LOMDA: Linear optimization for miRNA-disease association prediction

Yan-Li Lee[1], Ratha Pech[1*], Maryna Po[3], Dong Hao[1,2], Tao Zhou[1,2]

1 CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China.
2 Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China.
3 Department of Chemistry and Biochemistry, George Mason University, Virginia 22030, USA

* ratha.pech@gmail.com

Abstract

MicroRNAs (miRNAs) have been playing a crucial role in many important biological processes e.g., pathogenesis of diseases. Currently, the validated associations between miRNAs and diseases are insufficient comparing to the hidden associations. Testing all these hidden associations by biological experiments is expensive, laborious, and time consuming. Therefore, computationally inferring hidden associations from biological datasets for further laboratory experiments has attracted increasing interests from different communities ranging from biological to computational science. In this work, we propose an effective and efficient method to predict associations between miRNAs and diseases, namely linear optimization (LOMDA). The proposed method uses the heterogenous matrix incorporating of miRNA functional similarity information, disease similarity information and known miRNA-disease associations. Compared with the other methods, LOMDA performs best in terms of AUC (0.970), precision (0.566), and accuracy (0.971) in average over 15 diseases in local 5-fold cross-validation. Moreover, LOMDA has also been applied to two types of case studies. In the first case study, 30 predictions from breast neoplasms, 24 from colon neoplasms, and 26 from kidney neoplasms among top 30 predicted miRNAs are confirmed. In the second case study, for new diseases without any known associations, top 30 predictions from hepatocellular carcinoma and 29 from lung neoplasms among top 30 predicted miRNAs are confirmed.

Author summary

Identifying associations between miRNAs and diseases is significant in investigation of pathogenesis, diagnosis, treatment and preventions of related diseases. Employing computational methods to predict the hidden associations based on known associations and focus on those predicted associations can sharply reduce the experimental costs. We developed a computational method LOMDA based on the linear optimization technique to predict the hidden associations. In addition to the observed associations, LOMDA also can employ the auxiliary information (diseases and miRNAs similarity information) flexibly and effectively. Numerical experiments on global 5-fold cross validation show that the use of the auxiliary information can greatly improve the prediction performance. Meanwhile, the result on local 5-fold cross validation shows

that LOMDA performs best among the seven related methods. We further test the prediction performance of LOMDA for two types of diseases based on HDMMv2.0 (2014), including (i) diseases with all the known associations, and (ii) new diseases without known associations. Three independent or updated databases (dbDEMC, 2010; miR2Disease, 2009; HDMMv3.2, 2019) are introduced to evaluate the prediction results. As a result, most miRNAs for target diseases are confirmed by at least one of the three databases. So, we believe that LOMDA can guide experiments to identify the hidden miRNA-disease associations.

Introduction

MicroRNAs (miRNAs) are short for non-coding RNAs about 22 nucleotides that regulate gene expression of target post-transcriptional level [1–4]. In the last few decades, accumulative evidences show that miRNAs have strong relationships with many critical life processes including early cell growth, proliferation, apoptosis, differentiation and metabolism [5–9]. Moreover, miRNA dysregulation has also been shown to have close relation with many human complex diseases [10–14], including lung cancer [15] ,breast cancer [16, 17], cardiovascular diseases [18], and so on. Therefore, studying associations of miRNAs and diseases from biological datasets has become a significant problem in biomedical research which not only helps in investigation of pathogenesis [19], but also assists diagnosis, treatment, and preventions [20–23]. Biological experiments to verify new associations one by one would require a huge amount of time and labor, hence, an effective and efficient tool for selecting a small portion which are the most likely associations among a large pool of associations is needed for scientists to further experiments.

To investigate the miRNA-disease associations, comprehensive databases about miRNA-disease associations have been constructed, e.g., human miRNA disease database (HMDDv2.0) [24] collecting human miRNAs and diseases associations which are experiment-supported, dbDEMC [25] containing different expressions of miRNAs in human cancers detected by high-throughput methods, and miR2Disease [26] containing comprehensive resource of miRNA deregulations in various human diseases. These databases have facilitated researchers and scientists in understanding disease pathogenesis, furthermore, they are the main resources for association identification. Although there are rich collections about miRNA-disease association databases, these known associations are still limited comparing to all hidden miRNA-disease associations. Moreover, it is believed that one miRNA can be associated with multiple diseases and vice versa.

There is plenty of research [27] in predicting associations between miRNAs and diseases by using computational methods [28–32] and network-based methods [33–39]. Specifically, Chen *et al.* [34] proposed a method named random walk with restart for miRNA–disease associations (RWRMDA) to predict novel human miRNA-disease associations. In their proposed method, two datasets are used including the miRNA-disease network and miRNA functional similarities. Random walk cannot reach nodes that are isolated from the others. Therefore, it cannot predict associations between miRNAs and diseases that do not have any known associations. Similarly, Xuan *et al.* [40] proposed a method namely MIDP (short for miRNAs associated with Diseases Prediction), which is also based on the random walk technique, and the miRNA-disease network and miRNA functional similarities are used in their model. The difference between RWRMDA and MIDP is that the transition matrices are different. Chen *et al.* [41] developed a method by using the semi-supervised learning technique namely the regularized least squares for miRNA-disease association (RLSMDA) prediction, which integrates miRNA-disease associations, disease-disease similarities, and miRNA-miRNA

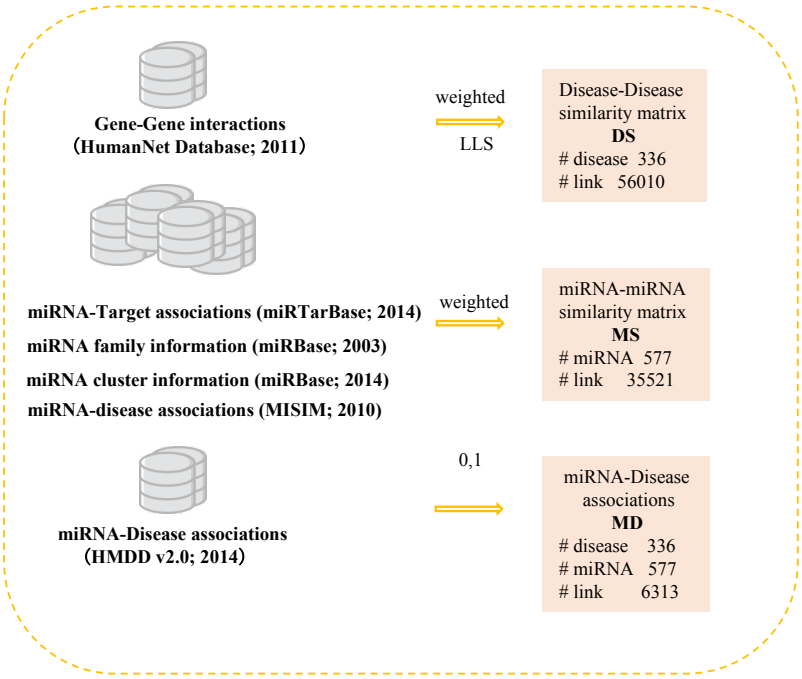

Fig 1. The illustration of the input datasets of LOMDA.

similarities. Treating association prediction problem as binary classification would suffer from one drawback such that known associations are used as positive labels, while some unknown associations are used as negative labels. However, these unknown associations could be the hidden associations which are actually with positive labels. Zeng *et al.* [42] proposed a method by using structural perturbation method (SPM) on the integration of miRNA-disease association network, miRNA similarity network, and disease similarity network. Zeng *et al.* [43] also proposed a method namely neural network model for miRNA-disease association prediction (NNMDA), which also uses heterogenous network by integrating neighborhood information. SPM and NNMDA can well integrate multiple biological data resources to do miRNAs-diseases association prediction.

In this work, we propose an effective and efficient method namely linear optimization for miRNA-disease associations prediction (LOMDA) to infer associations between miRNAs and diseases in both cases: either (i) diseases and miRNAs similarity information is available, or (ii) only known associations are provided. LOMDA can also predict associations between miRNAs and diseases that do not have any known associations, but have miRNAs functional similarity information or disease similarity information, respectively. The performance of LOMDA measured by AUC, precision, and accuracy has been shown much higher than other benchmarks. Moreover, case studies demonstrate the effectiveness of LOMDA on predicting novel associations between miRNAs and diseases. Results show that LOMDA would be a promising bioinformatics tool for biomedical researchers. The Matlab source code of the proposed method and datasets used in this work can be obtained at `https://github.com/rathapech/LOMDA`.

Materials and methods

Data description

The miRNA-disease associations and two types of auxiliary information will be introduced in this subsection.

The human miRNA–disease associations are obtained from HMDDv2.0 [24]. If a miRNA is related to a disease, there is one link between them. After removing duplications, there are 6313 associations between 577 miRNAs and 336 diseases. The corresponding adjacency matrix is denoted as **MD**.

The disease-disease similarity matrix is constructed based on the gene functional information. Genes with similar function regulate similar diseases with greater probability [37]. The functional similarity of genes here is characterized by the log-likelihood score (LLS) [44], which can be downloaded from the HumanNet database [45]. Furthermore, the similarity of diseases d_i and d_j is calculated as

$$\mathbf{DS}_{ij} = \begin{cases} \frac{\sum_{x \in S(d_i)} LLS(x, S(d_j)) + \sum_{y \in S(d_j)} LLS(y, S(d_i))}{|S(d_i)| + |S(d_j)|}, & |S(d_i)| + |S(d_j)| \neq 0 \\ 0, & otherwise \end{cases} \tag{1}$$

where $S(d_i)$ and $S(d_j)$ are gene sets that related to disease d_i and d_j, respectively. $|S|$ is the cardinality of set S. $LLS(x, S(d_i))$ is LLS between the gene x and the gene set $S(d_i)$, where $x \in S(d_j)$. There are 56010 links between 336 diseases, and they can be formulated as a symmetric weighted matrix **DS**.

The miRNA-miRNA functional similarity matrix is obtained from four sources of information, including verified miRNA-target associations (RST), which can be obtained from miRTarBase [46], miRNA family information (RSF, `ftp://mirbase.org/pub/mirbase/CURRENT/mi-Fam.dat.gz`), cluster information (RSC, `http://www.mirbase.org/`) and verified miRNA-disease associations (RSD, `http://www.cuilab.cn/files/images/cuilab/misim.zip`) [37]. RST_{ij} is the number of shared targets between miRNAs r_i and r_j. $RSF_{ij} = 1$ if two miRNAs belong to the same miRNA family, and $RSF_{ij} = 0$, otherwise. Similarly, $RSC_{ij} = 1$ if two miRNAs belong to the same cluster, and $RSC_{ij} = 0$, otherwise. RSD is calculated by MISIM (miRNA similarity) method [37]. Finally, $\mathbf{MS}(r_i, r_j)$ reads

$$\mathbf{MS}(r_i, r_j) = \eta.RST(r_i, r_j) + \beta.RSF(r_i, r_j) + \gamma.RSC(r_i, r_j) + \theta.RSD(r_i, r_j) \tag{2}$$

where η, β, γ and θ are parameters to adjust the four weights and are set as $\eta = 0.2$, $\beta = 0.1$, $\gamma = 0.2$ and $\theta = 0.5$ [42]. There are 35521 links between 577 miRNAs, and they can be formulated as a symmetric weighted matrix **MS**.

The above three matrices can be combined to get an heterogenous matrix as

$$\mathbf{A} = \begin{bmatrix} \mathbf{MS} & \mathbf{MD} \\ \mathbf{MD}' & \mathbf{DS} \end{bmatrix}, \tag{3}$$

where $\mathbf{MS} \in \mathbb{R}^{m \times m}$ is the miRNA functional similarity matrix in which m is the number of miRNAs; $\mathbf{DS} \in \mathbb{R}^{d \times d}$ is the disease similarity matrix in which d is the number of diseases; $\mathbf{MD} \in \mathbb{R}^{m \times d}$ is the known miRNA-disease association matrix. If only associations between miRNAs and diseases are available, we can set $\mathbf{A} = \mathbf{MD}$. The composition of heterogenous matrix **A** is shown in Fig 1.

LOMDA

Denoting the integration matrix by **A** as shown in Eq (3), we assume that the likelihood of associations between miRNAs and diseases can be written as a linear combination of

$\mathbf{A}$ and weighting matrix $\mathbf{Z}$ as

$$\mathbf{S} = \mathbf{AZ}. \tag{4}$$

Since $\mathbf{S}$ and $\mathbf{Z}$ are unknown, the problem of Eq (4) has infinite solutions. However, in order to obtain the likelihood $\mathbf{S}$ containing existing and predicted associations, $\mathbf{S}$ should be intuitively and reasonably close to $\mathbf{A}$. Then we can write

$$||\mathbf{A} - \mathbf{AZ}|| < \epsilon, \tag{5}$$

where ϵ is the threshold parameter. Moreover, to avoid the model to be overfitted and simultaneously to constrain the magnitudes of $\mathbf{Z}$, we can relax the Eq (5) as

$$\mathbf{E} = \alpha||\mathbf{A} - \mathbf{AZ}|| + ||\mathbf{Z}||, \tag{6}$$

where α is the positive free parameter greater than 0 and $||.||$ is the matrix norm. Without losing the generality [47], we use Frobenius norm and raise the two terms with power 2. We can have

$$\mathbf{E} = \alpha||\mathbf{A} - \mathbf{AZ}||_F^2 + ||\mathbf{Z}||_F^2, \tag{7}$$

where Frobenius norm is denoted as $||\mathbf{Z}||_F = \sqrt{\text{trace}(\mathbf{Z}^T\mathbf{Z})} = \sqrt{\sum_{i=1}^{min\{p,q\}} \sigma_i^2}$, e.g., σ_i is the singular value, p is the number of row, and q is the number of column of $\mathbf{Z}$. The expansion of Eq (7) reads

$$\begin{aligned} \mathbf{E} &= \alpha\mathbf{Tr}[(\mathbf{A} - \mathbf{AZ})^T(\mathbf{A} - \mathbf{AZ})] + \mathbf{Tr}(\mathbf{Z}^T\mathbf{Z}) \\ &= \alpha\mathbf{Tr}(\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{AZ} - \mathbf{Z}^T\mathbf{A}^T\mathbf{A} + \mathbf{Z}^T\mathbf{A}^T\mathbf{AZ}) + \mathbf{Tr}(\mathbf{Z}^T\mathbf{Z}), \end{aligned} \tag{8}$$

with its partial derivative being

$$\frac{\partial\mathbf{E}}{\partial\mathbf{Z}} = \alpha(-2\mathbf{A}^T\mathbf{A} + 2\mathbf{A}^T\mathbf{AZ}) + 2\mathbf{Z}. \tag{9}$$

Setting $\partial\mathbf{E}/\partial\mathbf{Z} = 0$, we can obtain the optimal solution of $\mathbf{Z}$ as

$$\mathbf{Z}^* = \alpha(\alpha\mathbf{A}^T\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}, \tag{10}$$

where $\mathbf{I}$ is the identity matrix. The likelihood matrix $\mathbf{S}$ can be obtained as

$$\mathbf{S} = \mathbf{AZ}^*. \tag{11}$$

Finally, we can obtain scores of unobserved miRNAs and diseases associations based on $\mathbf{S}$.

In addition, when only known association information of miRNAs and diseases is available, LOMDA can also effectively predict missing associations by using only $\mathbf{MD}$. We replace $\mathbf{A}$ by matrix $\mathbf{MD}$ in Eq (4) through Eq (11) and keep the derivation the same. When only matrix $\mathbf{MD}$ is used, we call the proposed method as LOMDA-MD. We illustrate the performance of these two derived methods in result section.

Results

Performance evaluation

To evaluate our proposed method against others, we adopt the 5-fold cross validation technique to test the prediction performance. Moreover, two different prediction tasks are considered: (i) predicting hidden associations for all the diseases simultaneously. Specifically, all the associations in $\mathbf{MD}$ are divided into 5 disjoint subsets. Four subsets are treated as training samples, and the remaining subset is treated as testing samples.

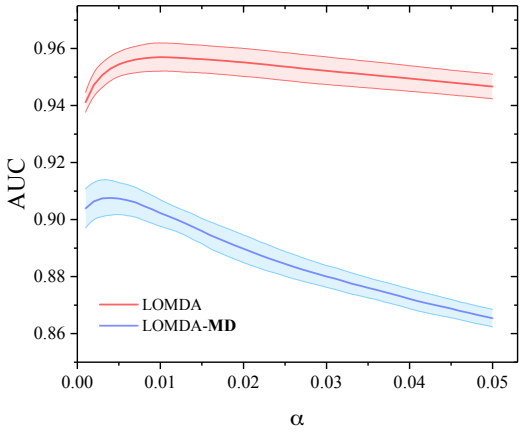

Fig 2. The contribution of auxiliary information to prediction results based on global 5-fold cross validation. The shaded area presents the standard deviation.

We call it as global 5-fold cross validation; (ii) predicting hidden associations for some specific diseases, i.e., for a specific disease, the related associations in matrix **MD** are divided into 5 disjoint subsets. One subset is considered as the testing set, and the remaining associations in **MD** are considered as training set. We call it as local 5-fold cross validation. For the two tasks, we repeatedly and independently do the simulation for five times until all five subsets are used as testing samples exactly once. This 5-fold cross validation is repeated by 20 times. That is to say, the final evaluation metrics introduced in the following are the average over 20*5 prediction results.

We introduce three evaluation metrics including AUC, precision, and accuracy to evaluate the prediction performance. AUC can be interpreted as a probability that a randomly chosen hidden association (i.e., associations in testing samples) is given a higher score than a randomly chosen unknown association (i.e., associations that have not been collected in training samples and testing samples) [48]. Notice that, in the second task for a specific disease, unknown associations denote associations between the target disease and miRNAs that have not been collected by the target disease according to the matrix **MD**. We randomly choose n pairs of associations from the testing samples and the unknown associations, respectively. n_1 denotes the number of pairs with higher score for hidden associations, and n_2 denotes the number of pairs with the same score between the hidden association and the unknown association. The score values are obtained based on the related method. Thus, AUC $= \frac{n_1+0.5n_2}{n}$, and n is set to 10000 in this work. Precision is defined as the ratio of right predicted associations (i.e., true positive associations) to all the predicted associations. We select top L associations in predicted scores and count them if they exist in testing data, and the hitting number is denoted as L_r, then precision $= \frac{L_r}{L}$. L is the number of testing samples in our work. Accuracy is obtained by the ratio of true positive associations and true negative associations to all the candidate associations, i.e., the union set of hidden associations and unknown associations.

Performance with auxiliary information

In this subsection, LOMDA and LOMDA-MD are used to test the contribution of auxiliary information to prediction results. The prediction is conducted for all the diseases simultaneously by the global 5-fold cross validation technique. For LOMDA, not only the training set but also the auxiliary information **MS** and **DS** are employed to do prediction, while for LOMDA-MD, only the training set is used. As shown in Fig

Table 1. Information included in seven methods. MS is the miRNA-miRNA functional similarity matrix, **DS** is the disease-disease similarity matrix, and **MD** is the known miRNA-disease association matrix.

Datasets	RWRMDA	RLSMDA	MIDP	SPM	NNMDA	LOMDA-MD	LOMDA
MS	Yes	Yes	Yes	Yes	Yes	No	Yes
DS	No	Yes	No	Yes	Yes	No	Yes
MD	Yes	Yes	Yes	Yes	Yes	Yes	Yes

2, LOMDA performs much better than LOMDA-MD. Meanwhile, we can find that LOMDA achieves its best performance on $\alpha = 0.01$, and $\alpha = 0.001$ for LOMDA-MD.

Comparison with other methods

A great number of methods are developed to predict hidden associations between miRNAs and diseases, including RWRMDA [34], RLSMDA [41], MIDP [40], SPM [42] and NNMDA [43]. Detailed information employed by each method is shown in Table 1. The auxiliary information (**MS** and **DS**) are employed according to Table 1. The prediction is conducted for some specific diseases based on local 5-fold cross validation. Table 2–Table 4 illustrate the AUC, precision, and accuracy of the proposed method and others on 15 different diseases, respectively. The highest values generating from any methods are shown in boldface. As shown in the bottom row of each table, average AUC values among the 15 diseases of RWRMDA, RLSMDA, MIDP, SPM, NNMDA, LOMDA-MD and LOMDA are 0.898, 0.950, 0.953, 0.786, 0.920, 0.938, 0.970, respectively. Average precision values on all the 15 diseases obtained from RWRMDA, RLSMDA, MIDP, SPM, NNMDA, LOMDA-MD and LOMDA are 0.207, 0.417, 0.392, 0.306, 0.442, 0.372 and 0.566, respectively. Average accuracy values on the 15 diseases from RWRMDA, RLSMDA, MIDP, SPM, NNMDA, LOMDA-MD and LOMDA are 0.944, 0.961, 0.959, 0.950, 0.938, 0.956 and 0.971, respectively. Overall speaking, LOMDA outperforms the other benchmarks on AUC, precision and accuracy.

Table 2. AUC of different methods by using local 5-fold cross validation on 15 diseases.

Disease	RWRMDA	RLSMDA	MIDP	SPM	NNMDA	LOMDA-MD	LOMDA
Breast neoplasms	0.951	0.973	0.975	0.841	0.968	0.947	**0.986**
Hepatocellular carcinoma	0.915	0.941	0.959	0.875	0.946	0.905	**0.982**
Renal cell carcinoma	0.925	0.938	**0.965**	0.683	0.915	0.927	0.962
Squamous cell carcinoma	0.874	0.955	0.954	0.719	0.935	0.937	**0.964**
Colorectal neoplasms	0.915	0.954	0.944	0.721	0.905	0.940	**0.965**
Glioblastoma	0.887	0.950	0.945	0.740	0.907	0.930	**0.954**
Heart failure	0.858	0.918	0.938	0.832	0.940	0.910	**0.983**
Acute myeloid leukemia	0.831	0.955	0.957	0.858	0.885	0.956	**0.981**
Lung neoplasms	0.907	0.952	0.958	0.794	0.927	0.942	**0.970**
Melanoma	0.915	0.960	0.952	0.821	0.955	0.954	**0.971**
Ovarian neoplasms	0.905	0.960	0.955	0.768	0.914	0.956	**0.966**
Pancreatic neoplasms	0.892	0.948	0.951	0.803	0.917	0.943	**0.972**
Prostatic neoplasms	0.872	0.939	0.937	0.739	0.879	0.937	**0.955**
Stomach neoplasms	0.931	0.952	0.961	0.847	0.858	0.938	**0.973**
Urinary bladder neoplasms	0.885	0.955	0.949	0.758	0.949	0.950	**0.966**
Average AUC	0.898	0.950	0.953	0.786	0.920	0.938	**0.970**

Table 3. Precision of different methods by using local 5-fold cross validation on 15 diseases.

Disease	RWRMDA	RLSMDA	MIDP	SPM	NNMDA	LOMDA-MD	LOMDA
Breast neoplasms	0.586	0.733	0.720	0.557	0.777	0.678	**0.799**
Hepatocellular carcinoma	0.478	0.676	0.603	0.624	0.739	0.608	**0.816**
Renal cell carcinoma	0.227	0.240	0.482	0.276	0.481	0.208	**0.516**
Squamous cell carcinoma	0.025	0.166	0.067	0.106	**0.389**	0.164	0.241
Colorectal neoplasms	0.301	0.493	0.411	0.321	0.406	0.429	**0.614**
Glioblastoma	0.101	0.427	0.239	0.258	0.312	0.347	**0.511**
Heart failure	0.116	0.220	0.362	0.248	0.523	0.175	**0.612**
Acute myeloid leukemia	0.002	0.257	0.182	0.185	0.377	0.267	**0.618**
Lung neoplasms	0.288	0.584	0.572	0.343	0.455	0.530	**0.626**
Melanoma	0.190	0.501	0.391	0.357	**0.552**	0.443	0.546
Ovarian neoplasms	0.119	0.432	0.317	0.242	0.354	0.435	**0.523**
Pancreatic neoplasms	0.100	0.424	0.311	0.234	0.330	0.366	**0.596**
Prostatic neoplasms	0.085	**0.315**	0.286	0.158	0.181	0.254	0.289
Stomach neoplasms	0.402	0.526	0.623	0.421	0.245	0.448	**0.704**
Urinary bladder neoplasms	0.089	0.260	0.314	0.263	**0.516**	0.228	0.473
Average *Precision*	0.207	0.417	0.392	0.306	0.442	0.372	**0.566**

Table 4. Accuracy of different methods by using local 5-fold cross validation on 15 diseases.

Disease	RWRMDA	RLSMDA	MIDP	SPM	NNMDA	LOMDA-MD	LOMDA
Breast neoplasms	0.926	0.951	0.949	0.919	0.943	0.941	**0.963**
Hepatocellular carcinoma	0.900	0.938	0.925	0.928	0.932	0.925	**0.965**
Renal cell carcinoma	0.961	0.962	0.975	0.964	0.942	0.961	**0.976**
Squamous cell carcinoma	0.976	0.980	0.976	0.977	0.956	0.980	**0.981**
Colorectal neoplasms	0.931	0.950	0.941	0.932	0.912	0.944	**0.962**
Glioblastoma	0.955	0.971	0.964	0.963	0.939	0.967	**0.976**
Heart failure	0.948	0.954	0.963	0.955	0.947	0.952	**0.977**
Acute myeloid leukemia	0.974	0.981	0.979	0.979	0.967	0.981	**0.990**
Lung neoplasms	0.931	0.960	0.959	0.937	0.932	0.955	**0.964**
Melanoma	0.928	0.956	0.945	0.942	0.933	0.950	**0.959**
Ovarian neoplasms	0.951	0.969	0.963	0.958	0.940	0.967	**0.972**
Pancreatic neoplasms	0.946	0.965	0.961	0.954	0.942	0.961	**0.976**
Prostatic neoplasms	0.946	**0.960**	0.958	0.949	0.921	0.958	**0.960**
Stomach neoplasms	0.923	0.940	0.951	0.924	0.903	0.929	**0.961**
Urinary bladder neoplasms	0.967	0.973	0.976	0.973	0.962	0.971	**0.980**
Average *AUC*	0.944	0.961	0.959	0.950	0.938	0.956	**0.971**

Case studies

In this subsection, two detailed case studies for five critical diseases including breast neoplasms, colon neoplasms, kidney neoplasms, hepatocellular carcinoma and lung cancer have been investigated by LOMDA. These five diseases have attracted widespread attention from the general public and academia. Breast neoplasms is the most common malignant tumor in women accounting for 25% [49] followed by prostate and colon cancer [50, 51]. Colon neoplasm is one of the common cancers which has high death rate [52]. The incidence of kidney cancer has increased 43% since 1973, and the risk of the disorder is higher in men than in women and increases with age [53]. Hepatocellular cancer is the third leading cause of cancer-related deaths worldwide [54]. Lung cancer is till a leading cause of cancer death in both men and women in the United States even though its extensive list of risk factors has been characterized [55].

In the first case study, we test the performance of LOMDA in predicting new associated miRNAs for breast neoplasms, colon neoplasms and kidney neoplasms based on **MD** (HMDDv2.0; 2014), **DS** (HumanNet database; 2011) and **MS** (miRTarBase; 2014, miRBase; 2003, MISIM; 2010). All the known associations in **MD** together with all the similarity information **MS** and **DS** are utilized to do prediction. Three independent or updated databases are introduced to verify the the prediction results, including dbDEMC (2010) [25], miR2Disease (2009) [26] and HMDDv3.2 (2019) [56]. First of all, for the target disease, we compute scores for all the candidate associations then sort them in descending order. After that, we select the top 30 association scores of the interested disease and manually verify the existences of the associations by the above three databases. Results for three diseases including breast neoplasms, colon neoplasms, and kidney neoplasms are shown in Table 5, Table 6, and Table 7, respectively. We can find that 30 out of the top 30 predicted breast neoplasms-related miRNAs are confirmed, 24 out of the top 30 predicted colon neoplasms-related miRNAs are confirmed, and 26 of the top 30 predicted kidney neoplasms-related miRNAs are confirmed.

Table 5. The top 30 predicted miRNAs associated with breast neoplasms. The first column records top 1-15 related miRNAs, and the third column records top 16-30 miRNAs.

miRNA	Evidence	miRNA	Evidence
hsa-mir-106a	dbDEMC; HMDDv3.2	hsa-mir-138	dbDEMC; HMDDv3.2
hsa-mir-142	HMDDv3.2	hsa-mir-574	HMDDv3.2
hsa-mir-130a	dbDEMC; HMDDv3.2	hsa-mir-15b	dbDEMC; HMDDv3.2
hsa-mir-99a	dbDEMC; HMDDv3.2	hsa-mir-19b	dbDEMC; HMDDv3.2
hsa-mir-150	dbDEMC; HMDDv3.2	hsa-mir-30e	HMDDv3.2
hsa-mir-378a	HMDDv3.2	hsa-mir-542	HMDDv3.2
hsa-mir-186	dbDEMC	hsa-mir-650	dbDEMC
hsa-mir-92b	dbDEMC; HMDDv3.2	hsa-mir-98	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-185	dbDEMC; HMDDv3.2	hsa-mir-370	dbDEMC; HMDDv3.2
hsa-mir-130b	dbDEMC; HMDDv3.2	hsa-mir-99b	dbDEMC
hsa-mir-192	dbDEMC; HMDDv3.2	hsa-mir-95	dbDEMC
hsa-mir-212	dbDEMC; HMDDv3.2	hsa-mir-32	dbDEMC; HMDDv3.2
hsa-mir-449b	dbDEMC	hsa-mir-196b	dbDEMC; HMDDv3.2
hsa-mir-372	dbDEMC; HMDDv3.2	hsa-mir-449a	dbDEMC; HMDDv3.2
hsa-mir-330	dbDEMC	hsa-mir-181c	dbDEMC; HMDDv3.2

Table 6. The top 30 predicted miRNAs associated with colon neoplasms. The first column records top 1-15 related miRNAs, and the third column records top 16-30 miRNAs.

miRNA	Evidence	miRNA	Evidence
hsa-mir-200a	dbDEMC; HMDDv3.2	hsa-mir-224	dbDEMC; HMDDv3.2
hsa-mir-92a	dbDEMC; HMDDv3.2	hsa-mir-429	dbDEMC; HMDDv3.2
hsa-mir-29b	dbDEMC; HMDDv3.2	hsa-mir-34b	*unconfirmed*
hsa-mir-222	dbDEMC; HMDDv3.2	hsa-mir-199a	*unconfirmed*
hsa-mir-34c	*unconfirmed*	hsa-mir-148a	dbDEMC; HMDDv3.2
hsa-mir-100	dbDEMC	hsa-mir-27a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-210	dbDEMC; HMDDv3.2	hsa-mir-181b	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-182	dbDEMC	hsa-mir-195	dbDEMC; HMDDv3.2
hsa-mir-375	dbDEMC; HMDDv3.2	hsa-mir-20b	dbDEMC
hsa-mir-1-2	dbDEMC	hsa-mir-150	dbDEMC; HMDDv3.2
hsa-mir-203	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-103a	*unconfirmed*
hsa-mir-181a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-181a-2	*unconfirmed*
hsa-mir-30d	dbDEMC; HMDDv3.2	hsa-mir-146b	dbDEMC
hsa-mir-99a	dbDEMC	hsa-mir-25	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-29c	dbDEMC	hsa-mir-151a	*unconfirmed*

Table 7. The top 30 predicted miRNAs associated with kidney neoplasms. The first column records top 1-15 related miRNAs, and the third column records top 16-30 miRNAs.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-34c	dbDEMC
hsa-mir-146a	dbDEMC	hsa-mir-30a	dbDEMC
hsa-mir-17	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-638	dbDEMC
hsa-mir-200a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-320a	dbDEMC
hsa-mir-451a	dbDEMC	hsa-mir-499a	*unconfirmed*
hsa-mir-200b	dbDEMC; miR2Disease	hsa-mir-1207	*unconfirmed*
hsa-mir-433	*unconfirmed*	hsa-let-7b	dbDEMC
hsa-mir-192	dbDEMC; HMDDv3.2	hsa-mir-362	dbDEMC; HMDDv3.2
hsa-mir-198	dbDEMC	hsa-mir-148b	*unconfirmed*; HMDDv3.2
hsa-mir-205	dbDEMC; miR2Disease	hsa-mir-199a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-423	dbDEMC	hsa-mir-15a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-200c	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-181a	dbDEMC
hsa-mir-30b	dbDEMC	hsa-mir-206	dbDEMC
hsa-mir-10a	dbDEMC	hsa-mir-182	dbDEMC; miR2Disease
hsa-mir-377	dbDEMC	hsa-mir-196b	dbDEMC

In the second case study, we verify the prediction performance of LOMDA on diseases without any known associations based on **MD** (HMDDv2.0; 2014), **DS** (HumanNet database; 2011) and **MS** (miRTarBase; 2014, miRBase; 2003, MISIM; 2010). Hepatocellular carcinoma and lung neoplasms are taken as examples. we remove all known associations of hepatocellular carcinoma and lung neoplasms from **MD** in **A**, respectively. Then we compute the likelihood scores of these diseases with all the miRNAs by using LOMDA. Finally, we select the top 30 candidates and manually check these candidates by the above three databases. All these predicted candidates belonging to hepatocellular carcinoma can be confirmed by at least one of the three databases and 29 among 30 predicted associations of lung neoplasms are also confirmed. The results are shown in Table 8 and Table 9, respectively.

Table 8. The top 30 miRNAs associated with hepatocellular carcinoma were predicted by LOMDA with hiding all known related miRNAs.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-18a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-155	dbDEMC; HMDDv3.2	hsa-mir-19b-1	dbDEMC; HMDDv3.2
hsa-mir-146a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-221	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-17	dbDEMC; HMDDv3.2	hsa-mir-29b-2	HMDDv3.2
hsa-mir-20a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-16-2	dbDEMC; HMDDv3.2
hsa-mir-125b-1	miR2Disease; HMDDv3.2	hsa-mir-223	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-34a	miR2Disease; HMDDv3.2	hsa-mir-19a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-29a	dbDEMC; HMDDv3.2	hsa-mir-29c	dbDEMC; HMDDv3.2
hsa-mir-125b-2	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-181a-1	miR2Disease; HMDDv3.2
hsa-mir-15a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-150	dbDEMC; HMDDv3.2
hsa-mir-29b-1	dbDEMC; HMDDv3.2	hsa-mir-1-1	HMDDv3.2
hsa-mir-92a-1	HMDDv3.2	hsa-mir-181a-2	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-126	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-92a-2	HMDDv3.2
hsa-mir-145	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-122	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-16-1	HMDDv3.2	hsa-mir-1-2	dbDEMC; miR2Disease; HMDDv3.2

Table 9. The top 30 miRNAs associated with lung neoplasms were predicted by LOMDA with hiding all known related miRNAs.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-146a	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-34c	dbDEMC; HMDDv3.2	hsa-mir-30e	dbDEMC; HMDDv3.2
hsa-mir-155	dbDEMC; HMDDv3.2	hsa-mir-152	dbDEMC; HMDDv3.2
hsa-mir-486	dbDEMC; HMDDv3.2	hsa-mir-219-2	*unconfirmed*
hsa-mir-15a	dbDEMC; HMDDv3.2	hsa-mir-223	HMDDv3.2
hsa-mir-34a	dbDEMC; HMDDv3.2	hsa-mir-520c	HMDDv3.2
hsa-mir-148b	dbDEMC; HMDDv3.2	hsa-mir-137	dbDEMC; HMDDv3.2
hsa-mir-126	dbDEMC; HMDDv3.2	hsa-mir-134	dbDEMC
hsa-mir-29a	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-125b	dbDEMC; miR2Disease; HMDDv3.2
hsa-mir-200b	dbDEMC; miR2Disease; HMDDv3.2	hsa-mir-204	dbDEMC; HMDDv3.2
hsa-mir-326	dbDEMC; HMDDv3.2	hsa-mir-374a	HMDDv3.2
hsa-mir-34b	dbDEMC; HMDDv3.2	hsa-mir-7b	miR2Disease; HMDDv3.2
hsa-mir-337	dbDEMC; HMDDv3.2	hsa-mir-375	dbDEMC; HMDDv3.2
hsa-mir-25	dbDEMC; HMDDv3.2	hsa-mir-92b	dbDEMC; HMDDv3.2
hsa-mir-122	dbDEMC HMDDv3.2	hsa-mir-28	dbDEMC; HMDDv3.2

Conclusion

Predicting novel associations between miRNAs and diseases helps scientists firstly focus on the most likely associations rather than blindly check on all possible associations which is extremely costly and laborious. Moreover, it can help researchers enhance their understanding toward molecular mechanisms of diseases at the miRNA level. This prediction also plays an important role in understanding the pathogenesis of human diseases at the early stage, therefore, it can help in diagnosis, treatment and prevention.

Motivated by the necessity of identifying novel associations between miRNAs and diseases, in this work we propose a computational method, namely linear optimization for miRNA-disease association (LOMDA) prediction. The proposed method utilizes the heterogenous matrix by integrating miRNA functional similarity information, disease similarity information, and known miRNA-disease associations. In case only known associations are available, the method can also be applied. Moreover, the method can also predict associations for new miRNAs (or diseases) by using miRNA functional similarity information (or disease similarity information). According to the cross validation evaluated by AUC, precision, and accuracy, the proposed method has been shown to perform very satisfactorily. Thus, LOMDA is an effective and efficient tool for predicting miRNA-disease associations.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11975071 and 61703074).

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–297.

2. Chatterjee S, Großhans H. Active turnover modulates mature microRNA activity in Caenorhabditis elegans. Nature. 2009;461(7263):546.

3. Llave C, Xie Z, Kasschau KD, Carrington JC. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science. 2002;297(5589):2053–2056.

4. Eulalio A, Huntzinger E, Izaurralde E. Getting to the root of miRNA-mediated gene silencing. Cell. 2008;132(1):9–14.

5. Zhu L, Zhao J, Wang J, Hu C, Peng J, Luo R, et al. MicroRNAs are involved in the regulation of ovary development in the pathogenic blood fluke Schistosoma japonicum. PLoS Pathogens. 2016;12(2):e1005423.

6. Fernando TR, Rodriguez-Malave NI, Rao DS. MicroRNAs in B cell development and malignancy. J Hematol Oncol. 2012;5(1):7.

7. Lize M, Pilarski S, Dobbelstein M. E2F1-inducible microRNA 449a/b suppresses cell proliferation and promotes apoptosis. Cell Death and Differ. 2010;17(3):452.

8. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. Nat Rev Cancer. 2006;6(4):259.

9. Duisters RF, Tijsen AJ, Schroen B, Leenders JJ, Lentink V, Made Ivd, et al. miR-133 and miR-30 regulate connective tissue growth factor: implications for a role of microRNAs in myocardial matrix remodeling. Circ Res. 2009;104(2):170–178.

10. Lynam-Lennon N, Maher SG, Reynolds JV. The roles of microRNA in cancer and apoptosis. Biol Rev. 2009;84(1):55–71.

11. Etheridge A, Lee I, Hood L, Galas D, Wang K. Extracellular microRNA: a new source of biomarkers. Mutat Res - Fund Mol M. 2011;717(1-2):85–90.

12. Espinosa CES, Slack FJ. Cancer issue: the role of microRNAs in cancer. Yale J Biol Med. 2006;79(3-4):131.

13. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006;6(11):857.

14. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–838.

15. Markou A, Tsaroucha EG, Kaklamanis L, Fotinou M, Georgoulias V, Lianidou ES. Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non–small cell lung cancer by quantitative real-time RT-PCR. Clin Chem. 2008;54(10):1696–1704.

16. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005;65(16):7065–7070.

17. Huang Q, Gumireddy K, Schrier M, Sage Cl, Nagel R, Nair S, et al. The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. Nat Cell Biol. 2008;10(2):202–210.

18. Latronico MV, Catalucci D, Condorelli G. Emerging role of microRNAs in cardiovascular biology. Circ Res. 2007;101(12):1225–1236.

19. Chen RW, Bemis LT, Amato CM, Myint H, Tran H, Birks DK, et al. Truncation in CCND1 mRNA alters miR-16-1 regulation in mantle cell lymphoma. Blood. 2008;112(3):822–829.

20. Xin F, Li M, Balch C, Thomson M, Fan M, Liu Y, et al. Computational analysis of microRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. Bioinformatics. 2008;25(4):430–434.

21. Xu J, Li CX, Lv JY, Li YS, Xiao Y, Shao TT, et al. Prioritizing candidate disease miRNAs by topological features in the miRNA target–dysregulated network: Case study of prostate cancer. Mol Cancer Ther. 2011;10(10):1857–1866.

22. Yu Z, Li Z, Jolicoeur N, Zhang L, Fortin Y, Wang E, et al. Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. Nucleic Acids Res. 2007;35(13):4535–4541.

23. Xiao Y, Guan J, Ping Y, Xu C, Huang T, Zhao H, et al. Prioritizing cancer-related key miRNA–target interactions by integrative genomics. Nucleic Acids Res. 2012;40(16):7653–7665.

24. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, et al. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 2014;42(D1):D1070–D1074.

25. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, et al. dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics. 2010;11(S4):S5.

26. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009;37(suppl_1):D98–D104.

27. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Brief Bioinform. 2015;17(2):193–203.

28. Pasquier C, Gardès J. Prediction of miRNA-disease associations with a vector space model. Sci Rep. 2016;6:27036.

29. Li JQ, Rong ZH, Chen X, Yan GY, You ZH. MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget. 2017;8(13):21187.

30. Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused LASSO on multiple data sets. Sci Rep. 2016;6:20533.

31. Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. Int J Data Min Bioinform. 2013;8(3):282–293.

32. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y, et al. RBMMMDA: predicting multiple types of disease-microRNA associations. Sci Rep. 2015;5:13877.

33. Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, et al. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol. 2013;7(1):101.

34. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA–disease associations. Mol Biosyst. 2012;8(10):2792–2798.

35. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. BMC Med Genet. 2013;6(1):12.

36. Sun D, Li A, Feng H, Wang M. NTSMDA: prediction of miRNA–disease associations by integrating network topological similarity. Mol Biosyst. 2016;12(7):2224–2232.

37. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010;26(13):1644–1650.

38. Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, et al. Prediction of microRNA-disease associations based on social network analysis methods. BioMed Res Int. 2015;2015.

39. Chen M, Lu X, Liao B, Li Z, Cai L, Gu C. Uncover miRNA-Disease Association by Exploiting Global Network Similarity. PloS ONE. 2016;11(12):e0166509.

40. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, et al. Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics. 2015;31(11):1805–1815.

41. Chen X, Yan GY. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014;4:5501.

42. Zeng X, Liu L, Lü L, Zou Q. Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics. 2018;34(14):2425–2432.

43. Zeng X, Wang W, Deng G, Bing, Quan Z. Prediction of potential disease-associated microRNAs by using neural networks. Mol Ther Nucleic Acids. 2019;16(14):566–575.

44. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genet. 2008;40(2):181–188.

45. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21(7):1109–1121.

46. Hsu SD, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2014;42(D1):D78–D85.

47. Pech R, Hao D, Lee YL, Yuan Y, Zhou T. Link prediction via linear optimization. Physica A. 2019;528:121319.

48. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.

49. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69–90.

50. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. Proc Natl Acad Sci USA. 2003;100(7):3983–3988.

51. Potter JD, Slattery ML, Bostick RM, Gapstur SM. Colon cancer: a review of the epidemiology. Epidemiol Rev. 1993;15(2):499–545.

52. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7–30.

53. Vogelzang NJ, Stadler WM. Kidney cancer. The Lancet. 1998;352(9141):1691–1696.

54. Parikh S, Hyman D. Hepatocellular cancer: a guide for the internist. Am J Med. 2007;120(3):194–202.

55. Torre LA, Siegel RL, Jemal A. Lung cancer statistics. In: Ahmad A, Gadgeel S, editors. Lung cancer and personalized medicine: current knowledge and therapies; 2016. p. 1–19.

56. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, et al. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res. 2019;47(D1):D1013–D1017.