PLOS ONE

RESEARCH ARTICLE

Predicting miRNA Targets by Integrating Gene Regulatory Knowledge with Expression Profiles

Weijia Zhang[1], Thuc Duy Le[1], Lin Liu[1], Zhi-Hua Zhou[2], Jiuyong Li[1]*

1 School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, South Australia, Australia, **2** National Key Labotorary, Nanjing University, Nanjing, Jiangsu, China

* jiuyong.li@unisa.edu.au

Citation: Zhang W, Le TD, Liu L, Zhou Z-H, Li J (2016) Predicting miRNA Targets by Integrating Gene Regulatory Knowledge with Expression Profiles. PLoS ONE 11(4): e0152860. doi:10.1371/journal.pone.0152860

Editor: Yun Zheng, Kunming University of Science and Technology, CHINA

Received: December 9, 2015

Accepted: March 21, 2016

Published: April 11, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work is supported by Australian Research Council Discovery Project DP130104090 (in part). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Motivation

microRNAs (miRNAs) play crucial roles in post-transcriptional gene regulation of both plants and mammals, and dysfunctions of miRNAs are often associated with tumorigenesis and development through the effects on their target messenger RNAs (mRNAs). Identifying miRNA functions is critical for understanding cancer mechanisms and determining the efficacy of drugs. Computational methods analyzing high-throughput data offer great assistance in understanding the diverse and complex relationships between miRNAs and mRNAs. However, most of the existing methods do not fully utilise the available knowledge in biology to reduce the uncertainty in the modeling process. Therefore it is desirable to develop a method that can seamlessly integrate existing biological knowledge and high-throughput data into the process of discovering miRNA regulation mechanisms.

Results

In this article we present an integrative framework, CIDER (**C**ausal mi**R**NA target **D**iscovery with **E**xpression profile and **R**egulatory knowledge), to predict miRNA targets. CIDER is able to utilise a variety of gene regulation knowledge, including transcriptional and post-transcriptional knowledge, and to exploit gene expression data for the discovery of miRNA-mRNA regulatory relationships. The benefits of our framework is demonstrated by both simulation study and the analysis of the epithelial-to-mesenchymal transition (EMT) and the breast cancer (BRCA) datasets. Our results reveal that even a limited amount of either Transcription Factor (TF)-miRNA or miRNA-mRNA regulatory knowledge improves the performance of miRNA target prediction, and the combination of the two types of knowledge enhances the improvement further. Another useful property of the framework is that its performance increases monotonically with the increase of regulatory knowledge.

Introduction

miRNAs are short non-protein coding RNAs that regulate gene expression by either marking their target mRNAs for degradation or repressing translation. miRNAs mainly identify their target mRNAs by binding to the 3'-untranslated region (3' UTR) or 5' UTR. Studies have shown that miRNAs play important roles in a broad range of biological processes, such as differentiation [1], development [2], apoptosis [3] and cellular signaling [4]. Because of their biological importance, miRNAs are related to a variety of diseases, such as cancer and cardiovascular diseases [5]. Therefore, precise identification of miRNA targets is critical to the understanding of the functions of miRNAs in both healthy and diseased tissues [6, 7].

Computational approaches are a necessary and promising way to help unveil the complete picture of miRNA regulatory relationships. Significant progress has been made in elucidating the relationships between miRNAs and their targets using wet-lab biological experiments [8–11]. However, it is unrealistic to hope for a complete picture of miRNA regulation mechanisms by relying solely on wet-lab experiments due to the huge number of possible relationships and high expenses of the experiments [12]. Therefore, dry-lab approaches have been considered as a cost-effective and promising alternative and have shown great promise in identifying putative miRNA targets [13–16].

Because of the large number of miRNAs and mRNAs involved in gene regulation, providing reliable predictions has always been a significant challenge for computational biology approaches. This problem is further exacerbated by the small number of available samples. Therefore researchers have to rely on the integration of biological knowledge and data driven discovery process to obtain a complete understanding of miRNA regulation mechanisms.

Bayesian network (BN) [17–22] provides an excellent platform for seamless integration of prior knowledge and data in the process of causal structure learning. Furthermore, the causal semantics of a BN makes it a preferred model for representing gene regulatory networks since the interactions among genes are causal relationships rather than statistical associations.

Valuable wet-lab validated knowledge cannot be effectively utilised with the existing methods [23–26]. These algorithms use prior knowledge to restrict their search space in the way that the knowledge is used to initialise the structure of a BN and the learning process is aimed at removing false positives from the initial structure [27–31]. Therefore the final structure is a sub-graph of the initial one and a miRNA-mRNA interaction will not be predicted if it is not included in the prior knowledge. Consequently such methods usually require users to have a large amount of knowledge which covers the complete or nearly complete knowledge of the network structure, and are not able to utilise the sparse and limited validated knowledge.

In this paper, we propose the CIDER framework to effectively utilise sparse wet-lab validated knowledge, including transcriptional miRNA-mRNA and post-transcriptional TF-miRNA regulatory knowledge [32]. Our method differentiates from the existing work in two aspects: first instead of using the regulatory knowledge to initiate the network structure and then remove false positive edges, we enforce the learning process to maintain the experimentally confirmed relationships without restricting the search space. Secondly the regulatory knowledge is used for the purpose of obtaining more accurate estimation of the causal effect of miRNAs on mRNAs, whereas existing methods use prior knowledge to learn the causal regulatory structure.

Our results on both real-world and simulated datasets demonstrate that a very small amount of validated regulatory knowledge improves the accuracy of predicted miRNA targets significantly, and the performance of CIDER increases monotonically with the increase of regulatory knowledge.

We show that when wet-lab validated knowledge is analysed together with expression profiles, CIDER discovers significantly more validated miRNA targets than using expression profiles alone. It is also shown that either TF-miRNA or miRNA-mRNA regulatory knowledge improves the performance, and the combination of the two types of knowledge enhances the performance further.

An important property of the framework is that the performance of miRNA target prediction improves monotonically with the amount of regulatory knowledge used. In other words CIDER makes more reliable discoveries from the data when the knowledge integrated into the framework increases. In Fig 1, we illustrate a promising knowledge discovery process based on this property. With the incorporation of regulatory knowledge in CIDER, the process becomes a feedback loop for the discovery of new biological hypotheses and it naturally combines dry-lab predictions with web-lab experiments.

Materials

Matched expression profiles

NCI-60 data for Epithelial to Mesenchymal Transition (EMT). The EMT [33] dataset includes the miRNA expression profiles for the NCI-60 panel cell lines from [34], and the dataset is available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26375. The mRNA expression profiles for NCI-60 were downloaded from ArrayExpress available at http://www.ebi.ac.uk/arrayexpress, accession number E-GEOD-5720. We use the cell lines categorized as epithelial (11 samples) and mesenchymal (36 samples) in this study.

Data of the 51 human breast cancer cell lines (BRCA). The BRCA dataset includes miRNA expression profiles from the breast cancer cell lines data provided by [35]. The mRNA expression profiles for these cell lines can be downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41313. 27 samples in the luminal group and 23 samples in the basal group are used.

Gene regulation databases

TF-miRNA interaction database. For transcriptional regulatory knowledge, we use TransmiR [36], a TF-miRNA regulatory relationships database including approximately 700 entries manually collected from relevant literatures. This database is available online at http://www.cuilab.cn/transmir.

Experimentally validated miRNA-mRNA interaction databases. The post-transcriptional regulatory knowledge is obtained from miRNA target databases Tarbase v6.0 [37], miRTarbase v4.5 [13] and miRWalk [38]. Tarbase and miRTarbase contain experimentally

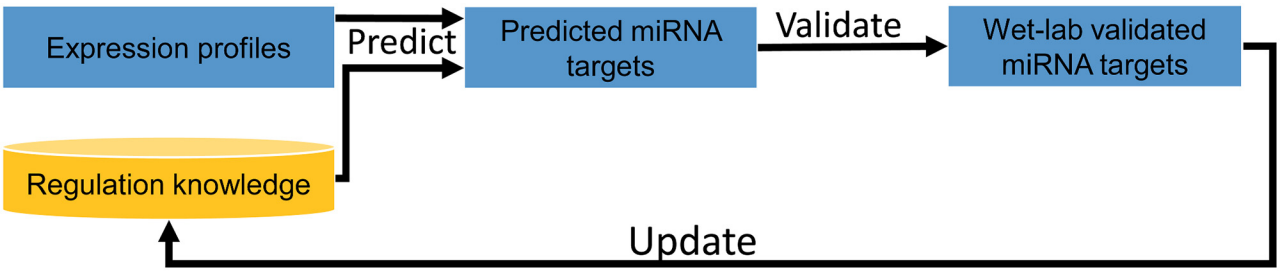

Fig 1. An iterative process of integrating and discovering miRNA regulatory relationships. Our proposed framework is one iteration of the above knowledge and data integrated discovery process. In the long run, wet-lab and dry-lab discoveries become an integrated feedback process for uncovering new biological insights. Bayesian network based causal reasoning provides an excellent platform for a seamless integration.

doi:10.1371/journal.pone.0152860.g001

confirmed miRNA target information manually collected from related literatures. miRWalk contains both predicted and validated miRNA targets, but we only utilise the experimentally validated targets in our experiments. The detailed information of experimentally validated miRNA-mRNA interactions retrieved from all these databases can be found in S3 File.

Predicted miRNA-mRNA interaction database. We also utilise TargetScan v6.2 [39], a commonly used miRNA target prediction database. TargetScan predicts miRNA targets by searching for binding sites that match the seed region of each miRNA. This database is available online at http://www.targetscan.org.

Methods

Notation

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ denote a *graph* where $\mathbf{V} = \{X_1, \ldots, X_l\}$ is a set of *vertices* and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is a set of *edges*. In our framework, the vertex set $\mathbf{V}$ represents a set of random variables corresponding to the expression levels of miRNAs and mRNAs (including TF coding mRNAs), and the edges represent the causal relationships between the variables.

We use $X_i \rightarrow X_j$ or $X_i \leftarrow X_j$ to represent a directed edge between X_i and X_j. $X_i - X_j$ is used to represent an undirected edge between X_i and X_j. The set of all parent nodes of X_j is denoted as pa_j. A *directed* graph is a graph in which all edges are directed. An *undirected* graph is a graph in which all edges are undirected. We say that a graph $\mathcal{G}$ is *acyclic* if and only if all its directed edges do not form any cycle in $\mathcal{G}$. In this article, we always assume the graph is acyclic.

The proposed CIDER framework

As illustrated in Fig 2, the CIDER framework consists of three steps. In the first step we perform differential gene expression analysis and query the databases for gene regulation knowledge. To identify the targets of a miRNA, we use *do-calculus* [18] to estimate the causal effects the miRNA have on all the mRNAs. In other words, *do-calculus* estimates how the expression values of the mRNAs change when the expression of the miRNA is intervened [41]. In order to apply *do-calculus*, we need to know the causal relationships between the variables. Therefore in Step 2 we construct the causal structure with the incorporation of regulatory knowledge, then we identify the miRNA targets using *do-calculus* in Step 3.

Step 1 (Data preparation). The differential expression analysis is performed as described in [42]. As a result for the EMT dataset, 35 miRNA probes and 1154 probes of mRNAs are identified as significantly differentially expressed. For the BRCA dataset, 92 miRNA probes and 1500 mRNA probes are identified. The detailed result can be found in S1 File.

After differential expression analysis, we extract the regulatory knowledge (i.e. TF-miRNA and miRNA-mRNA interactions) relevant to the differentially expressed expression profiles from the regulatory knowledge databases described previously.

Step 2 (Casual structure construction). Using both the gene regulation knowledge and gene expression data, we learn a causal Bayesian network (CBN) which models the structure of the gene regulatory network. A CBN consists of a pair $< \mathcal{G}, P >$, where $\mathcal{G}$ is a directed acyclic graph with the differentially expressed miRNAs and mRNAs as its vertices, and P is the joint probability function of the vertices. An edge in $\mathcal{G}$ indicates a causal relationship between the two vertices. For example, an edge directing from a miRNA to a mRNA means that the miRNA regulates the mRNA; and an edge directing from a TF coding mRNA to a miRNA indicates the TF regulates the miRNA.

A common way to learn the causal structure is to start from a completed graph, then update the graph according to the gene expression data. In order to integrate the regulatory knowledge, in CIDER we label all the edges given in the regulatory knowledge as *constant edges*,

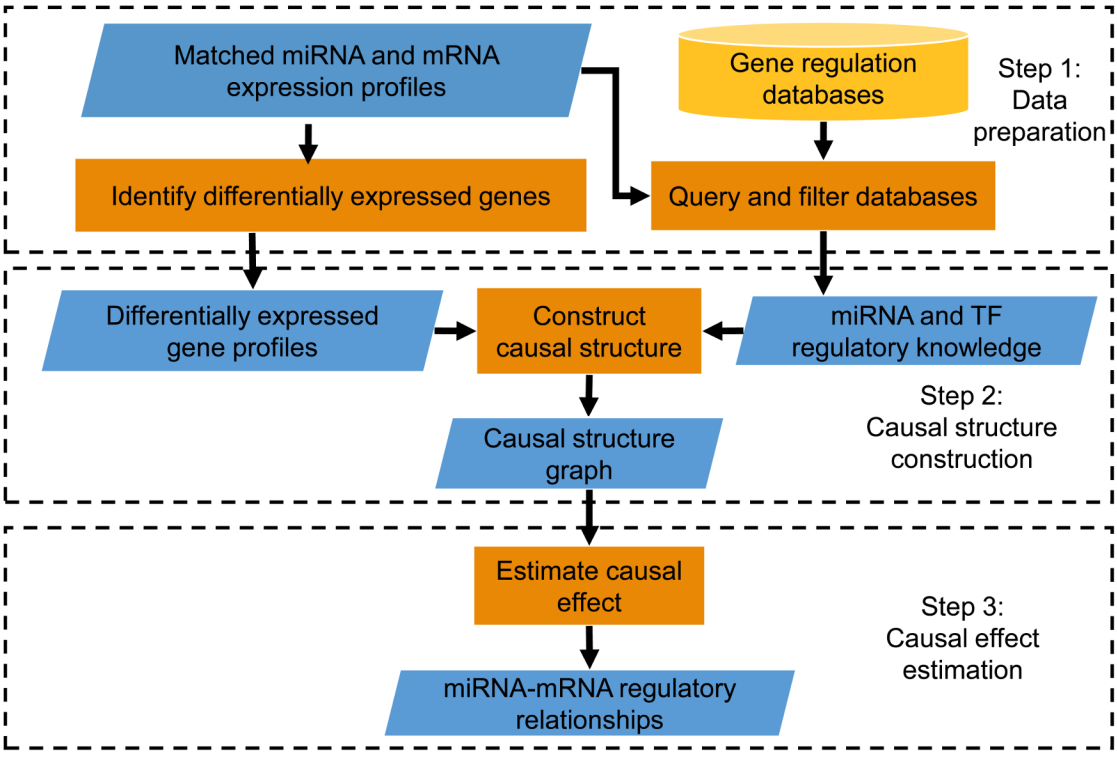

Fig 2. The proposed CIDER framework. First the differentially expressed miRNAs and mRNAs are selected in the expression profiles [40], then we query the regulatory databases for gene regulation knowledge. After that we build the causal structure according to the expression profiles and the knowledge, followed by the causal inference to identify miRNA-mRNA interaction pairs.

doi:10.1371/journal.pone.0152860.g002

which are never to be removed or altered (in terms of their directions) during the entire structure construction step.

Step 3 (Causal effect estimation). We estimate the causal effect that each miRNA has on all the mRNAs according to the causal structure and expression profiles. The causal effects measures when the expression level of a certain miRNA changes, how the expression level of other mRNA will change. For each miRNA, we choose the mRNAs with the largest causal effects as the predicted targets.

In the rest of this section, we discuss the details and intuitions of Step 2 and Step 3.

Causal structure construction

There are two steps involved in constructing a causal structure: determining the existence of edges between the nodes, and orienting the direction of the edges.

A common way [43, 44] to determine whether an edge exists between two nodes is conditional independence (CI) tests. More specifically, starting from a fully connected graph, we use CI tests to determine the dependency between all connected nodes pairs. If two nodes become independent when conditioned on any subsets of their neighbours, the edge between them is removed from the graph. Otherwise, the edge will remain in the causal structure.

During this procedure, edges may be incorrectly removed or maintained. Because the number of available samples is limited when comparing to the large number of variables in expression profiles, CI tests may declare two nodes are independent even if a dependency exists, thus the edge between them will be removed correctly. Furthermore, the incorrectly removed edges will not appear in the conditioning sets of later CI tests, which may lead to false positives (i.e.

two nodes would have been tested to be independent and their edge would have been removed if the incorrectly removed edges were kept and were in the conditioning set of the CI test).

In order to determine the orientation of edges we need to identify the *v*-structures in the causal structure defined as follows:

Definition 1 ([18]) *A triple* (X_i, X_j, X_k) *forms a v-structure in graph* $\mathcal{G}$ *if and only if it suffices both of the following conditions*:

1. X_i *and* X_j *as well as* X_j *and* X_k *are adjacent,* X_i *and* X_k *are not adjacent,*
2. X_i *and* X_k *are not independent when conditioned on* X_j.

When a *v*-structure (X_i, X_j, X_k) is identified, the edges can then be oriented as $X_i \rightarrow X_j \leftarrow X_k$ [18]. After all the *v*-structures have been identified and oriented, we can orient the remaining edges according to the principle of avoiding the creation of cycles and new *v*-structures [44].

Unfortunately, under most circumstances the above strategy can only orient some of the edges, leaving many undirected. Undirected edges introduce uncertainty in the next step, since the estimation has to be done on all possible orientations of the undirected edges and take the lower bounds as the inferred causal effects [31].

In our framework, we utilise regulatory knowledge to alleviate both the false edges and the undirected edges problems. We introduce the concept of *constant edge*. A *constant edge* is an edge between two nodes where their relationship are already validated via biological experiments, so the edge will never be removed no matter what result the CI tests are, and the direction of the edge can be correctly determined according to the knowledge. Now let us have a look at benefits of introducing constant edges with the following example.

With the introduction of constant edges, we are able to recover incorrectly removed edges and also remove some falsely discovered edges. Fig 3A shows a causal structure learned with CI tests only, which includes one falsely identified regulatory relationship (miR-200a to miR-200b) and two missed regulatory relationships (miR-200a to ZEB1 and miR-200b to ZEB1). Since it is has been experimentally confirmed that ZEB1 is a target of miR-200a, we mark the edge from miR-200a to ZEB1 as a constant edge and do not remove it when using CI tests (see Fig 3B). Because of the introduction of the edge from miR-200a to ZEB1, the falsely discovered edge from miR-200a to miR-200b is removed (see Fig 3C) as the result of the conditional independence test with ZEB1 being added to the conditioning set.

Constant edges can also help to orient more undirected edges. For example, in Fig 3C, although we have removed the false edge between miR-200a and miR-200b, the directions of the two edges (miR-200b/QKI and miR-429/ZEB1) still cannot be determined. However, when we have another constant edge that miR-200b regulates ZEB1 from the regulatory knowledge, we can orient the two edges as in Fig 3D otherwise a new v-structure (at ZEB1) or a cycle (miR-200b → ZEB1 → miR429 → QKI → miR-200b) will be introduced, either of which is not allowed acyclic assumption [43].

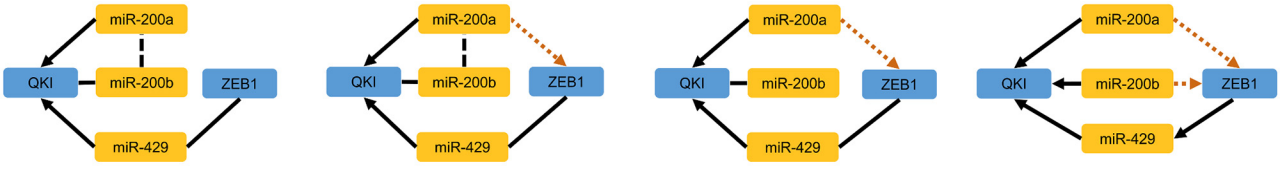

Fig 3. An illustration of how the prior knowledge helping the causal structure construction. Solid/dashed black lines indicate the edges correctly/incorrectly detected during the causal structure construction without the prior knowledge; Dotted brown lines indicate the edges added based on prior knowledge.

doi:10.1371/journal.pone.0152860.g003

As shown above, even when only one or two constant edge is introduced, the uncertainness in the causal structure can be significantly reduced. We briefly summarise the procedure of constructing the causal structure in Algorithm 1 (The details of the algorithm can be found in S6 File).

Algorithm 1 Construct the causal structure $\mathcal{G}$

Input: Gene expression profile, regulatory knowledge matrix.
Output: Constructed causal structure $\mathcal{G}$
  Initiate $\mathcal{G}$ as a fully connected graph
  //Mark constant edges
  Mark all constant edges in $\mathcal{G}$ according to the regulatory knowledge matrix.
  //Removes edges from $\mathcal{G}$ using CI tests
  Test conditional dependence among non-constant edges, remove an edge between two vertices if they are found independent.
  //Orient constant edges
  Orient constant edges according to regulatroy knowledge
  //Orient remaining edges
  Identify and orient all *v*-structures
  Orient remaining edges without creating new v-structure and cycle
return $\mathcal{G}$

Causal Effect Estimation

With the expression data and the causal structure among its variables, we need to infer the causal effects that a miRNA has on a mRNA. By assuming all variables in the expression profiles follow the multivariate Gaussian distribution, we can calculate the causal effects as follows:

Theorem 1 ([45]) *Let $X_1, \ldots, X_p, X_{p+1}, \ldots, X_{p+q}$ be jointly normal distributed. The causal effect of $X_i (i = 1, \ldots, p)$ on $X_j (j = p+1, \ldots, p+q)$, $ce(X_i, X_j)$ can be calculated as*:

$$ce(X_i, X_j) = \beta_{ij|pa_j} = \begin{cases} 0 & X_j \in pa_i \\ \beta_{ij} \text{ in } X_j \sim \beta_{ij} X_i + pa_j, & X_j \notin pa_i \end{cases} \quad (1)$$

where $X_j \sim \beta_{ij} X_i + pa_j$ is the shorthand for the linear regression of X_j on X_i and pa_j, and β_{ij} is the coefficient for X_i in the regression.

Given the above theorem, we are able to estimate the regulatory effect of each miRNA on all mRNAs in a dataset, and use the mRNAs with top ranked causal effects as the targets of the corresponding miRNA. Note that because the available regulatory knowledge is very sparse, some edges in the causal structure may still remain undirected. Therefore we use the minimum absolute value as the estimation of the lower bound of the causal effect. We briefly summarise this procedure in Algorithm 2. For more details, please refer to the S6 File.

Algorithm 2 Causal effects estimation

Input: Gene expression data $\mathbf{X}_{s \times n}$, causal structure $\mathcal{G}$.
Output: Causal effects matrix C where $C(i, j)$ is the causal effect of miRNA_i on mRNA_j.
  Initialize C as a zero matrix
  for All pairs of miRNA_i and mRNA_j **do**
    for All possible orientations of $\mathcal{G}$ **do**
      Calculate the causal effect with Theorem 1
    end for
    Let $C(i, j)$ be the causal effect with lowest absolute value
  end for
  return C

Evaluation methods

Evaluating miRNA target prediction methods is not an easy task. This is mainly because the current understanding of miRNA regulation mechanisms is still limited and experimentally validated target databases only contain information about frequently studied miRNAs. Therefore to evaluate the effectiveness of the CIDER framework, we use a number of different evaluation approaches described in the following:

1. We compare the predicted results to wet-lab validated miRNA target databases. Since CIDER needs access to regulatory knowledge, we reserve a part of the known regulatory relationships as the ground truth for evaluation. Specifically, when studying the performance of CIDER using TF-miRNA regulatory knowledge, we utilise the TF-miRNA interactions retrieved from TransmiR as the prior knowledge in constructing the causal structure and reserve the miRNA-mRNA interactions obtained from the miRNA target databases as the ground truth; when studying the effect of miRNA-mRNA regulatory knowledge, we utilise miRNA-mRNA interactions retrieved from TargetScan for causal structure construction and reserve the miRNA-mRNA interactions obtained from the experimentally validated miRNA target databases as the ground truth. In addition, if an interaction appears in the prior knowledge and the ground truth, we remove this entry from the knowledge and only use it for evaluation.

2. We compare the predicted targets to the results of miRNA transfection experiments. miRNA transfection is a technique that actively transfects a particular miRNA into cells, and by comparing the transfected expression profile to the controlled sample (same cell but without miRNA transfection), difference in mRNA expression level can be measured and mRNAs with top ranked logarithm fold change values can be considered as groundtruth miRNA targets [46].

3. We use gene pathway enrichment tools to analyse the functionality of predicted miRNA targets. It is often hypothesized that the predicted miRNA targets based on the expression profile should be closely related to the biological condition of the expression profiles. For example, the mRNAs targeted by miRNAs in the EMT dataset should be closely related to the epithelial to mesenchymal transition process. Therefore pathway functional analysis can be used to demonstrate the effectiveness of miRNA target prediction methods.

The above evaluations are used to demonstrate the effectiveness of CIDER for finding biologically relevant miRNA targets. To further demonstrate the performance CIDER when used with different amount of regulatory knowledge, we use the following simulation.

We simulate a gene regulatory networks and the corresponding gene expression profiles based on the linear structural equation model [47]. First we construct a directed graph where each node represents a miRNA or mRNA (including TF coding mRNA) in the regulatory network and the direction of an edge indicates that the parent node regulates the child node. Then we assign to each edge a weight w_i ($w_i \sim \mathcal{U}([-1,-0.1] \cup [0.1,1])$) which measures the amount of regulatory effect that the parent node has on the child node. Starting from the nodes without parents, we generate the expression value for each node following Gaussian distribution, with a non-Gaussian error terms added. Specifically the expression value of each gene is defined as follows:

$$x_i = b_i + \sum_{j \in pa(x_i)} w_j \cdot x_j + \epsilon_i, \tag{2}$$

where $pa(x_i)$ denotes the parent nodes of x_i, $w_j \cdot x_j$ is the regulatory effect of the j-th node has

on the i-th one, ϵ_i represents the non-Gaussian error term of the i-th node, and b_i represents the interception term. To alleviate the effect of randomness in the simulated data, in total 50 networks (each of the network has approximately 1000 nodes) are generated and the average results from these 50 networks are reported. For each network we generate two sets of expression profiles, containing 250 and 500 samples, respectively.

To evaluate the performance on simulated datasets, we use F-Score (the harmonic mean of precision and recall) to measure the performance of all methods, which is formulated as follows:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

We use F-Score to compare CIDER with a variety of popular miRNA target prediction methods, including Pearson correlation [42], Lasso [48], Z-Score [49]. Pearson correlation calculates the correlation coefficients between pairs of miRNAs and mRNAs, and use the strength of the correlations to measure the regulatory effect. Lasso is a popular regression method which also measures linear correlation, but uses the L1-norm to overcome the sparseness of the high dimensional expression profiles. Z-Score is a specifically designed method to infer gene regulatory network using data from gene knock-out experiments. Since only observational data is used in our study, we use the lowest expression value of each gene among all sample as the value of knocked-out gene expression.

Results and Discussions

Transcriptional knowledge improves miRNA-mRNA target prediction

In this section, we investigate the effect of transcriptional TF-miRNA regulatory knowledge on miRNA target prediction. We first apply CIDER to analyse only the expression profiles, then we allow CIDER to access both TF-miRNA regulatory knowledge and the expression data and compare the performance of these two settings. For each miRNA, we consider the mRNAs with Top 50 and Top 100 ranked causal effects as its targets and compare them with those in the combination of three experimentally confirmed miRNA-mRNA interaction databases: Tarbase, miRWalk and miRTarbase.

Although for both datasets only less than 20 of TF-miRNA interactions are integrated (the total number of possible edges is around 10^6), it is evident to see the benefit of TF-miRNA knowledge for predicting miRNA targets. As shown in Fig 4, with the help of TF-miRNA regulation knowledge, CIDER predicts more validated miRNA targets than using expression profiles alone.

Fig 5 illustrates a comparison of the miRNA targets predicted by CIDER with and without TF-miRNA knowledge from both datasets. For example, without the TF-miRNA knowledge of BMP2→miR-31, only three predicted targets of miR-31 agrees with the experimentally validated database. However, when the TF-miRNA regulation between BMP2 is incorporated, CIDER not only successfully uncovers the up-regulation effect between BMP2 and miR-31, but also identifies 9 experimentally validated targets.

We conduct pathway enrichment analysis of the predicted target genes with the focus on KEGG pathways (adjusted p-value$<$0.05). To determine whether the top predicted miRNA targets are related to respective biological processes (EMT and BRCA), we select the top 5 predicted targets for each miRNA. As shown in Table 1, the KEGG pathways are highly associated with the relevant biological process. For instance, epithelial tight junctions are closely related to EMT process and focal adhesion is shown to be related to breast cancer in previous research [50].

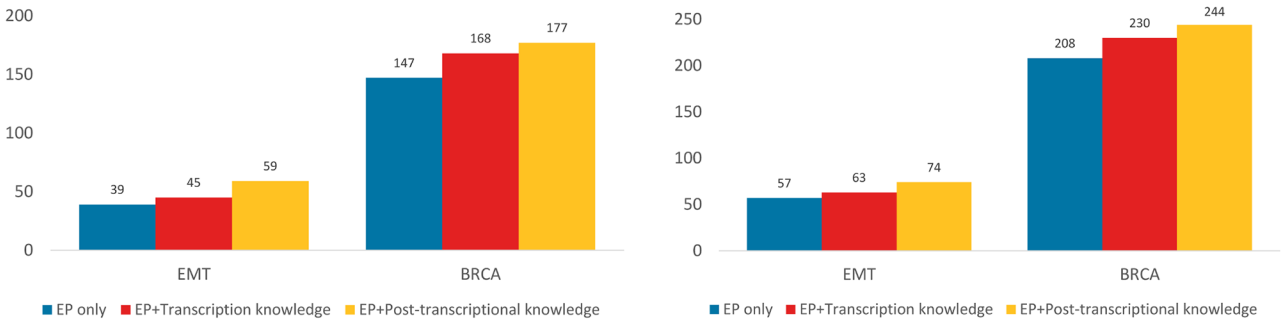

Fig 4. Number of experimentally validated miRNA targets (total number for all miRNAs) identified by CIDER when utilizing expression profiles (EP) only, EP + transcriptional regulatory knowledge, EP + post-transcriptional knowledge. (Left) Results for Top 100 predicted targets for each miRNA. (Right) Results for Top 150 predicted targets.

doi:10.1371/journal.pone.0152860.g004

Post-transcriptional knowledge improves miRNA target prediction

In this section we show that post-transcriptional miRNA-mRNA knowledge improves the performance of CIDER. Similar to the previous section, we first apply CIDER to analyse the expression profiles alone, then compare it to the results obtained by allowing CIDER to access both the regulatory knowledge and the expression profiles.

Since we need to keep the experimentally validated target databases to evaluate the performance, miRNA-mRNA regulatory relationships predicted by TargetScan are used as the regulatory knowledge.

We depict the number of experimentally validated miRNA targets found by CIDER using expression profiles only and using both post-transcriptional regulatory knowledge and expression profiles in Fig 4. CIDER is able to successfully utilise the post-transcriptional knowledge and find significantly more validated targets than using expression profiles alone, despite that the regulatory knowledge in TargetScan contains false positives. The results not only

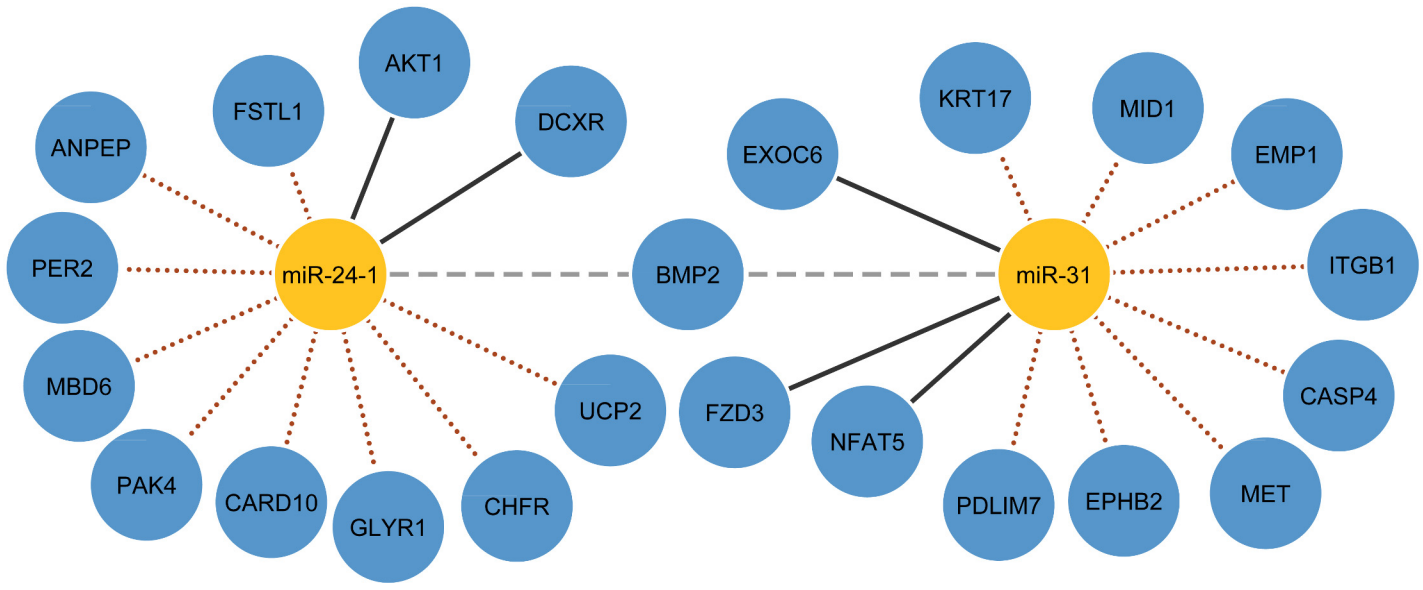

Fig 5. Comparison of miRNA targets identified by CIDER with and without TF-miRNA regulatory knowledge. Gray dashed lines indicate the TF-miRNA regulatory knowledge introduced from TransmiR. Black solid lines indicate miRNA-mRNA regulations found without knowledge. Brown dotted lines represent the additional miRNA-mRNA regulations found when TF-miRNA knowledge is utilised.

doi:10.1371/journal.pone.0152860.g005

Table 1. Top 10 enchriment KEGG pathways in the EMT and BRCA datasets. The p-values have been obtained through Hypergeometric analysis corrected by FDR method.

Datasets	Top 10 enrichment KEGG pathways	Adj-p-value
EMT	Epithelial tight junctions	5.95e-06
	Leukocyte transendothelial migration	1.82e-05
	Cell adhesion molecules	2.38e-04
	Arrhythmogenic right ventricular cardiomyopathy	3.23e-04
	Cell adhesion molecules	2.06e-03
	Melanogenesis	8.40e-03
	Regulation of actin cytoskeleton	9.74e-03
	Huntington's disease	3.30e-02
	Pathways in cancer	1.07e-02
	Amoebiasis	1.07e-02
BRCA	Pancreatic secretion	1.20e-03
	Leukocyte transendothelial migration	1.83e-03
	Focal adhesion	2.32e-03
	Amoebiasis	4.94e-03
	Purine metabolism	5.19e-03
	Regulation of actin cytoskeleton	5.30e-03
	Salivary secretion	5.58e-03
	Adherens junction	5.58e-03
	Pathways in cancer	6.03e-03
	Tight junction	6.09e-03

doi:10.1371/journal.pone.0152860.t001

demonstrate that CIDER is able to utilise post-transcriptional regulatory knowledge, but also indicate that CIDER can benefit from sequence-based prediction knowledge with false positives.

The reason behind the robustness of CIDER lies in the causal inference step. There the causal structure and expression profiles are analysed together to infer the amount of causal effects. If the false edges between miRNAs and mRNAs are not supported by the inference results, the noise introduced from false positive regulatory knowledge will be mitigated by the causal inference step.

When accessing all the experimentally validated miRNA target databases together with expression profiles, CIDER discovers more targets than accessing expression profiles alone. Since we use the databases as knowledge, other means are needed for evaluation. Therefore we compare the predicted targets for the EMT dataset to the transfection experiment on the MDA-MB-231 human cell line [41]. In this experiment, the gene expression level in the MDA-MB-231 samples transfected with hsa-miR-200a-3p/hsa-miR-200b-3p along with the expression level in those samples without hsa-miR-200a-3p and hsa-miR-200b-3p (control) were measured. (Please refer to S3 File for the detailed transfection experiment results). The differentially expressed genes from the controlled and transfected samples are used to validate the our computational predictions. Specifically, 345 and 533 genes are identified to be regulated by hsa-miR-200a-3p and hsa-miR-200b-3p, respectively.

The results demonstrate that with the help of post-transcriptional regulatory knowledge, CIDER identifies significantly more validated miRNA targets comparing to the miRNA targets predicted based only on expression profiles. Fig 6 shows that when equipped with the post-transcriptional miRNA-mRNA regulatory knowledge (brown dotted lines), CIDER is able to discover many novel miRNA-mRNA regulatory relationships that are missed by using expression data alone.

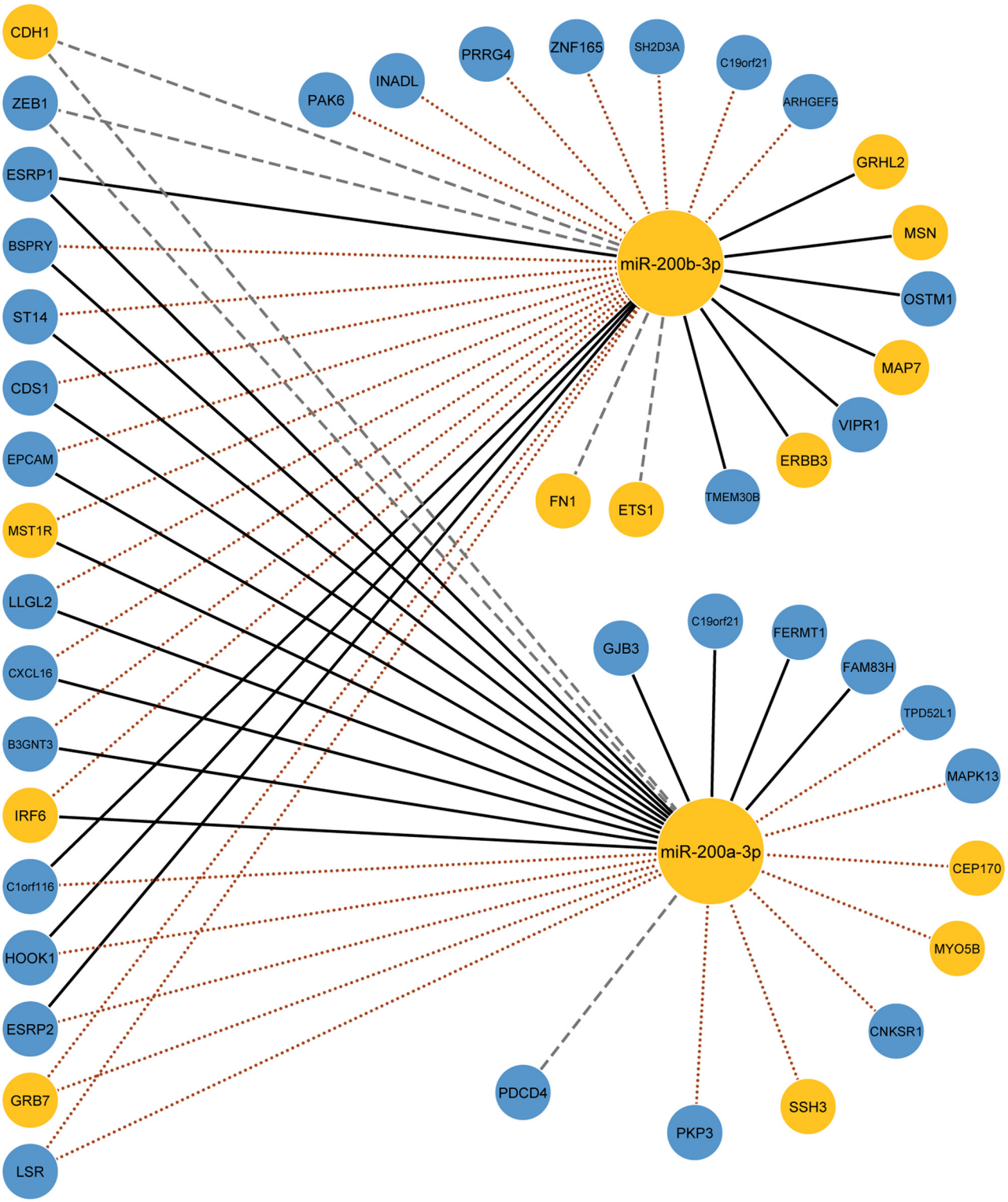

Fig 6. Comparison of validated regulatory relationships with/without regulatory knowledge on the EMT dataset. Black solid lines indicate validated interactions found with expression profiles; grey dashed lines indicate interactions provided by the regulatory knowledge; brown dotted lines indicate new interactions discovered by CIDER utilizing both expression profiles and regulatory knowledge, yellow shaded nodes are known oncogenes and oncomiRs according to [51].

doi:10.1371/journal.pone.0152860.g006

More prior knowledge leads to better predictions

It is important to know that how the framework works with different amounts and types of regulatory knowledge. In this section we study the performance of CIDER when utilizing different amounts and types of knowledge. Since currently the wet-lab validated knowledge is very sparse, we generate the simulated networks and expression profiles as described in the Evaluation Methods section for our analysis.

Even without knowledge, CIDER achieves comparable performance of state-of-the-art miRNA target prediction methods. As shown in Fig 7, when only utilizing the expression data, the performance of CIDER without prior knowledge is much better than Z-Score. Lasso and Pearson show similar performance regardless of the sparsity constraint added in Lasso. When comparing CIDER with Pearson and Lasso, even without using regulatory knowledge, CIDER shows slightly better performance than both methods because of CIDER utilised causation instead of correlation.

The performance of CIDER increases monotonically with the amount of knowledge. Combining post-transcriptional and transcriptional knowledge significantly boosts the performance of CIDER. To demonstrate this, we evaluate CIDER with three types of knowledge: miRNA-mRNA interactions, TF-miRNA interactions and the combination of these two. For each type of regulatory knowledge, starting from expression data only, we gradually increase the amount of knowledge available to CIDER from 0% to 50% (of the total amount of available knowledge of the type) by a 5% interval. As shown in Fig 8, both transcriptional and post-transcriptional knowledge separately improves the performance of CIDER significantly, and the combined knowledge leads to further improvement. For every type of regulatory knowledge, as the amount of utilised knowledge increases the performance of CIDER improves monotonically. With 50% of the combined knowledge, CIDER achieves very high accuracy.

In summary, CIDER is not only able to utilise either transcriptional or post-transcriptional regulatory knowledge to improve the performance of miRNA target prediction, but also able to utilise the combination of the two types of regulatory knowledge to further increase prediction accuracy. As the amount of regulatory knowledge increases, the performance of CIDER continuously improves. With this monotonic improvement, the miRNA target predicted by CIDER will become more accurate and reliable when our understanding of miRNA regulation improves and more knowledge is available for CIDER.

In return, CIDER can provide more precise guidance for selecting miRNA targets for wet-lab validation. Iteratively, as shown in Fig 1, CIDER will help to build a more and more complete gene regulation network.

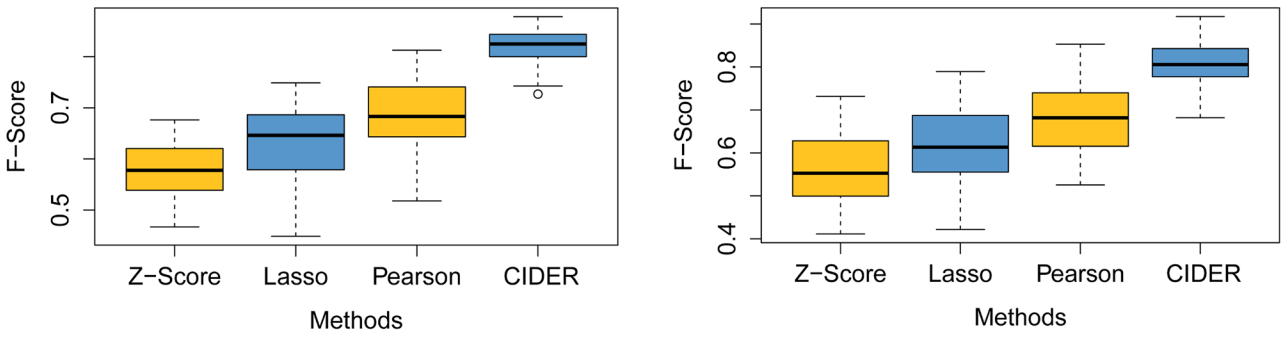

Fig 7. Comparing CIDER with Pearson, Lasso and Z-Score when only accessing expression profiles. Left: 250 samples; Right: 500 samples.

doi:10.1371/journal.pone.0152860.g007

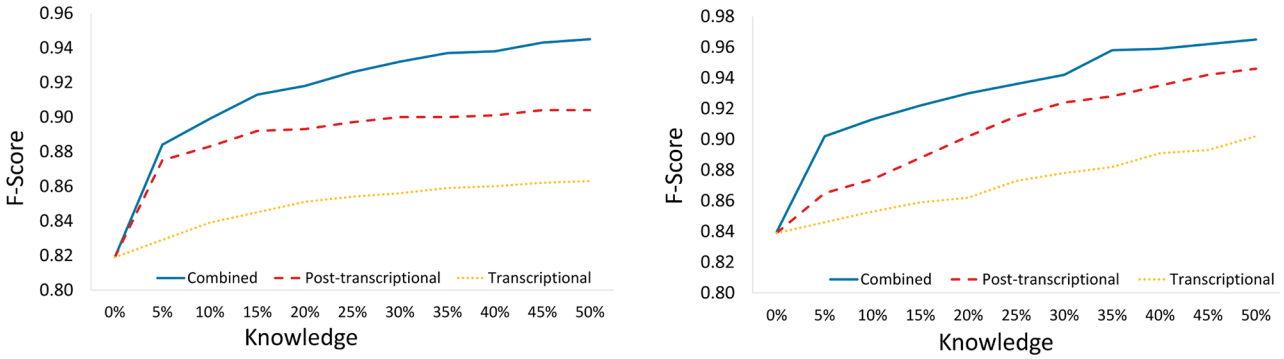

Fig 8. Performance of CIDER when utilizing different amounts and types of regulation knowledge. Sample size: 250 (left), 500 (right).

doi:10.1371/journal.pone.0152860.g008

Methods utilizing sequence bindings information are not suitable for integrating experimentally validated knowledge

Methods designed to utilise sequence based predictions are not suitable for utilizing validated regulatory knowledge. In this section we compare CIDER with ProMISe [30], a recently proposed method designed to utilise sequence binding information and expression profiles.

We compare two algorithms on the EMT and BRCA datasets. Both algorithms have access to the expression profiles, and exactly the same amount of regulatory knowledge, which contains the sequence binding interactions predicted by TargetScan, experimentally validated post-transcriptional knowledge in miRWalk and miRTarbase. Specifically, ProMISe uses the knowledge as sequence binding information, while CIDER uses it to initialise constant edges.

As can be seen in Fig 9, regardless of what threshold is selected for the miRNA targets, CIDER discovers more validated target than ProMISe. This results indicate that the top miRNA targets predicted by CIDER are consistently better than the ones predicted by ProMISe.

The reason is that instead of considering all possible miRNAs and mRNA pairs, ProMISe (and other similar algorithms) uses sequencing information to constrain their search space. In other words, a miRNA-mRNA interaction would not be considered unless the pair is included in the knowledge. Therefore when utilizing sequencing information, these algorithms will be misled by the false negatives; when utilizing experimentally validated knowledge, they will only predict interactions that are already included in the knowledge.

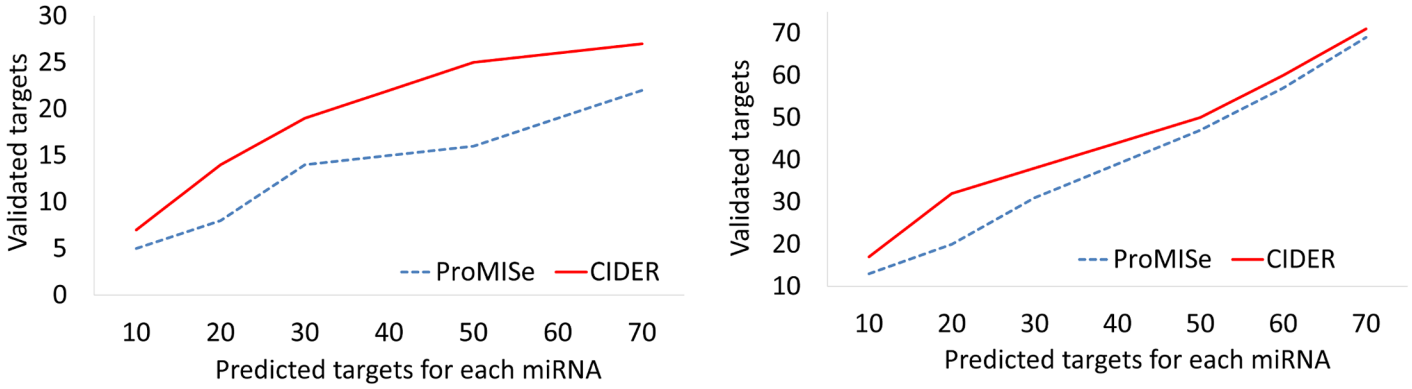

Fig 9. Performance comparison of CIDER and ProMISe when utilizing post-transcriptional regulation knowledge. Left: EMT dataset, right: 500 BRCA dataset.

doi:10.1371/journal.pone.0152860.g009

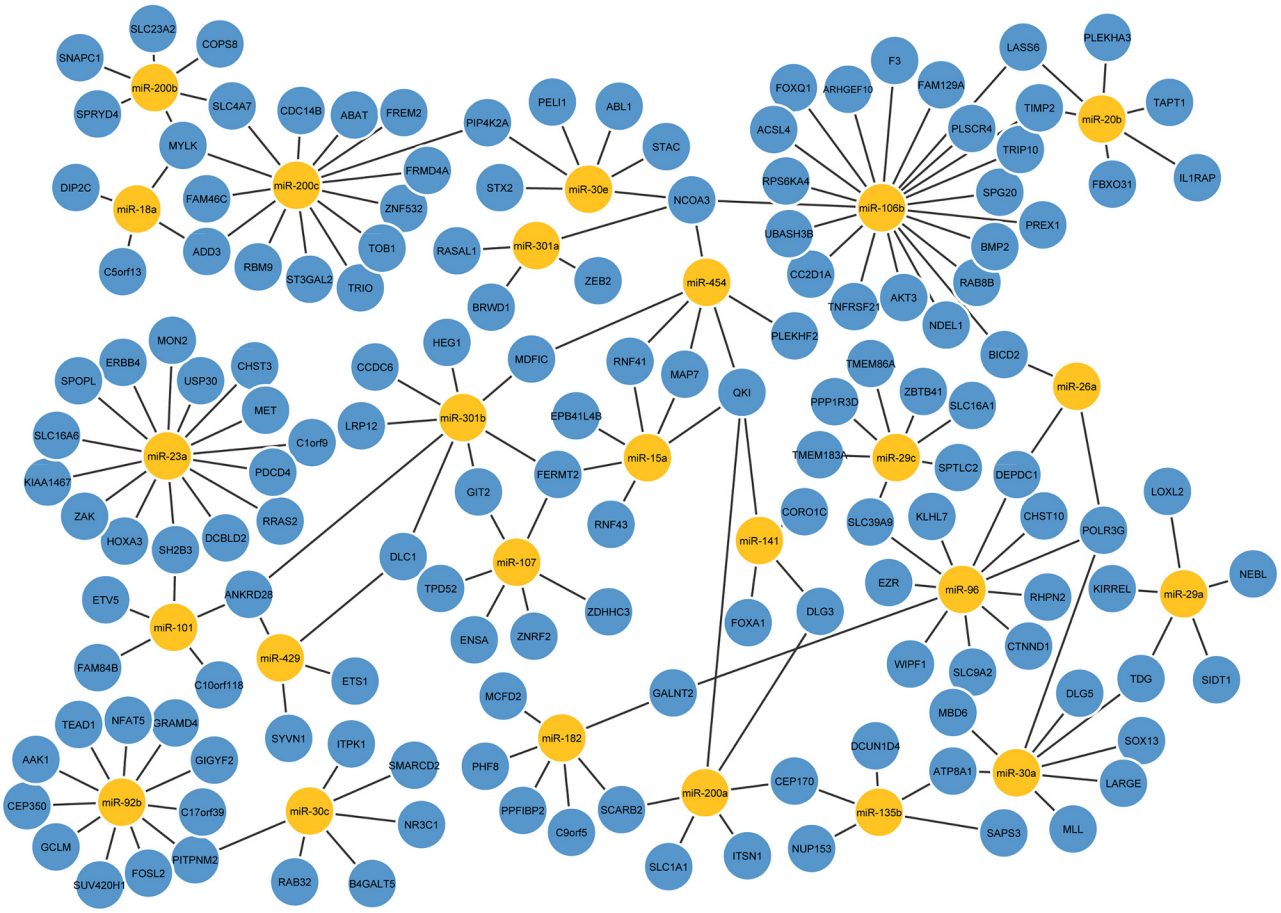

Fig 10. High confidence miRNA targets predicted by CIDER utilizing expression profiles, transcriptional and post-transcriptional knowledge. Only part of the interactions are shown for clarity of illustration, please refer to S5 File for the full results.

doi:10.1371/journal.pone.0152860.g010

Putative miRNA targets

In this section, we report the high-confidence miRNA targets predicted by CIDER in the EMT and BRCA datasets for biological researchers to explore. These predictions utilise expression profiles with both transcriptional and post-transcriptional regulatory knowledge. As we have shown in the previous section, CIDER performs better when utilizing the combined knowledge than using either type of regulatory knowledge separately. Therefore, we expect that the miRNA targets predicted by CIDER utilizing TF-miRNA interactions from TransmiR and miRNA-mRNA knowledge from Tarbase, miRTarbase, miRWalk, should provide valuable putative candidates for further biological wet-lab evaluation. To utilise sequence binding information to increase the confidence of the predicted targets, we intersect our discovery with miRNA target prediction from TargetScan.

These high-confidence predicted miRNA targets are presented in Fig 10, and we hope that a significant number of them will be validated by experiments in the future.

Conclusion

The future of biology is neither based on wet-lab experiments nor computational predictions alone, but on their combination. The progress of wet-lab experiments would be hampered

without the help of quality computational predictions, and the power of computational methods would be limited if accumulated biological knowledge were not integrated with the modeling process.

In this article, we present the CIDER framework that seamlessly integrates biological knowledge with high-throughput expression profiles for miRNA target prediction. We use a causal Bayesian network based method to explicitly exploit experimentally validated gene regulatory knowledge to improve the prediction of miRNA-mRNA interactions. Our results demonstrate that when utilizing transcriptional or post-transcriptional knowledge, CIDER discovers significantly more validated miRNA targets than using expression profile alone. Furthermore, when the amount of available regulatory knowledge increases, the performance of CIDER increases monotonically.

With the capability to improve prediction accuracy with the increment of gene regulatory knowledge, our causal discovery framework can serve as a promising tool for uncovering new biological insights using ever increasing regulatory knowledge and new high-throughput data.

Supporting Information

S1 File. Differential expression profiles of miRNAs and mRNAs for the EMT and BRCA datasets. The p-values are adjusted by Benjamini-Hochberg (BH) method.
(XLSX)

S2 File. R source code for the proposed CIDER framework.
(ZIP)

S3 File. Experimentally validated miRNA-mRNA regulatory knowledge. This file includes the miRNA-mRNA regulatory knowledge obtained from the following databases: TarBase, miRecords, miRWalk and miRTarBase.
(XLSX)

S4 File. miRNA transfection result on MDA-MB-231 samples. This file includes the transfection results for hsa-miR-200a and hsa-miR-200b, and control sample.
(XLS)

S5 File. High-confidence miRNA targets predicted by CIDER. This file includes the miRNA targets predicted by CIDER when utilizing post-transcriptional and transcriptional regulatory knowledge and expression profiles, these interactions are also predicted by TargetScan v7.0.
(XLSX)

S6 File. Detailed descriptions of Algorithm 1 and Algorithm 2, and additional validation results.
(PDF)

Acknowledgments

Funding This work is supported by Australian Research Council Discovery Project DP130104090 (in part).

Author Contributions

Conceived and designed the experiments: WZ TDL LL JL ZZ. Performed the experiments: WZ TDL. Analyzed the data: WZ TDL LL JL. Contributed reagents/materials/analysis tools: WZ TDL. Wrote the paper: WZ TDL LL JL ZZ.

References

1. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. Nature Reviews Cancer. 2006; 6(4):259–269. doi: 10.1038/nrc1840 PMID: 16557279
2. Jin P, Zarnescu DC, Ceman S, Nakamoto M, Mowrey J, Jongens TA, et al. Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. Nature Neuroscience. 2004; 7(113):113–117. doi: 10.1038/nn1174 PMID: 14703574
3. Xu C, Lu Y, Pan Z, Chu W, Luo X, Lin H, et al. The muscle-specific microRNAs miR-1 andmiR-133 produce opposing effects on apoptosis by targeting HSP60, HSP70 and caspase-9 in cardiomyocytes. Journal of Cell Science. 2007; 120(Pt 17):3045–3052. doi: 10.1242/jcs.010728 PMID: 17715156
4. Cui Q, Yu Z, Purisima EO, Wang E. Principles of microRNA regulation of a human cellular signaling network. Molecuar Systems Biology. 2006; 2(46).
5. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, et al. An analysis of human microRNA and Disease Associations. PLoS One. 2008; 3(10):e3420. doi: 10.1371/journal.pone.0003420 PMID: 18923704
6. Moqadam FA, Pieters R, den Boer M. The hunting of targets: challenges in miRNA targets. Leukemia. 2013; 27:16–23. doi: 10.1038/leu.2012.179
7. Liu B, Li J, Cairns MJ. Identifying miRNA, targets and functions. Briefings in Bioinformatics. 2012; 15 (1):1–19. doi: 10.1093/bib/bbs075 PMID: 23175680
8. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010; 466:835–840. doi: 10.1038/nature09267 PMID: 20703300
9. Mercatelli N, Coppola V, Bonci D, Miele F, Constantini A, Guadagnoli M. The inhibition of the highly expressed miR-221 and miR-222 impairs the growth of prostate carcinoma xenografts in mice. PLoS One. 2008; 3:4029. doi: 10.1371/journal.pone.0004029
10. Huang GT, Athanassiou C, Benos PV. mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Research. 2011; 39:W416–W423.
11. Liu B, Li J, Tsykin A, Liu L, Gaur AB, Goodall GJ. Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy. BMC Bioinformatics. 2009; 10(408):1–19.
12. Fisher RA. The Design of Experiments. Mamillan Publishers; 1971.
13. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Research. 2014; 42(D1):D78–D85. doi: 10.1093/nar/gkt1266 PMID: 24304892
14. Service RF. Biology's dry future. Science. 2013; 342:186–189. doi: 10.1126/science.342.6155.186 PMID: 24115420
15. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nature Methods. 2007; 4(12):1045–1049. doi: 10.1038/nmeth1130 PMID: 18026111
16. Bie TD, Tranchevent LC, van Oeffelen LMM, Moreau Y. Kernel-based data fusion for gene prioritization. Bioinformatics. 2007; 23:125–232. doi: 10.1093/bioinformatics/btm187
17. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology. 2000; 7:601–620. doi: 10.1089/106652700750050961 PMID: 11108481
18. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press; 2009.
19. Lauritzen SL. Graphical Models. Oxford University Press; 1996.
20. Kalisch M, Bühlmann P. Estimating High-Dimensional Directed Acylic Graphs with the PC-Algorithm. Journal of Machine Learning Research. 2007; 8:613–636.
21. Chickering DM, Heckerman D, Meek C. Large-sample learning of bayesian networks is NP-hard. Journal of Machine Learning Research. 2004; 5:1287–1330.
22. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning. 1995; 20:197–243. doi: 10.1007/BF00994016
23. Tai F, Pan W. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics. 2007; 23(14):1775–1782. doi: 10.1093/bioinformatics/btm234 PMID: 17483507
24. Tian Z, Hwang T, Kuang R. A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. Bioinformatics. 2009; 25(21):2831–2838. doi: 10.1093/bioinformatics/btp467 PMID: 19648139
25. Zhao Z, Wang J, Liu H, Ye J, Chang Y. Identifying Biologically Relevant genes via Multiple Heterogeneous Data Sources. In: Proceedings of The 14th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (K; 2008. p. 839–847.

26. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research. 2011; 39(suppl1):D152–D157. doi: 10.1093/nar/gkq1027 PMID: 21037258

27. Liu H, Yue D, Zhang L, Chen Y, Gao SJ, Huang Y. A Bayesian approach for identifying miRNA targets by combining sequence prediction and gene expression profiling. BMC Genomics. 2010; 11(Suppl 3): S12. doi: 10.1186/1471-2164-11-S3-S12 PMID: 21143779

28. Wang Z, Xu W, Zhu H, Liu Y. A Bayesian Framework to Improve microRNA Target Prediction by Incorporating External Information. Cancer Information. 2014; 13(Suppl 7):19–25.

29. Le TD, Liu L, Zhang J, Liu B, Li J. From miRNA regulation to miRNA-TF co-regulation: computational approaches and challenges. Briefings in Bioinformatics. 2014; p. 1–22.

30. Li Y, Liang C, Wong KC, Jin K, Zhang Z. Inferring probabilistic miRNA—mRNA interaction signatures in cancers: a role-switch approach. Nucleic Acids Research. 2014; 42(9):e76. doi: 10.1093/nar/gku182 PMID: 24609385

31. Le TD, Liu L, Liu B, Tsykin A, Goodall GJ, Satou K, et al. Inferring microRNA and transcription factor regulatory networks in heterogeneous data. BMC Bioinformatics. 2013; 14(92).

32. Chen CY, Chen ST, Fuh CS, Juan HF, Huang HC. Co-regulation of transcription factors and microRNAs in human transcriptional regulatory network. BMC Bioinformatics. 2011; 12(Suppl 1):S41. doi: 10.1186/1471-2105-12-S1-S41 PMID: 21342573

33. Gregory P, Bert A, Pateson E, Barry S, Farshid ATT, Vadas M, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature Cell Biology. 2008; 10(5):593–601. doi: 10.1038/ncb1722 PMID: 18376396

34. Søkilde R, Kaczkowski B, Podolska A, Cirera S, Gorodkin J, Møller S, et al. Global microRNA Analysis of the NCI-60 Cancer Cell Panel. Molecuar Cancer Therapeutics. 2011; 10(3):375–384. doi: 10.1158/1535-7163.MCT-10-0605

35. Riaz M, van Jaarsveld MT, Hollestelle A, van der Smissen WJP, Heine AA, Boersma AW, et al. miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs. Breast Cancer Research. 2013; 15(R33). doi: 10.1186/bcr3415 PMID: 23601657

36. Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcritrans factor-microRNA regulation database. Nucleic Acids Research. 2010; 38:119–122. doi: 10.1093/nar/gkp803

37. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Research. 2014. doi: 10.1093/nar/gku1215 PMID: 25416803

38. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk—Database: Prediction of possible miRNA binding sites by"walking" the genes of three genomes. Journal of Biomedical Informatics. 2011; 44(5):839–847. doi: 10.1016/j.jbi.2011.05.002 PMID: 21605702

39. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120(1):15–20. doi: 10.1016/j.cell.2004.12.035 PMID: 15652477

40. Smyth GK. Limma: linear models for microarray data. Springer Press; 2005.

41. Le TD, Liu L, Tsykin A, Goodall GJ, Liu B, Sun BY, et al. Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics. 2013; 29(6):765–771. doi: 10.1093/bioinformatics/btt048 PMID: 23365408

42. Le TD, Zhang J, Liu L, Li J. Ensemble Methods for MiRNA Target Prediction from Expression Data. PLoS One. 2015; 10(6):e0131627. doi: 10.1371/journal.pone.0131627 PMID: 26114448

43. Spirtes P, Glymour C., Scheines R. Causation, prediction, and search (Second Edition). The MIT Press; 2000.

44. Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. Journal of Machine Learning Research. 2014; 15:3741–3782.

45. Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. The Annals of Statistics. 2009; 37(6A):3133–3164. doi: 10.1214/09-AOS685

46. Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. Nat Biotechnol. 2009; 27(6):549–555. doi: 10.1038/nbt.1543 PMID: 19465925

47. Chu T, Mouillet JF, Hood BL, Conrads TP, Sadovsky Y. The assembly of miRNA-mRNA-protein regulatory networks using high-throughput expression data. Bioinformatics. 2015; 31(11):1870–1877. doi: 10.1093/bioinformatics/btv038

48. Lu Y, Zhou Y, Qu W, Deng M, Zhang C. A Lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics. 2011; 27(17):2406–2413. doi: 10.1093/bioinformatics/btr410 PMID: 21743061

49. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a rigorous assessment of systems biology models: the dream3 challenges. PloS One. 2010; 5:e9202. doi: 10.1371/journal.pone.0009202 PMID: 20186320

50. Luo M, Guan JL. Focal adhesion kinase: A prominent determinant in breast cancer initiation, progression and metastasis. Cancer Letters. 2010; 289(2):127–139. doi: 10.1016/j.canlet.2009.07.005 PMID: 19643531

51. An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. The Journal of Biological Databases and Curation. 2014. doi: 10.1093/database/bau015 PMID: 24608173