Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 16 (6): 7112–7125.
DOI: 10.3934/mbe.2019357
Received: 28 February 2019
Accepted: 29 July 2019
Published: 05 August 2019

Research article

Identification of lncRNAs-gene interactions in transcription regulation based on co-expression analysis of RNA-seq data

Sijie Lu[1], Juan Xie[2], Yang Li[1,2], Bin Yu[3], Qin Ma[2,]* and Bingqiang Liu[1,]*

[1] School of Mathematics, Shandong University, Jinan, Shandong 250100, China
[2] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA
[3] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

* **Correspondence:** Email: qin.ma@osumc.edu, bingqiang@sdu.edu.cn; Tel: 1-706-254-4293, 86-531-88363455.

Abstract: Long noncoding RNAs (lncRNA) play important roles in gene expression regulation in diverse biological contexts. Numerous studies have indicated that lncRNA-gene interactions are closely related to the occurrence and development of cancers. Thus, it is important to develop an effective method for the identification of target genes of lncRNA. Meanwhile, the high throughput sequencing data provide tremendous information about regulation correlation, by which the new target genes could be detected from known lncRNA regulated genes. In this study, we developed a method for elucidating lncRNA-gene interactions by using a biclustering approach, which allows for the identification of particular expression patterns across multiple datasets, indicating networks of lncRNA and gene interactions. A p-value strategy is followed to link co-expression patterns to certain lncRNAs. The method was applied on the breast cancer RNA-seq datasets along with a set of known lncRNA regulated genes. The evaluation indicated that the method can detect some new targets but fail to obtain higher coverage. We believe that this developed method will provide useful information for future studies on lncRNAs.

Keywords: lncRNA; Biclustering; RNA-seq; gene regulation; DNA motif

1. Introduction

With the advent of technologies allowing for large-scale, high throughput data, a much clearer understanding of the genomic mechanisms behind gene regulation have been gained. The scientists found that there are unexpected far more noncoding RNAs comparing with protein-coding genes and, and these noncoding regions play important roles in determining the complexity observed in the human genome [1,2]. Within these noncoding regions, long noncoding RNAs (lncRNAs), which are functionally defined as noncoding regions of RNA that are at least 200 base-pairs in length, have attracted lots of attention. Certain lncRNAs appear to act locally, while others have more distal regulatory effects, even acting across multiple chromosomes [3]. Many studies have identified specific functions of particular lncRNAs, including embryonic mechanisms, cell cycle functions, innate immunity, and disease processes. However, there are still thousands of lncRNAs have no identified functions [1,3–6]. Some studies have been performed that produce relatively few numbers of lncRNA functions [7], and have shown that the function of lncRNAs is highly cell-type-specific: one lncRNA may inhibit particular genes in one type of cell while promoting the same gene in another. This phenomenon makes it even more difficult to identify lncRNA functions on a large scale. Due to this specificity, researchers propose that future lncRNA studies should be performed on specific cell types to identify particular regulatory mechanisms.

One of the most prominent and intriguing applications of lncRNA regulatory investigation comes from cancer studies [8,9]. It has been shown that lncRNAs appear to have high connectivity with numerous diseases, especially cancer. Because of the highly cell type-specific nature of lncRNA regulatory functions and the irregularity of cancer cell genetic information, studying lncRNA regulation in specific cancer types may provide promising insight into specific genomic regulations of common cancer cells. In a few documented cases, specific lncRNAs have been shown to be significantly differentially expressed in specific cancer types, such as prostate cancer and breast cancer [1]. For these reasons, it seems appropriate to further investigate lncRNA-gene interactions in particular cancer cells.

The wealth of gene expression datasets available provides an opportunity to computationally identify co-expressed gene modules(CEMs), each of which is defined as a highly structured expression pattern on a specific gene set [10,11]. These CEMs tend to be functionally related or co-regulated by the same transcriptional regulatory signals (e.g., transcription factors, lncRNA and so on) under a specific condition or in a particular disease cell type. Overall, successful derivation of the CEMs may grant a higher-level interpretation of large-scale gene expression data, improve functional annotation of condition-specific gene activities, facilitate inference of gene regulatory relationships, hence, provide a better mechanism level understanding of complex diseases.

The computational identification of CEMs can be solved by a biclustering approach [12], which is a two-dimensional data mining technique that simultaneously identifies co-expressed genes under a subset of conditions. a high proportion of enriched biclusters on real datasets. Within this study, we try to identify new lncRNA-gene interactions and transcription factor-lncRNA partnerships from cancer RNA-seq data using a biclustering approach. The biclustering method will allow for the identification of particular expression patterns across multiple datasets, indicating networks of lncRNA and gene interactions. This developed method will also provide a framework for future lncRNA interaction studies. We applied this method on two sets of TCGA breast cancer RNA-seq data to generated CEMs based on known lncRNA-gene interactions. Then, the predicted CEMs are linked to lncRNA by a statistic p-value and the new lncRNA-gene relationship are generated. The evaluation on the predicted results showed that the pipeline can find some target genes for given lncRNA, and meanwhile the performance still has some space to be improved. We further conducted a TF motif analysis on the predicted CEMs and provide potential regulation cooperation between TFs and lncRNAs. The related original data with codes, results and supplementary data can be downloaded on https://github.com/IvesG/sGavin.git.

2. Materials and method

2.1. Data collection

Two sets of TCGA (The Cancer Genome Atlas) breast cancer RNA-seq data, one from the normal cell (referred as normal data) and the other from tumor cell (referred as tumor data) were downloaded from https://portal.gdc.cancer.gov/. The normal and tumor data consist of 113 and 1091 samples, respectively. And of the 113 normal samples, 112 of them are from the same patient among the tumors. Both datasets contain 60,483 genes, among which there are 19,824 protein-coding genes and 7,399 long intergenic noncoding RNAs (lincRNAs) genes. The RNA-seq data are all Upper Quartile normalized FPKM (UQ-FPKM) values.

A total of 1,081 experimentally validated lncRNA-associated regulatory entries were downloaded from LncReg [13], describing the comprehensive regulatory relationships among 258 lncRNAs and 571 genes. All these relationships were manually collected from PubMed with focus on the data generated by laboratory methods, and can be categorized into up/down/active/inactive based on regulatory relationships or transcription/post-transcription/translation/post-translation based on regulatory mechanisms.

2.2. Extract expression of target genes from RNA-seq dataset

As we focus on lncRNA-gene interactions, the relationships downloaded from LncReg were filtered to retain only relationships describing genes regulated by lncRNAs with specified species information (constrained to Homo sapiens and Mus musculus), resulting 925 relationships in total for the downstream analysis, covering interactions between 309 unique human genes and 103 human lncRNAs, as well as between 199 mouse genes and 100 mouse lncRNAs. It is noteworthy that these 925 relationships include 28 post-transcriptional regulations, 41 post-translational regulations, 714 transcriptional regulations, 23 translational regulations, 1 transcriptional &translational regulation and 118 unspecified relationships.

As the table from LncReg [13] only provides gene symbols, while the RNA-seq dataset uses Ensembl ID as gene's identifiers, we use Ensembl BioMart [14] to match gene symbols with Ensembl IDs for all the genes and lncRNAs. Then we got orthologous genes between mouse and human also using BioMart; we found orthologous human genes for all 199 mouse genes, and 38 overlapped with original human genes. For convenience, we recorded human genes, mouse genes that don't overlap with human genes, human lncRNAs and mouse lncRNAs that don't overlap with human lncRNAs as HG, MG, HL, and ML, respectively.

We combined the normal and tumor RNA-seq dataset together, then extracted expression values for all the HG, MG, HL, ML, protein-coding genes (PC, the remaining protein-coding genes except

HG and MG) and lincRNAs (linc, the remaining lincRNA except HL and ML). Taking the genes as rows and the conditions as columns, we obtained the RNA-seq expression matrix on which biclustering will be performed to detect CEMs.

2.3. Bi-clustering analysis

QUBIC is a biclustering analysis tool designed for co-expression analyses of genes based on their gene-expression patterns under multiple conditions. The software can generally identify all statistically significant groups, or biclusters, of genes with similar expression patterns under at least a specific number of experimental conditions, which tend to be more sensitive and more specific than other biclustering tools [15]. We use a quantile-based discretization method of QUBIC to generate a qualitative representing matrix for the RNA-seq expression matrix. Then we extracted the rows of known lncRNA regulated HG and MG from this representing matrix as seed 1 and HG, MG, HL, and ML rows as seed 2. Next bi-clustering analysis was performed on these two seeds to predict co-expressed gene modules (CEMs) in the qualitative representing matrix, respectively.

2.4. Predict the potential lncRNA-gene interactions

For an identified CEMs, we calculated the P-value of a bicluster enriched with genes regulated by a lncRNA using the hypergeometric function [16],

$$\Pr\,(r|N,\mathrm{K},n)=\frac{\binom{K}{r}\binom{N-\mathrm{K}}{n-r}}{\binom{N}{n}}$$

where *r* is the number of genes in a CEMs (with size *n*) that regulated by certain lncRNA, *N* is the total number of known lncRNA regulated genes in the whole genome, K is the number of genes regulated by that lncRNA in the whole genome.

We assumed that, if the known target genes of a given lncRNA are highly covered by a CEM with a significant p-value, the other genes in this CEM have high possibilities regulated by the given lncRNA. Thus, we used the smallest P-value for all possible lncRNAs as the p-value of the current bicluster and the relationships between lncRNA and genes in the bicluster are predicted.

2.5. Validation of the prediction

To evaluate the performance of the new methods on the prediction of new relationships between lncRNA and genes, we randomly separate seed2 into two parts with equal size named seedpart1 and seedpart2, for multiple times. Then bi-clustering analysis will be performed on seedpart2 to predict co-expressed gene modules (CEMs). For seedpart1 we find its part which is covered by co-expressed gene modules (CEMs) from seedpart2. We calculate the cover ratios by the size of seedpart1 to be divided by the size of the covered part by CEMs generated from seedpart2. Also, we calculate the p-values for the coverage rates to present the statistical significance of them.

2.6. Motif analysis on predicted CEMs

We choose several significant CEMs with sizes or conditions below 100, to conducted TF motif analysis. The promoter regions of the corresponding genes are inputted into the sub-routine findMotifs.pl of Homer [17], respectively. The script findMotifs.pl can firstly search for the upstream promoter sequences of a certain length automatically, and then perform motif finding on the promoters. For each run of findMotifs.pl on the datasets, we let the program output at most 5 top-ranking motifs, i.e. there will be up to 5 motifs discovered by findMotifs.pl for each CEMs. To evaluate the validity of the discovered motifs, findMotifs.pl automatically compares the similarity between the discovered motif profiles and the motif profiles archived in JASPAR [18] v2018 (http://jaspar.genereg.net/) under its default parameter setting. For each discovered motif having similarity with at least one motif archived in JASPAR, we present its motif logo as well as the information of its most similar motif in JASPAR.

3. Results

3.1. The known interactions of genes and lncRNAs

All the known interactions between lncRNAs and genes are showcased in Figure 1A. The related data can be download from https://github.com/IvesG/sGavin.git data/LncReg0419 and more details are written in data/readme.txt. In figure 1A, dark-blue nodes represent LncRNAs, light-blue nodes represent proteins, pink edges represent interactions documented in Homo, green edges represent interactions documented in Mus, orange edges represent interactions documented in both Homo and Mus. Meanwhile, there are some labels on the edges, categorized based on regulatory mechanisms including PTL (post-translational regulation), TC (transcriptional regulation), PTC (post-transcriptional regulation), TL (translational regulation), and NS (not sure). The distribution above is displayed in Figure 1B and nearly three fourth (714/925) of them are identified at the transcriptional level. Other labels on the edges are categorized based on regulatory relationships including down, up, active and inactive. The distribution above is displayed in Figure 1C. The down relationships (575) are more than up relationships (308), and the proportion of active/inactive is scarce (4.5%).

Figure 1E showed the distribution of a number of genes regulated by each lncRNA. It can be found that most lncRNA (~78%) regulate less than 5 genes. To show the specific details of the number of genes regulated by each lncRNA, Figure 1D is made, each point in the Figure 1D reflect the number of lncRNA (horizontal coordinate) that regulate certain number of genes (longitudinal coordinates) e.g. the point with coordinate (4,14) in Figure 1D indicate that there are 14 lncRNA and each of them regulate 4 genes. The lncRNA that regulate more genes in Figure 1D belongs to the more concentrated parts in Figure 1(A).

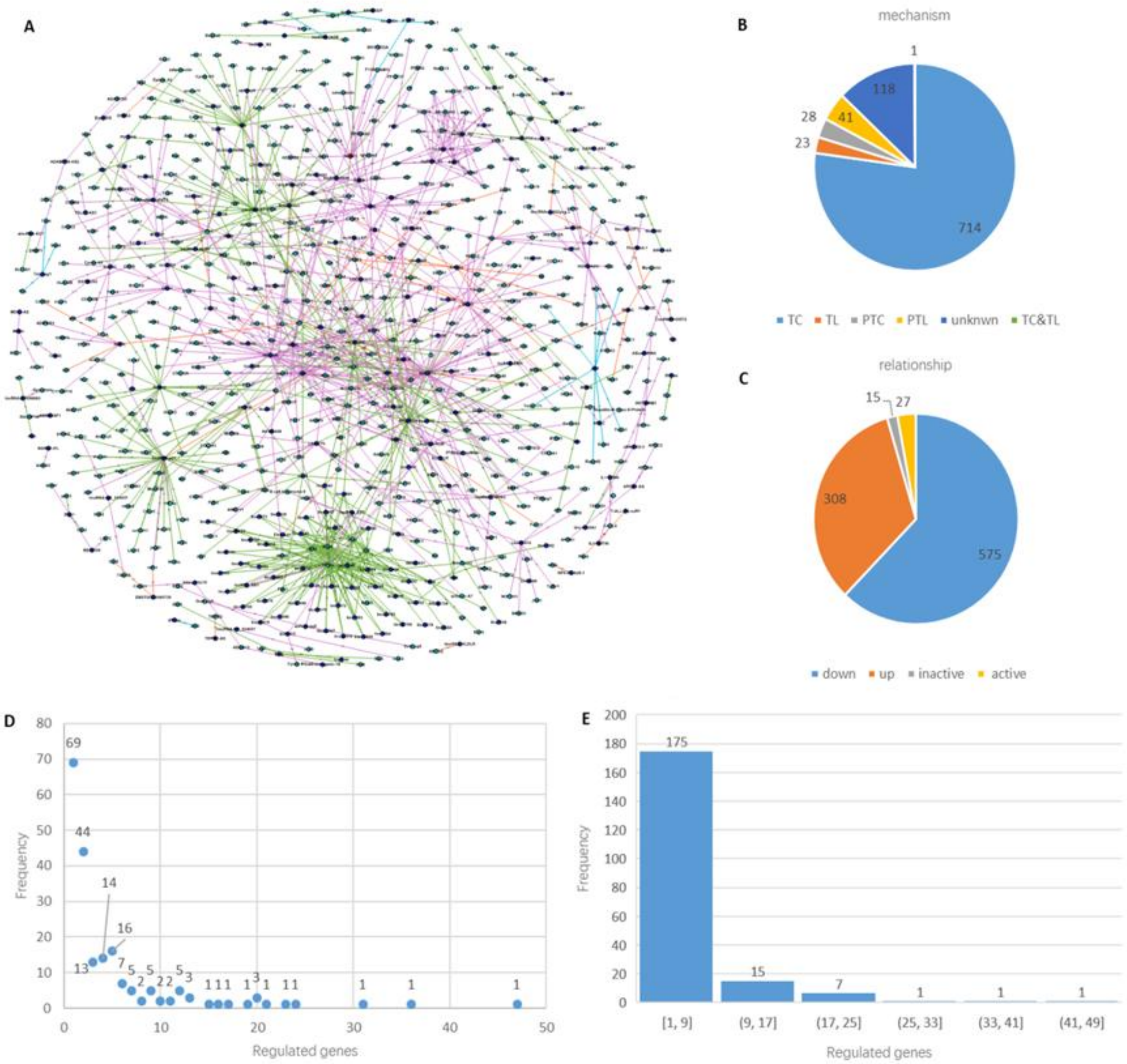

Figure 1. Correlation Analysis of the known interactions of genes and lncRNA. (A) The network of known interactions between genes and lncRNA. The distribution of the regulatory mechanisms of the network. (C) The distribution of the regulatory relationships of the network, (D&E) The number of lncRNA that regulate genes in the scatter diagram and bar chart respectively.

3.2. The predicted co-expressed gene modules (CEMs)

With the quantile-based discretization method and biclustering analysis, there are some co-expressed gene modules (CEMs) are found. The details of the way we identify CEMs are showcased in Figure 2C. Figure 2A shows the number of co-expressed gene modules(CEMs) we have got from seed1 and seed2 processed by max(min [19])-based (QUBIC1.0, [15,20]) and KL-based bi-clustering analysis (QUBIC2.0 [21]) respectively. And the distributions of numbers of genes and conditions for each CEM can be found in Figure 2D and Figure 2E. For instance, label seed1_1genes represent that Qubic1.0 is performed on seed1. To better illustrate the distributions, we have constrained the number of each gene and the number of each condition below 100 (in Figure 2B). It is found that the KL-based biclustering method tends to generate CEMs that contain fewer genes while more conditions than max(min [22])-based biclustering method from Figure 2D and Figure 2E. The difference between the results (whether the distribution of genes sizes and condition sizes or the number of CEMs that we have predicted) we get from seed1 and seed2 is subtle.

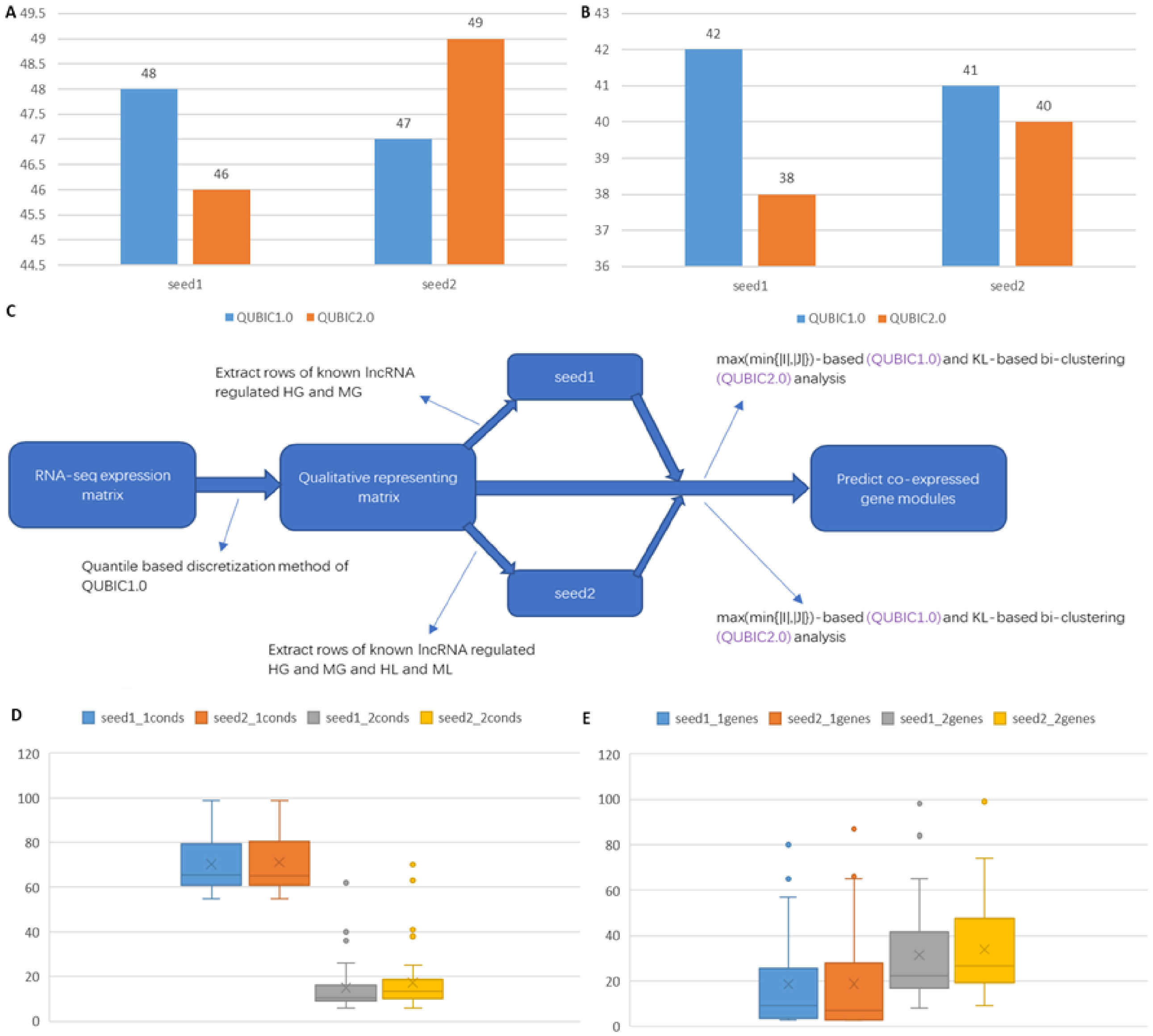

Figure 2. Correlation Analysis of the predicted co-expressed gene modules. (A) The number of CEMs that we have obtained. (B) The number of CEMs with size (both numbers of conditions and genes) constrained below 100. (C) The flow-process diagram describing the way we find CEMs. (D) The distribution of numbers of conditions of CEMs. (E) The distribution of the number of genes of CEMs.

3.3. Predicting potential interaction between lncRNAs and genes

The proportion of CEMs that have significant P-values (below a pre-selected P-value cutoff) as well as proportions of the number of unique enriched lncRNA in each bicluster that belong to certain categories (i.e., number of lncRNA = 2,3 or > 3) are calculated and shown in Figure 3A and Figure 3B. In the figures, Seed1_qubic1 represent the proportions from the results obtained using quantile discretization and using max(min [23])-based biclustering on seed 1, seed1_qubic2 represent using quantile discretization and using KL-based biclustering on seed1. In Figure 3A, it can be found that most CEMs have P-value more than 0.001 and seed1_qubic1 seems to have more significant P-value. Constrain P-value below 0.00001 and there is barely CEMs remained (less than 7%). In Figure 3B the majority of (more than70%) CEMs are with enriched lncRNA more than 3 and especially most (around 85%) of CEMs from seed1_qubic1.

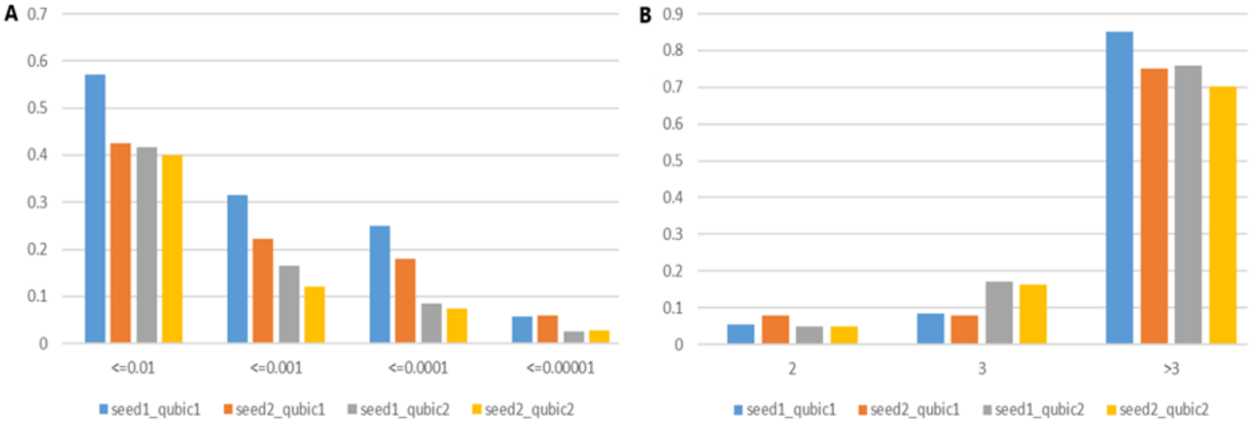

Figure 3. Correlation Analysis of the enriched lncRNA and P-value in CEMs. (A)Proportions of CEMs that significantly enriched with lncRNAs and proportions of the number of enriched lncRNA for seed1. (B)Proportions of CEMs that significantly enriched with lncRNAs and proportions of the number of enriched lncRNA for seed2.

3.4. Performance evaluation of the pipeline

For validation, we separated the HG + MG genes into two parts randomly and equally for 10 times and obtained 10 cover ratios correspondingly to check the accuracy of the previously predicted genes. The results of our validation are calculated and shown in Table 1. From Table 1 it can be found that all of the cover ratios are under 25% and the average ratio is 16.25%. We further calculated the p-value of the coverage rates. The results indicated that even the coverage date has a lot of space to be improved, the statistical significance of them are acceptable.

Table 1. Groups refer to genes that we extracted.

Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ratio	14.00%	19.60%	14.00%	15.70%	14.50%	21.30%	9.80%	24.30%	14.00%	15.30%
P-value	7.58E-07	5.14E-14	7.58E-07	8.11E-09	2.56E-07	1.24E-16	5.37E-03	1.27E-21	7.58E-07	2.65E-08

3.5. TF motif analysis on CEMs

Since lncRNA plays an important role in regulation, they should have cooperation with transcription factor [23,24]. Thus we conduct the analysis about the DNA binding sites of related to CEMs [19,22,25]. As described in the Method section, we choose five CEMs to conducted TF motif analysis. The corresponding gene list files each containing 3, 6, 14, 20 and 21 genes. The predicted motifs and the comparison between them and JASPAR motifs are listed in Table 2, along with the function of the target TFs. In Table 2, the second column has the name of lncRNA related with this CEMs and the p-value of their correlations; The third column contains the motif consensus by Homer; The fourth column provides TF names of the most similar motifs in JASPAR, along with the similarity scores in the fifth column. These TFs may have cooperation with corresponding lncRNAs. In the first column, all the P-value of the LncRNA from the CEM is below 0.01 and the least P-value is from LncRNA HOTAIR. The supplementary table S1 with more details, including and the logo of discovered motifs and the functions of corresponding TFs, can be downloaded by visiting the GitHub link.

Table 2. Comparison between discovered motifs and JASPAR motifs.

	lncRNA (p-value)	**Homer motifs**	**JASPAR TFs**	**Scores**
1	FOXCUT 3.6e-3	AACCAVTTHDCG	TFCP2	0.64
		TCCTATCACACR	MEIS2	0.62
		TTTTHAAAGGGG	CHR	0.67
		ARTGGTTGTWGA	FOXJ2	0.58
		GCAATCTCGC	IRF4	0.66
2	ANCR 1.1e-3	AGGGTGACAG	SPZ1	0.80
		GGTATCTTAC	GATA5	0.64
		CTCATAGGAG	GCM1	0.65
		TAAGTGAAAG	PRDM1	0.86
		CTTTTGGAAC	CHR	0.65
3	250-280 2.2e-4	WYTRTCTTTGCG	RXR	0.61
		TCTTACGG	ELK1	0.71
		GGCAAGGA	SD	0.76
		GAGGTATGTT	TEAD1	0.70
		TGCCGGGAGCGT	POL	0.64
4	HOXD-AS1 6.1e-3	CTCGAGTAGG	PB0114	0.63
		GCCCCCTGCA	PB0076	0.74
		ACGYMYATKYCC	GFY	0.59
		AGCGGGTT	PH	0.68
		AGGCGCCGCGCC	SP1	0.69
5	HOTAIR 5e-6	TGGCGCAGCGCG	PB	0.67
		GTACAACTTT	PB	0.66
		CMTSTGTCWCYK	NeuroG2	0.66
		GTGATCCATT	RHOXF1	0.68
		GGTMGRRGTGMW	TBX20	0.58

Table 3. Gene ontology information of selected CEMs.

LncRNA	**ID**	**Description**	**q-value**
FOXCUT	GO:0033613	transmembrane receptor protein tyrosine kinase activity	1.2937E-02
	GO:0033613	activating transcription factor binding	1.2937E-02
	GO:0019199	transmembrane receptor protein kinase activity	1.2937E-02
	GO:0001085	RNA polymerase II transcription factor binding	2.0767E-02
250-280	GO:0003735	structural constituent of ribosome	3.1600E-06
	GO:0003729	mRNA binding	1.1140E-03
	GO:0008483	transaminase activity	1.1140E-03
	GO:0048027	mRNA 5'-UTR binding	1.1140E-03
	GO:0016769	transferase activity, transferring nitrogenous groups	1.1140E-03
	GO:0045182	translation regulator activity	1.5826E-03
	GO:0030170	pyridoxal phosphate binding	1.5826E-03
	GO:0070279	vitamin B6 binding	1.5826E-03
	GO:0019843	rRNA binding	1.5826E-03
	GO:0019842	vitamin binding	3.1903E-03
HOXD-AS1	GO:0004714	transmembrane receptor protein tyrosine kinase activity	1.2937E-02
	GO:0033613	activating transcription factor binding	1.2937E-02
	GO:0019199	transmembrane receptor protein kinase activity	1.2937E-02
	GO:0001085	RNA polymerase II transcription factor binding	2.0767E-02
HOTAIR	GO:0005109	frizzled binding	1.9097E-03
	GO:0001227	transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding	1.9097E-03
	GO:0001664	G-protein coupled receptor binding	1.9686E-03
	GO:0001078	transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding	1.1196E-02
	GO:0008201	heparin binding	1.2885E-02
	GO:0005539	glycosaminoglycan binding	1.8790E-02
	GO:1901681	sulfur compound binding	1.9016E-02
	GO:0045236	CXCR chemokine receptor binding	2.1785E-02
	GO:0008301	DNA binding, bending	2.1785E-02
	GO:0001223	transcription coactivator binding	2.1785E-02
	GO:0042813	Wnt-activated receptor activity	2.1785E-02
	GO:0035198	miRNA binding	2.2807E-02
	GO:0017147	Wnt-protein binding	2.5258E-02
	GO:1990841	promoter-specific chromatin binding	2.5258E-02
	GO:0000982	transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	2.5258E-02
	GO:0001221	transcription cofactor binding	2.5587E-02

Table 4. KEGG pathway information of selected CEMs.

LncRNA	ID	Description	q-value
FOXCUT	hsa05216	Thyroid cancer	8.8587E-03
	hsa04510	Focal adhesion	8.8587E-03
	hsa05205	Proteoglycans in cancer	8.8587E-03
	hsa05218	Melanoma	1.7683E-02
	hsa05214	Glioma	1.7683E-02
	hsa04151	PI3K-Akt signaling pathway	1.8745E-02
	hsa05215	Prostate cancer	1.8745E-02
	hsa01522	Endocrine resistance	1.8745E-02
	hsa04919	Thyroid hormone signaling pathway	2.2317E-02
	hsa04152	AMPK signaling pathway	2.2317E-02
	hsa04068	FoxO signaling pathway	2.4442E-02
	hsa04550	Signaling pathways regulating pluripotency of stem cells	2.5129E-02
	hsa05224	Breast cancer	2.5129E-02
	hsa04218	Cellular senescence	2.7471E-02
	hsa05225	Hepatocellular carcinoma	2.7471E-02
	hsa04530	Tight junction	2.7471E-02
250-280	hsa03010	Ribosome	2.9900E-05
	hsa01210	2-Oxocarboxylic acid metabolism	3.7322E-03
	hsa00220	Arginine biosynthesis	3.7322E-03
	hsa00250	Alanine, aspartate and glutamate metabolism	4.7849E-03
	hsa01230	Biosynthesis of amino acids	7.9156E-03
HOXD-AS1	hsa05216	Thyroid cancer	8.8587E-03
	hsa04510	Focal adhesion	8.8587E-03
	hsa05205	Proteoglycans in cancer	8.8587E-03
	hsa05218	Melanoma	1.7683E-02
	hsa05214	Glioma	1.7683E-02
	hsa04151	PI3K-Akt signaling pathway	1.8745E-02
	hsa05215	Prostate cancer	1.8745E-02
	hsa01522	Endocrine resistance	1.8745E-02
	hsa04919	Thyroid hormone signaling pathway	2.2317E-02
	hsa04152	AMPK signaling pathway	2.2317E-02
	hsa04068	FoxO signaling pathway	2.4442E-02
	hsa04550	Signaling pathways regulating pluripotency of stem cells	2.5129E-02
	hsa05224	Breast cancer	2.5506E-02
	hsa04218	Cellular senescence	2.7471E-02
	hsa05225	Hepatocellular carcinoma	2.7471E-02
	hsa04530	Tight junction	2.7471E-02
HOTAIR	hsa04310	Wnt signaling pathway	5.7295E-03

3.6. Gene ontology terms and KEGG pathway information of selected CEMs

In order to further evaluate the biological significance of the identified CEMs, we tested the enrichment of the genes in each CEM in Gene ontology terms and KEGG pathways using clusterProfiler package of R project BioConductor under q-value cutoff 0.05, of which the description of the GO terms and KEGG pathways that the CEMs are enriched in are presented in Table 3 and Table 4 respectively. And the supplementary table S2 with more details, including original and adjusted P-value, proportion of the matched genes, gene's ID, etc., can be downloaded on GitHub link.

4. Discussion

Within this study, we have developed a method for elucidating lncRNA-gene and transcription factor-lncRNA interactions using a biclustering approach. The method was performed on 2 breast cancer RNA-seq datasets from TCGA. The bicluster method allows for the identification of particular expression patterns across multiple datasets, indicating networks of lncRNA and gene interactions. The developed method will also provide a way for future lncRNA interaction studies. Certainly, the predict performance still far from satisfactory, which is not unexpected since we only used RNA-Seq data. Actually, the interaction mechanism between lncRNA and genes are far more complex, and more data should be involved if we want to capture the whole picture of them. We are planning to include some other data, like proteomics and chromatin accessibility information, to improve the prediction. Besides, the evaluation on the relationship between lncRNA and predicted CEMs also has the potential to be improved, e.g. calculating the adjusted P-value or overall P-value in place of the original P-values used in this study. In view of the application, we will work on more specific examples of the regulatory functions of some particular lncRNAs and identify some hypothesized mechanisms of these regulatory functions. Also, the further analysis of the difference of lncRNA related genes between tumor and normal samples could provide more information for studying the process and mechanism of cancer occurrence and development, e.g. determination of the stage of developed tumors, which will be our concern in the future research.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (NSFC) [61772313 and 61432010], Young Scholars Program of Shandong University [YSPSDU, 2015WLJH19], the Innovation Method Fund of China (2018IM020200), and Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZHANGJIANGLAB. Qin Ma's work was supported by an R01 Award from the National Institute of General Medical Sciences of the National Institutes of Health [GM131399-01]. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation [ACI-1548562]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the National Science Foundation.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. V. S. Patil, R. Zhou and T. M. Rana, Gene regulation by non-coding RNAs, *Crit. Rev. Biochem. Mol. Biol.*, **49**(2014), 16–32.
2. J. Carlevaro-Fita, L. Liu, Y. Zhou, et al., LnCompare: gene set feature analysis for human long non-coding RNAs, *Nucleic Acids Res.*, **47**(2019), W523–W529.
3. K. W. Vance and C. P. Ponting, Transcriptional regulatory functions of nuclear long noncoding RNAs, *Trends Genet.*, **30**(2014), 348–355.
4. F. Ferre, A. Colantoni and M. Helmer-Citterich, Revealing protein-lncRNA interaction, *Brief Bioinform.*, **17**(2016), 106–116.
5. A. E. Kornienko, P. M. Guenzl, D. P. Barlow, et al., Gene regulation by the act of long non-coding RNA transcription, *BMC Biol.*, **11(**2013**)**, 59.
6. J. L. Rinn and H. Y. Chang, Genome regulation by long noncoding RNAs, *Annu. Rev. Biochem.*, **81**(2012), 145–166.
7. S. J. Liu, M. A. Horlbeck, S. M. Cho, et al., CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells, *Science*, **355**(2017), eaah7111.
8. Y. Meng and F. Yu, Long non coding RNA FAM3D-AS1 inhibits development of colorectal cancer through NF-kB signaling pathway, *Biosci. Rep.*, 39(2019), online published.
9. L. Peng, S. Gao, F. Bai, et al., LncRNA TPTE2P1 promotes the proliferation of thyroid carcinoma by inhibiting miR-520c-3p, *Panminerva Med.*, (2019), online published.
10. X. Chen, Q. Ma, X. Rao, et al., Genome-scale identification of cell-wall-related genes in switchgrass through comparative genomics and computational analyses of transcriptomic data, *BioEnergy Res.*, **9**(2016), 172–180.
11. S. Wang, Y. Yin, Q. Ma, et al., Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis, *BMC Plant Biol.*, **12**(2012), 138.
12. I. Ulitsky, A. Maron-Katz, S. Shavit, et al., Expander: From expression microarrays to networks and functions, *Nat. Protoc.*, **5**(2010), 303–322.
13. Z. Zhou, Y. Shen, M. R. Khan, et al., LncReg: A reference resource for lncRNA-associated regulatory networks, *Database* (*Oxford*)., (2015), bav083.
14. D. Smedley, S. Haider, S. Durinck, et al., The BioMart community portal: an innovative alternative to large, centralized data repositories, *Nucleic Acids Res.*, **43**(2015), W589-W598.
15. G. Li, Q. Ma, H. Tang, et al., QUBIC: A qualitative biclustering algorithm for analyses of gene expression data, *Nucleic Acids Res.*, **37**(2009), e101.
16. C. I. Castillo-Davis and D. L. Hartl, GeneMerge—post-genomic analysis, data mining, and hypothesis testing, *Bioinformatics*, **19**(2003), 891–892.
17. S. Heinz, C. Benner, N. Spann, et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, *Mol. Cell.*, **38**(2010), 576–589.
18. A. Khan, O. Fornes, A. Stigliani, et al., JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework, *Nucleic Acids Res.*, **46**(2017),

D260–D266.

19. J. Yang, X. Chen, A. McDermaid, et al., DMINDA 2.0: Integrated and systematic views of regulatory DNA motif identification and analyses, *Bioinformatics*, **33**(2017), 2586–2588.
20. Y. Zhang, J. Xie, J. Yang, et al., QUBIC: A bioconductor package for qualitative biclustering analysis of gene co-expression data, *Bioinformatics*, **33**(2017), 450–452.
21. J. Xie, A. Ma, Y. Zhang, et al., QUBIC2: A novel biclustering algorithm for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis, *bioRxiv*, (2018), 409961.
22. G. Li, B. Liu and Y. Xu, Accurate recognition of cis-regulatory motifs with the correct lengths in prokaryotic genomes, *Nucleic Acids Res*., **38**(2010), e12.
23. J. J. Quinn and H. Y. Chang, Unique features of long non-coding RNA biogenesis and function, *Nat. Rev. Genet*., **17**(2016), 47–62.
24. M. Rossi, G. Bucci, D. Rizzotto, et al., LncRNA EPR controls epithelial proliferation by coordinating Cdkn1a transcription and mRNA decay response to TGF-beta, *Nat. Commun*., **10**(2019), 1969.
25. G. Li, B. Liu, Q. Ma, et al., A new framework for identifying cis-regulatory motifs in prokaryotes, *Nucleic Acids Res*., **39**(2011), e42.