RESEARCH

Systematic exploration of autonomous modules in noisy microRNA-target networks for testing the generality of the ceRNA hypothesis

Danny Kit-Sang Yip[1], Iris K. Pang[2] and Kevin Y. Yip[1,3,4*]

*Correspondence:
kevinyip@cse.cuhk.edu.hk
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
Full list of author information is available at the end of the article

Abstract

Background: In the competing endogenous RNA (ceRNA) hypothesis, different transcripts communicate through a competition for their common targeting microRNAs (miRNAs). Individual examples have clearly shown the functional importance of ceRNA in gene regulation and cancer biology. It remains unclear to what extent gene expression levels are regulated by ceRNA in general. One major hurdle to studying this problem is the intertwined connections in miRNA-target networks, which makes it difficult to isolate the effects of individual miRNAs.

Results: Here we propose computational methods for decomposing a complex miRNA-target network into largely autonomous modules called microRNA-target biclusters (MTBs). Each MTB contains a relatively small number of densely connected miRNAs and mRNAs with few connections to other miRNAs and mRNAs. Each MTB can thus be individually analyzed with minimal crosstalk with other MTBs. Our approach differs from previous methods for finding modules in miRNA-target networks by not making any pre-assumptions about expression patterns, thereby providing objective information for testing the ceRNA hypothesis. We show that the expression levels of miRNAs and mRNAs in an MTB are significantly more anti-correlated than random miRNA-mRNA pairs and other validated and predicted miRNA-target pairs, demonstrating the biological relevance of MTBs. We further show that there is widespread correlation of expression between mRNAs in same MTBs under a wide variety of parameter settings, and the correlation remains even when co-regulatory effects are controlled for, which suggests potential widespread expression buffering between these mRNAs, which is consistent with the ceRNA hypothesis. Lastly, we also propose a potential use of MTBs in functional annotation of miRNAs.

Conclusions: MTBs can be used to help identify autonomous miRNA-target modules for testing the generality of the ceRNA hypothesis experimentally. The identified modules can also be used to test other properties of miRNA-target networks in general.

Keywords: Competing endogeneous RNA; MicroRNA-target bicluster; MicroRNA network

Background

MicroRNAs (miRNAs) are short endogenous RNAs that bind specific sites of messenger RNA (mRNA) targets called miRNA response elements (MREs) with partial or full sequence complementarity. The protein levels of the targets are regulated by the miRNAs through the promotion of RNA degradation or translational repres-

sion [16, 9, 72, 10]. Based on the distribution of MREs on different mRNAs, one miRNA could target multiple mRNAs, and multiple miRNAs could target the same mRNA, leading to a complex network of miRNA-mRNA interactions [26, 39].

While conventionally miRNAs are considered to regulate their mRNA targets, in theory mRNAs could also back-regulate their targeting miRNAs by affecting their availability in binding other mRNAs [28, 64]. If the expression level of an mRNA is increased, more copies of its targeting miRNAs will bind to it and become less available for binding other targets. These other targets will be de-repressed and their expression levels will increase. Similarly, if the expression level of an mRNA is decreased, more copies of its targeting miRNAs will become available. They will bind more to other targets and will decrease their expression levels. As a result, different targets of a miRNA can buffer each other [7, 25] and display a positive correlation of their expression levels [64]. In general, different transcripts (mRNAs and other non-coding RNAs) with MREs of the same miRNA may compete for the finite copies of the miRNA in a cell. This back-regulation mechanism and its *in vivo* functional roles have been coined the competing endogenous RNA (ceRNA) hypothesis [64].

One interesting example that supports the ceRNA hypothesis was found between the tumor suppressor gene PTEN and its pseudogene PTENP1 [61]. The MREs of some miRNAs that target PTEN, including miR-19b and miR-20a, are preserved in the truncated 3' end of the PTENP1 transcript, which allow it to act as a miRNA target decoy for PTEN. Indeed, the expression of both PTEN and PTENP1 was repressed by miR-19b and miR-20a in DU145 prostate cancer cells, and their expression levels exhibited a positive correlation across a large number of normal human tissues and prostate tumor samples. Functionally, PTENP1 was found to have tumor suppressive activity and was selectively lost in human cancer, which suggest a potential role of this pseudogene in the normal functioning of PTEN in tumor suppression. Additional evidence of the functional roles of ceRNA in human cancer was reported in the same study and a series of other studies [61, 17, 43, 66, 68]. Regulatory interactions between mRNAs that share common MREs had also been discovered in plants, a phenomenon known as "target mimicry" [28, 63].

At a more global scale, the idea that miRNA targets buffer each other has been used by a number of methods to study miRNA-target interactions. Some methods identify the subset of computationally predicted miRNA targets with a positive correlation of expression as the more reliable targets [42, 71, 31]. Some methods identify "sponge" modulators of miRNA-target interactions, which are RNAs whose expression is associated with changes in the mutual information between miRNAs and their targets [66]. All these methods assume a certain degree of generality of the ceRNA hypothesis, and require some high-throughput expression data as input.

In contrast, a transcriptome-wide systematic test of the ceRNA hypothesis has been lacking. It has not been certain whether the buffering between miRNA targets is sufficiently strong to be reflected by their expression levels in general. Conceptually, this can be tested by a two-step procedure, namely (1) gathering a list of miRNA-target pairs obtained by a method not considering their expression patterns, and (2) evaluating whether mRNAs targeted by same miRNAs are significantly more correlated in expression than other mRNAs (with proper control for effects due to

co-regulation, as discussed in detail below). While conceptually simple, there are a number of issues that make this kind of analysis practically difficult:

1. Noisy miRNA-target networks: Current computational methods for miRNA target prediction have limited accuracy and consistency [2], while the number of experimentally validated miRNA-target pairs is small [73]. False positives and false negatives in the miRNA target predictions would make it difficult to identify mRNAs with common targeting miRNAs.
2. Unshared targeting miRNAs: mRNAs targeted by a common miRNA may individually be targeted by other unshared miRNAs, which could affect their expression levels separately and lower their correlation.
3. Unshared mRNA targets: miRNAs that target a common set of mRNAs may individually have additional unshared targets, which could dilute the buffering effect of their common set of mRNA targets.
4. Partial effects at transcriptional level: The functional effects of miRNAs on their targets are only partially reflected by mRNA levels, while data about protein abundance are not as widely available.
5. Other gene regulatory mechanisms: Gene expression is regulated by a complex system that involves many other components. Even if two mRNAs are competing for their common targeting miRNAs, their expression levels may not appear correlated if they are individually affected by some other regulatory mechanisms. In addition, a miRNA may affect the expression level of a gene indirectly through its targets that directly or indirectly regulate the gene, leading to expression patterns more difficult to analyze.

In this study, we propose computational methods for studying noisy miRNA-target networks that can overcome the first three issues and tolerate the last two. The main idea is to identify small modules in the networks, which we call microRNA-target biclusters (MTBs), without using any expression data as input. It is a novel concept inspired by the related work on biclustering in the literature of gene expression data analysis [18]. Each MTB consists of a set of miRNAs and a set of mRNAs, where (1) the miRNAs target most of the mRNAs in the MTB but few other mRNAs and (2) the mRNAs are targeted by most of the miRNAs in the MTB but few other miRNAs. Each MTB represents a network module that potentially maintains a largely autonomous regulation sub-system. By tuning the level of interactions linking members of an MTB to non-members, and the level of missing intra-MTB interactions allowed, the impacts of false positives and false negatives in the interaction networks on the MTBs (issue 1) and the degree of independence of each MTB (issues 2 and 3) can be controlled.

To show the biological relevance of MTBs, we analyzed the expression patterns of the miRNAs and mRNAs in our MTBs using RNA-seq data from matched cell lines produced by the ENCODE consortium [23, 69]. We show that the MTBs identified by our methods contain miRNAs more anti-correlated in expression with the mRNAs in the same MTBs than both their other targets and random mRNA sets. As proposed in a series of previous studies, this strong anti-correlation observed indicates that the mRNAs in our MTBs are likely true targets of the miRNAs in the same MTBs [38, 32, 48, 50, 60, 57, 62, 54]. By using formal validation procedures and considering many different experimental settings, we show that our results are

statistically significant and robust. These results suggest that despite the incomplete reflection of the effects of miRNAs at the transcription level (issue 4) and the presence of other transcriptional regulatory mechanisms (issue 5), it is still possible to systematically analyze the effects of miRNAs on their targets using RNA-seq data.

Correspondingly, we observed stronger expression correlations among mRNAs in the same MTBs even if subtle effects due to co-regulation are controlled for. We also found that mRNAs and miRNAs in the same MTBs have related biological functions. Overall, our results suggest widespread expression buffering between mRNAs commonly targeted by the same miRNAs, which is in line with the ceRNA hypothesis.

In the literature of computational analysis of miRNAs, the two main focuses have long been on identifying miRNA-encoding regions from genomes [22, 6, 29, 33] and on predicting the targets of individual miRNAs [56, 70]. In line with the latest trend of studying the inter-related miRNA-target interactions from a network perspective [41, 48, 51, 12, 52, 77, 49], our work introduces a new way to study these miRNA-target networks by decomposing complex networks into simple modules that can be more easily analyzed.

Results and discussion

Defining MTBs and identifying them from miRNA-target networks

We collected computationally predicted human miRNA targets from 5 prediction methods. We combined these predictions to form a high-confidence set and a high-coverage set of miRNA-target predicted interactions, which consist of pairs predicted by at least one prediction method with high and moderate confidence, respectively. We also collected experimentally validated miRNA targets in human from a recent release of TarBase [73]. To study the effects of having validated interactions in these networks on our analyses, for both the high-confidence and high-coverage networks, we further considered either having only the computational predictions, or both the computational predictions and experimentally validated pairs combined, resulting in 4 integrated miRNA-target networks in total (Table 1).

Each of these networks can be represented either by a binary matrix or a bipartite graph (Figure 1). In the matrix representation, each row corresponds to an mRNA and each column corresponds to a miRNA. An element has value 1 if the miRNA represented by the column targets the mRNA represented by the row in the network, and 0 otherwise. In the graphical representation, each node in the first part represents an mRNA and each node in the second part represents a miRNA. There is an edge connecting a miRNA node and an mRNA node if the miRNA targets the mRNA.

An idealized definition of an MTB is a set of miRNAs and mRNAs in which each of these miRNAs targets all these mRNAs but not any other mRNAs, and each of these mRNAs are targeted by all these miRNAs not any other miRNAs (Figure 1, type R). In the matrix representation, it corresponds to a submatrix (i.e., a subset of rows and columns not necessarily adjacent to each other) containing only 1's, with all other elements on these rows and columns having value 0. In the graphical representation, it is a biclique (fully-connected bipartite subgraph) with no extra

edges incident on these nodes. If the miRNA-target network was free of false positive and false negative errors, MTBs of this type would be perfect cases for testing the ceRNA hypothesis since they represent totally autonomous modules isolated from the other parts of the network.

In practice, however, such ideal modules rarely exist in miRNA-target networks. Even if they do exist, they may not be observed in our integrated networks due to possible false positives and false negatives in the networks. We thus defined a number of other MTB types with less stringent requirements, by allowing some missing 1's in the submatrix and/or extra 1's in other elements on the defining rows and columns. We first defined three other types that retain the restrictive (R) requirement that the submatrix should contain all 1's (i.e., the miRNAs in an MTB should target all mRNAs in the MTB), but either only the defining columns are not allowed to have extra 1's (i.e., only the miRNAs (mi) are restricted from having extra interactions), or only the defining rows are not allowed to have extra 1's (i.e., only the mRNAs (m) are restricted from have extra interactions), or the general case (gen) that both are allowed. The corresponding MTB types are denoted as Rmi, Rm and Rgen, respectively. Analogously, we also defined four loose (L) types that allow 0's in the submatrix (i.e., the miRNAs in an MTB are not required to target all mRNAs in the MTB), resulting in the L, Lmi, Lm and Lgen types (Figure 1). Having different types of MTB enabled us to control the impacts of false positives and false negatives in the input network, and the amount of crosstalk between MTBs.

The different MTB types have drastically different numbers of possible occurrences in a network (Figure 1). For some types, there is at most a linear number of MTBs with respect to the number of mRNAs and miRNAs in the network (types R and L). For some other types, the maximum number of MTBs is exponential, but the number of maximal MTBs, i.e., MTBs not being submatrices of other MTBs, is linear (types Rmi, Rm, Lmi and Lm). Type Rgen could give an exponential number of maximal MTBs in theory, but in practice a tractable number is usually observed. Finally, type Lgen has an exponential number of MTBs, both in theory and in practice. Consequently, we developed a variety of algorithms to identify MTBs of the different types, from simple graph searching algorithms that can efficiently identify all MTBs of a certain type, to algorithms that only return a subset of MTBs with the highest scores based on the intra-MTB density of interactions.

Expression of miRNAs and mRNAs in the same MTBs are significantly more anti-correlated than general miRNA-target pairs

As a way to check whether the MTBs we identified represent modules with biological relevance, we examined the expression levels of the miRNAs and mRNAs in human cell lines obtained from RNA-seq experiments performed by the ENCODE Project Consortium [69, 23]. The details of the analysis pipeline are given in Materials and Methods (see also Figure 2 and Figure S1a). Briefly, for each miRNA-mRNA pair in an MTB, we calculated the Pearson correlation of their expression across the cell lines. For each MTB, we then counted the fraction of pairs having correlation values more negative than a certain threshold t, multiple values of which (from -0.1 to -0.7) were tested. A large fraction of pairs having expression correlations

more negative than the threshold would indicate that the regulatory effects of the miRNAs on the mRNAs were sufficiently strong to be observed in the expression data. To make sure that the negative correlations were not obtained by random chance, we compared these fractions with the corresponding fractions in random sets of expressed miRNAs and mRNAs of the same sizes as the identified MTBs. A p-value was then computed to determine if the fractions from the MTBs were significantly higher than those from the random background.

In addition, we wanted to check if the negative correlations were simply a general phenomenon among miRNAs and their targets regardless of their MTB memberships. We therefore repeated the above procedure using a second background set of miRNA-mRNA pairs that composed of miRNAs and their targets not participated in the same MTBs.

From the results for the expressed union set with TarBase interactions (Figure 3), we see that for moderate values of the correlation threshold (-0.1 to -0.4), for most MTB types, significantly more miRNA-mRNA pairs in the MTBs were anti-correlated in expression than random miRNA-mRNA pairs (panels a and b). For example, considering miRNA-mRNA pairs with expression correlation $<$ -0.1, all MTB types except type R had a significantly higher fraction of such pairs than random miRNA-mRNA pairs at the p=0.01 level.

Importantly, the miRNA-mRNA pairs in the MTBs were also significantly more anti-correlated in expression than miRNA-target pairs not in the same MTBs (Figure 3c,d), which suggests that the regulatory effects of miRNAs are either stronger or more clearly observed on their targets within the same MTBs than their other targets.

Significant p-values were obtained for both the MTBs from the high-confidence set (panels a and c) and the high-coverage set (panels b and d). We have also repeated our procedure for the networks without the validated miRNA-target interactions from TarBase (Figure S2), and when related miRNAs with the same miRNA number but different modifiers (such as 5p and 3p) were grouped (Figures S3 and S4). In all cases, the same general conclusion was drawn, that significantly more within-MTB miRNA-mRNA pairs were strongly anti-correlated in expression than random pairs and miRNA-target pairs not in the same MTBs. These consistent results show that MTB is a robust method for identifying miRNA-mRNA modules with strong expression relationships despite the fact that gene expression data were not used in defining the MTBs.

Figure 4 shows the distributions of fractions of miRNA-mRNA pairs satisfying the correlation threshold in an example setting. It can be seen that for some MTBs, almost all miRNA-mRNA pairs (with a fraction close to 1) had expression correlations more negative than threshold $t = -0.2$. More generally, about two-third of the MTBs had more than 20% of their miRNA-mRNA pairs satisfying this correlation threshold. In contrast, for both random groups of miRNAs and mRNAs, and other miRNA-target pairs, almost none of them had expression correlation more negative than -0.2.

Comparing the different MTB types, the general types that allow both the miRNAs to have extra-MTB targets and the mRNAs to be targeted by extra-MTB miRNAs (Rgen and Lgen) produced more MTBs as expected (Figures 5, S5-S7).

Interestingly, the MTBs of these two types also contained miRNAs and mRNAs with more significant anti-correlations, and over a broader range of correlation threshold values (Figure 3). In contrast, due to the rigid requirements of type R, no MTBs of this type could be discovered from the high-coverage set and few were identified from the high-confidence set. Even when MTBs of this type could be found, their miRNA-mRNA anti-correlations of expression were not statistically significant. These results confirm the importance of explicitly considering non-fully-connected miRNA-mRNA modules and possible errors in the input miRNA-mRNA interaction networks.

Non-expression features can be used to identify MTBs with strong miRNA-mRNA expression anti-correlation

While the MTBs in general contained a significantly higher fraction of miRNAs and mRNAs with strong expression anti-correlation, we were interested in knowing whether some simple features of the MTBs could help identify the subset of MTBs with particularly strong expression anti-correlation without referencing the expression data. This would be particularly useful in identifying the most interesting MTBs when expression data are not available. To explore this possibility, for each MTB we identified, we computed 7 non-expression features, including the number of mRNAs and miRNAs in it, the density of 1's in the MTB, in the same rows, columns or either but outside the MTB, and the MTB type. We then used these features to construct a Random Forest model [14] for predicting the fraction of miRNA-mRNA pairs within the MTB with expression correlation more negative than $t = -0.1$. Based on the results of 10-fold cross-validation, the average area under the receiver-operator characteristics (AUC) of ten equal-width fraction classes was 0.97, which is significantly higher than what would be expected for random predictions (AUC=0.5), indicating that the features were useful in identifying the MTBs with higher fractions of strong miRNA-mRNA expression anti-correlation.

We then looked for the features most important for identifying MTBs with strong expression anti-correlations between their member miRNAs and mRNAs. An exhaustive search of feature combinations identified two features that were consistently the most important in a 10-fold cross-validation procedure, namely the number of mRNAs in an MTB and the fraction of miRNAs outside an MTB that target the mRNAs in the MTB. Basically, MTBs with very strong anti-correlations between their miRNAs and mRNAs have a relatively small number of mRNAs and these mRNAs are targeted by few other miRNAs outside the MTBs, which are consistent with the intuition that MTBs with these properties are more autonomous, although the exact relationships of these two features with the fractions passing the correlation threshold are not linear in general.

Comparison with a previous method

To further check if MTBs represent novel miRNA-mRNA modules, we compared them with the miRNA regulatory modules (MRMs) identified by the Yoon and De Micheli method [75] from the same networks. It is one of the few methods in the literature that identify miRNA-mRNA modules from a miRNA-target network without requiring expression data as input. We applied the same procedure described

above to check the fraction of miRNA-mRNA pairs within each MRM identified by this method with expression correlation more negative than a threshold. We then collected all these fractions, and compared them with the corresponding fractions from the MTBs. Since type Lgen was found to be most biologically relevant in terms of miRNA-mRNA expression anti-correlation, in this and subsequent analyses we focus on this type of MTBs.

We found that for all threshold values between $t = -0.1$ and $t = -0.7$, there was constantly a higher fraction of miRNA-mRNA pairs within the MTBs passing the anti-correlation threshold than the MRMs identified by Yoon and De Micheli method as reflected by p-values < 0.5 (Figure 6). In many settings, the difference in these fraction values was statistically significant. For example, for all threshold values between $t = -0.1$ and $t = -0.5$, there was always a significantly higher fraction of miRNA-mRNA pairs in the MTBs passing the anti-correlation threshold than those in the MRMs at the $p = 0.01$ level based on the high-confidence network with TarBase interactions. These results further confirm that the MTBs successfully identified groups of miRNAs and mRNAs with strong expression relationships from the miRNA-target networks alone.

Potential widespread expression buffering between mRNAs in the same MTBs

After checking the biological relevance of MTBs, we then used them to study whether mRNAs commonly targeted by some miRNAs buffer each other in terms of their expression levels. We studied this question using three different methods.

First, we reasoned that if different mRNAs buffer each other, they should exhibit a positive correlation of expression levels across different cell types. To test if it was the case, we applied a procedure similar to the one we used for testing miRNA-mRNA anti-correlations described above. Specifically, we asked whether a significantly higher fraction of mRNA pairs in the same MTBs had expression correlation more positive than a threshold t, as compared to random mRNA pairs and mRNA pairs targeted by the same miRNA but not in the same MTBs.

The results (Figure 7) show that indeed significantly more mRNA pairs within type Lgen MTBs were strongly correlated in expression than both types of background mRNA pairs at various values of t from 0.1 to 0.4, no matter we considered the high-confidence or high-coverage set of miRNA target predictions, and whether experimentally validated pairs from TarBase were included or not. The p-values in the comparison with mRNA pairs targeted by same miRNAs but not in same MTBs as background were particularly significant (Figure 7b), indicating that MTBs helped discover mRNAs with strong expression correlations that could be hard to observe if all targets of a miRNA were considered together as a group.

We noticed that the positive correlations observed between mRNAs in the same MTBs are necessary but not sufficient for showing that they buffer each other. Since most mRNAs in the same MTB are expected to be targeted by the same miRNAs, a plausible alternative explanation is that the positive correlations were simply due to independent regulation by the same miRNAs without a feedback mechanism for the mRNAs to affect the expression level of each other. We argue that this co-regulation mechanism cannot fully explain the significant positive correlations observed, because mRNAs targeted by same miRNAs but not in same MTBs were

not as correlated in expression as those in the same MTBs. Also, one possible situation in which mRNAs cannot back-regulate their targeting miRNAs, and thus they cannot buffer other mRNA targets, is when the miRNAs have saturated expression levels across different cell types. This was also unlikely the case since we observed significant anti-correlations between miRNAs and their mRNA targets in the same MTBs.

Nonetheless, the above arguments cannot rule out the possibility that the main function of MTBs was to identify the more reliable miRNA-target pairs from the noisy interaction network, and thus co-regulation effects between mRNAs in the same MTBs were still stronger than other mRNAs targeted by the same miRNAs according to the network.

To more directly distinguish between co-regulation and buffering, we applied a second analysis method. The main idea is that if some mRNAs buffer each other, the expression level of one mRNA would provide some information for explaining the expression level of another mRNA, even when the expression level of the targeting miRNAs have already been considered. In other words, we wanted to test if one mRNA could help explain the expression level of another mRNA that could not be fully explained by the miRNA targets alone. This idea can be quantified by using partial correlation. Suppose R, T_1 and T_2 represent a miRNA regulator, target mRNA 1 and target mRNA 2, respectively. We define $f(R, T_1)$ as the correlation between R and T_1, and $f(R, T_1|T_2)$ as the expected correlation between R and T_1 given the level of T_2. The difference between them, $d(R, T_1, T_2) = f(R, T_1|T_2) - f(R, T_1)$ would be negative if T_2 provides additional information for explaining the expression anti-correlation between R and T_1, and it would be close to 0 if T_2 provides no additional information, such as when T_1 and T_2 were independently regulated by R. A similar method based on conditional mutual information was previously used to identify sponge modulators in miRNA-target networks [66].

Given R and T_1 from an MTB, we compared the partial correlation values using other mRNAs from the same MTB as T_2 with the values obtained by using other targets of R outside the MTB as T_2. The results (Figure 8) show that as expected, significantly more mRNAs from the same MTBs gave a strong negative value of $d(R, T_1, T_2)$ than other mRNA targets of the miRNAs, and the results were consistently obtained from all four miRNA-target networks. These results suggest that the mRNAs in an MTB do help explain the expression levels of each other in addition to what the regulating miRNAs can explain.

Finally, we reasoned that if two mRNAs buffer each other, they should have an expression correlation stronger than other mRNA pairs co-regulated by the same miRNA, even if we consider only those within the same MTBs. In other words, for any two mRNAs T_1 and T_2 from the same MTB, if $d(R, T_1, T_2)$ is strongly negative, $f(T_1, T_2)$ should be strongly positive. To test if this was the case, we picked the top x MTBs with most negative $d(R, T_1, T_2)$ values and bottom x MTBs with most positive $d(R, T_1, T_2)$ values. We then repeated the correlation analysis above (the first method) using either only the top MTBs or only the bottoms ones. For the 112 parameter settings we tested involving different input networks and different values of x and t, the top MTBs had equal or more significant p-values in 103 of the cases (92% of the 112 settings). This result confirmed our intuition that the top MTBs

with potentially stronger expression buffering among its member mRNAs had their expression levels more correlated.

Taken together, the results of the three methods show that the mRNAs in the same MTBs are significantly correlated in expression, and this cannot be explained purely by the fact that they are regulated by the same miRNAs. We propose that one likely alternative explanation is that these mRNAs buffer each other in terms of their expression levels.

mRNAs in same MTBs have related biological functions

In addition to expression correlations, another way to check the biological relevance of MTBs is to test whether the genes (mRNAs) in the same MTB are enriched in particular functional categories. We collected the Gene Ontology (GO) [8] annotation of the genes in each MTB, and computed the enrichment score of each GO term using both hypergeometric tests and EASE scores [37]. We then collected the most significant enrichment score of each MTB to form a distribution, and compared it with the corresponding distribution of a background set of mRNAs, where the background was either random sets of mRNAs with the same sizes as the MTBs, or mRNAs targeted by same miRNAs but not included in the same MTBs.

From the results (Figure 9 and Figure S8), it is seen that the genes in the MTBs were indeed more functionally related than both types of background mRNA sets. The results were largely unaffected by the exact way to compute enrichment scores (hypergeometric test p-values or EASE scores), although the results based on MTBs obtained from the high-coverage set of miRNA-target interactions were more significant.

Functional enrichment of mRNAs in same MTBs is not only due to co-expression

Since the mRNAs in an MTB were correlated in expression in general, we further tested whether co-expression alone was sufficient to explain the functional enrichment. To test it, we sampled random sets of mRNAs with similar sizes and expression correlation profiles as the MTBs, and computed the hypergeometric test p-values of the GO terms of the mRNAs in each set. We then compared the distribution of the most significant enrichment score from each of these sets with the scores from the MTBs.

The functional enrichment scores of the MTBs were found to be significantly stronger than the scores from the random sets of genes with similar levels of co-expression (Figure 10), especially when MTBs were identified from the two high-coverage miRNA-target networks. These results show that MTBs were able to identify groups of functionally related genes better than using co-expression information alone.

MTB as a way to annotate miRNA functions

Finally, we explored the potential application of MTBs in identifying functionally related miRNAs. Currently, functional annotation of miRNAs is far less complete than protein-coding genes. Since each MTB represents a largely autonomous module, we hypothesized that the miRNAs in an MTB were functionally related to one another and to the mRNAs in the same MTB. To check if this was the case, for

each MTB, we identified GO terms that were significantly enriched ($p<0.05$) based on the GO term annotations of the mRNAs. We then checked if these enriched GO terms were also related to the functions of the miRNAs in the same MTB. Table 3 shows several interesting examples we identified.

In the first example (MTB 1), the mRNA encoding AAK1, MAPK1 and PDK3 were annotated with the GO term "protein serine-threonine kinase activity". We found that several miRNAs in MTB 1 are able to target the activities of the MAPK family of serine-threonine protein kinases. miR-320a has been shown to directly target MAPK1 activity to control the expression of pro-inflammatory cytokines in patients with myasthenia gravis [19]. Other miRNAs in the same MTB, miR-17 [20, 74] and miR-20b [20] can both target the MAPK signaling cascades to regulate cell cycle phase transition [20] and keratinocyte differentiation [74]. In addition, miR-93 also directly modulates the activity of the protein serine-threonine kinase, LATS2 to control tumor angiogenesis and metastasis in human breast cancer cells [27].

In MTB 2, genes encoding CBX5, HMGA2 and RNF20 are annotated by the GO term "heterochromatin". The expression of HMGA2, a non-histone protein with important roles in chromosomal architecture and oncogenic transformation, is directly targeted by the let-7 miRNA [46, 55]. Interestingly, HMGA2 also functions as a ceRNA of Tgfbr3 through the let-7 miRNA family that commonly targets them, resulting in the promotion of lung cancer progression [45]. Furthermore, genome-wide chromatin-binding analysis suggested that let-7 and miR-185 are heterochromatin-bound miRNAs that can associate with AGO2 in the nucleus of senescent cells to mediate transcriptional gene silencing of proliferation-promoting genes [11]. Together, results from previously published studies supported our MTB classification of miR-185 and let-7 as miRNAs important in heterochromatin-binding and/or chromatin-remodeling.

We next examined the functions of the miRNAs in MTB 3, the mRNAs of which are annotated by the GO term "Notch signaling pathway". We found that miR-34a inhibits cell proliferation in part by directly targeting the expression of CDK6 [67], an important cell cycle regulator whose expression is dependent on the Notch signaling pathway in T cell development [40]. Likewise, the level of miR-497 has been shown to inversely correlate with CDK6 expression to regulate cell cycle progression [30]. Furthermore, both miR-34a and miR-449a have been shown to target the expression of Notch1 [15, 76], a member of the Notch family of receptors in human cancer cell lines. Two of the miRNAs in MTB 3 can also target ligands of the Notch receptors: miR-34a is known to directly target Delta-like 1 [21] whereas miR-15a targets the non-canonical notch ligand, Delta-like 1 homolog [4].

Finally, in MTB 4, the mRNA-associated GO term "proteasome-mediated ubiquitin-dependent protein catabolic process" is also functionally related to the three miRNAs in the MTB. miR-25 has been shown to directly target the E3 ubiquitin ligase, WWP2 [53] to control the reprogramming of somatic cells to induced pluripotent stem cells. miR-363 directly inhibits a ubiquitin-specific protease, USP28 to promote proteasome-mediated degradation of Myc in human hepatocellular carcinoma [36]. miR-93, which lies in the miR-106b-25 cluster, has been shown to target the expression of β-TRCP2, a component of the SCF ubiquitin ligase

complex important in the ubiquitination and subsequent proteasomal degradation of target proteins [65].

Collectively, these examples demonstrate that enrichment analysis based on the annotation of GO terms to the mRNAs in an MTB could be used as a way to annotate the functions of miRNAs in the same MTB.

Discussion

In this study, we have shown that mRNAs in the same MTBs have significant expression correlations that cannot be explained purely by the fact that they are regulated by the same miRNAs. We have used multiple methods to show the high possibility that these mRNAs buffer each other in terms of expression, which suggests that ceRNAs could play an important role in the regulation of many mRNAs. In order to fully test the generality of the ceRNA hypothesis, it is necessary to perform perturbation experiments to see how the alteration of the expression level of one mRNA could affect other mRNAs regulated by the same miRNAs. Without such experimental data, in this study we do not aim at completely proving or disproving the generality of the ceRNA hypotheses. Instead, we think the MTBs represent small miRNA-target modules that could be very useful in identifying candidate miRNAs and mRNAs of future experimental studies in testing hypotheses related to ceRNA.

The fact that more significant p-values were observed for the MTB types with higher error tolerance (such as Rgen and Lgen) suggests that analysis results could indeed be misled by the errors present in the networks, and that trading off the purity of modules with some error tolerance is a reasonable strategy to handle the current noisy miRNA-target networks. On the other hand, although the Rgen and Lgen types of MTBs had the most statistically significant results, there could also be interesting cases identified by the other types. For instance, type R MTBs theoretically represent fully autonomous modules with complete target sharing among its member miRNAs, which are ideal cases for studying the ceRNA hypothesis. When miRNA-target networks become more complete and accurate in the future, more statistically significant results may be obtained from this and the other types of MTB with more stringent definitions.

One aspect of MTBs that we have not yet explored in this study is their cell-type specificity. Since MTBs are defined purely based on the miRNA-target connections, the different miRNAs and mRNAs in an MTB may not be all expressed in the same cell types. It would be interesting to study whether different miRNAs in an MTB usually co-express in the same cell types and co-regulate the common mRNA targets, or express in different cell types and act as alternative regulators.

In this work, we have focused on the use of expression data from ENCODE, which include matched mRNA and miRNA expression data from the same cell lines. We have also tested mRNA-mRNA positive correlations within our identified MTBs using a larger data set originally obtained from more than 73,000 microarray experiments [58]. Statistically significant results were again observed, but within a narrower range of correlation threshold t. We will test the concept of MTBs using larger data sets in the future.

While we have defined eight different types of MTB, actually they can all be described by a general framework. A detailed discussion is given in the Supplementary

Materials. Briefly, a general MTB can be defined as a submatrix with an associated score, which is a combination of (1) the missing 1's in the MTB, (2) extra 1's in other rows of the defining columns, and (3) extra 1's in other columns of the defining rows. Each of the eight MTB types corresponds to a particular way to combine these three components. While this general model appears to be more of theoretical interests, it actually has a real application in helping define MTBs from weighted networks in which each miRNA-mRNA pair is given a weight that indicates how likely they interact. Using such a weighted network would provide more information for analysis and avoid defining arbitrary thresholds to form a binary network.

Conclusion

In this study, we have introduced microRNA-target biclusters (MTBs) as a method to systematically identify largely autonomous modules purely from the connections in a noisy miRNA-target network. To cater for modules involving miRNAs that do not target all mRNAs in the module, and the presence of false positives and false negatives in the network, we have defined eight different types of MTB with different levels of autonomy and error tolerance. We have shown that for some MTB types, especially those with higher error tolerance, the identified modules are biologically relevant by having significant anti-correlations between their member miRNAs and mRNAs as compared to both random miRNA-mRNA pairs and miRNA-target pairs not in the same MTBs. We have checked the robustness of our method using different input networks (high confidence or high coverage, with or without experimentally validated interactions), different values of the correlation threshold t in computing p-values, and whether to pre-group related miRNAs. The results were consistent across a wide spectrum of parameter settings.

The identified MTBs have enabled us to study how the expression patterns of their member mRNAs are related, with relatively small influence from other miRNAs and mRNAs outside the MTBs. Using three different analysis methods, namely direct expression correlation among the mRNAs, gain of miRNA-mRNA anti-correlation information by conditioning on another mRNA, and separate correlation analyses of MTBs with the strongest and weakest information gain, we have shown that there is strong correlation between the expression levels of mRNAs in the same MTBs that can well be explained by expression buffering as stated in the ceRNA hypothesis. These results show that although the regulatory effects of miRNAs are only partially reflected by the expression levels of their target mRNAs, and mRNA expression is affected by other regulatory mechanisms, it is still possible to use transcript levels to study the effects of miRNAs by decomposing a complex and noisy network of miRNA-target interactions into small modules that can be analyzed individually.

In the long term, the methods proposed in this study should be extended to model the hierarchical relationships between different MTBs and incorporate other regulatory mechanisms, to provide a more complete picture of the complex interactions between various types of biological objects in gene regulatory networks.

Materials and methods

Construction of miRNA-target networks

We collected experimentally validated human miRNA-target pairs from TarBase (v6.0) [73], which contained one of the most comprehensive sets of validated miRNA-

mRNA interactions. We considered only the experimental types that likely report direct miRNA-mRNA interactions, namely PCR, ReportGeneAssay and Sequencing.

In addition, we gathered computationally predicted human miRNA-mRNA interactions using 5 methods based on different prediction approaches, namely miRanda (Aug 2010) [39], miRGen (v2.0) [3], PicTar (Hg18) [5], PITA (v6) [44] and TargetScan (6.0) [47]. The dataset for miRanda used was the human predictions with "Good mirSVR score, Conserved miRNA", downloaded from `http://cbio.mskcc.org/microrna_data/human_predictions_S_C_aug2010.txt.gz`. The miRGen data file was downloaded from `http://diana.cslab.ece.ntua.gr/data/public/TF_GENEID_miRNA_sorted.txt`. The dataset for PicTar was the PicTar2 predicted target genes with conservation at the mammals' level based on RefSeq gene models and human hg18 reference assembly, downloaded from `http://dorina.mdc-berlin.de/rbp_browser/downloads/pictar_hg18_mammals.bulk_download.csv`. For PITA, the Human Top predictions of miRNA targets were downloaded from `http://genie.weizmann.ac.il/pubs/mir07/catalogs/PITA_targets_hg18_0_0_TOP.tab.gz`. For TargetScan, the data used were the predicted conserved targets, downloaded form `http://www.targetscan.org/vert_61/vert_61_data_download/Predicted_Targets_Info.txt.zip`.

The gene names in all data files were converted to official gene symbols using the lookup table in HGNC [34]. Records with unrecognized gene names were ignored.

A high-confidence interaction network was constructed by taking the union of the 1,000 highest-scoring predictions from each method (where the number for miRGen was slightly larger due to ties in prediction scores). A second network was constructed by adding to this network the experimentally validated interactions in TarBase. Similarly, two high-coverage interaction networks were constructed by taking the union of the 5,000 highest-scoring predictions from each method, one with TarBase interactions and one without.

Expression data

To study the expression levels of miRNAs and mRNAs across different cell types, we collected RNA-seq data in whole cells of human cell lines from ENCODE [69, 23], namely A549, AGO4450, BJ, GM12878, H1-hESC, HeLa-S3, K562, MCF7, NHEK and SK-N-SH, which contained the largest number of non-zero expression values for our mRNAs and miRNAs among all the human cell lines with RNA-seq data available from ENCODE at the time of collection. We used long PolyA+ RNA data to compute expression levels of mRNAs, and short total RNA data for miRNAs. Expression levels were computed by the number of reads mapped to each gene per kilobase per million reads (RPKM). We combined values from multiple replicates of the same experiment by taking their average.

As our goal was to study expression relationships between miRNAs and mRNAs, we focused on the set of mRNAs and miRNAs with non-zero expression values in at least 8 of the 10 cell lines. Considering only these miRNAs and mRNAs, we obtained the four integrated networks used in our analyses, namely the high-confidence/high-coverage expressed union network with/without TarBase interactions (Table 1).

Definitions of MTBs and identification algorithms

As described in the Results section, we defined eight MTB types that differ in whether missing 1's are allowed in the defining submatrix, and whether extra 1's are allowed in the defining rows and columns outside the MTB (Figure 1). Here we provide detailed definitions of the eight types, and describe the corresponding algorithms for identifying the MTBs of each type from a miRNA-target network. In our analyses, by default we considered only MTBs containing at least two mRNAs and at least two miRNAs. For the analysis of positive expression correlations between mRNA pairs, in order to avoid having only one correlation value per MTB, we further considered only MTBs with at least 3 mRNAs.

Type R

Type R is the most restrictive type that requires each participating miRNA to target all participating mRNAs but no other mRNAs, and each participating mRNA to be targeted by all participating miRNAs but no other miRNAs. In the matrix representation, an MTB of this type is a submatrix with all 1's, and all other elements on the same rows and columns are 0's. Since each row and each column can participate in at most one MTB, the total number of MTBs is at most $min(|R|, |C|)$, where R and C are the sets of all rows (i.e., mRNAs) and all columns (i.e., miRNAs), respectively, and the notation $|X|$ denotes the size of any set X.

We developed an algorithm to identify all MTBs of this type from a miRNA-target network in linear time. The basic idea is to use the columns with 1's in a row as its signature, and group all rows with the same signature together with the help of a hash table. Similarly, we defined the signature of a column as the rows at which it has 1's, and grouped all columns with the same signature together. For each group of rows, if the columns in its signature do not have 1's at other rows, it forms an MTB with these columns. Otherwise, by the definition of type R MTB, the whole group of rows cannot be members of any MTB. In this algorithm, whether there are other 1's in these columns can be efficiently checked by the following method. Suppose r is the group of rows, c is the set of columns defining its signature, and j is one of these columns. All columns in c do not have other 1's if and only if j belongs to a group with signature r for all $j \in c$.

The pseudocode of the whole algorithm is given in Algorithm 1.

Type Rmi

Type Rmi is the same as type R except that the mRNAs of an MTB are allowed to be targeted by additional miRNAs. In the matrix representation, extra 1's are allowed in other columns of the defining rows. There can be an exponential number of type Rmi MTBs, because if (r, c) is an MTB, then (r, c') is also an MTB for any set of columns $c' \subset c$. On the other hand, if we define a maximal MTB as one that is not a submatrix of another MTB, then each column can participate in at most one maximal MTB. Therefore the total number of maximal MTBs is at most $|C|$.

We modified the algorithm for type R to identify all maximal MTBs of type Rmi. For each column, we defined its signature as the rows at which it has 1's. We then grouped all columns with the same signature with the help of a hash table. Each resulting group of columns and the rows in their signatures form a maximal type Rmi MTB, with no additional checking required.

The pseudocode of the algorithm is given in Algorithm 2.

Type Rm

Type Rm is the transpose of type Rmi. It allows each miRNA of an MTB to target other mRNAs outside the MTB, but does not allow the mRNAs to have other targeting miRNAs. Using the same argument for type Rmi, there can be an exponential number of type Rm MTBs, but at most $|R|$ maximal MTBs.

The algorithm we used for identifying all maximal type Rm MTBs is analogous to the one for type Rmi, except that we grouped rows based on their signatures instead.

The pseudocode of the algorithm is given in Algorithm 3.

Type Rgen

Type Rgen maintains the requirement that all miRNAs participating in an MTB must target all participating mRNAs, but the miRNAs are allowed to have other targets and the mRNAs are allowed to be targeted by other miRNAs. This type of MTBs is best described by the graphical representation, where each MTB is a biclique, i.e., a complete subgraph with all the miRNA nodes connecting to all the mRNA nodes. Again, there can be an exponential number of MTBs, as each subgraph of a type Rgen MTB is also a type Rgen MTB. There can also be an exponential number of maximal type Rgen MTB. For example, if there are $2^{|C|}-1$ rows and the signature of each row is the same as its index, i.e., the first row has signature 000...001, the second row has signature 000...010, the third row has signature 000...011, and so on, then each of the $2^{|C|}-1$ non-empty column combinations participates in a different maximal MTB. Because of the exponential number of possible maximal MTBs, and the fact that finding maximal bicliques is NP hard [59], in theory it is infeasible to identify all maximal type Rgen MTBs in a miRNA-mRNA network.

In practice, however, both the number of maximal type Rgen MTBs and the size of each are small in the networks we studied. We therefore used an iterative algorithm to find all maximal type Rgen MTBs, based on the Apriori algorithm proposed for association rule mining [1, 13]. The basic idea is that if (r, c) is a type Rgen MTB, then for any $c' \subset c$, (r, c') must also be a type Rgen MTB. One can then iteratively discover MTBs with two columns, three columns, and so on, by testing k-column sets in the k-th iteration, constructed by merging two $(k-1)$-column sets in the previous iteration.

The pseudocode of the algorithm is given in Algorithm 4.

Type L

The definition of type L MTB involves three rules. First, each participating miRNA is allowed to target only some of the participating mRNAs, but it cannot target any other mRNAs. Second, each participating mRNA is allowed to be targeted by only some of the participating miRNAs, but it cannot be targeted by other miRNAs. Finally, in the graphical representation, the nodes that represent the participating rows and columns should all be connected, i.e., there should be a path between any two nodes. In other words, each type L MTB is a connected component. Since each row and each column can participate in at most one MTB, the total number of type L MTBs is at most $min(|R|, |C|)$.

We used a standard breadth-first search algorithm to find all connected components, i.e., all type L MTBs, in linear time.

The pseudocode of the algorithm is given in Algorithm 5.

Type Lmi

Type Lmi MTB differs from type L by allowing the participating mRNAs of an MTB to be targeted by additional miRNAs outside the MTB. Since each type Rmi MTB is also a type Lmi MTB, there is at maximum an exponential number of type Lmi MTBs in a miRNA-target network. Since each column can participate in at most one maximal MTB, there are no more than $|C|$ maximal type Lmi MTBs.

It is easy to see that the set of maximal type Lmi MTBs is exactly the same as the set of maximal type L MTBs. The algorithm for finding all maximal type L MTBs can thus be used for finding all maximal type Lmi MTBs. However, we did not adopt this approach for two reasons. First, by doing so it would be meaningless to define type L and type Lmi MTBs as two separate types. Second, a maximal type Lmi MTB is likely to have many member miRNAs and mRNAs not having interactions, leading to a low density of interactions within the MTB.

We therefore developed an algorithm for finding high-scoring type Lmi MTBs instead. The score of an MTB is defined as the density of 1's in the defining submatrix, where the density of 1's in an MTB is defined as the number of 1's divided by the total number of elements in the submatrix. The algorithm starts with the set of maximal type Rmi MTBs, which all have an interaction density of one by definition. We then removed MTBs that are too similar to another one. After that, for each column not in any MTB, we tested if it was reasonable to add it and all rows with a 1 in that column to the MTB. If the resulting density of 1's in the new MTB did not drop below a certain threshold (which we set to 0.3), we considered the addition of the column as reasonable. If there were multiple reasonable additions, we chose the one with the highest resulting density of 1's and repeated the process. Otherwise, the current MTB was returned as one of the high-scoring type Lmi MTBs.

The pseudocode of the algorithm is given in Algorithm 6.

Type Lm

Type Lm MTB is the transpose of type Lmi MTB. All the discussions about type Lmi MTBs can be applied to type Lm MTBs by swapping the rows and columns.

The pseudocode of the algorithm for finding all high-scoring type Lm MTBs is given in Algorithm 7.

Type Lgen

Type Lgen has the most relaxed definition among the eight types, and is likely the most practical one. Each miRNA in a type Lgen MTB is allowed to target only some of the mRNAs in the MTB, and is allowed to target other mRNAs. Likewise, each mRNA is allowed to be targeted by only some of the miRNAs in the MTB, and is allowed to be targeted by other miRNAs. To avoid having completely unrelated miRNAs and mRNAs in the same MTB, we maintained the connectedness requirement from type L. Since type Rgen is a special case of type Lgen, there can also be an exponential number of type Lgen MTBs. On the other hand, the number

of maximal MTBs is limited by the number of connected components in the network, which is at most $O(min(|R|, |C|))$.

Due to the exponential number of type Lgen MTBs, both theoretically and practically, it is infeasible to return all of them. On the other hand, it is also not meaningful to return all maximal MTBs, since they are usually very sparse and contain miRNAs and mRNAs that are only weakly connected. Therefore as in the cases of type Lmi and type Lm MTB, we adopted a different approach to return high-scoring MTBs, which are MTBs with a high density of 1's within the defining submatrices. We developed an algorithm to find these high-scoring MTBs, based on some ideas from a previously method proposed for finding communities in partite networks [24]. First, we used the algorithm for type Rgen MTBs to find all maximal bicliques, and called each of them a bicluster. We then removed biclusters that are too similar to another one. After this step, we iteratively added extra rows or columns to each MTB in ways similar to the algorithms for type Lmi and type Lm MTBs, except that when a column/row was added to an MTB, it was not required to also add the rows/columns with 1's in the adding column/row. For each bicluster, the best addition was kept. The process was repeated until no more mRNAs or miRNAs could be added without causing the density to drop below a threshold.

The pseudocode of the algorithm is given in Algorithm 8.

Workflow for expression correlation analyses

We used a unified workflow for studying the negative correlations between miRNAs and mRNAs in an MTB (Figure 2 and Figure S1a). Each time we considered one of the four integrated miRNA-target interaction networks of expressed miRNAs and mRNAs as input (Table 1, High-confidence/high-coverage expressed union with/without TarBase interactions). MTBs of the different types were identified from the network using the algorithms described above. For each MTB, we calculated the Pearson correlation between the expression levels of each pair of participating miRNA and mRNA across the human cell lines. We then summarized all these correlations by computing the fraction of them more negative than a correlation threshold t, multiple values of which (-0.1 to -0.7 with a step size of 0.1) were tested. After collecting all these fractions from the MTBs of a particular type, we compared them with the fractions from two backgrounds. The first one involved 1,000 random sets of expressed miRNAs and mRNAs with sizes matching the size distribution of the actual MTBs. The second one involved the same miRNAs and their other targets not included in the same MTBs as them. To quantify the comparisons, we used Wilcoxon rank-sum test to calculate a one-sided p-value for each MTB type at each value of t. A significant p-value would mean the fractions from the MTBs were significantly higher than the set of fractions in comparison. As our goal was to compare the results in various parameter settings rather than emphasizing on the significance of one particular set of results, the reported p-values were not corrected for multiple hypothesis testing. We remark that if one was to use the concept of MTB to identify one set of reliable miRNA-target modules for downstream analyses, the statistical significance of such modules should be carefully corrected taking into account the number of hypothesis tests performed.

We also repeated the analysis when different miRNAs with the same miRNA numbers but different modifiers (such as has-mir-121a and hsa-mir-121b) were grouped

together. The expression value of each group was defined as the average expression of the member miRNAs.

In the same way, we also tested the positive correlations between mRNAs in same MTBs, in which case we computed the fractions of pairs with expression correlation higher than a threshold t, where t took values from 0.1 to 0.7. The fractions obtained from mRNA pairs in same MTBs were first compared to fractions from random pairs of expressed mRNAs, and then to pairs of mRNAs targeted by the same miRNAs but were not in the same MTBs.

Predicting MTBs with strong miRNA-mRNA expression anti-correlation

We developed a method for predicting MTBs with strong miRNA-mRNA expression anti-correlation when expression data are unavailable. For each MTB, we constructed seven non-expression features, namely 1) its number of mRNAs, 2) its number of miRNAs, 3) the density of 1's in its submatrix, 4) the density of 1's in other rows of the defining columns, 5) the density of 1's in the other columns of the defining rows, 6) the density of 1's in the other rows of the defining columns or the other columns of the defining rows, and 7) the MTB type. Each MTB was thus represented by a vector of seven numeric values. The goal was to identify the anti-correlation class of each MTB, where ten equal-width classes were defined based on the distribution of average anti-correlation values of the MTBs. We then took 9/10 of the MTBs to train a Random Forest model using the implementation in Weka [35], and tested its accuracy using the remaining 1/10 of MTBs with their anti-correlation classes hidden. We repeated the process with 10 random sets of training-testing data, and reported their average area of the receiver operator characteristics (AUC).

Comparison with miRNA regulatory modules from Yoon and De Micheli

We compared the miRNA-mRNA expression anti-correlation with the miRNA regulatory modules (MRMs) from Yoon and De Micheli [75]. We implemented this method and applied it to find MRMs from each of our input miRNA-target networks. For each identified MRM, we computed the fraction of miRNA-mRNA pairs with expression correlation more negative than a threshold t. We then compared these fraction values with the fraction values from our type Lgen MTBs using a one-sided Wilcoxon rank-sum test. A significant p-value would indicate that the fraction values from the MTBs were significantly higher than the MRMs.

Workflow for testing whether correlated expression of mRNAs were more likely due to buffering than co-regulation

To test if the correlated expression of two mRNAs in the same MTB is due to buffering or co-regulation, we applied a method similar to the one in Sumazin et al. [66]. The idea is to compute $d(R, T_1, T_2) = f(R, T_1|T_2) - f(R, T_1)$, where R is a regulating miRNA, T_1 and T_2 are two mRNA targets of it, f is the Pearson correlation function, and $f(R, T_1|T_2)$ is defined as the expected correlation between R and T_1 after dividing the cell lines into two groups based on the expression value of T_2 (above mean and below mean). If $d(R, T_1, T_2)$ is negative, it would mean that the expression relationship between R and T_1 can be better explained when the

expression of T_2 is known, and thus T_1 and T_2 are not independently regulated by R, but they affect each other possibly due to buffering.

To globally test if the (R, T_1, T_2) combinations in our MTBs have significantly more negative $d(R, T_1, T_2)$ values than combinations involving the same R but T_1 and T_2 outside our MTBs, we used a procedure similar to checking the anti-correlations between miRNAs and targets in MTBs, but with the distribution of anti-correlation values replaced by these $d(R, T_1, T_2)$ values. The fraction of (R, T_1, T_2) combinations with a $d(R, T_1, T_2)$ value more negative than a threshold t was computed for each MTB, and the resulting distribution of fractions from all MTBs was compared to the background distribution with the same R's but T_1's and T_2's outside the MTBs using a one-sided Wilcoxon rank-sum test.

Based on the above calculations, we also collected x MTBs with the most negative $d(R, T_1, T_2)$ values and the x with most positive $d(R, T_1, T_2)$ values. We called the former set of MTBs the "top" MTBs and the latter set the "bottom" MTBs as the former set was expected to exhibit stronger expression buffering among the mRNAs in each of them. We then used our correlation pipeline to test if the mRNA-mRNA correlations were significantly stronger than other mRNA pairs targeted by the same miRNAs but were not in the same MTBs based on different values of the correlation threshold t. Considering the 4 input miRNA-target networks, 4 values of x (100, 200, 500 and 1000), and 7 values of t (0.1 to 0.7), we compared the p-values from the top MTBs and from the bottom MTBs under the $4 \times 4 \times 7 = 112$ parameter settings.

Workflow for functional enrichment analyses

We also setup a workflow for evaluating the functional relationships between the genes in same MTBs (Figure S1b). For each MTB, we collected the terms associated with each gene (mRNA) defined in Gene Ontology [8]. For each term, we then computed a p-value using hypergeometric test, to indicate the enrichment of the term in this set of genes as compared to the background set of all genes. To ensure robustness of our results, we also computed EASE scores as defined on the DAVID Web site [37], which can be considered a more stringent version of the p-values. The most significant p-value from each MTB was then collected to form a distribution, and it was compared to the most significant p-values from random sets of mRNAs of the same sizes of the MTBs. This comparison was quantified by a one-sided Wilcoxon rank-sum test, where a significant p-value would indicate that the genes in the MTBs were more enriched in same functional terms than random gene sets.

We also repeated the same analysis for sets of random mRNAs with a similar size and a similar level of co-expression as the MTBs.

Availability

The source code and compiled programs we used for our analyses are available at `http://yiplab.cse.cuhk.edu.hk/MTB/`.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KYY conceived the study. DK-SY and KYY designed the method. DK-SY prepared the data and ran all the tests. KYY, DK-SY and IKP interpreted the results. DK-SY, IKP and KYY wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

KYY is partially supported by the Hong Kong Research Grants Council Collaborative Research Fund CUHK8/CRF/11R. We thank Antonio Szeto and Kester Lee for helpful discussions.

Author details

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [2]School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [3]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [4]CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

References

1. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.
2. Panagiotis Alexiou, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. Lost in translation: An assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, 2009.
3. Panagiotis Alexiou, Thanasis Vergoulis, Martin Gleditzsch, George Prekas, Theodore Dalamagas, Molly Megraw, Ivo Grosse, Timos Sellis, and Artemis G. Hatzigeorgiou. miRGen 2.0: A database of microRNA genomic information and regulation. *Nucleic Acids Research*, 38:D137–D141, 2010.
4. Ditte C. Andersen, Charlotte H. Jensen, Mikael Schneider, Anne Yaël Nossent, Tilde Eskildsen, Jakob L. Hansen, Børge Teisner, and Søren P. Sheikh. Microrna-15a fine-tunes the level of delta-like 1 homolog (dlk1) in proliferating 3t3-l1 preadipocytes. *Experimental Cell Research*, 316:1681–1691, 2010.
5. Marianthi Kiriakidou andPeter T. Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes & Development*, 18(10):1165–1178, 2004.
6. Shay Artzi, Adam Kiezun, and Noam Shomron. miRNAminer: A tool for homologous microRNA gene search. *BMC Bioinformatics*, 9(1):39, 2008.
7. Aaron Arvey, Erik Larsson, Chris Sander, Christina S Leslie, and Debora S Marks. Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology*, 6(363), 2010.
8. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
9. David P. Bartel. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.
10. David P. Bartel. MicroRNAs: Target recognition and regulatory functions. *Cell*, 136:215–233, 2009.
11. Moussa Benhamed, Utz Herbig, Tao Ye, Anne Dejean, and Oliver Bischof. Senescence is an endogenous trigger for microrna-directed transcriptional gene silencing in human cells. *Nature Cell Biology*, 14(3):266–275, 2012.
12. Eric Bonnet, Marianthi Tatari, Anagha Joshi, Tom Michoel, Kathleen Marchal, Geert Berx, and Yves Van de Peer. Module network inference from a cancer gene expression data set identifies MicroRNA regulated modules. *PLOS ONE*, 5(e10162), 2010.
13. Christian Borgelt and Rudolf Kruse. Induction of association rules: Apriori implementation. In *Proceedings of The 15th Conference on Computational Statistics*, pages 395–400, 2002.
14. Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
15. Marina Capuano, Laura Iaffaldano, Nadia Tinto, Donatella Montanaro, Valentina Capobianco, Valentina Izzo, Francesca Tucci, Giancarlo Troncone, Luigi Greco, and Lucia Sacchetti. Microrna-449a overexpression, reduced notch1 signals and scarce goblet cells characterize the small intestine of celiac patients. *PLOS ONE*, 6(e29094), 2011.
16. Michelle A. Carmell, Zhenyu Xuan, Michael Q. Zhang, and Gregory J. Hannon. The Argonaute family: Tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes & Development*, 16(21):2733–2742, 2002.
17. Marcella Cesana, Davide Cacchiarelli, Ivano Legnini, Tiziana Santini, Olga Sthandier, Mauro Chinappi, Anna Tramontano, and Irene Bozzoni. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147:358–369, 2011.
18. Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.
19. Zhuoan Cheng, Shaobo Qiu, Lin Jiang, Anle Zhang, Wenjing Bao, Ping Liu, and Jianwen Liu. Mir-320a is downregulated in patients with myasthenia gravis and modulates inflammatory cytokines production by targeting mitogen-activated protein kinase 1. *Journal of Clinical Immunology*, 33:567–576, 2013.
20. Nicole Cloonan, Mellissa K Brown, Anita L Steptoe, Shivangi Wani, Wei Ling Chan, Alistair RR Forrest, Gabriel Kolle, Brian Gabrielli, and Sean M Grimmond. The mir-17-5p microrna is a key regulator of the g1/s phase cell cycle transition. *Genome Biology*, 9(R127), 2008.
21. Pasqualino de Antonellis, Chiara Medaglia, Emilio Cusanelli, Immacolata Andolfo, Lucia Liguori, Gennaro De Vita, Marianeve Carotenuto, Annamaria Bello, Fabio Formiggini, Aldo Galeone, Giuseppe De Rosa, Antonella Virgilio, Immacolata Scognamiglio, Manuela Sciro, Giuseppe Basso, Johannes H. Schulte, Giuseppe Cinalli,

Achille Iolascon, and Massimo Zollo. Mir-34a targeting of notch ligand delta-like 1 impairs cd15+/cd133+ tumor-propagating cells and supports neural differentiation in medulloblastoma. *PLOS ONE*, 6(e24584), 2011.

22. Tobias Dezulian, Michael Remmert, Javier F Palatnik, Detlef Weigel, and Daniel H Huson. Identification of plant microRNA homologs. *Bioinformatics*, 22(3):359–360, 2006.
23. Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Roder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.
24. Nan Du, Bai Wang, Bin Wu, and Yi Wang. Overlapping community detection in bipartite networks. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 176–179, 2008.
25. Margaret S. Ebert and Philip A. Sharp. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149:515–524, 2012.
26. Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(R1), 2003.
27. Ling Fang, William W. Du, Weining Yang, Zina Jeyapalan Rutnam, Chun Peng, Haoran Li, Yunxia Q. O'Malley, Ryan W. Askeland, Sonia Sugg, Mingyao Liu, Tanvi Mehta, Zhaoqun Deng, and Burton B. Yang. Mir-93 enhances angiogenesis and metastasis by targeting LATS2. *Cell Cycle*, 11(23):4352–4365, 2012.
28. José Manuel Franco-Zorrilla, Adrián Valli, Marco Todesco, Isabel Mateos, María Isabel Puga, Ignacio Rubio-Somoza, Antonio Leyva, Detlef Weigel, Juan Antonio García, and Javier Paz-Ares. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8):1033–1037, 2007.
29. Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, 26(4):407–415, 2008.
30. Mayuko Furuta, Ken ichi Kozaki, Kousuke Tanimoto, Shinji Tanaka, Shigeki Arii, Teppei Shimamura, Atsushi Niida, Satoru Miyano, and Johji Inazawa. The tumor-suppressive mir-497-195 cluster targets multiple cell-cycle regulators in hepatocellular carcinoma. *PLOS ONE*, 8(e60155), 2013.
31. Vincenzo Alessandro Gennarino, Giovanni D'Angelo, Gopuraja Dharmalingam, Serena Fernandez, Giorgio Russolillo, Remo Sanges, Margherita Mutarelli, Vincenzo Belcastro, Andrea Ballabio, Pasquale Verde, Marco Sardiello, and Sandro Banfi. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Research*, 22(6):1163–1172, 2012.
32. Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio, and Sandro Banfi. MicroRNA target prediction by expression analysis of host genes. *Genome Research*, 19(3):481–490, 2009.
33. Jan Gorodkin and Ivo L Hofacker. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Computational Biology*, 7(8):e1002100, 2011.
34. Kristian A. Gray, Louise C. Daugherty, Susan M. Gordon, Ruth L. Seal, Mathew W. Wright, and Elspeth A. Bruford. Genenames.org: The HGNC resources in 2013. *Nucleic Acids Research*, 41:D545–D552, 2013.
35. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11:10–18, 2009.
36. Han Han, Dan Sun, Wenjuan Li, Hongxing Shen, Yahui Zhu, Chen Li, Yuxing Chen, Longfeng Lu, Wenhua Li, Jinxiang Zhang, Yuan Tian, and Youjun Li. A c-myc-microrna functional feedback loop affects hepatocarcinogenesis. *Hepatology*, 57:2378–2389, 2013.
37. Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008.
38. Jim C Huang, Tomas Babak, Timothy W Corson, Gordon Chua, Sofia Khan, Brenda L Gallie, Timothy R Hughes, Benjamin J Blencowe, Brendan J Frey, and Quaid D Morris. Using expression profiling data to identify human MicroRNA targets. *Nature Methods*, 4(12):1045–1049, 2007.
39. Bino John, Anton J. Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S. Marks. Human microRNA targets. *PLoS Biology*, 2(e363), 2004.
40. Ila Joshi, Lisa M. Minter, Janice Telfer, Renée M. Demarest, Anthony J. Capobianco, Jon C. Aster, Piotr Sicinski, Abdul Fauq, Todd E. Golde, and Barbara A. Osborne. Notch signaling mediates g1/s cell-cycle progression in t cells via cyclin d3 and its dependent kinases. *Blood*, 113(8):1689–1698, 2008.
41. Je-Gun Joung and Zhangjun Fei. Identification of MicroRNA regulatory modules in arabidopsis via a probabilistic graphical model. *Bioinformatics*, 25(3):387–393, 2009.
42. Je-Gun Joung, Kyu-Baek Hwang, Jin-Wu Nam, Soo-Jin Kim, and Byoung-Tak Zhang. Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics*, 23(9):1141–1147, 2007.
43. Florian A. Karreth, Yvonne Tay, Daniele Perna, Ugo Ala, Shen Mynn Tan, Alistair G. Rust, Gina DeNicola, Kaitlyn A. Webster, Dror Weiss, Pedro A. Perez-Mancera, Michael Krauthammer, Ruth Halaban, Paolo Provero, David J. Adams, David A. Tuveson, and Pier Paolo Pandolfi. In vivo identification of tumor-suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell*, 147:382–395, 2011.

44. Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–1284, 2007.
45. Madhu S. Kumar, Elena Armenteros-Monterroso, Philip East, Probir Chakravorty, Nik Matthews, Monte M. Winslow, and Julian Downward. HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature*, 505(7482):212–217, 2014.
46. Yoong Sun Lee and Anindya Dutta. The tumor suppressor microrna let-7 represses the HMGA2 oncogene. *Genes & Development*, 21(9):1025–1030, 2007.
47. Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.
48. Marta Lionetti, Marta Biasiolo, Luca Agnelli, Katia Todoerti, Laura Mosca, Sonia Fabris, Gabriele Sales, Giorgio Lambertenghi Deliliers, Silvio Bicciato, Luigia Lombardi, Stefania Bortoluzzi, and Antonino Neri. Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. *Blood*, 114(25):e20–e26, 2009.
49. Bing Liu, Jiuyong Li, and Murray J. Cairns. Identifying miRNAs, targets and functions. *Briefings in Bioinformatics*, 2012. Advance Access.
50. Bing Liu, Jiuyong Li, and Anna Tsykin. Discovery of functional miRNA-mRNA regulatory modules with computational methods. *Journal of Biomedical Informatics*, 42:685–691, 2009.
51. Bing Liu, Jiuyong Li, Anna Tsykin, Lin Liu, Arti B Gaur, and Gregory J Goodall. Exploring complex miRNA-mRNA interactions with bayesian networks by splitting-averaging strategy. *BMC Bioinformatics*, 10(408), 2009.
52. Bing Liu, Lin Liu, Anna Tsykin, Gregory J. Goodall, Jeffrey E. Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA-mRNA regulatory modules with correspondence latent Dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, 2010.
53. Dong Lu, Matthew P. A. Davis, Cei Abreu-Goodger, Wei Wang, Lia S. Campos, Julia Siede, Elena Vigorito, William C. Skarnes, Ian Dunham, Anton J. Enright, and Pentao Liu. Mir-25 regulates wwp2 and fbxw7 and promotes reprogramming of mouse fibroblast cells to ipscs. *PLOS ONE*, 7(e40938), 2012.
54. Yiming Lu, Yang Zhou, Wubin Qu, Minghua Deng, and Chenggang Zhang. A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, 27(17):2406–2413, 2011.
55. Christine Mayr, Michael T. Hemann, and David P. Bartel. Disrupting the pairing between let-7 and hmga2 enhances oncogenic transformation. *Science*, 315(5818):1576–1579, 2008.
56. Hyeyoung Min and Sungroh Yoon. Got target?: Computational methods for microRNA target prediction and their extension. *Experimental & Molecular Medicine*, 42(4):233–244, 2010.
57. Juan Nunez-Iglesias, Chun-Chi Liu, Todd E. Morgan, Caleb E. Finch, and Xianghong Jasmine Zhou. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PLOS ONE*, 5(e8898), 2010.
58. Takeshi Obayashi, Yasunobu Okamura1, Satoshi Ito, Shu Tadaka, Ikuko N. Motoike, and Kengo Kinoshita. COXPRESdb: A database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, 41:D1014–D1020, 2013.
59. René Peeters. The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, 131(3):651–654, 2003.
60. Xinxia Peng, Yu Li, Kathie-Anne Walters, Elizabeth R Rosenzweig, Sharon L Lederer, Lauri D Aicher, Sean Proll, and Michael G Katze. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10(373), 2009.
61. Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J. Havenman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–1038, 2010.
62. William Ritchie, Stephane Flamant, and John E. J. Rasko. mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics*, 26(2):223–227, 2010.
63. Ignacio Rubio-Somoza, Detlef Weigel, José-Manuel Franco-Zorrilla, Juan Antonio García, and Javier Paz-Ares. ceRNAs: miRNA target mimic mimics. *Cell*, 147:1431–1432, 2011.
64. Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: The Rosetta stone of a hidden RNA language? *Cell*, 146:353–358, 2011.
65. Udainiya Savita and Devarajan Karunagaran. Microrna-106b-25 cluster targets β-trcp2, increases the expression of snail and enhances cell migration and invasion in h1299 (non small cell lung cancer) cells. *Biochemical and Biophysical Research Communications*, 434:841–847, 2013.
66. Pavel Sumazin, Xuerui Yang, Hua-Sheng Chiu, Wei-Jen Chung, Archana Iyer, David Llobet-Navas, Presha Rajbhandari, Mukesh Bansal, Paolo Guarnieri, Jose Silva, and Andrea Califano. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147:370–381, 2011.
67. Fang Sun, Hanjiang Fu, Qin Liu, Yi Tie, Jie Zhu, Ruiyun Xing, Zhixian Sun, and Xiaofei Zheng. Downregulation of CCND1 and CDK6 by mir-34a induces cell cycle arrest. *FEBS Letters*, 582:1564–1568, 2008.
68. Yvonne Tay, Lev Kats, Leonardo Salmena, Dror Weiss, Shen Mynn Tan, Ugo Ala, Florian Karreth, Laura Poliseno, Paolo Provero, Ferdinando Di Cunto, Judy Lieberman, Isidore Rigoutsos, and Pier Paolo Pandolfi. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, 147:344–357, 2011.
69. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.
70. Marshall Thomas, Judy Lieberman, and Ashish Lal. Desperately seeking microRNA targets. *Nature Structural & Molecular Biology*, 17(10):1169–1174, 2010.
71. Dang H Tran, Kenji Satou, and Tu B Ho. Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, 9(S5), 2008.

72. Marco Antonio Valencia-Sanchez, Jidong Liu, Gregory J. Hannon, and Roy Parker. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & Development*, 20(5):515–524, 2006.
73. Thanasis Vergoulis, Ioannis S. Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. TarBase 6.0: Capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*, 40:D222–D229, 2012.
74. Ning Wu, Eric Sulpice, Patricia Obeid, Sami Benzina, Frédérique Kermarrec, Stéphanie Combe, and Xavier Gidrol. The mir-17 family links p63 protein to MAPK signaling to promote the onset of human keratinocyte differentiation. *PLOS ONE*, 7(e45761), 2012.
75. Sungroh Yoon and Giovanni De Micheli. Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21(Suppl. 2):ii93–ii100, 2005.
76. Changcun Zhang, Ren Mo, Binde Yin, Libin Zhou, Yongchao Liu, and Jie Fan. Tumor suppressor microrna-34a inhibits cell proliferation by targeting notch1 in renal cell carcinoma. *Oncology Letters*, 7(5):1689–1694, 2014.
77. Shihua Zhang, Qingjiao Li, Juan Liu, and Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27:i401–i409, 2011.

Figures

Figure 1 Summary on the eight types of MTB. The different types are named according to (1) whether the defining submatrix of an MTB can only contain 1's (restrictive, R) or is allowed to contain 0's (loose, L), and (2) whether the miRNAs (mi) are not allowed to have extra targets, the mRNAs (m) are not allowed to be targeted by extra miRNAs, or the general case (gen) that both are allowed. In the formal mathematical definitions, r and c correspond to the sets of row and column indices defining an MTB, where each row corresponds to an mRNA and each column corresponds to a miRNA. a_{ij} is the value at row i and column j of the adjacency matrix. In the matrix representation, an example is shown for each type of MTB, where the sub-matrix enclosed by the rectangle corresponds to the example MTB. For visualization purpose, they are drawn to occupy consecutive rows and columns, but this is not required in the actual definitions of the MTB types. Values of irrelevant cells, i.e., those not on the rows and columns defining the MTB, are omitted. In the graphical representation, the red nodes are the mRNAs and miRNAs defining the example MTB, red lines are edges between them, and blue lines are edges connecting them to mRNAs or miRNAs outside the MTB. In the formulas for showing the number of MTBs of each type, R and C are the full sets of miRNAs and genes in the miRNA-target network, respectively, and $|R|$ and $|C|$ are their sizes. The function $connected(i, j)$ means the nodes in the graphical representation corresponding to row i and column j are connected, which can be formally defined as $\exists i_1, i_2, ..., i_k \in r, j_1, j_2, ..., j_k \in s, s.t. a_{ij_1} = a_{i_1 j_1} = a_{i_1 j_2} = a_{i_2 j_2} = ... = a_{i_k j_k} = a_{i_k j} = 1$.

Figure 2 Schematic figure explaining the workflow for testing the statistical significance of expression anti-correlation of miRNAs and mRNAs in the same MTBs. (a) An example MTB (submatrix in cells with red borders), corresponding random miRNA-mRNA pairs (cells with green borders) and other targets of the miRNAs that define the MTB (cells with blue borders). (b) The expression levels of the miRNAs and mRNAs from multiple cell lines were collected. The expression correlation between each miRNA-mRNA pair in the MTB was computed. Similar correlation values were also computed for the two background sets (not shown). (c) For each MTB and corresponding background sets, the computed correlation values were recorded. (d) These correlation values were compared against a threshold t (-0.1 for example), and the fraction of correlation values more negative than t was computed. The vector of these fractional values from the MTBs was then compared to the vectors from the two background sets by a statistical test.

Figure 3 Statistical significance of the negative correlations between the expression levels of miRNAs and mRNAs in the same MTBs. The p-values were computed based on the expressed union sets with TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set as compared to a random background sampled from all expressed mRNAs and miRNAs; and (c) the high-confidence set and (d) the high-coverage set as compared to a background consisting of miRNA-mRNA pairs with interactions in the input network but are not in same MTBs. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program. MTB types with no identified MTBs are omitted.

Tables

Additional Files

Additional file 1 — Supplementary materials
This file contains supplementary methods and supplementary figures.

Figure 4 Example fractions of miRNA-mRNA pairs satisfying the correlation threshold. Type Lgen MTBs were identified from the high-confidence expression union set of miRNA-target interactions with TarBase inputs. For each MTB, the fraction of miRNA-mRNA pairs with expression correlation more negative than t=-0.2 among the 10 cell lines was computed. The distribution of these fractional values is shown by a histogram. Also shown are the distributions of fraction values for random groups of miRNAs and mRNAs of the same sizes as the MTBs, and for groups of miRNAs and their target mRNAs not within same MTBs.

Figure 5 Statistics of the MTBs identified from the high-confidence integrated expressed union set with TarBase interactions. For each type of MTB, the average number of mRNAs per MTB, average number of miRNAs per MTB and the number of MTBs identified by our algorithm are shown.

Figure 6 Comparing MTBs with the miRNA regulatory modules (MRMs) identified by the Yoon and De Micheli method. The p-values were computed by comparing the fractions of correlations more negative than the threshold t of miRNA-mRNA pairs within MTBs, as compared to those from the MRMs. In the figure, 1E-16 represents the smallest p-value that could be outputted by our program.

Figure 7 Statistical significance of the positive Pearson correlations between the expression levels of mRNAs in the same type Lgen MTBs. The p-values were computed by comparing the fractions of correlations more positive than the threshold t of the mRNA pairs within MTBs, as compared to (a) a background consisting of random mRNA pairs, or (b) a background consisting of mRNA pairs targeted by a common miRNA in the input interaction network but not in same MTBs. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program.

Figure 8 Statistical significance of the extra information provided by an mRNA in our MTBs in explaining the relationship between a miRNA and another mRNA in the same MTB. The p-values were computed by comparing the fractions of $d(R, T1, T2)$ (see Materials and Methods) more negative than the threshold t of the $(R, T1, T2)$ combinations within MTBs, as compared to a background consisting of $(R, T1, T2)$ combinations where R comes from the same MTBs but $T1$ and $T2$ do not. In the figure, 1E-16 represents the smallest p-value that could be outputted by our program.

Figure 9 Statistical significance of the functional enrichment scores of the genes from same type Lgen MTBs. The p-values were computed based on the expressed union sets with TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program.

Figure 10 Comparison of the functional enrichment scores of the mRNAs in same MTBs with random co-expressed mRNAs. In the figure, 1E-16 represents the smallest p-value that could be outputted by our program.

Table 1 Summary statistics of the different datasets used in our study. The integrated datasets involve data from all the prediction sets with or without the experimentally validated miRNA-target pairs. Among them, the expressed union sets were formed by considering only the expressed miRNAs and mRNAs in the corresponding union sets. We used these four expressed union sets (with or without TarBase, high confidence or high coverage) in our analyses.

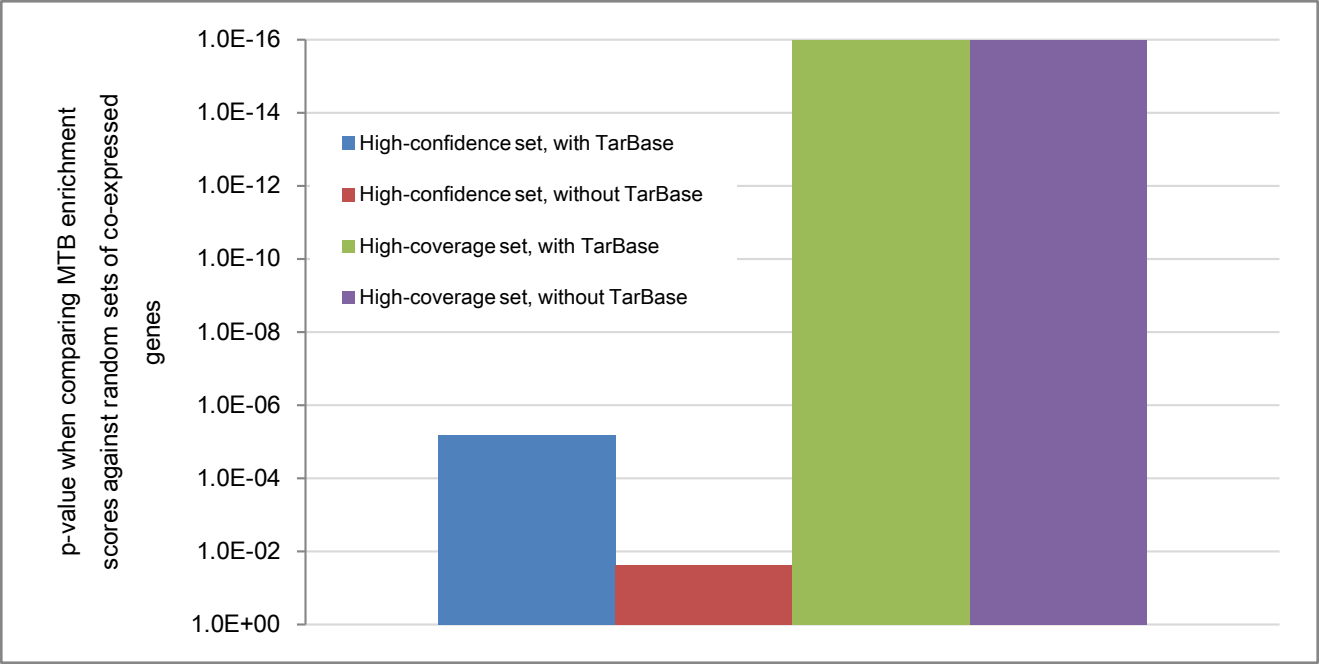
		High confidence			High coverage		
Dataset	Type	miRNAs	mRNAs	interactions	miRNAs	mRNAs	interactions
TarBase [73]	Validated	202	2,315	9,569	202	2,315	9,569
miRanda [39]	Predicted	409	567	1,000	856	2,633	5,000
miRGen [3]	Predicted	503	37	1,041	828	103	5,230
PicTar [5]	Predicted	91	438	1,000	164	1,541	5,000
PITA [44]	Predicted	290	760	1,000	582	2,708	5,000
TargetScan [47]	Predicted	29	214	1,000	40	948	5,000
Union (without TarBase)	Integrated	926	1,818	4,983	1,505	6,063	24,553
Expressed union (without TarBase)	Integrated	163	448	701	240	2,034	4,337
Union (with TarBase)	Integrated	1,063	3,711	14,548	1,631	7,208	34,111
Expressed union (with TarBase)	Integrated	181	605	1,020	256	2,188	4,653

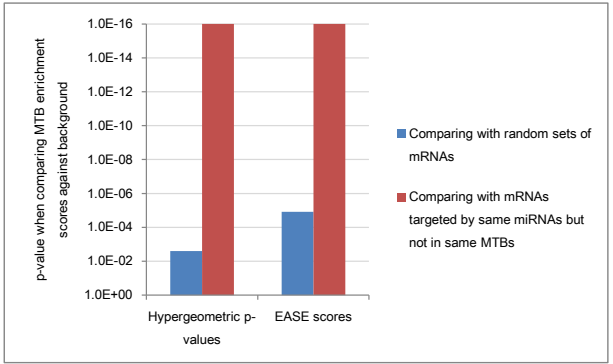
Table 2 Pearson correlation between $d(R, T_1|T2)$ and $f(T_1, T_2)$ for all combinations of miRNA R, mRNAs T_1 and T_2 from the same type Lgen MTBs.

Pearson correlation	High-confidence miRNA target predictions	High-coverage miRNA target predictions
With TarBase miRNA-target pairs	-0.32 (p<1E-16)	-0.30 (p<1E-16)
Without TarBase miRNA-target pairs	-0.19 (p=4.1E-5)	-0.30 (p<1E-16)

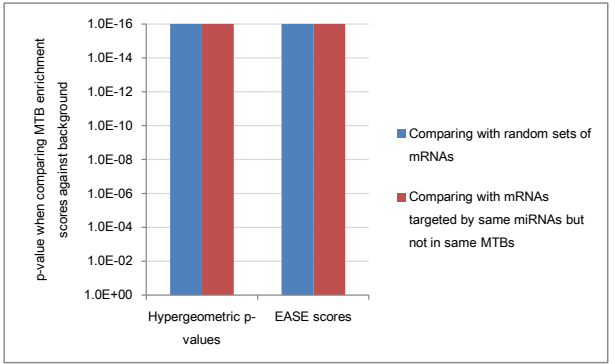
Table 3 Illustrating examples of using MTBs to functionally annotate miRNAs. Each row corresponds to one example MTB.

Enriched GO term			
MTB ID	Total number of mRNAs	mRNAs annotated with the GO term	miRNAs
GO:0004674 Protein serine/threonine kinase activity			
1	4	AAK1, MAPK1, PDK3	miR-17, miR-20b, miR-93, miR-320a
GO:0000792 Heterochromatin			
2	4	CBX5, HMGA2, RNF20	let-7b, let-7c, let-7d, let-7e, let-7g, miR-185
GO:0007219 Notch signaling pathway			
3	3	CDK6, TNRC6B	miR-15a, miR-34a, miR-449a, miR-497
GO:0043161 Proteasome-mediated ubiquitin-dependent protein catabolic process			
4	19	EDEM1, UBE2W, WWP2	miR-25, miR-32, miR-363

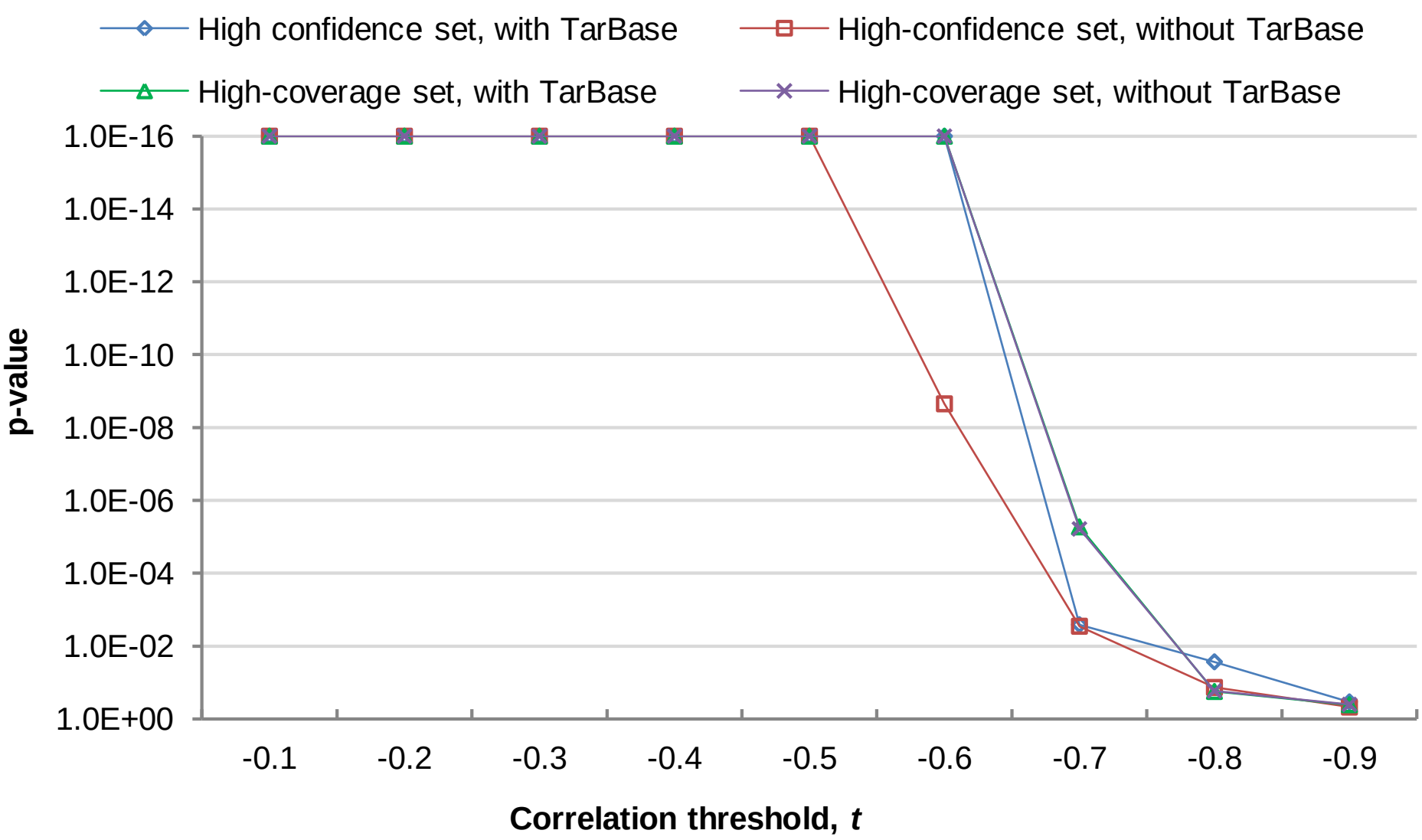

p-value when comparing MTB enrichment scores against random sets of co-expressed genes
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
High-confidence set, with TarBase
High-confidence set, without TarBase
High-coverage set, with TarBase
High-coverage set, without TarBase

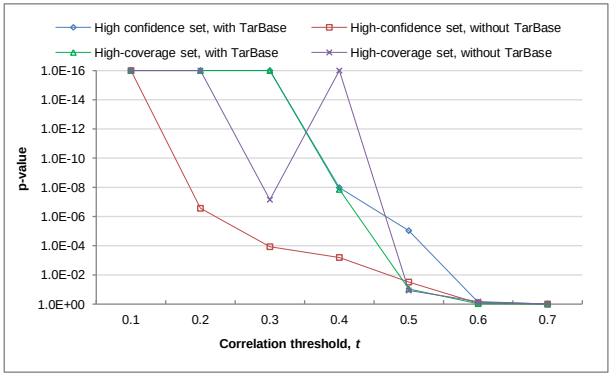

p-value when comparing MTB enrichment scores against background
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
Hypergeometric p-values
EASE scores
Comparing with random sets of mRNAs
Comparing with mRNAs targeted by same miRNAs but not in same MTBs

(a)

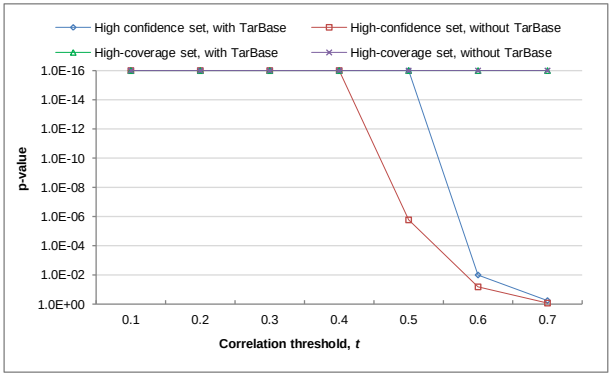

p-value when comparing MTB enrichment scores against background
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
Hypergeometric p-values
EASE scores
Comparing with random sets of mRNAs
Comparing with mRNAs targeted by same miRNAs but not in same MTBs

(b)

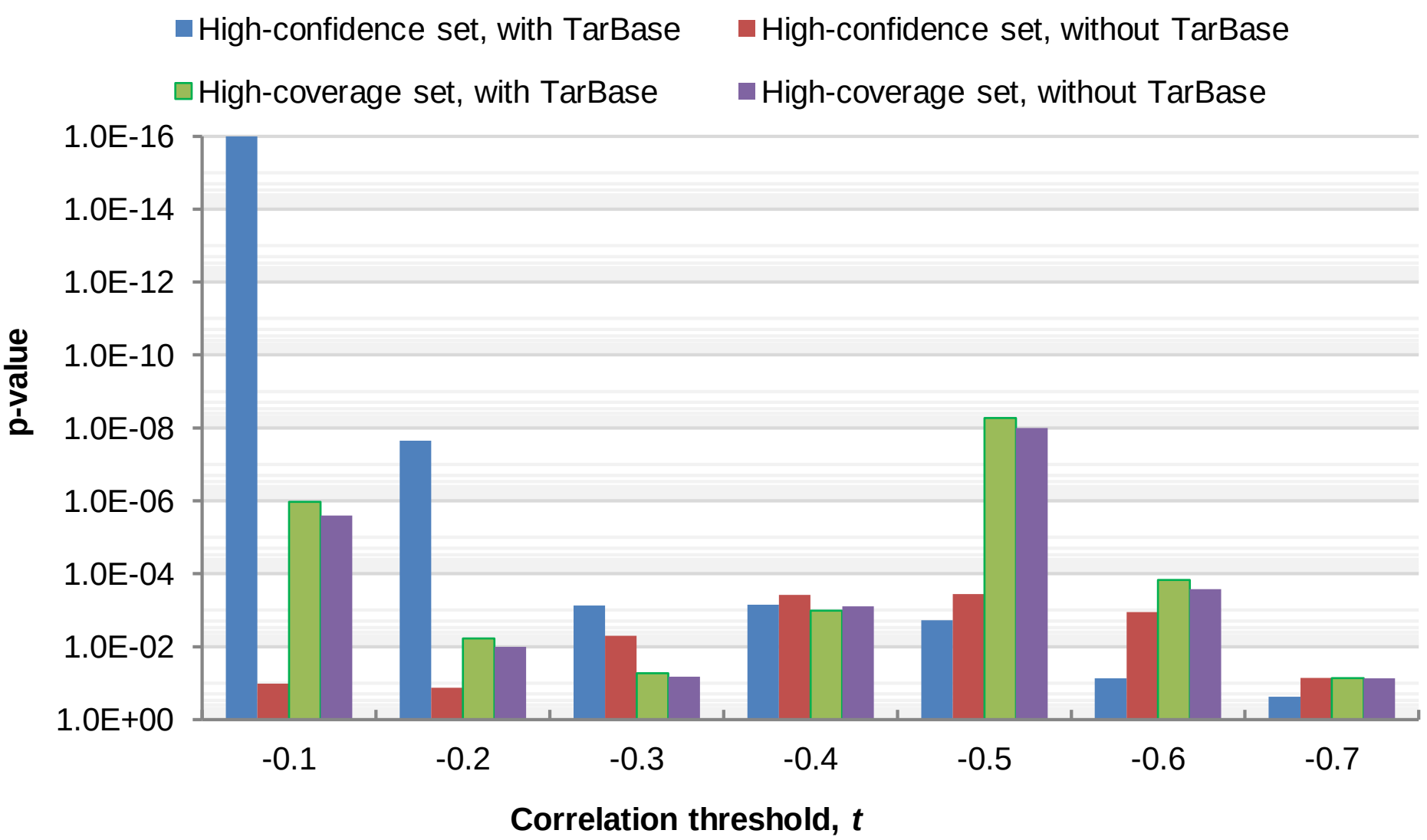

High confidence set, with TarBase
High-confidence set, without TarBase
High-coverage set, with TarBase
High-coverage set, without TarBase
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
p-value
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
Correlation threshold, t

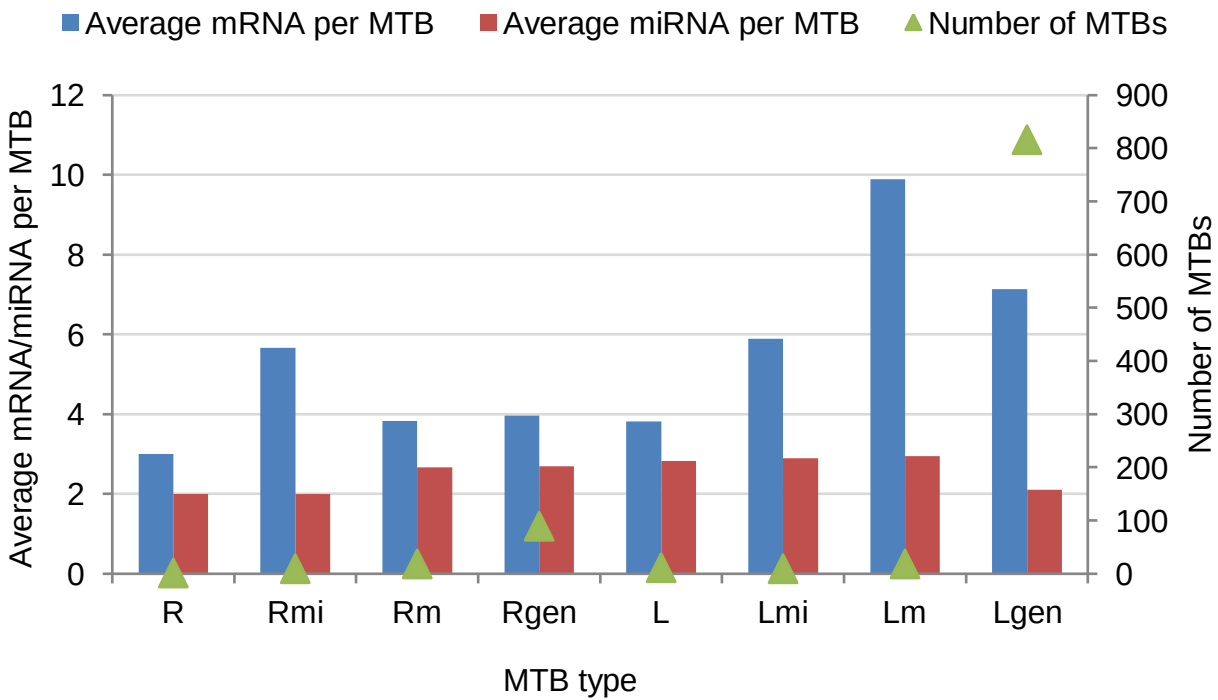

High confidence set, with TarBase
High-confidence set, without TarBase
High-coverage set, with TarBase
High-coverage set, without TarBase
p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Correlation threshold, t

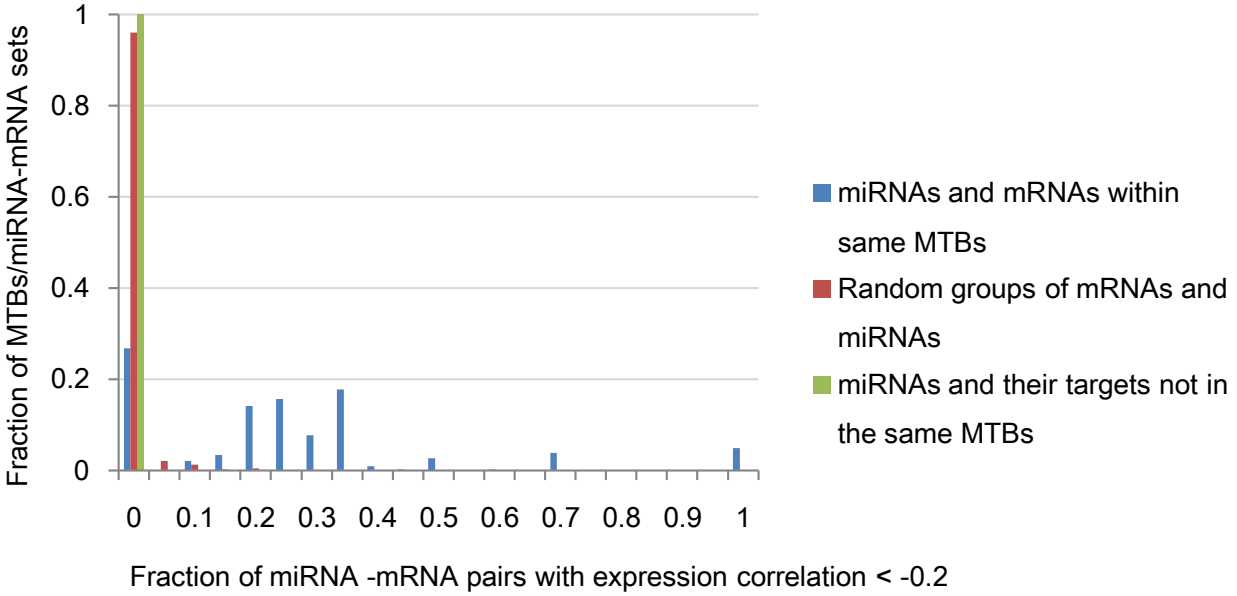
(a)

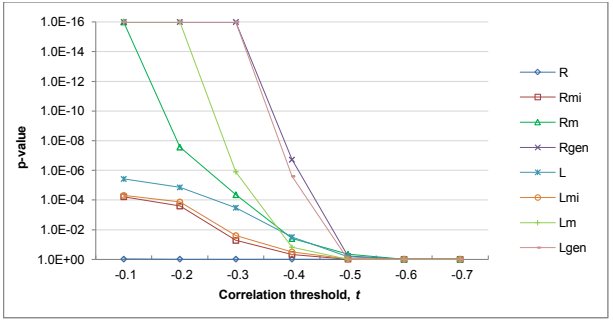

High confidence set, with TarBase
High-confidence set, without TarBase
High-coverage set, with TarBase
High-coverage set, without TarBase
p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Correlation threshold, t

(b)

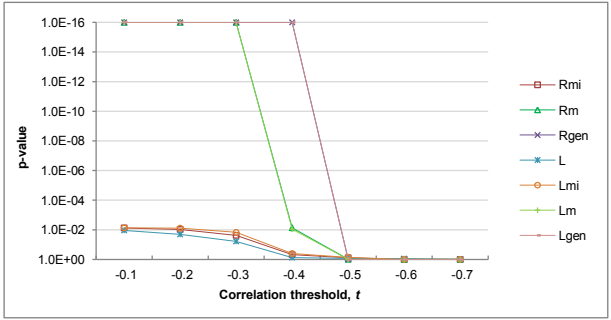

High-confidence set, with TarBase
High-confidence set, without TarBase
High-coverage set, with TarBase
High-coverage set, without TarBase
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
p-value
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Correlation threshold, t


Average mRNA per MTB
Average miRNA per MTB
Number of MTBs
Average mRNA/miRNA per MTB
Number of MTBs
0
2
4
6
8
10
12
0
100
200
300
400
500
600
700
800
900
R
Rmi
Rm
Rgen
L
Lmi
Lm
Lgen
MTB type

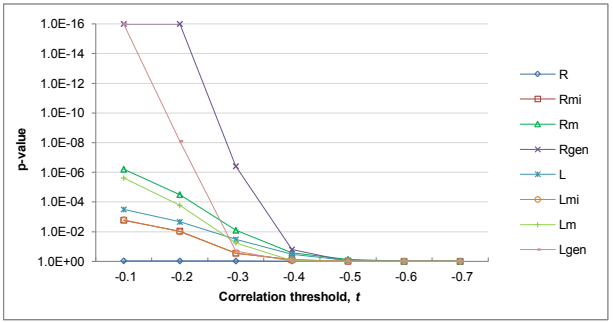

Fraction of MTBs/miRNA-mRNA sets
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Fraction of miRNA -mRNA pairs with expression correlation < -0.2
miRNAs and mRNAs within same MTBs
Random groups of mRNAs and miRNAs
miRNAs and their targets not in the same MTBs


p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Correlation threshold, t
R
Rmi
Rm
Rgen
L
Lmi
Lm
Lgen

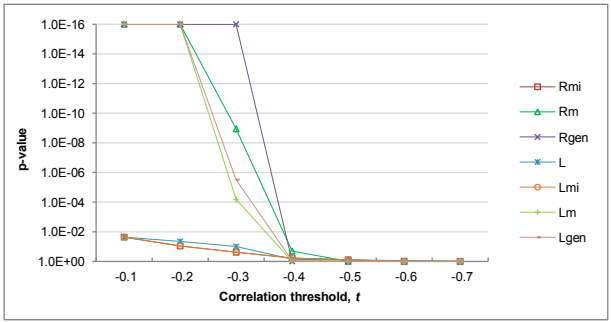
(a)


p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Correlation threshold, t
Rmi
Rm
Rgen
L
Lmi
Lm
Lgen

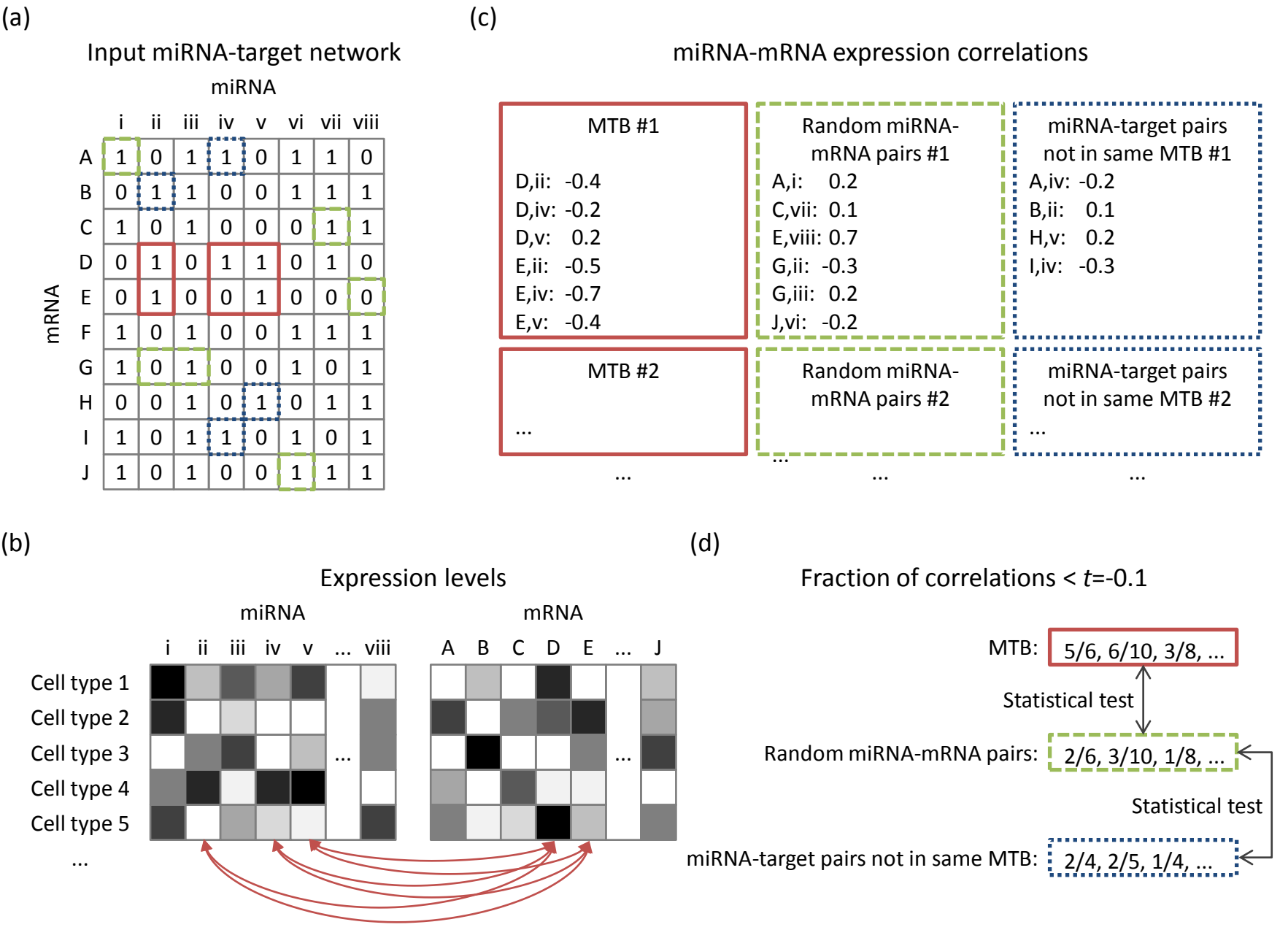
(b)

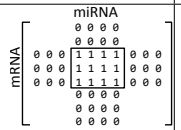
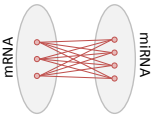
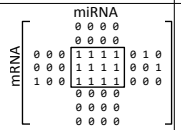
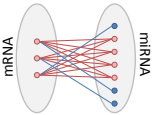
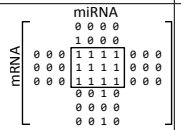
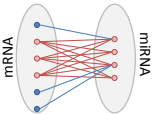
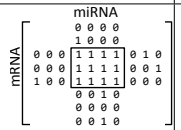
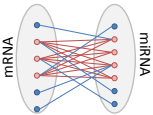
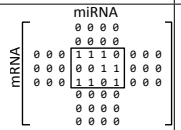
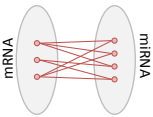
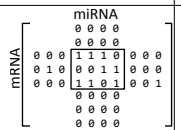
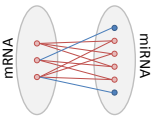
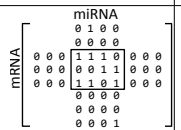
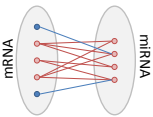
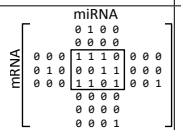
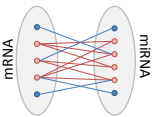

p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Correlation threshold, t
R
Rmi
Rm
Rgen
L
Lmi
Lm
Lgen

(c)


p-value
1.0E-16
1.0E-14
1.0E-12
1.0E-10
1.0E-08
1.0E-06
1.0E-04
1.0E-02
1.0E+00
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Correlation threshold, t
Rmi
Rm
Rgen
L
Lmi
Lm
Lgen

(d)


(a)
Input miRNA-target network
miRNA
i ii iii iv v vi vii viii
mRNA
A 1 0 1 1 0 1 1 0
B 0 1 1 0 0 1 1 1
C 1 0 1 0 0 0 1 1
D 0 1 0 1 1 0 1 0
E 0 1 0 0 1 0 0 0
F 1 0 1 0 0 1 1 1
G 1 0 1 0 0 1 0 1
H 0 0 1 0 1 0 1 1
I 1 0 1 1 0 1 0 1
J 1 0 1 0 0 1 1 1
(c)
miRNA-mRNA expression correlations
MTB #1
D,ii: -0.4
D,iv: -0.2
D,v: 0.2
E,ii: -0.5
E,iv: -0.7
E,v: -0.4
MTB #2
...
Random miRNA-mRNA pairs #1
A,i: 0.2
C,vii: 0.1
E,viii: 0.7
G,ii: -0.3
G,iii: 0.2
J,vi: -0.2
Random miRNA-mRNA pairs #2
...
miRNA-target pairs not in same MTB #1
A,iv: -0.2
B,ii: 0.1
H,v: 0.2
I,iv: -0.3
miRNA-target pairs not in same MTB #2
...
(b)
Expression levels
miRNA
i ii iii iv v ... viii
mRNA
A B C D E ... J
Cell type 1
Cell type 2
Cell type 3
Cell type 4
Cell type 5
...
(d)
Fraction of correlations < t=-0.1
MTB: 5/6, 6/10, 3/8, ...
Statistical test
Random miRNA-mRNA pairs: 2/6, 3/10, 1/8, ...
Statistical test
miRNA-target pairs not in same MTB: 2/4, 2/5, 1/4, ...

MTB type	Mathematical definition	Matrix representation	Graphical representation	Maximum number of MTBs	Maximum number of maximal MTBs	MTBs returned by our algorithm
R	$(r,c) : \forall i \in r, j \in c, p \notin r, q \notin c, a_{ij} = 1 \wedge a_{pj} = 0 \wedge a_{iq} = 0$			$min(\|R\|, \|C\|)$	$min(\|R\|, \|C\|)$	All
Rmi	$(r,c) : \forall i \in r, j \in c, p \notin r, a_{ij} = 1 \wedge a_{pj} = 0$			$2^{\|C\|}$	$\|C\|$	All maximal
Rm	$(r,c) : \forall i \in r, j \in c, q \notin c, a_{ij} = 1 \wedge a_{iq} = 0$			$2^{\|R\|}$	$\|R\|$	All maximal
Rgen	$(r,c) : \forall i \in r, j \in c, a_{ij} = 1$			$2^{min(\|R\|,\|C\|)}$	$2^{min(\|R\|,\|C\|)}$	All maximal
L	$(r,c) : \forall i \in r, j \in c, p \notin r, q \notin c, connected(i,j) \wedge a_{pj} = 0 \wedge a_{iq} = 0$			$min(\|R\|, \|C\|)$	$min(\|R\|, \|C\|)$	All
Lmi	$(r,c) : \forall i \in r, j \in c, p \notin r, q \notin c, connected(i,j) \wedge a_{pj} = 0$			$2^{\|C\|}$	$\|C\|$	Top-scored
Lm	$(r,c) : \forall i \in r, j \in c, q \notin c, connected(i,j) \wedge a_{iq} = 0$			$2^{\|R\|}$	$\|R\|$	Top-scored
Lgen	$(r,c) : \forall i \in r, j \in c, connected(i,j)$			$2^{min(\|R\|,\|C\|)}$	$min(\|R\|, \|C\|)$	Top-scored

Supplementary materials

Supplementary methods

Pseudocode of the MTB identification algorithms

Algorithm 1 Algorithm for identifying all type R MTBs from a miRNA-target interaction network

function TypeRMTB(R, C, A) ▷ R: the set of all row indices, each representing an mRNA ▷ C: the set of all column indices, each representing a miRNA ▷ A: the adjacency matrix of the miRNA-mRNA network
  Define $MapR$ as a (bit string → row index set) hash map, initialized to an empty map
  for each $i \in R$ **do** ▷ For each mRNA i
    $MapR$.put($A[i, .]$, i), where $A[i, .]$ represents the i-th row of matrix A
    ▷ Use the miRNAs targeting i as the key to get the existing set or create a new set in $MapR$, then add i to this set
  end for
  Define $MapC$ as a (bit string → column index set) hash map, initialized to an empty map
  for each $j \in C$ **do** ▷ For each miRNA j
    $MapC$.put($A[., j]$, j), where $A[., j]$ represents the j-th column of matrix A
    ▷ Use j's mRNA targets as the key to get the existing set or create a new set in $MapC$, then add j to this set
  end for
  Define $MTBs$ as the list of MTBs, initialized to an empty list
  for each $c \in MapR$.keys **do** ▷ For each key c of $MapR$, i.e., each group signature
    Define r as a bit vector corresponding to the row indices of $MapR$.get(c)
    ▷ r is the members of the group, where each mRNA in r is targeted by only and all of the miRNAs in c
    if r is a key of $MapC$ and $MapC$.get(r) $== c$ **then** ▷ The miRNAs in c also target only and all of the mRNAs in r
      $MTBs$.add((r, c)) ▷ The mRNAs and the miRNAs form an MTB
    end if
  end for
  Return $MTBs$
end function

Algorithm 2 Algorithm for identifying all maximal type Rmi MTBs from a miRNA-target interaction network

function TypeRmiMTB(R, C, A) ▷ R: the set of all row indices, each representing an mRNA
▷ C: the set of all column indices, each representing a miRNA
▷ A: the adjacency matrix of the miRNA-mRNA network
    Define Map as a (bit string → column index set) hash map, initialized to an empty map
    for each $j \in C$ **do** ▷ For each miRNA j
        Map.put($A[., j]$, j), where $A[., j]$ represents the j-th column of matrix A
        ▷ Use j's mRNA targets as the key to get the existing set or create a new set in Map, then add j to this set
    end for
    Define $MTBs$ as the list of MTBs, initialized to an empty list
    for each $r \in Map$.keys **do** ▷ For each key r of Map, i.e., each group signature
        $MTBs$.add((r, Map.get(r))) ▷ The group signature (mRNAs) and the group members (miRNAs) form an MTB
    end for
    Return $MTBs$
end function

Algorithm 3 Algorithm for identifying all maximal type Rm MTBs from a miRNA-target interaction network

function TypeRmMTB(R, C, A) ▷ R: the set of all row indices, each representing an mRNA
▷ C: the set of all column indices, each representing a miRNA
▷ A: the adjacency matrix of the miRNA-mRNA network
    Define Map as a (bit string → row index set) hash map, initialized to an empty map
    for each $i \in R$ **do** ▷ For each mRNA i
        Map.put($A[i, .]$, i), where $A[i, .]$ represents the i-th row of matrix A
        ▷ Use the miRNAs targeting i as the key to get the existing set or create a new set in Map, then add i to this set
    end for
    Define $MTBs$ as the list of MTBs, initialized to an empty list
    for each $c \in Map$.keys **do** ▷ For each key c of Map, i.e., each group signature
        $MTBs$.add((Map.get(c), c)) ▷ The group members (mRNAs) and the group signature (miRNAs) form an MTB
    end for
    Return $MTBs$
end function

Algorithm 4 Algorithm for identifying all maximal type Rgen MTBs from a miRNA-target interaction network

function TypeRgenMTB(R, C, A) ▷ R: the set of all row indices, each representing an mRNA ▷ C: the set of all column indices, each representing a miRNA ▷ A: the adjacency matrix of the miRNA-mRNA network
    Preprocess A to group all rows with identical signatures together and all columns with identical signatures together, using hash maps as in the algorithms for type R. Define the resulting matrix with no identical rows or columns as A', and its rows and columns as R' and C'
    Define $MTBs$, $MTBsCurr$ and $MTBsNext$ as the lists of all MTBs discovered, MTBs for the current iteration and MTBs for the next iteration, respectively, all initialized to an empty list
    for each $j \in C'$ **do** ▷ For each column of the pre-processed adjacency matrix
        $MTBsNext$.add(($A'[., j]$, j)), where $A'[., j]$ is the j-th column of A' ▷ (A'[.,j], j) is an MTB, but may or may not be maximal
    end for
    while $MTBNext$ is not empty **do** ▷ Beginning of the k-th iteration, where k starts with 1
        $MTBs$.addall($MTBsNext$) ▷ Add all MTBs in $MTBsNext$ to $MTBs$
        $MTBsCurr := MTBsNext$
        $MTBsNext := \emptyset$ ▷ All MTBs newly discovered in the previous iteration are to be worked on in this iteration
        for each pair of MTBs (r_1, c_1) and (r_2, c_2) in $MTBsCurr$, such that the $k-1$ smallest indexes of both sets are the same **do** ▷ Trying to merge these two MTBs to form a new MTB
            if $r_1 \cup r_2 \neq \emptyset$ **then** ▷ The two MTBs have some common rows
                Define $M = (r_1 \cup r_2, c_1 \cap c_2)$ as a new MTB
                Remove any MTBs (r, c) in $MTBs$ where $r \in r_1 \cup r_2$ and $c \in c_1 \cap c_2$ ▷ Remove any non-maximal MTBs
                $MTBsNext$.add(M)
            end if
        end for
    end while
    for each MTB in MTBs **do**
        Replace any grouped rows and columns with the original row and column indices in A
    end for
    Return $MTBs$
end function

Algorithm 5 Algorithm for identifying all type L MTBs from a miRNA-target interaction network

```
1: function TypeLMTB(R, C, A)        ▷ R: the set of all row nodes, each representing an mRNA
                                     ▷ C: the set of all column nodes, each representing a miRNA
                                     ▷ A: the miRNA-mRNA network in a (node → neighbor set) hash map format
2:     Define S := R ∪ C as the set of row and column nodes not in any MTB yet
3:     Define MTBs as the list of MTBs, initialized to an empty list
4:     while S ≠ ∅ do                ▷ While there are still row or column nodes not in any MTB
5:         Get any node x from S
6:         Define M as the nodes in the connected component of x, initialized to an empty set
7:         Define Q as the set of newly discovered nodes in the connected component of x, initialized to {x}
8:         while Q ≠ ∅ do            ▷ While there may still be undiscovered members of the connected component
9:             Define Q' as the set of nodes in the connected component to be discovered, initialized to an empty set
10:            for each node y ∈ Q do
11:                Q' := Q' ∪ A.get(y)        ▷ Add all neighbors of y to Q'
12:            end for
13:            M := M ∪ Q             ▷ Add all nodes discovered in the previous iterations to the MTB
14:            Q := Q' − M            ▷ Determine the set of newly discovered members of the connected component
15:        end while
16:        MTBs.add(M)                ▷ Add M as a new MTB
17:        S := S − M                 ▷ Redetermine the nodes not yet associated with any MTB
18:    end while
19:    Return MTBs
20: end function
```

Algorithm 6 Algorithm for identifying high-scoring type Lmi MTBs from a miRNA-target interaction network

function TypeLmiMTB(R, C, A, d, ρ) ▷ R: the set of all row indices, each representing an mRNA ▷ C: the set of all column indices, each representing a miRNA ▷ A: the adjacency matrix of the miRNA-mRNA network ▷ d: minimum distance between two initial MTBs both of which are to be kept ▷ ρ: minimum density of an MTB
  Define $MTBs :=$ TypeRmiMTB(R, C, A) as the starting set of MTBs
  for each pair of MTBs $M_1 = (r_1, c_1)$ and $M_2 = (r_2, c_2)$ in $MTBs$ **do** ▷ Remove initial MTBs that are too similar to some others
    if $dist(M_1, M_2) < d$ **then** ▷ Remove the smaller one if the two initial MTBs are too similar, checked in the same way as in Algorithm 1 of Du et al., 2008
      if $r_1 + c_1 < r_2 + c_2$ **then**
        $MTBs := MTBs$ - M_1
      else
        $MTBs := MTBs$ - M_2
      end if
    end if
  end for
  Define R' as the set of mRNAs not participating in any MTBs in $MTBs$
  while $R' \neq \emptyset$ **do** ▷ Try adding one of the rows not in any MTBs to one of the MTBs
    Define den as the highest density of the resulting MTB after any of the additions, initialized to 0
    for each $i \in R'$ **do**
      Define js as the columns with 1's in $A[i, .]$, the i-th row of A
      for each $M = (r, c) \in MTBs$ **do**
        if Density($(r \cup \{i\}, c \cup js)$) $> den$ **then** $den :=$ Density($(r \cup \{i\}, c \cup js)$)
        end if
      end for
    end for
    if $den > \rho$ **then** ▷ Perform the addition only if the resulting density is higher than the threshold
      Add the rows and columns to the MTB
      Remove the rows from R'
      if the resulting MTB is identical to another one in $MTBs$ **then**
        Remove it from $MTBs$
      end if
    else
      Break the while loop
    end if
  end while
  Return $MTBs$
end function

Algorithm 7 Algorithm for identifying high-scoring type Lm MTBs from a miRNA-target interaction network

function TypeLmMTB(R, C, A, d, ρ) ▷ R: the set of all row indices, each representing an mRNA ▷ C: the set of all column indices, each representing a miRNA ▷ A: the adjacency matrix of the miRNA-mRNA network ▷ d: minimum distance between two initial MTBs both of which are to be kept ▷ ρ: minimum density of an MTB
  Define $MTBs :=$ TypeRmMTB(R, C, A) as the starting set of MTBs
  for each pair of MTBs $M_1 = (r_1, c_1)$ and $M_2 = (r_2, c_2)$ in $MTBs$ **do** ▷ Remove initial MTBs that are too similar to some others
    if $dist(M_1, M_2) < d$ **then** ▷ Remove the smaller one if the two initial MTBs are too similar, checked in the same way as in Algorithm 1 of Du et al., 2008
      if $r_1 + c_1 < r_2 + c_2$ **then**
        $MTBs := MTBs$ - M_1
      else
        $MTBs := MTBs$ - M_2
      end if
    end if
  end for
  Define C' as the set of miRNAs not participating in any MTBs in $MTBs$
  while $C' \neq \emptyset$ **do** ▷ Try adding one of the columns not in any MTBs to one of the MTBs
    Define den as the highest density of the resulting MTB after any of the additions, initialized to 0
    for each $j \in C'$ **do**
      Define is as the rows with 1's in $A[., j]$, the j-th column of A
      for each $M = (r, c) \in MTBs$ **do**
        if Density($(r \cup is, c \cup \{j\})$) $> den$ **then** $den :=$ Density($(r \cup is, c \cup \{j\})$)
        end if
      end for
    end for
    if $den > \rho$ **then** ▷ Perform the addition only if the resulting density is higher than the threshold
      Add the rows and columns to the MTB
      Remove the columns from C'
      if the resulting MTB is identical to another one in $MTBs$ **then**
        Remove it from $MTBs$
      end if
    else
      Break the while loop
    end if
  end while
  Return $MTBs$
end function

Algorithm 8 Algorithm for identifying high-scoring type Lgen MTBs from a miRNA-target interaction network

function TypeLgenMTB(R, C, A, d, ρ) ▷ R: the set of all row indices, each representing an mRNA
▷ C: the set of all column indices, each representing a miRNA
▷ A: the adjacency matrix of the miRNA-mRNA network
▷ d: minimum distance between two initial MTBs both of which are to be kept
▷ ρ: minimum density of an MTB
  Define $MTBs :=$ TypeRgenMTB(R, C, A) as the starting set of MTBs
  for each pair of MTBs $M_1 = (r_1, c_1)$ and $M_2 = (r_2, c_2)$ in $MTBs$ **do**
  ▷ Remove initial MTBs that are too similar to some others
    if $dist(M_1, M_2) < d$ **then** ▷ Remove the smaller one if the two initial MTBs are too similar, checked in the same way as in Algorithm 1 of Du et al., 2008
      if $r_1 + c_1 < r_2 + c_2$ **then**
        $MTBs := MTBs$ - M_1
      else
        $MTBs := MTBs$ - M_2
      end if
    end if
  end for
  Define R' as the set of mRNAs not participating in any MTBs in $MTBs$
  Define C' as the set of miRNAs not participating in any MTBs in $MTBs$
  while $R' \neq \emptyset$ or $C' \neq \emptyset$ **do** ▷ Try adding one of the rows not in any MTBs to one of the MTBs
    Define $denR$ as the highest density of the resulting MTB after any of the additions, initialized to 0
    for each $i \in R'$ **do**
      for each $M = (r, c) \in MTBs$ **do**
        if Density($(r \cup \{i\}, c)$) $> denR$ **then** $denR :=$ Density($(r \cup \{i\}, c)$)
        end if
      end for
    end for
    if $denR > \rho$ **then** ▷ Perform the addition only if the resulting density is higher than the threshold
      Add the row to the MTB
      Remove the row from R'
      if the resulting MTB is identical to another one in $MTBs$ **then**
        Remove it from $MTBs$
      end if
    end if
    ▷ Try adding one of the columns not in any MTBs to one of the MTBs
    Define $denC$ as the highest density of the resulting MTB after any of the additions, initialized to 0
    for each $j \in C'$ **do**
      for each $M = (r, c) \in MTBs$ **do**
        if Density($(r, c \cup \{j\})$) $> denC$ **then** $denC :=$ Density($(r, c \cup \{j\})$)
        end if
      end for
    end for
    if $denC > \rho$ **then** ▷ Perform the addition only if the resulting density is higher than the threshold
      Add the column to the MTB
      Remove the column from C'
      if the resulting MTB is identical to another one in $MTBs$ **then**
        Remove it from $MTBs$
      end if
    end if
    if $denR \leq \rho$ and $denC \leq \rho$ **then**
      Break the while loop
    end if
  end while
  Return $MTBs$
end function

A unified general model of MTB

It is possible to generalize all eight types of MTB by one single unified model. A general MTB is defined as a submatrix of the input matrix of miRNA-mRNA interactions. Each MTB is associated with a score indicating its proximity to the ideal case. Specifically, let R and C be respectively the sets of all rows (mRNAs) and all columns (miRNAs) in the input matrix, r and c be the rows and columns involved in an MTB, a_{ij} represents the element at row i and column j of the matrix, and k_0, k_1 and k_2 are parameters with non-negative values. The score of an MTB (r, c), $f(r, c)$, is defined by the following formula:

$$f(r,c) \quad = \quad 1 - k_0 \frac{\sum_{i \in r, j \in c}(1 - a_{ij})}{|r||c|} - k_1 \frac{\sum_{p \in (R-r), j \in c} a_{pj}}{|R-r||c|} - k_2 \frac{\sum_{i \in r, q \in (C-c)} a_{iq}}{|r||C-c|}$$

The scoring formula consists of four parts. The first part is the constant value 1, which represents the maximum score of an MTB. The other three parts are penalty terms for missing 1's in the MTB, extra 1's on the same columns outside the MTB, and extra 1's on the same rows outside the MTB, respectively. The three parameters determine which penalty terms to apply and their relative weights. For the eight MTB types defined, the algorithms we used to identify the corresponding MTBs can be considered as algorithms for finding top-scoring MTBs, defined as some specific instantiations of this general model with different sets of parameter values, as shown in Table S1.

Table S1 Relationships between the unified general model and the MTBs identified by our algorithms for the eight types of MTB. *: For the L, Lmi, Lm and Lgen types, we also have an additional connectedness requirement between the rows and columns.

MTB type	k_0	k_1	k_2
R	∞	∞	∞
Rmi	∞	∞	0
Rm	∞	0	∞
Rgen	∞	0	0
L	0*	∞	∞
Lmi	1*	∞	0
Lm	1*	0	∞
Lgen	1*	0	0

For type R MTB, having any missing 1's in a submatrix or extra 1's on the same rows or columns outside the submatrix would result in a negative infinity score, which disqualifies it as a valid MTB. For type Rmi, the penalty terms for missing 1's in a submatrix and extra 1's on the same columns outside it still apply, but the penalty for extra 1's on the same rows outside the submatrix is waived. Similarly, for type Rm, extra 1's on the same rows are penalized, but extra 1's on the same columns are not. For type Rgen, both are not penalized and any submatrix having only 1's is qualified as a MTB. For type L, extra 1's on the same rows and columns outside a submatrix are penalized, but missing 1's within it is not. Similar arguments apply for types Lmi, Lm and Lgen, except that in these cases the within-MTB density of 1's is used to evaluate the quality of a MTB, but the allowed extra 1's on the same rows or columns are simply ignored.

In addition to the eight standard types, new types of MTB can be defined by using different combinations of values for the parameters k_0, k_1 and k_2. High-scoring

MTBs can be found by heuristic optimization algorithms. In some special cases, as with some of the eight standard types, it is also possible to derive efficient algorithms to return all MTBs or all maximal MTBs.

The scoring formula can also be written in another way. Suppose now r is a binary vector with $|R|$ entries in total, where an entry is 1 if the corresponding row is a member of the MTB of interest and 0 if not. Similarly, we redefine c as the binary vector with $|C|$ entries in total, where an entry is 1 if the corresponding column is a member of the MTB and 0 if not. The scoring formula can then be written in terms of these vectors and the whole miRNA-mRNA interaction matrix A:

$$f(r,c) \quad = \quad 1 - k_0 \frac{r^t r c^t c - r^t A c}{r^t r c^t c} - k_1 \frac{(\mathbf{1} - r)^t A c}{(R - r)^t (R - r) c^t c} - k_2 \frac{r^t A (\mathbf{1} - c)}{r^t r (C - c)^t (C - c)},$$

where $\mathbf{1}$ represents the binary vector of all 1's (with a length depending on the context).

The main reason to write the scoring formula using the vector and matrix representations is that the general MTB model can then be extended to handle non-binary interaction matrices, in which each element a_{ij} takes on a continuous value between 0 and 1 that represents the confidence of the miRNA targeting the mRNA. Correspondingly, one may also allow fractional values in the r and c vectors to represent fussy MTB memberships. With these changes, a whole series of other well-established optimization methods can be applied to identify MTBs of high scores.

Supplementary figures

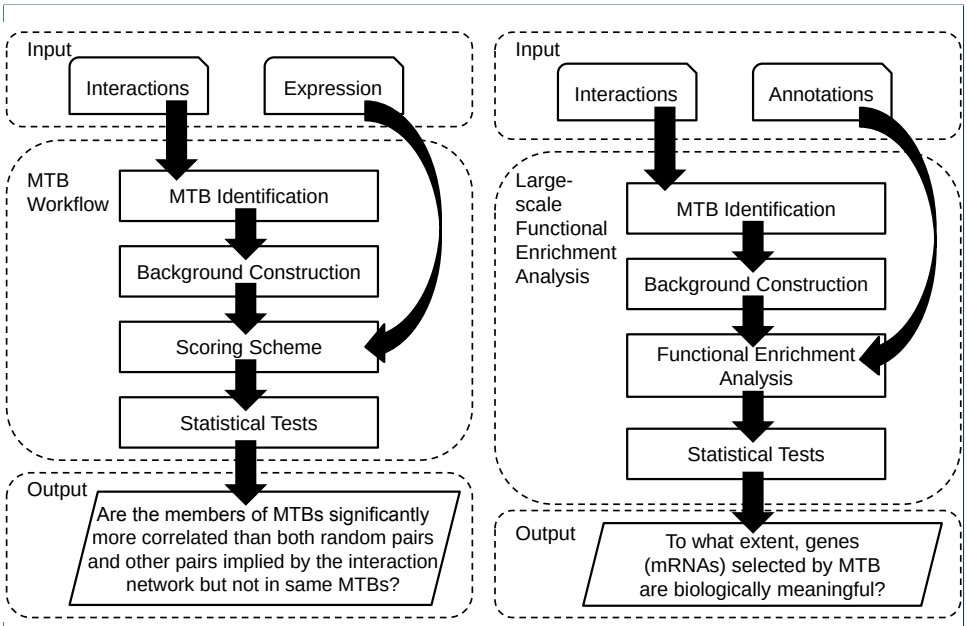


Figure S1 Workflow for studying (a) expression correlations between members of same MTBs and (b) functional relationships between genes in same MTBs.

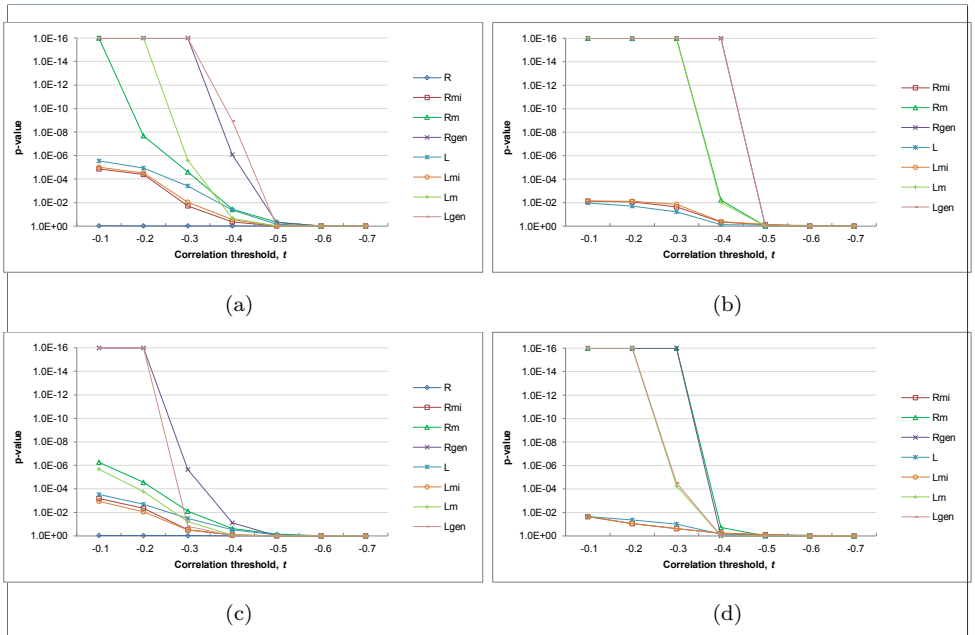


Figure S2 Statistical significance of the negative correlations between the expression levels of miRNAs and mRNAs in the same MTBs. The p-values were computed based on the expressed union sets without TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set as compared to a random background sampled from all expressed mRNAs and miRNAs; and (c) the high-confidence set and (d) the high-coverage set as compared to a background consisting of miRNA-mRNA pairs with interactions in the input network but are not in same MTBs. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program. MTB types with no identified MTBs are omitted.

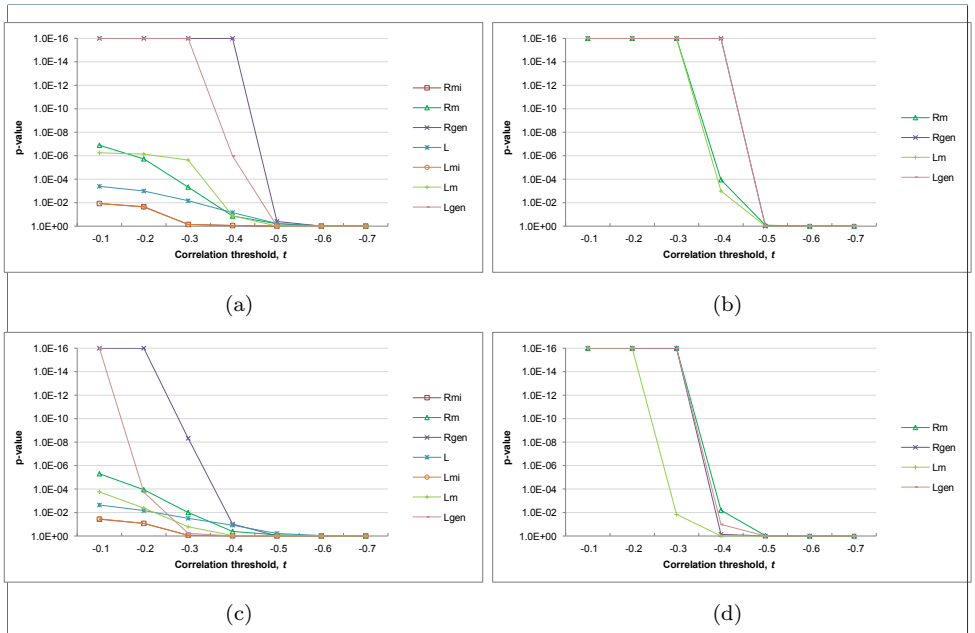


Figure S3 Statistical significance of the negative correlations between the expression levels of miRNAs and mRNAs in the same MTBs, when related miRNAs were grouped. The p-values were computed based on the expressed union sets with TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set as compared to a random background sampled from all expressed mRNAs and miRNAs; and (c) the high-confidence set and (d) the high-coverage set as compared to a background consisting of miRNA-mRNA pairs with interactions in the input network but are not in same MTBs. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program. MTB types with no identified MTBs are omitted.

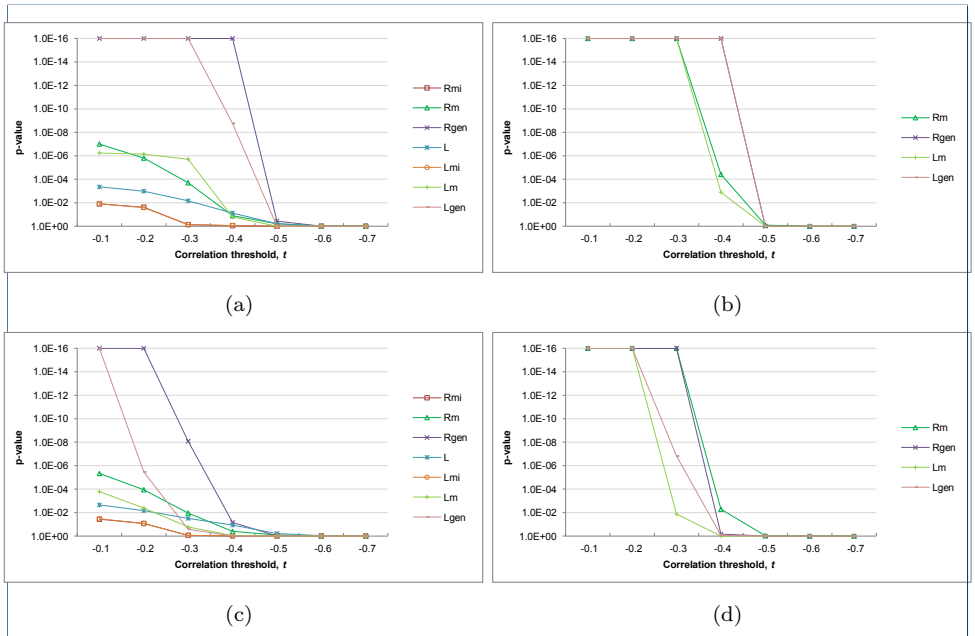


Figure S4 Statistical significance of the negative correlations between the expression levels of miRNAs and mRNAs in the same MTBs, when related miRNAs were grouped. The p-values were computed based on the expressed union sets without TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set as compared to a random background sampled from all expressed mRNAs and miRNAs; and (c) the high-confidence set and (d) the high-coverage set as compared to a background consisting of miRNA-mRNA pairs with interactions in the input network but are not in same MTBs. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program. MTB types with no identified MTBs are omitted.

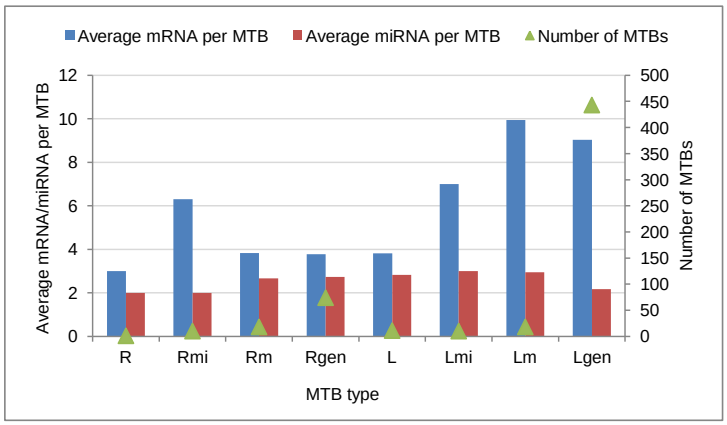


Figure S5 Statistics of the MTBs identified from the high-confidence expressed union set without TarBase interactions. For each type of MTB, the average number of mRNAs per MTB, average number of miRNAs per MTB and the number of MTBs identified by our algorithm are shown.

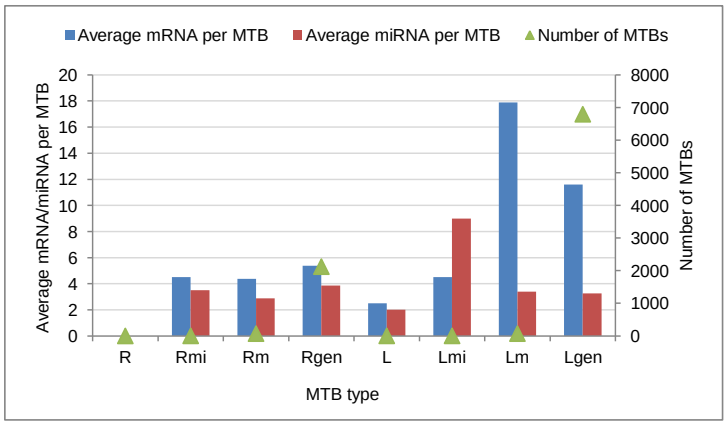


Figure S6 Statistics of the MTBs identified from the high-coverage integrated expressed union set with TarBase interactions. For each type of MTB, the average number of mRNAs per MTB, average number of miRNAs per MTB and the number of MTBs identified by our algorithm are shown.

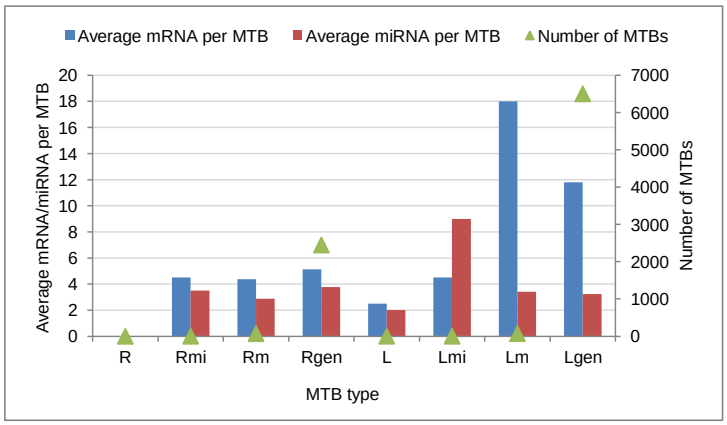
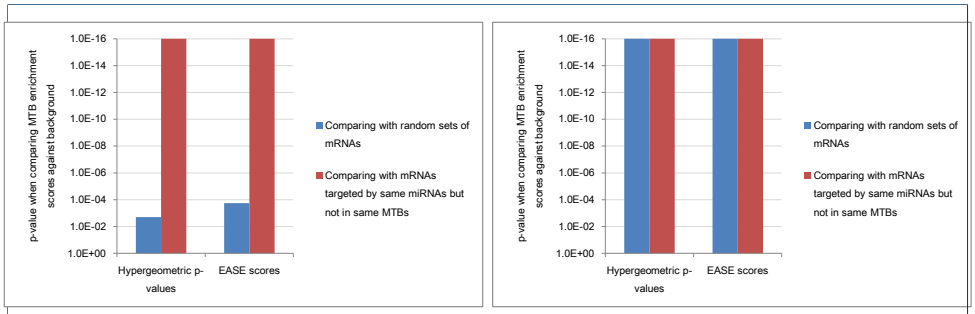


Figure S7 Statistics of the MTBs identified from the high-coverage integrated expressed union set without TarBase interactions. For each type of MTB, the average number of mRNAs per MTB, average number of miRNAs per MTB and the number of MTBs identified by our algorithm are shown.



(a) (b)

Figure S8 Statistical significance of the functional enrichment scores of the genes from same type Lgen MTBs. The p-values were computed based on the expressed union sets without TarBase interactions, for (a) the high-confidence set and (b) the high-coverage set. In the figures, 1E-16 represents the smallest p-value that could be outputted by our program.