Vol.2010-BIO-21 No.6
2010/6/18

Inference of microRNA transfection via target gene expression

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

MicroRNA (miRNA) is a new candidate to regulate gene expression in epigenetic manners. It turns out that each can really regulate expression of hundreds of genes in the manner dependent upon several situations, e.g., developmental stages, deceases, and external conditions. In spite of that, how and when miRNAs can regulate target genes' expression and what target genes are is unclear. In this paper, we have investigated regulation of gene expression via miRNA transfection, by using both gene expression profile and computationally predicted target genes. The miRNA, let-7a, transfected to tumors was listed at the top most plausible candidate by this analysis.

1. Introduction

MicroRNAs (miRNAs) are the single strand RNAs with the length from 20 to 25 nucleotide in cells. It is a member of non-coding RNAs which are supposed to control gene expression. They are believed to suppress expression of target genes, by aligned to complimentary seed sequence with the length of eight nucleotide, which is typically located at 3' untranslated region (UTR) of target genes. Although it is possible to validate target gene experimentally[1)], it is a very time/cost consuming process. Thus, usually, coputer oriented predictions are employed to list target genes of each miRNA.

Recently[2)], we have proposed the method which validates the accurency of miRNA target genes list only by the comparison between gene expression profile of target genes and that of others. We have validated our algorithm by analysis of gene expression profiles of miRNA-transfected cells and computer-predicted potential miRNA target lists; our method can correctly list the transfected miRNA as the most significant one. Since it is concept-oriented curated list, we can understand which aspect is important to confirm predicted miRNA target genes.

†1 Department of Physics, Chuo University, Tokyo, Japan
†2 COE Fellow, Graduate School of Medicine, Tohoku University, Sendai, Japan
∗1 Presently with Department of Cell Biology, The JFCR-Cancer Institute, Tokyo, Japan

2. Materials and Methods

2.1 Gene expression data for transfection experiment

We have downloaded transfection experiment[3)] data set, CBX79, which is deposited 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 trasfection 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 list of curated miRNA target genes (Tables S1-S7[7)]). This includes 50-70 miRNA families dependent upon concepts used for curate (**Table 1**). Then we have picked up genes which are the targeted by at least one miRNA. Then, from five hundreds to two thousands genes remain dependent upon concept used for curate (Table 1). Hereafter, we denote a set of these genes as G. Next, for each miRNA, m, we have listed target genes of it. We denote this set of genes as G_m, where m denotes one of miRNA families. Also we define a set of genes, $G'_m \equiv G \setminus G_m$. 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 two way t-test beween $\{\Delta x_g^{m_0} \mid g \in G_m\}$ and $\{\Delta x_g^{m_0} \mid g \in G'_m\}$. P-value, P_m, is computed for each miRNA, m. After applying FDR correction (BH method[5)]) to corresponding P-values, we have selected ms 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[6)].

2.3 Coincidence between biological replicates

We have also checked if two biological replicates satisfy reproducibility in three ways. First two stand for over all coincidence between P_ms, correlation coefficients between P_ms for two replicates. The first correlation coefficient is Pearson correlation coefficients between log transformed P_ms and the second is Spearman correlation coefficients between them. P-values for these are computed and 95 percentail significant interval for the form is also computed. The third is coincidence between significant miRNAs, ms. 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.

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

3. Results

First of all, during analysis, we noticed that there are no target gene information for mir-185. Furthermore, that for mir-107 is too few to give us significant results. For example, Table S1 includes no mir-185 target genes and only 8 mir-107 target genes. Thus, we have employed only let-7a tnasfected case. Independent of conditions, i.e., date and concepts, our method almost always gets non-empty set of significant miRNAs, ms (see **Table 2**). Thus, in principle, our method can detect miRNA regulation of gene expression. From **Table 3** to **Table 16** show which miRNA significantly regulates target genes. Most remarkably, P_m has the strong tendency to become smallest when $m = m_0$, let-7a, especially for one day after transfection. Thus, our method has not only ability to detect miRNA regulation of genes, but also that to infer transfected miRNA correctly, as the one regulate target genes mostly.

From **Table 17** to **Table 19** shows the results of several statistical tests for the coincidence between biological replicates. Almost all tests give us significant P-values < 0.05. Thus, biological replicates are good enough for inference of miRNA transfection.

Table 1 Number of genes in G and miRNAs included into each of concept-oriented target gene lists

Concepts for curate	Tables	Number of miRNAs considered	Number of genes in G
Pathway	S1	58	887
Coexpression module	S2	57	1997
KEGG	S3	51	1090
Protein complex	S4	70	620
GO BP	S5	64	1548
GO CC	S6	52	596
GO MF	S7	61	1142

Table 2 Numbers of significant miRNAs. The ranks of transfected micriRNA let-7a are shown in square brackets.

Concepts for curate	Transfected miRNA							
	let-7a							
	day 1				day 3			
	replicate 1		replicate 2		replicate 1		replicate 2	
Pathway	5	[1st]	16	[1st]	11	[4th]	10	[-]
Coexpression module	7	[-]	7	[-]	7	[-]	6	[-]
KEGG	9	[2nd]	11	[3rd]	5	[-]	4	[-]
Protein complex	8	[2nd]	15	[3rd]	16	[2nd]	16	[3rd]
GO BP	3	[1st]	8	[1st]	6	[1st]	10	[1st]
GO CC	2	[-]	6	[-]	8	[-]	10	[-]
GO MF	5	[2nd]	8	[1st]	8	[2nd]	5	[-]
Previous study[2)]	2	[1st]	2	[1st]	1	[1st]	33	[8th]

Table 3 Three top most significant miRNAs, one day after let-7a trasnfection. Bold characters are those transfected. Curated by pathway information.

Pathway, day 1			
replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
let-7/98	1.09×10^{-6}	**let-7/98**	1.24×10^{-10}
miR-34b	1.09×10^{-4}	miR-141/200a	9.19×10^{-9}
miR-141/200a	9.15×10^{-4}	miR-216	2.16×10^{-7}

Table 4 Three top most significant miRNAs, three days after let-7a trasnfection. Curated by pathway information.

Pathway, day 3			
replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-141/200a	5.37×10^{-6}	miR-216	6.22×10^{-6}
miR-15/16/195/424/497	2.32×10^{-5}	miR-204/211	7.02×10^{-6}
miR-196	8.10×10^{-5}	miR-34b	1.12×10^{-5}

Table 5 Three top most significant miRNAs, one day after let-7a trasnfection. Curated by coexpression module information.

Coexpression module day 1			
replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-141/200a	5.70×10^{-5}	miR-194	7.23×10^{-5}
miR-130/301	6.05×10^{-5}	miR-30-3p	2.14×10^{-4}
miR-93.hd/291-3p/294/295/302/372/373/520	2.10×10^{-4}	miR-93.hd/291-3p/294/295/302/372/373/520	4.11×10^{-4}

Table 6 Three top most significant miRNAs, three days after let-7a trasnfection. Curated by coexpression module information.

Coexpression module, day 3			
replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-122	8.85×10^{-4}	miR-124.2/506	2.76×10^{-5}
miR-93.hd/291-3p/294/295/302/372/373/520	2.31×10^{-3}	miR-33	1.10×10^{-3}
miR-200b/429	2.71×10^{-3}	miR-34/449	1.44×10^{-3}

Table 19 P-values are the probabilities of selecting miRNAs commonly between two replicates by chance. Bold numbers indicate significant P-values (< 0.05).

Concept for curate	time	# of significant miRNAs			P-value	
		common	replicate 1	replicate 2		
Pathway	day 1	5	5	16	**0**	$\mathbf{1.53 \times 10^{-3}}$
	day 3	4	11	10	$\mathbf{2.81 \times 10^{-2}}$	$\mathbf{2.64 \times 10^{-2}}$
Coexpression module	day 1	4	7	7	$\mathbf{4.72 \times 10^{-4}}$	$\mathbf{2.94 \times 10^{-2}}$
	day 3	2	7	6	$\mathbf{2.94 \times 10^{-2}}$	$\mathbf{2.78 \times 10^{-2}}$
KEGG	day 1	7	9	11	$\mathbf{3.41 \times 10^{-5}}$	$\mathbf{9.31 \times 10^{-5}}$
	day 3	3	5	4	$\mathbf{1.77 \times 10^{-4}}$	$\mathbf{9.23 \times 10^{-5}}$
Protein Complex	day 1	7	8	15	$\mathbf{4.44 \times 10^{-6}}$	$\mathbf{8.88 \times 10^{-5}}$
	day 3	8	16	16	$\mathbf{3.94 \times 10^{-3}}$	$\mathbf{3.94 \times 10^{-3}}$
GO BP	day 1	3	3	8	**0**	$\mathbf{2.90 \times 10^{-4}}$
	day 3	4	6	10	$\mathbf{4.86 \times 10^{-4}}$	$\mathbf{1.21 \times 10^{-3}}$
GO CC	day 1	2	2	6	**0**	$\mathbf{1.04 \times 10^{-3}}$
	day 3	7	8	10	$\mathbf{1.87 \times 10^{-6}}$	$\mathbf{1.05 \times 10^{-5}}$
GO MF	day 1	3	5	8	$\mathbf{1.32 \times 10^{-3}}$	$\mathbf{2.41 \times 10^{-3}}$
	day 3	3	8	5	$\mathbf{2.41 \times 10^{-3}}$	$\mathbf{1.32 \times 10^{-3}}$
Previous Study[2)]	day 1	2	2	2	**0**	**0**
	day 3	1	1	33	**0**	$\mathbf{1.7 \times 10^{-2}}$

Table 7 Three top most significant miRNAs, one day after let-7a trasnfection. Bold characters are those transfected. Curated by KEGG information.

KEGG day 1

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-30-3p	7.59×10^{-6}	miR-17-5p/20/93.mr/106/519.d	2.28×10^{-7}
let-7/98	2.28×10^{-5}	miR-21	3.08×10^{-7}
miR-17-5p/20/93.mr/106/519.d	5.23×10^{-5}	**let-7/98**	7.20×10^{-7}

Table 8 Three top most significant miRNAs, three days after let-7a trasnfection. Curated by KEGG information.

KEGG, day 3

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-21	7.27×10^{-7}	miR-21	5.34×10^{-6}
miR-15/16/195/424/49	5.99×10^{-5}	miR-29	9.14×10^{-5}
miR-17-5p/20/93.mr/106/519.d	1.88×10^{-3}	miR-142-3p	9.87×10^{-5}

Table 9 Three top most significant miRNAs, one day after let-7a trasnfection. Bold characters are those transfected. Curated by protein complex information.

Protein complex day 1

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-203.1	2.97×10^{-6}	miR-34/449	2.20×10^{-8}
let-7/98	1.29×10^{-5}	miR-130/301	3.39×10^{-6}
miR-196	1.50×10^{-5}	**let-7/98**	6.67×10^{-6}

Table 10 Three top most significant miRNAs, three days after let-7a trasnfection. Bold characters are those transfected. Curated by protein complex information.

Protein complex, day 3

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
let-7/98	2.44×10^{-5}	miR-221/222	1.33×10^{-8}
miR-15/16/195/424/49	5.99×10^{-5}	miR-125/351	3.66×10^{-6}
miR-122	3.25×10^{-5}	**let-7/98**	1.14×10^{-5}

Table 11 Three top most significant miRNAs, one day after let-7a trasnfection. Bold characters are those transfected. Curated by GO biological process information.

GO BP, day 1

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
let-7/98	6.70×10^{-7}	**let-7/98**	1.27×10^{-13}
miR-153	3.33×10^{-5}	miR-205	3.16×10^{-5}
miR-21	3.63×10^{-4}	miR-21	9.95×10^{-5}

Table 12 Three top most significant miRNAs, three days after let-7a trasnfection. Bold characters are those transfected. Curated by GO biological process information.

GO BP, day 3

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
let-7/98	4.47×10^{-7}	**let-7/98**	1.11×10^{-4}
miR-10	2.66×10^{-6}	miR-7	1.26×10^{-4}
miR-7	2.31×10^{-5}	miR-34b	3.15×10^{-4}

Table 13 Three top most significant miRNAs, one day after let-7a trasnfection. Curated by GO cellar components information.

GO CC, day 1

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-33	2.71×10^{-7}	miR-200b/429	3.74×10^{-5}
miR-22	1.11×10^{-4}	miR-22	9.41×10^{-5}
—	—	miR-203.1	1.26×10^{-4}

Table 14 Three top most significant miRNAs, three days after let-7a trasnfection. Curated by GO cellar components information.

GO CC, day 3

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-200b/429	2.32×10^{-6}	miR-25/32/92/363/367	2.91×10^{-7}
miR-199	2.21×10^{-5}	miR-29	6.97×10^{-7}
miR-33	2.41×10^{-5}	miR-199	2.52×10^{-4}

Table 15 Three top most significant miRNAs, one day after let-7a trasnfection. Bold characters are those transfected. Curated by GO molecular function information.

GO MF, day 1

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
let-7/98	1.57×10^{-5}	**let-7/98**	3.01×10^{-8}
miR-22	1.11×10^{-4}	miR-196	7.38×10^{-6}
miR-31	3.76×10^{-4}	miR-140	5.04×10^{-4}

Table 16 Three top most significant miRNAs, three days after let-7a trasnfection. Bold characters are those transfected. Curated by GO molecular function information.

GO MF, day 3

replicate 1		replicate 2	
miRNA	P_m	miRNA	P_m
miR-196	1.41×10^{-6}	miR-196	2.82×10^{-6}
let-7/98	1.67×10^{-6}	miR-140	1.00×10^{-4}
miR-138	1.70×10^{-4}	miR-142-5p	1.96×10^{-3}

Table 17 Pearson correlation coefficients between log transformed P_ms of two biological replicates. Bold numbers indicate signifcant P-values (< 0.05).

Concept for curate	time	Pearson	95 % confidence interval lower	upper	P-value
Pathway	day 1	0.73	0.59	0.83	$\mathbf{8.12 \times 10^{-11}}$
	day 3	0.36	0.11	0.56	$\mathbf{6.00 \times 10^{-3}}$
Coexpression module	day 1	0.51	0.29	0.68	$\mathbf{5.94 \times 10^{-5}}$
	day 3	0.30	0.05	0.53	$\mathbf{2.25 \times 10^{-2}}$
KEGG	day 1	0.65	0.45	0.79	$\mathbf{3.37 \times 10^{-7}}$
	day 3	0.51	0.27	0.69	$\mathbf{1.39 \times 10^{-4}}$
Protein complex	day 1	0.63	0.47	0.76	$\mathbf{3.83 \times 10^{-9}}$
	day 3	0.61	0.44	0.74	$\mathbf{1.90 \times 10^{-8}}$
GO BP	day 1	0.69	0.53	0.80	$\mathbf{3.75 \times 10^{-10}}$
	day 3	0.48	0.27	0.65	$\mathbf{5.75 \times 10^{-5}}$
GO CC	day 1	0.46	0.21	0.65	$\mathbf{6.70 \times 10^{-4}}$
	day 3	0.57	0.35	0.73	$\mathbf{1.46 \times 10^{-5}}$
GO MF	day 1	0.71	0.56	0.82	$\mathbf{1.06 \times 10^{-10}}$
	day 3	0.54	0.34	0.70	$\mathbf{6.10 \times 10^{-6}}$
Previous study[2)]	day 1	0.86	0.81	0.89	$\mathbf{< 2.2 \times 10^{-16}}$
	day 3	0.28	0.13	0.42	$\mathbf{2.9 \times 10^{-4}}$

Table 18 Spearman correlation coefficients between P_ms of two biological replicates. Bold numbers indicate signifcant P-values (< 0.05).

Concept for curate	time	Spearman	P-value
Pathway	day 1	0.46	$\mathbf{4.17 \times 10^{-4}}$
	day 3	0.31	$\mathbf{2.10 \times 10^{-2}}$
Coexpression module	day 1	0.42	$\mathbf{1.37 \times 10^{-3}}$
	day 3	0.33	$\mathbf{1.27 \times 10^{-2}}$
KEGG	day 1	0.52	$\mathbf{1.54 \times 10^{-4}}$
	day 3	0.28	5.03×10^{-2}
Protein Complex	day 1	0.40	$\mathbf{7.15 \times 10^{-4}}$
	day 3	0.34	$\mathbf{5.95 \times 10^{-3}}$
GO BP	day 1	0.23	7.22×10^{-2}
	day 3	0.34	$\mathbf{5.95 \times 10^{-3}}$
GO CC	day 1	0.47	$\mathbf{5.05 \times 10^{-4}}$
	day 3	0.40	$\mathbf{3.78 \times 10^{-3}}$
GO MP	day 1	0.42	$\mathbf{7.41 \times 10^{-4}}$
	day 3	0.21	1.00×10^{-1}
Previous study[2)]	day 1	0.28	$\mathbf{3.00 \times 10^{-4}}$
	day 3	0.13	1.00×10^{-1}

4. Discussion

Transfected let-7a is most frequently listed as the top-ranked (from 1st to 3rd) significant miRNA. Actually, it is ranked up to 3rd in 16 cases among total 27 cases. The second frequent miRNA, mir-21, is ranked only four times. Thus, curated lists surely detect let-7a as the most significant miRNA. On the other hand, in contrast to the previous study[2)] where we have always found let-7a in significant miRNAs, there are some cases where we failed to detect let-7a.

For example, lists curated by coexpression module or GO cellar components have never detected let-7a. Although one may thinks that it exhibits that our method does not work well, it may represent that it is important how we curate lists. These two concepts, coexpression module and GO cellar components, represent only common appearance of genes. The former represents that genes are expressed at the same time, while the later represents that genes are expressed at the same place. It is clear that it does not represent any direct biological relationships among genes. There are some possibilities that these two concepts fail to curate gene list in the biologically meaningful manner. On the other hand, if we can see Table 17 to Table 19 , we recognize that there are no evidences that these two result in especially bad biological reproducibility. Since we can attribute P-values to not only miRNAs but also each term/cluster included in each concept, we may be able to clarify why lists curated by these two concepts cannot detect let-7a. It will be the future works.

Remarkable feature of curated list compared with previous study is to detect let-7a at day 3 replicate 2, for which we could not detect let-7a as being highly enough ranked (only 8th) previously[2)]. Also, previous study gives us 33 significant miRNAs for day 3 replicate 2. It is too odd since other three cases give us at most two significant miRNAs. For curated list, day 3 replicate 2 has never expressed such a odd behaviour. Thus, we think that curated list is more trustable and robust than simple seed match.

Although there are many miRNAs other than let-7a, which are detected in the present analysis, it is rare that they are selected in more than two lists based upon distinct concept for curate. Thus, it is not plausible that simple overlaps of target genes has accidentally causes this side effect. Possibly, secondly regulation

of genes occurs and it is detected in distinct manner dependent upon the concept curate. This point also must be checked in future works.

Another problem of curated list is smallness of genes included into lists. This is mainly because authors[7] are interested in genes with known cellar functions. Although they include less than two thousands genes, seed match list includes more than thirteen thousands genes. Also, curated list lacks information of mir-185 and has very little information of mir-107 as mentioned in the above. This prevents us from applying our method to these two miRNA transfection experiments. We may need longer time to get curated list of target genes for any other non-popular miRNAs, too.

5. Conclusion

In this study, we have employed curated list of miRNA target genes instead of that based upon simple seed match which we have used in the previous study[2]. We have found that list curated based upon biological function, e.g., KEGG, GO biological process, GO molecular function, pathway and protein complex, can give us more stable and robust results, while those based upon simple simultaneous appearance, i.e., coexpression modules and GO cellar components, cannot detect transfected miRNA, let-7a, as top most significant miRNA. This may suggest that it is important by which concept list are curated. Not all curated list may work better than simple seed match. On the other hand, curated list give us more significant miRNAs which cannot be detected in the previous research[2]. Since they are not transfected directly, it is interesting in which aspects they are regarded as being significant. Since curated list is accompanied by terms/modules which are used for curate, we can check this point in the future study. This may enable us to have better curated list of miRNA target genes.

References

1) Chi, SW., Zang, JB., Mele, A., and Darnell, RB.: Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps, *Nature*, Vol.460, No.7254, pp.479–86 (2009).
2) Taguchi, Y-h., and Yasuda, J.: Inference of gene expression regulation via microRNA transfection, *Advanced Intelligent Computing Technology and Applications-ICIC2010*, (Huang, D.-S., et al (ed.)), *Lecture Notes in Computer Science* , Berlin / Heidelberg, Springer, in press (2010) .
3) Takahashi, Y., Forrest., A. A. R., Maeno1., E, Hashimoto., T, Daub., and 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*, Vol.4, e6677 (2009).
4) Ikeo, K., Ishii, J., Tamura, T., Gojobori, T., and Tateno, Y.: CIBEX: center for information biology gene expression database, *C R Biol.*, Vol.326, pp.1079–82 (2003).
5) Benjamini, Y., and Hochberg, Y,: Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J R Stat Soc B*, Vo.57, pp. 289–300, (1995).
6) R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (2009)
7) Tsang, J S, Ebert, M S, van Oudenaarden, A.: Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures, *Mol Cell.*, Vol.38, No.1, pp.140-53 (2010).