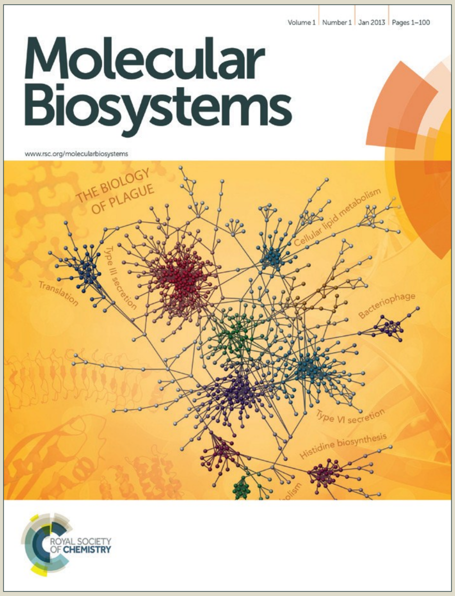
Molecular BioSystems

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

Molecular BioSystems

Dynamic Article Links ►

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/molecularbiosystems

PAPER

miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method†‡

Xing Chen*[a,b]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Increasing evidence has indicated that microRNAs (miRNAs) can functionally interact with environmental factors (EFs) to affect and determine human diseases. Uncovering the potential associations between diseases and miRNA-EF interactions could benefit the understanding of the underlying disease mechanism at miRNA and EF levels, miRNA signatures identification, and drug repurposing. In this study, based on the assumption that similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease and the framework of Random Walk with Restart (RWR), I developed a novel method of miREFRWR to uncover the hidden disease-related miRNA-EF interactions by implementing random walks on miRNA similarity network and EF similarity network, respectively. miREFRWR was evaluated by leave-one-out cross validation and achieved an AUC of 0.9500. It has been demonstrated that miREFRWR can effectively identify potential interactions in all the test classes, even if those test samples only share either EFs or miRNAs with the training samples. Furthermore, plenty of predictive results for acute promyelocytic leukemia and breast cancer (67 and 10 interactions out of the top 1% predictions, respectively) were verified by independent experimental studies. It is anticipated that miREFRWR could be a useful and important biological resources for the biomedical research.

Introduction

Phenotypes and diseases are often determined by the complex interactions between genetic factors (GFs) and environmental factors (EFs) [1-4]. As one class of important and newly identified GFs and one of the most important components of the cell, microRNAs (miRNAs) play critical roles in many important biological processes, including cell growth, proliferation, differentiation, apoptosis, signal transduction, viral infection and so on [5-11]. Accumulating evidences have indicated that miRNAs are associated with various diseases [5, 12-20]. One typical example is insulin secretion can be regulated by mir-375 [21, 22]. Numerous miRNAs have been linked with the initiation and development of various cancers [23]. For example, Huang et al (2008) confirmed that the upregulation of miR-373 and miR-520c to be significant players in tumour invasion and metastasis [24]. Therefore, the interactions between miRNAs and EFs may contribute to the development or treatment of many phenotypes and diseases.

Recently, increasing studies have indicated that miRNAs can functionally interact with plenty of EFs to affect and determine the phenotypes and diseases. Related EFs include drugs [25], alcohol [26], cigarette [27], stress [28], diet [29], virus [30], air pollution [31], radiation [32] and so on. For example, cigarette smoke condensate (CSC) could lead to cancer by dramatically increasing the expression level of mir-31 and hence activating LOC554202 in normal respiratory epithelia and lung cancer cells [33]. More importantly, the interactions between miRNA and EF also can benefit the disease treatment. For instance, during clinical treatment of ovarian cancer, Paclitaxel could significantly decrease the expression of mir-29c [34]. In the breast cancer treatment, 3,3'-Diindolylmethane (DIM) could inhibit the proliferation of breast cancer cell by increasing miR-21 expression and hence causing the downregulation of Cdc25A [35]. Therefore, identifying potential disease-related miRNA-EF interactions based on computational methods has become an important problem in the biomedical research and played critical roles in disease pathogenesis understanding at the miRNA-EF interactions level, miRNA signatures identification for given EFs, and new indication inference of approved drugs. Computational prediction has been an important complementary method for disease-related interactions identification, which can select promising disease-related interactions for further experiment validation, hence decrease the time and cost of biological experiments[36-42].

Yang et al (2011) manually collected experimentally supported disease-related miRNA-EF interactions and further constructed miREnvironment database, which included more than 2500 entries about ~800 miRNAs, ~260 EFs, ~180 phenotypes, and 17 species [43]. Qiu et al (2012) analyzed disease-related human miRNA-EF interactions in the miREnvironment database and obtained some important conclusions about the association patterns of miRNA-EF interactions [44]. Those conclusions indicated that miRNA-EF interactions had a significant correlation with the characteristics such as miRNA expression level, tissue specificity, conservation, and disease spectrum width. They further developed several methods for EF relationship characterization, cancer treatment result prediction, and novel EF-disease interactions inference. Although these proposed methods cannot predict ternary relationships among miRNAs, EFs, and diseases together simultaneously, they laid a theoretical foundation for disease-related miRNA-EF interactions prediction research. In my previous work, I proposed the similar nature of disease-related miRNA-EF interactions, i.e. similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease [45]. Based on this assumption and the framework of semi-supervised classifier, I developed a semi-supervised classifier based method (miREFScan) to predict potential disease-related interactions between miRNAs and EFs. Reliable performance has been obtained in both cross validation and case study about acute promyelocytic leukemia (APL) [45]. To my knowledge, miREFScan is the first computational tool for simultaneous ternary relationships prediction among miRNAs, EFs, and diseases.

However，little efforts have been made to analyze and predict potential disease-related miRNA-EF interactions from a network perspective. Network medicine could effectively predict the potential interactions among biological molecules, investigate how cellular systems induce different biological phenotypes under different conditions, and provide a novel approach to understand the complicated mechanism of disease and drug treatment[1]. Especially, network-based computational models have been widely used to predict disease-related genes, miRNAs, long non-coding RNAs (lncRNAs) and drug-target interactions[36, 46-50]. Therefore, it is fundamental and important to understand the mechanisms of complex diseases and identify new indications of drugs in a network-centric perspective. In this study, I developed a novel method of miREFRWR (miRNA-EF interactions inference based on the Random Walk with Restart) to infer potential disease-related miRNA–EF interactions by making full use of the tool of the network for data integration to predict potential associations. It consists of four steps: firstly, three networks (miRNA–miRNA similarity network, EF-EF similarity network, and known miRNA–EF interaction network for a given disease) are constructed; secondly, random walks are implemented on the miRNA similarity network and EF similarity network, respectively; Then, predictive results based on random walk on miRNA similarity network and EF similarity network are combined to obtain the final predictive results; finally, the most probable miRNA-EF pairs are selected according to the stable probability of the random walk. In the framework of leave-one-out cross validation (LOOCV), miREFRWR obtained the comparable performance (AUC=0.9500) with miREFScan in my previous work. It has also been demonstrated that miREFRWR had a reliable performance in all the test classes, even if the test samples only shared either EFs or miRNAs with the training samples. In the case studies about APL and breast cancer, miREFRWR further showed the advantages of making full use of network information to predict potential interactions. Especially in the APL-related miRNA-EF interactions prediction, sixty-seven interactions out of the top 1% predictions based on miREFRWR have been confirmed by experimental literatures. I further applied miREFRWR to predict potential novel miRNA-EF interactions for all the investigated diseases in the dataset. The top 100 interactions for each disease have been publicly released for further biological experiment validation.

Methods

Disease-related miRNA-EF interactions

Firstly, the whole dataset of known disease-related miRNA-EF interactions was downloaded from miREnvironment database (http://cmbi.bjmu.edu.cn/miren, Version of September, 2011) [43], including more than 2500 entries and each entry was composed of a miRNA name, an EF name, and their related phenotype/disease. This database is a very important and useful biological resource for the research about the mutual relationship among miRNAs, EFs, and diseases and lays the data foundation for disease-related miRNA-EF interactions identification. Secondly, human disease-related miRNA-EF interactions were extracted from the above dataset and double-checked to get rid of the entries with phenotype named ”n/a”. Furthermore, the names of diseases, miRNAs, and EFs were normalized and 862 distinct human disease-related miRNA-EF interactions were obtained, which contained the information about 418 miRNAs, 138 EFs, and 97 diseases (see Supplementary Table 1). This dataset was regarded as golden standard dataset in the cross validation and case studies for performance evaluation. Finally, disease-related miRNA-EF interaction adjacency matrix *A* was constructed for each given disease, where the entity *A(i,j)* in row *i* and column *j* is 1 if EF *j* could interact with miRNA *i* and their interaction could contribute to the given disease, otherwise 0.

Chemical structure similarity between EFs

In the previous drug research [45, 46, 51-58], chemical structure was widely applied to effectively evaluate drug similarity. Considering the fact that plenty of EFs are drugs in the known disease-related miRNA-EF interactions dataset, chemical structure similarity matrix *SCE* was constructed (here, C denotes chemical structure and E denotes EF) for EFs based on the tool of SIMCOMP [59] and the drug chemical structure information derived from various databases, such as KEGG database [60], PubChem [61], and ChemicalBook (http://www.chemicalbook.com/). The similarity score computed in this way is a global ratio between the size of common structures and union structures of two drugs based on a graph alignment algorithm. The entity *SCE(i,j)* in the row *i* column *j* is the chemical structure similarity score between EF *i* and *j* if they are both drugs, otherwise 0.

Functional similarity between miRNAs

Based on the assumption that miRNAs with similar functions are more likely to be related with similar diseases [12], Wang et al (2010) represented the relationships among different diseases by directed acyclic graph (DAG) and further inferred miRNA functional similarity by calculating the similarity between their associated disease DAGs [62]. Here, miRNA functional similarity scores were calculated by the tool of MISIM in May 2011 (http://cmbi.bjmu.edu.cn/misim/) [62]. Then, miRNA functional similarity matrix *SFM* was constructed (M denotes miRNA and F denotes functional similarity), where the entity *SFM(i,j)* in row *i* column *j* is the functional similarity score between miRNA *i* and *j*. Previous studies have shown that functional similarity scores calculated in this way coincided well with prior knowledge about miRNA function annotations [62]. In addition, miRNA functional similarity network has played critical roles in disease-related miRNAs and miRNA-EF interactions identification [36, 37, 45].

Network-based similarity for miRNAs and EFs

Considering the fact that some EFs are not drugs and some miRNAs don’t have any known associated diseases, hence the similarity scores for these EFs and miRNAs can’t be obtained based on aforementioned chemical structure similarity and functional similarity. Here, network-based similarity for miRNAs and EFs is proposed to improve the traditional similarity measure (see Figure 1). Disease-miRNA, disease-EF, and EF-miRNA interactions network can be obtained from known disease-related miRNA-EF interactions, respectively. Based on the observation that EFs interacting with more common miRNAs or diseases tend to be more similar, network-based EF similarity matrix *SME* and *SDE* (here, E denotes EF, M (D) indicates network-based similarity is obtained from EF-miRNA (disease) interactions) are constructed by extracting the information from disease-EF and EF-miRNA interactions networks, respectively, where the entity *SME(i.j)* (*SDE(i,j*) in row *i* column *j* is the number of miRNAs (diseases) shared by EF *i* and *j* in EF-miRNA (disease) interactions network. Correspondingly, from disease-miRNA and miRNA-EF interactions network, network-based miRNA similarity matrix *SDM* and *SEM* (here, M denotes miRNA, D (E) indicates network-based similarity is obtained from disease (EF)-miRNA interactions) can be obtained in the similar way, where the entity *SDM (i,j)* (*SEM (i,j)*) in row *i* column j is the number of diseases (EFs) shared by miRNA *i* and *j*. Network-based similarity must be normalized. Taking *SME* as an example, corresponding normalized matrix is defined as follows:

$$\overline{SME} = (DME)^{-1/2}\, SME (DME)^{-1/2}$$

where diagonal matrix *DME* is defined such that *DME(i,i)* is the sum of the *ith* row of *SME*. Other three network-based similarity matrix are also normalized in the similar way. To avoid circular design and optimistic prediction performance report of LOOCV, network-based miRNA similarity and EF similarity was recalculated when each cross validation run was implemented, i.e. the information of tested disease-related miRNA-EF interactions was discarded from known disease-related miRNA-EF interaction network and only current training dataset was used to calculate network-based similarity.

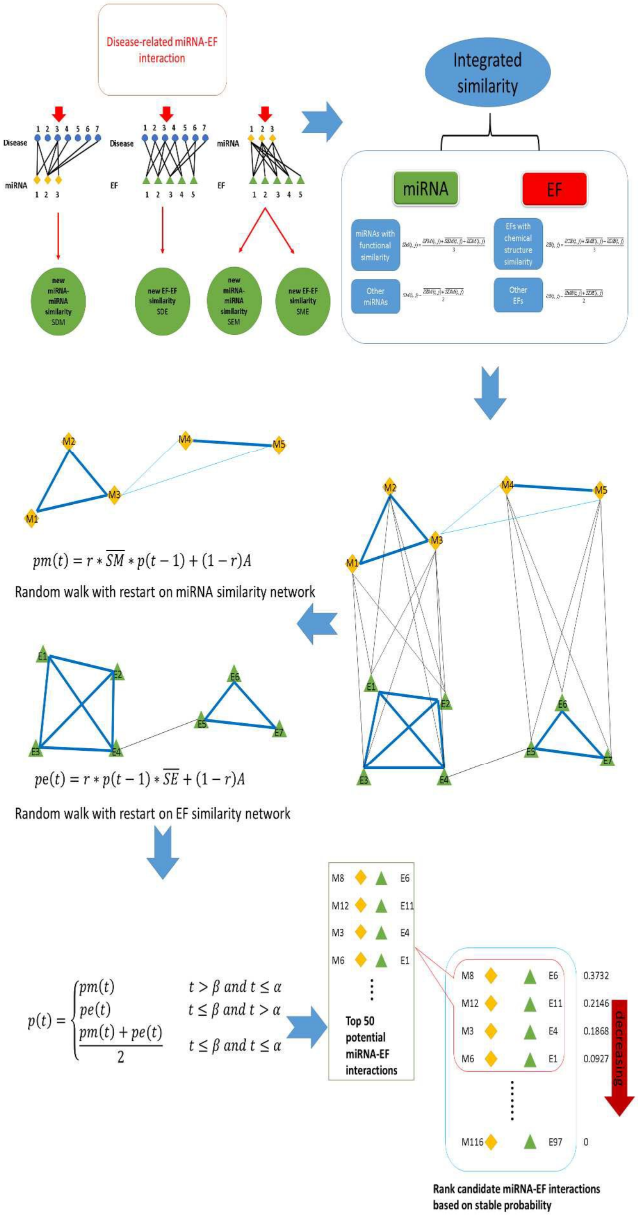

Disease-related miRNA-EF interaction
Integrated similarity
Disease
miRNA
1 2 3 4 5 6 7
1 2 3
Disease
EF
1 2 3 4 5 6 7
1 2 3 4 5
miRNA
EF
1 2 3
1 2 3 4 5
new miRNA-miRNA similarity SDM
new EF-EF similarity SDE
new miRNA-miRNA similarity SEM
new EF-EF similarity SME
miRNA
EF
miRNAs with functional similarity
Other miRNAs
EFs with chemical structure similarity
Other EFs
M1
M2
M3
M4
M5
$pm(t) = r * \overline{SM} * p(t-1) + (1-r)A$
Random walk with restart on miRNA similarity network
E1
E2
E3
E4
E5
E6
E7
$pe(t) = r * p(t-1) * \overline{SE} + (1-r)A$
Random walk with restart on EF similarity network
$p(t) = \begin{cases} pm(t) & t > \beta \text{ and } t \le \alpha \\ pe(t) & t \le \beta \text{ and } t > \alpha \\ \frac{pm(t)+pe(t)}{2} & t \le \beta \text{ and } t \le \alpha \end{cases}$
M8 E6
M12 E11
M3 E4
M6 E1
Top 50 potential miRNA-EF interactions
M8 E6 0.3732
M12 E11 0.2146
M3 E4 0.1868
M6 E1 0.0927
M116 E97 0
decreasing
Rank candidate miRNA-EF interactions based on stable probability

Fig.1 Flowchart of miREFRWR. This flowchart provides a brief description of new method developed in this paper. Step 1: Calculating network-based miRNA similarity and EF similarity. Step 2: Calculating integrated miRNA similarity and EF similarity. Step 3: Constrcting heterogeneous network. Here, a simple example is provided. The upper network is the miRNA similarity network and the lower network is the EF similarity network. They are connected into a heterogeneous network by known miRNA-EF interactions. Step 4: Implementing random walks on miRNA similarity network and EF similarity network and introducing two parameters to restrict the iteration steps of random walks on these two networks, respectively. Step 5: Combining predictive results based on random walks on miRNA similarity network and EF similarity network to obtain final predictive results. Step 6: Ranking all the candidate miRNA-EF interactions based on stable probability and selecting potential disease-related miRNA-EF interactions for experimental validation.

Integrated similarity for miRNAs and EFs

Based on aforementioned drug chemical structure similarity and network-based EF similarity, integrated similarity matrix *SE* for EFs can be constructed based on trivial combinatorial coefficients (see Figure 1), where the entity *SE(i,j)* in row *i* column *j* is defined as follows:

$$SE(i,j) = \begin{cases} \dfrac{SCE(i,j) + \overline{SME(i,j)} + \overline{SDE(i,j)}}{3} & i,j \in IE \\[2ex] \dfrac{\overline{SME(i,j)} + \overline{SDE(i,j)}}{2} & otherwise \end{cases}$$

where *IE* is the set of drugs among all the EFs investigated in this paper.

Also, integrated similarity matrix *SM* for miRNAs can be defined in the similar way (see Figure 1), where the entity *SM(i,j)* in row *i* column *j* is defined as follows:

$$SM(i,j) = \begin{cases} \dfrac{SFM(i,j) + \overline{SEM(i,j)} + \overline{SDM(i,j)}}{3} & i,j \in IM \\[2ex] \dfrac{\overline{SEM(i,j)} + \overline{SDM(i,j)}}{2} & otherwise \end{cases}$$

where *IM* is the set of miRNAs which have known functional similarity with other miRNAs investigated in this paper. The fact must be pointed out is that combinatorial coefficients could be better selected according to further cross validation. For simplicity, the trivial combinatorial coefficients have been adopted here according to previous studies, where similar operations of combining different similarity measures into an integrated similarity have been adopted [45, 46]. In these previous studies, reliable predictive performance has been obtained and the robustness of predictive accuracy to combinatorial coefficients selection has been illustrated [45, 46].

miREFRWR

In this paper, based on the assumption that similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease and the framework of Random Walk with Restart (RWR) [36, 46, 48, 50], I developed a novel method of miREFRWR to infer potential disease-related miRNA–EF interactions. As we all know, traditional RWR has been widely applied to plenty of biological problems, such as disease genes prioritization[48, 50], drug-target interactions prediction[46], disease-related miRNAs inference[36], and so on. However, traditional RWR has some critical limitations. miREFRWR has significant differences from traditional RWR. miREFRWR could make full use of the tool of the network for data integration to predict potential associations. The information of known disease-related miRNA-EF interactions, drug chemical structure similarity and miRNA functional similarity would be integrated in the framework of miREFRWR.

Based on above basic ideas, miREFRWR was developed as follows (see Figure 1, motivated by literature [63]). Firstly, some matrices are normalized before random walk is implemented on the network. Taking miRNA similarity matrix *SM* as an example, corresponding normalized matrix is defined as follows:

$$\overline{SM} = (DM)^{-1/2} SM (DM)^{-1/2}$$

where diagonal matrix *DM* is defined such that *DM(i,j)* is the sum of the *ith* row of *SM*. EF similarity matrix *SE* is also normalized in the similar way. For the disease-related miRNA-EF interaction adjacency matrix *A*, the entity *A(i,j)* in row *i* and column *j* is divided by the sum of elements in the matrix *A*. Hence, normalized miRNA similarity matrix $\overline{SM}$, normalized EF similarity network $\overline{SE}$, and normalized interaction adjacency matrix $\overline{A}$ (for brief description in the following equation, I set $p(0) = \overline{A}$) have been constructed, respectively. Secondly, considering the fact that there are different topologies and network structures in the miRNA similarity network and EF similarity network and hence the optimal iteration steps might be different on the two networks, two parameters were introduced to restrict the iteration steps of random walk on these two networks, respectively (motivated by literature [63]). Here, the parameters α and *β* are denoted as the numbers of maximal iterations in the miRNA similarity network and EF similarity network, respectively. Furthermore, the restart of random walk in every time step at source nodes can be allowed with probability *r* ($0<r<1$). Thirdly, two random walks would be implemented on miRNA similarity network and EF similarity network, respectively. The random walks in these two networks will finally converge to a unique solution after some steps of iterations. The predictive results from these two random walks would be combined to give the final prediction. miREFRWR is defined as follows (motivated by literature [63]):

```
$for\ t = 1\ to\ max(\alpha, \beta)$
$if\ t \le \alpha$
$pm(t) = r * \overline{SM} * p(t-1) + (1-r)A$
$if\ t \le \beta$
$pe(t) = r * p(t-1) * \overline{SE} + (1-r)A$
```

$$p(t) = \begin{cases} pm(\mathrm{t}) & \mathrm{t} > \beta \text{ and } \mathrm{t} \le \alpha \\ pe(\mathrm{t}) & \mathrm{t} \le \beta \text{ and } \mathrm{t} > \alpha \\ \dfrac{pm(\mathrm{t}) + \mathrm{pe}(\mathrm{t})}{2} & \mathrm{t} \le \beta \text{ and } \mathrm{t} \le \alpha \end{cases}$$

where *p(t)* is a matrix with the entity in row *i* and column *j* as the probability of arriving in the pair consisting of miRNA *i* and EF *j* at time step *t*. The value of matrix *p* could be updated based on this iteration equation and the current value of matrix *p*. Finally, when the number of iterations exceeds the maximum of α and *β*, the random walk would be terminated. Candidate miRNA-EF pairs are ranked according to corresponding values in final probability matrix *p* to select potential disease-related miRNA-EF interactions. The high-scored interactions can be expected to have a high probability to be associated with the given disease and will have priority to be tested in the biological experiments.

Results

Leave-one-out cross validation

Parameter α =4, β=4 and r=0.8 was chosen according to previous studies [63]. Actually, these parameters can be better selected based on further cross validation and the influence of parameters selection on the predictive results would be discussed in the following section.

LOOCV was implemented to evaluate the performance of miREFRWR. Considering the fact that miREFRWR cannot rank candidate miRNA-EF interactions for all the diseases simultaneously, LOOCV was implemented for each disease, respectively. In the known disease-related miRNA-EF interaction dataset, only about 8.89 miRNA-EF interactions have been associated with each disease on average, which means little difference between LOOCV and 10-fold cross validation. Furthermore, 32, 17, 12, 9, 3 out of all the 97 diseases have 1, 2, 3, 4, 5 known related interactions, respectively, which means multi-fold cross validation also cannot be implemented for most of the diseases. For these reasons, LOOCV was selected for performance validation.

In the LOOCV schema, each known interaction associated with the given disease is taken in turn as test sample and other known interactions associated with this disease are taken as training samples. Therefore, if this disease has only one known miRNA-EF interaction, LOOCV cannot be implemented. Aforementioned, network-based miRNA similarity and EF similarity matrices were recalculated when each cross validation run was implemented to avoid circular design and optimistic prediction performance of LOOCV. The performance of miREFRWR is evaluated based on the rank of this test sample in the candidate samples, which are composed of known left-out interaction and miRNA-EF pairs without the known associations with the given disease. Furthermore, ROC curve (plotting true positive rate (TPR, sensitivity) versus false positive rate (FPR, 1-specificity) at different cutoffs) was drawn and AUC was calculated (area under ROC curve). AUC= 1 shows perfect performance and 0.5 indicates random performance.

miREFRWR was compared with miREFScan (see Figure 2), which is the first disease-related miRNA-EF interaction prediction method. As a result, miREFRWR achieved an AUC of 0.9500, which showed the comparable performance with miREFScan. However, as a network-based method for miRNA-EF interactions prediction, it would bring a novel network perspective for the current research and promote the progression of developing network-based methods for miRNA-EF interactions prediction in the future.

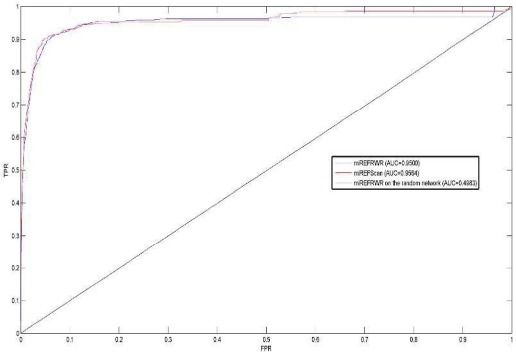


Fig.2 Performance Comparison. Comparison between miREFRWR and the first disease-related miRNA-EF interaction prediction method

miREFScan in terms of ROC curve and AUC based on LOOCV. As a result, miREFRWR achieved an AUC of 0.9500, which showed the comparable performance with miREFScan.

To evaluate whether the results of LOOCV by miREFRWR were likely to be obtained by chance, 100 random disease-related miRNA-EF interaction networks were generated. LOOCV procedure was implemented over these random networks and the mean FPR and mean TPR were obtained to plot the ROC curve and calculate AUC. As a result, AUC of 0.4983 demonstrated that observed excellent performance of miREFRWR cannot be achieved by chance, and hence prediction results by miREFRWR would be of biological significance and reflect some mechanisms of human complex diseases (see Figure 2).

Furthermore, miREFRWR was compared with some similar versions of miREFRWR which either ignored the use of network-based similarity or implemented miREFRWR only on the single network (see Supplementary Figure 1). As a result, miREFRWR significantly improved other methods, demonstrating the reasonability of introducing network-based similarity (AUC comparison between miREFRWR and miREFRWR without introducing network-based similarity in Supplementary Figure 1) and implementing miREFRWR on both miRNA similarity network and EF similarity network (AUC comparison between miREFRWR and miREFRWR implemented only on the single network in Supplementary Figure 1).

Also, when LOOCV was implemented for each disease, the ROC curve and the corresponding AUC can be obtained to assess how well the known miRNA-EF interactions of this disease were ranked relative to the candidate pairs (see Supplementary Table 2). The performance of the miREFRWR was evaluated by counting how many diseases had an AUC larger than different cutoffs (See Figure 3).

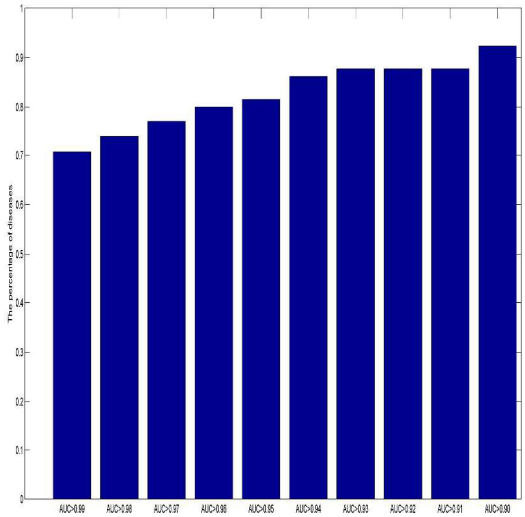


Fig.3 Performance of miREFRWR based on AUC. The performance of miREFRWR was evaluated by counting how many diseases had an AUC larger than different cutoffs.

Parameter effects on the performance of miREFRWR

There are three parameters in miREFRWR, including the numbers of maximal iterations in the miRNA similarity network and EF similarity network, and restart probability. To investigate the parameter effects on the performance of miREFRWR, various values are assigned to three parameters (the numbers of maximal iteration were taken to be between 1 steps to 5 steps and the restart probability was chosen from 0.1 to 0.9) and corresponding AUC of miREFRWR was calculated in the framework of LOOCV (see Supplementary Table 3). The result demonstrated that miREFRWR can obtain excellent performance in almost all the parameters selection (See Figure 4 and Figure 5).

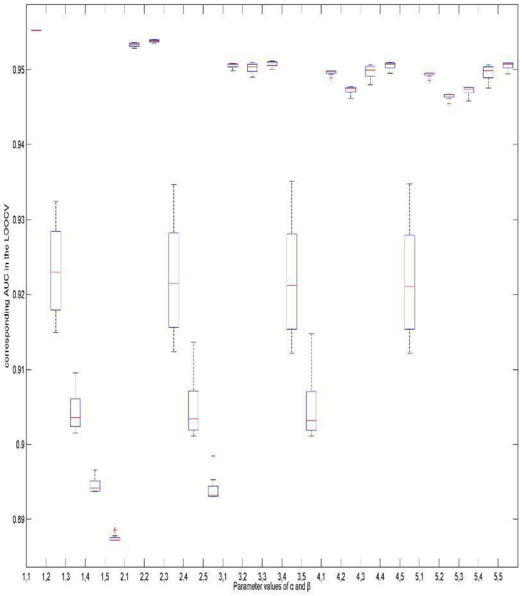


Fig.4 Performance of miREFRWR based on different selections of the numbers of maximal iteration. To investigate the parameter effects on the performance of miREFRWR, various values were assigned to α and β and corresponding AUC of miREFRWR was calculated in the framework of LOOCV. The result demonstrates that miREFRWR can obtain excellent performance in almost all the parameters selection. From this figure, it could be easily found that miREFRWR tends to show better performance when parameter α is greater than or equal to β and worse performance when α is less than β.

For the selection of parameter α and β, $\alpha=4$ and $\beta=4$ were reported to produce the best performance in the previous research about disease genes prioritization [63]. However, interesting conclusions can be obtained from the results in current disease-related miRNA-EF interactions prediction. Box plot for the AUCs in the framework of LOOCV corresponding to different parameter values of α and β was shown in Figure 4. From this figure, it could be easily found that miREFRWR tends to show better performance when parameter α is greater than or equal to β. Further confirmation can be obtained from the box plot for the AUCs in the framework of LOOCV when α is greater than or equal to β and α is less than β (see Supplementary Figure 2). This observation may arise from the fact that most of EFs show little similarity to other EFs and only local network information can be obtained for the random walk on the EF similarity network. Instead, the edges in the miRNA similarity network are denser than edges in EF similarity network. Therefore, the number of random walk steps on the EF similarity network should be less than steps on the miRNA similarity network.

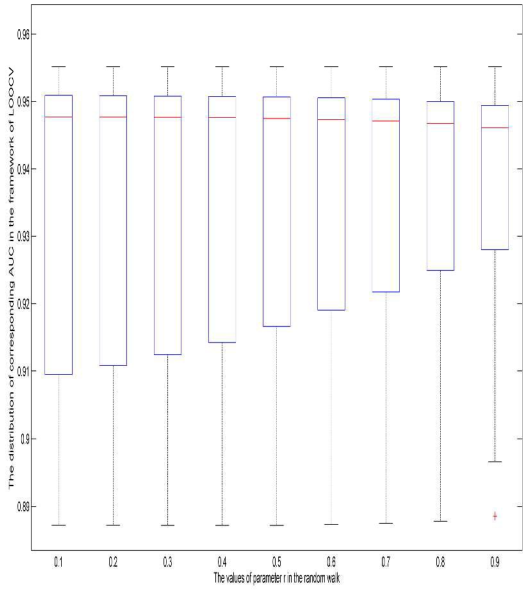


Fig.5 Performance of miREFRWR based on different selections of the restart probability. To investigate the performance of miREFRWR based on the different selection of restart probability, I set various value of r ranging from 0.1 to 0.9 and calculated AUC in the framework of LOOCV when different parameter values of α and β were chosen. Box plot for the AUCs corresponding to different parameter values of r is shown. It could be observed that the performance of miREFRWR is stable based on any selection of parameter values.

It has been demonstrated the predictive results of random walk are robust to the restart probability in the previous research about disease-related genes identification and disease-miRNA association inference[36, 48, 50]. As mentioned before, to investigate the performance of miREFRWR based on the different selections of restart probability, various values of *r* ranging from 0.1 to 0.9 were adopted and corresponding AUCs were calculated in the framework of LOOCV when different parameter values of α and β were chosen. Box plot for the AUCs corresponding to different parameter values of *r* is shown in Figure 5. It could be observed that the performance of miREFRWR is stable based on any selection of parameter values.

LOOCV in the new framework

The flaw of evaluation procedure for the pair-input computational prediction problems based on the cross validation have been pointed out in the recent literature [64]. The paired nature of inputs causes a natural partitioning of test samples. Normally, pair-input computational methods achieve different predictive performances for distinct test classes [64]. Based on this new validation framework, the test pairs of disease related miRNA-EF interactions are classified into four distinct classes: C1 is composed of the test samples sharing both EFs and miRNAs with the training samples; C2 is composed of the test samples sharing only miRNAs with the training samples; C3 is composed of the test samples sharing only EFs with the training samples; C4 is composed of the test samples sharing neither EFs nor miRNAs with the training samples. LOOCV is implemented for these four test classes and corresponding performance of miREFRWR has been shown in Figure 6 (AUC of 0.9931 in C1, 0.7929 in C2, 0.9548 in C3, 0.6803 in C4). Results demonstrated that RLSMDA has a reliable performance in all the test classes, even if the test samples only share either EFs or miRNAs with the training samples.

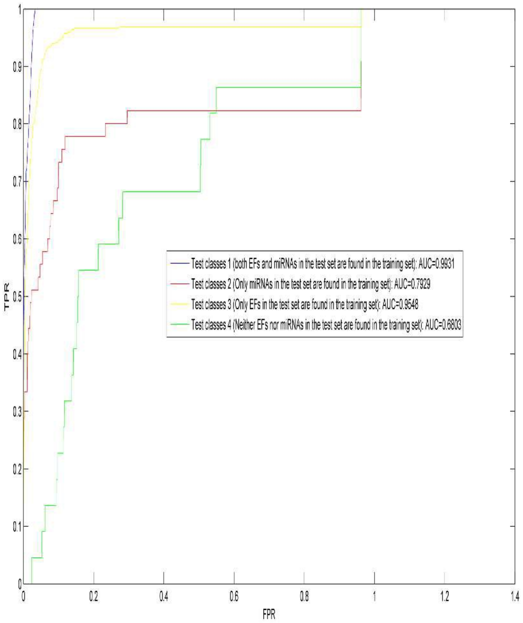


Fig.6 Performance evaluatin of miREFRWR based on the LOOCV in the new framework. To further investigate the performance of miREFRWR LOOCV is implemented for four test classes and corresponding performance of miREFRWR has been shown (AUC of 0.9931 in C1, 0.7929 in C2, 0.9548 in C3, 0.6803 in C4). Results demonstrated that RLSMDA has a reliable performance in all the test classes, even if the test samples share neither EFs nor miRNAs with the training samples.

Case studies

APL, a subtype of acute myelogenous leukemia, is regarded as the most malignant form of acute leukemia with a severe bleeding tendency and a highly fatal course of only weeks [65, 66]. Many studies have shown that the combined action of miRNAs and EFs would contribute to the development of effective therapy ways for APL. For example, four known APL related miRNA-EF interactions have been provided in the training dataset. All-trans retinoic acid (ATRA) can benefit the treatment of APL by suppressing the regulation of let-7a, mir-15a, and mir-16 [67]. The interaction between mir-21 and arsenic trioxide (ATO) may have a great curative effect on APL by regulating ATO-induced cell death [68]. Therefore, identifying disease-related miRNA-EF interactions could play a great role during clinical treatment.

Here, potential APL-related miRNA-EF interactions were predicted based on miREFRWR. As a result, 67 out of top 1% candidate interactions have been confirmed by latest experimental literatures [65, 69, 70] (see Supplementary Table 4). Previous method, miREFScan, only found 53 confirmed interactions. Fourteen confirmed interactions predicted by miREFRWR cannot be obtained by miREFScan, while all the confirmed interactions predicted by miREFScan can be obtained by miREFRWR. Moreover, all the confirmed interactions always obtain better ranking in the predictive list of miREFRWR than miREFScan (see Figure 7). For the top 0.5% and 0.1% of predictive list, 40 and 5 interactions based on miREFRWR have been confirmed, respectively. However, miREFScan only found 12 and 2 confirmed interactions. Above comparisons between miREFRWR and miREFScan fully demonstrated superior performance of new proposed methods and its potential value for disease diagnosis and treatment.

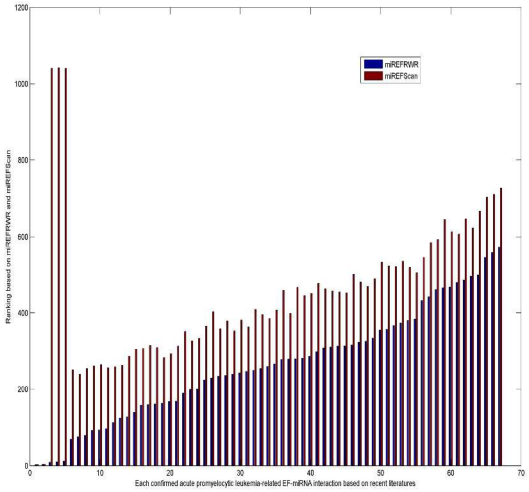


Fig.7 Case study of APL. For APL-related miRNA-EF interactions prediction, 67 out of top 1% candidate interactions have been confirmed by latest experimental literatures. Previous method, miREFScan, only found 53 confirmed interactions. Moreover, all the confirmed interactions always obtain better ranking in the predictive list of miREFRWR than miREFScan.

In recent literature [70], authors found the upregulation of miR-15a, miR-16, and let-7a in APL patients and cell lines treated by ATRA based on a miRNA microarrays platform and quantitative real time–polymerase chain reaction (qRT–PCR). The interactions between ATRA and these three miRNAs were ranked the 9th, 10th, and 12th in the predictive list based on miREFRWR, respectively. These interactions were ranked the 1041st, 1042nd, and 1040th by miREFScan, respectively. In another experimental literature [65], authors demonstrated that mir-16 and let-7a were significantly differentially expressed after ATO treatment in APL cell NB4. These two APL-related miRNA-EF interactions were ranked the 3rd and 4th in predictive list based on miREFRWR among more than 50, 000 candidate interactions.

To further evaluate the performance of miREFRWR on independent dataset, case study about breast cancer was implemented. Breast cancer is one of the most commonly occurring female cancers and makes up about 22% of all cancers in women. Recent biological experiments confirmed that miRNA let-7g was affected by Trastuzumab treatment in BT474 human breast cancer cells based on miRNA microarray profiling [71]. The interaction between let-7g and Trastuzumab was ranked 7th in the predictive list for breast cancer based on miREFRWR. By contrast, In the potential breast cancer-related miRNA-EF interactions list predicted by miREFScan, this interaction was only ranked 121st. Ichikawa et al. (2012) confirmed that miR-30b and miR-26a were upregulated in breast cancer cells after Trastuzumab treatment [72]. In the predictive list based on miREFRWR, these two interactions were ranked 75th and 88th. They are ranked 143rd and 165th by previous method miREFScan. In the top 1% predictive list based on miREFRWR, 10 interactions associated with breast cancer were confirmed [23, 73-75] (see Supplementary Table 5), while only 8 interactions can be found in the predictive list produced by miREFScan. Similar to the results in the case study of APL, two confirmed interactions predicted by miREFRWR cannot be obtained by miREFScan, while all the confirmed interactions predicted by miREFScan can be obtained by miREFRWR. Moreover, all the confirmed interactions always obtain better ranking in the predictive list of miREFRWR than miREFScan.

Predicting novel human disease-related miRNA-EF interactions

After confirming reliable predictive accuracy of miREFRWR based on LOOCV and the case studies about APL and breast cancer, miREFRWR was further applied to predict novel disease-related miRNA-EF interactions for all the 97 diseases investigated in this article. The top 100 potential miRNA-EF interactions associated with each disease were publicly released to facilitate further experimental validation from biologists (see Supplementary Table 6). Reliable performance demonstrated in previous LOOCV and case studies leads us to believe that these predicted novel relationships among miRNAs, EFs, and human diseases could benefit the diagnosis and treatment of diseases.

Discussion

The reliable performance of miREFRWR could be mainly attributed to the combination of two factors as follows. One is that I analyze and predict potential disease-related miRNA-EF interactions from a network perspective. Network-based methods could effectively identify biological properties at a network level and predict potential interactions among biological molecules. More importantly, global network information was adopted here, whose advantages over local network information methods have been demonstrated in many previous studies. The other is that known experimentally verified disease-related miRNA-EF interactions were used as the seed dataset to capture the potential associations between diseases and miRNA-EF interactions. Furthermore, drug chemical structure similarity, miRNA functional similarity, and networked-based similarity have also been integrated into miREFRWR. These two factors also constitute the novelties of miREFRWR. In conclusion, miREFRWR could be a novel, important and effective biomedical tool in the computational biology research.

Although excellent performance has been obtained in the both cross validation and case studies, it should be noted that some limitations still exist in the current version of miREFRWR. Firstly, although miREFRWR can obtain excellent performance in almost all the parameters selections, how to decide the parameters values is not still solved well. Secondly, I plan to introduce more reliable similarity measures into this computational model, such as disease phenotypical similarity, drug side-effect similarity, and miRNA functional similarity based on miRNA-target interactions. Also, how to integrate different similarity measures is an interesting and important problem in the computational biology. Furthermore, the current version of miREFRWR cannot be applied to the diseases without any known related miRNA-EF interactions. The performance of miREFRWR could be further improved when more experimentally confirmed human disease-related miRNA-EF interactions have been obtained in the future. Finally, the relationship between miRNA-EF interactions and cancer hallmark would be a very important problem for the future reseach. Specially, cancer hallmark network could be constructed at the miRNA and EF levels to effectively evaluate cancer risks[76].

Conclusions

Disease-related miRNA-EF interactions prediction is an important goal of biomedical research and plays a critical role in the understanding of disease pathogenesis at the miRNA and EF levels and the design of specific molecular tools for the prognosis, diagnosis, treatment and prevention of human disease. In this paper, a novel method of miREFRWR was developed to predict potential disease-related miRNA-EF interactions. LOOCV and case studies about APL and breast cancer demonstrated that miREFRWR can effectively identify potential disease-related miRNA-EF interactions on a large scale by integrating the information of known disease-related miRNA-EF interactions, drug chemical structure, and miRNA functional similarity. Especially, miREFRWR has a reliable predictive accuracy in different test datasets according to the evaluation methods proposed in a recent article. Furthermore, the top 100 interactions associated with each disease have been publicly released to guide future biological experiments. It is anticipated that miREFRWR could be an effective and important biological tool for the research of non-coding RNAs, complex diseases, and drug design in the future.

Acknowledgements

This work was supported by the National Natural Science of Foundation of China under Grant No. 11301517 and National Center for Mathematics and Interdisciplinary Sciences, CAS.

Notes and references

[a] *National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn.*

[b] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn.*

† Electronic Supplementary Information (ESI) available. See DOI: 10.1039/b000000x/

‡ **Authors' contributions.** XC conceived the prediction method, developed the prediction method, conceived, designed and implemented the experiments, analyzed the result, and wrote the paper. All authors read and approved the final manuscript.

1. A. L. Barabási, N. Gulbahce and J. Loscalzo, *Nat Rev Genet*, 2011, **12**, 56-68.
2. W.-H. Chow, L. M. Dong and S. S. Devesa, *Nat Rev Urol*, 2010, **7**, 245-257.
3. U. N. Das, *Nutrition*, 2010, **26**, 459-473.
4. A. M. Soto and C. Sonnenschein, *Nat. Rev. Endocrinol.*, 2010, **6**, 363-370.
5. A. Esquela-Kerscher and F. J. Slack, *Nat Rev Cancer*, 2006, **6**, 259-269.
6. X. Karp and V. Ambros, *Science*, 2005, **310**, 1288.
7. A. M. Cheng, M. W. Byrom, J. Shelton and L. P. Ford, *Nucleic Acids Res*, 2005, **33**, 1290-1297.
8. E. A. Miska, *Curr Opin Genet Dev*, 2005, **15**, 563-568.
9. P. Xu, M. Guo and B. A. Hay, *Trends Genet*, 2004, **20**, 617-624.
10. Q. Cui, Z. Yu, E. O. Purisima and E. Wang, *Mol Syst Biol*, 2006, **2**, 46.
11. D. P. Bartel, *Cell*, 2004, **116**, 281-297.
12. M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, *PLoS One*, 2008, **3**, e3420.
13. M. V. G. Latronico, D. Catalucci and G. Condorelli, *Circ Res*, 2007, **101**, 1225-1236.
14. Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu and Y. Wang, *BMC Syst Biol*, 2010, **4**, S2.
15. G. A. Calin and C. M. Croce, *Nat Rev Cancer*, 2006, **6**, 857-866.
16. R. F. Duisters, A. J. Tijsen, B. Schroen, J. J. Leenders, V. Lentink, I. van der Made, V. Herias, R. E. van Leeuwen, M. W. Schellings and P. Barenbrug, *Circ Res*, 2009, **104**, 170-178.
17. A. Markou, E. G. Tsaroucha, L. Kaklamanis, M. Fotinou, V. Georgoulias and E. S. Lianidou, *Clin Chem*, 2008, **54**, 1696-1704.
18. T. E. Miller, K. Ghoshal, B. Ramaswamy, S. Roy, J. Datta, C. L. Shapiro, S. Jacob and S. Majumder, *J Biol Chem*, 2008, **283**, 29897-29903.
19. F. J. Slack and J. B. Weidhaas, *N Engl J Med*, 2008, **359**, 2720-2722.
20. M. S. Weinberg and M. J. A. Wood, *Hum Mol Genet*, 2009, **18**, R27-R39.
21. M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. MacDonald, S. Pfeffer, T. Tuschl, N. Rajewsky and P. Rorsman, *Nature*, 2004, **432**, 226-230.
22. H. H. G. van Es and G. J. Arts, *Drug Discov Today*, 2005, **10**, 1385-1391.
23. F. Xin, M. Li, C. Balch, M. Thomson, M. Fan, Y. Liu, S. M. Hammond, S. Kim and K. P. Nephew, *Bioinformatics*, 2009, **25**, 430-434.
24. Q. Huang, K. Gumireddy, M. Schrier, C. Le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang and A. J. Klein-Szanto, *Nat Cell Biol*, 2008, **10**, 202-210.
25. R. T. Lima, S. Busacca, G. M. Almeida, G. Gaudino, D. A. Fennell and M. H. Vasconcelos, *Eur. J. Cancer*, 2011, **47**, 163-174.
26. Y. Ladeiro, G. Couchy, C. Balabaud, P. Bioulac‐Sage, L. Pelletier, S. Rebouissou and J. Zucman‐Rossi, *Hepatology*, 2008, **47**, 1955-1963.
27. A. Izzotti, P. Larghero, C. Cartiglia, M. Longobardi, U. Pfeffer, V. E. Steele and S. De Flora, *Carcinogenesis*, 2010, **31**, 894-901.
28. Y. Gidron, M. De Zwaan, K. Quint and M. Ocker, *Mol Med Rep*, 2010, **3**, 455.
29. A. Alisi, L. Da Sacco, G. Bruscalupi, F. Piemonte, N. Panera, R. De Vito, S. Leoni, G. F. Bottazzo, A. Masotti and V. Nobili, *Lab. Invest.*, 2010, **91**, 283-293.
30. Z. Lin and E. K. Flemington, *Cancer Lett*, 2011, **305**, 186-199.
31. M. J. Jardim, *Mutat Res*, 2011, **717**, 38-45.
32. O. M. Niemoeller, M. Niyazi, S. Corradini, F. Zehentmayr, M. Li, K. Lauber and C. Belka, *Radiat Oncol*, 2011, **6**, 29.
33. S. Xi, M. Yang, Y. Tao, H. Xu, J. Shan, S. Inchauste, M. Zhang, L. Mercedes, J. A. Hong and M. Rao, *PLoS One*, 2010, **5**, e13764.
34. T. Boren, Y. Xiong, A. Hakam, R. Wenham, S. Apte, G. Chan, S. G. Kamath, D.-T. Chen, H. Dressman and J. M. Lancaster, *Gynecol. Oncol.*, 2009, **113**, 249-255.
35. Y. Jin, X. Zou and X. Feng, *Anticancer Drugs*, 2010, **21**, 814-822.
36. X. Chen, M. X. Liu and G. Yan, *Mol Biosyst*, 2012, **8**, 2792-2798.
37. X. Chen and G.-Y. Yan, *Sci. Rep.*, 2014, **4**, 5501.
38. X. Chen, *Sci. Rep.*, 2015, **5**, 13186.
39. X. Chen, C. C. Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang and Q. Dai, *Sci. Rep.*, 2015, **5**, 13877.
40. X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang and Q. Dai, *Sci. Rep.*, 2015, **5**, 11338.
41. X. Chen and G.-Y. Yan, *Bioinformatics*, 2013, **29**, 2617-2624.
42. X. Chen, *Sci. Rep.*, 2015, **5**, 16840.
43. Q. Yang, C. Qiu, J. Yang, Q. Wu and Q. Cui, *Bioinformatics*, 2011, **27**, 3329-3330.
44. C. Qiu, G. Chen and Q. Cui, *Sci. Rep.*, 2012, **2**, 318.
45. X. Chen, M. X. Liu, Q. H. Cui and G. Y. Yan, *PLoS One*, 2012, **7**, e43425.
46. X. Chen, M. X. Liu and G. Yan, *Mol BioSyst*, 2012, **8**, 1970-1978.
47. J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu and M. Zhou, *Mol Biosyst*, 2014, **10**, 2074-2081.
48. S. Köhler, S. Bauer, D. Horn and P. N. Robinson, *Am J Hum Genet*, 2008, **82**, 949.
49. M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou and J. Sun, *Mol Biosyst*, 2015, **11**, 760-769.
50. X. Chen, G. Y. Yan and X. P. Liao, *OMICS*, 2010, **14**, 337-356.

51. T. van Laarhoven, S. B. Nabuurs and E. Marchiori, *Bioinformatics*, 2011, **27**, 3036-3043.
52. Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics*, 2010, **26**, i246-i254.
53. K. Bleakley and Y. Yamanishi, *Bioinformatics*, 2009, **25**, 2397-2403.
54. Y. C. Wang, Z. X. Yang, Y. Wang and N. Y. Deng, *Lett Drug Des Discov*, 2010, **7**, 370-378.
55. Z. Xia, L. Y. Wu, X. Zhou and S. T. C. Wong, *BMC Syst Biol*, 2010, **4**, S6.
56. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics*, 2008, **24**, i232-i240.
57. W. Yu, X. Cheng, Z. Li and Z. Jiang, *Drug Develop Res*, 2011, **72**, 219-224.
58. A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin and R. Sharan, *Mol Syst Biol*, 2012, **8**, 592.
59. M. Hattori, Y. Okuno, S. Goto and M. Kanehisa, *J Am Chem Soc*, 2003, **125**, 11853-11865.
60. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, *Nucleic Acids Res*, 2006, **34**, D354-D357.
61. E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, *Annu. Rep. Comput. Chem.*, 2008, **4**, 217-241.
62. D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, *Bioinformatics*, 2010, **26**, 1644-1650.
63. M. Xie, T. Hwang and R. Kuang, in *Advances in Knowledge Discovery and Data Mining*, Springer2012, pp. 292-303.
64. Y. Park and E. M. Marcotte, *Nat. Methods*, 2012, **9**, 1134-1136.
65. S. H. Ghaffari, D. Bashash, A. Ghavamzadeh and K. Alimoghaddam, *Tumour Biol.*, 2012, **33**, 157-172.
66. J. H. Martens, A. B. Brinkman, F. Simmer, K.-J. Francoijs, A. Nebbioso, F. Ferrara, L. Altucci and H. G. Stunnenberg, *Cancer cell*, 2010, **17**, 173-185.
67. C. D. Davis and S. A. Ross, *Nutr. Rev.*, 2008, **66**, 477-482.
68. J. Gu, X. Zhu, Y. Li, D. Dong, J. Yao, C. Lin, K. Huang, H. Hu and J. Fei, *Med. Oncol.*, 2011, **28**, 211-218.
69. Y. Wu, X. Li, J. Yang, X. Liao and Y. Chen, *Zhonghua Xue Ye Xue Za Zhi*, 2012, **33**, 546.
70. R. Garzon, F. Pichiorri, T. Palumbo, M. Visentini, R. Aqeilan, A. Cimmino, H. Wang, H. Sun, S. Volinia and H. Alder, *Oncogene*, 2007, **26**, 4148-4157.
71. X.-F. Le, M. I. Almeida, W. Mao, R. Spizzo, S. Rossi, M. S. Nicoloso, S. Zhang, Y. Wu, G. A. Calin and R. C. Bast Jr, *PLoS One*, 2012, **7**, e41170.
72. T. Ichikawa, F. Sato, K. Terasawa, S. Tsuchiya, M. Toi, G. Tsujimoto and K. Shimizu, *PLoS One*, 2012, **7**, e31422.
73. S. L. Tilghman, M. R. Bratton, H. C. Segar, E. C. Martin, L. V. Rhodes, M. Li, J. A. McLachlan, T. E. Wiese, K. P. Nephew and M. E. Burow, *PLoS One*, 2012, **7**, e32754.
74. M. Jansen, E. Reijm, A. Sieuwerts, K. Ruigrok-Ritstier, M. Look, F. Rodríguez-González, A. Heine, J. Martens, S. Sleijfer and J. Foekens, *Breast Cancer Res. Treat.*, 2012, **133**, 937-947.
75. E. J. Jung, L. Santarpia, J. Kim, F. J. Esteva, E. Moretti, A. U. Buzdar, A. Di Leo, X. F. Le, R. C. Bast and S. T. Park, *Cancer*, 2012, **118**, 2603-2614.
76. E. Wang, N. Zaman, S. Mcgee, J.-S. Milanese, A. Masoudi-Nejad and M. O'Connor-McCourt, *Semin. Cancer Biol.*, 2015, **30**, 4-12.