Efficient framework for predicting MiRNA-disease associations based on improved hybrid collaborative filtering

Ru Nie1,2†, Zhengwei Li1,2,3,4*,†, Zhu-hong You5*, Wenzheng Bao6 and Jiashu Li1,2

From Fifteenth International Conference on Intelligent Computing (ICIC 2019) Nanchang, China. 3-6 August 2019

Abstract

Background: Accumulating studies indicates that microRNAs (miRNAs) play vital roles in the process of development and progression of many human complex diseases. However, traditional biochemical experimental methods for identifying disease-related miRNAs cost large amount of time, manpower, material and financial resources.

Methods: In this study, we developed a framework named hybrid collaborative filtering for miRNA-disease association prediction (HCFMDA) by integrating heterogeneous data, e.g., miRNA functional similarity, disease semantic similarity, known miRNA-disease association networks, and Gaussian kernel similarity of miRNAs and diseases. To capture the intrinsic interaction patterns embedded in the sparse association matrix, we prioritized the predictive score by fusing three types of information: similar disease associations, similar miRNA associations, and similar disease-miRNA associations. Meanwhile, singular value decomposition was adopted to reduce the impact of noise and accelerate predictive speed.

Results: We then validated HCFMDA with leave-one-out cross-validation (LOOCV) and two types of case studies. In the LOOCV, we achieved 0.8379 of AUC (area under the curve). To evaluate the performance of HCFMDA on real diseases, we further implemented the first type of case validation over three important human diseases: Colon Neoplasms, Esophageal Neoplasms and Prostate Neoplasms. As a result, 44, 46 and 44 out of the top 50 predicted disease-related miRNAs were confirmed by experimental evidence. Moreover, the second type of case validation on Breast Neoplasms indicates that HCFMDA could also be applied to predict potential miRNAs towards those diseases without any known associated miRNA.

Conclusions: The satisfactory prediction performance demonstrates that our model could serve as a reliable tool to guide the following research for identifying candidate miRNAs associated with human diseases.
Introduction

MicroRNAs (miRNAs) are endogenous small non-coding RNAs (19–22 nucleotides) that could regulate gene expression by base-pairing to partially complementary mRNAs [1]. Since the first miRNA, lin-4, was discovered by Lee et al. in 1993 [2], more than 38,000 miRNA sequences from 271 organisms have been accumulated to date [3]. Plenty of evidence indicates that miRNAs play critical roles in many fundamental and important biological processes, such as immune response, transcription, proliferation and differentiation [4]. The mutation and dysregulated expression of miRNAs may be connected with the development and progression of many diseases [5, 6]. For instance, miR-155 downregulated target gene TP53INP1 whose expression was strongly reduced in pancreatic ductal adenocarcinoma development [7]. Besides, induction of endogenous miR-340 expression was capable to suppress tumor cell migration and invasion, whereas miR-340 knockdown led to breast cancer cell migration and invasion [8]. Moreover, three most up-regulated miRNAs (miR-221, 222, and 146) distinguished unequivocally between papillary thyroid carcinoma and normal thyroid [9]. Therefore, exploring the relationships between miRNAs and diseases could not only provide novel insights into disease pathogenesis at the molecular level, but also benefit the design of specific molecular tools for disease diagnosis, treatment and prevention [10, 11].

However, traditional in vivo or biochemical experiment for identifying disease-related miRNA candidates have multiple bottlenecks, such as long operation time, extremely high cost and false positive results [12, 13]. Consequently, quickly and automatically identifying these associations with in silico methods is a useful supplement for future experimental validation and could substantially reduce the cost and effort [14–16]. Actually, based on the generally accepted assumption that functionally similar miRNAs are likely to be associated with phenotypically similar diseases and vice versa, a large number of computational models have been proposed for identifying potential disease-related miRNAs in recent years. For example, Jiang et al. [17] explored a network-based computational model through hypergeometric distribution to prioritize disease-related miRNAs. Shi et al. [18] focused on the functional connections between miRNA targets and disease genes in protein–protein interaction (PPI) networks and presented a novel method to identify disease-related miRNAs. Xu et al. [19] introduced an approach based on MTDN for prioritizing putative miRNAs associated with diseases by combining paired miRNA and mRNA expression data. In addition, Chen et al. [20] constructed HGIMDA framework to identify potential disease-related miRNAs by combing multiple source information, e.g., experimentally validated miRNA-disease relationship, disease semantic similarity, miRNA functional similarity, Gaussian interaction profile kernel similarity. Based on the same multiple source data, Chen et al. [21] also presented PRMDA to infer potential disease-related miRNAs by personalized recommendation-based algorithm. Although HGIMDA and PRMDA could be applied to those diseases without experimentally validated miRNA, the predictive accuracy needs to be further enhanced. In addition, Marissa et al. [22] presented an in-silico method named MAP for predicting putative miRNA-disease associations through network diffusion on multi-omics biological data including miRNA-gene associations, protein–protein interactions, and gene-disease associations, and so on. Yu et al. [23] proposed TCRWMDA for miRNA-disease association prediction through three-layer heterogeneous network combined with unbalanced random walk. The case study results indicate that TCRWMDA is an effective tool to predict the potential miRNA-disease associations. Li et al. [24], developed a novel method named NIM-CGNC which employs graph convolutional networks to extract feature representations and a neural inductive matrix completion model to generate association matrix completion. Experimental results indicate this method could be used for predicting those diseases without any known related miRNAs.

Collaborative filtering aims at predicting the user interest for a given item based on a collection of user profiles and there are already some basic applications in miRNA-disease associations (MDAs) prediction [25–27]. However, these approaches generally fail to achieve satisfactory results. In this work, we developed a computational approach named hybrid collaborative filtering for predicting miRNA-disease associations (HCFMDA) to infer putative associations between diseases and miRNAs. By fusing experimentally verified MDAs, disease similarity, and miRNA functional similarity to mine intrinsic discriminative information embedded in the correlations between diseases and miRNAs, HCFMDA could be applied for identifying
potential miRNAs for those diseases without any known related miRNA. In the leave-one-out cross-validation (LOOCV), HCFMMDA achieved AUC (area under the curve) value of 0.8379 and demonstrated reliable predictive performance. In addition, we also used HCFMMDA to carry out two types of case validation on four important human complex diseases (Colon Neoplasms, Esophageal Neoplasms, Prostate Neoplasms and Breast Neoplasms). As a result, 88%, 92%, 88% and 92% out of the top 50 putative miRNAs for those 4 diseases were confirmed by experiment evidence. All the results indicate HCFMMDA is effective and reliable for the prediction of MDAs.

Materials and methods

Human miRNA-disease associations

The experimentally verified human miRNA-disease associations for HCFMMDA were retrieved from HMDD v2.0 database [28]. After the data preprocessing and verification, we obtained altogether 5430 experimentally validated associations between 383 diseases and 495 miRNAs. Then, we constructed an adjacency matrix tally validated associations between 383 diseases and their respective children. The semantic value of disease pair $d_i$ and $d_j$ could be calculated as

$$DSV(d_i) = \sum_{d_j \in V(d_i)} D_{d_i}(d_j)$$

where $D_{d_i}(d_j)$, as the semantic contribution value of disease $d_i$ to $d_j$, could be calculated as

$$D_{d_i}(d_j) = \begin{cases} 1, & \text{if } d_j = d_i \\ \max\{ \rho \cdot D_{d_i}(d_j'), d_j' \in \text{children of } d_j \}, & \text{otherwise} \end{cases}$$

where $\rho$ is a contribution factor. The semantic contribution value of disease $d_i$ to $d_j$ is inversely proportional to the distance between them in the DAG.

Based on the idea that two diseases will be more similar if their DAGs overlaps more nodes, we constructed semantic similarily matrix $SD$ for those diseases. Each element of $SD$ denotes the semantic similarity of disease pair $\langle d_k, d_i \rangle$, which could be calculated as following:

$$SD(d_k, d_i) = \frac{\sum_{t \in V(d_k) \cap V(d_i)} (D_{d_k}(t) + D_{d_i}(t))}{DSV(d_k) + DSV(d_i)}$$

Gaussian kernel similarity for diseases and miRNAs

Gaussian kernel similarity comes from the topological distribution of the experimentally verified MDAs. Herein, we introduced binary vector $BV(d_i)$ as the interaction profile for disease $d_i$, which is the $ith$ row of the adjacent matrix $X$. Hence, Gaussian kernel similarity of disease pair $\langle d_i, d_j \rangle$ could be expressed by

$$GaussianKernel(d_i, d_j) = e^{-\frac{\|BV(d_i) - BV(d_j)\|^2}{2\sigma^2}}$$
which could be generated through averaging the interaction profiles of all diseases.

\[
\delta_d = \frac{1}{N_d} \sum_{i=1}^{N_d} BV(d_i)
\]

In the same manner, Gaussian kernel similarity between miRNA pair \( \langle m_i, m_j \rangle \) could be defined as follows:

\[
GD(d_i, d_j) = \exp(-\delta_d BV(d_i) - BV(d_j))
\]

where \( \delta_d \) is a parameter for adjusting kernel bandwidth, which could be generated through averaging the interaction profiles of all diseases.

\[
\delta_d = \frac{1}{N_d} \sum_{i=1}^{N_d} BV(d_i)
\]

In the same manner, Gaussian kernel similarity between miRNA pair \( \langle m_i, m_j \rangle \) could be defined as follows:

\[
GM(m_i, m_j) = \exp(-\delta_m BV(m_i) - BV(m_j))
\]

where \( BV(m_i) \) is the interaction profile for miRNA \( m_i \) and \( \delta_m \) is used to control kernel bandwidth.

Integrated similarities for miRNAs and diseases

Following the previous steps, the miRNA functional similarity, disease semantic similarity and Gaussian kernel similarity were generated separately. To cope with data sparsity and effectively utilize all kinds of similarities and correlations, we further constructed an integrated similarity matrix \( ID \) for diseases and \( IM \) for miRNAs respectively, which could be expressed as follows:

\[
ID(d_i, d_j) = \begin{cases} 
SD(d_i, d_j), & \text{if } SSD(d_i, d_j) > 0 \\
GD(d_i, d_j), & \text{otherwise}
\end{cases}
\]

\[
IM(m_i, m_j) = \begin{cases} 
FM(m_i, m_j), & \text{if } FSM(m_i, m_j) > 0 \\
GM(m_i, m_j), & \text{otherwise}
\end{cases}
\]

To ensure the equal importance, each similarity value should be normalized to the same interval before integration. The specific approach is to subtract the mean value and divide it by the standard deviation of the corresponding matrix.

HCFMDA

In this work, by integrating heterogeneous data including the miRNA functional similarity, disease semantic similarity, known MDA networks, Gaussian kernel similarity of miRNAs and diseases, we proposed a pipeline named improved hybrid collaborative filtering to predict miRNA-disease associations (HCFMDA). Figure 1 illustrates the flowchart of the entire process of HCFMDA. In order to fully exploit highly discriminative feature information embedded in the sparse MDAs, we incorporated and fused three different association sources: different diseases associated with the same miRNA, different miRNAs associated with the same disease, and ‘not-so-similar’ diseases or miRNAs, which could make the model more robust to data sparsity.

Disease-based collaborative filtering predicted the score \( \hat{x}_{ij} \) of a test miRNA \( m_i \) for a disease \( d_j \) based on top-\( N \) most similar diseases towards \( d_j \). Consequently, the corresponding recommended score \( \hat{x}_{ij} \) could be represented by

\[
\hat{x}_{ij} = \frac{1}{N} \sum_{d_i \in T_d(d_j)} s_d(d_i, d_j)x_{u=d_i} | T_d(d_i) | = N
\]

where \( T_d(d_i) \) and \( s_d(d_i, d_j) \) represent a set of top-\( N \) most similar diseases towards disease \( d_j \) and the integrated similarity value of disease pair \( \langle d_i, d_j \rangle \), respectively. It could be seen from Fig. 2a that this method only exploits the known associations between the test miRNA and similar diseases of \( d_j \), which accounts for only a small part in the matrix. We denoted this predictive source as the set of similar disease associations (SDA):

\[
SDA_{ij} = \{ X_{u=d_i} | d_u \in T_d(d_j) \}
\]

Similarly, the predictive score \( \hat{x}_{ij} \) of test pair \( \langle d_i, m_j \rangle \) of miRNA-based collaborative filtering could also be calculated by averaging the associations of similar miRNAs related to the same disease \( d_i \).

\[
\hat{x}_{ij} = \frac{1}{N} \sum_{m_v \in T_m(m_j)} s_m(m_j, m_v)x_{i=v} | T_m(m_j) | = N
\]

where \( T_m(m_j) \) and \( s_m(m_j, m_v) \) denote a set of top-\( N \) most similar miRNAs towards miRNA \( m_j \) and the integrated similarity between miRNA \( m_j \) and \( m_v \), respectively. As illustrated in Fig. 2b, Eq. (15) only exploits the known similar miRNAs associated with the test disease for prediction. We refer to these predictive sources as the set of similar miRNA associations (SMA):

\[
SMA_{ij} = \{ X_{i=v} | m_v \in T_m(m_j) \}
\]

In practice, solely relying on such SDA or SMA is undesirable, particularly when the association adjacent matrix \( X \) is very sparse. The predictive accuracy could be improved by incorporating more associations from those ‘not-so-similar’ diseases or miRNAs. As illustrated in Fig. 2c, those associations from ‘not-so-similar’ diseases or miRNAs could provide additional information to improve the prediction. In this work, we refer to this
predictive source as similar disease-miRNA associations (SDMA):

\[ SDMA_{ij} = \{ X_{u,v} | d_u \in T_d(i_j), m_v \in T_m(m_j), u \neq i, v \neq j \} \]

\[ \hat{x}_{i,j} = \frac{1}{K} \sum_{X_{u,v} \in T_{d,m}(X_{i,j})} s_{d,m}(X_{i,j}, X_{u,v})X_{u,v} | T_{d,m}(X_{i,j})| = K \]

(16)

(17)

where \( T_{d,m}(X_{i,j}) \) denotes a set of top-K most similar miRNA-disease pairs. Here, we constructed \( s_{d,m}(X_{i,j}, X_{u,v}) \) as the similarity between entity \( X_{i,j} \) and \( X_{u,v} \).

\[ s_{d,m}(X_{i,j}, X_{u,v}) = \frac{1}{\sqrt{\left(\frac{1}{s_d(d_i,d_u)}\right)^2 + \left(\frac{1}{s_m(m_j,m_v)}\right)^2}} \]

(18)

Each element of the matrix \( X \) was employed as a separate predictor, whose confidence could be calculated according to its similarity towards the test association. We then predicted the expected value of the test association by averaging the individual predictions weighted by their confidence.

Finally, we calculated the expected value of the unknown test association \( X_{i,j} \) by the following equation:
where $P(X_{ij}|SDA, SMA, SDMA)$ denotes the estimating conditional probability depending on the predictors coming from the pool of SDA, SMA and SDMA. Likewise, $P(X_{ij}|SDA)$, $P(X_{ij}|SMA)$, and $P(X_{ij}|SDMA)$ represent the pool of SDA, SMA and SDMA predictors, respectively. $\alpha$ and $\beta$ were used to control the selection (sampling) of data from those three different sources. If $\beta$ is equal to 1, HCFMDA only uses SDMA recommend score to predict potential miRNAs for given diseases. In addition, to remove noise and accelerate the operation speed, singular value decomposition (SVD) technique was applied in HCFMDA.

**Performance evaluation**

In practice, there are only 5430 experimentally verified MDAs (i.e., known associations) from HMDD V2.0 [28] and therefore most elements ($383 \times 495 - 5430 = 184155$) in adjacent matrix $X$ are zeros, which indicates that $X$ is very sparse and so it is not feasible to adopt multi-fold cross-validation to test the performance of our method. As demonstrated in a series of studies [31–37], leave-one-out cross-validation (LOOCV) is more rigorous and objective than independent dataset test and K-fold cross-validation. Therefore, we implemented LOOCV to validate the performance of HCFMDA. In the LOOCV, for a designated disease $d_i$, each known $d_i$-related association was left out in turn as a test sample and all other known associations (in total 5429) were used to train the model. Therefore, other miRNAs irrelevant to the disease $d_i$ along with the test miRNA were treated as candidate miRNAs. Then, we sorted all candidate miRNAs by the predictive scores derived from our model in descending order. If the rank of the test association exceeded a given threshold, we could view it as a successful identification. Receiver-operating characteristics (ROC) curve is a fundamental evaluation tool to illustrate diagnostic ability of a binary classifier. The ROC curve is generated by plotting true positive rate (TPR) against false positive rate (FPR) at different cut-off points. The corresponding formulas are as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where $TP$, $FN$, $FP$ and $TN$ represents true positive, false negative, false positive and true negative, respectively. More specifically, $TP$ represents the number of known MDAs (positive samples) predicted correctly, and $FN$ is the number of positive samples that are falsely predicted to unknown MDAs (negative samples). Similarly, $FP$ denotes the number of negative samples incorrectly predicted to positive samples while $TN$ stands for the number of negative samples predicted correctly.

In addition, we calculated AUC to evaluate the predictive performance of HCFMDA. The AUC is equivalent to the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one. Specifically, AUC = 1 represents a perfect test, while AUC = 0.5 means a worthless test.
Results

Experimental results performed by HCFMDA
To verify the performance of HCFMDA to identify disease-related miRNAs, LOOCV was employed as testing strategy based on HMDD V2.0 dataset. For parameters $\alpha$ and $\beta$, we adopted grid search strategy to search their optimal values ($\alpha = 0.3, \beta = 0.1$). Figure 3 illustrates the AUC values performed by HCFMDA on HMDD 2.0 by varying $\alpha$ from 0.1 to 1.0 with the step of 0.2, where $\beta$ is set to 0.1. It indicates that the fusion of multi-similarity measurement could enhance the prediction performance of HCFMDA model.

Then, we compared HCFMDA with other five state-of-the-art computational approaches: HGIMDA [20], WBSMDA [38], EGBMMDA [39], PRMDA [21] and DRMDA [40]. Figure 4 illustrates the AUC values comparison of those prediction models in the same framework of LOOCV. Our model achieved AUC of 0.8379, while the AUC values from HGIMDA, WBSMDA, EGBMMDA, PRMDA and DRMDA were 0.8077, 0.8031, 0.8221, 0.8315 and 0.8339, respectively. Because HCFMDA could capture high discriminative information embedded in the correlations between miRNAs and diseases, it achieved the superior predictive performance compared with the other five methods. Besides, the ROC curve of HCFMDA is smoother than those of other methods, which reflects HCFMDA is more robust and accurate. In addition, the time complexity of our model is obvious lower than the other methods. We attribute it to the introduction of SVD matrix decomposition that could reduce the impact of noise and improve the predictive speed. In conclusion, HCFMDA demonstrates reliable performance for predicting MDAs.

Case validation 1: for diseases with known associated miRNAs
Apart from the validation of LOOCV, we also carried out case studies for several important human diseases to further verify the performance of HCFMDA. Predictive results of HCFMDA were confirmed by matching experimentally verified MDAs (i.e., known associations) from two other independent databases: miR2Disease [41] and dbDEMC [42]. Herein, we adopted two types of case validations. At first, we applied HCFMDA over three different diseases, i.e. Colon Neoplasms (CN), Esophageal Neoplasms (EN), and Prostate Neoplasms (PN). For a given disease, each time a known association in HMDD v2.0 was left out as a test sample and all unknown associations were taken as the candidate miRNAs, while all other known associations were used for training samples. In the second type of case validation for Breast Neoplasms (BN), we removed all known associations related to BN in HMDD v2.0 and then employed HCFMDA to predict potential BN-related miRNAs. The key point of this type of case validation is to ensure that the prioritization of putative miRNAs related to BN only makes use of the correlations of other phenotypic diseases similar to BN. Therefore, it could be used to demonstrate whether our model is applicable to those isolated diseases without any known related miRNAs.
As the most common type of gastrointestinal cancer, CN is the second-leading cause of cancer-related deaths in the USA. It was estimated that about 140,250 people was diagnosed with CN, and 50,630 died from the disease in 2018 [43]. Since patients in early stage of CN have only mild symptoms and are hard to be detected, there is an urgent demand of novel diagnostic biomarkers for its early detection. Fortunately, a significant number of CN-related miRNAs have been successfully identified in the past decades. For instance, Ma et al. [44] reported that the up-regulated miRNAs in CN including miR-182, miR-17, miR-106a, miR-93, miR-200c, miR-92a, let-7a and miR-20a (FDR value < 5%), while the down-regulated miRNAs were miR-195, miR-143 and miR-145 (FDR value < 5%). Moreover, Marta et al. [45] identified and validated a signature of 6 miRNAs (miRNA19a, miRNA19b, miRNA15b, miRNA29a, miRNA335, and miRNA18a) as biomarkers that could differentiate significantly CN patients from those healthy people. In the case validation for CN, we adopted HCFMDA to prioritize the top 50 miRNAs from candidate miRNAs (Table 1). We found that 9 out of the top 10 candidate miRNAs could be supported to be associated with CN by the experimental evidence. Besides, 88% of top 50 prioritized miRNAs were validated to be involved with CN. For example, many experiments [46] reported that the expression level of has-miR-20a (ranked No.1 in Table 1) was significantly higher in CN tissue than those in the normal adjacent mucosa and suggests that it can be taken as a novel prognostic marker and therapeutic target for CN. Also, miR-18a (ranked No.2 in Table 1) was also confirmed by experiments and could play an important role in CN pathogenesis [47]. Actually, some of the rest unconfirmed miRNAs in Table 1 were also confirmed by recent studies. For example, has-miR-92 directly targeted the anti-apoptotic molecule BCL-2-interacting mediator of cell death (BIM) in colon cancer tissues and was recently proposed as a key oncogenic component of miR-17–92 cluster through targeting and down-regulating the proapoptotic protein Bim in CN [48]. In addition, Antonio et al. [49] reported that has-miR-101 expression was differentially impaired in CN specimens and it might function as a tumor suppressor in CN and that its pharmacological restoration might hamper the aggressive behavior of CRC in vivo.

Esophageal neoplasms (EN), or esophageal cancer, occupies the sixth position among malignant tumors worldwide with regard to mortality and ranks fourth in China [50]. Due to lack of effective clinical diagnosis approaches for EN, it is often diagnosed at a more advanced stage and its overall 5-year survival rate is only about 25% [51]. Therefore, investigating the mechanism of EN is seriously essential to improve its diagnosis, treatment and prognosis. Numerous recent studies have indicated that aberrant expression of miRNAs is involved in EN. For instance, Hu et al. [50] identified that miR-375 was downregulated in tumor tissue and cell line EC109 of EN samples when compared with normal tissues and cells. Experiments confirmed that as a tumor suppressor

| Table 1 Predictive results of the top 50 prioritized miRNAs related to CN based on experimentally verified associations in HMDD v2.0 database |
| miRNA | Evidence | miRNA | Evidence |
|-------|----------|-------|----------|
| hsa-mir-20a | dbDEMC, miR2D-disease | hsa-mir-92a | Unconfirmed |
| hsa-mir-18a | dbDEMC, miR2D-disease | hsa-mir-141 | dbDEMC, miR2D-disease |
| hsa-mir-19b | dbDEMC, miR2D-disease | hsa-mir-214 | dbDEMC |
| hsa-mir-191 | dbDEMC, miR2D-disease | hsa-mir-30c | dbDEMC, miR2D-disease |
| hsa-mir-143 | dbDEMC, miR2D-disease | hsa-mir-93 | dbDEMC, miR2D-disease |
| hsa-mir-132 | miR2Disease | hsa-mir-34c | miR2Disease |
| hsa-mir-29b | dbDEMC, miR2D-disease | hsa-mir-21 | dbDEMC, miR2D-disease |
| hsa-mir-19a | dbDEMC, miR2D-disease | hsa-mir-25 | dbDEMC, miR2D-disease |
| hsa-mir-34a | dbDEMC, miR2D-disease | hsa-mir-194 | dbDEMC, miR2D-disease |
| hsa-mir-101 | Unconfirmed | hsa-mir-32 | dbDEMC, miR2D-disease |
| hsa-let-7e | dbDEMC | hsa-mir-92b | Unconfirmed |
| hsa-let-7d | dbDEMC | hsa-mir-205 | dbDEMC |
| hsa-let-7a | dbDEMC, miR2D-disease | hsa-let-7 g | dbDEMC, miR2D-disease |
| hsa-mir-200b | dbDEMC | hsa-mir-222 | dbDEMC |
| hsa-mir-127 | dbDEMC, miR2D-disease | hsa-mir-203 | dbDEMC, miR2D-disease |
| hsa-mir-125b | dbDEMC | hsa-mir-146a | dbDEMC |
| hsa-mir-199a | Unconfirmed | hsa-mir-34b | dbDEMC, miR2D-disease |
| hsa-mir-223 | dbDEMC, miR2D-disease | hsa-mir-16 | dbDEMC |
| hsa-let-7b | dbDEMC, miR2D-disease | hsa-mir-429 | dbDEMC |
| hsa-mir-125a | dbDEMC, miR2D-disease | hsa-mir-221 | dbDEMC, miR2D-disease |
| hsa-mir-155 | dbDEMC, miR2D-disease | hsa-mir-200a | Unconfirmed |
| hsa-let-7c | dbDEMC | hsa-mir-146b | Unconfirmed |
| hsa-mir-106b | dbDEMC, miR2D-disease | hsa-mir-29a | dbDEMC, miR2D-disease |
| hsa-let-7f | dbDEMC, miR2D-disease | hsa-mir-95 | dbDEMC, miR2D-disease |
| hsa-let-7i | dbDEMC | hsa-mir-373 | dbDEMC |

The column 1 and 3 list the top 1–25 and top 26–50 CN-related miRNAs, respectively.
in EN cells, miR-375 inhibited cell proliferation and invasion by repressing the expression of its direct target MTDH, an oncogene associated with tumorigenesis in EN. Herein, we took EN as a case validation and prioritized the candidate miRNAs of the disease. As illustrated in Table 2, all the top 10 predicted miRNAs associated with EN were successfully verified by experimental evidence collected from the two independent databases. Meanwhile, 46 out of the top 50 predicted miRNAs were also validated to be related to EN. For example, recent studies indicated that miR-195 was down-regulated in EN tissues compared with normal esophageal tissues ($P = 0.05$) and experimental results indicated that Cdc42 protein was reduced after miR-195 mimics transfected ($P = 0.01$) [52]. In addition, Zhang et al. [53] first presented that tanshinone IIA inhibited human EN cell growth through miR-122-mediated Pyruvate kinase M2 (PKM2) down-regulation pathway.

Prostate neoplasms (PN) is the most common malignancy and the third leading cancer-related cause of death among men in the western world. Although the 5-year survival rate of PN is higher in early-stage after treatment with surgical resection or androgen deprivation therapy, one-third of treated PN patients will experience disease recurrence and progress into castration-resistant PN, a more aggressive disease [54]. Therefore, an impressing need exists to identify novel miRNAs as tools or biomarkers for the prediction of aggressive PN. In the case validation for PN by HCFMDA, 8 of top 10 miRNAs and 44 out of top 50 candidate PN-associated miRNAs were validated by the two independent databases (see Table 3). Moreover, 4 of the rest 6 unsupported miRNAs

### Table 2: Predictive results of the top 50 predicted miRNAs related to EN based on known associations in HMDD v2.0 database

| miRNA     | Evidence | miRNA     | Evidence |
|-----------|----------|-----------|----------|
| hsa-mir-17 | dbDEMC   | hsa-mir-24 | dbDEMC   |
| hsa-mir-18a| dbDEMC   | hsa-mir-10b| dbDEMC   |
| hsa-mir-19b| dbDEMC   | hsa-mir-30c| dbDEMC   |
| hsa-mir-125b| dbDEMC  | hsa-mir-30a| dbDEMC   |
| hsa-mir-221| dbDEMC   | hsa-mir-181a| dbDEMC  |
| hsa-mir-16 | dbDEMC   | hsa-mir-15b| dbDEMC   |
| hsa-mir-29a| dbDEMC   | hsa-mir-93 | dbDEMC   |
| hsa-mir-200b| dbDEMC  | hsa-mir-106a| dbDEMC  |
| hsa-mir-106b| dbDEMC  | hsa-mir-18b | dbDEMC   |
| hsa-let-7d  | dbDEMC   | hsa-mir-132 | dbDEMC   |
| hsa-mir-1   | dbDEMC   | hsa-mir-23b | dbDEMC   |
| hsa-let-7i  | dbDEMC   | hsa-mir-122 | unconfirmed |
| hsa-let-7f  | unconfirmed | hsa-mir-194 | dbDEMC,miR2Disease |
| hsa-let-7e  | dbDEMC   | hsa-mir-7   | dbDEMC   |
| hsa-mir-222 | dbDEMC   | hsa-mir-218 | unconfirmed |
| hsa-mir-29b | dbDEMC   | hsa-mir-127 | dbDEMC   |
| hsa-mir-429 | dbDEMC   | hsa-mir-302c| dbDEMC   |
| hsa-mir-181b| dbDEMC   | hsa-mir-199b| dbDEMC   |
| hsa-mir-142 | dbDEMC   | hsa-mir-135a| dbDEMC   |
| hsa-mir-125a| dbDEMC   | hsa-mir-193b| dbDEMC   |
| hsa-mir-182 | dbDEMC   | hsa-mir-20b | dbDEMC   |
| hsa-let-7g  | dbDEMC   | hsa-mir-302b| dbDEMC   |
| hsa-mir-195 | dbDEMC   | hsa-mir-107 | dbDEMC,miR2Disease |
| hsa-mir-146b| dbDEMC   | hsa-mir-204 | unconfirmed |
| hsa-mir-9   | dbDEMC   | hsa-mir-23a | dbDEMC   |

The column 1 and 3 list the top 1–25 and top 26–50 EN-related miRNAs, respectively.

### Table 3: Predictive results of the top 50 prioritized miRNAs related to PN based on known associations in HMDD v2.0 database

| miRNA     | Evidence | miRNA     | Evidence |
|-----------|----------|-----------|----------|
| hsa-mir-21 | dbDEMC,miR2Disease | hsa-mir-200a | dbDEMC   |
| hsa-mir-155| dbDEMC   | hsa-mir-23a | dbDEMC,miR2Disease |
| hsa-let-7a | dbDEMC,miR2Disease | hsa-mir-106b | dbDEMC   |
| hsa-mir-146a| miR2Disease | hsa-mir-19b | dbDEMC,miR2Disease |
| hsa-mir-17  | miR2Disease | hsa-mir-24 | dbDEMC,miR2Disease |
| hsa-mir-20a | miR2Disease | hsa-let-7b | dbDEMC,miR2Disease |
| hsa-mir-143 | dbDEMC,miR2Disease | hsa-mir-223 | dbDEMC,miR2Disease |
| hsa-mir-18a | 24752237 | hsa-mir-34a | dbDEMC,miR2Disease |
| hsa-let-7c  | dbDEMC,miR2Disease | hsa-mir-15a | dbDEMC,miR2Disease |
| hsa-mir-92a | 29568403 | hsa-let-7i | dbDEMC   |
| hsa-let-181b | dbDEMC,miR2Disease | hsa-mir-200b | dbDEMC   |
| hsa-let-7f  | dbDEMC,miR2Disease | hsa-mir-25 | dbDEMC,miR2Disease |
| hsa-mir-19a | dbDEMC   | hsa-mir-142 | unconfirmed |
| hsa-mir-1   | dbDEMC,miR2Disease | hsa-mir-141 | miR2Disease |
| hsa-mir-9   | dbDEMC   | hsa-mir-222 | dbDEMC,miR2Disease |
| hsa-mir-126 | dbDEMC,miR2Disease | hsa-let-7g | dbDEMC,miR2Disease |
| hsa-let-7e  | dbDEMC   | hsa-mir-29c | dbDEMC   |
| hsa-mir-221 | dbDEMC,miR2Disease | hsa-mir-125a | dbDEMC,miR2Disease |
| hsa-let-7d  | dbDEMC,miR2Disease | hsa-mir-203 | 21159887 |
| hsa-mir-16  | dbDEMC,miR2Disease | hsa-mir-106a | dbDEMC,miR2Disease |
| hsa-mir-150 | dbDEMC   | hsa-mir-133a| dbDEMC   |
| hsa-mir-29a | dbDEMC,miR2Disease | hsa-mir-34b | dbDEMC   |
| hsa-mir-93  | 26124181 | hsa-mir-34c | dbDEMC   |
| hsa-mir-210 | miR2Disease | hsa-mir-27a | dbDEMC,miR2Disease |
| hsa-mir-200c | dbDEMC   | hsa-mir-15b | dbDEMC   |

The column 1 and 3 list the top 1–25 and top 26–50 PN-related miRNAs, respectively. The evidences for the associations are either database studies or PMIDs of other experimental literatures.
were verified by recent studies. For example, Williams et al. [55] identified miR-200b as a downstream target of androgen receptor and linked its expression to decreased tumorigenicity and metastatic capacity of the prostate cancer cells. In addition, as an "antimetastatic" miRNA in PN, miR-203 expression is specifically attenuated in bone metastatic prostate cancer suggesting a fundamental antimetastatic role for this miRNA [56].

Besides, we further made the comparison between HCFMDA and other five aforementioned MDA prediction models in terms of the case studies of diseases CN, EN and PN. It could be seen from Table 4 that HCFMDA ranks second best among all predictive models for those three diseases. The predictive hit rate of HCFMDA is only lower than that of PRMDA in the case study of CN and EN, and that of EGBMMDA in the case study of PN, which fully demonstrates that HCFMDA could be used as a reliable tool for predicting disease-related miRNAs.

**Case validation 2: for diseases without known associated miRNAs**

To further validate the predictive performance of HCFMDA for those diseases without any known related miRNA, we also implemented another type of

| miRNA | Evidence  | miRNA | Evidence  |
|-------|-----------|-------|-----------|
| hsa-mir-367 | dbDEMC,HMDD | hsa-mir-608 | dbDEMC,HMDD |
| hsa-mir-302c | dbDEMC,HMDD | hsa-mir-638 | dbDEMC,HMDD |
| hsa-mir-302a | dbDEMC,HMDD | hsa-mir-518b | unconfirmed |
| hsa-mir-302b | dbDEMC,HMDD | hsa-mir-602 | dbDEMC |
| hsa-mir-488 | HMDD | hsa-mir-612 | dbDEMC |
| hsa-mir-215 | dbDEMC,HMDD | hsa-mir-615 | dbDEMC |
| hsa-mir-302d | dbDEMC,HMDD | hsa-mir-637 | dbDEMC |
| hsa-mir-218 | dbDEMC,HMDD | hsa-mir-657 | dbDEMC |
| hsa-mir-383 | dbDEMC,HMDD | hsa-mir-185 | dbDEMC |
| hsa-let-7d | dbDEMC,miR2Disease,HMDD | hsa-mir-518c | dbDEMC |
| hsa-let-7f | dbDEMC,miR2Disease,HMDD | hsa-mir-622 | dbDEMC |
| hsa-let-7c | dbDEMC,miR2Disease,HMDD | hsa-mir-583 | dbDEMC |
| hsa-mir-19a | dbDEMC,HMDD | hsa-mir-557 | dbDEMC |
| hsa-mir-153 | dbDEMC,HMDD | hsa-mir-600 | dbDEMC |
| hsa-let-7b | dbDEMC,HMDD | hsa-mir-601 | dbDEMC |
| hsa-let-7i | dbDEMC,miR2Disease,HMDD | hsa-mir-611 | unconfirmed |
| hsa-mir-296 | dbDEMC,HMDD | hsa-mir-654 | dbDEMC |
| hsa-mir-17 | dbDEMC,HMDD | hsa-mir-662 | dbDEMC |
| hsa-mir-799 | dbDEMC,miR2Disease,HMDD | hsa-mir-799 | unconfirmed |
| hsa-mir-429 | dbDEMC,miR2Disease,HMDD | hsa-mir-18a | dbDEMC,miR2Disease,HMDD |
| hsa-mir-338 | dbDEMC,HMDD | hsa-mir-486 | dbDEMC,HMDD |
| hsa-let-7g | dbDEMC,HMDD | hsa-mir-629 | dbDEMC,HMDD |
| hsa-mir-20a | miR2Disease,HMDD | hsa-mir-596 | unconfirmed |
| hsa-mir-19b | dbDEMC,HMDD | hsa-mir-17 | miR2Disease,HMDD |
| hsa-mir-324 | HMDD | hsa-mir-339 | dbDEMC,HMDD |

The column 1 and 3 list the top 1–25 and top 26–50 BN-related miRNAs, respectively.
case validation for Breast Neoplasms (BN) by removing all the known BN-related associations in HMDD v2.0. That is to say, we only utilized the known associations of other diseases except BN and adopted the indirect way to predict BN-related miRNAs. We then ranked all the 495 candidate miRNAs by their predictive scores and verified the top 50 ones according to the databases of dbDEMC, miR2Disease and HMDD v2.0. The predictive results (Table 5) indicated that all top 10 and 46 out of the top 50 prioritized miRNAs were confirmed by those databases. The achieved results indicate that HCFMDA could also be applied to predict novel miRNAs for those isolated diseases.

Discussion
Identification of novel disease-related miRNAs is beneficial for understanding disease pathogenesis at the molecular level, and developing effective disease diagnostic biomarkers and therapeutic tools. In this work, we proposed an efficient computational framework, HCFMDA, to improve the predictive performance of MDAs by integrating heterogeneous information: miRNA functional similarity, disease semantic similarity, known MDA networks, Gaussian kernel similarity of miRNAs and diseases. HCFMDA employs not only traditional disease-based and miRNA-based associations, but also associations from other ‘not-so-similar’ diseases and miRNAs to smooth the predictions. We then implemented LOOCV and two types of case validations over four important human cancers. The achieved results demonstrate that HCFMDA is indeed robust against data sparsity, which is better than other five state-of-the-art models, i.e., HGIMDA, WBSMDA, EGBMMDA, PRMDA, and DRMDA.

Conclusions
The excellent performance of HCFMDA mainly attributes to the following aspects. First, many kinds of heterogeneous data including miRNA functional similarity, disease semantic similarity, and known MDAs were integrated into our model, which contains highly discriminative information. Second, by fusing three kinds of similar associations including disease-based associations, miRNA-based associations, and other ‘not-so-similar’ diseases and miRNAs associations our model could fully mine and capture the intrinsic associations between miRNAs and diseases even if the MDA matrix is very sparse. Although some favorable results have been made, there still exists several limitations in HCFMDA. First, there are only 5430 known MDAs among 383 diseases and 495 miRNAs, and therefore the corresponding MDA matrix is very sparse and needs to be further enriched. Second, although we have integrated some heterogeneous data into our model, there is still room for improving the performance of HCFMDA by integrating more effective data sources which could provide more useful information for predicting MDAs. Moreover, we will improve the efficiency of our model by introducing graph-based recommendation filtering algorithms in the future.

Abbreviations
MDA: MiRNA-disease association; HCF: Hybrid collaborative filtering; LOOCV: Leave-one-out cross-validation; BV: Binary vector; DAG: Directed acyclic graph; SDA: Similar disease associations; SMA: Similar miRNA associations; SDMA: Similar disease-miRNA associations; SVD: Singular value decomposition; ROC: Receiver operating characteristic; AUC: Area under curve; FPR: False positive rate; TPR: True positive rate; TP: True positive; FN: False negative; FP: False positive; TN: True negative.

Acknowledgements
We thank the editor and the anonymous reviewers for their comments and suggestions.

About the supplement
This article has been published as part of BMC Medical Informatics and Decision Making Volume 21 Supplement 1: Proceedings of the 2019 International Conference on Intelligent Computing (ICIC 2019): medical informatics and decision making. The full contents of the supplement are available online at https://bmcmedicalinformaticsanddecisionmaking.biomedcentral.com/articles/supplements/volume-21-supplement-1.

Authors’ contributions
RN and ZL designed the study, prepared the data sets and wrote the manuscript. ZY, WB, and JL designed, performed and analyzed experiments. All authors have read and approved the final manuscript.

Funding
This work is supported by the Science and technology innovation 2030—”new generation artificial intelligence” major project (No. 2018aa0100) and the National Natural Science Foundation of China (Nos. 61873270, 61732012, 61902337). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank the anonymous reviewers for their constructive advices.

Availability of data and materials
The datasets used and/or analyzed during the current study are available at https://github.com/ivantsinghua/HCFMDA.

Declarations
Ethics approval and consent to participate
Not applicable.
Consent for publication
Not applicable.
Competing interests
The authors declare that they have no competing interests.

Author details
1Engineering Research Center of Mine Digitalization of Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China. 2School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. 3Institute of Machine Learning and Systems Biology, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China. 4KUNPAND Communications (Kunshan) Co., Ltd., Suzhou 215300, China. 5School of Computer Science, Northwestern
44. Qin M, Lie Y, Cun W, Yong-Yang Y, Bin Z, Zhong-Guang Z. Differential expression of colon cancer microRNA in microarray study. J Sichuan Univ. 2011;42(3):344–8.
45. Herreros Villanueva M, Duran S, Carmen Martin A, Pérez-Palacios R, WlANavarro E, Marcuello M, et al. Plasma MicroRNA signature validation for early detection of colorectal cancer. 2019.
46. Zhang GJ, Li Y, Zhou H, Xiao HK, Zhou T. miR20a is an independent prognostic factor in colorectal cancer and is involved in cell metastasis. Mol Med Rep. 2014;10(1):283–91.
47. Anna Brunet V, Carles P, Irene M, Anna F, Emma D, Aleydis P, et al. microRNA expression profile in stage III colorectal cancer: circulating miR-18a and miR-29a as promising biomarkers. Oncol Rep. 2013;30(1):320–6.
48. Schee K, Boye K, Abrahamsen TW, Fodstad Ø, Flatmark K. Clinical relevance of microRNA miR-21, miR-31, miR-92a, miR-101, miR-106a and miR-145 in colorectal cancer. BMC Cancer. 2012;12(1):505.
49. Antonio S, Maria Chiara V, Pasquale S, Cinzia C, Annamaria S, Laura V, et al. Loss of miR-101 expression promotes Wnt/β-catenin signalling pathway activation and malignancy in colon cancer cells. J Pathol. 2013;229(3):379–89.
50. Hu C, Lv L, Peng J, Liu D, Wang X, Zhou Y, et al. MicroRNA-375 suppresses esophageal cancer cell growth and invasion by repressing metadherin expression. Oncol Lett. 2017;13(6):4769–75.
51. Guo Y, Chen Z, Zhang L, Zhou F, Shi S, Feng X, et al. Distinctive MicroRNA profiles relating to patient survival in esophageal squamous cell carcinoma. Can Res. 2008;68(1):26–33.
52. Fu M, Li S, Yu T, Qian L, Gao S, Hong Z, et al. The expression of miR-195 in esophageal cancer and its effect on proliferation of esophageal cancer cell lines. Acta Univ Med Nanjing. 2013.
53. Zhang H-S, Zhang F-J, Li H, Liu Y, Du G-Y, Huang Y-H. Tanshinone IIA inhibits human esophageal cancer cell growth through miR-122-mediated PKM2 down-regulation. Arch Biochem Biophys. 2016;598:50–6.
54. Xin Z, Jin Feng W. Prognostic role of microRNA-145 in prostate cancer: a systems review and meta-analysis. Prost Int. 2015;3(3):71–4.
55. Williams LT, Dorina V, Elena V, Volpert OV. miR-200b inhibits prostate cancer EMT, growth and metastasis. PLoS ONE. 2013;8(12):e83991.
56. Saini S, Majid S, Yamamura S, Tabatabai L, Suh SO, Shahryari V, et al. Regulatory role of mir-203 in prostate cancer progression and metastasis. Clin Cancer Res. 2011;17(16):5287–98.

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