Research Article

A Filtering Method for Identification of Significant Target mRNAs of Coexpressed and Differentially Expressed MicroRNA Clusters

Su Yeon Lee,[1] Soo-Yong Shin,[2] Young Jo Yoon,[3] and Yu Rang Park [iD][4]

[1]*Bioinformatics Team, Samsung SDS, Seoul, Republic of Korea*
[2]*Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea*
[3]*Office of Clinical Research Information, Asan Medical Center, Seoul, Republic of Korea*
[4]*Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea*

Correspondence should be addressed to Yu Rang Park; yurang.park@amc.seoul.kr

MicroRNA (miRNA) binding is primarily based on sequence, but structure-specific binding is also possible. Various prediction algorithms have been developed for predicting miRNA target genes; the results, however, have relatively high levels of false positives, and the degree of overlap between predicted targets from different methods is poor or null. We devised a new method for identifying significant miRNA target genes from an extensive list of predicted miRNA target gene relationships using hypergeometric distributions. We evaluated our method in statistical and semantic aspects using a common miRNA cluster from six solid tumors. Our method provides statistically and semantically significant miRNA target genes. Complementing target prediction algorithms with our proposed method may have a significant synergistic effect in finding and evaluating functional annotation and enrichment analysis for miRNA.

1. Introduction

MicroRNAs (miRNAs) are a class of small RNAs that regulate gene expression at the transcript level, protein level, or both [1–4]. miRNAs modulate gene activity and are aberrantly expressed in most types of cancers [5]. Due to their small size and stability, miRNAs can also be measured in biologic fluids such as plasma and serum and can serve as circulating biomarkers [6–9]. In spite of the continuous attempts to identify miRNAs and to elucidate their basic mechanisms of action, little is understood about their biological functions.

Because of the regulatory role of miRNAs and lack of direct functional annotation to miRNAs, functional enrichment methods for miRNAs rely on their target gene's functional annotations [10–12]. For instance, if the target genes of a specific miRNA are significantly enriched with a set of Gene Ontology (GO) terms, it is reasonable to infer that the miRNA is also involved in the same GO annotations. Several studies on miRNAs have used "predicted target-genes' functional annotation-based" miRNA function prediction strategy [13, 14]; these methods, however, are limited in that they do not consider the many-to-many-to-many tripartite network topology among miRNAs, target genes, and GO annotation [15–17]. In our previous work, we proposed three types of measures (miRNA-centric, target gene-centric, and target link-centric) and a novel index for calculating the functional enrichment of miRNA. Among the three measures, the miRNA-centric measurements showed the best performance [18]. We also found that the miRNA's intrinsic properties of multiplicity and cooperability may be correctly modeled by combined hypergeometric distributions.

Most of the miRNA-to-mRNA target links are estimated by prediction algorithms. However, these algorithms generate a relatively high level of false positives [19], and the

degree of overlap between predicted targets from different methods is often poor or null [13, 20]. Studies in this field have developed multiple databases with enormous amount of miRNA-to-target mRNA relationships computed using diverse algorithms [21], whereas only a few experimentally validated targets are available [22, 23]. In light of this circumstance, there is an unmet need for a method for identifying a significant miRNA target from a copious amount of predicted miRNA-target mRNA pairs.

According to miRNA characteristics (multiplicity and cooperatively activities), we employed the hypergeometric distribution to identify significant miRNA target genes from the extensive list of miRNA target genes. We also evaluated the performance of our method in two aspects: statistical significance and functional enrichment.

2. Methods

2.1. Computational Methods. To find significant target mRNAs from the input miRNA cluster, we first searched for target mRNAs from all miRNA members within the input cluster from the miRNA target database. For each targeted mRNAs, we then calculated the numbers of miRNAs that have target relationships (p_i) with the mRNA and those that do not (p_j) using the two-by-two contingency table. We also calculated the numbers of miRNAs not in the input miRNA cluster by dividing those that have a target relationship (p_k) with the mRNA and those that do not (p_l), as shown in Table 1. Functional enrichment was tested from this contingency table using a hypergeometric distribution. The hypergeometric distribution applies to sampling without replacement from a finite population whose elements can be classified into two mutually exclusive categories: has/does not have a target relationship.

We then calculated the adjusted p values using the Bonferroni correction. Finally, for evaluating our methods, 10,000 simulated mRNA sets of the same size were also randomly sampled from the target mRNAs of the input miRNA cluster.

Using hypergeometric distribution, we assumed that the coordinated function among miRNAs within a cluster is valid when these miRNAs are regulated or annotated by common factors such as same target mRNA, Gene Ontology, or pathway.

2.2. Data Set: miRNA Clusters. We obtained an miRNA set created by Volinia et al. [24] that has differentially expressed sets of up- or downregulated miRNAs in six solid tumor samples. Among the miRNA clusters, we selected an miRNA cluster composed of 57 miRNAs by prediction analysis of microarray (PAM) in six solid tumor samples versus normal tissues. The complete list of 57 miRNAs is in Additional File 1.

2.3. Creating Variations of the miRNA-mRNA Target Pair. To build the miRNA-mRNA target pair, we chose three representative miRNA databases: TarBase (Data release 6.0,

Table 1: 2×2 contingency table of miRNA frequency calculated for each target mRNA.

Target mRNA	Input miRNA cluster	
	In cluster	Not in cluster
Has a target relationship	p_i	p_k
Does not have a target relationship	p_j	p_l

February 28th, 2014) [23], MirTarBase (Data release 4.5, February 28th, 2014) [22], and mirDIP (Data release 1.0, February 12th, 2014) [14]. The TarBase and MirTarBase databases provide experimentally validated miRNA-target interaction data and evidence level (strong and less strong) of each interaction. The mirDIP database provides *in silico*-predicted miRNA-target interaction data from six established target prediction algorithms and 12 miRNA prediction databases. GO annotation of miRNA-target-mRNA was obtained from the Entrez Gene database. We excluded GO associations with ND (no biological data) and NR (not recorded) evidence code. Detailed processes are provided in Figure 1.

2.4. Statistical and Semantic Evaluation Measurement. To evaluate our methods, we compared the performance in terms of statistical significances between a significant 317 mRNA cluster and randomly simulated 10,000 clusters. Each randomly generated cluster had the same size as the significant mRNA set. GO functional enrichment analysis was performed for all mRNA sets using GO annotations retrieved from the NCBI Entrez Gene database. We filtered out 339 GO terms that were greater than 0.05. The resulting lists of 377 GO terms are shown in Additional File 2. To reduce the number of GO terms, enriched GO terms and p values were submitted to REduce and Visualize GO (REViGO).

We computed the average log p values of ranked GO term sets from functional enrichment analysis of the significant mRNA set and the randomly simulated 10,000 mRNA sets. These average log p values were then used for comparing the performance. Functional enrichment was performed using GO annotations of mRNA from NCBI Entrez gene [25]. Average log p values of ranked GO terms were based on the general assumption that highly significant GO terms are more desirable because it means the members of the cluster are highly correlated to each other. For semantic evaluation of the significant mRNA set, we used REViGO, which is a web-based system that summarizes a list of GO terms by finding a representative subset of the terms using the semantic similarity-based clustering algorithm [26].

3. Results

3.1. Statistical Significances. For evaluating our methods, we compare performance in terms of statistical significances between a significant 317 mRNA cluster and randomly
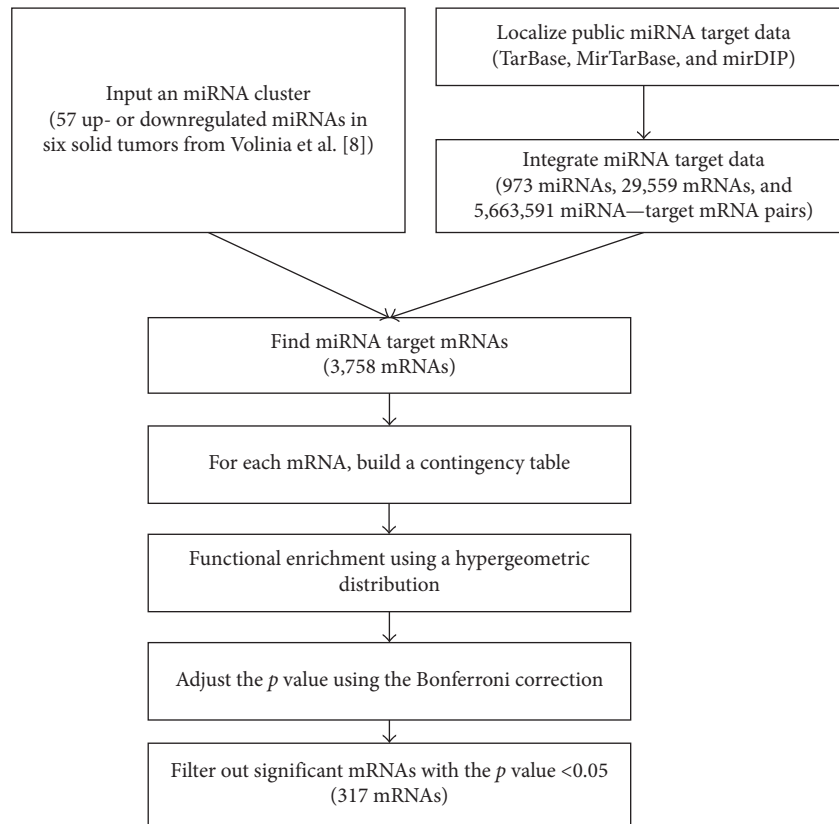
Input an miRNA cluster
(57 up- or downregulated miRNAs in
six solid tumors from Volinia et al. [8])

Localize public miRNA target data
(TarBase, MirTarBase, and mirDIP)

Integrate miRNA target data
(973 miRNAs, 29,559 mRNAs, and
5,663,591 miRNA—target mRNA pairs)

Find miRNA target mRNAs
(3,758 mRNAs)

For each mRNA, build a contingency table

Functional enrichment using a hypergeometric
distribution

Adjust the p value using the Bonferroni correction

Filter out significant mRNAs with the p value <0.05
(317 mRNAs)

FIGURE 1: Flowchart of the computational method for identifying significant miRNA target genes.

simulated 10,000 clusters. Each randomly generated cluster has the same size with the significant mRNA cluster. All mRNA cluster performed GO functional enrichment analysis using GO annotation from the NCBI Entrez gene. Figure 2 shows the distributions of average log p values for the rank of GO terms which belong to the biological process category. The significant mRNA set is shown as the red dotted graph. Randomly simulated 10,000 clusters are shown as box plots. The significant mRNA set showed a higher average log p value than random clusters did, which indicated that the members of the cluster highly correlated and meaningfully composed.

3.2. Gene Ontology Analysis of Significant miRNA Target Genes. Using the UniProt database as background and the default semantic measure (SimRel), our analysis clearly showed that biological processes associated with cancer metabolism, regulation of cell death and apoptotic process, and negative regulation of autophagy were significantly overrepresented.

Figure 3 shows the REViGO scatter plot represented in a two-dimensional space derived by applying multidimensional scaling to a matrix of GO terms semantic similarities. The resulting lists of 339 GO terms along with their p values were further summarized by the REViGO reduction analysis tool that condenses the GO description by removing redundant terms. The remaining
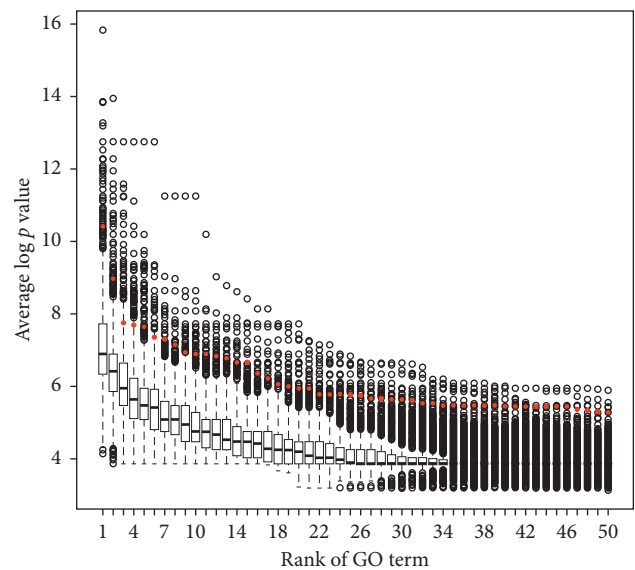

FIGURE 2: Evaluation of statistical significance across thresholds. The significant mRNA set and randomly simulated 10,000 clusters are shown as red dotted graphs and box plots, respectively.

terms after the redundancy reduction were plotted in a two-dimensional space. Bubble color indicates the p value (legend in the upper right-hand corner): the two ends of the colors are red and blue, which represent lower and higher p values, respectively. Size indicates the
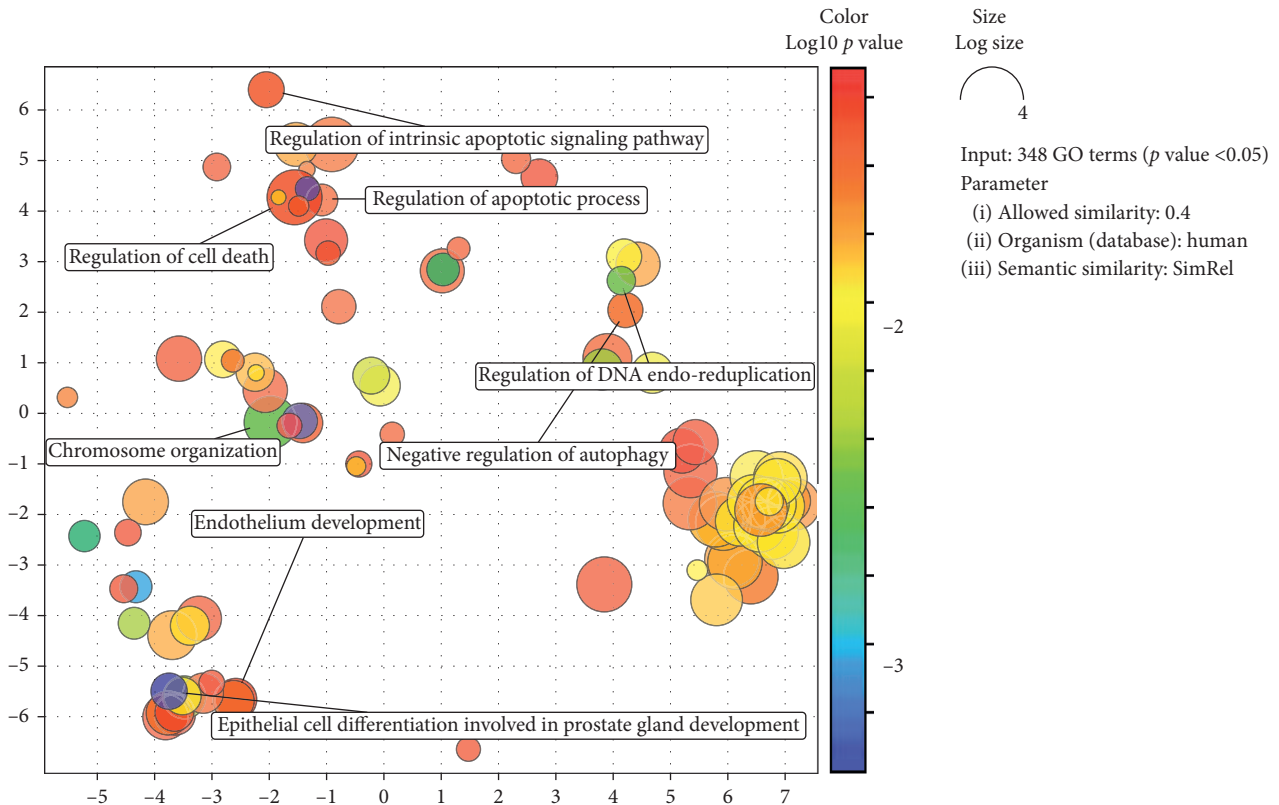
Figure 3: REViGO scatter plot for the significant mRNA set.

4. Discussion

Functional enrichment studies for miRNA expression are performed in three steps: (1) selecting differentially expressed miRNAs, (2) finding their target mRNA, and (3) carrying out analysis of mRNA set overrepresentation [27]. Functional enrichment studies for miRNAs are mostly based on the annotation of target mRNA; however, due to improvements in the miRNA target prediction algorithms, a large number of target mRNAs are predicted. Considering this, filtering out significant mRNAs using a stable statistical method is of great importance. In this study, we proposed a method for identifying the significant miRNA target mRNA from the miRNA cluster. The proposed method was verified by functional enrichment analysis of differentially expressed or coexpressed miRNA clusters.

Inaccurate functional enrichment methods are a hindrance in increasing clinical utility for miRNAs, such as miRNA-based biomarkers or predictors [28, 29]. Several tools have been recently established for direct prediction of miRNA functions [10, 30]; however, these methods do not consider the regulatory or indirect functions of miRNAs, such as regulation or inhibition of target genes [31]. The

intrinsic properties of multiplicity and cooperative activities of miRNAs should be considered while annotating the miRNA function. miRGator v3.0 is a tool created considering these characteristics and allows the user to manually select miRNAs and target mRNAs [32]. However, such tools are only useful when the number of miRNA and mRNA pairs is small.

The limitation of the proposed method is that the hypergeometric distribution has a significant effect when members belonging to an miRNA cluster are regulated by common factors such as the target mRNA, GO, and pathway. The proposed method constructs a target mRNA set with statistical significance by receiving miRNA clusters with similar expression characteristics. The assumption of hypergeometric is well suited to this problem because the cluster-received input already has similar characteristics.

The miRNA target prediction algorithms were modified to generate more accurate results based on the expanding understanding of the molecular mechanism of miRNA regulation. Nevertheless, identifying significant target mRNAs from the numerous, uncurated miRNA target links remain as a problem. Our method is based on computationally identifying statistically significant mRNAs using predicted or experimentally validated target relationships. Complementing target prediction algorithms with our proposed method may have significant synergistic effects in finding and evaluating functional annotation and enrichment analysis for miRNA.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Su Yeon Lee and Soo-Yong Shin have contributed equally to this work.

Acknowledgments

Supplementary Materials

Supplementary 1. Additional File 1: the complete list of 57 miRNAs.

Supplementary 2. Additional File 2: the resulting lists of 377 GO terms.

References

[1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] P. T. Nelson, D. A. Baldwin, L. M. Scearce, J. C. Oberholtzer, J. W. Tobias, and Z. Mourelatos, "Microarray-based, high-throughput gene expression profiling of microRNAs," *Nature Methods*, vol. 1, no. 2, pp. 155–161, 2004.

[3] E. C. Lai, "microRNAs: runts of the genome assert themselves," *Current Biology*, vol. 13, no. 23, pp. R925–R936, 2003.

[4] K. Sun and E. C. Lai, "Adult-specific functions of animal microRNAs," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 535–548, 2013.

[5] M. V. Iorio and C. M. Croce, "MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review," *Embo Molecular Medicine*, vol. 9, no. 6, p. 852, 2017.

[6] X. Chen, Y. Ba, L. Ma et al., "Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases," *Cell Research*, vol. 18, no. 10, pp. 997–1006, 2008.

[7] P. S. Mitchell, R. K. Parkin, E. M. Kroh et al., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10513–10518, 2008.

[8] G. Sozzi, M. Boeri, M. Rossi et al., "Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study," *Journal of Clinical Oncology*, vol. 32, no. 8, pp. 768–773, 2014.

[9] R. Belzeaux, R. X. Lin, and G. Turecki, "Potential use of microRNA for monitoring therapeutic response to antidepressants," *CNS Drugs*, vol. 31, no. 4, pp. 253–262, 2017.

[10] I. Ulitsky, L. C. Laurent, and R. Shamir, "Towards computational prediction of microRNA function and activity," *Nucleic Acids Research*, vol. 38, no. 15, p. e160, 2010.

[11] Y. Q. Wu, D.-J. Chen, H.-B. He et al., "Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs," *PLoS One*, vol. 7, no. 1, Article ID e30988, 2012.

[12] Y. Xiao, C. Xu, J. Guan et al., "Discovering dysfunction of multiple microRNAs cooperation in disease by a conserved microRNA co-expression network," *PLoS One*, vol. 7, no. 2, Article ID e32201, 2012.

[13] S. Ekimler and K. Sahin, "Computational methods for microRNA target prediction," *Genes*, vol. 5, no. 3, pp. 671–683, 2014.

[14] E. A. Shirdel, W. Xie, T. W. Mak, and I. Jurisica, "NAVi-GaTing the micronome–using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs," *PLoS One*, vol. 6, no. 2, Article ID e17429, 2011.

[15] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan, "Inference of miRNA targets using evolutionary conservation and pathway analysis," *BMC Bioinformatics*, vol. 8, no. 1, p. 69, 2007.

[16] J. Xu and C. Wong, "A computational screen for mouse signaling pathways targeted by microRNA clusters," *RNA*, vol. 14, no. 7, pp. 1276–1283, 2008.

[17] Y Gusev, "Computational methods for analysis of cellular functions and pathways collectively targeted by differentially expressed microRNA," *Methods*, vol. 44, no. 1, pp. 61–72, 2008.

[18] S. Y. Lee, K. A. Sohn, and J. H. Kim, "MicroRNA-centric measurement improves functional enrichment analysis of co-expressed and differentially expressed microRNA clusters," *BMC Genomics*, vol. 13, no. 7, p. S17, 2012.

[19] L. S. Hon and Z. Zhang, "The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression," *Genome Biology*, vol. 8, no. 8, p. R166, 2007.

[20] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nature Methods*, vol. 3, no. 11, pp. 881–886, 2006.

[21] W. Ritchie and J. E. Rasko, "Refining microRNA target predictions: sorting the wheat from the chaff," *Biochemical and Biophysical Research Communications*, vol. 445, no. 4, pp. 780–784, 2014.

[22] S. D. Hsu, F.-M. Lin, W.-Y. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D163–D169, 2011.

[23] T. Vergoulis, I. S. Vlachos, P. Alexiou et al., "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support," *Nucleic Acids Research*, vol. 40, no. D1, pp. D222–D229, 2012.

[24] S. Volinia, G. A. Calin, C.-G. Liu et al., "A microRNA expression signature of human solid tumors defines cancer gene targets," *Proceedings of the National Academy of Sciences*, vol. 103, no. 7, pp. 2257–2261, 2006.

[25] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 43, no. D1, 2014.

[26] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO summarizes and visualizes long lists of gene ontology terms," *PLoS One*, vol. 6, no. 7, Article ID e21800, 2011.

[27] F. Garcia-Garcia, J. Panadero, J. Dopazo, and D. Montaner, "Integrated gene set analysis for microRNA studies," *Bioinformatics*, vol. 32, no. 18, pp. 2809–2816, 2016.

[28] T. Bleazard, J. A. Lamb, and S. Griffiths-Jones, "Bias in microRNA functional enrichment analysis," *Bioinformatics*, vol. 31, no. 10, pp. 1592–1598, 2015.

[29] W. Ritchie, S. Flamant, and J. E. Rasko, "Predicting micro-RNA targets and functions: traps for the unwary," *Nature Methods*, vol. 6, no. 6, pp. 397-398, 2009.

[30] M. Lu, B. Shi, J. Wang, Q. Cao, and Q. Cui, "TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs," *BMC Bioinformatics*, vol. 11, no. 1, p. 419, 2010.

[31] L. P. Lim, N. C. Lau, P. Garrett-Engele et al., "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs," *Nature*, vol. 433, no. 7027, pp. 769–773, 2005.

[32] S. Cho, I. Jang, Y. Jun et al., "miRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting," *Nucleic Acids Research*, vol. 41, no. D1, pp. D252–D257, 2013.