

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

Junpeng Zhang¹, Thuc Duy Le², Lin Liu², Bing Liu³, Jianfeng He⁴, Gregory J Goodall⁵ and Jiuyong Li^{2,*}

¹Faculty of Engineering, Dali University, Dali, CN.

²School of Information Technology and Mathematical Sciences, University of South Australia, AU.

³Children's Cancer Institute Australia, Randwick NSW 2301, AU.

⁴Kunming University of Science and Technology, Kunming, CN.

⁵Centre for Cancer Biology, SA Pathology, SA 5000, AU.

In this file, we provide supplementary materials discussed in the Results sections.

1. Results

1.1. Overview of the validation

To assess the results obtained by using our method with the EMT and MCC data, the following validations and analysis are conducted.

From (Khan *et al.*, 2009), we obtain the transfection experimental data in 7 different cancer cell types, involving more than 20 different miRNAs and 40 unique siRNAs. The overlap between the miRNAs of either the EMT or the MCC dataset and transfected miRNAs from (Khan *et al.*, 2009) are 18 unique miRNAs.

Firstly, we use the transfection experimental result to validate the identified significant miRNA-mRNA causal interactions, specifically the target genes involved in the significant causal interactions. The validation is done in comparison with the target genes predicted by MicroCosm, as well as the target genes discovered by using the non condition-specific analysis based on IDA without considering the difference in the behavior of miRNAs across conditions. We also compare our method with five correlation methods (Pearson, Spearman, Kendall, Lasso and Elastic-net) in the number of validated targets.

Secondly, function, pathway and literature based analyses are done for the identified active miRNAs.

To show the effectiveness of our method, we also compare the results of active miRNAs obtained using our method with five non-causal methods (Pearson, Spearman, Kendall, Lasso and Elastic-net) and other five existing approaches (DIANA-mirExTra, Sylamer, MIR, miReduce and cWords).

Finally, we examine the patterns of the regulatory behavior of the identified active miRNAs and how they differ in different conditions.

1.2. Validation in comparison with MicroCosm

When the LFC cutoff is set to 1.0, the comparison result is shown in Figure S1. For the EMT dataset, for 8 out of the 11 miRNAs (i.e. excluding miR-141, miR-200a and miR-215), our method produces higher rate of experimentally confirmed target genes than MicroCosm does. For the MCC dataset, for 10 out of the 11 miRNAs (i.e. excluding miR-215), our method outperforms MicroCosm in the rate of

experimentally confirmed target genes.

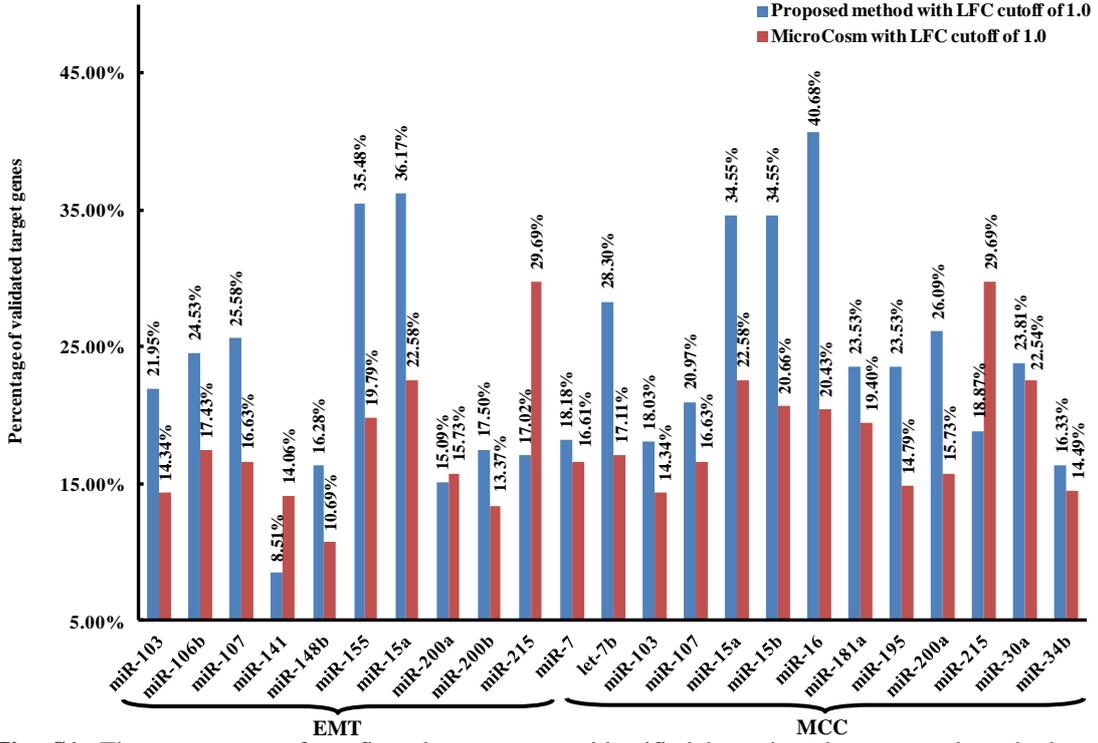


Fig. S1. The percentage of confirmed target genes identified by using the proposed method and MicroCosm in the EMT and MCC dataset (LFC=1.0).

We also conceive a hypergeometric (HG) statistic test to assess the significance of the validated miRNA-mRNA interactions. Let S be the number of possible target genes for n miRNAs in the dataset, K be the number of miRNA-mRNA interactions from transfection experiments for these miRNAs, N be the number of miRNA-mRNA interactions predicted by our method for these miRNAs, and x be the number of validated miRNA-mRNA interactions by transfection experiments for these miRNAs. The p -value of the validation results, which is the probability of the random method to achieve the same or better results than the proposed method, is calculated using the cumulative hypergeometric test formula:

$$p(X \geq x) = \sum_{i=x}^N \frac{\binom{K}{i} \binom{S-n-K}{N-i}}{\binom{S-n}{N}} \quad (1)$$

We use the Matlab function *hygepdf* to compute statistical significance level $p(X \geq x)$ of the validated miRNA-mRNA interactions. The validation of miRNA-mRNA interactions is statistically significant when $p(X \geq x)$ is less than, e.g., 0.05.

As shown in Table S1, the validated miRNA-mRNA interactions are all statistically significant (p -value < 0.05) in both the EMT and MCC datasets.

Table S1. Statistically significant level of validated miRNA-mRNA interactions for the 11 and 12 miRNAs in the EMT and MCC dataset, respectively.

Dataset	S	K_1/K_2	N	x_1/x_2	p_1/p_2
EMT($n=11$)	1126	4152/1791	489	214/103	9.8512E-07/3.5097E-05
MCC($n=12$)	1318	6361/2952	654	324/169	5.2288E-07/2.3328E-06

K_1 and K_2 are the number of miRNA-mRNA interactions from transfection experiments with LFC cutoff of 0.5 and 1.0, respectively. x_1 and x_2 denote the number of validated miRNA-mRNA interactions with LFC cutoff of 0.5 and 1.0 respectively. p_1 and p_2 represent significant level p -value with respect to LFC cutoff of 0.5 and 1.0, respectively.

1.3. Validation in comparison with non condition-specific analysis

As Illustrated in Figure S2, the cutoffs value of the \log_2 fold change in transfection experiments is set to 1.0. The comparison results also indicate that considering the difference of the condition of interest and the other conditions can help improve the prediction of miRNA-mRNA regulatory relationships.

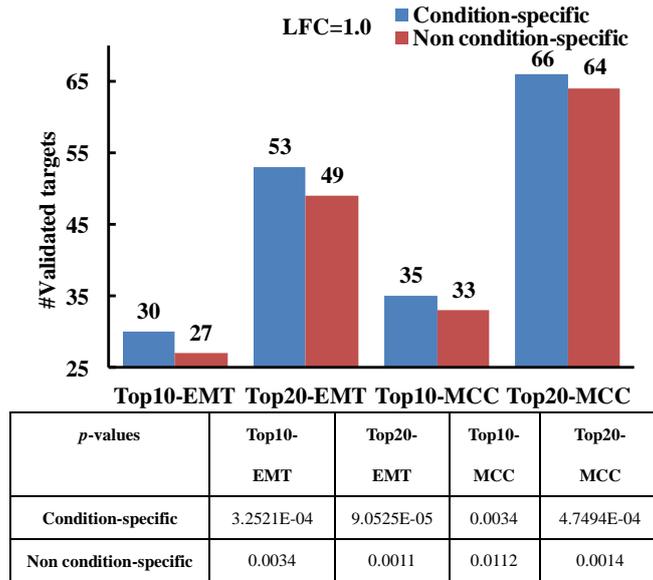


Fig. S2. Comparison between condition-specific and non condition-specific analyses in the number of validated targets, with LFC=1.0. The p -values of the validated targets are calculated using cumulative hypergeometric test.

1.4. Validation in comparison with five correlation methods

As Illustrated in Figure S3, the cutoffs value of the \log_2 fold change in transfection experiments is set to 1.0. The comparison results also show that our proposed method outperforms all the five correlation methods (Pearson, Spearman, Kendall, Lasso and Elastic-net) in the number of validated targets for all cases (Top10-EMT, Top20-EMT, Top10-MCC and Top20-MCC).

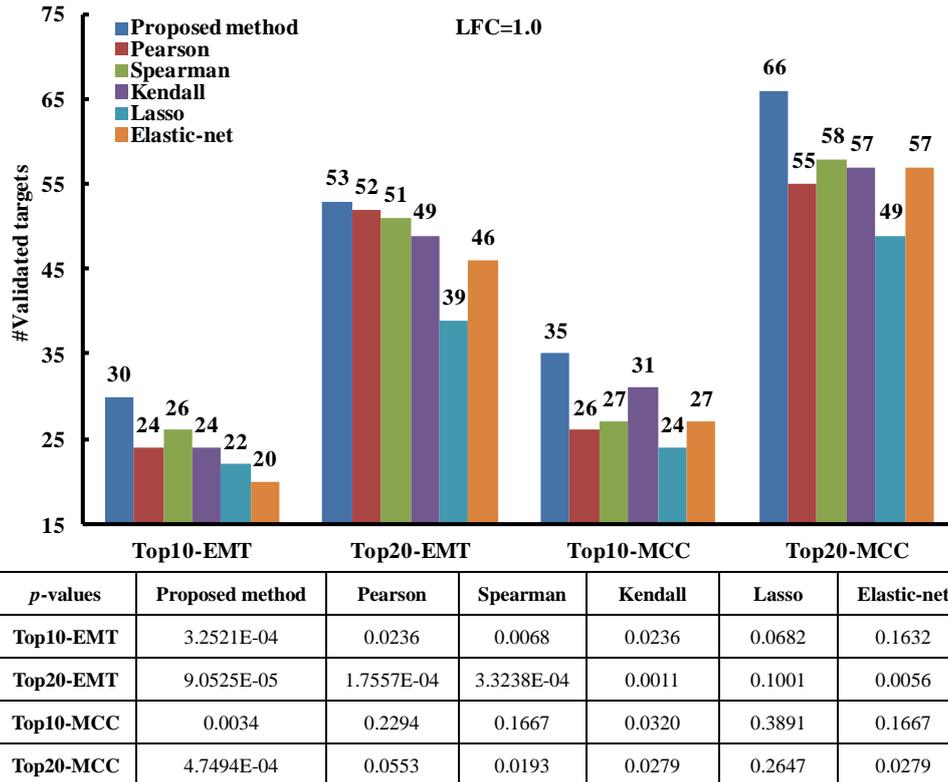


Fig. S3. Comparison between the proposed method and five correlation methods in the number of confirmed targets, with LFC=1.0. The *p*-values of the validated targets are calculated using cumulative hypergeometric test.

1.5. Correlations for finding active miRNAs

We use other five of correlation methods (Pearson, Spearman, Kendall, Lasso and Elastic-net) for finding active miRNAs. The R function used for the Pearson, Spearman and Kendall method is *cor* (package *stats*) with parameter *method*="pearson", "spearman", and "kendall", respectively. The *stats* package can be downloaded at <http://www.r-project.org>. The R function for Lasso and Elastic-net is *glmnet* (Friedman *et al.*, 2010) with parameter *alpha*=1 and 0.5, respectively.

1.6. Other existing methods for detecting active miRNAs

We compare our proposed method with five existing approaches, namely DIANA-mirExTra (Alexiou *et al.*, 2010), Sylamer (van Dongen *et al.*, 2008), MIR (Cheng and Li, 2008), miReduce (Sood *et al.*, 2006) and cWords (Rasmussen *et al.*, 2013). The web link of DIANA-mirExTra is <http://diana.cslab.ece.ntua.gr/hexamers/>, and it identifies overrepresented 6nt long motifs (hexamers) on the 3'UTR sequences of deregulated genes. The inputs of DIANA-mirExTra contain two lists: a list of differentially expressed mRNAs and a list of non-differentially expressed (background) mRNAs. The Java Graphical Interface of Sylamer can be obtained from <http://www.ebi.ac.uk/research/enright/software/sylamer>, and the input of it is a list of ranked differential mRNAs by adjusted *p*-value from *limma* (Smyth, 2005). The C++ program for MIR is available at <http://www.dartmouth.edu/~chaocheng/software/InferMiRNA/infermir.html>, and two input files are needed: the differentially expressed mRNAs with log fold change file from *limma* and the

miRNA-gene binding affinity data file based on the predicted binding energy by miRanda (Betel *et al.*, 2008). The Perl script of miReduce implementation can be obtained at https://www.mdc-berlin.de/10615841/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/mireduce. The three inputs of it are differentially expressed mRNAs with log fold change from *limma*, the 3'UTR sequences of these differentially expressed mRNAs and miRNA sequences of differentially expressed miRNAs. The online link of cWords is from <http://servers.binf.ku.dk/cwords/>, and the input is a ranked list of differentially expressed mRNAs by adjusted *p*-value from *limma*. For Sylamer, miReduce and cWords, we combine the results of seed region from 6nt to 8nt as final results of them. For all methods, we set the *p*-value cutoff of 0.05 to detect active miRNAs within differentially expressed miRNAs in the EMT and MCC datasets.

References

- Alexiou,P. *et al.* (2010) The DIANA-mirExTra web server: from gene expression data to microRNA function, *PLoS One*, **5**, e9171.
- Betel,D. *et al.* (2008) The microRNA.org resource: targets and expression, *Nucleic Acids Res.*, **36**, D149-D153.
- Cheng,C. and Li,L.M. (2008) Inferring microRNA activities by combining gene expression with microRNA target prediction, *PLoS One*, **3**, e1989.
- Friedman,J. *et al.* (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.*, **33**, 1-22.
- Khan,A.A. *et al.* (2009) Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs, *Nat. Biotechnol.*, **27**, 549-55.
- Rasmussen,S.H. *et al.* (2013) cWords - systematic microRNA regulatory motif discovery from mRNA expression data, *Silence*, **4**, 2.
- Sood,P. *et al.* (2006) Cell-type-specific signatures of microRNA on target mRNA expression, *Proc. Natl. Acad. Sci. USA*, **103**, 2746-2751.
- Smyth,G.K. (2005) *Limma: Linear Models for Microarray Data*. Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, New York, 397-420.
- van Dongen, S. *et al.* (2008) Detecting microRNA binding and siRNA off-target effects from expression data, *Nat. Methods*, **5**, 1023-1025.