CeModule: an integrative framework for discovering regulatory patterns from genomic data in cancer

Qiu Xiao, Jiawei Luo*, Cheng Liang, Jie Cai, Guanghui Li and Buwen Cao

Abstract

Background: Non-coding RNAs (ncRNAs) are emerging as key regulators and play critical roles in a wide range of tumorigenesis. Recent studies have suggested that long non-coding RNAs (lncRNAs) could interact with microRNAs (miRNAs) and indirectly regulate miRNA targets through competing interactions. Therefore, uncovering the competing endogenous RNA (ceRNA) regulatory mechanism of lncRNAs, miRNAs and mRNAs in post-transcriptional level will aid in deciphering the underlying pathogenesis of human polygenic diseases and may unveil new diagnostic and therapeutic opportunities. However, the functional roles of vast majority of cancer specific ncRNAs and their combinational regulation patterns are still insufficiently understood.

Results: Here we develop an integrative framework called CeModule to discover lncRNA, miRNA and mRNA-associated regulatory modules. We fully utilize the matched expression profiles of lncRNAs, miRNAs and mRNAs and establish a model based on joint orthogonality non-negative matrix factorization for identifying modules. Meanwhile, we impose the experimentally verified miRNA-lncRNA interactions, the validated miRNA-mRNA interactions and the weighted gene-gene network into this framework to improve the module accuracy through the network-based penalties. The sparse regularizations are also used to help this model obtain modular sparse solutions. Finally, an iterative multiplicative updating algorithm is adopted to solve the optimization problem.

Conclusions: We applied CeModule to two cancer datasets including ovarian cancer (OV) and uterine corpus endometrial carcinoma (UCEC) obtained from TCGA. The modular analysis indicated that the identified modules involving lncRNAs, miRNAs and mRNAs are significantly associated and functionally enriched in cancer-related biological processes and pathways, which may provide new insights into the complex regulatory mechanism of human diseases at the system level.

Keywords: Regulatory pattern, Module discovery, microRNA, lncRNA function, ceRNA, Cancer, Machine learning

Background

MicroRNAs (miRNAs) are small (~ 22 nt), endogenous, single-stranded and non-coding RNA molecules, which play crucial roles in post-transcriptional regulation by repressing mRNA translation or destabilizing target mRNAs [1]. Many studies have revealed that the mutation and dysregulated miRNA expression may cause various human diseases [2, 3]. MiRNAs act as essential components of complex regulatory networks and are involved in many different biological processes, such as cell proliferation, metabolism, and oncogenesis [4–6]. Therefore, understanding the functional roles and regulatory mechanisms of miRNAs will greatly facilitate the diagnosis and treatment of human diseases [7, 8].

Recently, a competing endogenous RNA (ceRNA) hypothesis has been presented by Salmena et al. [9], which has dramatically shifted our understanding of miRNA regulatory mechanism. The complex ceRNA post-transcriptional regulatory mechanism reported that by sharing common miRNA response elements (MREs), several types of competing endogenous RNAs or miRNA sponges (e.g. lncRNAs, pseudogenes and circRNAs) compete with protein-coding RNAs for binding to miRNAs, thereby...
relieving miRNA-mediated target repression. Numerous convincing evidence has been discovered in a variety of species by biological experiments [10, 11]. For example, the study found that IncRNA HULS plays an important role in liver cancer, which serves as an endogenous sponge by reducing miR-372-mediated translational repression of PRKACB [12]. IPS1 overexpression has also been reported to increase the expression of PHO2 by competitively interacting with miR-399 in arabidopsis [13]. In addition, numerous studies have shown that ceRNA crosstalk exists in a variety of cellular behaviors, and many diseases are affected by their disturbances [14, 15]. However, the cooperative regulation mechanisms and the roles of ceRNA–associated activities in physiologic and pathologic conditions are in their infancy, and thus require further research.

The development of high-throughput techniques has made a vast amount of omics data to be publicly available, thereby enabling systematic investigation of the complex regulatory networks. Great efforts have been made to decipher the interaction mechanism of numerous biomolecules in a transcriptional or post-transcriptional level, such as co-regulatory motif discovery [16], miRNA–miRNA regulatory module identification [17, 18], miRNA and TF (transcription factor) co-regulation inference [19]. Meanwhile, other methods have been developed to prioritize cancer-related biological molecules, such as miRNAs [20, 21]. Undoubtedly, all these studies provide a global perspective for the study of combinatorial effects and human complex diseases.

In recent years, IncRNAs as a class of ncRNAs and miRNA sponges have been identified in many human cancers [22]. Some systematic studies on many diseases have been carried out [23–25]. In addition, some tools related to IncRNA, such as DIANA-LncBase [26], Linc2GO [27] and LncRNA2Disease [28], have been developed. However, the functions and modular organizations of most of IncRNAs are still not clear, and the novel regulatory mechanism based on ceRNA hypothesis requires comprehensive investigation. To the best of our knowledge, little effort has been devoted to methods that are specifically designed to investigate the cancer-specific regulatory patterns involved in miRNA and miRNA sponges on a large scale.

In this study, we develop a novel integrative framework called CeModule to systematically detect regulatory patterns involving IncRNAs, miRNAs, and mRNAs. The proposed method fully exploits the IncRNA/miRNA/mRNA expression profiles, the experimentally determined miRNA-IncRNA interactions, the verified miRNA-mRNA interactions, and the weighted gene-gene functional interactions. Here, inspired by [29–31], we adopt a model with joint orthogonality non-negative matrix factorization to detect these modules. In addition, both network-regularized constraints and sparsity penalties are incorporated into the model for helping to discover and characterize the IncRNA-miRNA-mRNA associated regulatory modules.

Finally, we apply the proposed method to ovarian cancer (OV) and uterine corpus endometrial carcinoma (UCEC) datasets downloaded from TCGA [32]. The results indicate that CeModule could be effectively applied to the discovery of biologically function modules, which greatly advances our understanding of the coordination mechanisms on a system level.

**Methods**

In the following sections, we will first introduce the mathematical formulation of CeModule. Afterwards, the modules are identified based on the decomposed matrix components. Finally, several experiments and literature surveys are performed to systematically evaluate these modules.

**The CeModule algorithm for identifying modules by integrating massive genomic data**

**Joint orthogonal non-negative matrix factorization**

In this study, we identify the IncRNA, miRNA and mRNA-associated regulatory modules by a non-negative matrix factorization (NMF)-based framework. The corresponding objective function of standard NMF [31, 33] is formulated as follows:

\[
\min_{W,H} \|X - WH^T\|_F^2 \quad \text{s.t.} \quad W \geq 0, H \geq 0
\]  

(1)

where \(\|.\|_F\) denotes the Frobenius norm.

Existing studies have indicated that orthogonality NMF could produce a better modularity interpretation [6, 30, 34]. Therefore, we present a integrative framework using joint orthogonality NMF to determine the module regulation and membership through simultaneously integrating multiple data sources. To clearly describe the problem, let \(X \in R^{S \times N1}\), \(X \in R^{S \times N2}\), and \(X \in R^{S \times N3}\) denote the IncRNA, miRNA, and mRNA expression matrices, respectively. Subsequently, we define an objective function of joint orthogonality NMF as follows:

\[
\min_{W,H_1,H_2,H_3} \sum_{i=1,2,3} \left( \|X_i - WH_i^T\|_F^2 + \frac{1}{2a}\|H_i^T H_i - I\|_F^2 \right) \quad \text{s.t.} \quad W \geq 0, H_i \geq 0
\]  

(2)

where \(W(\text{size}:S \times K)\) denotes the common basic matrix; coefficient matrices \(H_1, H_2\), and \(H_3\) have dimensions \(N1 \times K\), \(N2 \times K\), and \(N3 \times K\), respectively; \(a\) is the hyper-parameter that controls the trade-off of \(H_i\); dimension \(K\) represents the desired number of modules.

However, many data sources often contain noise, and several investigations of NMF have been conducted to improve the performance [35]. To obtain sparse solutions and regulatory modules with better biological interpretation, the sparse constraints were incorporated into this model.
similar to that suggested by Hoyer [36], which can effectively make matrices $H_i$ sparse. The objective function of joint orthogonality NMF with sparsity penalties can be written as follows:

$$
\min_{w,H_1,H_2,H_3} \sum_{i=1,2,3} \left( \left\| X_i - WH_i^T \right\|_F^2 + \frac{1}{2} \alpha \left\| H_i^T H_i - I \right\|_F^2 \right) \\
y_1 \left\| W \right\|_F^2 + y_2 \sum_{i=1,2,3} \left\| H_i \right\|_1 \\
\text{s.t. } W \geq 0, H_i \geq 0
$$

(3)

where $y_1$ and $y_2$ are the regularization coefficients.

**The mathematical formulation of CeModule**

Apart from the expression profiles, the data sources including miRNA-lncRNA interactions, miRNA-mRNA interactions and gene-gene network have also been fully utilized to improve the performance. Here, to improve the quality of identified modules, the network-based penalties are imposed on this computational model based on Hoyer’s work [6, 36] and make sure that those tightly linked lncRNAs/miRNAs/mRNAs are forced to assign into the same module.

Let $A \in \mathbb{R}^{N_2 \times N_1}$ and $B \in \mathbb{R}^{N_2 \times N_1}$ denote the adjacency matrices of miRNA-lncRNA and miRNA-mRNA interaction networks, respectively, $Ce \in \mathbb{R}^{N_3 \times N_3}$ is the matrix of gene-gene functional interaction network. For the miRNA-lncRNA interaction network, we perform the network-based constraints according to the objective function as follows:

$$
O_1 = \sum_{ij} a_{ij} h_1^{(2)} (h_1^{(1)})^T = \text{Tr}(H_2^T A H_1)
$$

(4)

where $a_{ij}$ is the entity of $A$; $h_1^{(2)}$ and $h_1^{(1)}$ represent the $i$th and $j$th rows of $H_2$ and $H_1$, respectively. Similarly, the corresponding objective functions of two other networks can be obtained as follows:

$$
O_2 = \sum_{ij} b_{ij} h_2^{(2)} (h_2^{(3)})^T = \text{Tr}(H_2^T B H_3)
$$

(5)

$$
O_3 = \sum_{ij} c_{ij} h_3^{(2)} (h_3^{(3)})^T = \text{Tr}(H_3^T C H_3)
$$

(6)

Then, combining the function in Eq. (3) with three network-based regularization terms, we can mathematically formulate the optimization problem of CeModule as follows:

$$
\min_{w,H_1,H_2,H_3} \sum_{i=1,2,3} \left( \left\| X_i - WH_i^T \right\|_F^2 + \frac{1}{2} \alpha \left\| H_i^T H_i - I \right\|_F^2 \right) \\
-\lambda_1 \text{Tr}(H_2^T A H_1) - \lambda_2 \text{Tr}(H_2^T B H_3) - \lambda_3 \text{Tr}(H_3^T C H_3) \\
y_1 \left\| W \right\|_F^2 + y_2 \sum_{i=1,2,3} \left\| H_i \right\|_1 \\
\text{s.t. } W \geq 0, H_i \geq 0
$$

(7)

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are the regularization parameters. In the following, we adopt an iterative updating method [37] to obtain local optimal solution for the optimization problem.

Let $\Phi = [\phi_{ik}], \Psi = [\psi_{ik}], \Omega = [\omega_{pk}], \Theta = [\theta_{pk}]$ be the Lagrange multipliers for constrain $w_{ik} \geq 0$, $h_{ik}^{(1)} \geq 0$, $h_{pk}^{(2)} \geq 0$, and $h_{pk}^{(3)} \geq 0$, respectively. We can obtain the Lagrange function of Eq. (7) as follows:

$$
L_f = \sum_{i=1}^{3} \left[ \text{Tr}(X_i X_i^T) - 2 \text{Tr}(X_i H_i W^T) + \text{Tr}(W H_i^T H_i W^T) + \frac{1}{2} \alpha (\text{Tr}(H_i^T H_i H_i - I)) \right] \\
- \lambda_1 \text{Tr}(H_2^T A H_1) - \lambda_2 \text{Tr}(H_2^T B H_3) - \lambda_3 \text{Tr}(H_3^T C H_3) \\
y_1 \text{Tr}(W W^T) + y_2 \sum_{i=1}^{3} \text{Tr}(E_i^T H_i) + \text{Tr}(\Phi W^T) \\
+ \text{Tr}(\Psi H_i^T) + \text{Tr}(\Omega H_i^T) + \text{Tr}(\Theta H_i^T)
$$

(8)

where $E_i \in \{1\}^{N_1 \times K}$, $E \in \{1\}^{N_2 \times K}$, and $E \in \{1\}^{N_3 \times K}$. The partial derivatives of the above function for $W$ and $H_i$ are:

$$
\frac{\partial L_f}{\partial W} = \sum_{i=1}^{3} \left[ -2X_i H_i + 2W H_i^T H_i \right] + 2y_1 W + \Phi \\
\frac{\partial L_f}{\partial H_1} = -2X_1^T W + 2H_1^T W^T + \frac{1}{2} \alpha (A H_1 H_1^T A - A H_1) \\
- \lambda_1 A^T H_2 + y_2 E_1 + \Psi \\
\frac{\partial L_f}{\partial H_2} = -2X_2^T W + 2H_2^T W^T + \frac{1}{2} \alpha (A H_2 H_2^T H_2 - A H_2) \\
- \lambda_1 A H_1 - \lambda_2 B H_3 + y_2 E_2 + \Omega \\
\frac{\partial L_f}{\partial H_3} = -2X_3^T W + 2H_3^T W^T + \frac{1}{2} \alpha (A H_3 H_3^T H_3 - A H_3) \\
- \lambda_2 B^T H_2 - \lambda_3 C H_3 + y_2 E_3 + \Theta
$$

(9)

Using the KKT conditions [38, 39] $\phi_{ik} w_{ik} = 0$, $\psi_{ik} h_{ik}^{(1)} = 0$, and $\theta_{ik} h_{pk}^{(3)} = 0$, we obtain the following equations for $w_{ik}$, $h_{ik}^{(1)}$, $h_{pk}^{(3)}$, and $h_{pk}^{(3)}$:

$$
-2 \sum_{i=1}^{3} (X_i H_i)_{ik} w_{ik} + 2 \sum_{i=1}^{3} \left[ W H_i^T H_i + (y_1 W) \right]_{ik} = 0 \\
(2X_1^T W - 2a_1 A H_1 - \lambda_1 A^T H_2)_{ik} h_{ik}^{(1)} = 0 \\
(2X_2^T W - 2a_2 A H_1 - \lambda_2 B H_3)_{ik} h_{pk}^{(2)} = 0 \\
(2X_3^T W - 2a_3 C H_3 - \lambda_3 C^T H_2)_{ik} h_{pk}^{(3)} = 0
$$

(10)

Finally, we determine the multiplicative update rules for $W$ and $H_i$ as follows:
The four non-negative matrices \( W, H_1, H_2 \) and \( H_3 \) are updated according to the above rules until convergence. More details about the derivations and proof for the convergence of the optimization problem are provided in the Additional file 1.

Determining ceRNA modules

The obtained coefficient matrices \( H_1, H_2 \), and \( H_3 \) will guide us to detect ceRNA-associated regulatory modules. Here, similar to the way for identifying co-modules developed by Chen et al. [40], we obtain a z-score for each element based on the columns of \( H_1 \), \( H_2 \), and \( H_3 \) as follows:

\[
z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},
\]

where \( \mu_j \) denotes the average value of lncRNA (or miRNA, mRNA) \( i \) in \( H_1 \) (or \( H_2, H_3 \)), and \( \sigma_j \) is the standard deviation. Subsequently, we assign lncRNA (or miRNA, mRNA) \( i \) into module \( j \) if \( z_{ij} \) exceeds a given threshold \( T \), and then all the ceRNA-associated modules can be obtained. The overall workflow of the proposed CeModule framework for identifying regulatory module is shown in Fig. 1.

Experimental setup and module validation

We systematically evaluate the performance of CeModule by conducting a functional enrichment analysis for genes in each module. We downloaded the GO (Gene

![Fig. 1 Overall workflow of CeModule for detecting lncRNA, miRNA, and mRNA-associated regulatory patterns](image-url)
Ontology) terms in biological process from http://www.geneontology.org/, and obtained the canonical pathways from MSigDB [41]. We removed the GO terms with evidence codes equal to NAS (Non-traceable Author Statement), ND (No biological Data available) or EA (Electronic Annotation) and those with fewer than 5 genes similar to Li et al. [18]. The hypergeometric test was used to calculate the statistical significance for genes in each module with respect to each GO term or pathway. Meanwhile, we used TAM [42], which is a free online tool for annotations of human miRNAs, to perform enrichment analysis for miRNAs in the identified modules.

We also investigate the miRNA cluster/family enrichment for each module, and obtained the miRNA cluster information and miRNA families from miRBase (http://www.mirbase.org/) (release 21) [43]. Furthermore, to determine whether these modules related to specific cancer, we acquired those known cancer-related lncRNAs from LncRNADisease [28] and Lnc2Cancer [44]. The verified disease-related miRNAs and genes were collected from HMDD v2.0 [45], and DisGeNET [46], respectively.

Additionally, the method contains several parameters, more detailed information about them are illustrated in Additional file 1. Here, we determined the values of reduced dimension $K$ on the basis of a miRNA cluster analysis. The results show that the miRNAs used in this study covered 69/76 miRNA clusters with an average of about 2.7/2.3 miRNAs per cluster for OV/UCEC dataset. Therefore, we set $K$ to 70 in the two cancer datasets, which is approximately equal to the number of miRNA clusters.

**Results**

**Data sources and preprocessing**

We applied CeModule to ovarian cancer (OV) and uterine corpus endometrial carcinoma (UCEC) genomic data and downloaded the matched mRNA and lncRNA expression profiles from http://www.larssonlab.org/tcga-lncrnas/ [47]. Due to the expression values of many lncRNAs/mRNAs in the original data source are all zeros or close to zeros, as done in [48], we removed some lncRNAs/mRNAs in the expression profiles with a variance less than the percentile specified by a cutoff (30%) and filter those lncRNAs/mRNAs with overall small absolute values less than another percentile cutoff (60%). The corresponding Matlab functions are `genevarfilter` and `genelowvalfilter`, respectively. We obtained the miRNA expression profiles of OV/UCEC from the TCGA data portal (http://cancergenome.nih.gov/) and removed the rows (or miRNAs) where all the expression values are zeros. These expression data were further log2-transformed. Finally, the datasets contain 7982(8056) lncRNAs, 415(505) miRNAs, and 10,618(10308) mRNAs across 385(183) matched samples for OV (UCEC), which were represented in three matrices $X_\mu$, $X_\alpha$ and $X_\beta$, and then the method in [49] is adopted to ensure non-negative constraints.

The experimentally verified interactions between miRNAs and lncRNAs were downloaded from DIANA-LncBase [26] and starBase v2.0 [50]. We obtained the miRNA targets from three experimentally verified databases, including miRecords (version 4.0) [51], TarBase (version 6.0) [52], and miRTarBase (version 6.1) [53]. After filtering out duplicate interactions or interactions involving lncRNAs, miRNAs, and mRNAs that were absent in the expression profiles, 12,969/6165 miRNA-lncRNA and 20,848/25447 miRNA-mRNA interactions were finally retained for OV/UCEC dataset. The weighted gene-gene network is derived from HumanNet [54], which is a probabilistic functional gene network. After filtering those genes absent from the expression data, 536,698/252021 interactions are retained for OV/UCEC. Finally, we obtained the miRNA-lncRNA matrix $A$, the miRNA-mRNA matrix $B$ and the gene-gene matrix $C$.

**Topological characteristics analysis**

We identified modules in ovarian cancer and uterine corpus endometrial carcinoma by integrating multiple heterogeneous data sources, and obtained 70 modules for OV/UCEC (Additional file 2: Table S1) with an average of 68.2/46.1 lncRNAs, 6.3/5.5 miRNAs, and 55.5/48.1 mRNAs per module. The distributions of number of lncRNAs, miRNAs, and mRNAs for the identified modules for OV and UCEC datasets are displayed in Additional file 1: Figure S1 and S2.

According to the constructed regulatory networks by merging those modules identified by our method, we found that a small number of nodes are more likely to be hubs or act as bridges, and tend to be involved in more competing interactions and participate in more human diseases. For instance, Fig. 2a presents a global view of the regulatory network for OV, which demonstrated that the network was densely connected and a small fraction of the nodes presented significantly higher degree, betweenness centrality, and closeness centrality than other nodes. The top 10 lncRNAs/miRNAs/mRNAs for each dimension (degree, closeness, and betweenness) in the networks of OV and UCEC datasets are listed in Table 1 and Additional file 1: Table S2, and there are substantial overlaps exist across the three dimensions (Fig. 2b and Additional file 1: Figure S3 and S4). Meanwhile, as shown in Fig. 2c and Additional file 1: Table S2, we found that all the top 10 lncRNAs (MALAT1, NEAT1, GAS5, H19, SNHG1, TUG1, FGDS-AS1, SNHG5, XIST, MEG3) and 8 out of the top 10 miRNAs (MAL2, XIST, SCAMP1, C17orf76-AS1, MALAT1, C11orf95, SEC22B, UBXN8) with the highest degree participate in at least 5 or more modules in OV and UCEC datasets, respectively. The
number distributions of modules for all the module members (lncRNAs/miRNAs/mRNAs) are provided in Additional file 2: Table S1.

On the other hand, most of the above lncRNAs are supported to be associated with different cancers by public databases or literature. For example, MALAT1 was found to be overexpressed in many solid tumors such as hepatocellular carcinoma [55] and lung cancer [56]. The downregulation of MEG3 is related to poor prognosis and promotes cell proliferation in gastric cancer [57] and bladder cancer [58]. Moreover, MALAT1, NEAT1, GAS5, H19 and XIST have been experimentally validated to be ovarian cancer-related lncRNAs [44], which were identified as hubs that connect 26, 15, 22, 20 and 9 modules in OV dataset, respectively. Additionally, MALAT1 also has been supported to be related to uterine corpus endometrial carcinoma and connected 7 modules in UCEC dataset. The above observations indicate that these lncRNAs can control communication among different functional components in the two datasets. Meanwhile, 8 (let-7b, mir-99b, mir-10b, mir-30a, mir-182, mir-183, mir-200c, mir-25) and 5 (mir-141, mir-10a, mir-200a, let-7b, mir-200b) of the 10 miRNAs with the highest degree are confirmed to be the well-known OV-related and UCEC-related miRNAs by HMDD [45]. We also found that these miRNAs are significantly enriched in cell cycle-related biological processes (Fig. 3a). In addition, we performed the same analysis for mRNAs and also came to the similar observations.

**Functional enrichments of modules**

To investigate the functional significance of the identified modules in ovarian cancer and uterine corpus
endometrial carcinoma datasets, we perform GO biological process and KEGG pathway enrichment analyses using hypergeometric test for coding genes in each of the modules (FDR < 0.05). The enriched GO terms and KEGG pathways of all the identified modules for OV and UCEC datasets are listed in Additional file 3: Table S3 and Additional file 4: Table S4. The results show that about 88.6%/91.4% of the modules in OV/UCEC are significantly enriched in at least one GO term, and 110/129 different enriched pathways are discovered for the identified modules. The most frequently enriched biological processes contain cell adhesion, immune response, signal transduction, cell cycle and inflammatory response. For instance, Table 2 lists the representative enriched GO terms for the selected modules in OV dataset, and we found that these modules are involved in

![Fig. 3 a](image-url) Functional enrichment analysis for the 10 miRNAs with the highest degree using TAM in OV.  
 ![Fig. 3 b](image-url) Pathway enrichment analysis of the module 15 in OV dataset.  
 ![Fig. 3 c](image-url) Pathway enrichment analysis of the module 17 in OV dataset. The area proportion of each pathway presents the number of genes enriched in this pathway.

---

**Table 1** The top 10 lncRNAs, miRNAs and mRNAs with the highest degree, closeness centrality, and betweenness centrality in OV

| Rank | Degree | miRNAs | mRNAs | Betweenness | miRNAs | mRNAs | Closeness | miRNAs | mRNAs | mRNAs |
|------|--------|--------|--------|-------------|--------|--------|-----------|--------|--------|--------|
| 1    | MALAT1 | let-7b  | RPS16  | MALAT1      | mir-10a | TCF7L1  | LINC00240 | mir-155 | NMES3  |
| 2    | NEAT1  | mir-10a | RPS11  | NEAT1       | let-7b  | SNRPF   | RP11-403I13.8 | mir-506 | HIF3A  |
| 3    | GASS   | mir-99b | RPS5   | H19         | mir-30a | PTP4A   | MALAT1    | mir-206 | TCF7L1  |
| 4    | H19    | mir-10b | RPS18  | GASS        | mir-146a| PRRX2   | NEAT1     | mir-223 | LRRC6  |
| 5    | SNHG1  | mir-30a | RPS8   | TUG1        | mir-375 | PNI3R   | FGDS-AS1  | mir-10a | ACTG1  |
| 6    | TUG1   | mir-143 | SRGN   | FGDS-AS1    | mir-149 | NRS3I   | H19       | mir-30a | PNI3R  |
| 7    | FGDS-AS1| mir-182 | TYROBP | SNHGS       | mir-99b | LRRC6   | TUG1      | let-7b  | PRRX2  |
| 8    | SNHG5  | mir-183 | RPL11  | XIST        | mir-183 | HIF3A   | XIST      | mir-197 | PTP4A3 |
| 9    | XIST   | mir-200c| ALOX5AP| SNHG1       | mir-143 | CTSD    | SNHG1     | mir-146a| CTSD   |
| 10   | MEG3   | mir-25  | RPL3   | SNHG3       | mir-320a| ACTG1   | SNHG14    | mir-25  | SNRPF  |
many biological processes or pathways that related to cancers [59, 60]. For example, module 2 is enriched in regulation of cell activation (GO:0050865) and immune system process (GO:0002376), and modules 7 and 15 are enriched in p53 signaling pathway (KEGG: hsa04115) and Focal adhesion (KEGG: hsa04510), respectively. As shown in Fig. 3b and c, we also found that some enriched pathways are shared by several modules, and

| Module | GO term | Description | q-value | Cancer IncRNAs | Cancer miRNAs | Cancer mRNAs |
|--------|---------|-------------|---------|----------------|---------------|--------------|
| 2      | GO:0002376 | immune system process | 1.04E-12 | MALAT1, MIR155HG, NEAT1, PVT1 | mir-10a | APOC1, APOE, BTG3, C1QA, C1QB, CBS, CCL2, etc |
|        | GO:0009605 | response to external stimulus | 2.31E-07 | | | |
|        | GO:0006954 | inflammatory response | 2.76E-04 | | | |
|        | GO:0050865 | regulation of cell activation | 2.25E-03 | | | |
|        | GO:0007154 | cell communication | 2.25E-03 | | | |
| 7      | GO:0032502 | developmental process | 1.32E-06 | DLEU2, DNM3OS, GASS, HOTAIRM1, MALAT1, SNHG1, SNHG3, SNHG5, TP53TG1 | mir-196b, mir-199b | CHST2, CLDN11, COX6B1, MGP, DACT3, DCHS1, DLK1, etc |
|        | GO:0030154 | cell differentiation | 1.62E-05 | | | |
|        | GO:0060284 | regulation of cell development | 1.06E-04 | | | |
|        | GO:0010942 | positive regulation of cell death | 2.89E-04 | | | |
|        | GO:0007275 | multicellular organismal development | 7.77E-07 | | | |
| 15     | GO:0007155 | cell adhesion | 2.57E-06 | GASS, H19, MEG3, SNHG5 | mir-202, mir-506, mir-508, mir-513c | FSTL1, LHX1, MEST, MFAP2, CDH3, NRS4A1, MMP2, etc |
|        | GO:0022610 | biological adhesion | 2.64E-06 | | | |
|        | GO:0009968 | negative regulation of signal transduction | 1.38E-03 | | | |
|        | GO:0042698 | ovulation cycle | 3.10E-04 | | | |
|        | GO:0050896 | response to stimulus | 2.54E-05 | | | |
| 17     | GO:0022411 | cellular component disassembly | 1.43E-20 | DNMT3OS, GASS, H19, LINCO0467, MEG3, RMRP, RP11-304 L19.5, RP11-385 J12, SNHG5 | mir-127, mir-134, mir-379, mir-370, mir-382, mir-409, mir-410, mir-431, mir-432, mir-433, mir-485, mir-493, mir-654, mir-758 | GPC3, SPARC, LHX1, LUM, MEST, MFAP2, IGF2BP2, etc |
|        | GO:0009968 | negative regulation of signal transduction | 7.65E-04 | | | |
|        | GO:0060284 | regulation of cell development | 8.80E-04 | | | |
|        | GO:0045595 | regulation of cell differentiation | 5.91E-04 | | | |
|        | GO:0006413 | translational initiation | 8.31E-21 | | | |

Note: The bold letters represent the IncRNAs/miRNAs/mRNAs related to ovarian cancer; q-value represents the corrected p-value using the Benjamini-Hochberg method.
some of them have been reported to be involved in OV [61]. Interestingly, these two modules contain three common mRNAs (EMILIN1, COL1A2, ENC1) and one of them (COL1A2) is related to cancer, suggesting that these modules (e.g. modules 15 and 17, modules 31 and 32 in OV) with many overlaps of mRNAs are more likely to have similar biological functions.

Accumulating evidence has demonstrated that miRNAs located in the same cluster or belonging to the same family are likely to function synergistically or are related to the same diseases [42]. In this study, we also conducted a miRNA cluster/family enrichment analysis for the identified modules based on TAM (http://www.cuilab.cn/tam) [42]. The results indicated that 35/27 of the identified modules are significantly enriched in at least one miRNA cluster or miRNA family for OV/UCEC (p-value< 0.05) (Additional file 5: Table S5). For instance (see Table 3), module 1 in OV contains 9 miRNAs, 4 of which (mir-362, mir-532, mir-500, mir-501) belong to the miR-188 cluster, and three miRNAs (mir-362, mir-532, mir-501) have been supported to be associated with cancer by HMDD. Moreover, two miRNAs (mir-200b, mir-200c) in this module, which belong to the miRNA family MIPF0000019, have been shown to be related to OV [45], while another two miRNAs (mir-500, mir-501) also belong to the miRNA family MIPF0000139. As another example, two of 8 miRNAs (let-7c, mir-99a) in module 20 are from the let-7c cluster and have been shown to be dysregulated in various cancers [17]. All the findings indicate the capability of CeModule in discovering cancer-specific modules.

Co-expression analysis of lncRNA-miRNA-mRNA regulatory modules

We also performed an analysis to evaluate the statistical significance of (anti)-correlations between lncRNAs, miRNAs and mRNAs within modules for both datasets. We expect that the molecules within those modules identified by CeModule are more (anti)-correlated than random sets of genes. Here, we define a correlation evaluation score to quantify the strength of competition in any given module \( C_v \) as follows:

\[
S(C_v) = \frac{\sum \mid corr_{\text{liRm}} \mid + \sum \mid corr_{\text{miRm}} \mid + \sum \mid corr_{\text{lmR}} \mid}{N}
\]  

(12)

which is defined as the average absolute values of PCCs (Pearson correlation coefficients) for all lncRNA-miRNA, miRNA-mRNA, and lncRNA-mRNA pairs, where \( N \) is the number of all the possible pairs for the three types of relationships in \( C_v \). \( corr \) is a function for calculating the pair-wise PCC based on the corresponding expression data.

To investigate the statistical significance, we adopt a permutation test by shuffling these lncRNAs, miRNAs and mRNAs according to those identified modules, and then compute the average competing evaluation score for them. As shown in Fig. 4a, the correlation evaluation scores of our method ranged from 0.072 to 0.352 for OV, and ranged from 0.100 to 0.489 for UCEC, they exhibit significantly higher correlation than the random modules (p-value = 1.20e-20 for OV, p-value = 3.03e-17 for UCEC, Wilcoxon rank sum test). We can also obtain the same conclusions on the two examples for modules 1 (p-value = 2.70e-06, Student’s t-test) and 2 (p-value = 1.04e-09) (Fig. 4b). Here, the correlation evaluation scores of these identified modules are generally weak, this is mainly due to the fact that the vast majority of Pearson correlation coefficients (PCCs) of lncRNA-miRNA, miRNA-mRNA and lncRNA-mRNA pairs were weak in the used datasets of OV and UCEC (Table 4).

Regulatory modules are strongly implicated in cancer

Based on the fact that the input data included the lncRNA, miRNA and mRNA expression profiles of OV and UCEC samples, we expect the modules indentified by our method to be related to cancers, especially OV/UCEC. Here, we obtained 82/265/4288 (116/322/4721) cancer-related lncRNAs/miRNAs/mRNAs that are involved in the expression profiles as the benchmark sets for OV (UCEC), and collected 11/5 lncRNAs, 83/75 miRNAs and 73/158 mRNAs related to OV/UCEC from several reliable databases as mentioned in the Section of Methods.

| Module | Overlap miRs & | p-value | Overlap miRs & | p-value |
|--------|----------------|---------|----------------|--------|
| 1      | mir-362,mir-532, mir-500, mir-501 | 1.22e-06 | mir-200b,mir-200c | 7.33e-04 |
|        | –              | –       | mir-500,mir-501  | 2.40e-03 |
| 18     | mir-99b,mir-152a | 9.15e-04 | mir-100,mir-99b | 9.15e-04 |
| 20     | let-7c, mir-99a | 1.03e-04 | mir-200a,mir-200b | 1.01e-03 |
|        | mir-200a, mir-200b | 3.07e-04 | –              | –       |
| 30     | –              | –       | let-7b,let-7c   | 4.96e-03 |
| 70     | mir-516a,mir-519a, mir-522,mir-518e | 1.45e-03 | mir-516a,mir-519a, mir-522,mir-518e | 7.66e-04 |

Note: a/b represent the miRNAs that overlap between modules and miRNA clusters as well as families.
As shown in Fig. 5a, 45.7% (92.9%), 71.4% (90.0%) and 22.9% (100%) of all the identified modules in OV dataset contained at least two OV-related (cancer-related) lncRNAs, miRNAs and mRNAs, respectively. Meanwhile, the corresponding ratios in UCEC dataset are 1.4% (62.9%), 64.3% (91.4%) and 10.0% (100%) for uterine corpus endometrial carcinoma-related (cancer-related) lncRNAs, miRNAs and mRNAs. The significant level of overlap between every module and cancer (OV/UCEC) lncRNAs/miRNAs/mRNAs is evaluated by hypergeometric test, and Table 5 lists the OV-related and cancer-related lncRNAs for several representative modules. For example, module 66 in OV dataset contains 58 lncRNAs, 9 of which are cancer lncRNAs and 6 of them are ovarian cancer lncRNAs. To take another example, module 51 in UCEC dataset contains 61 lncRNAs, 8 of which are cancer lncRNAs and 3 of them are uterine corpus endometrial carcinoma-related lncRNAs. We provided all the cancer (OV/UCEC) related modules for both datasets in Additional file 6: Table S6.

For OV (UCEC) dataset, the identified modules involve 1258/171/2172 (1252/172/2498) different lncRNAs/miRNAs/mRNAs. In the results of OV, as shown in Fig. 5b, 43 IncRNAs belong to the benchmark set of cancer IncRNAs (p-value = 1.18e-14, hypergeometric test), and 8 of them are relevant to ovarian cancer (p-value = 3.93e-05). In UCEC, 47 IncRNAs in those modules belong to the corresponding benchmark set (p-value = 6.05e-11) and 3 of which are UCEC specific IncRNAs (p-value = 2.93e-02). For miRNAs, 64.9%/77.3% of the 171/172 miRNAs are known to be involved in cancer in both datasets, and 51/43 miRNAs are specifically associated with OV/UCEC (p-value = 2.70e-05 for OV, p-value = 6.29e-06 for UCEC). Meanwhile, 1058/1186 mRNAs have been verified to be related to cancer, and 27/29 mRNAs are confirmed to be associated with ovarian cancer and uterine corpus endometrial carcinoma in OV and UCEC datasets, respectively. All the cancer-related and OV (UCEC) related molecules in those modules for both datasets are listed in Additional file 6: Table S6.

We also performed a differential expression analysis by two-sample t-test for those OV-related miRNAs (83 miRNAs) to investigate the cancer-specific abnormal changes in expression profile data. As a result, we identified 13 differentially expressed miRNAs (mir-200c, mir-99b, mir-183, mir-187, mir-10b, mir-625, mir-92b, mir-182, mir-449b, mir-107, mir-134, mir-98, mir-141, Additional file 7: Table S7) from those miRNAs, and found that 62.9% (44/70, Additional file 7: Table S7) of the modules contain at least one miRNAs that are differential expression. There are four modules (modules 13, 57, 60, and 69) are significantly enriched in ovarian cancer related differentially expressed miRNAs (hypergeometric test, FDR < 0.05, Additional file 7: Table S7). For example, module 57 contains 5 OV-related miRNAs (mir-182, mir-183, mir-200c, mir-625, mir-99b) and all of them are differential expression (FDR = 2.40e-05). The above observations imply that the lncRNAs/miRNAs/mRNAs in the identified modules are involved in various cancers, which confirm that the proposed method has a potential capability to discover modules related to cancers.

**Discussion**

Increasing evidence indicates that a novel competitive endogenous RNA (ceRNA) regulatory mechanism exists between non-coding RNAs and protein-coding RNAs.
LncRNAs and miRNAs are two kinds of crucial regulators and participate in many important biological processes. The aberrant expression of lncRNAs and miRNAs often contribute to tumorigenesis. To utilize the tremendous amounts of heterogeneous omics data and investigate the synergistic and cooperative mechanisms involve in lncRNAs, miRNAs, and mRNAs, our method integrates lncRNA/miRNA/mRNA expression profile data in an NMF framework, and simultaneously incorporates interaction networks in a regularized manner. The results of both (OV/UCEC) datasets indicate that the modules identified by CeModule contain many lncRNAs/miRNAs/mRNAs with specific topological patterns that are involved in some crucial biological processes and may cause cancers. Meanwhile, we further investigated whether the discovered modules were associated with the survival of ovarian cancer patients. The clinical data are downloaded from TCGA, and 383 samples are retained after removing those not included in the expression data or those with unavailable survival time. Kaplan-Meier survival analysis also indicates the ability of the method to discover modules that provide useful information for the prediction of cancer prognosis (Additional file 1).

Conclusions

In this study, we systematically investigate the efficiency of CeModule in identifying biologically functional modules that related to specific biological processes or cancers. We applied our method on the lncRNA/miRNA/mRNA expression data with matched samples of ovarian cancer and uterine corpus endometrial carcinoma from TCGA, and finally obtained 70 regulatory modules in both datasets. The observations indicate that these modules are densely connected and show specific topological characteristics. Meanwhile, these modules are significantly associated with many disease-related biological processes and pathways. Furthermore, a large number of lncRNAs/miRNAs/mRNAs in the modules are involved in various human complex diseases, such as ovarian cancer. All the results fully demonstrate the capability of CeModule for identifying of biologically functional modules. As a large number of sample-matched lncRNAs/miRNAs/mRNAs expression profile data become available, we believe that CeModule can serve as a potential tool for revealing condition-specific ceRNA regulatory patterns for cancer.

### Table 5

| Module | Cancer lncRNAs | OV lncRNAs | q-value |
|--------|----------------|------------|---------|
| 2      | MALAT1,MIR155HG,NEAT1,PVT1 | MALAT1,NEAT1,PVT1 | 1.05e-02 |
| 7      | DLEU2,DNM3OS,GASS,HOTAIRM1,MALAT1,SNHG1,SNHG3,SNHG5,TPS3TG1 | DNM3OS,GASS,MALAT1 | 3.08e-02 |
| 12     | MALAT1,RMRP,RP11-385 J1.2,XIST | MALAT1,XIST | 3.0e-03 |
| 31     | GASS,MALAT1,NEAT1,RP11-304 L19.5,SNHG3,SNHG5,TPS3TG1,UC1 | GASS,MALAT1,NEAT1,UC1 | 3.95e-05 |
| 41     | DLEU2,GASS,LINC00467,MALAT1,NEAT1,SNHG1,SNHG3 | GASS,MALAT1,NEAT1 | 3.91e-04 |
| 62     | DNM3OS,H19,HOTAIRM1,LINC00152,MALAT1,MEG3,NEAT1,PVT1,RMRP,RP11-304 L19.5,RP11-40199.4,SNHG5,UC1,XIST | DNM3OS,MALAT1,PVT1,NEAT1,UC1,XIST | 3.91e-04 |
| 66     | H19,MALAT1,MEG3,NEAT1,PVT1,SNHG1,SNHG3,UC1,XIST | H19,MALAT1,PVT1,NEAT1,UC1,XIST | 6.58e-09 |

Note: Num⁰ and Numⁱ are the ratios of lncRNAs that associated with cancer and OV in these modules. q-value is the FDR-corrected p-value after multiple testing correction.
Consent for publication
Not applicable.

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

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

Author details
1College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. 2Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China. 3College of Information Science and Engineering, Shandong Normal University, Jinan 250000, China.

Received: 6 October 2017 Accepted: 24 January 2019
Published online: 07 February 2019

References
1. Jopling CL, Yi MK, Lancaster AM, Lemon SM, Samow P. Modulation of hepatitis C virus RNA abundance by a liver-specific microRNA. Science. 2005; 309(5740):1577–81.
2. Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, Teschendorff AE. dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. 2017;45(D1):D812–8.
3. Jin D, Lee H. A computational approach to identifying gene-microRNA modules in cancer. PLoS Comput Biol. 2015;11(11):e1004942.
4. Karp X, Ambros V. Developmental biology. Encountering microRNAs in cell fate signaling. Science. 2005;310(5752):1288–9.
5. Cheng AM, Byrom MW, Shelton J, Ford LP. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucleic Acids Res. 2005;33(4):1290–7.
6. Xiao Q, Luo J, Liang C, Li G, Cai J, Ding P, Liu Y. Identifying IncRNA and miRNA co-expression modules from matched expression data in ovarian Cancer. IEEE/ACM Trans Comput Biol Inform. 2018;15:91.
7. Zeng XX, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Brief Bioinform. 2016;17(2):193–203.
8. Luo J, Xiao Q. A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. J Biomed Inform. 2017;66:194–203.
9. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta stome of a hidden RNA language? Cell. 2011;146(3):353–8.
10. Poliseno L, Salmena L, Zhang J, Panariello R, Grave de Peralta J, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465(7301):1033–40.
11. Ebert MS, Neilson JR, Sharp PA. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods. 2007;4(9):721–6.
12. Wang JY, Liu XF, Wu HC, Li PH, Gu ZD, Qiao YX, Chen N, Sun FY, Fan QS. CREB up-regulates long non-coding RNA HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. 2010;38(16):5366–83.
13. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, Garcia JA, Paz-Ares J. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet. 2007;39(8):1033–7.
14. Chiu CY, Wang LJ, Lu TP, Hsiao TH, Chuang EY, Chen Y. Differential correlation analysis of glioblastoma reveals immune ceRNA interactions predictive of patient survival. BMC Bioinformatics. 2017;18(1):132.
15. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014;505(7483):344–52.
16. Liang C, Li Y, Luo JW, Zhang ZL. A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microRNA co-regulatory networks in human. Bioinformatics. 2015;31(14):2348–55.
17. Liang C, Li Y, Luo J. A novel method to detect functional microRNA regulatory modules by Biclques merging. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(3):549–56.

Abbreviations
cerNA: Competing endogenous RNA; EA: Electronic Annotation; GO: Gene Ontology; IncRNAs: Long non-coding RNAs; miREs: microRNA response elements; NAS: Nontraceable Author Statement; ncRNAs: Non-coding RNAs; ND: No biological Data available; NMF: Non-negative Matrix Factorization; OV: Ovarian Cancer; TCGA: The Cancer Genome Atlas

Acknowledgements
Not applicable.

Funding
This work was supported by the National Natural Science Foundation of China (Grant nos. 61873089, 61527120, 61620223, 61862025), Shandong Provincial Natural Science Foundation, China (Grant no. ZR2016FB10), Hunan Provincial Science and Technology Project Foundation, China (Grant no. 2018TP1018), Jiangxi Provincial Natural Science Foundation, China (Grant no. 20181BAB211016), Hunan Provincial Natural Science Foundation, China (Grant no. 2018102024), and Key Project of the Education Department of Hunan Province, China (Grant no. 17A037). The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets supporting the conclusions of this article are included within the article and its additional files. The code used in the current study is available at https://github.com/xiaoqjx2018/CoModule.

Authors’ contributions
QX and JL initiated the study. QX conducted the experiments and wrote the manuscript. QX, LJ, CL, GL, BC and JC conceived and designed the study, supervised the analysis, and interpreted the results. All authors approved the final version.

Ethics approval and consent to participate
Not applicable.
