery today* **22**, 424–432 (2017).
7. Li, Y. *et al*. Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research* **42**, D1070–D1074 (2014).
8. Jiang, Q. *et al*. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research* **37**, D98–D104 (2009).
9. Yang, Z. *et al*. dbdemc: a database of differentially expressed mirnas in human cancers. *BMC genomics* **11**, S5 (2010).
10. Wen, X., Shao, L., Xue, Y. & Fang, W. A rapid learning algorithm for vehicle classification. *Information Sciences* **295**, 395–406 (2015).
11. Jiang, Q. *et al*. Prioritization of disease microrna through a human phenome-microrna network. *BMC systems biology* **4**, S2 (2010).
12. Shi, H. *et al*. Walking the interactome to identify human mirna-disease associations through the functional link between mirna targets and disease genes. *BMC systems biology* **7**, 101 (2013).
13. Mørk, S., Pletscher-Frankild, S., Caro, A. P., Gorodkin, J. & Jensen, L. J. Protein-driven inference of mirna–disease associations. *Bioinformatics* btt677 (2013).
14. Xu, J. *et al*. Prioritizing candidate disease mirnas by topological features in the mirna target–dysregulated network: Case study of prostate cancer. *Molecular cancer therapeutics* **10**, 1857–1866 (2011).
15. Jiang, Q., Wang, G., Jin, S., Li, Y. & Wang, Y. Predicting human microrna-disease associations based on support vector machine. *International journal of data mining and bioinformatics* **8**, 282–293 (2013).
16. Jiang, Q., Wang, G. & Wang, Y. An approach for prioritizing disease-related microrna based on genomic data integration. In *Biomedical Engineering and Informatics* (*BMEI*), *2010 3rd International Conference On*, vol. 6, 2270–2274 (IEEE, 2010).
17. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **31**, i221–i229 (2015).
18. Chen, X. & Yan, G.-Y. Semi-supervised learning for potential human microrna-disease associations inference. *Scientific reports* **4**, 5501 (2014).
19. Zou, Q., Li, J., Song, L., Zeng, X. & Wang, G. Similarity computation strategies in the microrna-disease network: a survey. *Briefings in functional genomics* elv024 (2015).
20. Xuan, P. *et al*. Prediction of microrna associated with human diseases based on weighted k most similar neighbors. *PloS one* **8**, e70204 (2013).
21. Chen, H. & Zhang, Z. Prediction of associations between omim diseases and microrna by random walk on omim disease similarity network. *The Scientific World Journal* **2013** (2013).
22. Li, J.-Q., Rong, Z.-H., Chen, X., Yan, G.-Y. & You, Z.-H. Mcmda: Matrix completion for mirna-disease association prediction. *Oncotarget* **8**, 21187 (2017).
23. Chen, X. mirefrwr: a novel disease-related microrna-environmental factor interactions prediction method. *Molecular BioSystems* **12**, 624–633 (2016).
24. Chen, X., Liu, M.-X., Cui, Q.-H. & Yan, G.-Y. Prediction of disease-related interactions between microrna and environmental factors based on a semi-supervised classifier. *PloS one* **7**, e43425 (2012).
25. Chen, H. & Zhang, Z. Similarity-based methods for potential human microrna-disease association prediction. *BMC medical genomics* **6**, 12 (2013).
26. Liu, Y., Zeng, X., He, Z. & Zou, Q. Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2016).
27. Chen, X. *et al*. Rbmmmda: predicting multiple types of disease-microrna associations. *Scientific reports* **5**, 13877 (2015).
28. Chen, X. *et al*. Wbsmda: within and between score for mirna-disease association prediction. *Scientific reports* **6** (2016).
29. Chen, X. *et al*. Hgimda: Heterogeneous graph inference for mirna-disease association prediction. *Oncotarget* **7**, 65257–65269 (2016).
30. You, Z.-H. *et al*. Pbmda: A novel and effective path-based computational model for mirna-disease association prediction. *PLoS computational biology* **13**, e1005455 (2017).
31. Chen, X. *et al*. A novel computational model based on super-disease and mirna for potential mirna-disease association prediction. *Molecular BioSystems* (2017).
32. Chen, X., Wu, Q.-F. & Yan, G.-Y. Rknnmda: Ranking-based knn for mirna-disease association prediction. *RNA biology* 1–11 (2017).
33. Fan, J., Wang, W. & Zhu, Z. Robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315* (2016).

34. Kueng, R., Rauhut, H. & Terstiege, U. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis* **42**, 88–116 (2017).
35. Xu, L. & Davenport, M. Dynamic matrix recovery from incomplete observations under an exact low-rank constraint. In *Advances in Neural Information Processing Systems*, 3585–3593 (2016).
36. Chen, P. & Suter, D. Recovering the missing components in a large noisy low-rank matrix: Application to sfm. *IEEE transactions on pattern analysis and machine intelligence* **26**, 1051–1063 (2004).
37. Xu, B. H. *et al*. Video restoration based on patchmatch and reweighted low-rank matrix recovery. *Multimedia Tools and Applications* **75**, 2681–2696 (2016).
38. Fawcett, T. An introduction to roc analysis. *Pattern recognition letters* **27**, 861–874 (2006).
39. Yan, H.-J., Ma, J.-Y., Wang, L. & Gu, W. Expression and significance of circulating microrna-31 in lung cancer patients. *Medical science monitor: international medical journal of experimental and clinical research* **21**, 722 (2015).
40. Le, H.-B. *et al*. Evaluation of dynamic change of serum mir-21 and mir-24 in pre-and post-operative lung carcinoma patients. *Medical oncology* **29**, 3190–3197 (2012).
41. Leidinger, P., Keller, A. & Meese, E. Micrornas–important molecules in lung cancer research. *Frontiers in genetics* **2**, 104 (2012).
42. Roa, W. *et al*. Identification of a new microrna expression profile as a potential cancer screening tool. *Clinical & Investigative Medicine* **33**, 124–132 (2010).
43. Finkelstein, M., Boulard, M. & Wilk, N. Epidemiology and etiology. *Regul Toxicol Pharmacol* **12**, 224–37 (1990).
44. Chen, C.-P. *et al*. mir-340 suppresses cell migration and invasion by targeting myo10 in breast cancer. *Oncology reports* **35**, 709–716 (2016).
45. Wu, X. *et al*. Comprehensive expression analysis of mirna in breast cancer at the mirna and isomir levels. *Gene* **557**, 195–200 (2015).
46. Ma, L., Li, G.-z., Wu, Z.-s. & Meng, G. Prognostic significance of let-7b expression in breast cancer and correlation to its target gene of bsg expression. *Medical Oncology* **31**, 773 (2014).
47. Subramanian, M. *et al*. A mutant p53/let-7i-axis-regulated gene network drives cell migration, invasion and metastasis. *Oncogene* **34**, 1094–1104 (2015).
48. Lin, Z., Chen, M. & Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055* (2010).
49. Wang, E. *et al*. Cancer systems biology in the genome sequencing era: Part 1, dissecting and modeling of tumor clones and their networks. In *Seminars in cancer biology*, vol. 23, 279–285 (Elsevier, 2013).
50. Wang, E. *et al*. Cancer systems biology in the genome sequencing era: Part 2, evolutionary dynamics of tumor clonal networks and drug resistance. In *Seminars in cancer biology*, vol. 23, 286–292 (Elsevier, 2013).
51. Wang, E. *et al*. Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. In *Seminars in cancer biology*, vol. 30, 4–12 (Elsevier, 2015).
52. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* **26**, 1644–1650 (2010).
53. Fang, E. X., Liu, H., Toh, K.-C. & Zhou, W.-X. Max-norm optimization for robust matrix recovery. *arXiv preprint arXiv:1609.07664* (2016).
54. Gu, C., Liao, B., Li, X. & Li, K. Network consistency projection for human mirna-disease associations inference. *Scientific reports* **6** (2016).
55. Su, X. & Khoshgoftaar, T. M. A survey of collaborative filtering techniques. *Advances in artificial intelligence* **2009**, 4 (2009).
56. Kozomara, A. & Griffiths-Jones, S. mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic acids research* gkq1027 (2010).
57. Bocci, C., Carlini, E. & Kileel, J. Hadamard products of linear spaces. *Journal of Algebra* **448**, 595–617 (2016).

Acknowledgements

This work is in part supported by National Natural Science Foundation of China (61672356, 61572188), by Hunan Natural Science Foundation (2017JJ2239), by Scientific Research Fund of Hunan Provincial Education Department (15B216), and by Scientific Research Project of Hunan University of Science and Technology (E51697).

Author Contributions

L.P. conceived the project, developed the prediction method, designed and implemented the experiments, analyzed the result, and wrote the paper. M.M.P. and B.L. analyzed the result, and wrote the paper. G.H.H. and W.L. implemented the experiments, and analyzed the result. K.Q.L. analyzed the result. All authors reviewed the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-06201-3

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.