Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data

Junpeng Zhang[a], Thuc Duy Le[b], Lin Liu[b], Bing Liu[c], Jianfeng He[d], Gregory J Goodall[e], Jiuyong Li[b,*]

[a]*Faculty of Engineering, Dali University, Dali, CN*
[b]*School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, AU*
[c]*Children's Cancer Institute Australia, Randwick NSW 2301, AU*
[d]*Kunming University of Science and Technology, Kunming, CN*
[e]*Centre for Cancer Biology, SA Pathology, SA 5000, AU*

Abstract

Motivation: microRNAs (miRNAs) play crucial roles in complex cellular networks by binding to the messenger RNAs (mRNAs) of protein coding genes. It has been found that miRNA regulation is often condition-specific. A number of computational approaches have been developed to identify miRNA activity specific to a condition of interest using gene expression data. However, most of the methods only utilize the data in a single condition, thus the activity discovered may not be unique to the condition of interest. Additionally, these methods are based on statistical associations between the gene expression levels of miRNAs and mRNAs, so they may not be able to reveal real gene regulatory relationships, which are causal relationships.

*Corresponding author: Tel: +61 8 83023898; Fax: +61 8 83023381.
Email addresses: zhangjunpeng_411@yahoo.com (Junpeng Zhang), leyty017@mymail.unisa.edu.au (Thuc Duy Le), lin.liu@unisa.edu.au (Lin Liu), BLiu@ccia.unsw.edu.au (Bing Liu), jfenghe@kmust.edu.cn (Jianfeng He), greg.goodall@health.sa.gov.au (Gregory J Goodall), jiuyong.li@unisa.edu.au (Jiuyong Li)

August 2, 2014

Results: We propose a novel method to infer condition-specific miRNA activity by considering: (1) the difference between the regulatory behavior that a miRNA has in the condition of interest and its behavior in the other conditions; (2) the causal semantics of miRNA-mRNA relationships. The method is applied to the epithelial-mesenchymal transition (EMT) and multi-class cancer (MCC) datasets. The validation by the results of transfection experiments shows that our approach is effective in discovering significant miRNA-mRNA interactions. Functional and pathway analysis and literature validation indicate that the identified active miRNAs are closely associated with the specific biological processes, diseases and pathways. More detailed analysis of the activity of the active miRNAs implies that some active miRNAs show different regulation types in different conditions, but some have the same regulation types and their activity only differs in different conditions in the strengths of regulation.

Availability: The R and Matlab scripts are in the Supplementary materials.

---

1. Background

microRNAs (miRNAs) are a family of short non-coding RNA molecules (usually 19-25nt) that regulate gene expression via the full degradation of the target mRNA transcript or the translational repression of it [4]. miRNAs have been found to be involved in most biological processes, including developmental timing, cell proliferation, metabolism, differentiation, apoptosis, stress responses, cellular signaling and even various human cancers [1, 3, 9].

miRNA target prediction is a vital step towards the understanding of miRNA activity. Since experimental methods are limited by their low efficiency and high cost, computational approaches have become a key alternative for predicting miRNA activity. Several tools have been developed to identify miRNA targets, such as MicroCosm [12], PicTar [18], TargetScan [10] and miRanda [5]. However, the predictions are based on sequence complementarity and/or structural stability of the putative duplex and thus have a high rate of false positives and false negatives [28]. Furthermore, sequence data are static and they do not change in different conditions or at different times. Thus, from sequence data alone we are unable to identify the effect of miRNAs on their targets' expression in specific biological conditions, while miRNA regulation or activity is often condition-specific [19].

Some recent work has combined sequence data and gene expression data to infer miRNA activity. Cheng and Li [7] employed an enrichment score used by the Gene Set Enrichment Analysis (GSEA) [37] to infer miRNA activity. They identified the activity enhancement of miRNAs in miRNA transfected HeLa cells. Madden *et al.* [26] combined correspondence analysis, between group analysis and co-inertia analysis (CIA) to detect miRNA activity using microarray datasets. They produced a ranked list of miRNAs associated with a specific splitting in the samples, by combining miRNA target predictions with gene expression levels. Volinia *et al.* [41] proposed T-REX to build miRNA activity map. They used the effect of miRNAs over their targets for detecting miRNA activity with mRNA expression profiles. Some tools also have been developed, including miReduce [33], Sylamer [38], BIRTA [42], DIANA-mirExTra [2], mirAct [21], miTEA [36], and cWords [30], to infer

miRNA activity. The first three are stand-alone applications and the last four provide online services.

Although these methods were successfully applied to infer condition-specific miRNA activity, most of them only look at a specific condition (e.g. cancer), without considering the difference in miRNA activity between conditions. Therefore, the miRNA activity found based on the information in one condition of interest may contain regulatory relationships that are not unique to the condition. Moreover, when considering only one specific condition, the number of samples that can be used is smaller, worsening the over-fitting problem with high-dimensional gene expression data.

Additionally, most existing methods use statistical correlations or associations to identify miRNA-mRNA interactions. However, associations may not reveal gene regulatory relationships which are indeed causal relationships. For example, the expression levels of a miRNA and a gene can be strongly correlated, but the correlation may not indicate a regulatory relationship between the miRNA and the gene, because the strong correlation may be the consequence of the regulation of a common regulator of them.

To address the above limitations, in this paper, we propose a novel approach to discovering condition-specific miRNA activity.

To identify miRNA activity that is specific to a condition of interest, our method exploits the difference between the regulatory behavior of miRNAs in the condition of interest and in the other condition. We divide matched samples of miRNA and mRNA expression into groups according to sample conditions, e.g. cancer and normal. Then miRNA-mRNA causal interactions are examined using each group of samples respectively, but only those

interactions showing significant difference in their strengths in different conditions (called *significant causal interactions*) are retained. These significant causal interactions are then used to find out *active miRNAs* with respect to the condition of interest, i.e. miRNAs that have significantly different causal interactions with mRNAs in different conditions. The significant causal interactions associated with an active miRNA are the condition-specific activity of this miRNA.

To capture the causal semantics of miRNA-mRNA regulatory relationships, in the above procedure, we use IDA [24, 25], a causal inference method to estimate the strengths of miRNA-mRNA interactions. With observational data, IDA simulates an intervention process (e.g. a gene knock-down experiment) and predicts the causal effects of the intervention. It is proved to be an effective method for predicting the causal regulatory effects that a miRNA has on a mRNA [20].

To validate the proposed method, we apply it to two gene expression datasets: epithelial-mesenchymal transition (EMT) and multi-class cancer (MCC), respectively. The identified miRNA activity is validated by using the miRNA transfection experiments data, as well as by functional analysis, pathway analysis and the information from literature. The results show that the proposed method can effectively infer condition-specific miRNA activity.

2. Methods

2.1. Overview of the proposed method

As illustrated in Figure 1, the method comprises the following steps:

(1) Data preparation. Given the matched miRNA and mRNA expression

profiles, a list of differentially expressed miRNAs and mRNAs are identified. The expression profiles of the differentially expressed miRNAs and mRNAs are then split into sample groups according to the conditions (phenotypes) of the samples. In each condition, the miRNA and mRNA samples are matched and integrated into one dataset (matrix) as an input of the next step.

(2) Using IDA to learn a causal structure and to calculate the causal effects of each miRNA on each mRNA. This is done for each condition separately. To overcome the over-fitting problem of high-dimensional data, we use bootstrapping to improve the stability of the estimation of causal effects.

(3) Extracting significant miRNA-mRNA causal interactions. Kolmogorov-Smirnov (KS) test is used to evaluate the significance of the difference of the causal effects of a miRNA-mRNA causal interaction in different conditions. miRNA target binding information is used as a constraint in this step for extracting significant causal interactions, and an interaction that passes the KS test and is implied by the target binding information is selected as a significant miRNA-mRNA causal interaction.

(4) Detecting active miRNAs and condition-specific miRNA activity. For each miRNA, the difference of its significant causal interactions across all conditions is assessed using the KS test. If the test result is significant for the miRNA, then it is an active miRNA with respect to the condition of interest, and its activity in the condition (significant causal interactions with the mRNAs) is considered specific to the condition.

In the following, we will present the key steps in detail.

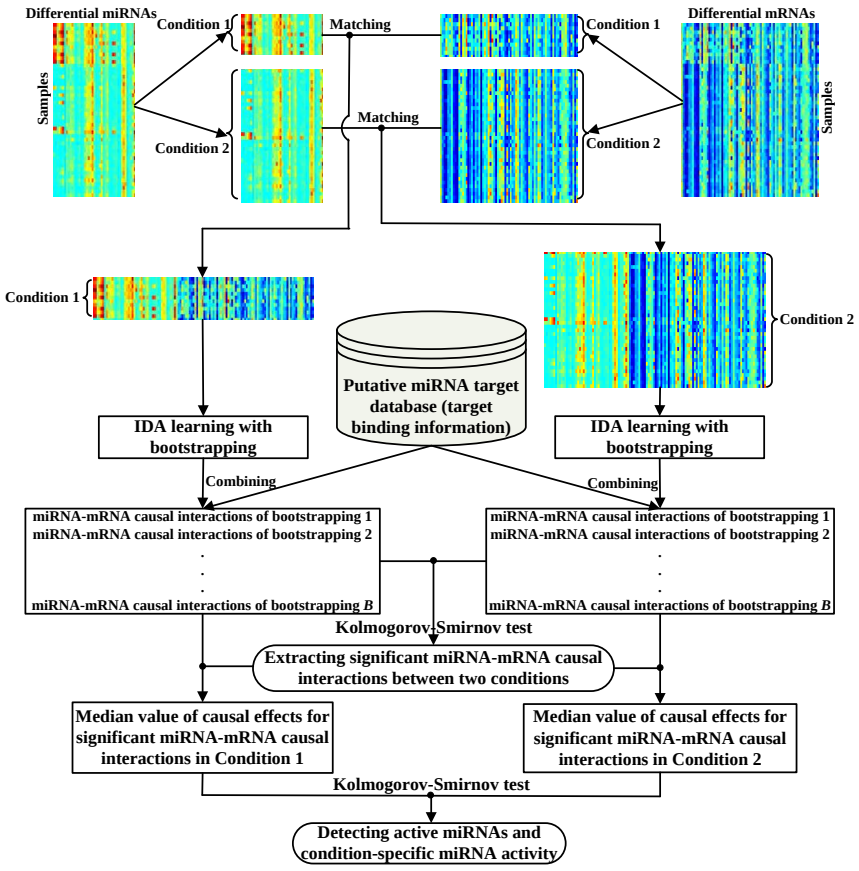


Figure 1: A flowchart of the proposed method. The expression profiles of differentially expressed miRNA and mRNAs are firstly split according to the conditions (phenotypes) of the samples, and in each condition the matched miRNA and mRNA expression samples are integrated. Then, we identify miRNA-mRNA causal relationships with IDA learning by combining target binding information, in each condition respectively. Bootstrapping is used to improve the result of IDA with high-dimensional data. Kolmogorov-Smirnov (KS) test is then conducted to identify significant miRNA-mRNA causal interactions. To determine if a miRNA is an active miRNA, in each condition, we obtain the median value of all causal effects obtained in all the B runs of bootstrapping for each significant miRNA-mRNA causal interaction associated with this miRNA, and use KS test to evaluate the significance of the difference of this miRNA's activity in the two conditions.

2.2. Causal inference with IDA

The application of IDA [24, 25] to matched miRNA and mRNA expression data can be divided into two steps: (1) learning a causal structure from expression data, and (2) calculating causal effects [20].

In step (1), the expression levels of miRNAs and mRNAs are represented by a set of random variables. The PC algorithm [35] is used to learn the causal structure of the variables in the form of a directed acyclic graph (DAG), where the nodes represent the random variables (miRNAs or mRNAs) and the edges denote causal relationships between these variables.

The PC algorithm is based on conditional dependence tests. Since different DAGs may encode the same conditional independencies in a given dataset, the output of the PC algorithm is an *equivalence class* of DAGs, which can be uniquely described by a *completed partially directed acyclic graph* (CPDAG) [24]. Learning a CPDAG from high-dimensional data is computational expensive, and we need to select an efficient conditional independent test to implement it. The PC algorithm with partial correlation test [14] is proved to be uniformly consistent in the high-dimensional context, thus we can use it to learn causal structures with gene expression data in inferring gene causal regulatory networks. In this paper, we use the R-package *pcalg* [15] which implements the PC algorithm with partial correlation test and set the significant level of the conditional independence test α=0.01.

In step (2), we simulate the controlled experiments with do-calculus [13], to estimate the causal effect that each miRNA has on a mRNA. Given a DAG, do-calculus can estimate the causal effect of a node on any other node in the DAG from observational data. For a miRNA-mRNA causal interaction, we

calculate the causal effect *ef*(miRNA, mRNA) based on each of the DAGs represented by the CPDAG learnt by the PC algorithm, respectively. We then use the minimum absolute value of the obtained causal effects as the final result of this step, to get a lower bound on the estimated strength of miRNA-miRNA causal interaction. For example, *ef*(miRNA, mRNA) $\in$ $\{0.75, 0.55, -0.7, 0.65\}$, then the final result *ef* (miRNA, mRNA) is 0.55. It suggests that the causal effect of the miRNA on the target gene is at least 0.55. Details of how the causal effects are calculated are out of the scope of this paper and interested readers are referred to [24, 25, 20] for more information.

Unstable estimation caused by the small number of samples is a challenge to the proposed method, and the problem may get more serious with high-dimensional gene expression data. To tackle this problem, we use a bootstrapping strategy. The above described IDA procedure is carried out in each run of the bootstrapping, and all the results will be used in the next step for identifying significant miRNA-mRNA causal interactions.

2.3. Identifying significant miRNA-mRNA interactions

As mentioned previously, to identify condition-specific miRNA activity, significant miRNA-mRNA causal interactions, i.e. those that vary in their strengths in the condition of interest and the other conditions should be the focus of the examination. The interactions that do not change much across different conditions are not unique to the condition of interest.

In order to evaluate the significance of a miRNA-mRNA causal interaction, we compare its causal effect, *ef*, calculated in the two different conditions using a two-sample Kolmogorov-Smirnov (KS) statistic test. The KS test can

assess whether the distribution of *ef* in the samples of one condition is significantly shifted compared to the distribution in the samples of the other condition. We choose to use KS test because it has the following advantages: (1) it is non-parametric and hence does not rely on any assumptions about the distribution of the changes of causal effects; (2) it does not rely on arbitrary thresholds; and (3) it measures significant shifts between the entire distribution rather than just comparing the tails.

Suppose that $F_j(ef) = \frac{1}{B}\sum_{i=1}^{B} I_{ef_i \leq ef}$ is the empirical cumulative distribution function (*cdf*) of *ef* in the two groups of samples, where $j \in \{1, 2\}$, B is the number of bootstrapping runs, and I is the indicator function whose value is 1 when $ef_i \leq ef$ and 0 otherwise. Then the KS test is the maximum difference (D) between the two groups of samples in value of the *cdfs*, i.e. $D = sup_{ef}|F_1(ef) - F_2(ef)|$, where sup_{ef} is the supremum of the set of difference.

We use the Matlab function *kstest2* to calculate the KS test statistic and the asymptotic p-value (adjusted by Benjamini-Hochberg (BH) method) of each miRNA-mRNA causal interaction. The miRNA-mRNA causal interaction with adjusted p-value less than 0.05 is regarded as a significant miRNA-mRNA causal interaction between the two conditions.

In the implementation, before conducting the KS test, for each miRNA-mRNA causal interaction, we check if the mRNA is a predicted target of the miRNA by using miRNA target binding information, and the interaction is undergoing the KS test only if the interaction is confirmed by the target binding information.

2.4. Inferring condition-specific miRNA activity

Generally, a single significant miRNA-mRNA interaction only shows the partial activity of the miRNA regarding the condition of interest, as the miRNA may be involved in multiple significant miRNA-mRNA interactions. Thus, to obtain a complete picture of the condition-specific activity of a miRNA we need to investigate all the significant interactions in which the miRNA is involved. Note that our definition of an active miRNA is specific to the condition of interest. The overall causal effect that the active miRNA has on all the mRNAs significantly interacting with it must have changed significantly between the condition of interest and the other condition.

To infer such condition-specific active miRNAs or their activity, firstly, for each identified significant miRNA-mRNA causal interaction, we find out the median value of its causal effects calculated during the B times of bootstrapping, in each condition respectively.

Then for each miRNA, we examine the difference of the distributions of the median causal effects of all its associated significant causal interactions in the condition of interest and the other condition, using the KS test. We also use the Matlab function *kstest2* to calculate the KS test statistic and the asymptotic p-value (adjusted by Benjamini-Hochberg (BH) method) for the miRNA. If the adjusted p-value is less than 0.05, then this miRNA is regarded as an active miRNA specific to the condition of interest.

3. Results

3.1. Data sources and preparation

To demonstrate our method, we apply it to the matched miRNA and mRNA expression profiles from the epithelial-mesenchymal transition (EMT) and multi-class cancer (MCC) datasets.

The miRNA expression profiles of EMT are from Søkilde *et al.* [34] (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26375`). They were profiled from the 60 cancer cell lines of the drug screening panel of human cancer cell lines at the National Cancer Institute (NCI-60). The mRNA expression profiles of EMT for NCI-60 were obtained from ArrayExpress (`http://www.ebi.ac.uk/arrayexpress`, accession number E-GEOD-5720). Samples of the EMT data categorized as epithelial (11 samples) and mesenchymal (36 samples) were used for this work.

The miRNA expression profiles of MCC were obtained from Lu *et al.* [22] (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2564`). The mRNA expression profiles of MCC are from Ramaswamy *et al.* [29] (`http://www.broad.mit.edu/cancer/pub/migcm`). Samples of the MCC data classified as normal (21 samples) and tumor (67 samples) were used in this work.

We perform differential gene expression analysis on the gene expression profiles to identify differentially expressed miRNAs and mRNAs between the two conditions in each dataset using the *limma* package [32] of Bioconductor. Genes with more than 10 missing values are removed. As a result, 46 probes of miRNAs and 1612 probes of mRNAs for the EMT dataset and 66 probes of miRNAs and 1318 probes of mRNAs for the MCC dataset are identified to be differentially expressed at significant level (adjusted p-value < 0.05,

adjusted by BH method). The detailed results of the differentially expressed miRNAs and mRNAs are in Supplementary Material 1.

We use the putative miRNA target information in MicroCosm v5 [12] as the constraint when identifying significant miRNA-mRNA causal interactions (see Figure 1). Note that the putative target information is an independent component in our method and any database of miRNA target information can be used. We choose MicroCosm to illustrate the method.

The number of bootstrapping, B is set to 100.

3.2. Validations by transfection experiment result

In this section, we validate identified significant miRNA-mRNA interactions using the transfection experimental data from [16], and the data included in the transfection data is listed in Supplementary Material 2. The 18 unique miRNAs overlap with 11 and 12 miRNAs in the EMT and MCC datasets respectively, which enables the validation of the significant interactions involving these miRNAs. Differentially expressed genes between the control and miRNA transfected samples are considered as targets of the miRNA, and are used as the ground truth to validate the predicted miRNA targets (significant miRNA-mRNA interactions). In the transfection experiment, mRNA differential expression levels are calculated by comparing mRNA expression levels between transfection and control samples via $\log_2$ fold change (LFC). The larger the absolute value of the LFC is, the more significant the mRNA differential expression level is. The commonly used fold change (FC) cutoffs are 1.5 and 2.0 [8], so we use the logarithm of the FC cutoffs and round them to 0.5 and 1.0 in this work. The following validations are done using the differentially expressed genes (ground truth)

obtained with both LFCs respectively.

3.2.1. Validation in comparison with MicroCosm

To assess how well the proposed method enriches the putative miRNA target information (MicroCosm), we compare the performance of the method with MicroCosm. For each dataset, we retrieve all the target genes of each miRNA predicted by the method and calculate the percentage of confirmed targets. As shown in Figure 2, our method overall performs better than MicroCosm in most miRNAs in both datasets. When the LFC cutoff is set to 1.0, our method also performs better than MicroCosm in terms of the number of validated targets (see Figure S1 in Supplementary Material 3). The results suggest that the method significantly enriches the putative target information used in the model. We also conceive a cumulative hypergeometric (HG) test to assess the statistical significance of the number of validated miRNA-mRNA interactions in our method. We found that the number of validated miRNA-mRNA interactions are statistically significant (p-value < 0.05) in both the EMT and MCC datasets (see Table S1 in Supplementary Material 3 for details).

3.2.2. Validation in comparison with non condition-specific approach

To evaluate the effectiveness of the condition-specific approach, we compare the performance of the proposed method with its non condition-specific variant that does not consider sample categories. The non condition-specific approach will not split the dataset into different conditions and it simply applies IDA to the whole dataset to enrich the putative target information.

We extract the top 10 and 20 predicted targets for the all transfected

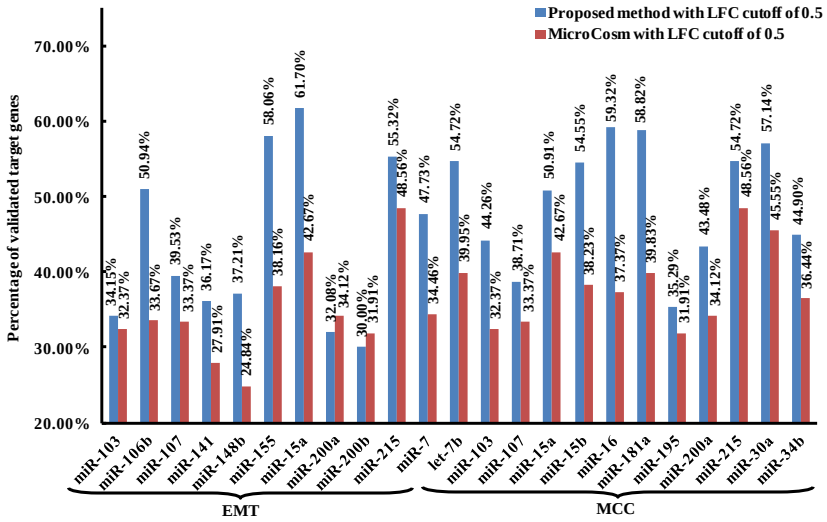
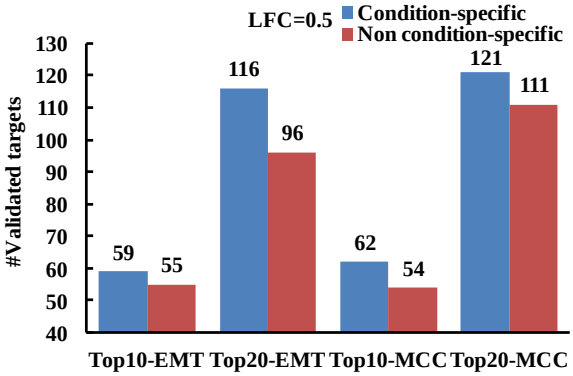


Figure 2: The percentage of confirmed target genes identified by using the proposed method and MicroCosm in the EMT and MCC dataset (LFC=0.5).

miRNAs in the EMT and MCC datasets (11 for EMT and 12 in MCC) and compare the total number of validated targets for the two approaches. Figure 3 shows that the proposed condition-specific approach predicts more confirmed targets than that by the non condition-specific approach in all cases (Top10-EMT, Top20-EMT, Top10-MCC and Top20-MCC). When the LFC cutoff in transfection experiments is set to 1.0, the proposed method also performs better than non condition-specific method in terms of the number of validated targets (see Figure S2 in Supplementary Material 3).

3.2.3. Validation in comparison with the cases using correlation methods

In order to show the advantage of causal inference, in this section, we replace step (2) of our proposed method (see Section 2.1) with each of the five correlation methods, Pearson, Spearman, Kendall, Lasso and Elastic-net

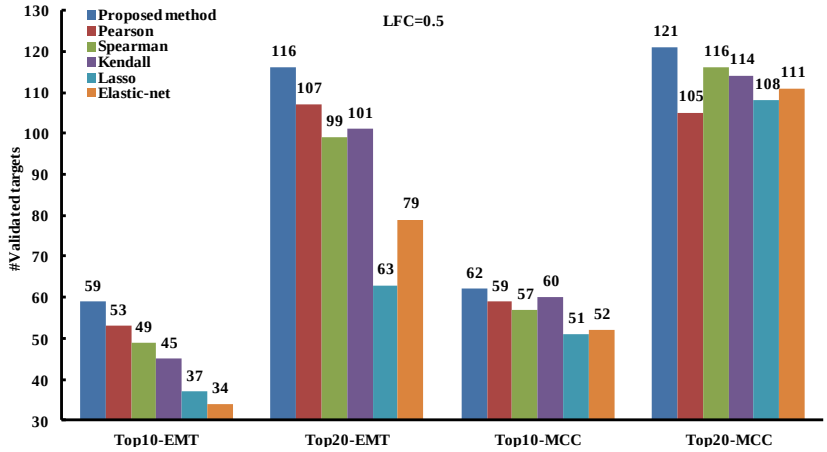


p-values	Top10-EMT	Top20-EMT	Top10-MCC	Top20-MCC
Condition-specific	1.0364E-05	2.6406E-09	0.0071	8.0926E-04
Non condition-specific	2.4648E-04	0.0010	0.1638	0.0325

Figure 3: Comparison between condition-specific and non condition-specific analyses in terms of the number of validated targets (LFC=0.5). *p*-values of validated targets are calculated using cumulative hypergeometric test.

(called the correlation methods), and compare their results with the results obtained by the proposed method in terms of the number of validated miRNA targets.

We extract the top 10 and 20 predicted targets for all transfected miRNAs in the EMT and MCC datasets (11 in EMT and 12 in MCC), and compare the total number of validated targets obtained by different methods. As illustrated in Figure 4, our method performs better than all the correlation methods in all cases (Top10-EMT, Top20-EMT, Top10-MCC and Top20-MCC). When the LFC cutoff is set to 1.0, the proposed method also outperforms all the five correlation methods in the number of validated targets (see Figure S3 in Supplementary Material 3).



p-values	Proposed method	Pearson	Spearman	Kendall	Lasso	Elastic-net
Top10-EMT	1.0364E-05	9.7664E-04	0.0102	0.0625	0.5259	0.7512
Top20-EMT	2.6406E-09	2.0966E-06	2.3565E-04	8.0484E-05	0.9490	0.2455
Top10-MCC	0.0071	0.0286	0.0627	0.0185	0.3362	0.2714
Top20-MCC	8.0926E-04	0.1451	0.0062	0.0126	0.0733	0.0325

Figure 4: Comparison between the proposed method and the correlation methods in terms of the number of confirmed targets (LFC=0.5). p-values of validated targets are calculated using cumulative hypergeometric test.

3.3. Inferring condition-specific active miRNAs

Using our method, we have identified 18 and 41 active miRNAs in the EMT and MCC datasets, respectively. The identified miRNA activity with the KS tests and the box plots of causal effects of the active miRNAs in both datasets are provided in Supplementary Material 4.

3.3.1. Causal effects vs. correlations for inferring condition-specific active miRNAs

To show the effectiveness of using causal effects as the measure of the strength of miRNA-mRNA interactions, we also use the five correlation methods (Pearson, Spearman, Kendall, Lasso and Elastic-net) for detecting condition-specific miRNA activity. That is, in Step (2) of our method (Section 2.1), instead of using IDA, we use one of the correlation methods to compute the strength of an interaction, and the obtained values instead of casual effects are used in the next two steps.

Since no benchmarks are available, to compare the performance of each method in inferring condition-specific miRNA activity, we use the number of identified active miRNAs out of the differentially expressed miRNAs as the criterion. If a method identifies the largest number of condition-specific active miRNAs, the method performs the best. As shown in Table 1, our method significantly outperforms all the five correlation methods in detecting active miRNAs suggesting that causal effect is a useful measure to detect active miRNAs. The results of condition-specific miRNA activity using the five correlation methods are provided in Supplementary Material 5.

Table 1: Comparing the number of active miRNAs found by the proposed method and the five correlation methods.

Dataset	Proposed method	Pearson	Spearman	Kendall	Lasso	Elastic-net
EMT	18	6	2	2	0	0
MCC	41	38	26	24	1	5

3.3.2. Comparing with existing methods in inferring condition-specific active miRNAs

We evaluate the performance of our method by comparing it with other five existing methods: DIANA-mirExTra [2], Sylamer [38], MIR [7], miReduce [33] and cWords [30]. Similarly, a miRNA with p-value less than 0.05 is regarded as an active miRNA. We also use the number of identified active miRNAs out of the differentially expressed miRNAs as the criterion. As illustrated in Table 2, for the EMT dataset, our method is comparable with cWords and performs better than the other four existing methods. For the MCC dataset, our method outperforms all the five existing methods. The results of condition-specific active miRNAs (p-value$<$0.05) using the five existing methods are provided in Supplementary Material 6.

3.4. Validation of active miRNAs

We use the TAM [23] software to conduct the functional analysis of the active miRNAs found by our method. Significant biological functions and associated diseases are identified for an active miRNA with the adjusted p-value (adjusted by BH method) of 0.05. The analysis of the molecular pathways that the active miRNAs are potentially involved is performed with

Table 2: Comparing the number of active miRNAs found by the proposed method and the five existing methods.

Dataset	Proposed method	DIANA-mirExTra	Sylamer	MIR	miReduce	cWords
EMT	18	5	10	13	3	18
MCC	41	35	20	10	0	6

mirPath [40], and TarBase 6.0 [39] is regarded as a reference database to mine significantly enriched pathways.

For the EMT dataset, the 47 samples are closely related to 9 human cancer cell lines, Breast, Cardiovascular Nervous System, Colon, Leukemia, Lung, Melanoma, Ovarian, Prostate and Renal. Here, we only discuss EMT in the biological functions and diseases associated with these 9 human cancer cell lines.

As illustrated in Figure 5, 9 out of the 18 active miRNAs identified using our method are significantly associated with epithelial-mesenchymal transition, and 11 miRNAs are closely related to Breast Neoplasms. As shown in Figure 6, out of the 18 active miRNAs, most of them are significantly enriched in the top five KEGG pathways, including Pathways in cancer, Chronic myeloid leukemia, Cell cycle, HTLV-I infection and Colorectal cancer.

Our method has found a number of literature-confirmed active miRNAs in EMT, including 4 members (miR-141, miR-200a, miR-200c, and miR-429) of the miR-200 family and 2 members (miR-192 and miR-215) of the miR-192 family. Previous studies [11, 27] have revealed that members of the miR-200 family play a critical role in the suppression of EMT, tumor cell adhesion,

migration, invasion and metastasis, and may have therapeutic implications for the treatment of metastatic and drug-resistant tumors. The miR-200 family and miR-192 family are critical mediators of p53-regulated EMT [17].

The MCC samples are closely associated with 11 human cancer lines, including Bladder, Breast, Colon, Lung, Melanoma, Mesothelioma, Ovarian, Pancreas, Prostate, Renal and Uterus. In the results, out of the 41 miRNAs indentified to be active between the normal and tumor samples, 20 miRNAs are shown to be significantly associated with miRNA tumor suppressors in Figure 5. Furthermore, many active miRNAs are significantly associated with Carcinoma of Renal Cell, Colonic Neoplasms, Lung Neoplasms, Melanoma, Ovarian Neoplasms, Pancreatic Neoplasms and Prostatic Neoplasms.

The pathway analysis indicates that more than half of the 41 active miRNAs are significantly enriched in the top five KEGG pathways including Prostate cancer, Pathways in cancer, Chronic myeloid leukemia and Bladder cancer, except RNA transport (see Figure 6).

Our method has also identified that 6 members (let-7a, let-7b, let-7c, let-7d, let-7f and let-7g) of the let-7 family, 2 members (miR-181a and miR-181c) of the miR-181 family and 2 members (miR-29a and miR-29c) of the miR-29 family are active in the process of tumor. Recent research [6] has found out that let-7 and its family members are highly conserved across species in sequence and function, and misregulation of the let-7 family leads to a less differentiated cellular state and the development of cell-based diseases such as cancer. The miR-181 family has been demonstrated to play an important role in occurrence and progression of malignant tumors such as lung cancer, pancreatic cancer, prostate cancer and breast cancer [43]. The miR-29

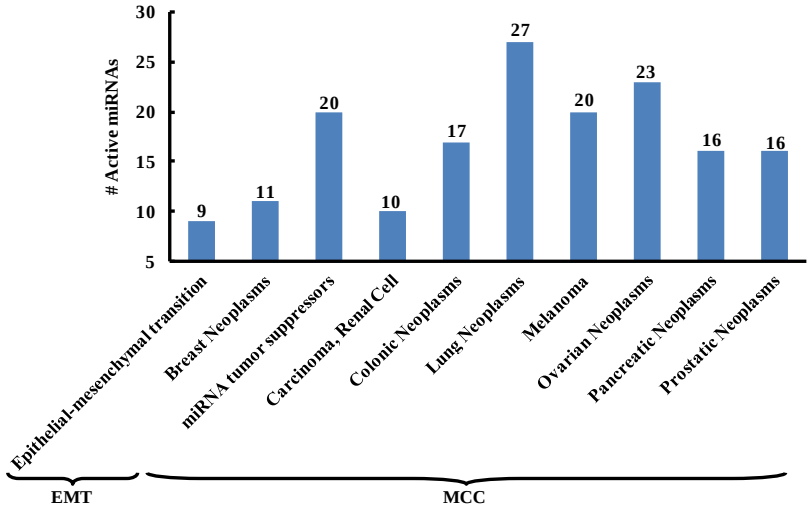


Figure 5: Functional analysis of active miRNAs. The results are generated by TAM, and significant biological functions and associated diseases are identified with a p-value cutoff of 0.05.

family has also been shown to be silenced or down-regulated in many different types of cancer and have subsequently been attributed predominantly tumor-suppressing properties [31].

3.5. Condition-specific miRNA activity

To understand the types of regulation of active miRNAs on their targets in different conditions, we compare the number of positive and negative effect of each active miRNA on their targets. With our method, a positive (negative) causal effect indicates up-regulation (down-regulation) of the miRNA on its interacting mRNA. If the number of negative effects of an active miRNA on its targets is more than that of positive effects on its targets in one condition, the active miRNA dominantly down-regulates its target genes in the condition, and vice versa. When an active miRNA has the same

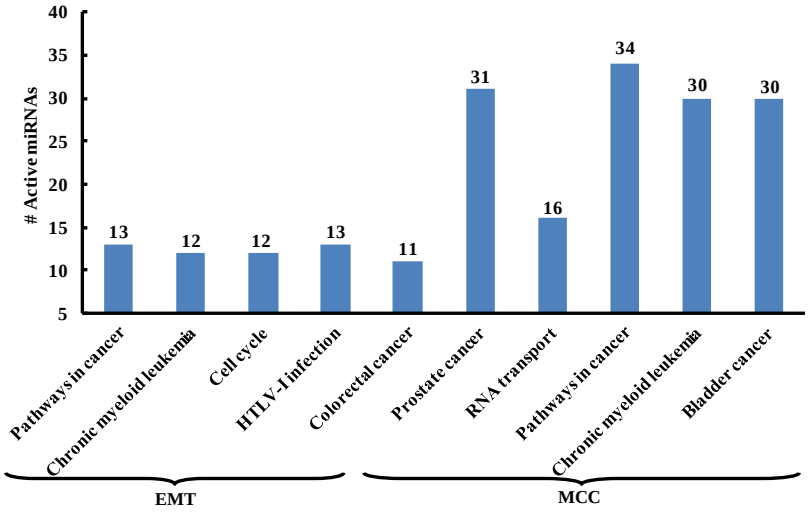


Figure 6: Pathway analysis of active miRNAs. The results are generated by mirPath, the top five KEGG pathways are listed for both datasets.

number of negative effects and positive effects on its targets in a condition, the regulation type of the active miRNA is uncertain in the condition.

As shown in Table 3, most active miRNAs identified using our method (13 out of 18) down-regulate their target genes in class E (epithelial), but most active miRNAs (10 out of 18) up-regulate their targets in class M (mesenchymal). Most active miRNAs (11 out of 18) have different regulation types between E and M. For the MCC dataset, most active miRNAs (30 out of 41) up-regulate their target genes in class N (normal), but most active miRNAs (23 out of 41) down-regulate their targets in class T (tumor). In total, 17 active miRNAs have different regulation types between N and T.

The results indicate that in each sample condition there is a dominant regulation type. The results also show that some active miRNAs behave differently in different conditions in the types of regulation (up or down), but

Table 3: The number of active miRNAs showing down-regulation (Down), up-regulation (Up) and different-regulation (Dif) in EMT and MCC dataset. E: Epithelial; N: Normal; M: Mesenchymal; T: Tumor. In the MCC dataset, 4 active miRNAs have uncertain regulation type in N and 1 in T.

Dataset	#Down in E(N)	#Up in E(N)	#Down in M(T)	#Up in M(T)	#Dif in both conditions
EMT	13	5	8	10	11
MCC	7	30	23	17	17

some active miRNAs have the same regulation type in different conditions and the difference across conditions is just the strengths of their regulation. If we only look at the whole dataset without considering the difference between conditions, we may miss the interactions in a specific condition (e.g. cancer).

4. Conclusions

miRNAs have been regarded as the main regulators at the post-transcriptional level. Identifying the targets of miRNAs is a fundamental task in predicting miRNA functions. Great efforts have been made to elucidate miRNA functions and regulatory mechanism. One stream of the research is focused on miRNA activity specific to a condition of interest. However, most of the studies only utilize samples obtained in the specific condition, without examining the difference of miRNA behavior in the specific condition and the other conditions, thus the miRNA activity discovered may not be unique to the specific condition. Furthermore, most computational methods only use associations or correlations in predicting miRNA-mRNA regulation while the

regulation is in fact causal relationships.

In this study, we have proposed an alternative method to reveal 'truly' condition-specific miRNA activity with the consideration of the causal semantics of miRNA-mRNA relationships.

We have applied our method to the EMT and MCC datasets. The validation with transfection experiment data illustrates that our method is more efficient than MicroCosm v5 in identifying the miRNA targets, and considering the difference across different sample conditions improves the number of validated interactions.

The comparison with five correlation methods demonstrates that causal effects provide a better measure than correlations in modeling the strengths of miRNA-mRNA interactions, leading to more effective discovery of active miRNAs.

As the main aim of the paper is to identify condition-specific active miRNAs and their activity, we conduct function and pathway analysis of the active miRNAs detected using our method. The results have shown that a significant number of the identified active miRNAs are closely related to the biological functions associated with the conditions of samples in the EMT and MCC datasets, and play a vital role in the potential pathogenesis of complex diseases. Furthermore, to understand the activity of the active miRNAs, we investigate how these miRNA behave differently in different conditions. It was found out that some active miRNAs show different regulation types in different conditions and some active miRNAs have the same regulation types and their activity only differs in different conditions in terms of the strengths of regulation.

In conclusion, the validation and analysis results indicate that the proposed method can be an effective method to detect condition-specific miRNA activity.

Acknowledgement

This work has been partially supported by the Applied Basic Research Foundation of Science and Technology of Yunnan Province (No: 2013FD038), the Australian Research Council Discovery grant DP130104090, and the Science Research Foundation for Youth Scholars of Dali University (No: KYQN201203).

References

[1] Ambros,V. (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing, *Cell*, **113**, 673-6.

[2] Alexiou,P. *et al.* (2010) The DIANA-mirExTra web server: from gene expression data to microRNA function, *PLoS One*, **5**, e9171.

[3] Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, **116**, 281-97.

[4] Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions, *Cell*, **136**, 215-33.

[5] Betel,D. *et al.* (2008) The microRNA.org resource: targets and expression, *Nucleic Acids Res.*, **36**, D149-D153.

[6] Boyerinas,B. *et al.* (2010) The role of let-7 in cell differentiation and cancer, *Endocr. Relat. Cancer*, **17**, F19-36.

[7] Cheng,C. and Li,L.M. (2008) Inferring microRNA activities by combining gene expression with microRNA target prediction, *PLoS One*, **3**, e1989.

[8] Dalman,M.R. *et al.* (2012) Fold change and p-value cutoffs significantly alter microarray interpretations, *BMC Bioinformatics*, **13**, S11.

[9] Du,T. and Zamore,P.D. (2007) Beginning to understand microRNA function, *Cell Res.*, **17**, 661-3.

[10] Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs, *Genome Res.*, **19**, 92-105.

[11] Gregory,P.A. *et al.* (2008) MicroRNAs as regulators of epithelial-mesenchymal transition, *Cell Cycle*, **7**, 3112-8.

[12] Griffiths-Jones,S. *et al.* (2008) miRBase: tools for microRNA genomics, *Nucleic Acids Res.*, **36**, D154-8.

[13] Judea,P. (2000) Causality: Models, Reasoning, and Inference, Cambridge University Press, New York, USA.

[14] Kalisch,M. and Bühlmann,P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm, *J. Mach. Learn. Res.*, **8**, 613-636.

[15] Kalisch,M. *et al.* (2012) Causal Inference Using Graphical Models with the R Package pcalg, *J. Stat. Softw.*, **47**, 1-26.

[16] Khan,A.A. *et al.* (2009) Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs, *Nat. Biotechnol.*, **27**, 549-55.

[17] Kim,T. *et al.* (2011) p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2, *J. Exp. Med.*, **208**, 875-83.

[18] Krek,A. *et al.* (2005) Combinatorial microRNA target predictions, *Nat. Genet.*,, **37**, 495-500.

[19] Le,H.S. and Bar-Joseph,Z. (2013) Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation, *Bioinformatics*, **29**, i89-97.

[20] Le,T.D. *et al.* (2013) Inferring microRNA-mRNA causal regulatory relationships from expression data, *Bioinformatics*, **29**, 765-771.

[21] Liang,Z. *et al.* (2011) mirAct: a web tool for evaluating microRNA activity based on gene expression data, *Nucleic Acids Res.*, **39**, W139-44.

[22] Lu,J. *et al.* (2005) MicroRNA expression profiles classify human cancers, *Nature*, **435**, 834-8.

[23] Lu,M. *et al.* (2010) TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs, *BMC Bioinformatics*, **11**, 419.

[24] Maathuis,H.M. *et al.* (2009) Estimating high-dimensional intervention effects from observational data, *Ann. Stat.*, **37**, 3133-3164.

[25] Maathuis,H.M. *et al.* (2010) Predicting causal effects in large-scale systems from observational data, *Nat. Methods*, **7**, 247-249.

[26] Madden,S.F. *et al.* (2010) Detecting microRNA activity from gene expression data, *BMC Bioinformatics*, **11**, 257.

[27] Mongroo,P.S. and Rustgi,A.K. (2010) The role of the miR-200 family in epithelial-mesenchymal transition, *Cancer Biol. Ther.*, **10**, 219-22.

[28] Rajewsky,N. (2006) microRNA target predictions in animals, *Nat. Genet.*, **38**, S8-13.

[29] Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures, *Proc. Natl. Acad. Sci. USA*, **98**, 15149-54.

[30] Rasmussen S.H. *et al.* (2013) cWords - systematic microRNA regulatory motif discovery from mRNA expression data, *Silence*, **4**, 2.

[31] Schmitt,M.J. *et al.* (2013) MiRNA-29: a microRNA family with tumor-suppressing and immune-modulating properties, *Curr. Mol. Med.*, **13**, 572-85.

[32] Smyth,G.K. (2005) Limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman (Ed.). New York: Springer, 397-420.

[33] Sood,P. *et al.* (2006) Cell-type-specific signatures of microRNA on target mRNA expression, *Proc. Natl. Acad. Sci. USA*, **103**, 2746-2751.

[34] Søkilde,R. *et al.* (2011) Global microRNA analysis of the NCI-60 cancer cell panel, *Mol. Cancer Ther.*, **10**, 375-84.

[35] Spirtes,P. *et al.* (2000) Causation, Prediction, and Search, MIT Press, Cambridge, MA.

[36] Steinfeld,I. (2013) miRNA target enrichment analysis reveals directly active miRNAs in health and disease, *Nucleic Acids Res.*, **41**, e45.

[37] Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. USA*, **102**, 15545-50.

[38] van Dongen, S. *et al.* (2008) Detecting microRNA binding and siRNA off-target effects from expression data, *Nat. Methods*, **5**, 1023-1025.

[39] Vergoulis,T. *et al.* (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support, *Nucleic Acids Res.*, **40**, D222-9.

[40] Vlachos,I.S. *et al.* (2012) DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways, *Nucleic Acids Res.*, **40**, W498-504.

[41] Volinia,S. *et al.* (2010) Identification of microRNA activity by Targets' Reverse EXpression, *Bioinformatics*, **26**, 91-7.

[42] Zacher,B. *et al.* (2012) Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data, *Bioinformatics*, **28**, 1714-20.

[43] Zhu,Y.K. *et al.* (2012) Advances in research on miR-181 family members and malignant tumors, *Tumor*, **32**, 837-841.