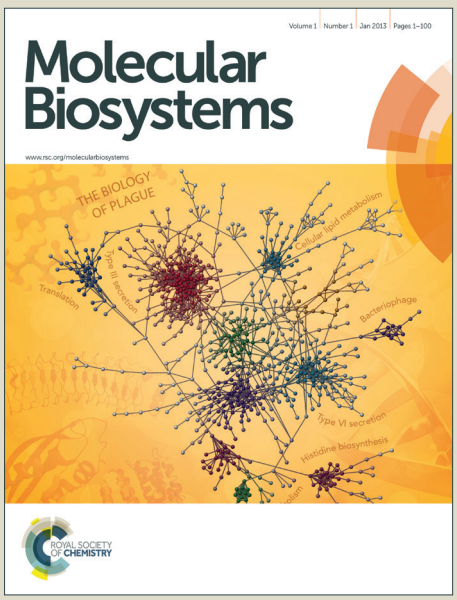
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems RSCPublishing

METHOD

MIROR: A METHOD FOR CELL-TYPE SPECIFIC MICRORNA OCCUPANCY RATE PREDICTION

Cite this: DOI: 10.1039/x0xx00000x

Received XXth XXXX XXXX,
Accepted XXth XXXX XXXX

DOI: 10.1039/x0xx00000x

www.rsc.org/XXXX

Peng Xie,[‡a] Yu Liu,[‡a] Yanda Li,[a] Michael Q. Zhang,[*ab] Xiaowo Wang[*a]

MicroRNA (miRNA) regulation is highly cell-type specific. It is sensitive to both the miRNA-mRNA relative abundance and the competitive endogenous RNA (ceRNA) effect. However, almost all existing miRNA target prediction methods neglected the influence of cellular environment when analyzing miRNA regulation effects. In this study, we proposed a method, MIROR (MiRNA Occupancy Rate predictor), to predict miRNA regulation intensity in a given cell type. The major considerations were the miRNA-mRNA relative abundance and the endogenous competition between different mRNA species. The output of MIROR is the predicted miRNA occupancy rates of each target site. The predicted results significantly correlated with Ago HITS-CLIP experiment that indicated miRNA binding intensities. When applied to the analysis of breast invasive carcinoma dataset, MIROR identified a number of differentially regulated miRNA-mRNA pairs with significant miRNA occupancy rate changes between tumor and normal tissues. Many of the predictions were supported by previous researches, including the ones without significant change in mRNA expression level. These results indicate that MIROR provides a novel strategy to study the miRNA differential regulation in different cell types.

Availability: MIROR is freely available at http://bioinfo.au.tsinghua.edu.cn/member/xwwang/MIROR.

Introduction

MicroRNAs (miRNAs) are a class of small noncoding RNA molecules, which play important regulatory roles in a variety of crucial biological processes[1-3]. Abnormal expression of miRNA is linked to physiological disorders and cancers[4]. MiRNAs play their regulatory function by targeting mRNAs through the interaction with the Argonaute (Ago) family protein and the formation the RNA-induced silencing complex (RISC). In animals, the major determinant for target recognition is the miRNA seed sequence, which is a 6~7 nucleotide sequence at the 5'end of a miRNA. The short "seed" regions cannot guarantee highly specific target binding and it is believed that one miRNA usually can targets hundreds of potential binding sites[5]. Despite their biological importance, current understanding of the functions of miRNAs is still limited. Till now, more than one thousand human miRNAs have been reported[6], hundreds of thousands of miRNA-mRNA interactions have been predicted[5, 7, 8]. However, only less than two thousand miRNA-mRNA regulation events have been verified through experiments[9].

Bioinformatics approaches facilitate miRNA target identification at genome-wide scale and are fundamental for the study of miRNA functions. Most target prediction algorithms are sequence-based, which score miRNA-target pairs according to base-pairing, complex stability, site accessibility, etc[5, 8, 10]. Other features, such as cross species conservation, relative positions and flanking sequence contexts in 3'UTRs, have also been considered to improve prediction accuracy[10]. However, current algorithms still suffer from high false positive rates and the consistency between different algorithms is limited[11, 12]. More recently, several methods have been proposed to improve the reliability of existing target predictions[13-15]. These methods usually take predictions from one or several sequence-based algorithms as input, examine statistical significance of each predicted miRNA-target pair by their expression correlation in a large amount of different cell types/samples and output highly correlated pairs to form a more reliable target prediction set. However, these methods could only identify the targets significantly affect by miRNA at RNA expression level, but not the ones mainly regulated at translational level.

In addition, all the predictions made by these algorithms are static – they provide potential miRNA-mRNA interaction pairs regardless of a certain specific cell type. In fact, miRNA-mRNA interactions are highly dynamic and cell type specific, and should be measured in a quantitative way[16].

In this study, we tried to quantify the interaction between each miRNA-mRNA pair by the "miRNA occupancy rate" – what proportion of a specific species of mRNA is bound by a miRNA species. This occupancy rate can change in different cell types and is affected in several ways. First, the relative abundance of miRNAs and target sites changes in different cell types. Second, the interaction between a miRNA-mRNA pair can be significantly influenced by other mRNAs, since there can be hundreds of RNAs (including mRNAs and ncRNAs) competing for one miRNA simultaneously[17]. These mutually influencing RNAs are called competing endogenous RNAs (ceRNAs) and their impact on each other has been reported recently[18]. Attempts considering this effect have been made to improve target prediction and infer ceRNA pairs[19]. However, to our knowledge, a method for cell type specific and quantitative description of miRNA-mRNA interaction has not been reported.

Here we propose a method, MIROR, to predict miRNA occupancy rates of different target sites in a certain cellular environment (given miRNA and mRNA expression levels). It can be used to predict cell type specific miRNA occupancy rate based on the consideration of miRNA-mRNA relative abundance and endogenous competing effect. MIROR prediction correlated well with HITS-CLIP data from mouse brain samples. In application, MIROR can help to predict miRNA occupancy rate changes between cell types (for example, from normal to cancer cells). When applied to TCGA breast cancer data, MIROR showed higher sensitivity than other methods in identifying tumor-related miRNA-mRNA pairs. In sum, we proposed a method to quantify the cell type specific miRNA regulatory effect and provided a novel way to study the differential miRNA regulation between different cell states.

Materials and Methods

Data sources

i. HITS-CLIP Data

The mouse brain HITS-CLIP data[20] was downloaded from the website of Darnell's lab at Rockefeller University (http://ago.rockefeller.edu/). Five replicates (from A to E) were included in this data set. Two different antibodies were used for Ago immunoprecipitation (2A8 for replicates A/B/C, 7G1-1* for replicates D/E). The authors provided an Ago-miRNA-mRNA Ternary map, which was estimated by integrating the five replicates. The integrated HITS-CLIP peak heights represented miRNA binding intensity. miRNA expression data was downloaded from the same website. mRNA expression data was downloaded from the GEO database under accession number GSE16338. All HITS-CLIP reads were aligned to mm9 genome using bowtie, allowing 1 mismatch (bowtie -f -m 1--best --strata). We further transformed the alignment to the wiggle (WIG) format and matched them with HITS-CLIP reads clusters provided by the authors[20]. Data were visualized with UCSC genome browser (http://genome.ucsc.edu/).

ii. Breast Invasive Carcinoma Dataset

Tumor and matched normal tissue samples including miRNA and mRNA expression data were downloaded from The Cancer Genome Atlas data portal (July 2012). The downloaded data set contained 170 tumor and normal samples, with BCGSC_Illumina GA(/IlluminaHiSeq)_miRNASeq platform for miRNAs expression and UNC_IlluminaHiSeq_RNASeqV2 platform for mRNAs expression.

iii. Input miRNA target prediction list

For the ease of comparing MIROR with other methods, we used TargetScan (V5.2) prediction as putative miRNA-target list for all the methods. Totally, the putative miRNA-target list included 272,534 pairs.

iv. GenMiR++, Magia and TaLasso predictions

For GenMiR++, Magia, and TaLasso analysis, we used the miRNA and mRNA expression files as the same as ours, and used default parameters of each algorithm. GenMiR++ was run in Matlab platform with the given code, Magia and TaLasso were run on their website.

Modeling the miRNA binding process

We modeled the miRNA binding process at the thermodynamic equilibrium state of the miRNA-mRNA interacting system. In this model, the ith species of miRNA and the jth species of target site were represented as miR_i (i=1, …, m) and TS_j (j=1, …, n), respectively. The reaction of miRi binding to the target site TS_j could be formulized as:

$$miR_i + TS_j \rightleftharpoons miR_i : TS_j$$

Where the $miR_i:TS_j$ represents the miRNA-mRNA complex. Under the assumption of chemical equilibrium, the molecular concentrations will follow this equation:

$$K_{ij} = \frac{[miR_i:TS_j]}{[miR_i][TS_j]} \quad (1)$$

K_{ij} is the equilibrium constant, which is a function of the Gibbs free energy:

$$K_{ij} = \exp\left(-\frac{\Delta G_{ij}}{RT}\right) \quad (2)$$

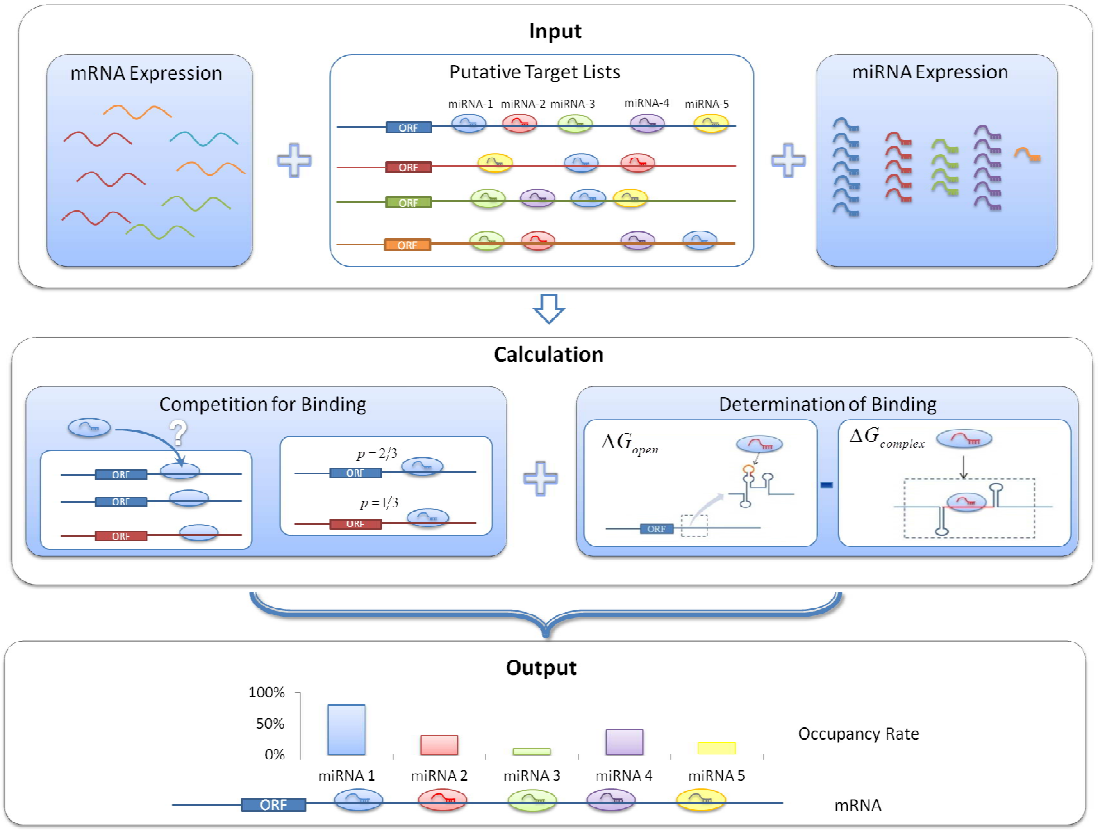


Fig. 1 Workflow of MIROR. Input: mRNA and miRNA expression data and a putative miRNA target list (e.g., predicted miRNA targets of TargetScan or PITA). Calculation: solve the thermodynamic equilibrium problems of miRNA binding processes (Methods). Output: miRNA occupancy rate at each target site.

Where R and T represent ideal gas constant and absolute temperature, respectively. $\Delta G_{ij}=G_{ij}^{bound}- G_{ij}^{unbound}$, which is the change of Gibbs free energy. In this work, we used the function in PITA package[8] to estimate the value of the energy change. It has been recently reported that the value of equilibrium constant K_{ij} for miRNA or small RNA binding in vivo are different from that directly estimated by RNA hybridization, but is still linearly correlated with the log transformed Gibbs free energy[21, 22]. So we added a scaling factor k to modify the calculation of equilibrium constant and re-formulized Equation (2).

$$K_{ij} = \exp\left(-\frac{\Delta G_{ij}}{kRT}\right) \quad (3)$$

Deriving from Equation (1) we can easily get the percentage of miR_i that bind to the TS_j as:

$$P_{ij} = \frac{[miR_i{:}TS_j]}{\sum_{k=1}^{n}[miR_i{:}TS_k]+[miR_i]} = \frac{K_{ij}[TS_j]}{\sum_{k=1}^{n}K_{ik}[TS_k]+1} \quad (4)$$

The occupancy rate of miR_i at TS_j can be derived as:

$$OC_{ij} = \frac{[miR_i{:}TS_j]}{\sum_{k=1}^{m}[miR_k{:}TS_j]+[TS_j]} = \frac{K_{ij}[miR_i]}{\sum_{k=1}^{m}K_{kj}[miR_k]+1} \quad (5)$$

In principle, by solving these equations, we can get the occupation rate at the equilibrium state of the system. However, as each miRNA species typically could bind to hundreds of target sites species, it's not possible to solve these equations directly. So we used a numerical approach to solve this problem in a step-by-step way: in each step, we substituted the numbers of free molecules got from the last step into Equation (4) to calculate P_{ij}; a small proportion of miRNA was distributed to target sites according to P_{ij}. We iterated this process until all miRNAs had been distributed or all biding sites had been occupied by miRNA. In the case where more than one miRNA species (e.g. different members of the same miRNA family) could bind to the same target site species and the free target site number was less than the total number of miRNA molecules distributed to it, we divided the free molecules of this target site species to miRNA species according to Equation (5). In this model, there were three parameters that needed to be trained. The first parameter was the relative ratio between total number of effective miRNAs and that of mRNAs. This is necessary because miRNA and mRNA expression levels are measured by different experimental approaches that their "expression values" are not directly comparable. The second parameter was the scaling factor k. The third parameter was the closest distance for two target sites to be bound simultaneously (neighboring target sites with distances smaller than the threshold will be merged as one site). We evaluated the performance of each group of parameters by the correlation between the predicted miRNA occupancy and the heights of HITS-CLIP clusters. The best parameter set trained when using TargetScan predictions as the input also applied well when

using PITA prediction as the input, with Spearman correlation coefficient (SCC) equal to 0.35.

The step length for distributing the miRNAs also influences the result. In principle, we get more accurate solution when decrease step length and increase the number of calculation steps. We tested different total number of steps from 100 to 1000 and found that the predicted results stabilized when this number was over 500 (Supplementary Table S1). So we chose 500 as the default number of total steps to balance the accuracy and the computational cost.

Data, scripts and user manual of MIROR are freely available at http://bioinfo.au.tsinghua.edu.cn/member/xwwang/MIROR.

Comparison of predicted miRNA binding intensity with Ago HITS-CLIP data

At all miRNA binding sites supported by both TargetScan and HITS-CLIP results, we compared miRNA binding intensity predicted by MIROR and intensity measured by HITS-CLIP experiment. Binding sites were first ranked by expression levels of corresponding mRNA and then divided into sliding windows (each window contained 500 binding sites, sliding step was 1 site). In each window, SCC was calculated between miRNA binding intensity predicted by MIROR and intensity measured by HITS-CLIP experiment. As a control, we calculated in each window the SCC between binding site abundance (expression levels of corresponding mRNA) and HITS-CLIP measured miRNA binding intensity.

Results

miRNA Occupancy Rate Prediction

We proposed a thermodynamic model to predict miRNA-mRNA occupancy rates of target sites. This model takes the expression levels of both miRNA and mRNA and a putative miRNA target list (e.g. TargetScan predictions) as the input, considers the thermodynamic equilibrium of the binding process, and outputs the predicted miRNA-mRNA occupancy rate (Fig 1).

The putative miRNA-target list for MIROR is flexible according to the needs of users. For the ease of comparison with other methods, in this paper we took TargetScan predicted targets as the input list. We did parameter training and evaluated parameter sets according to the correlation between predicted occupancy rate and HITS-CLIP binding intensity (Supplementary Table S1) the mouse brain dataset[20]. Best performance was achieved when we set total number of miRNA at 5 to 6 times of that of mRNAs and RT-scaling factor k equal to 20 (Supplementary Table S1). The model was not sensitive to the minimum distance constrain

The putative miRNA-target list for MIROR is flexible according to the needs of users. For the ease of comparison with other methods, in this paper we took TargetScan predicted targets as the input list. We did parameter training and evaluated parameter sets according to the correlation between predicted occupancy rate and HITS-CLIP binding intensity (Supplementary Table S1) the mouse brain dataset[20]. Best performance was achieved when we set total number of miRNA at 5 to 6 times of that of mRNAs and RT-scaling factor k equal to 20 (Supplementary Table S1). The model was not sensitive to the minimum distance constrain

The best distance constrain turned out to be 20-40nt (Supplementary Table S1), which was consistent with the reported size of Ago footprint[20].

The predicted occupancy rate is useful in that it provides information for the following questions: in a certain cell type (given the expression level of miRNAs and mRNAs), 1. to what extent a concerned mRNA species is regulated by miRNAs; 2. which miRNA is the primary regulator of a concerned mRNA species. In addition, comparison of miRNA occupancy rates between different cell types, for example normal and cancer cells, can help us to discover biologically relevant miRNAs and mRNAs by identifying pairs with significant occupancy rate changes. This quantitative measurement provides a novel strategy for the discovery of key molecules in the miRNA regulation processes.

MIROR-Predicted miRNA Occupancy Correlated with HITS-CLIP AGO Binding Intensities

AGO HITS-CLIP (high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation) experiment, which uses antibody against Ago proteins to capture miRNA-

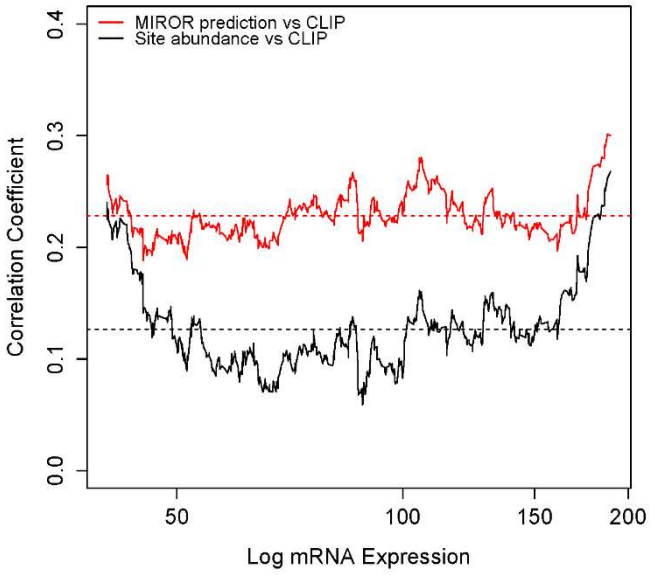


Fig. 2 SCC curve. HITS-CLIP reported miRNA binding sites were ranked according to mRNA expression levels and divided into sliding windows. In each window, we calculated the SCC between MIROR-predicted and HITS-CLIP-measured miRNA binding intensity (red curve). Same calculation was done between binding site abundance (mRNA expression levels) and HITS-CLIP intensity.

mRNA-Ago complex and performs RNA-seq to read out the corresponding miRNA and mRNA species, have been applied to identify miRNA binding sites in different cell types[20, 23]. Sequencing reads from HITS-CLIP experiments cluster at miRNA target sites and reads abundance correlates with the relative concentration of miRNA-mRNA duplex. Thus, HITS-CLIP data could serve as a benchmark to validate our model. We chose the data set of mouse brain tissue[20], since corresponding mRNA and miRNA expression data were also available.

In order to evaluate the performance of MIROR, we compared the predicted number of miRNAs bound to target sites with HITS-CLIP miRNA binding profile. The Spearman Correlation Coefficient (SCC) between prediction and HITS-CLIP data was 0.39 (P value $< 2.2\times10^{-16}$). To normalize the effect of binding site abundance and present the performance of MIROR in higher resolution, we ranked the target sites according to mRNA expression levels, divided them into sliding windows and calculated the SCC between predicted and HITS-CLIP miRNA binding intensities in each window (Fig 2). The height of SCC curve by comparing the MIROR prediction with HITS-CLIP data evidently surpassed that of binding sites abundance with HITS-CLIP profile (Fig 2).

Since this performance evaluation would be limited by the consistency of experimental replicates, we also analyzed the correlation of the five replicates in this dataset (Fig S1 a, b). For each replicate, we pooled the other four replicates as a whole and compared the single replicate with the pooled results. As shown in Fig S1a, our predictions achieved over 70% of the average SCC of the best experimental replicate. This result indicates that MIROR could effectively predict the quantitative nature of miRNA regulation.

We also tested whether the result was biased by the input putative target set. We used the parameters trained by TargetScan predictions, and the PITA predicted miRNA-mRNA pairs as the input for MIROR. The SCC between MIROR predictions and HITS-CLIP profile was 0.35, which is comparable to the results of using TargetScan prediction set as the input.

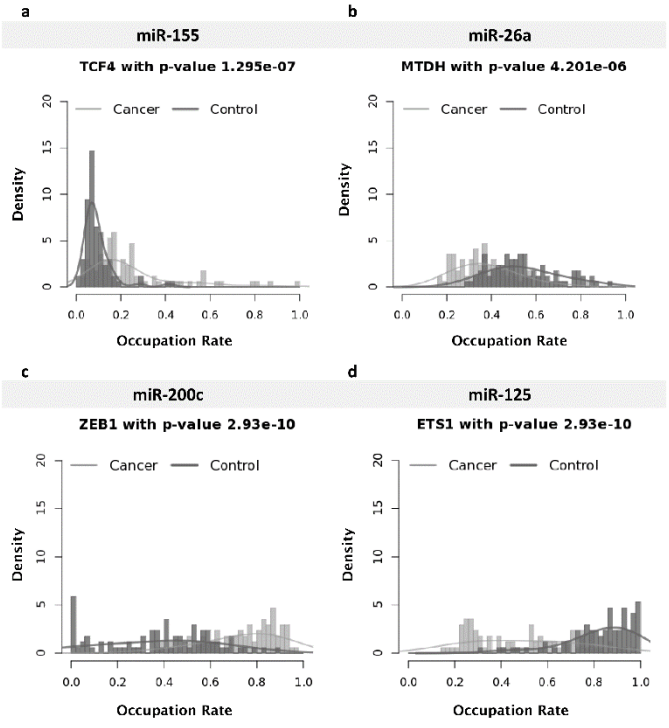


Fig. 3 Example of miRNA occupancy rate changes of a few known breast cancer related miRNA-mRNA pairs.

MIROR Predictions Indicated Significant Differential miRNA Regulations in Breast Cancer Cells

It has been reported that miRNAs play an important role in tumor generation and progression. Several studies indicated that some miRNAs could act as oncogenes or tumor suppressors through the interaction with target genes. Here we applied MIROR to The Cancer Genome Atlas (TCGA) breast invasive carcinoma dataset to predict the prominent miRNA-target pairs between tumor and normal samples. We predicted miRNA occupancy rates for each miRNA-target pair in cancer and normal samples respectively. For each miRNA-target pair, we generated two twenty-bin histograms of the occupancy rate (one for normal samples, one for cancer samples) and performed histogram based KS test (Frank Porter, 2008) to examine the consistency of the OC distribution between the normal and the cancer samples. The pairs with significant OC distribution changes may indicate a general trend of differential miRNA regulation between cancer and normal (Fig 3). We ranked these pairs according to their FDR[24] of the test of occupancy rate changes. The largest FDR of top 10000 pairs (~4% of all putative pairs) was 2.39e-3, suggesting considerable difference between the miRNA-mRNA interaction maps of normal and tumor tissues. Among these top pairs, we found a number of well-known breast cancer related miRNA-mRNA pairs (Fig 3). For example, previous work showed that the regulation of oncomir miR-155 on tumor suppressor gene TCF4 was strongly associated with breast cancer formation[25, 26]; miR-26a has been reported to antagonize human breast carcinogenesis by targeting MTDH[27]; miR-200c can repress the epithelial-to-mesenchymal transition (EMT) in breast cancer by targeting ZEB1[28, 29]. MIROR predicted that three miRNA families, mir-8 (including miR-141, miR-200a/b/c and miR-429), let-7 and mir-182 dominate the top OC change pairs by taking up about 3000 in the top 4000 pairs (Supplementary Fig S3). These three families were extensively studied tumor suppressors of breast cancer[30-32]. We also analyzed the miRNAs included in the top 10000 pairs by ranking them according to their target pair occurrence. All of the Top 20 miRNAs (Supplementary Table S2) were reported to be breast cancer related[33-35].

MIROR Effectively Predicted Breast-Cancer-Related miRNA-target pairs

One important task of bioinformatics tools is to infer key molecules that are biologically relevant. In the field of miRNA

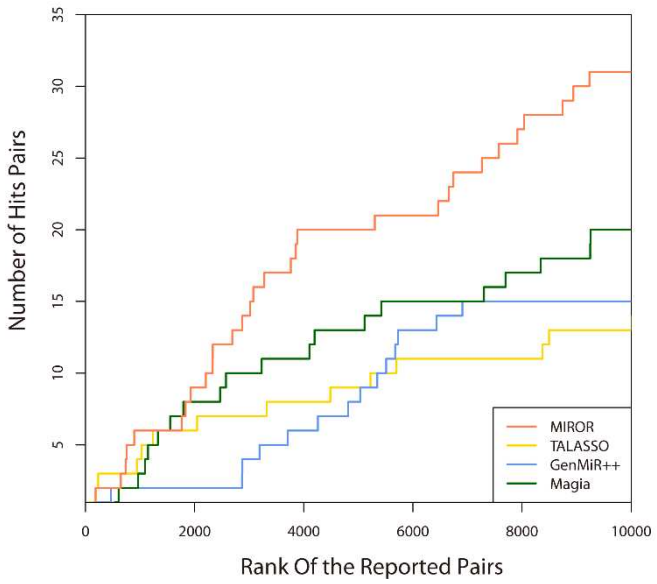


Fig. 4 Literature supported pairs in top predictions. miRNA-mRNA pairs reported by four methods were ranked according to FDR (Competition model, Magia) or algorithm-defined scores (TaLasso, GenMiR++). Top 10000 predictions were shown in the figure. X-axis shows the rank of reported pairs and y-axis shows the number of hits with breast cancer related pairs from five review papers, for each method.

research, several approaches can be applied for the purpose of predicting critical miRNA-mRNA pairs in given biological processes, using both miRNA and mRNA expression data, for example, the Bayesian network based method GeneMiR++[14], the regression based method TaLasso[13] and the integrative method Magia[15]. In this study, we set out to infer breast-cancer-related miRNAs and mRNAs based on MIROR-predicted miRNA occupancy changes. We also applied other methods to analyze the TCGA breast invasive carcinoma dataset and compared their performance with MIROR. For the sake of fairness, we chose TargetScan predicted pairs (272,534) as the candidate set of miRNA-target pairs for all approaches. We collected breast cancer related miRNA-target pairs from five review papers as the positive set[36-40]. These reviews included 133 miRNA-mRNA pairs, 78 of which were included in TargetScan predictions. We ranked predicted miRNA-mRNA pairs of each method according to the statistical significance and examined the occurrence of the 78 reported pairs. Among the top 10,000 predictions (top ~4%), MIROR predictions hit 31 reported pairs (P value 1.65e-25, Hypergeometric test), while Magia hit 20, GenMiR++ hit 15, and TALASSO hit 13 reported pairs (Fig 4). A detailed comparision of the number of hits in top 100, 500, 1000, 2000, 5000 predicted pairs were listed in Supplementary Table S4. These results indicated that the MIROR could effectively enrich breast-cancer-related miRNA-mRNA interaction pairs and provide a more reliable candidate list for further studies.

Another advantage of MIROR is its capability of identifying potential cancer-related miRNAs-mRNAs interactions without significant mRNA expression changes. These pairs might be neglected by other methods, such as Magia and TaLasso, which are based on the assumption that significant interactions should notably affect the amount of target mRNAs. In the TCGA breast cancer dataset, we identified 5,762 pairs with mRNA fold change < 1.5 in the top 10,000 predictions. Among these pairs, we found the well-known oncogene ETS1 and tumor suppressor miRNA miR-125b (Fig 3d). Previous work showed that ETS1 was targeted by miR-125b but the regulation was only reflected at translational level[25]. This suggests that there are a number of functional miRNA-target regulations without significant alteration in mRNA level, which would be missed by previous methods could be identified by MIROR.

Discussion

The miRNA-directed gene regulation system is highly dynamic and is cell type specific. However, most existing algorithms predict miRNA binding regardless of cellular environment. In this work, we tried to model the process of mRNAs competing for miRNAs and to predict miRNA occupancy rates in a cell-type specific manner. Based on the assumption of thermodynamic equilibrium, this model considered the influence of relative abundance of miRNA and mRNA and the competition between target sites. The output of this model is the predicted miRNA occupancy rate, which is a quantitative measurement of miRNA regulatory intensity. Predicted results were consistent with the HITS-CLIP experiment results. Based on the predicted occupancy rates, we proposed a new strategy for the discovery of cancer related miRNAs and target mRNAs: to find miRNA-target pairs with significant occupancy rate changes between cancer and matched-normal tissue samples. We presented an example of this kind of application on the TCGA breast cancer dataset. Comparing with three existing algorithms, MIROR showed higher sensitivity in discovering literature-supported breast cancer related miRNA-mRNA pairs.

Accurate description of the miRNA-mRNA interaction network is an important but challenging task. In this study and many others of the similar kind, a set of pre-defined miRNA-target pairs need to be provided in the first place. Usually this set is predicted by sequence-based algorithm, like TargetScan, PITA, PicTar, etc. In this paper, we used the target set predicted by TargetScan as input set. A comparison between TargetScan prediction and HITS-CLIP miRNA binding map showed that the overlap between these two sets were 1,321 miRNA binding sites, which was about one third of the total number of HITS-CLIP supported sites. We expect that collecting a more comprehensive and reliable putative miRNA-mRNA interaction set should be able to further improve the performance of MIROR.

In our model, the relative abundance between mRNA and miRNA is a key parameter that affects the predicted result

(Supplementary Table S1). We did parameter fitting for the ratio of total amounts of miRNA and target sites, so that the predicted miRNA occupancy best fitted the HITS-CLIP data. We used this ratio as the default parameter of our program and showed that it performed well when using different miRNA prediction algorithms as the input data. However, as we couldn't find other Ago HITS-CLIP data sets with matched mRNA and miRNA expression data to test our model, it's possible that this ratio may not best fit some other cell types. But a recent work indicated that the total miRNA/mRNA ratio is roughly consistent across species: in mouse and in fruit fly, the number of additional miRNA biding sites needed to dilute miRNA repression effect to 50% was similar[22]. In the near future, we are trying to use a synthetic biology approach to build a system where relative amount of miRNA and mRNA will be experimentally measured. In this system, miRNA regulation will be measured under different conditions so that we can construct a function that determines the miRNA binding process.

Studying miRNA regulation in a quantitative and systematic manner is a necessary step towards deepening our understanding of the miRNA regulatory system [16, 19]. Recently, Coronnello et al. also proposed a thermodynamic model named ComiR, [41] to predict the miRNA binding probability of a mRNA in a specific cell type. The major difference between our model and theirs is: (1) we considered both miRNA and mRNA expression levels while ComiR only considered miRNA expression levels; (2) our model considered the experimental validated ceRNA effect, but ComiR assumed independence between each miRNA-mRNA binding process. We didn't included ComiR in our comparison as it was implemented as an online service that did not allow the large amount of calculation for TCGA breast cancer data analysis.

There are several possible ways to further improve miRNA occupancy rate prediction. First, alternative splicing can result in 3' UTRs with variable lengths. The percentage of isoforms with different 3' UTRs can change with cell types, which may affect the competitiveness for miRNA binding of mRNAs. Second, we currently only consider target sites provided by sequence feature based miRNA target prediction algorithms (e.g. TargetScan). These algorithms usually neglect ncRNAs such as pseudogene transcripts and circular RNAs, which have been shown to modulate miRNA regulation[18, 42]. Third, mRNAs may interact with RNA binding proteins and change the property of the miRNA target sites (e.g. the PUF protein[43]). With the development of techniques, like RIP-seq etc, the influence of other mRNA binding proteins could be considered, which may in turn explain the false positive prediction of current algorithms.

Conclusions

Taken together, we see MIROR as a progress towards the goal of comprehensively describing cell status and unraveling the mechanism of miRNA regulation in a quantitative way. This may be especially important in the study of the mechanisms of cell differentiation and different disease.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work is supported by the NBRPC grant 2012CB316503, NSFC grant (61322310, 31371341, 91019016), FANEDD grant 201158, and Outstanding Tutors for doctoral dissertations of S&T project in Beijing no. 20111000304

Notes and references

[a] Bioinformatics Division, Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China. Fax: 86-10-62783552; Tel: 86-10-62794294 ext.808; E-mail: xwwang@tsinghua.edu.cn

[b] Department of Molecular and Cell Biology Center for Systems Biology, The University of Texas, Dallas 800 WestCampbell Road, RL11 Richardson, TX 75080-3021, USA. Fax: 972-883-5710; Tel: 972-883-2523; E-mail: michael.zhang@utdallas.edu

† Electronic Supplementary Information (ESI) available: [Supplementary Figures 1-3, Supplementary Tables 1-4]. See DOI: 10.1039/b000000x/

‡ These authors contributed equally.

1 D. P. Bartel, *Cell*, 2009, **136**, 215-233.
2 L. He and G. J. Hannon, Nature reviews. *Genetics*, 2004, **5**, 522-531.
3 N. Rajewsky, *Nature genetics*, 2006, **38** Suppl, S8-13.
4 J. Manikandan, J. J. Aarthi, S. D. Kumar and P. N. Pushparaj, *Bioinformation*, 2008, 2, 330-334.
5 A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky, *Nature genetics*, 2005, **37**, 495-500.
6 A. Kozomara and S. Griffiths-Jones, *Nucleic acids research*, 2011, **39**, D152-157.
7 D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson and D. P. Bartel, *Nature structural & molecular biology*, 2011, **18**, 1139-1146.
8 M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul and E. Segal, *Nature genetics*, 2007, **39**, 1278-1284.
9 F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li, *Nucleic acids research*, 2009, **37**, D105-110.
10 A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim and D. P. Bartel, *Molecular cell*, 2007, **27**, 91-105.
11 M. Thomas, J. Lieberman and A. Lal, *Nature structural & molecular biology*, 2010, **17**, 1169-1174.
12 W. Ritchie, S. Flamant and J. E. Rasko, *Nature methods*, 2009, **6**, 397-398.
13 A. Muniategui, R. Nogales-Cadenas, M. Vazquez, X. L. Aranguren, X. Agirre, A. Luttun, F. Prosper, A. *Pascual-Montano and A. Rubio, PloS one*, 2012, **7**, e30766.
14 J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey and Q. D. Morris, *Nature methods*, 2007, **4**, 1045-1049.
15 A. Bisognin, G. Sales, A. Coppe, S. Bortoluzzi and C. Romualdi, *Nucleic acids research*, 2012, **40**, W13-21.

16 S. Mukherji, M. S. Ebert, G. X. Zheng, J. S. Tsang, P. A. Sharp and A. van Oudenaarden, *Nature genetics*, 2011, **43**, 854-859.
17 L. Salmena, L. Poliseno, Y. Tay, L. Kats and P. P. Pandolfi, *Cell*, 2011, **146**, 353-358.
18 F. A. Karreth, Y. Tay, D. Perna, U. Ala, S. M. Tan, A. G. Rust, G. DeNicola, K. A. Webster, D. Weiss, P. A. Perez-Mancera, M. Krauthammer, R. Halaban, P. Provero, D. J. Adams, D. A. Tuveson and P. P. Pandolfi, *Cell*, 2011, **147**, 382-395.
19 P. Sumazin, X. Yang, H. S. Chiu, W. J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva and A. Califano, *Cell*, 2011, **147**, 370-381.
20 S. W. Chi, J. B. Zang, A. Mele and R. B. Darnell, *Nature*, 2009, **460**, 479-486.
21 Y. Hao, Z. J. Zhang, D. W. Erickson, M. Huang, Y. Huang, J. Li, T. Hwa and H. Shi, *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**, 12473-12478.
22 L. M. Wee, C. F. Flores-Jasso, W. E. Salomon and P. D. Zamore, *Cell*, 2012, **151**, 1055-1067.
23 J. Wen, B. J. Parker, A. Jacobsen and A. Krogh, *RNA*, 2011, **17**, 820-834.
24 Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**, 289-300.
25 Y. Zhang, L. X. Yan, Q. N. Wu, Z. M. Du, J. Chen, D. Z. Liao, M. Y. Huang, J. H. Hou, Q. L. Wu, M. S. Zeng, W. L. Huang, Y. X. Zeng and J. Y. Shao, *Cancer research*, 2011, **71**, 3552-3562.
26 X. Xiang, X. Zhuang, S. Ju, S. Zhang, H. Jiang, J. Mu, L. Zhang, D. Miller, W. Grizzle and H. G. Zhang, *Oncogene*, 2011, **30**, 3440-3453.
27 B. Zhang, X. X. Liu, J. R. He, C. X. Zhou, M. Guo, M. He, M. F. Li, G. Q. Chen and Q. Zhao, *Carcinogenesis*, 2011, **32**, 2-9.
28 U. Burk, J. Schubert, U. Wellner, O. Schmalhofer, E. Vincan, S. Spaderna and T. Brabletz, *EMBO Rep*, 2008, **9**, 582-589.
29 E. N. Howe, D. R. Cochrane and J. K. Richer, *Breast Cancer Res*, 2011, **13**, R45.
30 C. V. Pecot, R. Rupaimoole, D. Yang, R. Akbani, C. Ivan, C. Lu, S. Wu, H. D. Han, M. Y. Shah, C. Rodriguez-Aguayo, J. Bottsford-Miller, Y. Liu, S. B. Kim, A. Unruh, V. Gonzalez-Villasana, L. Huang, B. Zand, M. Moreno-Smith, L. S. Mangala, M. Taylor, H. J. Dalton, V. Sehgal, Y. Wen, Y. Kang, K. A. Baggerly, J. S. Lee, P. T. Ram, M. K. Ravoori, V. Kundra, X. Zhang, R. Ali-Fehmi, A. M. Gonzalez-Angulo, P. P. Massion, G. A. Calin, G. Lopez-Berestein, W. Zhang and A. K. Sood, *Nature communications*, 2013, **4**, 2427.
31 X. Sun, S. Qin, C. Fan, C. Xu, N. Du and H. Ren, *Oncology reports*, 2013, **29**, 2079-2087.
32 R. Lei, J. Tang, X. Zhuang, R. Deng, G. Li, J. Yu, Y. Liang, J. Xiao, H. Y. Wang, Q. Yang and G. Hu, *Oncogene*, 2013, DOI: 10.1038/onc.2013.65.
33 M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, *PloS one*, 2008, **3**, e3420.
34 B. Xie, Q. Ding, H. Han and D. Wu, *Bioinformatics*, 2013, **29**, 638-644.
35 M. Ouzounova, T. Vuong, P. B. Ancey, M. Ferrand, G. Durand, F. Le-Calvez Kelm, C. Croce, C. Matar, Z. Herceg and H. Hernandez-Vargas, *BMC genomics*, 2013, **14**, 139.
36 J. Krell, A. E. Frampton, J. Jacob, L. Castellano and J. Stebbing, *Pharmacogenomics*, 2012, **13**, 709-719.
37 R. Garzon, G. A. Calin and C. M. Croce, *Annual review of medicine*, 2009, **60**, 167-179.
38 D. Luo, J. M. Wilson, N. Harvel, J. Liu, L. Pei, S. Huang, L. Hawthorn and H. Shi, *Journal of translational medicine*, 2013, **11**, 57.
39 E. O'Day and A. Lal, *Breast cancer research : BCR*, 2010, **12**, 201.
40 C. A. Andorfer, B. M. Necela, E. A. Thompson and E. A. Perez, *Trends in molecular medicine*, 2011, **17**, 313-319.
41 C. Coronnello, R. Hartmaier, A. Arora, L. Huleihel, K. V. Pandit, A. S. Bais, M. Butterworth, N. Kaminski, G. D. Stormo, S. Oesterreich and P. V. Benos, *PLoS computational biology*, 2012, **8**, e1002830.
42 T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard and J. Kjems, *Nature*, 2013, **495**, 384-388.
43 K. Friend, Z. T. Campbell, A. Cooke, P. Kroll-Conner, M. P. Wickens and J. Kimble, *Nature structural & molecular biology*, 2012, **19**, 176-183.