Molecular Therapy
Nucleic Acids
Original Article

Prediction of Potential Disease-Associated MicroRNAs by Using Neural Networks

Xiangxiang Zeng,[1,2,6] Wen Wang,[1,6] Gaoshan Deng,[3,6] Jiaxin Bing,[1,6] and Quan Zou[4,5]

[1]Shenzhen Research Institute of Xiamen University, Xiamen University, Shenzhen 518000, Guangdong, China; [2]Department of Information Science and Technology, Xiamen University, Xiamen 361005, Fujian, China; [3]Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA; [4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610000, China; [5]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610000, China

Identifying disease-related microRNAs (miRNAs) is an essential but challenging task in bioinformatics research. Much effort has been devoted to discovering the underlying associations between miRNAs and diseases. However, most studies mainly focus on designing advanced methods to improve prediction accuracy while neglecting to investigate the link predictability of the relationships between miRNAs and diseases. In this work, we construct a heterogeneous network by integrating neighborhood information in the neural network to predict potential associations between miRNAs and diseases, which also consider the imbalance of datasets. We also employ a new computational method called a neural network model for miRNA-disease association prediction (NNMDA). This model predicts miRNA-disease associations by integrating multiple biological data resources. Comparison of our work with other algorithms reveals the reliable performance of NNMDA. Its average *AUC* score was 0.937 over 15 diseases in a 5-fold cross-validation and *AUC* of 0.8439 based on leave-one-out cross-validation. The results indicate that NNMDA could be used in evaluating the accuracy of miRNA-disease associations. Moreover, NNMDA was applied to two common human diseases in two types of case studies. In the first type, 26 out of the top 30 predicted miRNAs of lung neoplasms were confirmed by the experiments. In the second type of case study for new diseases without any known miRNAs related to it, we selected breast neoplasms as the test example by hiding the association information between the miRNAs and this disease. The results verified 50 out of the top 50 predicted breast-neoplasm-related miRNAs.

INTRODUCTION

The first microRNA (miRNA) named lin-4 was discovered 20 years ago by Victor Ambros.[1] Since then, thousands of currently annotated miRNAs have been discovered in various species of plants, animals, and viruses.[2] The expression of mRNAs is suppressed in a sequence-specific manner by miRNAs that consist of small endogenous noncoding RNAs.[3,4] Many studies indicated that miRNAs are important cell components and have vital roles in multiple stages of biological processes, such as cell growth,[5] cell development,[5] cell cycle regulation,[6] cell apoptosis,[7] stress responses,[8] and tumor invasion.[9] Furthermore, the strong associations between miRNAs and diseases have been verified by numerous biological studies.[10,11] The accumulating knowledge of disease-related miRNAs could contribute to pathological classifications, individualized diagnoses, and disease treatments.[12–14] However, exploring the underlying miRNA-disease associations still remains a challenge for biologists.[15–18] Powerful computational approaches that could effectively reveal miRNA-disease associations must be urgently developed.

In recent years, many computational prediction methods have been used to identify reliable disease-miRNA candidates for further experimental studies[19–30] and achieved excellent performance. Based on the assumptions that miRNAs that have similar functions are more likely associated with similar disease and vice versa, Jiang et al.[31] estimated the similarity between miRNAs by measuring the similarity of their target genes. The miRNA network based on targets was combined with a disease phenotype network to infer the correlation scores between miRNAs and diseases. In addition, they improved the score calculation by further integrating the similarities of miRNAs with the phenotype similarities of diseases.[32] Li et al.[33] collected miRNA targets and measured the function consistency score (FCS) between the target genes and the disease-related genes. However, this method ignored the topological structure when calculating the FCS. Xu et al.[34] focused on extracting features from the miRNA-disease network data, which were constructed under two considerations, namely, a feature set primarily related to miRNA information and disease phenotype information. HDMP[35] predicted disease-related miRNAs by weighting the most similar neighbors.

In addition, Chen et al.[36] presented the random walk with restart for miRNA-disease association (RWRMDA) model to identify potential miRNA-disease pairs by adopting random walks on the miRNA functional similarity network. Shi et al.[37] improved the RWRMDA by considering miRNA-target associations, disease-gene associations,

Received 9 January 2019; accepted 11 April 2019;
https://doi.org/10.1016/j.omtn.2019.04.010.

[6]These authors contributed equally to this work.

Correspondence: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610000, China.
E-mail: zouquan@nclab.net

and protein-protein interaction networks. Xuan et al.[21] developed the method miRNAs associated with diseases prediction (MIDP), which utilizes the features of different nodes on the basis of random walks with a restart. Afterward, an extension method, named MIDPE, was proposed by constructing a miRNA-disease bilayer network. This approach was developed because nearly all the previous methods based on random walks could not be applied without any known related miRNA.

Moreover, machine-learning methods have also been considered for identifying miRNA-disease associations. Chen and Yan[38] presented regularized least-squares for miRNA-disease association (RLSMDA), in which data from known miRNA-disease associations, disease-disease similarity datasets, and miRNA-miRNA functional similarity networks were integrated. Zou et al.[26] introduced two methods to predict miRNA-disease association. CATAPULT[26] is a biased support vector machine (SVM) that was trained to classify miRNA-disease pairs. The other method, the KATZ method,[26] denotes the associations on the basis of the paths of different lengths in the miRNA-disease network. Based on transduction learning, Luo et al.[39] adopted a strategy of collective prediction based on transduction learning (CPTL) to infer potential miRNA-disease associations. Yu et al.[40] first reconstructed the similarity matrices for miRNAs and diseases and then used label propagation to predict the possible links between miRNAs and diseases. The model of within and between score for miRNA-disease association prediction (WBSMDA)[41] uncovers potential miRNA-disease associations according to the within and between scores for many complex diseases, which could predict the potential related miRNAs of new diseases and new miRNAs without known association information. Chen and Huang[42] presented a computational model named Laplacian regularized sparse subspace learning for miRNA-disease association prediction (LRSSLMDA), which used Laplacian regularization to preserve the local structures of the training data. Li et al.[43] designed the matrix completion for miRNA-disease association prediction model (MCMDA), which could efficiently update the low-rank miRNA-disease matrix to identify their associations. Meanwhile, the path-based miRNA-disease association (PBMDA)[22] prediction model is an effective model to predict miRNA-disease association. This model adopts the depth-first search algorithm by integrating the disease semantic similarity, miRNA functional similarity, known human miRNA-disease associations, and Gaussian interaction profile kernel similarity for miRNAs and diseases. inductive matrix completion for miRNA-disease association prediction (IMCMDA)[19] is a matrix computational algorithm that could efficiently update the low-rank miRNA-disease matrix to identify their associations. Chen et al.[44] presented a computational model of matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction (MDHGI) to find new miRNA-disease associations by integrating the predicted association probability obtained from matrix decomposition through sparse learning method.

All mentioned methods have their own strengths, and these methods can be categorized into five aspects: (1) neighborhood-based methods, such as HDMP[35] and CPTL;[39] (2) random walk-based methods, such as RWRMDA,[36] Shi's method,[37] MIDP, and MIDPE;[21] (3) machine-learning-based methods, such as Xu et al.'s method[34] and RLSMDA;[38] (4) path-based methods, such as KATZ[26] and PBMDA;[22] and (5) matrix-based methods, such as MCMDA,[43] IMCMDA,[19] and MDHGI[44].

Inspired by popular neural-network-based approaches[45] and the latest advances in network embedding technologies,[46] we employ NNMDA, which could accurately and efficiently predict miRNA-disease associations by integrating neighborhood information based on neural networks. Specifically, network embedding is an effective approach that aims at converting the network into a low-dimensional space while preserving the structural information of the network.[46] In this way, nodes and associations of the network can be represented as compacted yet informative vectors in the embedding space.[46] In the experiment, we use two evaluation methods, namely, leave-one-out cross-validation (LOOCV) and 5-fold cross-validation (5-fold CV), to verify the performance of our method. Compared with existing approach, our method achieves an outstanding performance in identifying potential miRNA-disease associations. For further verification, we used case studies to analyze the performance of NNMDA. Experimental results show that our method has reliable performance on detecting novel associations. We also found that some special associations and corresponding miRNAs require further attention.

RESULTS

In this section, we analyze the performance of NNMDA from several aspects. Evaluation criteria and methods used in this paper are introduced. The performance of NNMDA was compared with those of other methods in identifying potential associations between miRNAs and diseases. Finally, case studies were utilized to further evaluate the reliability of NNMDA.

Evaluation Criteria and Methods

In this paper, area under the curve (AUC), precision (PRE), and recall (REC) were used as evaluation criteria for the performance of models.

AUC is the area under the receiver operating characteristic (ROC) curve and is established by plotting the true positive rate (TPR) against false positive rate (FPR) at various threshold settings.

PRE (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, whereas REC is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The equations are as follows:

$$PRE = \frac{TP}{TP + FP} \qquad \text{(Equation 1)}$$

where TP and FP are the numbers of true positive and false positive samples, respectively, with respect to a specific disease. A large PRE value indicates good prediction accuracy.

Table 1. Comparison of Various Computational Approaches' *AUC* Values through 5-Fold Cross-Validation

Method	RWRMDA	HDMP	IMCMDA	RLSMDA	MIDP	SPM	NNMDA
Breast neoplasm	0.785	0.801	0.812	0.832	0.838	0.932	0.968
Hepatocellular carcinoma	0.749	0.759	0.744	0.794	0.807	0.918	0.966
Renal cell carcinoma	0.815	0.833	0.793	0.839	0.862	0.901	0.912
Squamous cell carcinoma	0.819	0.820	0.837	0.849	0.870	0.899	0.924
Colorectal neoplasm	0.793	0.802	0.766	0.831	0.845	0.885	0.927
Glioblastoma	0.680	0.700	0.781	0.714	0.786	0.840	0.911
Heart failure	0.722	0.770	0.924	0.738	0.821	0.950	0.945
Acute myeloid leukemia	0.839	0.858	0.861	0.853	0.915	0.957	0.916
Lung neoplasm	0.827	0.835	0.841	0.855	0.876	0.892	0.943
Melanoma	0.784	0.790	0.761	0.807	0.837	0.951	0.949
Ovarian neoplasm	0.882	0.884	0.875	0.909	0.923	0.949	0.928
Pancreatic neoplasm	0.871	0.895	0.894	0.887	0.945	0.954	0.954
Prostatic neoplasm	0.823	0.854	0.775	0.841	0.882	0.928	0.936
Stomach neoplasm	0.779	0.787	0.783	0.797	0.821	0.859	0.955
Urinary bladder neoplasm	0.821	0.850	0.813	0.845	0.897	0.898	0.920
Average AUC	0.799	0.816	0.817	0.826	0.862	0.914	0.937

$$REC = \frac{TP}{TP + FN}, \qquad \text{(Equation 2)}$$

where *FN* is the number of false negative samples with respect to a specific disease.

We evaluated the performance of NNMDA to predict potential disease-related miRNAs by using two evaluation methods (LOOCV and 5-fold CV).

5-fold CV is often used to evaluate the ability of a model to predict potential associations. For a specific disease *d*, *d*-related relationships are randomly divided into five subsets, four of which are used as known information, whereas the remaining one is used for testing.

LOOCV is also a widely used evaluation method. For the disease *d(i)* in our experiment, each known miRNA-disease pair (take miRNA-disease pair (*m(j)-d(i)*) as an example) was selected as the test sample, whereas all the other known miRNA-disease pairs were considered as training samples. First, we artificially changed the known miRNA-disease pairs (*m(j)-d(i)*) into unverified miRNA-disease pairs *d(i)* that were considered as candidate samples. We then ranked the predicted score of the test miRNA-disease pair (*m(j)-d(i)*) with the candidate samples. If the rank of the test miRNA-disease pair (*m(j)-d(i)*) exceeded the given threshold, then the model successfully predicted the miRNA-disease pair (*m(j)-d(i)*).

5-Fold CV

In 5-fold CV, we randomly divided the associations of each disease into five subsets of equal sizes that were used as testing sets.

We compared our method with the following widely applied miRNA-disease prediction algorithms: (1) RWRMDA,[36] (2) HDMP,[35] (3) IMCMDA,[19] (4) RLSMDA,[38] (4) MIDP,[21] and (5) SPM.[23] Table 1 shows the prediction performance measured as *AUC* for different diseases.

Among the 15 algorithms, NNMDA achieved the best performance. Table 1 shows that the average *AUC* scores of RWRMDA, HDMP, IMCMDA, RLSMDA, SPM, and NNMDA were 79.9%, 81.6%, 81.7%, 82.6%, 86.2%, 91.4%, and 93.6%, respectively. The average *AUC* score of NNMDA was higher than that of the other methods by 13.8%, 12.1%, 12.0%, 11.0%, 7.5%, and 2.3%, respectively.

In terms of *AUC* score, NNMDA achieved the highest averaged value but did not exhibit the best performance among all the diseases, particularly in heart failure. Hence, we compared the performance of NNMDA with those of *PRE* and *REC*. For a specific disease, we ranked the related candidates according to their scores.

We measured the *PRE* and *REC* within the top 20, 40 ..., 80, and 100 candidates in the rank list because the top portion of the prediction links is important. *PRE* indicates the ratio of positive samples in the top-*k* samples, whereas *REC* measures how many positive samples are correctly identified within the top-*k*.

Figures 1A and 1B plot the performance of the three methods that achieved the top three *AUC* scores in heart failure. We found that NNMDA outperformed the other two methods in terms of *PRE* (Figure 1A) and *REC* (Figure 1B), indicating the competitiveness of this approach. We also showed that with the increase in *k*, *REC* increased but *PRE* declined. This finding reveals that the links

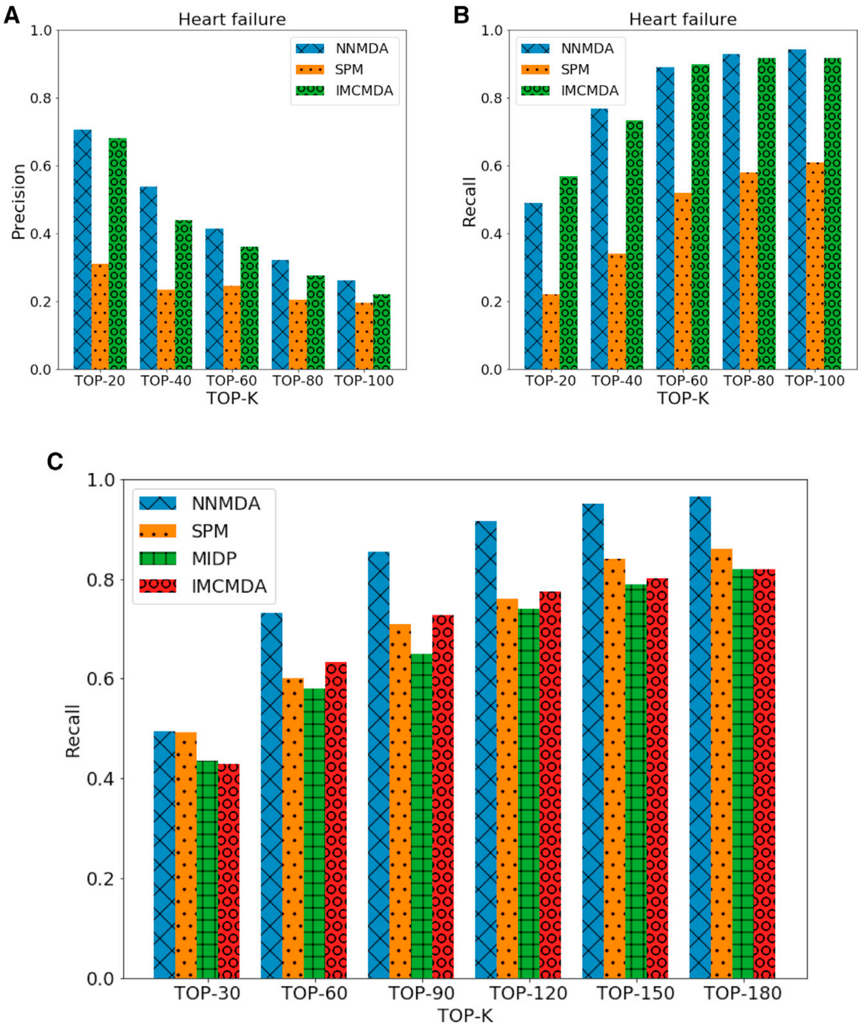

Figure 1. Performances on 5-Fold Cross-Validation Precision

(A) Precision on disease heart failure. (B) Recall on disease heart failure. (C) Average recalls for the 15 tested diseases on four methods (NNMDA, IMCMDA, MIDP, and SPM), which contain the diseases breast neoplasm, hepatocellular carcinoma, renal cell carcinoma, squamous cell carcinoma, colorectal neoplasm, glioblastoma, heart failure, acute myeloid leukemia, lung neoplasm, melanoma, ovarian neoplasm, pancreatic neoplasm, prostatic neoplasm, stomach neoplasm, and urinary bladder neoplasm.

ranked the top places have a high probability of being potential associations.

Figure 1C shows the average *REC* for the 15 tested diseases. Within the top 30, the average RECs of NNMDA, MIDP, SPM, and IMCMDA for all 15 diseases were 49.5%, 43.5%, 49.4%, and 43.0%, respectively. This finding indicates that NNMDA performs slightly better than the other three methods. With the increment of *k*, the performance of NNMDA remarkably increased for the top 60 to 120 predictions. NNMDA outperformed the other four methods.

LOOCV

In this section, a ROC curve was plotted by using the results of LOOCV. The x axis of the ROC graph is the *TPR*, whereas the y axis is the *FPR*. The ROC curve based on LOOCV is plotted in Figure 2. On the basis of the ROC curve, *AUC* could be calculated as an evaluation metric for the model.

Based on the LOOCV results, we compared NNMDA with other four methods, namely, IMCMDA, RWRMDA, HDMP, and RLSMDA. The results showed that NNMDA, IMCMDA, RWRMDA, HDMP, and RLSMDA had obtained AUCs of 0.8432, 0.8034, 0.7891, 0.7702, and 0.6953, respectively. NNMDA achieved the best performance among all these models. Therefore, we can intuitively observe the improved performance of NNMDA in predicting miRNA-disease associations.

Case Study

Two different types of case studies were implemented to validate the performance and evaluate the accuracy of NNMDA for miRNA-disease association prediction.

In the first case study, all the associations between miRNAs and diseases were used to uncover potential associations. For a special disease, we extract the top 30 candidate associations of this disease to determine whether or not these associations can be confirmed by miR2Disease and dbDEMC V2.0 databases. The number of known miRNA-disease associations that are not included in HMDD are used to estimate the performance of NNMDA. Table 2 shows the prediction results of the top 30 predicted lung neoplasm-related miRNAs.

As shown in Table 2, 9 out of the top 10 and 26 out of the top 30 predicted lung-neoplasm-related miRNAs were included. Therefore, most of the potential associations were confirmed by the miR2Disease and dbDEMC V2.0 databases.

An important criterion for evaluating the usefulness of a model is whether or not it can be used to predict potential related miRNAs for a new disease. In the second case study, we evaluated the performance of NNMDA when it was implemented to the new disease without any known related miRNAs. Breast neoplasms were used as an example in our experiment. Therefore, we hid the association information between miRNAs and breast neoplasms by setting all

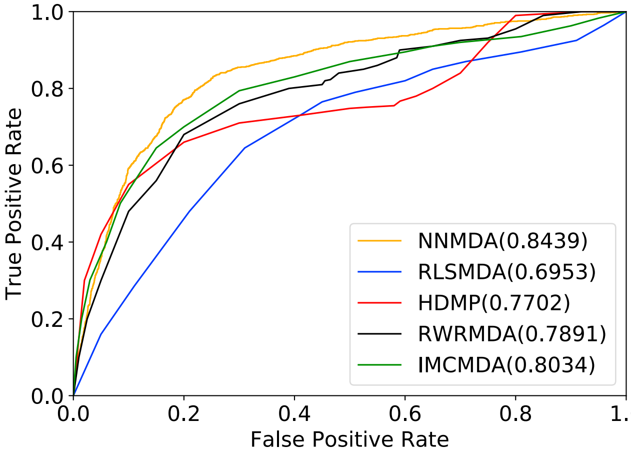

Figure 2. Comparison of Performance among NNMDA and Baseline Methods (NNMDA, IMCMDA, RWRMDA, HDMP, and RLSMDA)

their known associations as unknown ones. We then implemented NNMDA to obtain the ranking list of the association prediction scores for miRNA-breast neoplasms. We analyzed in detail the prediction accuracy on breast neoplasm and mainly focused on the top 50 miRNA candidates. The results for breast neoplasms are represented in Table 3.

DISCUSSION

Identifying potential disease-related miRNAs could provide new insights into the role of miRNA for its impact on clinical measure, diagnosis, and treatment. However, relying on traditional experimental-based methods, predicting the associations between miRNA and disease seems inefficient. Consequently, great numbers of computational methods have been proposed to solve this challenging problem in recent years. In this paper, we apply a neural-network-based model to predict miRNA-disease associations, which aggregates the neighbor information during the process and preserves the topology of the original network at the same time. After that, to comprehensively verify the performance of our method, 5-fold CV and LOOCV are implemented to evaluate NNMDA in comparison with other state-of-the-art approaches. Compared to the state-of-the-art method, NNMDA performs better in terms of *AUC* values on the dataset and is able to retrieve more correct associations. In addition, case studies on two common diseases also gave a strong confirmation to the prediction ability of our method. Results show that NNMDA could be a useful tool for studying the miRNA-disease relationship. The success of our method is mainly due to the following two reasons. First, the constructed similarity networks for both miRNAs and diseases are well integrated in the neural network. Second, the imbalance of datasets that we take into consideration helped improve the efficiency. Nonetheless, more informative data sources should be integrated into our model to further improve the prediction performance. The future work may further take more optimization methods into account to accurately uncover associations between miRNAs and diseases.

Table 2. Prediction Results of the Top 30 Predicted Lung Neoplasm-Related miRNAs Based on Known Associations in HMDD V1.0

miRNA	Evidence	miRNA	Evidence
hsa-let-7g	dbDEMC mir2disease	hsa-mir-18b	dbDEMC
hsa-mir-135b	dbDEMC	hsa-mir-17	dbDEMC
hsa-mir-133b	dbDEMC	hsa-mir-21	dbDEMC mir2disease
hsa-mir-200b	dbDEMC mir2disease	hsa-mir-148a	dbDEMC mir2disease
hsa-let-7d	dbDEMC mir2disease	hsa-mir-18a	dbDEMC mir2disease
hsa-mir-181b-1	unverified	hsa-mir-30e	dbDEMC
hsa-mir-29c	dbDEMC mir2disease	hsa-mir-101-1	mir2disease
hsa-mir-98	dbDEMC mir2disease	hsa-mir-30c-2	unverified
hsa-mir-221	dbDEMC mir2disease	hsa-mir-125a	dbDEMC mir2disease
hsa-mir-186	dbDEMC	hsa-mir-200c	dbDEMC mir2disease
hsa-mir-142	unverified	hsa-mir-126	dbDEMC mir2disease
hsa-mir-146a	dbDEMC	hsa-mir-31	dbDEMC mir2disease
hsa-mir-146b	dbDEMC mir2disease	hsa-mir-30c-1	unverified
hsa-mir-101-1	mir2disease	hsa-mir-30a	dbDEMC
hsa-let-7b	dbDEMC	hsa-mir-192	mir2disease dbDEMC

The first column contains the top 1–15 related miRNAs, whereas the third column shows the top 16–30 related miRNAs.

MATERIALS AND METHODS

miRNA-Disease Network

Data of known human miRNA-disease associations used in this paper were retrieved from the human miRNA-disease database (HMDDv2.0) to construct the miRNA-disease network.[47] If a disease is associated with a miRNA, then an edge is added to link them. The miRNA-disease association matrix is asymmetric and binary, i.e., each entry of the association matrix could only be 0 or 1. A total of 6,441 associations between 577 miRNAs and 336 diseases were obtained after duplications were removed.

Disease Functional Similarity

Functionally similar genes exhibit a great probability of regulating similar diseases. Therefore, we used gene functional information to construct a disease similarity network. The data can be downloaded from the HumanNet database,[33] which contains an associated log-likelihood score (LLS) of each interaction between two genes or gene sets. Similarity $DS(d_i, d_j)$ between diseases d_i and d_j is based on the gene functional information and was calculated as follows:

$$\mathrm{DS}\left(\mathrm{d_i},\mathrm{d_j}\right)=\begin{cases}\dfrac{\sum\limits_{x\epsilon \mathrm{S}(\mathrm{d_i})} LLS(x,\mathrm{S}(\mathrm{d_j}))+\sum\limits_{y\epsilon \mathrm{S}\left(\mathrm{d_j}\right)} LLS\left(y,\mathrm{S}\left(\mathrm{d_i}\right)\right)}{\left|\left|\mathrm{S}(\mathrm{d_i})\right|\right|+\left|\left|\mathrm{S}\left(\mathrm{d_j}\right)\right|\right|}, & \left|\left|\mathrm{S}(\mathrm{d_i})\right|\right| \\ 0, & otherwise\end{cases}$$
$$+\left|\left|\mathrm{S}\left(\mathrm{d_j}\right)\right|\right|\neq 0, \qquad \text{(Equation 3)}$$

where $S(d_i)$ and $S(d_j)$ represent the gene sets that are related to diseases , $\mathrm{d_i}$ and d_j, respectively. $S(d_i)$ and $S(d_j)$ are the cardinalities of

Table 3. Prediction Results of the Top 50 Predicted Breast Neoplasm-Related miRNAs When the Known Associations of Breast Neoplasms Were Considered as Unknown Ones

miRNA	Evidence	miRNA	Evidence
hsa-mir-155	dbDEMC HMDD	hsa-mir-19b-1	HMDD
hsa-mir-21	dbDEMC HMDD	hsa-mir-1-1	HMDD
hsa-mir-146a	dbDEMC HMDD	hsa-mir-145	dbDEMC HMDD
hsa-mir-29b-1	HMDD	hsa-mir-29c	dbDEMC HMDD
hsa-mir-125b-1	HMDD	hsa-mir-199a-2	HMDD
hsa-mir-29b-2	HMDD	hsa-mir-223	dbDEMC HMDD
hsa-mir-34a	dbDEMC HMDD	hsa-mir-126	dbDEMC HMDD
hsa-mir-15a	dbDEMC HMDD	hsa-mir-133a-2	HMDD
hsa-mir-125b-2	HMDD	hsa-mir-19a	dbDEMC HMDD
hsa-mir-20a	dbDEMC HMDD	hsa-mir-199a-1	HMDD
hsa-mir-16-1	HMDD	hsa-let-7b	dbDEMC HMDD
hsa-mir-16-2	HMDD	hsa-mir-26a-1	HMDD
hsa-mir-221	dbDEMC HMDD	hsa-let-7c	dbDEMC HMDD
hsa-mir-29a	dbDEMC HMDD	hsa-mir-142	HMDD
hsa-let-7a-2	HMDD	hsa-mir-146b	HMDD
hsa-mir-26a-2	HMDD	hsa-mir-150	dbDEMC HMDD
hsa-mir-1-2	HMDD	hsa-mir-210	dbDEMC HMDD
hsa-let-7a-1	HMDD	hsa-mir-196a-2	HMDD
hsa-let-7a-3	HMDD	hsa-let-7i	dbDEMC HMDD
hsa-mir-17	dbDEMC HMDD	hsa-let-7d	dbDEMC HMDD
hsa-mir-31	dbDEMC HMDD	hsa-mir-195	dbDEMC HMDD
hsa-mir-92a-1	HMDD	hsa-mir-222	dbDEMC HMDD
hsa-mir-18a	dbDEMC HMDD	hsa-mir-92a-2	HMDD
hsa-mir-122	dbDEMC HMDD	hsa-mir-24-1	HMDD
hsa-mir-133a-1	HMDD	hsa-mir-133b	dbDEMC HMDD

The first column contains the top 1–25 related miRNAs, whereas the third column shows the top 26–50 related miRNAs.

gene sets , $S(d_i)$ and $S(d_j)$, respectively. $LLS(x, \mathrm{S}(\mathrm{d_i}))$ is the LLS between gene x and gene set $S(d_i)$, where $x\varepsilon\, \mathrm{S}(\mathrm{d_i})$. If $DS(d_i, d_j) > 0$, then it can be considered as the weight of the link connecting diseases d_i and d_j. Hence, we obtained a weighted disease similarity network containing 112,896 similar associations among 336 diseases.

miRNA Similarity

The miRNA similarity network was constructed by employing four main miRNA similarities, which are based on verified miRNA-target associations, family information, cluster information, and verified miRNA-disease associations. The verified miRNA-target associations can be downloaded from the miRTarBase,[48] a database of miRNA-target interactions (http://mirtarbase.mbc.nctu.edu.tw/php/index.php) validated by reporter assays and next-generation sequencing experiments. Two miRNA nodes are considered as connected if they share common targets. The edge weight, miRNA similarity based on target, represents the number of shared targets between miRNAs.

We can obtain the family information and cluster information from miRBase.[49] If two miRNAs belong to the same miRNA family, then the value of miRNA similarity based on family would be set as 1, otherwise 0. We obtained 153 clusters of miRNAs. In terms of miRNA similarity based on cluster, the value would be set to 1 if the two miRNAs belonged to the same cluster. These two matrices were found to both be Boolean types. According to literature,[50,51] functionally similar miRNAs tend to connect with similar diseases and vice versa. We downloaded functional similarity data from http://www.cuilab.cn/files/images/cuilab/misim.zip from a previous study.[50] With these data, we constructed matrix FMS to represent the miRNA functional similarity. The element $FMS(m(i),\ m(j))$ denotes the functional similarity between miRNA $m(i)$ and $m(j)$. After a simple combination introduced in Zeng et al.[23], a weighted miRNA similarity network containing 332,928 similar associations among 577 miRNAs was obtained.

Gaussian Interaction Profile Kernel Similarity for Diseases and miRNAs

Considering that similar diseases tend to be related to functionally similar miRNAs and vice versa,[50,52] we calculated Gaussian interaction profile kernel similarities for the miRNAs and diseases' similarity. First, we used $A(i, j)$ to represent the interaction between disease $d(i)$ and miRNA $m(j)$, where A is the miRNA-disease association matrix. Gaussian interaction kernel similarity between disease $d(i)$ and $d(j)$ was calculated as follows:

$$\mathrm{GS}\left(\mathrm{d_i},\ \mathrm{d_j}\right) = \exp\left(-\gamma_\mathrm{d}\sum_{k=1}^{k=nm}\left(A(i,k)-A(j,k)\right)^2\right), \qquad \text{(Equation 4)}$$

where γ_d is used to control the kernel bandwidth that is obtained by normalizing a new bandwidth parameter $\gamma_\mathrm{d}^{'}$ by the average number of associations with miRNAs for all the diseases. γ_d is defined as follows:

$$\gamma_\mathrm{d} = \gamma_\mathrm{d}^{'}\Big/\left[\frac{1}{nm}\sum_{k=1}^{k=nm}\sum_{s=1}^{s=nd}A(k,s)^2\right]. \qquad \text{(Equation 5)}$$

Gaussian interaction profile kernel similarity between miRNA $m(i)$ and $m(j)$ is defined in a similar way:

$$\mathrm{GS}\left(\mathrm{m_i},\mathrm{m_j}\right) = \exp\left(-\gamma_\mathrm{m}\sum_{k=1}^{k=nd}\left(A(k,i)-A(k,j)\right)^2\right), \qquad \text{(Equation 6)}$$

$$\gamma_\mathrm{m} = \gamma_\mathrm{m}^{'}\Big/\left[\frac{1}{nd}\sum_{k=1}^{k=nm}\sum_{s=1}^{s=nd}A(k,s)^2\right]. \qquad \text{(Equation 7)}$$

Schematic Overview

As shown in Figure 3, the framework consists of four major steps: (1; Figure 3A) construct a heterogeneous network based on three miRNA similarity interactions, two disease similarity interactions,

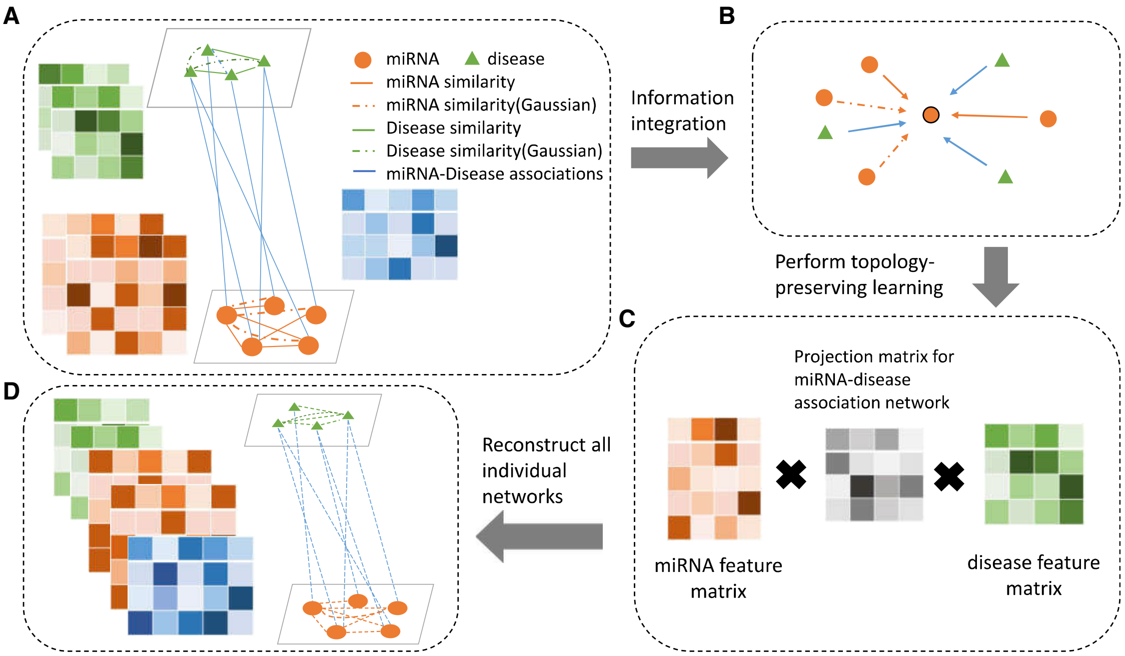
A
miRNA
disease
miRNA similarity
miRNA similarity(Gaussian)
Disease similarity
Disease similarity(Gaussian)
miRNA-Disease associations
Information integration
B
Perform topology-preserving learning
C
Projection matrix for miRNA-disease association network
miRNA feature matrix
disease feature matrix
Reconstruct all individual networks
D

Figure 3. Flowchart of NNMDA

(A) NNMDA uses several individual miRNA-related or disease-related networks to construct a heterogeneous network (details of the used datasets are introduced in Materials and Methods). In a heterogeneous network, different types of nodes are connected by distinct types of edges. Two nodes can be connected by more than one edge (e.g., the solid link between diseases representing disease functional similarity and the chain line between them representing disease Gaussian similarity). (B) Each node adopts a neighborhood information aggregation operation to extract information from the neighborhood. Each arrow represents a specific aggregation function with respect to a specific edge type. Each node then updates its feature representation by integrating its current representation with the aggregated information. (C) NNMDA learns the topology-preserving node features that are useful for miRNA-disease interaction prediction by enforcing the node features to reconstruct the original individual networks. (D) Reconstruction of all individual matrices.

and miRNA-disease associations. The similarity matrices are symmetric, whereas the miRNA-disease association matrix is asymmetric and binary, i.e., each entry of the association matrix could be only 0 or 1. (2; Figure 3B) Integrate the neighborhood information of miRNAs and diseases and further embed them into low-dimensional representations in neural network. (3; Figure 3C) Reconstruct miRNA-disease association matrix by using extracted feature vectors and minimize the loss between the new reconstructed matrices and the observed matrices. This step aims to enforce the learned representations as much as it could from the original matrices. (4; Figure 3D) Predict the miRNA-disease associations by ranking and selecting the values in a decreasing order in the reconstructed association matrix.

Heterogeneous Network

Given a heterogeneous network G = (V, E), V is a node set that contains two kinds of node type, NT = {miRNA, disease}, and E is an edge set with edge types ET = {miRNA-miRNA, miRNA-miRNA-Gaussian, disease-disease-functional, disease-disease-Gaussian, miRNA-disease}. In our framework, each node only belongs to a single node type, whereas the same two nodes can be linked by more than one edge, e.g., two diseases can be simultaneously associated to a disease-disease-functional edge and a disease-disease-Gaussian edge. For each matrix, normalization is first implemented before further processing after data preparation. If $A^{'}$ is the corresponding normalized matrix of the original matrix A, then it can be formulated as follows:

$$A^{'}(\mathrm{i},\mathrm{j}) = \frac{A(i,j)}{\sum_{k=1}^{k=Col(A)} A(i,k)}, \qquad \text{(Equation 8)}$$

where $Col(A)$ is the size of A column dimension. A heterogeneous network can be generated using the normalized matrices as association weight.

Neighborhood Information Aggregation and Node Embedding

To develop a network topology-preserving embedding model that can be used to predict *miRNA-disease* interactions, we adopted the neighborhood information aggregation strategy. For each node u with type $t\in NT$(each node only belongs to a single node type), its features could be aggregated from its neighbors:

$$EM_t^{'}(u) = concat\left(\sum_{s\in ET}\sum_{e=(u,v)\in E_s} A_s^{'}(;,v)\cdot\sigma_s, EM_t(u)\right), \qquad \text{(Equation 9)}$$

where $EM_t^{'}$ is the embedding of node type t (miRNA or disease), the initial representations of nodes (EM) are randomly set, v is the neighbor of u with node type t$^{'}$ (t$^{'}\in NT$), and σ_s is defined as follows:

$$\sigma_s = \sigma(\mathrm{EM}_{t^{'}} \bullet \mathrm{W}_s + \mathrm{b}_s), \qquad \text{(Equation 10)}$$

where W_s $(W_s \in R^{d \times d})$ and $b_s (b_s \in R^d)$ are the parameters trained in neural network, W_s is the weights parameter, and b_s is the bias term. $\sigma(\cdot)$ (implemented as *RELU(x) = max(x, 0)*) is activation function in the neural network. In this step, we further learned node representations into lower dimensional vectors and implement normalization:

$$E^{'}(u) = \frac{\sigma(EM^{'}(u) \bullet W_0 + b_0)}{||\sigma(EM^{'}(u) \bullet W_0 + b_0)||^2}, \qquad \text{(Equation 11)}$$

where $EM^{'}(u)$ is the embedding of node u, W_0 $(W_0 \in R^{2d \times d})$ is the weights, b_0 $(b_0 \in R^d)$ is the responding bias term, and $\sigma(\cdot)$ (implemented as *RELU(x) = max(x, 0)*) is activation function. Therefore, the new embedding was normalized by its l_2-norm. Through neighborhood aggregation, the final neighborhood information is the summation of neighborhood information aggregation with respect to every edge type. We then obtained the representation of each node considering its neighbor information and its own features and learned structural and topological information as the feature vectors.

Topology-Preserving Learning of the Node Embedding

Given the embedding of nodes $E(\cdot)$, topology-preserving learning of the node embedding is defined as:

$$min_{W,b,P,G} \sum_{t \in ET} \left(A_t - E_u^{'} P_t G_t^T E_v^{'\mathrm{T}}\right)^2, \qquad \text{(Equation 12)}$$

where $P, G \in R^{d+k}$ functions are projection matrices that can be used to extract the principle features from node representations. $E_u^{'}$ and $E_v^{'}$ are the embeddings of miRNA or disease with $u, v \in NT$. After projections of $E_u^{'}$ and $E_v^{'}$ by P and G, the inner product of the two projected vectors should reconstruct the original edge weight. For a symmetric matrix reconstruction (i.e., *miRNA-miRNA* or *disease-disease* similarity matrix), matrix P = G was used to enforce symmetry of the recovery. Here, the summation of the squared reconstruction errors was minimized for all edges with respect to all unknown parameters. Given that all operations are differentiable and subdifferentiable, parameters can be trained in an end-to-end manner by performing gradient descent.

After training, each interaction confidence score between miRNA and disease could be predicted by the reconstructed miRNA-disease association matrix. A high score indicates a large probability for the potential association:

$$RD_{reconstruct} = E_r^{'} P_{RD} G_{RD}^T E_d^{'\mathrm{T}}, \qquad \text{(Equation 13)}$$

where $E_r^{'}$ represents miRNA feature matrix and $E_d^{'}$ represents disease feature matrix. In this sense, we can consider our prediction task as a matrix factorization or completion problem. However, our method incorporates a deeper learning model to construct the feature matrices by explicitly defining the construction processes. Through these steps, our method incorporates the prior knowledge of network topology, after which the loss minimization procedure is implemented to prevent the network from arbitrarily factorized.

To further improve the performance of NNMDA, the imbalance of datasets is also taken into consideration. In the process of recovering the associations between miRNAs and diseases, we calculated the loss between prediction matrix and original matrix. To our intuition, the loss obtained from incorrectly predicting a verified entry as an unverified entry (FN) should be different from the loss obtained from wrongly indicating an unknown entry as an verified entry (FP). Because the unknown entries should be considered as unlabeled instead of negative, we redefine loss as follows:

$$Loss = Loss_{FN} + \alpha Loss_{FP}. \qquad \text{(Equation 14)}$$

In our experiment, labeled data are regarded more important than unlabeled. To balance the datasets, in the process of recovering the miRNA-disease matrix, we set parameter α to the ratio of number of entries to size of the matrix, which was finally set to be $\alpha = 6441 \div 577 \div 336 = 0.03$ for the experiments. As a result, the method obtained performance improvement in identifying miRNA-disease associations.

AUTHOR CONTRIBUTIONS

X.Z. conceived the project, designed the experiments, and edited the final paper. W.W. wrote the paper and drafted the figures. G.D. and J.B. contributed to materials and data analysis. W.W., G.D., and J.B. performed the experiments. Q.Z. wrote the paper and edited the paper.

CONFLICTS OF INTEREST

The authors declare no potential conflicts of interest.

ACKNOWLEDGMENTS

The work was supported by the Basic Research Program of Science and Technology of Shenzhen (JCYJ20180306172637807), National Natural Science Foundation of China (grant nos. 61472333, 61772441, 61472335, 61272152, and 41476118), the Project of Marine Economic Innovation and Development in Xiamen (no. 16PFW034SF02), the Natural Science Foundation of the Higher Education Institutions of Fujian Province (no. JZ160400), the Natural Science Foundation of Fujian Province (no. 2017J01099), and the President Fund of Xiamen University (no. 20720170054). X.Z. is supported by a Juan de la Cierva position (code, IJCI-2015-26991).

REFERENCES

1. Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854.

2. Jopling, C.L., Yi, M., Lancaster, A.M., Lemon, S.M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science *309*, 1577–1581.

3. Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

4. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.

5. Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell *113*, 673–676.

6. Carleton, M., Cleary, M.A., and Linsley, P.S. (2007). MicroRNAs and cell cycle regulation. Cell Cycle *6*, 2127–2132.

7. Petrocca, F., Visone, R., Onelli, M.R., Shah, M.H., Nicoloso, M.S., de Martino, I., Iliopoulos, D., Pilozzi, E., Liu, C.G., Negrini, M., et al. (2008). E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer. Cancer Cell *13*, 272–286.

8. Leung, A.K.L., and Sharp, P.A. (2010). MicroRNA functions in stress responses. Mol. Cell *40*, 205–215.

9. Ma, L., Teruya-Feldstein, J., and Weinberg, R.A. (2007). Tumour invasion and metastasis initiated by microRNA-10b in breast cancer *449*, 682–688.

10. Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

11. Hammond, S.M. (2007). MicroRNAs as tumor suppressors. Nat. Genet. *39*, 582–583.

12. Weidhaas, J. (2010). Using microRNAs to understand cancer biology. Lancet Oncol. *11*, 106–107.

13. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37*, D98–D104.

14. Liao, Z., Li, D., Wang, X., Li, L., and Zou, Q. (2018). Cancer Diagnosis Through IsomiR Expression with Machine Learning Method. Curr. Bioinform. *13*, 57–63.

15. Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics *34*, 1953–1956.

16. Hu, Y., Zhou, M., Shi, H., Ju, H., Jiang, Q., and Cheng, L. (2017). Measuring disease similarity and predicting disease-related ncRNAs by a novel method. BMC Med. Genomics *10* (*Suppl 5*), 71.

17. Chen, X., Xie, D., Zhao, Q., and You, Z.H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. *20*, 515–539.

18. Jiang, L., Ding, Y., Tang, J., and Guo, F. (2018). MDA-SKF: Similarity Kernel Fusion for Accurately Discovering miRNA-Disease Association. Front. Genet. *9*, 618.

19. Chen, X., Wang, L., Qu, J., Guan, N.N., and Li, J.Q. (2018). Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics *34*, 4256–4265.

20. Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H., and Liu, H. (2018). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. Bioinformatics *34*, 3178–3186.

21. Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., Zhang, Z., and Ding, J. (2015). Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics *31*, 1805–1815.

22. You, Z.H., Huang, Z.A., Zhu, Z., Yan, G.Y., Li, Z.W., Wen, Z., and Chen, X. (2017). PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

23. Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics *34*, 2425–2432.

24. Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Brief. Bioinform. *17*, 193–203.

25. Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. Int. J. Data Min. Bioinform. *8*, 282–293.

26. Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., and Ju, Y. (2015). Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods. BioMed Res. Int. *2015*, 810514.

27. Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. Brief. Funct. Genomics *15*, 55–64.

28. Tang, W., Liao, Z., and Zou, Q. (2016). Which statistical significance test best detects oncomiRNAs in cancer tissues? An exploratory analysis. Oncotarget *7*, 85613–85623.

29. Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis. *9*, 3.

30. Chen, X., Cheng, J.Y., and Yin, J. (2018). Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. RNA Biol. *15*, 1192–1205.

31. Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y., and Wang, Y. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst. Biol. *4* (*Suppl 1*), S2.

32. Jiang, Q., Hao, Y., Wang, G., Zhang, T., and Wang, Y. (2010). Weighted Network-Based Inference of Human MicroRNA-Disease Associations. In 2010 Fifth International Conference on Frontier of Computer Science and Technology, I. Stojmenovic, G. Farin, M. Guo, H. Jin, K. Li, L. Hu, X. Wei, and X. Che, eds. (IEEE Xplore), pp. 431–435.

33. Li, X., Wang, Q., Zheng, Y., Lv, S., Ning, S., Sun, J., Huang, T., Zheng, Q., Ren, H., Xu, J., et al. (2011). Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. Nucleic Acids Res. *39*, e153.

34. Xu, J., Li, C.X., Lv, J.Y., Li, Y.S., Xiao, Y., Shao, T.T., Huo, X., Li, X., Zou, Y., Han, Q.L., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol. Cancer Ther. *10*, 1857–1866.

35. Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., and Huang, Y. (2013). Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. PLoS ONE *8*, e70204.

36. Chen, X., Liu, M.X., and Yan, G.Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. Mol. Biosyst. *8*, 2792–2798.

37. Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., Zhao, Z., Jiang, W., Guo, Z., and Li, X. (2013). Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst. Biol. *7*, 101.

38. Chen, X., and Yan, G.Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. Sci. Rep. *4*, 5501.

39. Luo, J., Ding, P., Liang, C., Cao, B., and Chen, X. (2017). Collective Prediction of Disease-Associated miRNAs Based on Transduction Learning. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 1468–1475.

40. Yu, S.P., Liang, C., Xiao, Q., Li, G.H., Ding, P.J., and Luo, J.W. (2018). GLNMDA: a novel method for miRNA-disease association prediction based on global linear neighborhoods. RNA Biol. *15*, 1215–1227.

41. Chen, X., Yan, C.C., Zhang, X., You, Z.H., Deng, L., Liu, Y., Zhang, Y., and Dai, Q. (2016). WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. Sci. Rep. *16*, 21106.

42. Chen, X., and Huang, L. (2017). LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. PLoS Comput. Biol. *13*, e1005912.

43. Li, J.Q., Rong, Z.H., Chen, X., Yan, G.Y., and You, Z.H. (2017). MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget *8*, 21187–21199.

44. Chen, X., Yin, J., Qu, J., and Huang, L. (2018). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. PLoS Comput. Biol. *14*, e1006418.

45. Wan, F., Hong, L., Xiao, A., Jiang, T., and Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. Bioinformatics *35*, 104–111.

46. Su, C., Tong, J., Zhu, Y., Cui, P., and Wang, F. (2018). Network embedding in biomedical data science. Brief. Bioinform. Published online December 10, 2018. https://doi.org/10.1093/bib/bby117.

47. Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. *42*, D1070–D1074.

48. Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., et al. (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. *42*, D78–D85.

49. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic Acids Res. *31*, 439–441.

50. Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics *26*, 1644–1650.

51. Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. (2008). An analysis of human microRNA and disease associations. PLoS ONE *3*, e3420.

52. Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M.Q. (2010). Development of the human cancer microRNA network. Silence *1*, 6.