SCIENTIFIC REPORTS



OPEN

SUBJECT AREAS:
MIRNAS
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
SYSTEMS BIOLOGY
CANCER GENOMICS

Received
7 May 2014

Accepted
13 June 2014

Published
30 June 2014

Correspondence and requests for materials should be addressed to X.C. (xingchen@amss.ac.cn) or G.Y.Y. (yangy@amt.ac.cn)

Semi-supervised learning for potential human microRNA-disease associations inference

Xing Chen[1,2] & Gui-Ying Yan[1,2]

[1]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing, 100190, China, [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China.

MicroRNAs play critical role in the development and progression of various diseases. Predicting potential miRNA-disease associations from vast amount of biological data is an important problem in the biomedical research. Considering the limitations in previous methods, we developed Regularized Least Squares for MiRNA-Disease Association (RLSMDA) to uncover the relationship between diseases and miRNAs. RLSMDA can work for diseases without known related miRNAs. Furthermore, it is a semi-supervised (does not need negative samples) and global method (prioritize associations for all the diseases simultaneously). Based on leave-one-out cross validation, reliable AUC have demonstrated the reliable performance of RLSMDA. We also applied RLSMDA to Hepatocellular cancer and Lung cancer and implemented global prediction for all the diseases simultaneously. As a result, 80% (Hepatocellular cancer) and 84% (Lung cancer) of top 50 predicted miRNAs and 75% of top 20 potential associations based on global prediction have been confirmed by biological experiments. We also applied RLSMDA to diseases without known related miRNAs in golden standard dataset. As a result, in the top 3 potential related miRNA list predicted by RLSMDA for 32 diseases, 34 disease-miRNA associations were successfully confirmed by experiments. It is anticipated that RLSMDA would be a useful bioinformatics resource for biomedical researches.

MicroRNAs (miRNAs) are a class of small endogenous single-stranded non-coding RNAs (~22 nt), which normally post-transcriptionally suppress gene expression and protein production by base pairing to the 3′ untranslated regions (UTRs) of their target messenger RNAs (mRNAs)[1–4]. In some cases, miRNAs may also function as positive regulators[5,6]. It has been demonstrated that many miRNAs are highly conserved[7]. Especially, some of them are even lineage specific. After the discovery of the first two well-known miRNAs (Caenorhabditis elegans (C. elegans) lin-4 and let-7 by conventional forward genetic screens[8–10]), thousands of miRNAs (for example, more than 1400 miRNAs in human according to miRBase[11]) have been discovered in eukaryotic organisms ranging from nematodes to humans in the past few years[12]. It is estimated that 1–4% genes in the human genome are miRNAs[13]. MiRNAs recognize their target primarily through sequence complementarity between the seed region of the miRNA and the binding sites on its target mRNAs[14]. It has been conjectured that a single miRNA can regulate as many as 200 mRNAs[13] and about one thirds of human gene can be targeted by miRNAs[12,15]. Therefore, one miRNA can regulate many target genes and one target gene can be targeted by multiple miRNAs[15]. These miRNA-mRNA interactions construct an important post-transcriptional regulatory network which plays critical roles in various biological processes[16–19]. It has been observed that miRNA-mediated regulations are evolutionarily conserved[19–21] and hence typically rare sequence variants that disrupt miRNA regulations are often related to human diseases[19,22–24].

Accumulating evidences indicates that miRNA is one of the most important components of the cell, playing critical roles in many significant biological processes, including the development[25], proliferation[26], differentiation[27], and apoptosis[28] of the cell, signal transduction[16], viral infection[27] and so on. Therefore, the dysregulation of the miRNAs are related to plenty of the diseases, playing important roles in the development, progression[13,29,30], prognosis, diagnosis, and treatment response evaluation of human disease[31–38].

Especially in the last few years, many studies have demonstrated that numerous miRNAs are associated with initiation and development of various cancers and cancer-related processes[39–42]. Abnormality of miRNAs leads to the dysfunction of downstream target genes, which can lead to the development of cancer in turn[42]. MiRNAs have been important part of the field of human molecular oncology[40]. Another well-known example is that mir-375 can regulate insulin secretion[43,44]. Therefore, identifying disease-related miRNAs is one of the most important goals of

biomedical research, which can benefit the understanding of disease pathogenesis at the molecular level, molecular tools design for disease diagnosis, treatment and prevention[31–34,36,45,46]. Searching for disease-miRNA associations form experimental methods is expensive and time-consuming[45,46]. Encouragingly, plenty of biological data about miRNAs has been generated. Therefore, there is strong incentive to develop powerful computational methods for predicting potential disease-related miRNAs on a large scale[47]. Computational methods are an essential complementary means for disease-related miRNAs prioritization, which can benefit the understanding of miRNAs function, decrease the number of biological experiments, and select most promising miRNAs for further experimental validation[45,47].

To provide a comprehensive resource of experimentally verified miRNA-disease associations, Lu, et al.[30] and Jiang, et al.[48] successively constructed two publicly available and manually curated databases, i.e. Human MicroRNA Disease Database (HMDD) and miR2Disease. Focusing on cancer-related miRNAs, Yang, et al.[49] developed a manually curated database of Differentially Expressed MiRNAs in human Cancer (dbDEMC). The establishment of these disease-related miRNAs databases laid a solid data fundament for predictive research. Lu, et al.[30] integrated and analyzed these disease-miRNA associations to obtain some important patterns between human diseases and miRNAs, which not only benefited the understanding of human diseases at miRNA level, but also laid the solid theoretical fundament for the identification of novel disease-related miRNAs. The most important conclusion in this paper is that miRNAs related to phenotypically similar diseases tend to be functionally related, which have been treated as the basic assumption of many current disease-miRNAs associations predication methods[30].

Some bioinformatics methods have been developed for predicting novel disease-miRNA associations mostly based on aforementioned assumption in literature[30]. Jiang, et al.[45] extended logically previous disease genes prioritization methods and developed a computational model based on hypergeometric distribution to prioritize the entire microRNAome for disease of interest. This method integrated the miRNA functional interactions network, disease similarity network, and known phenome-microRNAome network constructed based on miR2Disease. However, this method only adopts local similarity measure and strongly relies on the predicted miRNA-target interactions, which have a high rate of false-positive and high false-negative results. Other limitations lie in the construction of miRNA functional similarity network (two miRNAs may be functionally related when target genes are located in the same functional modules or pathways, rather than significantly share common target genes) and the use of disease phenotypical similarity network (Only used the information whether or not two phenotype are similar, rather than similarity scores). As a result, the prediction accuracy of this method is not high. Based on the assumption that most of miRNAs associated with given disease regulates genes associated with this disease, or functionally related genes with these known disease genes, Jiang, et al.[50] proposed a computational method based on genomic data fusion in the framework of naïve Bayes. Recently, Shi et al.[51] developed a computational framework to identify miRNA-disease associations by focusing on the functional link between miRNA targets and disease genes in protein-protein interaction networks. These two methods strongly relied on known disease-genes association and miRNA-target interactions. However, the molecular bases for as many as 60% of human disease are unknown. The problem of miRNA-target interactions has also limited the application of this method.

Jiang, et al.[46] and Xu, et al.[40] extracted different feature vectors and developed the support vector machine classifier to distinguish positive disease miRNAs from negative ones, respectively. As we all known, selecting negative disease-related miRNAs is currently difficult or even impossible. Hence, these methods selected unlabeled disease-miRNAs interactions as negative samples, which would largely influence the predictive accuracy. Based on the assumption that global network similarity measures are better suited to capture the associations between diseases and miRNAs than traditional local network similarity, Chen, et al.[47] first adopted global network similarity and developed the method of Random Walk with Restart for MiRNA–Disease Association (RWRMDA). Also, Xuan et al.[52] developed the new prediction method of HDMP based on weighted k most similar neighbors by calculating the functional similarity between miRNAs from the information content of disease terms and phenotype similarity between diseases and assigning higher weight to members of miRNA family or cluster. RWRMDA and HDMP obtained excellent predictive accuracy based on cross validation and case studies. However, they does not work for disease without any known associated miRNA. Furthermore, the selection of parameter k is critical to the performance of HDMP and we should have different values of this parameter when different diseases are investigated. Recently, Chen and Zhang[53] adopt the method of Network-Consistency-Based Inference (Net-CBI) to infer potential disease-miRNA associations based on the idea of network consistency and the integration of miRNA functional similarity network, disease similarity network and known miRNA disease associations. Although Net-CBI can work for diseases not linked with any known miRNAs, the performance is significantly worse than RWRMDA based on the validation of cross validation.

Taken together, the limitations of previous methods are summarized as follows. Firstly, some methods strongly relies incomplete and inaccuracy datasets such as miRNA-target interactions, disease-related genes; secondly, some methods need negative disease-miRNA associations; thirdly, although methods such as RWRMDA have obtained reliable predictive accuracy, they can't predict novel miRNAs for diseases which do not have any known associated miRNAs; finally, methods such as Net-CBI can work for disease without known related miRNAs, but unsatisfactory performances have been obtained. To solve these problems, we developed the method of Regularized Least Squares for MiRNA-Disease Association (RLSMDA) by integrating known disease-miRNA associations, disease-disease similarity dataset, and miRNA-miRNA functional similarity network to uncover potential disease-miRNA associations. RLSMDA can predict novel miRNAs for diseases which do not have any known related miRNAs. More importantly, it is developed in the framework of semi-supervised classifier, so it does not need negative miRNA-disease associations. Furthermore, different from RWRMDA, RLSMDA is a global approach which can reconstruct the missing associations for all the diseases simultaneously. Cross validations, Case studies about several important diseases, global prediction for all the diseases simultaneously, and independent prediction for diseases without any known related miRNAs have fully demonstrated the superior performance of RLSMDA to previous methods.

Results

Leave-one-out cross validation. Here, we implemented LOOCV on known experimentally verified miRNA-disease associations to evaluate the predictive performance of RLSMDA. To our knowledge, RWRMDA[47], HDMP[52], and the global network algorithm developed by Shi et al.[51] are the-state-of-art approaches in the computational research about disease-related miRNA prediction. However, the global network algorithm developed by Shi et al.[51] focused on the functional connectivity between miRNA targets and disease genes in PPI network. Therefore, this method integrated the information of disease gene associations, miRNA-target interactions, and protein interactions, which were totally different from the dataset used in RLSMDA. Furthermore, this method did not use the information of known disease-miRNA associations and cross validation by splitting known samples into test samples and training samples implemented in this paper

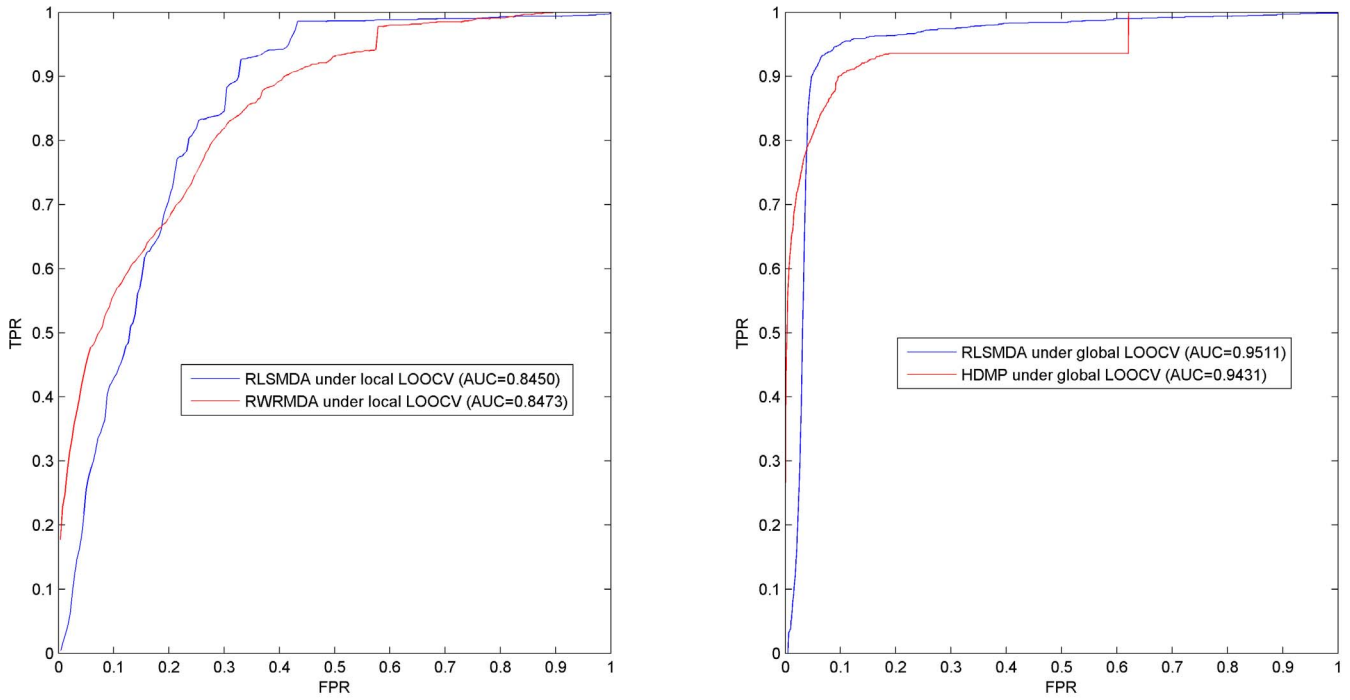


Figure 1 | Method comparison: (left) Comparison between RLSMDA and RWRMDA proposed by Chen, et al.[47] in terms of ROC curve and AUC based on local leave-one-out cross validation on 1394 known experimentally verified miRNA–disease associations. RLSMDA obtained comparable performance in the local LOOCV as RWRMDA, while RWRMDA cannot predict disease-related miRNAs for diseases without known related miRNAs and all the diseases simultaneously. RLSMDA can successfully solve these two critical shortcomings of RWRMDA. (right) Comparison between RLSMDA and HDMP in the term of global LOOCV. RLSMDA and HDMP obtained the AUC of 0.9511 and 0.9431, respectively. Although only slight improvement has been obtained here, RLSMDA can predict the potential miRNAs for diseases which do not have known related miRNAs, which has solved the most critical limitation of HDMP. The performance of RLSMDA could be further improved by introducing the information of miRNA family and cluster as what has been done in the method of HDMP.

cannot be implemented for this method. Therefore, the performance of this method and RLSMDA could not be compared in a fair and reasonable way. Based on the above consideration, we will compare the performance of RLSMDA with RWRMDA and HDMP.

For simplicity, we choose $\eta_M = 1$, $\eta_D = 1$ for trade-off parameters in the cost functions according to previous literatures[54] and weight parameter $w = 0.9$ in the final classifier considering the fact that miRNA functional similarity has played a critical role in disease-related miRNA prediction, as what have shown in the method of RWRMDA. Both trade-off parameters in the cost function and weight parameter in the final classifier can be better selected by further cross validation.

LOOCV can be implemented in the following two ways: (1) For the *ith* disease, each known miRNA associated with disease *i* was left out in turn as test miRNA. Entity *F(i,j)* in row *i* column *j* of the matrix *F* reflect the probability that miRNA j is related to the disease i. How well this test miRNA was ranked relative to the candidate miRNAs was evaluated based on the *ith* line of the matrix *F* (seed miRNAs: other known disease-miRNA associations; candidate miRNAs: all the miRNAs which do not have the evidence to show their association with disease *i*). If the rank of test miRNA exceeds the given threshold, the model was considered to successfully predict this miRNA–disease association. We called the LOOCV in this way as local LOOCV. (2) Unlike LOOCV, we did not give a fixed disease, where all the diseases were considered simultaneously. Each known disease-miRNA association was left out in turn as test association and how well this test association was ranked relative to the candidate associations was evaluated based on matrix *F* (seed associations: other known disease-miRNA associations; candidate associations: all the disease-miRNA pairs which do not have the evidence to confirm the association). If the rank of test association exceeds the given threshold, the model was considered to successfully predict this association. We called the LOOCV in this way as global LOOCV. The difference between local and global LOOCV is whether we considered all the diseases simultaneously. From the aforementioned fact that RWRMDA cannot uncover the missing associations for all the diseases simultaneously, we cannot implement global LOOCV for RWRMDA. For the HDMP, global LOOCV can be implemented. As a global predictive approach, RLSMDA can be checked in both local and global LOOCV.

Receiver-operating characteristics (ROC) curve was drawn and Area under the curve (AUC) was calculated to evaluate the performance of predictive methods. ROC curve plots true positive rate (sensitivity) versus false positive rate (1-specificity) at different thresholds. Sensitivity refers to the percentage of the test samples whose ranking is higher than a given threshold. Specificity refers to the percentage of samples that are below the threshold. AUC = 1 indicates perfect performance and AUC = 0.5 indicates random performance.

According to literature[47], the AUC of RWRMDA is 0.8617, which has significantly improved the performance of previous computational method based on the hypergeometric distribution[45]. However, for diseases which only have 1 known miRNA, LOOCV can't be implemented. To be fair, we think left-out known association obtained the random rank in that case, i.e. for *N* candidate miRNAs, we regard the rank of left-out known miRNA as *(N+1)/2*. Recalculated AUC for RWRMDA was 0.8473. For global LOOCV, HDMP obtained an AUC of 0.9431. For RLSMDA, AUC in local and global LOOCV is 0.8450 and 0.9511, respectively (see Figure 1). We can reach the conclusion that the performance of RLSMDA is comparable to RWRMDA and slightly better than HDMP. However, RWRMDA and HDMP cannot predict the potential miRNAs for diseases which do not have known related miRNAs, which is the

Table 1 | The top 50 potential Hepatocellular cancer (HCC) related miRNAs predicted by RLSMDA and the confirmation for their associations by various databases are listed here (1st column: top 1–25; 2nd column: top 26–50). Forty of top 50 miRNAs have been confirmed to be related with HCC

Name	Evidence	Name	Evidence
hsa-mir-155	HMDD,dbDEMC,miR2Disease	hsa-mir-29c	HMDD,DbDEMC
hsa-mir-24	HMDD,miR2Disease	hsa-mir-146b	HMDD
hsa-mir-107	HMDD,dbDEMC,miR2Disease	hsa-mir-194	dbDEMC,miR2Disease
hsa-mir-29b	HMDD,DbDEMC	hsa-let-7d	HMDD,miR2Disease
hsa-mir-126	HMDD,dbDEMC,miR2Disease	hsa-mir-135b	Unconfirmed
hsa-let-7i	HMDD,DbDEMC	hsa-mir-497	HMDD,DbDEMC
hsa-mir-183	HMDD,miR2Disease	hsa-mir-204	Unconfirmed
hsa-mir-214	HMDD,dbDEMC,miR2Disease	hsa-let-7b	HMDD,miR2Disease
hsa-mir-34c	HMDD	hsa-mir-25	HMDD,dbDEMC,miR2Disease
hsa-mir-31	HMDD,miR2Disease	hsa-mir-32	Unconfirmed
hsa-mir-191	HMDD,DbDEMC	hsa-mir-196b	Unconfirmed
hsa-mir-181b	HMDD,dbDEMC,miR2Disease	hsa-mir-378	Unconfirmed
hsa-let-7f	HMDD,miR2Disease	hsa-mir-142	HMDD,miR2Disease
hsa-mir-103	dbDEMC,miR2Disease	hsa-mir-95	Unconfirmed
hsa-let-7g	HMDD,miR2Disease	hsa-mir-148b	HMDD,dbDEMC,miR2Disease
hsa-mir-132	miR2Disease	hsa-mir-210	HMDD,DbDEMC
hsa-mir-128b	miR2Disease	hsa-mir-205	HMDD,miR2Disease
hsa-mir-151	miR2Disease	hsa-mir-199b	HMDD,miR2Disease
hsa-mir-451	dbDEMC	hsa-mir-498	HMDD
hsa-mir-150	HMDD,dbDEMC,miR2Disease	hsa-mir-182	HMDD,miR2Disease
hsa-let-7c	HMDD,dbDEMC,miR2Disease	hsa-mir-421	HMDD
hsa-mir-34b	Unconfirmed	hsa-mir-93	HMDD,dbDEMC,miR2Disease
hsa-mir-141	HMDD,miR2Disease	hsa-mir-340	Unconfirmed
hsa-mir-29a	HMDD,DbDEMC	hsa-mir-193b	Unconfirmed
hsa-mir-658	Unconfirmed	hsa-mir-30c	HMDD,miR2Disease

major defect of their methods. Furthermore, RWRMDA is a local approach which cannot uncover the missing associations for all the diseases simultaneously, i.e. we cannot compare the scores between one miRNA and two different diseases. Although there is no significant improvement in the way of AUC, RLSMDA can successfully solve aforementioned these two problems. Furthermore, HDMP introduce additional information of miRNA family and cluster, which benefit the performance of their method. It is much likely that the performance of RLSMDA would be further improved after introducing the information of miRNA family and cluster into its model. Excellent performance demonstrates RLSMDA can recover known experimentally verified miRNA–disease associations and hence has the potential to predict potential associations.

Parameter effect. In the above cross validation, we want to place more emphasis on miRNA space classifier (this classifier is based on the dataset of miRNA functional similarity dataset) in the final classifier based on the fact that miRNA functional similarity has played a critical role in disease-related miRNA prediction. However, we cannot totally rely on the results from miRNA space, because in that way we cannot predict potential miRNAs for diseases which do not have any known related miRNAs. Therefore, we chose weight parameter $w = 0.9$ in the final classifier. We also assigned the different weights for the classifier constructed in the miRNA space and calculated corresponding AUCs. The result has been shown in Supplementary Figure 1 and it could be observed that a higher weight can improve the final performance of RLSMDA.

Case studies. It has been demonstrated that many miRNAs are associated with various human cancers[12,13,38,55–57] and almost half of miRNAs are located in cancer-associated genomic regions or fragile sites[12,55]. Here, case studies about several important diseases were implemented to evaluate the independent predictive ability of RLSMDA. Predictive results were confirmed based on the update of HMDD and the datasets in miR2disease and dbDEMC.

Hepatocellular cancer (Hepatocellular carcinoma, malignant hepatoma, HCC) is the third leading cause of cancer deaths worldwide nowadays, with over 500,000 people affected (http://emedicine.medscape.com/article/197319-overview). As the most common type of liver cancer, the most affected people of HCC come from Asia and Africa, where high prevalence of hepatitis B and hepatitis C strongly leads to the development of chronic liver disease and HCC (http://emedicine.medscape.com/article/197319-overview). In the gold-standard data, 34 miRNAs have been related to the development of HCC. For example, independent experimental observations showed that the expression of miRNAs let-7e, 125a and 99b were quite lower in HCC compared to normal liver[58]. MiRNAs without the known relevance to HCC were prioritized based on the predictive results of RLSMDA. Among the top 50 predicted HCC-related miRNAs, 40 miRNAs have been confirmed by aforementioned various databases. Especially, top 20 potential miRNAs are all confirmed. The top 50 potential HCC related miRNAs and evidences for the associations with HCC were listed (See Table 1). Unconfirmed potential miRNA with the highest rank is the miR-34b (ranked 22th). However, the recent findings in the literature[59] showed that the potentially functional SNP rs4938723 in the promoter region of pri-miR-34b/c may lead to the development of HCC in the investigated Chinese population, which established the connection between HCC and miR-34b. All the datasets used in this paper is generated before the publication of this paper. Therefore, this successful independent literature validation gave a further strong support to the reliable performance demonstration of RLSMDA. We did not further check whether the associations between other unconfirmed potential miRNAs and HCC can be verified based on recent experimental literatures. However, the excellent performance of RLSMDA based on cross validation and previous case study makes us believe that RLSMDA can predict more disease-related miRNAs.

In our previous paper about the method of RWRMDA[47], 98% (Breast cancer), 74% (Colon cancer), and 88% (Lung cancer) of top 50 predicted miRNAs are confirmed by published experiments. It seems that the predictive accuracy for Breast cancer and Lung cancer has been much satisfactory. Hence, we implemented the case study about Colon cancer here to see whether RWRMDA can further

Table 2 | The top 20 potential disease related miRNAs predicted by RLSMDA in the global ranking and the confirmation for their associations by various databases are listed here. Fifteen of top 20 disease-miRNA associations have been confirmed

Ranking	Diseases	miRNAs	Evidence
1	Colonic Neoplasms	hsa-mir-222	dbDEMC
2	Stomach Neoplasms	hsa-mir-451	miR2Disease
3	Ovarian Neoplasms	hsa-mir-15a	
4	Colorectal Neoplasms	hsa-mir-19a	HMDD,miR2Disease
5	Muscular Disorders, Atrophic	hsa-mir-206	
6	Colonic Neoplasms	hsa-mir-203	dbDEMC,miR2Disease
7	Stomach Neoplasms	hsa-mir-19b	
8	Breast Neoplasms	hsa-let-7e	HMDD,dbDEMC
9	Colonic Neoplasms	hsa-mir-92b	
10	Carcinoma, Hepatocellular	hsa-mir-155	HMDD,dbDEMC,miR2Disease
11	Colorectal Neoplasms	hsa-mir-125b	HMDD
12	Breast Neoplasms	hsa-let-7b	HMDD,dbDEMC
13	Adenocarcinoma	hsa-mir-200b	HMDD
14	Colonic Neoplasms	hsa-mir-183	dbDEMC,miR2Disease
15	Breast Neoplasms	hsa-mir-92a	HMDD
16	Ovarian Neoplasms	hsa-mir-143	miR2Disease
17	Breast Neoplasms	hsa-mir-223	HMDD,dbDEMC
18	Neoplasms	hsa-mir-15a	HMDD
19	Breast Neoplasms	hsa-mir-16	HMDD,dbDEMC
20	Stomach Neoplasms	hsa-mir-92a	

improve the performance of our method in the case study of Colon cancer. As the third most common cancer in the world, more than half of the people who die of Colon cancer come from developed countries (http://en.wikipedia.org/wiki/Colonic_cancer). Usually colon cancer strikes without symptoms, therefore, it's important to get a colon cancer screening test. If the colon cancer is found early, the doctor can use surgery, radiation, and/or chemotherapy for effective treatment (http://www.webmd.com/colorectal-cancer/default.htm). There are thirty-seven known colon cancer related miRNAs in the golden standard dataset. For example, miR-200b and miR-141 have been shown to be highly overexpressed in colon carcinoma[60]. Candidate miRNAs were prioritized in the term of scores obtained from the method of RLSMDA. Forty-two out of top fifty predicted colonic cancer related miRNAs have been confirmed by various databases and literatures[12,61,62]. The top 50 potential colonic cancer related miRNAs and confirmation evidences for the associations were listed (See Supplementary Table 1). A typical example is miR-18b, which is ranked 24th in the predictive list. Recent experimental literature confirm its connection to colonic cancer[62]. In that paper, the expression of miR-18b was upregulated in colonic cancer tissues, compared with the para-cancerous control. Therefore, miR-18b is expected to participate in the process of colonic cancer and play a critical role in the carcinogenesis of colon. As mentioned, the dataset used in this paper for potential miRNAs prediction is generated before the publication of this paper. Another independent validation further supports the excellent performance of RLSMDA.

As mentioned, RLSMDA can reconstruct the missing associations for all the diseases simultaneously. The top 20 potential disease-miRNA associations predicted by RLSMDA and the confirmation based on various databases are listed in the Table 2. Fifteen of top 20 potential disease-miRNAs associations have been confirmed. Also, the top 100 potential disease-miRNA associations were shown in Supplementary Table 2 and verified based on various databases and literatures[12,61]. These 100 potential associations involved various diseases, including breast cancer, colonic cancer, brain cancer, type 2 diabetes and so on. As a result, 61 out of top 100 potential associations were confirmed.

Applicability of RLSMDA to diseases without any known related microRNAs. To demonstrate that RLSMDA is applicable to diseases without any known associated miRNAs, we implemented case studies for the diseases discussed in the above section by removing all the known verified miRNAs which have been shown to be related to this disease. This operation made sure that prioritizing candidate miRNAs for the given disease only made use of the information of other diseases having known related miRNAs and similarity information. The fact must be pointed out we select the same candidate miRNA set as normal case study for a given disease, i.e. abandoned known seed miRNAs were not regarded as candidate miRNAs.

For the Hepatocellular cancer, we removed 34 known HCC related miRNAs to prioritize candidate miRNAs based on the predictive result of RLSMDA. Among the top 50 potential prediction, 36 miRNAs have been confirmed by various databases. The top 50 potential HCC related miRNAs when the information about known HCC related miRNAs are removed and evidences for the associations with HCC were listed (See Supplementary Table 3). The aforementioned successful independent literature validation example about HCC and miR-34b were also ranked in the top 50 predictive list. For the colon cancer, after removing 37 known seed miRNAs, RLSMDA was implemented to uncover potential connection between colon cancer and candidate miRNAs. As a result, 36 out of top 50 miRNAs are confirmed by various databases and literatures[12,61,62]. Top 50 potential miRNAs and the evidences were listed (See Supplementary Table 4). Surprisingly, successful independent predictive example of miR-18b and colon cancer is ranked 1st by RLSMDA when known colon cancer related miRNAs are removed.

Except for above simulation experiments, RLSMDA was also applied to diseases without any known related miRNAs in our golden standard dataset. In this way, when we prioritize candidate miRNAs for the given disease, only the disease-miRNA associations of other diseases and similarity information between these diseases have been used. The prediction result was verified based on recent experimental literatures. As a result, in the top 3 potential related miRNA list predicted by RLSMDA for 32 diseases investigated here, 34 disease-miRNA associations were successfully confirmed by biological experiments[63–95] (See Table 3).

For example, hsa-mir-21 has been shown to play a critical role in various cellular processes including maturation, migration, proliferation, and survival. Accumulated evidences has linked mir-21 to many complex human diseases and its associations with many diseases have been collected in the golden standard dataset, such as Breast cancer, Brain cancer, Lung cancer, Stomach cancer, and so on. Here, we predicted mir-21 as the most likely related miRNAs for

Table 3 | Confirmed disease-miRNA associations predicted by RLSMDA for diseases without known related miRNAs in our golden standard dataset

Ranking	Diseases	miRNAs	PMID
1	Acute Coronary Syndrome	hsa-mir-1	21806992
1	Aortic Aneurysm, Abdominal	hsa-mir-21	22357537
1	Aortic Aneurysm, Thoracic	hsa-mir-21	22010139
1	Arthritis, Psoriatic	hsa-mir-146a	20500689
1	Crohn Disease	hsa-mir-16	22386737
1	Laryngeal Neoplasms	hsa-mir-205	22605671
1	Leukemia, Myelogenous, Chronic, BCR-ABL Positive	hsa-mir-181a	22442671
1	Liver Failure	hsa-mir-221	21400558
1	Lupus Erythematosus, Systemic	hsa-mir-146a	21529448
1	Mesothelioma	hsa-mir-18a	21358347
1	Osteosarcoma	hsa-mir-15a	22922827
1	Retinoblastoma	hsa-mir-181b	21373755
1	Sezary Syndrome	hsa-mir-21	21525938
1	Vascular Diseases	hsa-mir-21	20560046
2	Amyloidosis	hsa-mir-16	21834602
2	Antiphospholipid Syndrome	hsa-mir-20a	21794077
2	Aortic Valve Stenosis	hsa-mir-21	22882958
2	Atrial Fibrillation	hsa-mir-223	22944230
2	Creutzfeldt-Jakob Syndrome	hsa-mir-146a	22043907
2	Endometrial Neoplasms	hsa-mir-194	21851624
2	Huntington Disease	hsa-mir-200c	22906125
2	Lichen Planus, Oral	hsa-mir-21	21943223
2	Mesothelioma	hsa-mir-20a	21358347
2	Lymphoma, Non-Hodgkin	hsa-mir-21	22487708
2	Osteosarcoma	hsa-mir-16	22922827
3	Colitis, Ulcerative	hsa-mir-143	21557394
3	Cystic Fibrosis	hsa-mir-155	21282106
3	Endometrial Neoplasms	hsa-mir-155	21176560
3	Fibrosis	hsa-mir-29c	21784902
3	Hyperlipidemias	hsa-mir-122	22587332
3	Keratoconus	hsa-mir-184	21996275
3	Mycosis Fungoides	hsa-let-7a	21966986
3	Neoplasms, Squamous Cell	hsa-mir-181a	21244495
3	Osteoporosis	hsa-mir-133a	22506038

Abdominal Aortic Aneurysm (AAA), Thoracic Aortic Aneurysm (TAA), Sezary Syndrome (SS), and Vascular Diseases. These predictions were all confirmed by biological experiments. Maegdefessel et al identified mir-21 as a key modulator of proliferation and apoptosis of vascular wall smooth muscle cells during development of AAA and provided a new therapeutic pathway that could be targeted to treat AAA[95]. Jones et al observed decreased expression of mir-21 in TAA compared to normal aortic samples and further identified a significant relationship between its expression level and aortic diameter[65]. Narducci et al profiled the expression of miRNAs in a cohort of 22 SS patients and identified differential expression of mir-21 between SS and controls[75]. Cheng and Zhang pointed out mir-21 plays important roles in biological processes, such as vascular smooth muscle cell proliferation and apoptosis, cardiac cell growth and death, and cardiac fibroblast functions, and so on. Furthermore, they showed that mir-21 is proven to be involved in the pathogenesis of the cardiovascular diseases[76]. These successful predictive examples fully demonstrates that RLSMDA has the potential to provide high-quality disease-miRNA associations for the diseases without any known related miRNAs, which solved the critical deficiency existing in the previous methods.

Predicting novel human miRNAs-disease associations. Here, we further applied RLSMDA to predict potential human disease-miRNAs associations after confirming the reliable performance of RLSMDA in the term of cross validation and case studies. All the known disease-miRNA associations in the gold-standard dataset were used as positive samples. We publicly released potential human disease-miRNA association list to facilitate the biological experimental validation (see Supplementary Table 5). It is anticipated that potential disease-miRNA associations predicted here could be validated by further biological experiments and useful for biomedical research.

Discussions

Identifying potential disease-miRNA associations is critical for understanding the pathogenesis of disease at the miRNA level and further improving human medicine. In this paper, RLSMDA was developed to identify disease-related miRNAs by integrating disease-disease semantic similarity information, miRNA-miRNA functional similarity information, and known human miRNA-disease associations on a large scale. RLSMDA was motivated in the framework of regularized least squares and the basic assumption that functionally related miRNAs tend to be related to phenotypically similar diseases. Compared with previous methods, RLSMDA can identify related miRNAs for diseases without any known associated miRNAs. Furthermore, RLSMDA does not need negative samples selection and reconstruct the missing associations for all the diseases simultaneously. Cross validation and case studies about Hepatocellular cancer and Lung cancer have fully demonstrated the reliable performance of RLSMDA. Furthermore, we implemented simulated case studies for Hepatocellular cancer and Lung cancer after removing all the known verified miRNAs which have been shown to be related to this disease. Plenty of prediction results were confirmed by various databases and literature. More importantly, when we applied RLSMDA to diseases without any known related miRNAs in our golden standard dataset, 34 disease-miRNA associations, ranked in the top 3 potential related miRNA list predicted by RLSMDA for 32

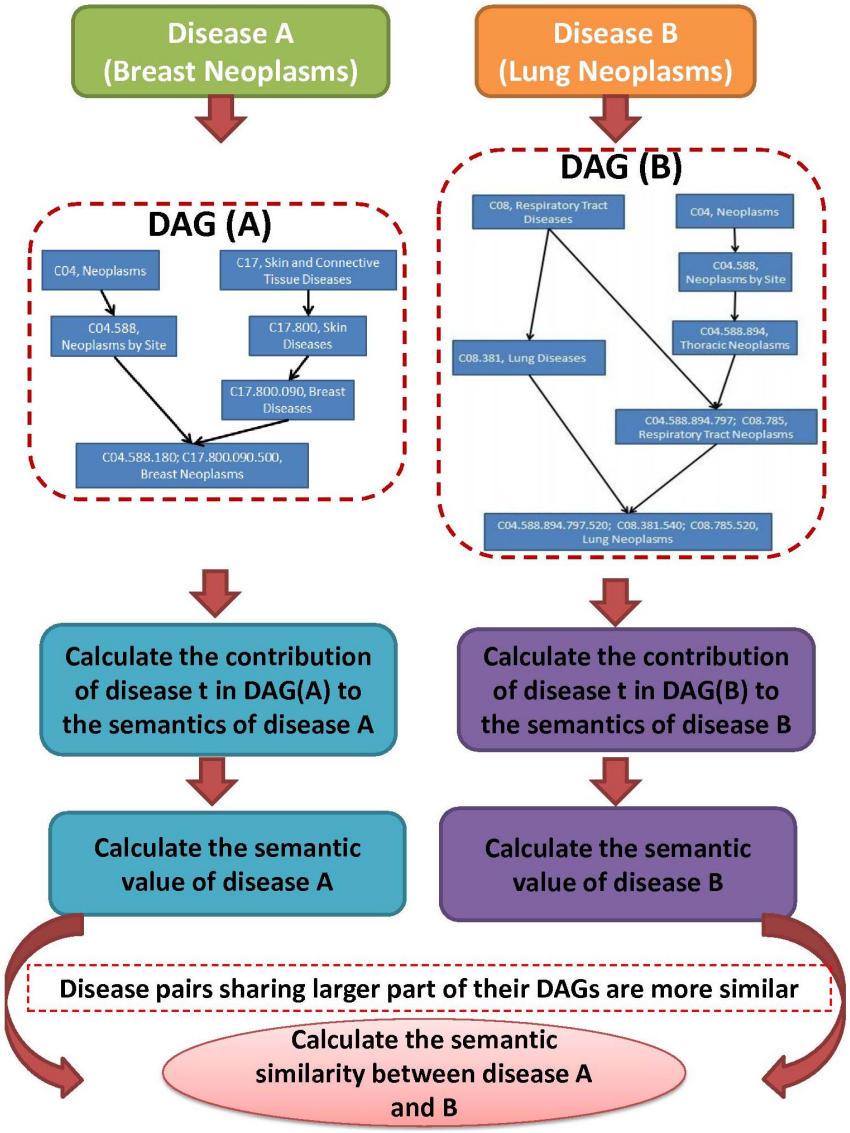


Figure 2 | **The basic idea of disease semantic similarity calculation.**

diseases investigated here, were successfully confirmed by biological experiments.

These excellent examples fully demonstrated that RLSMDA is applicable to diseases without any known associated miRNAs. Considering the fact that RLSMDA can reconstruct the missing associations for all the diseases simultaneously, we applied it to implement global prediction for all the diseases simultaneously. As a result, 15 of top 20 potential disease-miRNAs associations have been confirmed. Also, out of the top 100 potential disease-miRNA associations, 61 potential associations were confirmed, involved various diseases including breast cancer, colonic cancer, brain cancer, type 2 diabetes and so on. We publicly released potential miRNA lists for 137 diseases investigated in this paper to guide biological experiments. It is anticipated that RLSMDA would be a useful resource for researches about the associations between miRNAs and human diseases.

The reliable performance of RLSMDA could largely be attributed to several factors as follows. Firstly, heterogeneous datasets (known disease-miRNA associations, miRNA functional similarity, and disease semantic similarity) were integrated to capture the potential associations between disease and miRNA. Especially, RLSMDA can predict potential related miRNAs for diseases without any known associated miRNAs by introducing the information of disease similarity. Secondly, RLSMDA is a semi-supervised method, which overcomes the difficulties in obtaining negative disease-miRNA associations samples in the practical problems. Finally, RLSMDA is a global approach, which can predict the scores between miRNAs and diseases for all the diseases simultaneously. These three critical success factors also constitute the novelties of RLSMDA. Hence, RLSMDA represents a novel, useful, and important biomedical resource for miRNA-disease association identification.

Although there are several important novelties in the method development of RLSMDA, some limitations also exist. Firstly, how to decide the parameters values in the RLSMDA is not still solved well. Especially, we need to integrate predictive result from disease space and miRNA space by weight parameters. How to directly obtain a single classifier or reasonably integrate results from different spaces would be a critical problem for future research. Secondly, more reliable construction of disease similarity and miRNA similarity would further improve the predictive ability. We plan to integrate more biological relevant information to define miRNA similarity and disease similarity. Thirdly, more available experimentally verified human disease-miRNA associations would promote the development and the performance of computational human disease-miRNA identification methods.

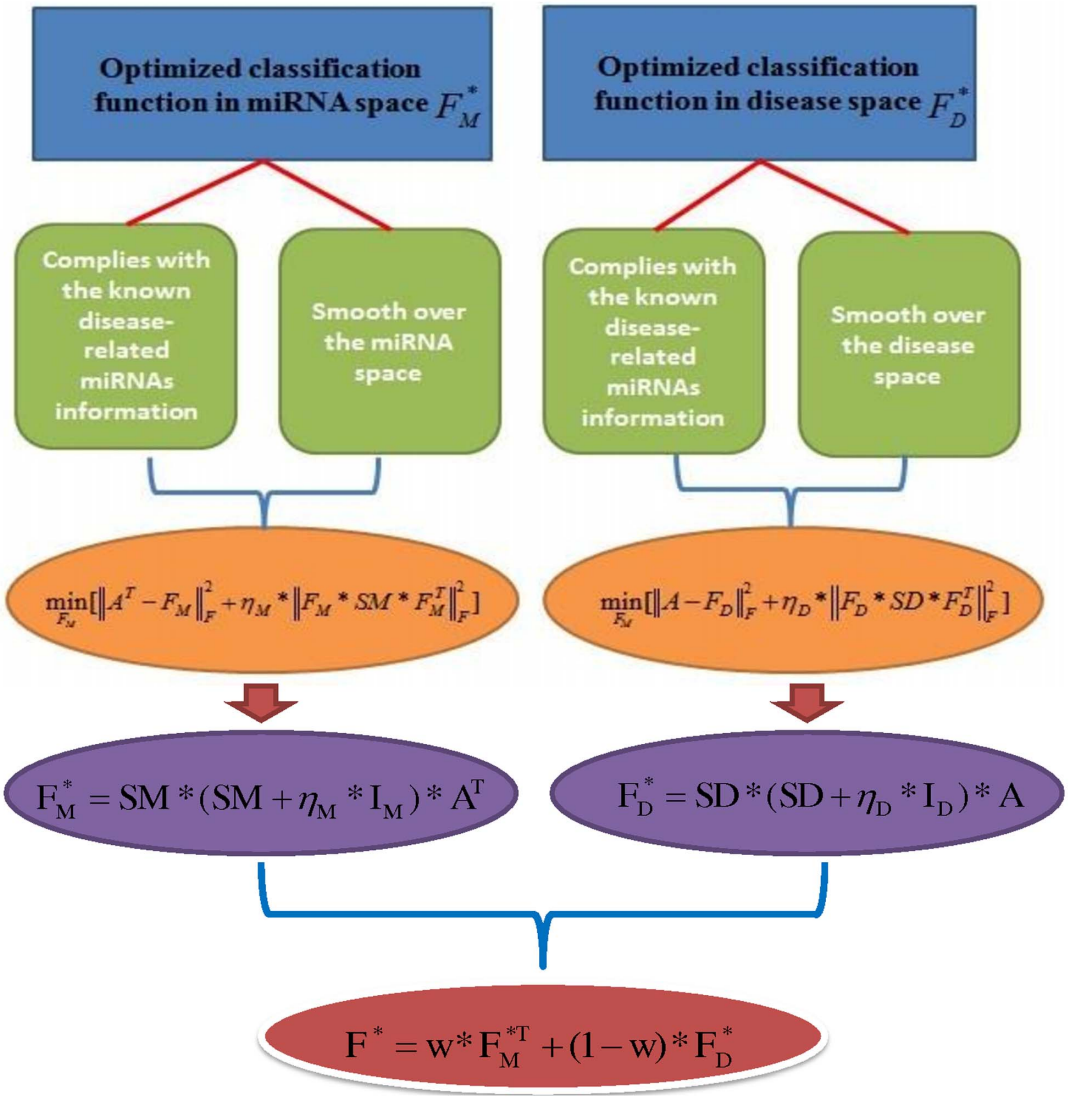


Figure 3 | The flowchart of RLSMDA includes three steps: solving optimization problem; obtaining the optimal classifier in the disease and miRNA space, respectively; combining classifiers in the disease and miRNA space to obtain final predictive result.

Methods

Human miRNA-disease associations. The human miRNA-disease association dataset used as gold standard dataset in this paper was downloaded from the supplementary material of literature[96] (obtained from HMDD in September, 2009). We want to confirm our prediction list based on the update of HMDD and the datasets in other datasets, so we did not use the newest association dataset in HMDD and the datasets in the other databases. The gold standard in this paper includes 1616 distinct high-quality experimentally verified human miRNA-diseases associations. After implementing the operations such as merging different miRNA copies which produce the same mature miRNA and unifying the name of mature miRNAs and diseases, 1395 miRNA–disease associations, including 271 miRNAs and 137 diseases, were used in this paper (see Supplementary table 6). We use *nd* as the number of diseases and *nm* as the number of miRNAs. Matrix *A* is denoted as the adjacency matrix of disease-miRNA associations, where the entity *A(i,j)* in row *i* column *j* is 1 if miRNA *j* is related to the disease *i*, otherwise 0.

MiRNA functional similarity. In the literature[96], functional similarity score for each miRNA pair was calculated based on the assumption that miRNAs with similar functions tend to be related with similar diseases. We downloaded the miRNA functional similarity scores from http://cmbi.bjmu.edu.cn/misim/ in January 2010 (see Supplementary table 7). Matrix *SM* is denoted as the miRNA functional similarity matrix, where the entity *SM(i,j)* in row *i* column *j* is the functional similarity between miRNA *i* and *j*. MiRNA functional similarity used here has been used to predict disease-related miRNAs and environmental factor-miRNA combination interactions and excellent performance have been obtained[47,97].

Disease semantic similarity. Here, we calculated the disease similarity in the same way as literature[96]. The basic idea of disease semantic similarity calculation is illustrated in Figure 2. We can obtain the relationship between diseases from MeSH database (http://www.ncbi.nlm.nih.gov/), which provided a strict system for disease classification. Disease can be described as a DAG, where the nodes represent disease itself and its ancestor diseases and the link from a parent node to a child node represents the relationship between these two nodes. For example, disease *A* can be described as a graph DAG(*A*)=(*A*,T(*A*),E(*A*)), where T(*A*) is the node set including node *A* itself and all ancestor nodes of *A* and E(*A*) is the corresponding links set. The contribution of disease t in DAG(*A*) to the semantics of disease *A* is defined as follows:

$$\begin{cases} D_A(A) = 1 \\ D_A(t) = \max\{\Delta * D_A(t') | t' \in \text{children of } t\} & \text{if } t \neq A \end{cases}. \quad (1)$$

where Δ is the semantic contribution factor. The contribution of disease *A* to its own semantic value is one, while the contributions of other ancestor diseases to the semantic value of disease *A* decrease with the distance between this disease and disease *A*. Therefore, we can define the semantic value of disease *A* based on the contribution of ancestor diseases and disease *A* itself, i.e.

$$DV(A) = \sum_{t \in T(A)} D_A(t). \quad (2)$$

Based on the assumption that disease pairs sharing larger part of their DAGs are more similar, we defined the semantic similarity between two diseases *A* and *B* as follows:

$$SD(A,B) = \frac{\sum_{t \in T(A) \cap T(B)} (D_A(t) + D_B(t))}{DV(A) + DV(B)}. \quad (3)$$

Matrix *SD* is denoted as the disease semantic similarity matrix, where the entity *SD(i,j)* in row *i* column *j* is the disease semantic similarity between disease *i* and *j* (see Supplementary table 8).

Regularized Least Squares for MiRNA-Disease Association (RLSMDA). Based on the underlying assumption that miRNAs associated with more similar diseases are more similar, and vice versa, here we developed the method of Regularized Least Squares for MiRNA-Disease Association (RLSMDA) to uncover the potential miRNAs associated with various diseases (See Figure 3). RLSMDA is designed to construct a continuous classification function which can reflect the probability that each miRNA is related to a given disease. We hope the function can meet the following two criterions: (1) it complies with the known disease-related miRNAs information; (2) it is smooth over the miRNA space and disease space, i.e. for a given disease (miRNA), similar miRNAs (diseases) would obtain similar scores, which meet the basic assumption of our methods. Considering the difficulties of obtaining negative sample, a semi-supervised classifier is constructed under the framework of Regularized Least Squares (RLS), which is obtained by defining a cost function and minimizing this cost function. Cost functions can be developed in miRNA space and disease space, respectively. Taking miRNA space and as an example, optimal classification function can be obtained by solving the following optimization problem:

$$\min_{F_M}[\left\|A^T - F_M\right\|_F^2 + \eta_M * \left\|F_M * SM * F_M^T\right\|_F^2] \quad (4)$$

where $\|\bullet\|_F$ is the Frobenius norm and η_M is the trade-off parameter. The solution of this optimization problem is:

$$F_M^* = SM * (SM + \eta_M * I_M) * A^T \quad (5)$$

where I_M is the identity matrix with the same size as matrix *SM*.

In the similar way, we can obtain the optimal classification function in the disease space as follows:

$$F_D^* = SD * (SD + \eta_D * I_D) * A \quad (6)$$

where I_D is the identity matrix with the same size as matrix *SD*.

Finally, the optimal classifier in two different spaces will be combined to give the final solution based on a simple weighted average operation, i.e.

$$F^* = w * F_M^{*T} + (1 - w) * F_D^* \quad (7)$$

where the entity *F(i,j)* in row *i* column *j* reflect the probability that miRNA *j* is related to the disease *i*.

1. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
2. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
3. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343–349 (2004).
4. Ambros, V. microRNAs: tiny regulators with great potential. *Cell* **107**, 823–826 (2001).
5. Jopling, C. L., Yi, M. K., Lancaster, A. M., Lemon, S. M. & Sarnow, P. Modulation of Hepatitis C Virus RNA Abundance by a Liver-Specific MicroRNA. *Science* **309**, 1577–1581 (2005).
6. Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931–1934 (2007).
7. Cuperus, J., Fahlgren, N. & Carrington, J. Evolution and functional diversification of MIRNA genes. *Plant Cell* **23**, 431–442 (2011).
8. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
9. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901–906 (2000).
10. Pasquinelli, A. E. & Ruvkun, G. Control of developmental timing by microRNAs and their targets. *Annu Rev Cell Dev Biol* **18**, 495–513 (2002).
11. Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140–D144 (2006).
12. Bandyopadhyay, S., Mitra, R., Maulik, U. & Zhang, M. Q. Development of the human cancer microRNA network. *Silence* **1**, 6 (2010).
13. Esquela-Kerscher, A. & Slack, F. J. Oncomirs—microRNAs with a role in cancer. *Nat Rev Cancer* **6**, 259–269 (2006).
14. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
15. Yang, H. *et al.* Evaluation of genetic variants in microRNA-related genes and risk of bladder cancer. *Cancer Res* **68**, 2530 (2008).
16. Cui, Q., Yu, Z., Purisima, E. O. & Wang, E. Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol* **2**, 46 (2006).
17. Friedman, R. C., Farh, K. K. H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105 (2009).
18. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**, 522–531 (2004).
19. Li, J. *et al.* Evidence for Positive Selection on a Number of MicroRNA Regulatory Interactions during Recent Human Evolution. *PLoS Genet* **8**, e1002578 (2012).
20. Chen, K. & Rajewsky, N. Deep conservation of microRNA-target relationships and 3′UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harb Symp Quant Biol* **71**, 149–156 (2006).
21. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
22. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* **38**, 1452–1456 (2006).
23. Saunders, M. A., Liang, H. & Li, W. H. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci USA* **104**, 3300 (2007).
24. Sethupathy, P. & Collins, F. S. MicroRNA target site polymorphisms and human disease. *Trends Genet* **24**, 489–497 (2008).
25. Karp, X. & Ambros, V. Enhanced: encountering microRNAs in cell fate signaling. *Science* **310**, 1288–1289 (2005).
26. Cheng, A. M., Byrom, M. W., Shelton, J. & Ford, L. P. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res* **33**, 1290–1297 (2005).
27. Miska, E. A. How microRNAs control cell division, differentiation and death. *Curr Opin Genet Dev* **15**, 563–568 (2005).
28. Xu, P., Guo, M. & Hay, B. A. MicroRNAs and the regulation of cell death. *Trends Genet* **20**, 617–624 (2004).
29. Latronico, M. V. G., Catalucci, D. & Condorelli, G. Emerging role of microRNAs in cardiovascular biology. *Circ Res* **101**, 1225–1236 (2007).
30. Lu, M. *et al.* An analysis of human microRNA and disease associations. *PLoS One* **3**, e3420 (2008).
31. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**, 857–866 (2006).
32. Duisters, R. F. *et al.* miR-133 and miR-30 Regulate Connective Tissue Growth Factor. *Circ Res* **104**, 170–178 (2009).
33. Markou, A. *et al.* Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non–small cell lung cancer by quantitative real-time RT-PCR. *Clin Chem* **54**, 1696–1704 (2008).
34. Miller, T. E. *et al.* MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. *J Biol Chem* **283**, 29897–29903 (2008).
35. Slack, F. J. & Weidhaas, J. B. MicroRNA in cancer prognosis. *N Engl J Med* **359**, 2720–2722 (2008).
36. Weinberg, M. S. & Wood, M. J. A. Short non-coding RNA biology and neurodegenerative disorders: novel disease targets and therapeutics. *Hum Mol Genet* **18**, R27–R39 (2009).
37. Huang, Q. *et al.* The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. *Nat Cell Biol* **10**, 202–210 (2008).
38. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
39. Xin, F. *et al.* Computational analysis of microRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. *Bioinformatics* **25**, 430–434 (2009).
40. Xu, J. *et al.* Prioritizing Candidate Disease miRNAs by Topological Features in the miRNA Target–Dysregulated Network: Case Study of Prostate Cancer. *Mol Cancer Ther* **10**, 1857–1866 (2011).
41. Yu, Z. *et al.* Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. *Nucleic Acids Res* **35**, 4535–4541 (2007).
42. Xiao, Y. *et al.* Prioritizing cancer-related key miRNA–target interactions by integrative genomics. *Nucleic Acids Res* **40**, 7653–7665 (2012).
43. Poy, M. N. *et al.* A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226–230 (2004).
44. Van Es, H. H. G. & Arts, G. J. Biology calls the targets: combining RNAi and disease biology. *Drug Discov Today* **10**, 1385–1391 (2005).
45. Jiang, Q. *et al.* Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst Biol* **4**, S2 (2010).
46. Jiang, Q., Wang, G., Jin, S., Li, Y. & Wang, Y. Predicting human microRNA–disease associations based on support vector machine. *Int J Data Min Bioinform* **8**, 282–293 (2013).
47. Chen, X., Liu, M. X. & Yan, G. RWRMDA: predicting novel human microRNA-disease associations. *Mol Biosyst* **8**, 2792–2798 (2012).
48. Jiang, Q. *et al.* miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* **37**, D98–D104 (2009).
49. Yang, Z. *et al.* dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC genomics* **11**, S5 (2010).
50. Jiang, Q., Wang, G. & Wang, Y. An approach for prioritizing disease-related microRNAs based on genomic data integration. *BMEI* **6**, 2270–2274 (2010).
51. Shi, H. *et al.* Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *BMC Syst Biol* **7**, 101 (2013).
52. Xuan, P. *et al.* Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. *PLoS One* **8**, e70204 (2013).
53. Chen, H. & Zhang, Z. Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med Genomics* **6**, 12 (2013).
54. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043 (2011).

55. Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci USA* **101**, 2999–3004 (2004).
56. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci USA* **103**, 2257–2261 (2006).
57. Calin, G. A. *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* **353**, 1793–1801 (2005).
58. Feitelson, M. A. & Lee, J. Hepatitis B virus integration, fragile sites, and hepatocarcinogenesis. *Cancer Lett* **252**, 157 (2007).
59. Xu, Y. *et al.* A potentially functional polymorphism in the promoter region of miR-34b/c is associated with an increased risk for primary hepatocellular carcinoma. *Int J Cancer* **128**, 412–417 (2011).
60. Cahill, S. *et al.* Effect of BRAFV600E mutation on transcription and post-transcriptional regulation in a papillary thyroid carcinoma model. *Mol Cancer* **6**, 21 (2007).
61. Baffa, R. *et al.* MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. *J Pathol* **219**, 214–221 (2009).
62. Wang, Y. X. *et al.* Initial study of microRNA expression profiles of colonic cancer without lymph node metastasis. *J Dig Dis* **11**, 50–54 (2010).
63. Widera, C. *et al.* Diagnostic and prognostic impact of six circulating microRNAs in acute coronary syndrome. *J Mol Cell Cardiol* **51**, 872–875 (2011).
64. Maegdefessel, L. *et al.* MicroRNA-21 blocks abdominal aortic aneurysm development and nicotine-augmented expansion. *Sci Transl Med* **4**, 122ra122–122ra122 (2012).
65. Jones, J. A. *et al.* Selective MicroRNA Suppression in Human Thoracic Aneurysms Relationship of miR-29a to Aortic Size and Proteolytic Induction. *Circ Cardiovasc Genet* **4**, 605–613 (2011).
66. Chatzikyriakidou, A., Voulgari, P., Georgiou, I. & Drosos, A. The Role of microRNA-146a (miR-146a) and its Target IL-1R-Associated Kinase (IRAK1) in Psoriatic Arthritis Susceptibility. *Scand J Immunol* **71**, 382–385 (2010).
67. Paraskevi, A. *et al.* Circulating MicroRNA in inflammatory bowel disease. *J Crohns Colitis* **6**, 900–904 (2012).
68. Cao, P. *et al.* Comprehensive expression profiling of microRNAs in laryngeal squamous cell carcinoma. *Head Neck* **35**, 720–728 (2013).
69. Fei, J., Li, Y., Zhu, X. & Luo, X. miR-181a post-transcriptionally downregulates oncogenic RalA and contributes to growth inhibition and apoptosis in chronic myelogenous leukemia (CML). *PLoS One* **7**, e32834 (2012).
70. Sharma, A. D. *et al.* MicroRNA-221 regulates FAS-induced fulminant liver failure. *Hepatology* **53**, 1651–1661 (2011).
71. Hai-yan, W., Yang, L., Mei-hong, C. & Hui, Z. Expression of MicroRNA-146a in peripheral blood mononuclear cells in patients with systemic lupus erythematosus. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* **33**, 185–188 (2011).
72. Balatti, V. *et al.* MicroRNAs dysregulation in human malignant pleural mesothelioma. *J Thorac Oncol* **6**, 844–851 (2011).
73. Cai, C.-K. *et al.* miR-15a and miR-16-1 downregulate CCND1 and induce apoptosis and cell cycle arrest in osteosarcoma. *Oncol Rep* **28**, 1764–1770 (2012).
74. Xu, X. *et al.* Microarray-based analysis: identification of hypoxia-regulated microRNAs in retinoblastoma cells. *Int J Oncol* **38**, 1385–1393 (2011).
75. Narducci, M. *et al.* MicroRNA profiling reveals that miR-21, miR486 and miR-214 are upregulated and involved in cell survival in Sezary syndrome. *Cell Death Dis* **2**, e151 (2011).
76. Cheng, Y. & Zhang, C. MicroRNA-21 in cardiovascular disease. *J Cardiovasc Transl Res* **3**, 251–255 (2010).
77. Weng, L. *et al.* Dysregulation of miRNAs in AL amyloidosis. *Amyloid* **18**, 128–135 (2011).
78. Teruel, R. *et al.* Identification of miRNAs as potential modulators of tissue factor expression in patients with systemic lupus erythematosus and antiphospholipid syndrome. *J Thromb Haemost* **9**, 1985–1992 (2011).
79. Villar, A. V. *et al.* Myocardial and circulating levels of microRNA-21 reflect left ventricular fibrosis in aortic stenosis patients. *Int J Cardiol* **167**, 2875–2881 (2013).
80. Wang, J. *et al.* [Differential expressions of miRNAs in patients with nonvalvular atrial fibrillation]. *Zhonghua Yi Xue Za Zhi* **92**, 1816–1819 (2012).
81. Lukiw, W., Dua, P., Pogue, A., Eicken, C. & Hill, J. Upregulation of micro RNA-146a (miRNA-146a), a marker for inflammatory neurodegeneration, in sporadic Creutzfeldt–Jakob disease (sCJD) and Gerstmann–Straussler–Scheinker (GSS) syndrome. *J Toxicol Environ Health A* **74**, 1460–1468 (2011).
82. Dong, P. *et al.* MicroRNA-194 inhibits epithelial to mesenchymal transition of endometrial cancer cells by targeting oncogene BMI-1. *Mol Cancer* **10**, 99 (2011).
83. Jin, J. *et al.* Interrogation of brain miRNA and mRNA expression profiles reveals a molecular regulatory network that is perturbed by mutant huntingtin. *J Neurochem* **123**, 477–490 (2012).
84. Danielsson, K., Wahlin, Y.-B., Gu, X., Boldrup, L. & Nylander, K. Altered expression of miR-21, miR-125b, and miR-203 indicates a role for these microRNAs in oral lichen planus. *J Oral Pathol Med* **41**, 90–95 (2012).
85. Thapa, D. R. *et al.* B-cell activation induced microRNA-21 is elevated in circulating B cells preceding the diagnosis of AIDS-related non-Hodgkin lymphomas. *AIDS* **26**, 1177 (2012).
86. Pekow, J. R. *et al.* miR-143 and miR-145 are downregulated in ulcerative colitis: Putative regulators of inflammation and protooncogenes. *Inflamm Bowel Dis* **18**, 94–100 (2012).
87. Bhattacharyya, S. *et al.* Elevated miR-155 promotes inflammation in cystic fibrosis by driving hyperexpression of interleukin-8. *J Biol Chem* **286**, 11604–11615 (2011).
88. Tan, Z., Liu, F., Tang, H. & Su, Q. [Expression and its clinical significance of hsa-miR-155 in serum of endometrial cancer]. *Zhonghua Fu Chan Ke Za Zhi* **45**, 772–774 (2010).
89. Qin, W. *et al.* TGF-β/Smad3 signaling promotes renal fibrosis by inhibiting miR-29. *J Am Soc Nephrol* **22**, 1462–1474 (2011).
90. Gao, W. *et al.* Plasma levels of lipometabolism-related miR-122 and miR-370 are increased in patients with hyperlipidemia and associated with coronary artery disease. *Lipids Health Dis* **11**, 55 (2012).
91. Hughes, A. E. *et al.* Mutation altering the miR-184 seed region causes familial keratoconus with cataract. *Am J Hum Genet* **89**, 628–633 (2011).
92. Maj, J., Jankowska-Konsur, A., Sadakierska-Chudy, A., Noga, L. & Reich, A. Altered microRNA expression in mycosis fungoides. *Br J Dermatol* **166**, 331–336 (2012).
93. Yang, C. C. *et al.* miR-181 as a putative biomarker for lymph-node metastasis of oral squamous cell carcinoma. *J Oral Pathol Med* **40**, 397–404 (2011).
94. Wang, Y. *et al.* MiR-133a in human circulating monocytes: a potential biomarker associated with postmenopausal osteoporosis. *PLoS One* **7**, e34641 (2012).
95. Maegdefessel, L. *et al.* MicroRNA-21 blocks abdominal aortic aneurysm development and nicotine-augmented expansion. *Sci Transl Med* **4**, 122ra122 (2012).
96. 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).
97. Chen, X., Liu, M. X., Cui, Q. H. & Yan, G. Y. Prediction of Disease-Related Interactions between MicroRNAs and Environmental Factors Based on a Semi-Supervised Classifier. *PloS one* **7**, e43425 (2012).

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 11301517, 10531070, 10721101, 11371355 and National Center for Mathematics and Interdisciplinary Sciences, CAS.

Author contributions

X.C. and G.Y.Y. conceived the prediction method and wrote the paper. X.C. developed the prediction method, conceived, designed and implemented the experiments, analyzed the result. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chen, X. & Yan, G.-Y. Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* **4**, 5501; DOI:10.1038/srep05501 (2014).

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