# MiRaGE: Inference of Gene Expression Regulation via MicroRNA Transfection II

Y-h. Taguchi[1] and Jun Yasuda[2]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan
[2] COE Fellow, Graduate School of Medicine, Tohoku University, Sendai 980-8575, Japan
jun.yasuda@jfcr.or.jp, tag@granular.com

**Abstract.** How each microRNA regulates gene expression is unknown problem. Especially, which gene is targeted by each microRNA is mainly depicted via computational method, typically without biological/ experimental validations. In this paper, we propose a new computational method, MiRaGE, to detect gene expression regulation via miRNAs by the use of expression profile data and miRNA target prediction. This method is tested to miRNA transfection experiments to tumor cells and succeeded in inference of transfected miRNA as only one miRNA with significant $P$-values for the first time.

**Keywords:** MicroRNA, target genes, tumor, computational inference.

## 1 Introduction

MicroRNAs (miRNAs) are post-transcriptional regulators of gene expression. It binds to target messenger RNAs (mRNAs) through complementary sequences in the three prime untranslated regions (3 UTRs) of the mRNA, and consequently suppresses the expression of the mRNAs. miRNAs are short ($19 \sim 22$ bases) RNA molecules and abundant in many human cells. The human genome may encode over 1,000 miRNAs, and the coverage by all possible miRNAs may be about 60 % of mammalian genes.

On the other hand, how a miRNA regulates its target genes and which genes are regulated by a miRNA is unclear. Especially, the later is mainly depicted via computational prediction[1], without any biological/experimental validations. There are some direct ways to investigate the bindings of mRNAs to the miRNA-protein complexes, e.g., HITS-CLIP[2] but capability of these methods for identification of miRNA-mRNA relationship is limited since it is unlikely that all the potential target mRNAs for a miRNA simultaneously express in a cell.

Another experimental way to detect miRNA target genes is to analyze the difference of gene expression profiles with or without the transfection of the miRNA to a cell line. However, it is unrealistic to test all the miRNAs with this method because it is time and money consuming.

In this paper, we describe a computational method to detect miRNAs which regulate the cellular transcriptomes in a cell in response to extracellular stimuli by analyzing the difference of gene expression profiles and computational

D.-S. Huang et al. (Eds.): ICIC 2011, LNBI 6840, pp. 129–135, 2012.

miRNA target predictions. By this retrograde method, we can narrow down the miRNAs that actually play some biological roles. Simple expression profiling of miRNAs may not be efficient to find the critical ones. Moreover, we may deduce potentially important miRNAs from the large accumulation of gene expression profile databases such as GEO by our method. It seems to be difficult to go back to the RNA samples for obtaining the miRNA expression profiles corresponds to those old profile data.

We would like to obtain the proof of concept for our methods through the analyses of the gene expression profiles with or without the transfection of single miRNAs, which actually induced cell cycle arrest in the human lung cancer cell lines, and our algorithm can quite frequently predict the transfected miRNA.

## 2 Materials and Methods

### 2.1 Gene Expression Data for Transfection Experiment

We have downloaded transfection experiment[3] data set, CBX79, which is deposited and revised at CIBEX data base[4] at Center for Information Biology and DNA Data Bank of Japan (DDBJ), National Institute of Genetics (Mishima, Japan). It includes two biological replicates of negative, mir-107, 185, and let-7a transfection experiments, one day and three days after the transfection. Expression of 45015 genes (probes) are listed.

Since our method is robust for the random noise of gene expression variance and the overall distribution of gene expression between technical replicates should be within the acceptable range, we did not apply any normalization procedure.

### 2.2 Inference of miRNA which Regulates Target Genes Significantly

The way to detect miRNA whose target genes are significantly differently expressed between negative control and treated one is as follows. First, we have downloaded a list of conserved seed match in 3' UTRs of genes to each miRNA[1][5]. This includes 162 miRNA families. The reason why we do not use major target gene list, e.g., targetScan[6], PITA[7], pictar[8], miranda[9], and others, but use seed match is because Alexiou *et al*[10] recently reported that simple seed match often resuits in higher $F$-measure than more complicated estimations of target genes (see e. g. Supplementary Data[10]). Then we have picked up genes which has at least one seed match for any miRNAs in those 3' UTRs. Then, 13270 genes remain.

Hereafter, we denote a set of these genes as $G$. Next, for each miRNA, $m$, we have listed genes which has at least one seed match in 3' UTRs. We denote this set of genes as $G_m$, where $m$ denotes one of miRNA families. Also we define a set of genes, $G'_m \equiv G \setminus G_m$, which is a set of genes included into $G$, but not into

[1] `http://hollywood.mit.edu/targetrank/hsa_conserved_miR_family_ranked_targets.txt`

$G_m$. After denoting expression of gene $g$ under transfection of miRNA $m_0$, $m_0$ is one of mir-107, 185, let-7a, and Negative Control (NC), as $x_g^{m_0}$, we compute gene expression difference between post-miRNA transfection and NC,

$$\Delta x_g^{m_0} \equiv \log x_g^{m_0} - \log x_g^{NC}.$$

Then we apply paired $t$-test between $\{\Delta x_g^{m_0} \mid g \in G_m\}$ and $\{\Delta x_g^{m_0} \mid g \in G'_m\}$. $P$-value, $P_m$, is computed for each miRNA, $m$. In the previous work[11], we have employed two sided $t$-test. However, in this study, we employ one-sided $t$-test, which checks if expression of genes in $G_m$ is significantly *suppressed* than that of genes in $G'_m$. After applying FDR correction (BH method[12]) to 162 $P$-values, we have selected $m$s whose FDR corrected $P$-value is less than 0.05 as miRNA which regulates target genes significantly. For $t$-test, we have used t.test module in base package in R[13].

### 2.3 Coincidence between Biological Replicates

We have also checked if two biological replicates satisfy reproducibility in three ways. Firstly, we employed Pearson correlation coefficients between log transformed $P_m$s and secondly, Spearman correlation coefficients between them. $P$-values for these are computed as well as 95 percentile significant interval for the form. Thirdly, we analyzed coincidence between significant miRNAs, $m$s between two biological replicates. If the first(second) replicates have $m_1(m_2)$ significant miRNAs and $m_{12}$ miRNAs are selected for both replicates, $P$-value computed by binomial distribution $P(m_1,\mathrm{N},m_2/N)$ or $P(m_2,\mathrm{N},m_1/N)$, where $P(x, N, p)$ is the probability that $x$ among $N$ is selected when the probability of selection is $p$. $N$ is the number of genes in $G$.

We have used cor.test module in base package of R for $P$-values of correlation coefficients and pbinom module for binomial distribution.

### 2.4 Significant Overlap between Target Genes

To compute $P$-values of accidental agreement between target genes, $P^O_{m,m'}$ of miRNAs m and m', we have employed binomial distribution $P(n_{mm'}, n_m, n_{m'}/N)$, where $n_{mm'}$ is the number of co-target genes, $n_m(n_{m'})$ is the number of target genes of m(m').

## 3 Results

Independent of conditions, i.e., date and transfected miRNA, our method almost always gets non-empty set of significant miRNAs, $m$s (see Table 1). Thus, in principle, our method can detect miRNA regulation of gene expression. Table 1 shows which miRNA significantly regulates target genes. Most remarkably, $P_m$ has the strong tendency to become smallest when $m = m_0$. If we compare the number of the significant miRNAs found in the previous study[11], it is clear that the drastic improvement took place. For example, all analyses of mir-107 transfection experiment gave us several tens of significant miRNAs in the

**Table 1.** All of significant miRNAs, one day or three days after trasnfection. Bold characters are those transfected. The one in parentheses means not significant but with the smallest $P$-values.

| | | replicate 1 | | replicate 2 | |
|---|---|---|---|---|---|
| transfection day | | miRNA | $P_m$ | miRNA | $P_m$ |
| mir-107 | 1 | **miR-103/107** | $1.32 \times 10^{-8}$ | **miR-103/107** | $4.96 \times 10^{-9}$ |
| | 3 | **miR-103/107** | $3.31 \times 10^{-5}$ | — | — |
| mir-185 | 1 | **miR-185** | $4.32 \times 10^{-12}$ | **miR-185** | $8.74 \times 10^{-18}$ |
| | | miR-326 | $1.57 \times 10^{-6}$ | miR-326 | $2.42 \times 10^{-6}$ |
| | | miR-491 | $1.99 \times 10^{-4}$ | miR-491 | $1.05 \times 10^{-4}$ |
| | | miR-124.2/506 | $3.78 \times 10^{-4}$ | miR-34b | $1.68 \times 10^{-4}$ |
| | | miR-7 | $6.15 \times 10^{-4}$ | miR-122 | $2.79 \times 10^{-4}$ |
| | | miR-485-5p | $7.68 \times 10^{-4}$ | miR-124.2/506 | $6.77 \times 10^{-4}$ |
| | | miR-339 | $1.57 \times 10^{-6}$ | miR-331 | $9.34 \times 10^{-4}$ |
| | | — | — | miR-34/449 | $1.41 \times 10^{-3}$ |
| | | — | — | miR-485-5p | $1.51 \times 10^{-3}$ |
| | 3 | [**miR-185** ] | $5.48 \times 10^{-4}$ | — | — |
| let-7a | 1 | **let-7/98** | $2.18 \times 10^{-14}$ | **let-7/98** | $4.77 \times 10^{-16}$ |
| | | miR-196 | $1.97 \times 10^{-4}$ | miR-196 | $1.40 \times 10^{-4}$ |
| | 3 | **let-7/98** | $2.66 \times 10^{-8}$ | **let-7/98** | $5.47 \times 10^{-5}$ |

previous study but only the transfected one in the present study, indicating that the improvement in the present study does not deteriorate the sensitivity. Hereafter, we name our present method as MiRaGE, which is, at present, only method to detect transfected miRNA as unique miRNA whose target genes are expressed significantly compared with non-target genes.

MiRaGE has two unique features compared with the previous methods.

1. Comparison between $G_m$ and $G'_m$. In other method, comparison is done between $G_m$ and all other genes whose expression is measured.
2. Target gene table is obtained based upon only seed match.

The reason why these two are important will be discussed later.

Table 2 shows the results of several statistical tests for the coincidence between biological replicates. For all six cases, at least two out of three tests give the significant $P$-values $< 0.05$. For most of cases, $P$-values is almost 0 within numerical accuracy ($P < 2.2 \times 10^{-16}$). Thus, biological replicates are good enough for inference of miRNA transfection.

## 4 Discussion

Although $P_{m_0}$ is mostly the smallest, $P_m$ with $m \neq m_0$ also can sometimes take the value as small as $P_{m_0}$. For example, for two replicates at one day after mir-185 transfection (see Table 1) there are seven and nine significant miRNAs respectively. The materials analyzed in this study are experimental. Hence it is

**Table 2.** Comparison of two biological replicates. Bold numbers indicate significant $P$-values ($< 0.05$). Bold asterisks (*****) indicate $P < 2.2 \times 10^{-16}$.

| Transfection | mir-107 | | mir-185 | | let-7a | |
|---|---|---|---|---|---|---|
| Time | day 1 | day 3 | day 1 | day 3 | day 1 | day 3 |
| Pearson | 0.94 | 0.25 | 0.89 | 0.54 | 0.95 | 0.74 |
| | 95 % confidence interval | | | | | |
| lower | 0.92 | 0.39 | 0.92 | 0.64 | 0.96 | 0.81 |
| upper | 0.96 | 0.10 | 0.86 | 0.42 | 0.93 | 0.67 |
| $P$-value | * | **0.0016** | * | **$1.66 \times 10^{-13}$** | * | * |
| Spearman | 0.63 | 0.76 | 0.65 | 0.52 | 0.55 | 0.32 |
| $P$-value | * | * | * | * | * | **$4.73 \times 10^{-5}$** |
| | # of significant miRNAs | | | | | |
| common | 1 | 0 | 5 | 0 | 2 | 1 |
| replicate 1 | 1 | 1 | 7 | 1 | 2 | 1 |
| replicate 2 | 1 | 0 | 9 | 0 | 2 | 1 |
| $P$-value | * | — | **$1.96 \times 10^{-7}$** | — | * | * |

clearly distinguishable the true and false positives. The seed sequences of false positive miRNAs are different from that of the transfected one. The causes of false positives may be the secondary effect of transfected miRNA. The downregulation of target messengers by miRNA transfection could cause downregulation of other miRNAs targets or upregulation of false positive miRNAs. Alternatively, co-targeting between transfected miRNA and other miRNAs could cause the false positives: the miRNAs with overlapped targets with transfected one would show smaller $P$-value than other endogenous miRNAs. Theoretically false positives by co-targeting are unavoidable in our method. However, it will be avoidable when the expression profiles of miRNA are available: the unexpressed miRNAs showing small $P$-value can be filtered out from the list.

One may wonder that co-targeting among miRNAs results in more significant regulation of target genes of non-transfected miRNAs than that of the transfected miRNA. We calculated $P$-values of significant overlap of target genes by $P(n_{mm_0}, n_m, n_{m_0}/N)$. As a result, even after correction considering multiple comparison, 156, 157 and 156 miRNAs among in total 162 miRNAs have significant overlap ($P < 0.05$) with transfected miRNA of mir-107, mir-185 and let-7a respectively. This means, almost all of non-transfected miRNAs have significant large number of common target genes with those transfected miRNA. However, the correlation coefficient between $P_m$ and $P^O_{m,m_0}$, do not show significant correlations 11 out of 12 (i.e., two biological replicates $\times$ two time points (day 1 or day 3) $\times$ three transfection) cases (see Table 3). This means, significant regulation of target genes of non-transfected miRNA cannot be explained by the accidental target gene overlap with those of transfected miRNA, $m_0$. Only one exception among total 12 cases is for let-7 day 3 replicate 2. Since the effects of transfected miRNA become weaker at the day 3, the secondary effect may become apparent and the some fluctuations in experimental conditions might

**Table 3.** Significance of correlation between $P_m$ and $P^O_{m,m_0}$. Bold numbers are significant ($P < 0.05$).

| Transfection | mir-107 | | | | mir-185 | | | | let-7a | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | day 1 | | day 3 | | day 1 | | day 3 | | day 1 | | day 3 | |
| Replicates | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| correlation | -0.049 | -0.086 | -0.11 | -0.12 | 0.050 | 0.098 | 0.13 | -0.11 | -0.012 | -0.14 | -0.092 | -0.24 |
| $P$-value | 0.54 | 0.28 | 0.15 | 0.13 | 0.52 | 0.22 | 0.10 | 0.16 | 0.88 | 0.07 | 0.25 | **0.0019** |

cause the significance in this case. Actually, $P_{m_0}$ for three days after transfection sometimes does not have small enough $P$ -value (mir-107 replicate 2 and mir-185 replicate 2, see Table 1).

It is also interesting that miRNA target genes are generally more expressed than the other genes in the present datasets (see Fig. 1 in the previous work [11], there are peaks around $\log x_g \simeq 7$, which is far from origin). This tendency cannot be seen in genes not targeted by any miRNA. This fact may also be important to understand how each miRNA regulate genes in cancer formation/suppression.

The reasons why MiRaGE works better than previous methods are worth mentioning. For example, T-REX[14] can detect transfected miRNA as those with the smallest $P$-value, but cannot avoid having other miRNAs with significantly small $P$-value. As T-REX, most of such methods compare target genes of a miRNA with all others, while MiRaGE compares target genes with the genes targeted by all the other miRNAs. Since genes targeted by miRNAs are significantly different from genes not targeted by any miRNAs as mentioned above, it is important to employ genes targeted by other miRNAs as negative set. Other critical factor to improve the specificity is to employ simple seed match as miRNA target gene table. Since this table also decides the negative set, highly curated thus those with smaller genes mimic the size of negative set. The success of MiRaGE demonstrates that choice of the negative set is critical to estimate the important miRNAs during the biological processes.

## 5 Conclusion

In this paper, we have shown that gene expression profile combined with miRNA target genes predicted computationally can often correctly infer transfected miRNA. This suggests that we may be able to infer miRNA regulation of genes solely from gene expressions without considering any other information than computationally predicted target genes.

**Acknowledgement.** This work was supported by KAKENHI (23300357).

## References

1. Brennecke, J., Stark, A., Russell, R.B., Cohen, S.M.: Principles of MicroRNA-targetRecognition. PLoS Biol. 3, 85 (2005)
2. Chi, S.W., Zang, J.B., Mele, A., Darnell, R.B.: Argonaute HITS-CLIP Decodes MicroRNA-mRNA Interaction Maps. Nature 460(7254), 479–486 (2009)

3. Takahashi, Y., Forrest, A.A.R., Maeno, E., Hashimoto, T., Daub, C.O., Yasuda, J.: MiR-107 and MiR-185 Can Induce Cell cycle Arrest in Human Non Small Cell Lung Cancer Cell Lines? PLoS One 4, e6677 (2009)
4. Ikeo, K., Ishii, J., Tamura, T., Gojobori, T., Tateno, Y.: CIBEX: center for information biology gene expression database. C R Biol. 326, 1079–1082 (2003)
5. Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., Burge, C.B.: Determinants of Targeting by Endogenous and Exogenous MicroRNAs and siRNAs. RNA 13, 1894–1910 (2007)
6. Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P.: Most Mammalian mRNAs are Conserved Targets of MicroRNAs. Genome Res. 19, 92–105 (2009)
7. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E.: The Role of Site Accessibility in MicroRNA Target Recognition. Nat. Genet. 39, 1278–1284 (2007)
8. Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., Rajewsky, N.: Combinatorial MicroRNA Target Predictions. Nat. Genet. 37, 495–500 (2005)
9. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., Marks, D.S.: Human MicroRNA Targets. PLoS Biol. 2, e363 (2004)
10. Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M., Hatzigeorgiou, A.G.: Lost in Translation: An Assessment and Perspective for Computational MicroRNA Target Identification. Bioinformatics 25, 3049–3055 (2009)
11. Taguchi, Y.-h., Yasuda, J.: Inference of gene expression regulation via microRNA transfection. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2010. LNCS, vol. 6215, pp. 672–679. Springer, Heidelberg (2010)
12. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B 57, 289–300 (1995)
13. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0, `http://www.R-project.org`
14. Volinia, S., Visone, R., Galasso, M., Rossi, E., Croce, C.: Identification of MicroRNA Activity by Targets' Reverse EXpression. Bioinformatics 26, 91–97 (2010)