The Author(s) *BMC Systems Biology* 2016, **10**(Suppl 4):117
DOI 10.1186/s12918-016-0357-1

BMC Systems Biology

RESEARCH Open Access

Gene-microRNA network module analysis for ovarian cancer

Shuqin Zhang[1*] and Michael K. Ng[2]

From The 27th International Conference on Genome Informatics
Shanghai, China. 3-5 October 2016

Abstract

Background: MicroRNAs (miRNAs) are involved in many biological processes by regulating post-transcriptional gene expression. The alterations of the regulatory pathways can cause different diseases including cancer. Although many works have been done to study the gene-miRNA regulatory network, the intertwined relationship is far from being fully understood. The objective of this study is to integrate both gene expression data and miRNA data so as to explore the complex relationships among them.

Methods: By integrating the networks consisting of gene coexpression, miRNA coexpression, gene-miRNA coexpression, and the known gene-miRNA interactions, we aim to find the most connected network modules so as to study their functions and properties. In this paper, we proposed an optimization model for identification of the modules in the integrated networks. This model tries to find both the modules in the gene-gene and miRNA-miRNA coexpression networks and the densely connected gene-miRNA subneworks. An approximation computational method was developed to solve the optimization problem.

Results: We applied the method to 556 human ovarian cancer samples with both gene expression data and miRNA expression data. The identified modules are significantly enriched by miRNA clusters, GO-BPs, and KEGG pathways. We compared our method with some existing methods and showed the better performance of our method. We also showed that the miRNAs and genes in our identified modules are associated with cancers, especially ovarian cancer.

Conclusions: This study provides strong support that the subnetworks consisting of genes and miRNAs with close interactions contribute the cancers. The proposed computational method can be applied to other studies that are related to different types of networks.

Keywords: Gene-miRNA network, Module identification, Data integration

Background

MicroRNAs (miRNAs) are small ('22 nucleotides) non-coding RNAs that have emerged as key gene regulators in diverse plant and animal genomes. Typically, miRNAs regulate the genes by base pairing with the complementary sequences of the corresponding mRNAs, either inhibiting translation or degrading the mRNAs [1, 2]. MiRNAs are involved in many biological processes, such as development, differentiation, apoptosis and proliferation [3–6]. Each miRNA is potentially able to regulate around 100 or more mRNA targets and over 30% of all human genes are supposed to be regulated by miRNAs [3, 6, 7]. The alterations in the regulatory pathways can cause different diseases, including cancer, heart disease, cardiovascular disease, and matabolc disorders [8–13]. The disruption of the miRNA functions will contribute to these diseases. Therefore, identification and validation of miRNA targets is essential, which may lead to new therapeutic methods [6, 14–16].

MiRNA targets prediction has attracted much attention in recent years. Although many experimental tools for miRNA target validation are available [17–22], the lack of high-throughput and low-cost methods makes the

*Correspondence: zhangs@fudan.edu.cn
[1] Center for Computational Systems Biology, School of Mathematical Sciences, Fudan University, No.220 Handan Road, 200433 Shanghai, China
Full list of author information is available at the end of the article

development of computational techniques necessary. The computational methods are mainly based on expression data of both miRNAs and mRNAs and the sequence-based putative interactions between them. Roughly, these methods can be divided into three groups. The first group includes methods that compute the pairwise correlations or mutual information between miRNAs and mRNAs [23–30]. The second group is mainly related to the linear regression methods [31–35]. And the third group is mainly based on Bayesian methods [36–39]. A review of these methods can be found in [6].

Although huge advances have been made for miRNA target prediction, a lot of works are still left to do. Most of the current methods mainly considered the down-regulatory effects from miRNAs. However, as shown in [6], by considering the enhancement effects from miRNAs, more interactions are identified and biologically sound. Also, as shown in [40], the base pairing between an miRNA and its target not only results in repressed target expression, but also have an impact on the levels of the miRNA. And as one miRNA may regulate many mRNAs, and be regulated by several miRNAs, the intertwined relationship between miRNAs and mRNAs becomes very complex. Therefore, the systems tools such as networks should be more appropriate for studying the relationships between miRNAs and mRNAs.

Recently, a few papers have been published to study the complex interactions between genes and miRNAs from the system point of view by using networks [41–43]. One essential property of many different types of networks is the module structure, which describes the densely connected subnetworks. The members in the same module may function as a whole in the system. By identifying the modules, we can do gene prediction, gene function annotation and so on. In the networks composed of both genes and miRNAs, a good module should include both genes and miRNAs with connections. To better identify the gene-miRNA modules, Zhang et al. developed a framework of SNMNMF, which is based on non-negative matrix factorization and utilized a variety of data, including gene-gene interaction (GGI) and transcription factor binding sites (TFBS) [41]. However, SNMNMF tends to pay more attention to the gene modules while overlook the connections between miRNAs. Also, its computational speed may limit the practical use of the method. Then, using similar datasets, Le et al. described a regression-based method, PIMiM36 (Protein Interaction based MiRNA Modules) [42], but using a non-convex algorithm. This method may result in unstable outcomes because of its random initialization. Then, Li et al. developed a two-stage overlap clustering method, Mirsynergy [43]. This method improves the efficiency substantially, and importantly, facilitates the setting of predefined parameters. However, this method does not consider the relations between miRNAs, and thus finds the modules with a large number of genes/miRNAs, which are shown in their enrichment analysis.

In this paper, we establish networks to explore gene-miRNA relationships. We first integrate gene expression and miRNA expression data by measuring the distance between genes, miRNAs and gene-miRNAs with Pearson correlation coefficient, thus transferring all the relations into edges in networks. Based on these networks, we propose our module identification method, and study the gene-miRNA interactions. We also include the known gene-miRNA interactions in our study. In the second section, we will present our method for identifying the modules in the integrated networks. Then we apply the method to the ovarian cancer data to show its performance. We also compared our method with Mirsynergy. Finally, we give our conclusion.

Methods

Assume we have gene expression data of N_g genes, and miRNA expression data of N_m miRNAs for N samples. The first step is to build the coexpression networks. We use Pearson correlation coefficient to measure the coexpressions. After computing the correlations of genes, miRNAs, and gene-miRNAs, we construct the adjacency matrix by hard thresholding. If the absolute value of the Pearson correlation coefficient between genes(miRNAs, gene-miRNAs) is greater than some given value, we assign an edge between them; otherwise, there is no edge. For gene coexpression network and miRNA coexpression network, we try different thresholds and compute the linear regression coefficient between the $\log_{10}$ transformed degree frequency of degree d ($\log_{10} f(d)$) and d ($\log_{10} d$) to make the network has approximately scale free property as described in [44]. The threshold for constructing the gene-miRNA network depends on the AUCs for the known gene-miRNAs interactions being clustered in the same module. Given a fixed threshold, we use our method proposed in the following to get the score of one gene and one miRNA in the same module. We rearrange the gene-miRNA interactions in the descending order according to their scores and compute the AUC as AUC $= \frac{\sum_{i=1}^{q} R_i - q(q+1)/2}{pq}$, where $\{R_i\}$ is the rank of the i_{th} known interacting gene-miRNA pair ranking from the smallest, p is the number of known non-interacting gene-miRNA pairs, and q is the number of known interacting gene-miRNA pairs. We choose the threshold that achieves the highest AUC.

We consider the constructed gene coexpression network G_g and the miRNA coexpression network G_m. The adjacency matrix for network G_g is A_g, where $A_g(i,j) = 1$ represents there is an edge between gene i and gene j. Similarly, we define the adjacency matrix A_m for the miRNA coexpression network. We use D_g, D_m to denote the diagonal matrix with the diagonal entries being the degree of the corresponding gene/miRNA, where the degree of gene i is defined as $d_g(i) = \sum_{j=1}^{N_g} A_g(i,j)$. The interaction matrix between the genes and miRNAs is denoted as $C_{N_g \times N_m}$, where $C(i,j) = 1$ represents there is a connection between the corresponding gene and miRNA. In our study, C includes two parts: the gene-miRNA coexpression network, and the known gene-miRNA interaction network. The integrated network is composed of the above four networks. Here, we assume the integrated network is connected. If the network is unconnected, we may divide it into connected parts with some classical methods like spectral clustering. We define the modules of the integrated network to be the densely connected subnetworks consisting of both genes and miRNAs. Assume there are K modules in the network. We let S_g be the assignment of the N_g genes into K modules for the network G_g,

$$S_g(i,k) = \begin{cases} 1, \text{ if vertex } i \in V_k, \\ 0, \text{ otherwise,} \end{cases}$$

where $i = 1, 2, \cdots, N_g; k = 1, 2, \cdots, K$, V_k denotes the k-th module. Similarly, we define the assignment of the N_m miRNAs into K modules for the network G_m as S_m.

For each of the two coexpression networks, we may use a module identification method [45] to cluster the genes/miRNAs separately. This method has shown to outperform most updated methods including spectral clustering in module identification. Taking our considered network G_g as an example, we define

$$\Psi_g(S_g) = \sum_{k=1}^{K} \frac{S_g(.,k)^T (2A_g - D_g) S_g(.,k)}{S_g(.,k)^T S_g(.,k)}, \tag{1}$$

where $S_g(.,k)$ denotes the k-th column of matrix S_g for the network G_g. Then the optimization problem for identifying the modules is formulated as:

$$\begin{aligned} \max \quad & \Psi_g(S_g) \\ s.t. \quad & S_g(i,k) \in \{0,1\},\ i = 1, 2, \cdots, N_g, k = 1, 2, \cdots, K, \\ & \sum_{k=1}^{K} S_g(.,k) = \mathbf{1}, \end{aligned} \tag{2}$$

where $\mathbf{1}$ is a vector with all the entries being 1. By letting $\tilde{S}_g(.,k) = \frac{S_g(.,k)}{\|S_g(.,k)\|_2}$, the problem is relaxed to:

$$\max\ \tilde{\Psi}_g(\tilde{S}_g) = \text{Tr}\left(\tilde{S_g}^T (2A_g - D_g) \tilde{S}_g\right) \quad s.t.\ \tilde{S_g}^T \tilde{S}_g = I_K.$$

Let $L_g = 2A_g - D_g$, we can use the standard procedure of spectral clustering to get the module label for the network.

Similarly, we can define the optimization problem for clustering the miRNAs in the miRNA coexpression network. We use $\Psi_m(S_m)$ to denote the objective function when doing module identification for miRNAs, and define $\tilde{S}_m(.,k) = \frac{S_m(.,k)}{\|S_m(.,k)\|_2}$. To find the modules in the integrated network, besides considering the connections in both gene coexpression and miRNA coexpression networks, we expect that the genes and miRNAs with dense connections are clustered into one module. That is, we want to maximize $S_g{}^T(.,k) C S_m(.,k)$. To balance the size of genes and miRNAs in different modules , we divide the term $S_g^T(.,k) C S_m(.,k)$ by $\|S_g(.,k)\|_2 \|S_m(.,k)\|_2$. By putting all these terms together, our objective becomes:

$$\begin{aligned} \Psi(S_g, S_m) = \Psi_g(S_g) + \Psi_m(S_m) \\ + \lambda \sum_{k=1}^{K} \frac{S_g^T(.,k) C S_m(.,k)}{\|S_g{}^T(.,k)\|_2 \|S_m(.,k)\|_2}, \end{aligned}$$

where λ controls the contributions of the connections within each coexpression network and those between the two networks. The optimization problem is formulated as:

$$\begin{aligned} \max \quad & \Psi(S_g, S_m) \\ s.t. \quad & S_g(i,k) \in \{0,1\}, i = 1, 2, \cdots, N_g, k = 1, 2, \cdots, K, \\ & S_m(j,k) \in \{0,1\}, j = 1, 2, \cdots, N_m, k = 1, 2, \cdots, K, \\ & \sum_{k=1}^{K} S_g(\cdot,k) = \mathbf{1}, \sum_{k=1}^{K} S_m(\cdot,k) = \mathbf{1}. \end{aligned}$$

We define $L_g = 2A_g - D_g, L_m = 2A_m - D_m, L_w = diag(L_g, L_m), L_b = \begin{pmatrix} \mathbf{0} & C \\ C^T & \mathbf{0} \end{pmatrix}$, $\tilde{S} = \begin{pmatrix} \tilde{S}_g \\ \tilde{S}_m \end{pmatrix}$, and $L = L_w + \lambda L_b$.

With the same technique as in (2), the above optimization problem can be relaxed to:

$$\max\ \tilde{\Psi}(\tilde{S}) = \text{Tr}(\tilde{S}^T L \tilde{S}), \quad s.t.\ \tilde{S}^T \tilde{S} = 2I_K.$$

In this formulation, the constant coefficient 2 can be put into L, such that each column of $\tilde{S}$ has the norm 1. We take $\tilde{S}$ as a data set composed of N_g+N_m nodes and do k-means clustering to get the assignment label for each node.

The algorithm is summarized in the following.

Algorithm:

Input: Adjacency matrix A_g, A_m, C, and K, which is the number of modules.

```
1  Compute the matrices $L_g, L_m$;
2  Construct the matrix $L$;
3  Compute the $K$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$
   corresponding to the $K$ largest eigenvalues of matrix
   $L$;
4  Construct a new matrix $T \in R^{(N_g+N_m) \times K}$, with
   columns $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K$;
5  Cluster the points constructed from each row of
   matrix $T$ with $k$-means clustering into $K$ clusters;
```

Output: Index of nodes in each module.

With this algorithm, we can identify the modules in the integrated network. Here, K is prespecified as the number of modules in the integrated network. Since any node is assigned to a module, in practice, the nodes in some module may not connect densely. Additionally, some module may consist of genes or miRNAs only. Therefore, after implementing our algorithm, we need to check the structure of the clusters to make sure the identified modules are all densely connected subnetworks with both genes and miRNAs.

Results

Data sets

We downloaded the level 2 gene expression and miRNA expression data for ovarian cancer from The Cancer Genome Atlas (TCGA). The gene expression data is generated with UNC AgilentG4502A_07_03, and the miRNA expresssion data is generated with UNC miRNA_8x15kv2. For the gene expression data, we averaged the expression data for different probes corresponding to the same gene, and used the average of different samples to represent the missing data for a specific gene. After the preprocessing, we have 22747 genes with annotations for the 562 samples. We did the same process for the miRNA expression data, and finally we have 590 miRNAs for the 595 samples. We chose the data for the common 556 samples. We downloaded the gene-miRNA interaction data from miRTarBase (*http://mirtarbase.mbc.nctu.edu.tw*). There are a total of 39110 gene-miRNA interactions.

Network construction

We computed the variance of the expression values for all genes across the considered samples, and selected those genes with large variance. Here, we selected the first 3200 genes with the largest expression variance, which corresponds to the variance greater than 1. We used the method described in the "Methods" section to build the gene coexpression and miRNA coexpression network. We chose a threshold of 0.60 such that the degree of both coexpression networks follows power law distribution, and the main correlations are kept. The average degree of the gene coexpression network and miRNA coexpression network is 13.25 and 8.80, repectively. Here, we set $\lambda = 1$. By choosing the threshold for building gene-miRNA coexpression network, we aim to get a good clustering of genes and miRNAs. We choose the threshold from 0.1 to 0.9 with a stepsize 0.1. For each value, we build the gene-miRNA coexpression network. Then we run our algorithm by setting C being this network adjacency matrix. After we get the matrix T as shown in our algorithm, we normalize each row of T denoted as $\tilde{T}$ and compute $\tilde{T}\tilde{T}^T$. It is easy to figure out that the *ij*-th score in $\tilde{T}\tilde{T}^T$ describes the possibility of the *i*-th subject and the *j*-th subject in the same module. We set the number of modules K to be 100 to 200 with a stepsize 10, and compute all the AUCs. Figure 1 shows the average AUCs for different K. When the threshold is 0.3, we can cluster the known gene-miRNAs in the same module with the highest AUC of value 0.59. Thus we choose the threshold to be 0.3. The mean of the absolute value of correlation coefficients is 0.07(gene-miRNA), 0.10(gene-gene), and 0.12(miRNA-miRNA). We searched the corresponding interactions between the 3200 genes and the 590 miRNAs in our downloaded interaction data set, and totally there are 2648 interations.

Then the matrix C is composed of two types of elements: the known gene-miRNA interactions and the gene-miRNA coexpression network. We note that the best λ and threshold can be chosen by changing the values of λ and the threshold alternatively, and those achieving the highest AUC are selected.

Experimental results

We applied our proposed method to the integrated network. For the number of modules, we selected different values starting from 100 to 200. These different values correspond to the modules on different connection levels. We choose those clusters which satisfy: (1) both genes and miRNAs are in the cluster, (2) all of gene-gene, miRNA-miRNA, and gene-miRNA connections appear in the cluster, as our identified modules. With the different choices of number of modules, one gene/miRNA may belong to different modules. We combine those modules that have an overlapping percentage larger than 90%. Finally, we got 46 modules. The full list of identified modules is in Additional file 1.

MiRNA module enrichment analysis

We downloaded miRNA cluster data from the miRBase website (*http://www.mirba se.org/*), with the inter-miRNA distance cutoff of 10 kb. This criterion resulted in 153 clusters containing from 2 to 46 miRNAs. In this section, 'cluster' means these clusters.

We compared our identified miRNA modules that are included in the gene-miRNA modules with the downloaded clusters. We did enrichment for both clusters and modules to see whether the modules are enriched by the clusters, and whether the clusters are enriched by the modules. We use hypergeometric distribution to do the test, and then use Bonferroni correction to adjust the p-values. There are a total of 14 modules enriched by clusters, and 8 clusters enriched by modules with the overlap size between clusters and modules being at least 3. Table 1 shows the information of the enriched modules by different clusters. The "No." of modules is the corresponding column in Additional file 1. The column "MiRNAs" lists

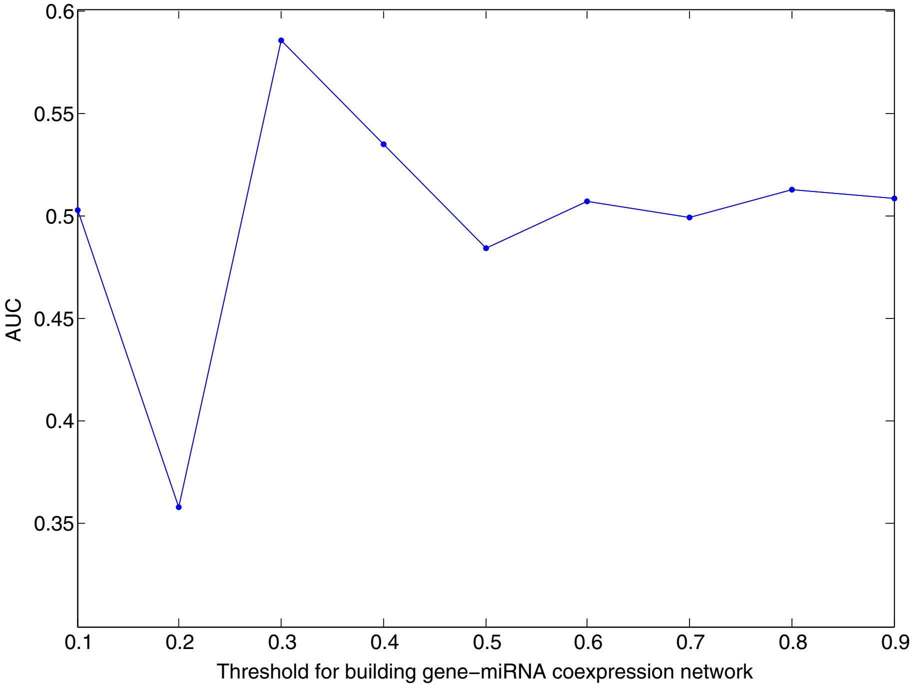

Fig. 1 AUCs for different cutoffs when building the gene-miRNA coexpression network

the overlapping miRNAs with the corresponding clusters. The loci of the clusters are also given. One typical example is our identified module 22. All the nodes in the same cluster belong to this module. The total distance for all these miRNAs is around 6000 bp. Taking as another example, 5 of 12 miRNAs in module 35 belong to a cluster with size 6 in Chr13. The total distance of this cluster is about 700 bp.

Gene module enrichment analysis

To evaluate the performance of our proposed method for gene module identification, we did enrichment analysis for Gene Ontology biological process (GO-BP) terms and KEGG pathways with DAVID [46, 47]. By taking the cutoff of the Benjamini *p*-values as 0.05, 15 modules are enriched by GO-BP terms and 7 modules are enriched by KEGG pathways significantly. Table 2 listed the enriched KEGG pathways. All the *p*-values are Benjamin *p*-values from DAVID. The KEGG pathway enrichment results of the modules in Table 1 are also listed. Two typical modules are module 41 and module 18, which have significant enrichment of GO-BP, KEGG pathways and miRNAs clusters. From the KEGG pathway enrichment results, it can be seen that these two modules are quite related to cancers. We put all these enrichment results in Additional file 3.

Gene-miRNA modules are strongly associated with cancers

We checked the 14 modules that are enriched by miRNA clusters, of which 7 modules have enrichment of cancer related pathways. From the KEGG pathway enrichment results, we can directly see that module 18, 41 are associated with different cancers. For module 21, the 'focal adhesion' pathway is known to be involved in tumour formation and progression [48]. The pathway 'ECM-receptor interaction' enriched by module 22 is also identified to be linked to carcinogenesis in multiple cancers [49]. Since the discovery of the MAPK signalling pathway, which is enriched in module 37, the enormous role of perturbed MAPK signaling in cancer biology has become evident. Specifically, more than 30% of human cancers include mutations in genes encoding proteins in this pathway [50]. Such evidence further shows that the modules enriched by miRNA clusters are likely to be enriched by cancer related pathways.

We checked the cancer related miRNAs from the website: *http://mircancer.ecu.edu*. There are 295 different miRNAs related to cancer, of which 122 are in our identified modules. 57 of the 295 miRNAs are related to ovarian cancer, of which 29 are in our identified modules, which achieves a *p*-value 0.0386. This suggests that the modules we identified are related to ovarian cancer significantly.

For the modules listed in Table 1, all of them have cancer related miRNAs. We listed the number of cancer associated miRNAs in Table 3. 13 of the 14 modules are enriched by cancer associated miRNAs significantly. In module 1 and module 37, all the involved miRNAs are associated with cancers. Figure 2 shows our constructed network for module 37. 156 genes are regulated by the 12 miRNAs. There are a total of 47 known regulations. Figure 3 shows

Table 1 MiRNA module enrichment results

No.	*p*-value	MiRNAs	Loci
38	2.01E-18	miR-411, miR-299, miR-758, miR-329-1, miR-543, miR-495, miR-654, miR-376b, miR-376a-1, miR-381, miR-487b, miR-539, miR-487a, miR-382, miR-154, miR-377, miR-409, miR-369, miR-376c,miR-889,miR-410	Chr14 101022066-101066801
45	1.41E-14	miR-379, miR-411, miR-299, miR-758, miR-329-1, miR-543, miR-376c, miR-654, miR-376b, miR-376a-1, miR-381, miR-487a, miR-382, miR-154, miR-377, miR-409, miR-369, miR-495, miR-487b, miR-539, miR-410	Chr14 101022066--101066801
5	1.29E-15	miR-411, miR-758, miR-329-1, miR-543, miR-495, miR-376b, miR-376a-1, miR-487b,, miR-539, miR-889, miR-382 miR-154, miR-409, miR-369, miR-654,miR-487a, miR-410	Chr14 101022066--101066801
2	2.19E-02	miR-379, miR-299, miR-376c, miR-376a-1, miR-381, miR-377	Chr14 101022066--101062118
10	1.40E-03	miR-379, miR-299, miR-376c, miR-376a-1, miR-381, miR-377	Chr14 101022066--101062118
38	2.70E-05	miR-493, miR-337, miR-433, miR-127, miR-432, miR-136	Chr14 100869060--100884783
12	1.40E-03	miR-379, miR-299, miR-376c, miR-376a-1, miR-381, miR-377	Chr14 101022066--101062118
21	1.26E-02	miR-379, miR-299, miR-376c, miR-376a-1, miR-381, miR-377	Chr14 101022066--101062118
35	1.17E-06	miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1	Chr13 91350605--91351391
45	5.66E-03	miR-493, miR-337, miR-433, miR-127, miR-136	Chr14 100869000--100885000
18	2.13E-04	miR-424, miR-503, miR-542, miR-450a-1	ChrX 134546614--134540262
1	4.45E-05	miR-200b, miR-200a, miR-429	Chr1 1167104--1169087
10	2.02E-02	miR-337, miR-127, miR-136	Chr14 100869060--100884783
11	1.78E-03	miR-18b, miR-20b, miR-363	ChrX 134170198--134169452
22	2.83E-03	miR-508, miR-507, miR-506	Chr X 147236913--147230843
35	9.44E-04	miR-106b, miR-93, miR-25	Chr 7 100093993--100093643
37	9.44E-04	miR-106b, miR-93, miR-25	Chr 7 100093993--100093643
41	1.40E-02	miR-17, miR-19a, miR-20a	Chr 13 91350605--91351135

Table 2 Enriched KEGG pathways for the gene modules

No.	Enriched Pathways	*p*-value
17	Cytokine-cytokine receptor interaction	7.40E-05
	NOD-like receptor signaling pathway	1.20E-03
	Chemokine signaling pathway	5.80E-03
	Hematopoietic cell lineage	3.30E-02
	Complement and coagulation cascades	1.80E-01
	Systemic lupus erythematosus	2.70E-01
18	p53 signaling pathway	3.50E-03
	Small cell lung cancer	2.70E-03
	Cell cycle	4.00E-03
	Pathways in cancer	2.10E-02
	Non-small cell lung cancer	8.20E-02
	Glioma	8.00E-02
	Melanoma	7.70E-02
	Pancreatic cancer	6.90E-02
	Chronic myeloid leukemia	6.40E-02
	Prostate cancer	6.80E-02
24	Cytokine-cytokine receptor interaction	4.40E-05
	NOD-like receptor signaling pathway	9.40E-04
	Chemokine signaling pathway	4.30E-03
	Hematopoietic cell lineage	2.70E-02
	Complement and coagulation cascades	1.60E-01
	Systemic lupus erythematosus	2.40E-01
28	Antigen processing and presentation	4.00E-03
	Cytokine-cytokine receptor interaction	1.20E-02
	Natural killer cell mediated cytotoxicity	9.20E-02
	Hematopoietic cell lineage	1.30E-01
	Graft-versus-host disease	1.70E-01
	Chemokine signaling pathway	1.40E-01
	NOD-like receptor signaling pathway	2.70E-01
	Viral myocarditis	3.00E-01
31	Systemic lupus erythematosus	1.30E-02
41	Small cell lung cancer	1.30E-03
	Chronic myeloid leukemia	3.10E-02
	Pathways in cancer	2.40E-02
	Colorectal cancer	1.90E-02
	Cell cycle	3.40E-02
	Thyroid cancer	1.30E-01
	Bladder cancer	1.60E-01
	Endometrial cancer	1.80E-01
	Non-small cell lung cancer	1.60E-01
	Acute myeloid leukemia	1.60E-01
	Glioma	1.60E-01
	p53 signaling pathway	1.50E-01
	Melanoma	1.50E-01
	Pancreatic cancer	1.40E-01
	Prostate cancer	1.60E-01
43	ECM-receptor interaction	1.20E-09
	Focal adhesion	6.70E-06
	Vascular smooth muscle contraction	1.60E-01
21	Focal adhesion	8.80E-01
22	ECM-receptor interaction	8.30E-01
35	Acute myeloid leukemia	8.20E-01
	p53 signaling pathway	6.40E-01
	Chronic myeloid leukemia	5.20E-01
37	Hypertrophic cardiomyopathy (HCM)	7.00E-02
	Gap junction	2.70E-01
	Dilated cardiomyopathy	2.10E-01
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	4.60E-01
	MAPK signaling pathway	7.80E-01
45	Pathways in cancer	7.30E-01
	Basal cell carcinoma	5.00E-01
	Hedgehog signaling pathway	3.70E-01

the known regulatory interactions of genes and miRNAs in module 37. From this known network and our module information, we may predict other regulations in this module. In module 1, all the miRNAs are associated with ovarian cancer. The genes in this module take part in the process of transcription, gene expression etc.. The complex gene-miRNA regulatory relations may be related to ovarian cancer. In module 41, 5 miRNAs are associated with ovarian cancer. By checking the GO-BP terms, we found that the most enriched term is 'sexual reproduction', which has a *p*-value 7.50E-06, and Benjamini *p*-value 5.40E-03. This module also enriches the GO-term: 'gamete generation', 'male gamete generation', and 'spermatogenesis' significantly. All these show that this module should be very important in ovarian cancer development.

Comparison with Mirsynergy

There have been some other methods proposed for studying gene-miRNA modules. SNMNMF is the first paper to address this problem [41], and Mirsynergy works the best till now, to the best of our knowledge. As shown in [43], Mirsynergy works better than SNMNMF, and it runs much faster. Thus here we only compare our method with Mirsynergy. This method operates in two steps: it first detects the miRNA modules based on gene-miRNA relationship, then expands each miRNA module by greedily including (excluding) mRNAs into (from) the miRNA module to maximize the synergy score, which is a function of gene-miRNA and gene–gene interactions. Different

Table 3 Cancer associated miRNAs in the modules shown in Table 1

Module No.	1	2	5	10	11	12	18
Total No. of miRNAs	5	13	21	9	6	9	12
No. of cancer miRNAs	5	9	8	6	5	6	10
p-value	0	1.80E-04	4.82E-02	3.40E-03	1.77E-03	3.40E-03	4.74E-06
Module No.	21	22	35	37	38	41	45
Total No. of miRNAs	11	11	12	12	29	11	39
No. of cancer miRNAs	7	2	10	12	13	10	15
p-value	2.17E-03	7.00E-01	4.74E-06	0	2.29E-03	9.58E-07	6.43E-03

from our method, it did not consider the coexpressions of miRNAs. We use the same gene coexpression network, and the same gene-miRNA interaction data including both the known interactions and the gene-miRNA coexpressions. We directly applied the R package Mirsynergy to test the data. Table 4 shows the results.

With our proposed method, we identified 46 modules, while with Mirsynergy we identified 18 modules. We first did miRNA enrichment analysis for both the modules enriched by the miRNA clusters, and the miRNA clusters enriched by the modules. 14 of 46 modules identified with our method were enriched (Bonferroni corrected *p*-value<0.05) and 8 clusters were enriched by the modules by setting the overlap of the clusters and modules being 3 or larger. In contrast, there is one module enriched by miRNA clusters, and one cluster enriched by modules for the modules identified by Mirsynergy. We also did Gene Ontology biological process (GO-BP) terms and KEGG pathway enrichment analysis. 15 modules are enriched by GO-BPs and 7 modules are enriched by KEGG pathways with our method, while 4 modules are enriched by GO-BPs and one is enriched by KEGG pathway. This may be because the interactions of genes and miRNAs are very sparse in our data set, which results in similar synergy scores of many genes/miRNAs and thus one module may consist of many genes/miRNAs, while other modules have very small size. As shown in Table 4, although the average number of genes and the average number of miRNAs of the modules identified by Mirsynergy are larger than that of our method, there is one module having 1152 genes. Such cases have been addressed in [43]. By taking into account the coexpressions of miRNAs, the miRNAs that may compose modules are more densely connected, which can be identified with our method with

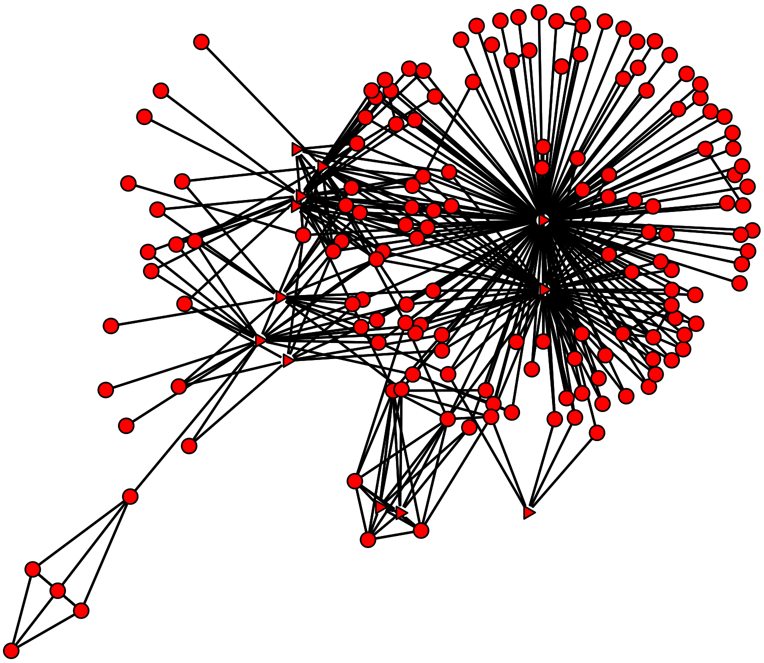

Fig. 2 Network structure of module 37. Triangle represents the miRNAs and circle represents the genes

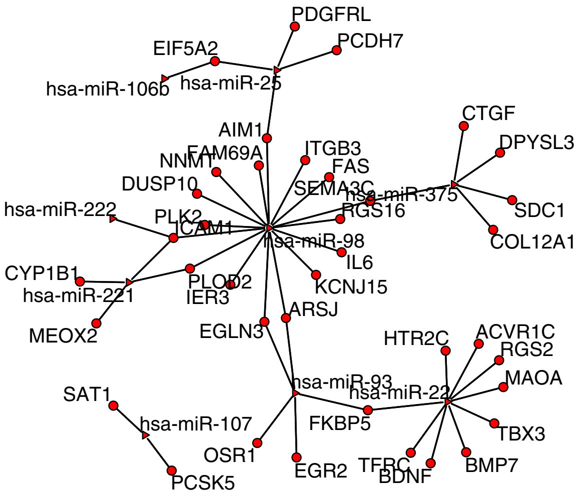

Fig. 3 Known interaction network for module 37. Triangle represents the miRNAs and circle represents the genes

a high accuracy. The identified modules by Mirsynergy are in Additional file 2, and the enrichment results are in Additional file 4.

Discussion

MiRNAs are actively involved in many biological processes by regulating the post-transcriptional gene expression. Increasing evidence shows that miRNAs play critical roles in many diseases including cancer and have a potential clinical value in diagnosis, treatment and prognosis. Although many works have been done to identify the targets of miRNAs and elucidate their complex regulatory networks, the complex relationships between miRNAs and genes are not fully understood. In this paper, we integrated the gene expression and miRNA expression data to study their complex interactions. By computing the pairwise Pearson correlation coefficients, we transformed the two data sets into networks. Then we proposed an optimization model to identify the modules in the integrated networks. We define the modules as subnetworks composed of genes, miRNAs, gene-gene interactions, miRNA-miRNA interactions, and gene-miRNA interactions. With such definitions, we found the interaction patterns of genes and miRNAs in the complex network. An approximate numerical algorithm is developed to solve the optimization problem. Compared to the existing methods, our method considers both the interactions within gene-gene, miRNA-miRNA networks, and the interactions between gene and miRNAs. By tuning the parameters for intra- and inter- networks, our method can give a good balance of all the interactions. The proposed method can be extended to study the modules in more networks with inter-connections. One weakness of our method is that the number of modules K should be given. To find the consistent results, we should try different KŠs, which may waste some computational time. Also, in other real applications, the identification accuracy may be related to the density of the intra-network connections. Thus we may need to add more tuning parameters to balance the intra-network connections. We applied our proposed method to an ovarian cancer data set. 14 modules are enriched by the miRNA clusters with overlap size being at least 3, 15 modules are enriched by GO-BP terms, and 7 modules are enriched by KEGG pathways significantly. In the identified modules, 122 miRNAs are cancer associated and 29 miRNAs are related to ovarian cancer, which has a p-value 0.039. These results show that the genes and miRNAs act together to contribute to the cancers. To find the omarkers of cancers, or develop therapy methods for cancers, we should take into account their interactions. From the module structures, we can also predict the unknown gene-miRNA interactions based on the

Table 4 Module enrichment performance of Mirsynergy and our method

Method	N_{module}	$\bar{N}_g$	$\bar{N}_m$	$N_{\text{en-module}}$	$N_{\text{en-cluster}}$	N_{GO}	N_{KEGG}
Mirsynergy	18	88.4	21.9	1	1	4	1
Our method	46	42.6	10.3	14	8	15	7

'N_{module}' denotes the total number of modules identified. '$\bar{N}_g$' and '$\bar{N}_m$' denote the mean number of genes and miRNAs in the modules. '$N_{\text{en-module}}$' denotes the number of enriched modules by clusters. '$N_{\text{en-cluster}}$' denotes the number of enriched clusters by modules. 'N_{GO}', 'N_{KEGG}' denote the number of modules enriched by GO-BP and KEGG pathway known gene-miRNA interactions. These predicted results may give some theoretical basis for further experimental validations. Although with our current method we can get more information on gene-miRNA interactions, their complex relationships are far from being fully known. To understand the biological system better, we need to add more elements into the model. Integrating with other data sets, such as DNA methylation, histone modification, is left as one of our research topics.

Conclusions

Our proposed method provides a way for studying the module structures in the complex gene-miRNA interaction network. The experimental results show that the modules composed of both genes, miRNAs, and their interactions are very likely to be related to cancers. These identified modules provide important information for further cancer studies, and are worth experimental validations.

Additional files

Additional file 1: The identified modules. We listed all the 46 modules identified with our proposed method. Each module includes both genes and microRNAs. (CSV 51 kb)

Additional file 2: The identified modules with Mirsynergy. We listed all the modules identified with Mirsynergy. Each module includes both genes and microRNAs. (CSV 87 kb)

Additional file 3: Enrichment results of the modules in Additional file 1. We presented the GO-BP enrichment results and KEGG pathway enrichment results of modules in Additional file 1. (XLSX 120 kb)

Additional file 4: Enrichment results of the modules in Additional file 2. We presented the GO-BP enrichment results and KEGG pathway enrichment results of modules in Additional file 2. (XLSX 68 kb)

Acknowledgments
S. Zhang's research is supported in part by NSFC grant No. 11471082, and Shanghai Natural Science Foundation 13ZR1403600, and M. Ng's research is supported in part by Hong Kong Research Grant Council GRF Grant No. 12302715.

Declarations
This article has been published as part of BMC Systems Biology Volume 10 Supplement 4, 2016: Proceedings of the 27th International Conference on Genome Informatics: systems biology. The full contents of the supplement are available online at http://bmcsystbiol.biomedcentral.com/articles/supplements/volume-10-supplement-4.

Funding
The publication of this work is supported by NSFC grant No. 11471082.

Availability of data and materials
The level 2 gene expression and miRNA expression data for ovarian cancer are downloaded from The Cancer Genome Atlas (TCGA). The gene expression data set is generated with UNC AgilentG4502A_07_03, and the miRNA expresssion data set is generated with UNC miRNA_8x15kv2. The gene-miRNA interaction data are downloaded from miRTarBase (*http://mirtarbase.mbc.nctu.edu.tw*). The miRNA cluster data are downloaded from the miRBase website (*http://www.mirba se.org/*).

Authors' contributions
SZ and MN designed the study. SZ did the experiments and drafted the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Center for Computational Systems Biology, School of Mathematical Sciences, Fudan University, No.220 Handan Road, 200433 Shanghai, China. [2]Department of Mathematics, Hongkong Baptist University, Kowloon Tong, Hongkong, Hongkong.

Published: 23 December 2016

References

1. Lee RC, Feinbaum RL, Ambros V. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. Cell. 1993;75:843–54.
2. Kozomara A, Griffiths-Jones S. mirbase: integrating microrna annotation and deep-sequencing data. Nucleic Acids Res. 2011;39:152–7.
3. Nilsen TW. Mechanisms of microrna-mediated gene regulation in animal cells. Trends Genet. 2007;23:243–9.
4. Lynam-Lennon N, Maher SG, Reynolds JV. The roles of microrna in cancer and apoptosis. Biol Rev Camb Philos Soc. 2009;84:55–71.
5. Dai X, Zhuang Z, Zhao PX. Computational analysis of mirna targets in plants: current status and challenges. Brief Bioinf. 2011;12:115–21.
6. Muniategui A, Pey J, Planes FJ, Rubio A. Joint analysis of mirna and mrna expression data. Brief Bioinform. 2013;14(3):263–78. doi:10.1093/bib/bbs028.
7. Flynt AS, Lai EC. Biological principles of microrna-mediated regulation: shared themes amid diversity. Nat Rev Genet. 2008;9:831–42.
8. Sayed D, Abdellatif M. Micrornas in development and disease. Physiol Rev. 2011;91:827–87.
9. Pencheva N, Tavazoie SF. Control of metastatic progression by microrna regulatory networks. Nat Cell Biol. 2013;15:546–54.
10. Zhang W, Zang J, Jing X, Sun Z, Yan W, Yang D, Shen B1, Guo F. Identification of candidate mirna biomarkers from mirna regulatory network with application to prostate cancer. J Transl Med. 2014;12. doi:10.1186/1479-5876-12-66.
11. Taft RJ, Pang KC, Mercer TR, Mattick JS. Non-coding rnas: regulators of disease. J Pathol. 2010;220:126–39.
12. Calin GA, Croce CM. Microrna signatures in human cancers. Nat Rev Cancer. 2006;6:857–66.
13. Huang Y, Shen XJ, Zou Q, et al. Biological functions of micrornas: a review. J Physiol Biochem. 2011;67:129–39.
14. Pfeifer A, Lehmann H. Pharmacological potential of rnai-focus on mirna. Pharmacol Therap. 2010;126:217–27.
15. Gentner B, Visigalli I, Hiramatsu H, et al. Identification of hematopoietic stem cell-specific mirnas enables gene therapy of globoid cell leukodystrophy. Sci Trans Med. 2010;2(58):58ra84.
16. Brown BD, Naldini L. Exploiting and antagonizing microrna regulation for therapeutic and experimental applications. Nat Rev Genet. 2009;10: 578–85.
17. Thomas M, Lieberman J, Lal A. Desperately seeking microrna targets. Nat Struct Mol Biol. 2010;17:1169–1174.
18. Saito T, Saetrom P. Micrornas-targeting and target prediction. New Biotechnol. 2010;27:243–9.
19. Maziere P, Enright AJ. Prediction of microrna targets. Drug Discov Today. 2007;12:452–8.
20. Chi SW, Zang JB, Mele A, et al. Argonaute hits-clip decodes microrna-mrna interaction maps. Nature. 2009;460:479–86.
21. Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. Cell. 2010;141:129–41.

22. Jin H, Tuo W, Lian H, et al. Strategies to identify microrna targets: new advances. New Biotechnol. 2010;27:734–8.
23. Huang G, Athanassiou C, Benos P. Mirconnx: condition-specific mrna-microrna network integrator. Nucleic Acids Res. 2011;39:416.
24. Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C. Magia, a web-based tool for mirna and genes integrated analysis. Nucleic Acids Res. 2010;38:352–9.
25. Bandyopadhyay S, Mitra R. Targetminer: microrna target predic- tion with systematic identification of tissue-specific negative examples. Bioinformatics. 2009;25:2625.
26. Gamazon ER, Im HK, Duan S, Lussier YA, Cox NJ, Dolan ME, Zhang W. Exprtarget: an integrative approach to predicting human microrna targets. PLoS ONE. 2010;5:13534.
27. Nam S, Kim B, Shin S, Lee S. Mirgator: an integrated system for functional annotation of micrornas. Nucleic Acids Res. 2008;36:159–64.
28. Hausser J, Berninger P, Rodak C, Jantscher Y, Wirth S, Zavolan M. Mirz: an integrated microrna expression atlas and target prediction resource. Nucleic Acids Res. 2009;37:266–72.
29. Ritchie W, Flamant S, Rasko J. Mimirna: a microrna expression profiler and classification resource designed to identify functional correlations between micrornas and their targets. Bioinformatics. 2010;26:223–7.
30. Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, Cutillo L, Ballabio A, Banfi S. Microrna target prediction by expression analysis of host genes. Genome Res. 2009;19:481–90.
31. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD. Using expression profiling data to identify human microrna targets. Nat Methods. 2007;4:1045–1049.
32. Lu Y, Zhou Y, Qu W, Deng M, Zhang C. A lasso regression model for the construction of microrna-target regulatory networks. Bioinformatics. 2011;27:2406–413.
33. Ritchie W, Rajasekhar M, Flamant S, et al. Conserved expression patterns predict microrna targets. PLoS Computat Biol. 2009;5(9):e1000513.
34. Jayaswal V, Lutherborrow M, Ma DDF, et al. Identification of micrornas with regulatory potential using a matched microrna-mrna time-course data. Nucleic Acids Res. 2009;37(8):e60.
35. Ragan C, Zuker M, Ragan MA. Quantitative prediction of mirna-mrna interaction based on equilibrium concentrations. PLoS Computat Biol. 2011;7(2):e1001090.
36. Liu B, Li J, Tsykin A, Liu L, Gaur AB, Goodall GJ. Exploring complex mirna-mrna interactions with bayesian networks by splitting averaging strategy. BMC Bioinforma. 2009;10:408.
37. Nam A, Li M, Choi K, Balch C, Kim S, Nephew K. Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression. Nucleic Acids Res. 2009;37:356–62.
38. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–15550.
39. Mootha VK, Lindgren CM, Eriksson KF, et al. Pgc-1alpha-responsive genes involved in oxidative phos- phorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267–73.
40. Pasquinelli AE. Micrornas and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev. 2012;13:271–82.
41. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. Bioinformatics. 2011;27:401–9.
42. Le HS, Bar-Joseph Z. Integrating sequence, expression and interaction data to determine condition-specific mirna regulation. Bioinformatics. 2013;29:89–97.
43. Li Y, Liang C, Wong KC, Luo J, Zhang C. Mirsynergy: detecting synergistic mirna regulatory modules by overlapping neighbourhood expansion. Bioinformatics. 2014;30:1–9.
44. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Gen Mol Biol. 2005;4:17.
45. Zhang S, Zhao H. Community identification in networks with unbalanced structure. Phys Rev E. 2012;85:066114.
46. Huang D, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2009;4:44–57.
47. Huang D, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
48. McLean GW, Carragher NO, Avizienyte E, Evans J, Brunton VG, Frame MC. The role of focal-adhesion kinase in cancer-a new therapeutic opportunity. Nat Rev Cancer. 2005;5(7):505–15.
49. Krupp M, Maass E, Marquardt JU, et al. The functional cancer map: A systems-level synopsis of genetic deregulation in cancer. BMC Med Genet. 2011;4(53):. doi:10.1186/1755-8794-4-53.
50. Gelb BD, Tartaglia M. Ras signaling pathway mutations and hypertrophic cardiomyopathy: getting into and out of the thick of it. J Clin Invest. 2011;121:844–7.