Journal of Bioinformatics and Computational Biology
Vol. 8, No. 1 (2010) 99–115

DOI: 10.1142/S0219720010004574

Imperial College Press
www.icpress.co.uk

IDENTIFYING CO-REGULATING MICRORNA GROUPS

JIYUAN AN

The National Centre for Adult Stem Cell Research
The Eskitis Institute for Cell and Molecular Therapies
Griffith University, Nathan, QLD, 4111, Australia
j.an@griffith.edu.au

KWOK PUI CHOI

Department of Statistics and Applied Probability
National University of Singapore, Singapore
stackp@nus.edu.sg

CHRISTINE A. WELLS

The National Centre for Adult Stem Cell Research
The Eskitis Institute for Cell and Molecular Therapies
Griffith University, Nathan, QLD, 4111, Australia
c.wells@griffith.edu.au

YI-PING PHOEBE CHEN

Faculty of Science and Technology
Deakin University, Australia
phoebe.chen@deakin.edu.au

Received 21 November 2009
Revised 15 December 2009
Accepted 15 December 2009

Background: Current miRNA target prediction tools have the common problem that their false positive rate is high. This renders identification of co-regulating groups of miRNAs and target genes unreliable. In this study, we describe a procedure to identify highly probable co-regulating miRNAs and the corresponding co-regulated gene groups. Our procedure involves a sequence of statistical tests: (1) identify genes that are highly probable miRNA targets; (2) determine for each such gene, the minimum number of miRNAs that co-regulate it with high probability; (3) find, for each such gene, the combination of the determined minimum size of miRNAs that co-regulate it with the lowest p-value; and (4) discover for each such combination of miRNAs, the group of genes that are co-regulated by these miRNAs with the lowest p-value computed based on GO term annotations of the genes. **Results:** Our method identifies 4, 3 and 2-term miRNA groups that co-regulate gene groups of size at least 3 in human. Our result suggests some interesting hypothesis on the functional role of several miRNAs through a "guilt by association" reasoning. For example, miR-130, miR-19 and miR-101 are known neurodegenerative diseases associated miRNAs. Our 3-term miRNA table shows that miR-130/19/101 form a co-regulating group of rank 22 (p-value $= 1.16 \times 10^{-2}$).

Since miR-144 is co-regulating with miR-130, miR-19 and miR-101 of rank 4 (p-value = 1.16×10^{-2}) in our 4-term miRNA table, this suggests hsa-miR-144 may be neurodegenerative diseases related miRNA. **Conclusions:** This work identifies highly probable co-regulating miRNAs, which are refined from the prediction by computational tools using (1) signal-to-noise ratio to get high accurate regulating miRNAs for every gene, and (2) Gene Ontology to obtain functional related co-regulating miRNA groups. Our result has partly been supported by biological experiments. Based on prediction by TargetScanS, we found highly probable target gene groups in the Supplementary Information. This result might help biologists to find small set of miRNAs for genes of interest rather than huge amount of miRNA set. **Supplementary Information:** https://www.deakin.edu.au/~phoebe/JBCBAnChen/JBCB.htm

Keywords: microRNA; co-regulating; target gene.

1. Background

MicroRNAs (miRNAs) are a type of endogenous regulatory RNAs approximately 22 nucleotides in length. miRNAs perform post-transcriptional inhibition on target genes. More than 4000 miRNAs have been identified in eukaryotes and one-third of human genes are likely to be regulated by miRNAs.[1] Due to difficulty in identifying target genes of miRNAs experimentally, computational methods are often used to predict target genes. Prominent among these methods are TargetScanS,[2,3] MiRanda,[1] DIANA-microT,[4] PicTar,[5] TarBase[6] and MicroTar.[7] The following three main criteria are employed in these methods: (1) a miRNA should have complementary binding sites in the 3′UTR of its target genes; (2) the interaction of a miRNA and its target gene should result in lower free energy than a threshold value; and (3) a target gene should have conserved regions in the 3′UTR sequences of near species. Note that MicroTar[7] does not consider evolutionary conservation: it uses only criteria (1) and (2) to predict miRNA's target genes. Furthermore, these prediction tools may produce different target sets[8,9] for the same given set of miRNAs.

Several miRNAs may work together to suppress a group of functionally related genes.[9,10] A group of miRNAs are said to be co-regulating if they regulate some genes in common. The interaction between miRNAs and mRNA involves a complex process.[11] The regulation of mRNA by miRNAs could happen in two ways, as illustrated in Fig. 1: (1) several miRNAs regulate mRNAs coordinately to perform one function; or (2) different miRNAs regulate mRNAs to perform more than one function simultaneously.

Prediction tools such as TargetScanS and MiRanda can be used to identify potential target genes for all miRNAs. Yoon and Micheli in Ref. 12 have proposed a biclique-based method to find co-regulating groups of miRNAs and mRNAs. Joung *et al.* in Ref. 13 have described a population-based co-evolutionary learning method to find miRNA–mRNA models. There is room for improvement in the precision of these prediction tools. For example, TargetScanS has a signal-to-noise ratio of 2.4:1 on human sequences with mouse, rat, chicken and dog orthologs, which is equal to a precision of $(2.4 - 1)/2.4 = 58.3\%$. If we predict co-regulating miRNA groups

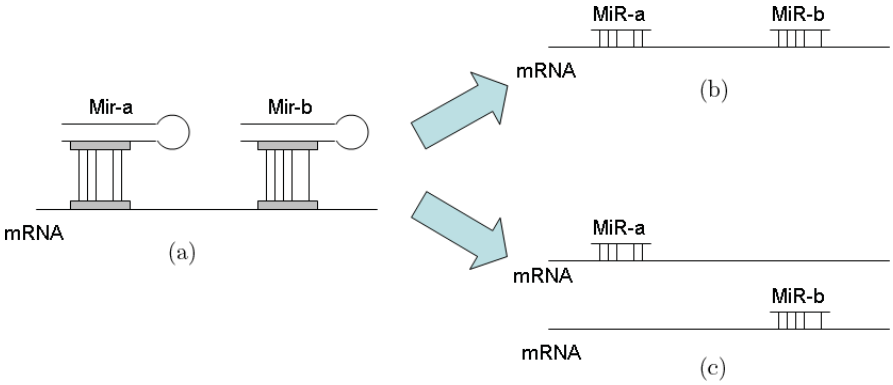

Fig. 1. Multiple miRNAs regulate one mRNA with complementary binding. (a) The two miRNAs have complementary binding sites in an mRNA. The shadow rectangles represent binding parts. (b) Two active miRNAs regulate an mRNA coordinatively. (c) Two miRNAs regulate the same mRNA simultaneously.

based on combinations of miRNAs individually predicted at such a precision level, the precision of the predicted co-regulating groups is going to be low. For example, the precision of predicted miRNA combinations of size 3 becomes $58.3\%^3 = 19.8\%$.

In this paper, we start from a set of human miRNA–gene pairs predicted by TargetScanS. Then we develop a sequence of statistical analyses to (1) select genes that have at least 99.9% probability of being targets of some miRNAs; (2) determine for each selected gene the minimum number of miRNAs that have at least 99.9% probability of co-regulating it; (3) identify for each selected gene the combination of miRNAs of the determined minimum size that co-regulate the gene at the lowest p-value; and (4) find for each such combination of co-regulating miRNAs the group of genes that are co-regulated by these miRNAs at a p-value less than 0.01 based on coherence of the Gene Ontology (GO) terms annotated to these genes. As a result of our analysis, we produce human miRNA co-regulating tables in terms of the number of miRNAs. Four tables (4^+-term, 3-term, 2-term and 1-term) can be found in the Supplementary Information available at https://www.deakin.edu.au/~phoebe/JBCBAnChen/JBCB.htm, which is a Microsoft Excel file. The tables are shown in different Excel Sheets named "miRNA & targets(4+)", "miRNA & targets(3)", "miRNA & targets(2)" and "miRNA & targets(1)". Note that highly homologous genes have been removed to avoid forming co-regulated gene groups among themselves because they have similar 3′UTRs.

Our proposed procedure has a number of advantages over existing methods. For example, while Joung *et al.*[13] take into account expression of miRNAs and mRNAs, they do not consider biological function. More importantly, most miRNAs inhibit protein production but do not change the mRNA expression levels. Moreover, gene expression levels are tissue-specific. Therefore, to find real targets of miRNAs, Gene Ontology information is more suitable than mRNA expression data, because gene association construction in Gene Ontology is based on protein expression level. In this paper, we take biological function into account when we

compute the p-value of our co-regulated gene groups based on the coherence of the GO terms annotated to the genes. As another example, the biclique-based MRM method of Ref. 12 finds miRNA groups heuristically and uses GO to validate the finding. Due to its heuristic nature, it may miss miRNA groups or gene groups that have high probability of being co-regulated. In contrast, our procedure retains all highly probable co-regulation groups.

2. Method

A simple-minded approach to genome-wide identification of miRNA–gene co-regulation group pairs is as follows. First, apply a computational miRNA target prediction tool like TargetScanS on a genome-wide basis to obtain an initial miRNA–gene pairs map $R = \{(Y, X)|$ the miRNA Y is predicted by TargetScanS to target gene $X\}$, which can be derived straightforwardly from the 3227×162 matrix mentioned in the following subsection"Data". Then compute the anti-chain $S = \max\{(M, G)|$ for all $Y_1 \in M$, for all $Y_2 \in M$, for all $X_1 \in G$, for all $X_2 \in G$, $(Y_1, X_1) \in R$, $(Y_1, X_2) \in R$, $(Y_2, X_1) \in R$, $(Y_2, X_2) \in R\}$. (Here we have used standard set comprehension notations: M ranges over subsets of miRNAs, and G ranges over subsets of target genes predicted by TargetScanS. Recall an anti-chain of an ordered set S is a subset A containing the maximal elements of S.)

However, given the poor signal-to-noise performance of TargetScanS, the false-positive level of miRNA–gene co-regulation group pairs derived in this simple-minded way is unacceptably high. In fact, given such a co-regulation group pair (M, G), its probability of containing at least one pair $(Y, X) \in M \times G$ that is not co-regulated is $1 - p^{|M \times G|}$, where $p = 58.3\%$ is the reported precision of TargetScanS.[3] For example, if $|M| = |G| = 4$, the chance of (M, G) containing at least one miRNA–gene pair that is not co-regulated is $1 - 58.3\%^{16} = 99.98\%$.

Therefore, we need a more stringent approach to derive the miRNA-gene co-regulation group pairs from the basic miRNA-gene pairs map R produced by TargetScanS. We introduce below a sequence of statistical tests for this purpose.

2.1. *Data*

The miRNA sequences are obtained from Rfam.[a] MiRNAs conserved in human, mouse, rat, chicken and dog are clustered into 162 families based on miRNA seed region (nucleotides 2–8). Some families consist of many miRNAs. For example, miR-15/16/195/424/497 have the same seed "AGCAGCA". Human gene sequences are obtained from RefSeq, and orthologous sequences in human, mouse, rat, chicken and dog are taken from the UCSC genome browser multiZ mulitple genome alignments.[14]

The predicted miRNA targets from TargetScanS[2,3] have 15,825 pairs between miRNAs and target genes for conserved miRNAs and genes in human, mouse, rat,

[a]http://www.sanger.ac.uk/software/Rfam

chicken and dog. This data can be downloaded from TargetScanS.[b] We rearranged the data into a 3227×162 matrix, whose rows correspond to genes and columns correspond to miRNA families. The elements of the matrix represent the number of binding sites. If there is no regulation relationship between a miRNA family and a gene, the corresponding element in the matrix is set to be 0; otherwise, the element is set to be the number of binding sites.

2.2. *Architecture of our method*

Figure 2 shows the procedure of identifying co-regulating miRNA groups. Ovals represent datasets and rectangles represent the procedures. The Perl source code of TargetScanS was kindly provided to us by Lewis laboratory.

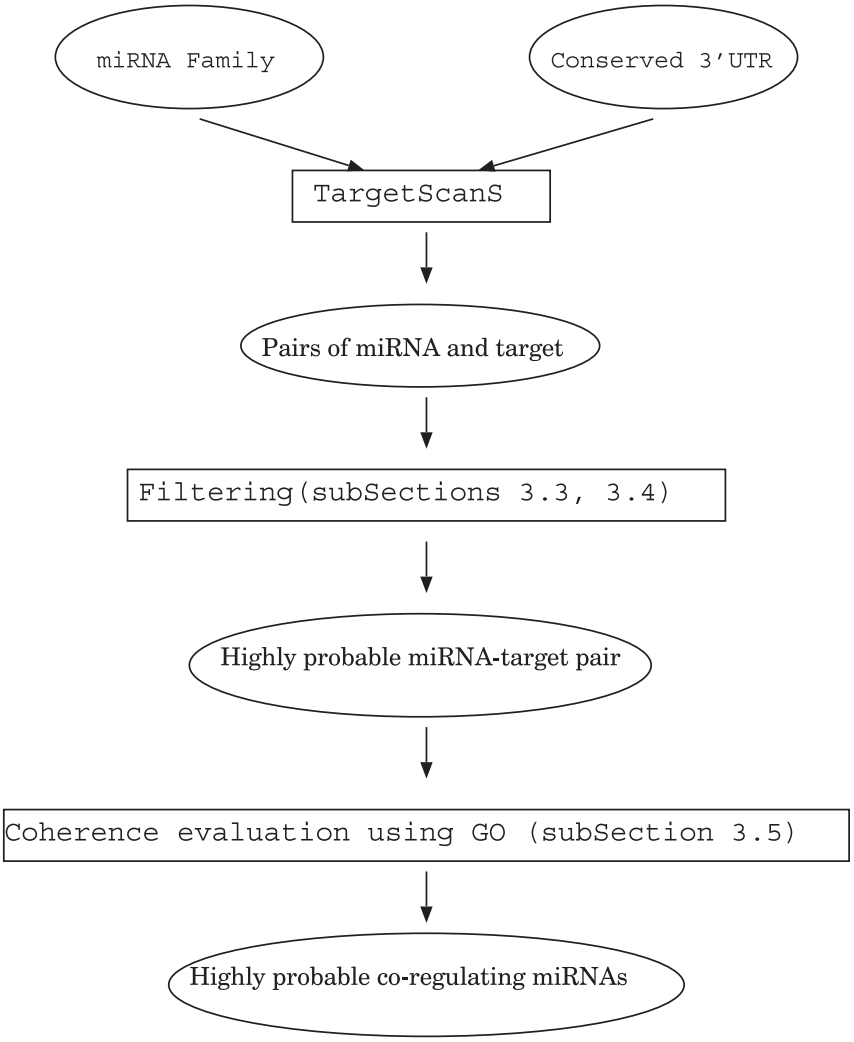

Fig. 2. A flow chart for identifying highly probable co-regulation miRNA–gene group pair.

[b]http://genes.mit.edu/tscan/targetscanS2005.html

By executing TargetScanS, target genes are predicted for each miRNA. Since the target genes include a large amount of false positives, a filtering procedure, as described in the following Step 1.1 and 1.2, is needed. After the filtering procedure, the miRNAs and their targets are guaranteed at 99.9% accuracy. This procedure can be divided into two steps. Firstly, for each gene we estimate the number of miRNAs that are really targeting the gene as described in the following Step 2. Secondly, to find most probable real regulating miRNAs for each gene, we have to find the "best" one. Here, best is in the sense that it is the most unlikely combination to occur by chance, i.e. it has the smallest p-value whose details can be found in the following Step 3. In the last procedure, we find the combinations of miRNAs that are regulating common target genes. These common target genes are validated using GO. It is based on the hypothesis that miRNA's co-regulated genes usually have coherent functions. The identified co-regulating miRNAs have not only high accuracy in individual miRNA and its target genes, but also the target genes are functionally related.

Step 1.1. How many predicted regulating miRNAs are needed for us to believe a gene to be a real miRNA target?

The signal-to-noise ratio of TargetScanS on orthologous human, mouse, rat, chicken and dog sequences is 2.4:1.[3] That is, there is 1 false-positive prediction for every 2.4 predictions on average. So when TargetScanS predicts that a miRNA Y regulates a gene X, the probability p that the gene X is really regulated by the miRNA Y is $(2.4 - 1)/2.4 = 58.3\%$. It follows that, when TargetScanS predicts that a gene X is regulated by a group of n miRNAs, the probability that the gene X is really the target of at least one of these n miRNAs is $P_n = \text{Prob}(X \text{ is a miRNA target} | n \textit{ miRNAs match } X) = 1 - (1 - p)^n$. For example, $P_1 = 58.3\%$, $P_2 = 82.6\%$, $P_3 = 92.7\%$, $P_4 = 97\%$, $P_5 = 98.7\%$ and $P_8 = 99.9\%$.

Therefore, a human gene X with orthologs in mouse, rat, chicken and dog that is predicted by TargetScanS to have at least eight distinct miRNA target sites has a 99.9% chance of being a real miRNA target. Although some miRNAs have multiple target sites in one target gene, we consider one target gene for multiple target sites conservatively. So if we define $R_X = \{Y | (Y, X) \in R\}$, we first obtain a more reliable set $R' = \{(Y, X) \in R || R_X| \geq 8\}$ of miRNA–gene pairs by restricting the initial miRNA–gene pairs map R to those genes that are the target of at least eight distinct miRNAs.

Step 1.2. How many predicted regulating miRNAs are needed for us to believe a gene to be really co-regulated by k miRNAs?

If TargetScanS predicts n miRNAs to target a human gene X with orthologs in mouse, rat, chicken and dog, the probability that gene X is indeed co-regulated by at least two of these n miRNAs is $P_{2,n} = \text{Prob}(X \text{ is co-regulated by}$

at least 2 miRNAs$|n$ *miRNAs match* $X) = 1 - (1-p)^n - np(1-p)^{n-1}$, where $p = 58.3\%$ is the precision of TargetScanS mentioned in Step 1.1. For example, $P_{2,2} = 34\%$, $P_{2,3} = 62.3\%$, $P_{2,4} = 80.1\%$, $P_{2,12} = 99.95\%$. Thus a human gene X with orthologs in mouse, rat, chicken and dog that is matched by 12 miRNAs is likely to be co-regulated by at least two miRNAs with probability 99.9%. However, we do not know which two miRNAs among the 12 miRNAs really co-regulate X.

In general, if TargetScanS predicts n miRNAs to target a human gene X with orthologs in mouse, rat, chicken and dog, the probability that gene X is indeed co-regulated by at least k of these n miRNAs is $P_{k,n}$ = Prob (X is co-regulated by at least k *miRNAs*$|n$ *miRNAs match* $X) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i(1-p)^{n-i}$. Given a threshold value for $P_{k,n}$, we can estimate the maximum number k of miRNAs that co-regulate a gene X that has n predicted regulating miRNAs.

For this paper, we set the threshold on $P_{k,n}$ to 99.9%. Then we compute for each gene X selected in Step 1.1 that is predicted to have n co-regulating miRNAs by TargetScanS, the number k of co-regulating miRNAs that X can be assumed to have at 99.9% probability.

Step 2. Which k miRNAs most likely co-regulate a given gene?

It is reasonable to postulate that a pair of miRNAs Y_1 and Y_2 that are predicted to regulate a large number of genes in common are more likely to be co-regulating. We can assess the chance of such a co-regulation by a hypergeometric p-value, $pval(Y_1, Y_2) = \sum_{z \geq z_0} P(z|\ n, y_1, y_2) = \sum_{z \geq z_0} \binom{n}{z}\binom{n-z}{y_1-z}\binom{n-y_1}{y_2-z} / \binom{n}{y_1}\binom{n}{y_2}$, where z_0 is the number of genes that are predicted as targets of both Y_1 and Y_2, and y_i is the number of genes that are predicted as targets of Y_i $(i = 1, 2)$, and n is the total number of genes considered that are miRNA targets.[c]

To generalize the hypergeometric p-value to more than two miRNAs is complicated. The probability that k miRNAs $\{Y_1, \ldots, Y_k\}$ sharing at least z_0 common targets out of $y_1, \ldots, y_k$ targets can be shown to be no greater than $\binom{n}{z_0} \prod_{i=1}^{k} \binom{n-z_0}{y_i-z_0} / \binom{n}{y_i}$. We denote this expression by $Pval(Y_1, \ldots, Y_k)$ which estimates the actual p-value conservatively. Furthermore, the same expression was also used by Wu *et al.* (2003)[15] to provide a p-value for genome phylogenetic profiles.

Of n miRNAs that are predicted to regulate a given gene X from Step 1.1, we want to obtain k miRNAs that are most likely to co-regulate gene X. The value of k is calculated as the largest value so that $P_{k,n} \geq 99.9\%$, as per Step 1.2. After the value of k is calculated, we obtain the most likely co-regulating k miRNAs

[c]A possible variation is to further correct the counts n, z_0, y_1, and y_2 for predicting noise of TargetScanS. As noted earlier, TargetScanS has precision $p = 58.3\%$. Thus if TargetScanS predicts that Y_1 has y_1' targets, Y_2 has y_2' targets, Y_1 and Y_2 has z_0' common targets, and a total of n' genes to be target of some miRNAs, we need to set $y_1 = py_1'$, $y_2 = py_2'$, $z_0 = p^2 z_0'$, and $n = pn'$. This variation is better suited to the case of simultaneous co-regulation.

from the predicted list of n miRNAs that target the gene X, by enumerating the combinations of k miRNAs and picking the combination with the lowest p-value.

Step 3. Which are the genes that are most likely to be co-regulated by a given group of miRNAs?

Given a group M of miRNAs obtained in Step 2 that are co-regulators of a gene X, the list R' of miRNA-gene pairs from Step 1.1 can be used to identify a set $G_M = \{X \mid \text{for all } Y \in M,\ (Y, X) \in R'\}$ of all high-probability miRNA target genes that are co-regulated by this co-regulating miRNA group M.

While each gene in G_M has a probability of at least 99.9% of being the target of some miRNA, it may not necessarily be a high probability target of all miRNAs in M, due to the limitation of the statistics used in Step 1.1. Therefore, we propose a separate p-value assessment of G_M based on functional homogeneity.

We use Gene Ontology (GO)[16] annotations to check the functional coherence of gene groups. The more GO terms that are annotated to a gene group, the higher the probability that the miRNAs co-regulate this gene group. To take into consideration of hierarchical information of GO in this work, we use parent–child approach[17,18] to evaluate homogeneity of co-regulated genes. The GO is a hierarchical scheme. It is categorized into three name spaces: biological process (P), molecular function (F), and cellular component (C). The universal term and three name space root terms (GO:0003674, GO:0008150 and GO:000557) are not considered to evaluate the coherence of the co-regulated gene groups.

Suppose $G_M = \{X_1, \ldots, X_h\}$ are predicted to form a co-regulated gene group. Suppose the GO terms $t_1, \ldots, t_l$ are the GO terms used to annotate $X_1, \ldots, X_h$. Suppose GO term $t \in \{t_1, \ldots, t_l\}$ is annotated to gene set V in human, while t has parents $t_{p_1}, t_{p_2}, \ldots$, which are annotated to gene sets $u_{t_{p_1}}, u_{t_{p_2}}, \ldots$ in human as shown in Fig. 3. According to Ref. 18, we use the *union* of these genes as population set $U = u_{t_{p1}} \cup u_{t_{p2}} \cup \cdots$. Therefore the chance of t being annotated to z_0 genes in G_M is given by the hypergeometric p-value $pval(t, G_M) = \sum_{z \geq z_0} \binom{u}{z}\binom{u-z}{v-z}\binom{u-v}{w-z} / \binom{u}{v}\binom{u}{w}$, where $u = |U|$, $v = |V|$ and $w = |U \cap G_M|$.

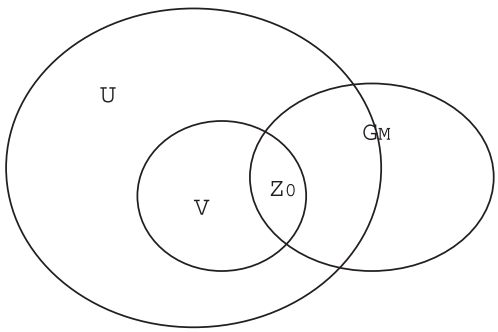

Fig. 3. Parent–child approach for validation of co-regulated target genes in terms of GO gene annotations.

As ℓ informative GO terms are annotated to G_M, we make a conservative Bonferroni correction to the p-value above by multiplying it by ℓ. We consider a gene group G_M to be functionally homogeneous if it has a sufficiently small Bonferroni-corrected p-value. Then we propose (M, G_M) to be a co-regulation miRNA–gene group pair. The Bonferroni-corrected p-value threshold is set to 0.01 in this work.

3. Results

We obtain target genes for human miRNAs using TargetScanS.[3] miRNAs having the same seed region are viewed as one family as they are considered to have the same predicted target genes. In this work, each family is represented by one miRNA. We collate the raw predictions from TargetScanS to a matrix whose columns are 162 miRNAs (miRNA families) and rows are 3227 genes; corresponding to the miRNA–gene pair map R mentioned in the preamble of Sec. 2. A gene has an average of 4.9 regulating miRNAs. A miRNA, on average, regulates 98.9 genes. Every element in the matrix represents the number of seed binding sites between corresponding miRNA and gene. If the element is non-zero, the corresponding gene and miRNA have regulatory association.

Figure 4 shows the frequency distribution of the number of the miRNAs that are predicted to regulate a gene. The vertical axis represents the number of miRNA per gene predicted by TargetScanS. The horizontal axis represents the number of genes. More than half of the genes have 1, 2 or 3 miRNAs that are predicted to regulate the genes. There are several exceptions. The nuclear factor I/B (NFIB) gene has

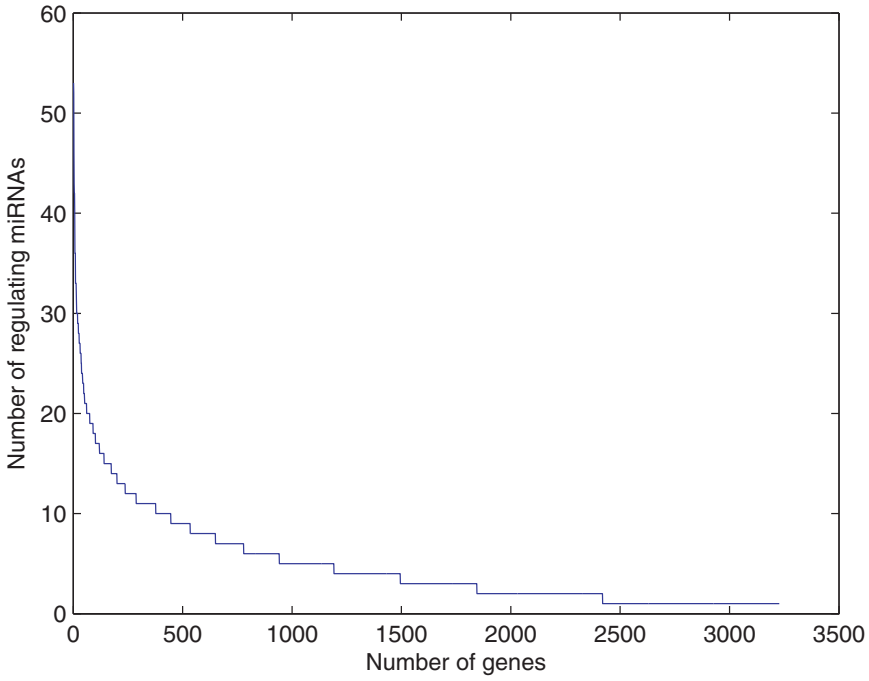

Fig. 4. Frequency distribution of predicted miRNAs per gene.

the largest number of predicted regulating miRNAs (53 miRNAs). NFIB has 6562 nucleotides in its 3′UTR. As another example, the trinucleotide repeat containing 6B (TNRC6B) gene has 52 predicted miRNAs and its 3′UTR has 12,685 nucleotides. In total, 19 genes have more than 30 predicted regulating miRNAs. And 445 genes have more than 10 predicted miRNAs.

In Step 1.2 we derive the probability $P_{k,n}$ for a gene to be a target for a group of co-regulating miRNAs of size at least k if it is predicted to be the target of n miRNAs by TargetScanS. We set a threshold of 99.9% for $P_{k,n}$ and obtain the largest value of k for each gene, from among the list of reliable miRNA target genes computed in Step 1.2, that is predicted to be the target of n miRNAs by TargetScanS. This k represents the largest number of miRNAs which co-regulate the gene with high probability. Figure 5 shows the largest number of high probability co-regulating miRNAs (vertical axis) versus the number of predicted miRNAs (horizontal axis). As mentioned earlier, one gene (NFIB) is targeted by 53 predicted miRNAs. Based on our calculation, this gene is co-regulated by at least 30 miRNAs among the 53 predicted miRNAs with high probability. We enumerate all possible combinations of 30 miRNAs from the 53 miRNAs and calculate their p-values, as per Step 2, to identify the most likely 30 co-regulating miRNAs from the list of 53 miRNAs. The miRNA combination with the lowest p-value is considered to be the real co-regulating miRNA group for the gene. We have also performed the same analysis for other high probability genes. As this is a time-consuming task, we select here only up to a maximum of 12 miRNAs for each gene.

Finally, as per Step 3, for each co-regulating miRNA group, we combine all the target genes of the miRNAs in the group to find the genes that they co-regulate in common. We get three co-regulating miRNA tables in terms of number of co-regulating miRNAs. Altogether, we have identified 12 4^+-term miRNA groups as shown in Table 1. We have also identified $1-$, $2-$ and 3-term miRNA groups whose p-values are less than 0.01. These tables can be found in Supplementary Information. Each group corresponds to one Excel Sheet. They have the same format as shown in Table 1: 1st column is miRNA group, 2nd column is its target genes and 3rd column shows the p-value.

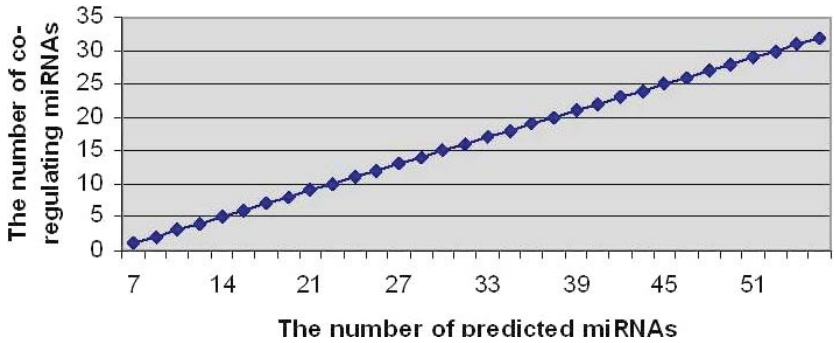

Fig. 5. A plot of $\text{argmax}_k P_{k,n}$ versus n.

Table 1. Four-term miRNA and targeting gene groups.

MiRNA groups	Gene groups	*p*-value
miR-181,miR-27,miR-25/32/92/363/367,miR-30-5p	NOVA1,CPEB4,QKI	3.59E-03
miR-144,miR-101,miR-27,miR-128	TNPO1,ARID2,FLRT3,RNF38 MEIS2,NR5A2,PDS5B CDH11,FBXW7,KBTBD8	4.25E-03
miR-381,miR-15/16/195/424/497,miR-9,miR-30-5p	CPEB3,RAP2C,CPEB2	8.19E-03
miR-144,miR-101,miR-19,miR-130/301	ATXN1,RNF38,ROBO2,ZEB2 NEUROD1,BCL2L11,ERBB4	1.16E-02
miR-381,miR-9,miR-25/32/92/363/367,miR-30-5p	CPEB3,CPEB2,CPEB4	1.22E-02
miR-144,miR-101,miR-19,miR-25/32/92/363/367	RNF38,ROBO2,ZEB2,NLK PCDH10	1.40E-02
miR-181,miR-9,miR-19,miR-25/32/92/363/367	CPEB4,DDX3X,RAP1B	2.62E-02
miR-181,miR-19,miR-25/32/92/363/367,miR-30-5p	TNRC6B,CPEB4,RAP1B	2.62E-02
miR-144,miR-101,miR-27,miR-19	RNF111,RNF38,ZEB2,NLK	3.28E-02
miR-381,miR-144,miR-101,miR-27,miR-128	ZFHX3,MEIS2,NR5A2	4.62E-02
miR-144,miR-101,miR-27,miR-26	TNPO1,ARID2,DYRK1A NLK,KBTBD8	4.81E-02
miR-144,miR-27,miR-128,miR-26	TNPO1,ARID2,ARFGEF1 ANK2,KBTBD8	4.81E-02

3.1. *Can we find co-regulating miRNAs from shuffled miRNAs?*

To observe how many co-regulating miRNA groups are generated by chance, we artificially construct miRNAs that correspond to original miRNAs to see their co-regulating miRNA groups. In this work, to keep nucleotide ingredient, we shuffle all original miRNA sequences and pass them through miRNA prediction tool TargetScanS. Then we use Gene Ontology to extract functional co-regulating miRNA groups as described in Step 3.

We shuffle original 648 human miRNA sequences. TargetScanS takes these sequences to find target genes from non-homolog genes. It is because if we do not remove homolog genes, those genes become co-regulated by miRNAs due to them having the same 3′UTRs. For example, genes PCDHA2, PCDHA3, PCDHA4 , . . . , PCDHA8 have the same 3'UTR sequences. Therefore, only one gene PCDHA3 is left for the homolog group. The co-regulating miRNAs that have less than 0.05 *p*-value are selected. We repeated the procedure 10 times. The comparisons of the number of selected co-regulating miRNA groups between shuffled and original miRNAs are shown in Fig. 6. The ratio of the number of generated co-regulating miRNAs from shuffled miRNAs to original miRNAs is very small. In 2, 3 and 4 miRNA combinations, the ratios are 0.28, 0.11 and 0.05 respectively, which means the functional homogeneity of the groups of target genes from the shuffled miRNAs is significantly less than the original miRNAs.

In order to keep the same number of total target genes, we shuffle miRNAs in term of each gene to get randomized pairs of miRNA and target genes. This randomization strategy was used in Ref. 19: the miRNAs and their target genes form a matrix, whose rows and columns represent miRNAs and target genes, respectively.

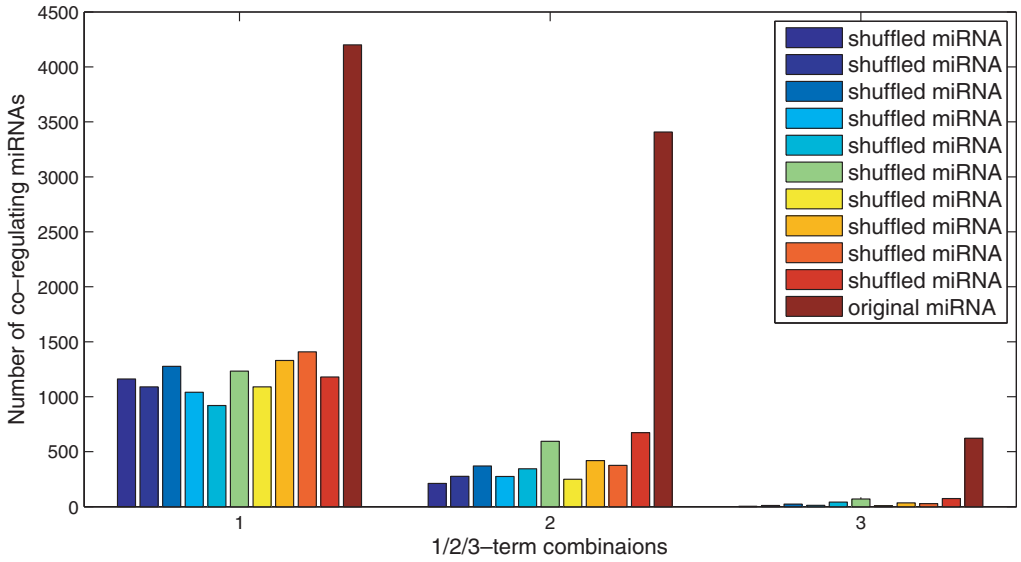

Fig. 6. Comparison of the number of generated co-regulating miRNA groups.

The shuffling is carried out in every column. That is, the regulating miRNAs are randomly re-placed. The total number of regulating miRNAs for each gene is kept unchanged, but the number of target genes for each miRNA is dramatically changed as shown in Fig. 7. The average number of original target genes (blue bar) is reduced as the number of terms of miRNAs increases. The average number of randomized

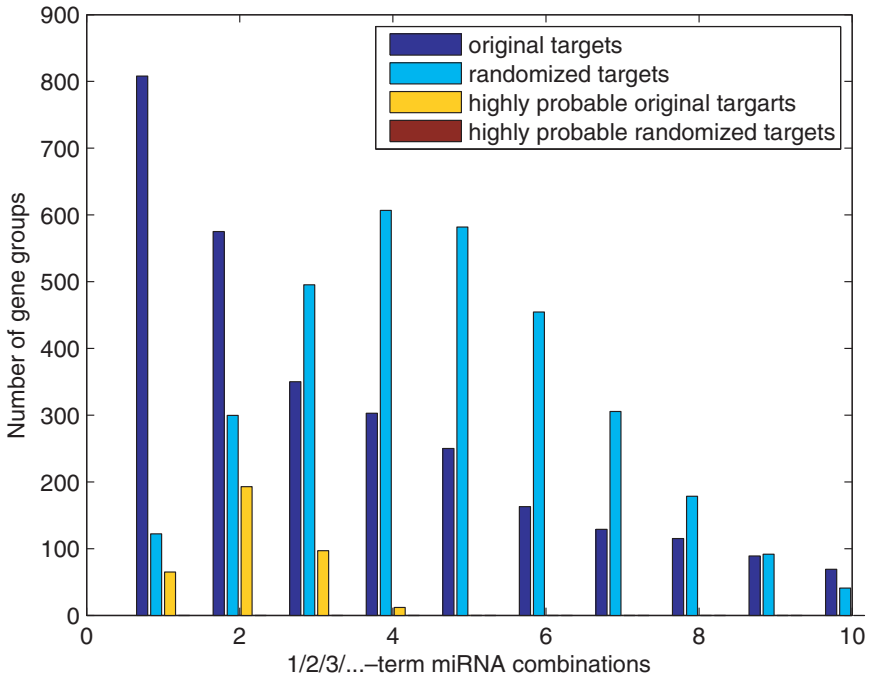

Fig. 7. Identifying co-regulating miRNA using randomized target genes.

target genes (green bar) is shown to be near normal distribution. The result is in agreement with that of Ref. 19: genes are mainly co-regulated by 3, 4, 5 miRNA combinations. Yellow/magenta bars represents the number of highly probable original/randomized targets in term-$1, 2, \ldots$ using GO evaluation method of Step 3. We can see that there is no highly probable randomized targets (magenta bar) at all. It means that we cannot find any co-regulating miRNAs from randomized target genes.

4. Discussion

Functionally related genes are usually regulated by a group of miRNAs instead of one miRNA. For example, breast cancer genes are regulated by 29 miRNAs.[20] It is a challenging task to find all genes that are related to a particular disease.[21–23] Recent research has demonstrated that many miRNAs are related to several diseases.[24] Finding a group of miRNAs that regulate functionally related genes is one of the key issues in the miRNA field.

Current miRNA target prediction tools have the common problem that their false positive rate is quite high.[2,3] This complicates reliable identification of co-regulating groups of miRNAs and target genes. In this study, we attempt to find highly probable co-regulating miRNAs and the corresponding co-regulated gene groups. We derived, based on the signal-to-noise ratio of TargetScanS, the minimum number of miRNAs that are predicted by TargetScanS to target a gene in order for that gene to have at least 99.9% probability of being regulated by a miRNA. Specifically, when a gene is predicted by TargetScanS to have five or more regulating miRNAs, then there is at least 99.9% probability that at least one of the miRNAs is really regulating the gene, even though we cannot pinpoint the exact miRNA that regulates the gene. We have also developed a sequence of hypergeometric p-values that allow us to rank combinations of miRNAs that are likely to co-regulate a given gene. Since experimental identification of miRNA target is still in the infant stage, providing such high probability miRNA-targets information should be helpful to the biologists.

Recent studies show that some genes are regulated by many miRNAs, while some genes may not be regulated by any miRNA. For example, hub proteins have more regulating miRNAs than other proteins.[25] To validate the identified co-regulating miRNAs and their corresponding co-regulated gene groups, we check whether the co-regulated genes have coherent GO term annotations.[12] In this work, we filter out miRNA groups if their co-regulated gene groups have p-values larger than 0.01.

Table 1 shows the 4^{+}-term co-regulating miRNA groups. (The 1-, 2- and 3- term miRNA groups can be found in Supplementary Information). These tables are useful to suggest some hypothesis on the functional role of several miRNAs, through a "guilt by association" reasoning similar to that of genome phylogenetic profiling.[15] For example:

Hsa-miR-144, hsa-miR-101, hsa-miR-19 and hsa-miR-130/301 have similar co-regulated gene groups. They co-regulate seven genes: "ATXN1", "BNF38",

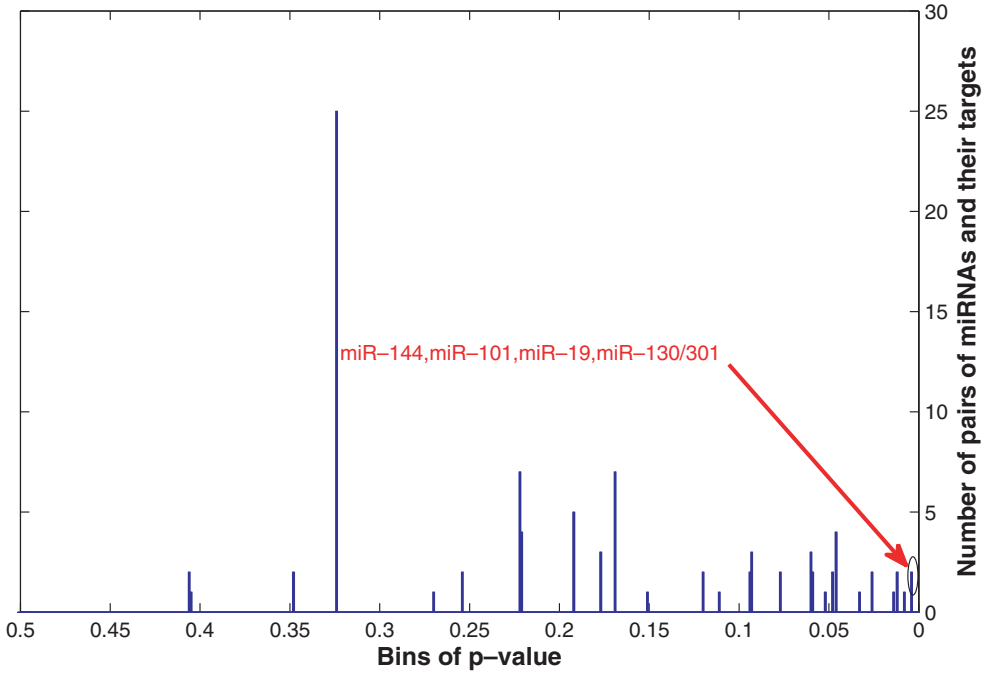

Fig. 8. The distribution of p-values for 4-term miRNA groups.

"ROBO2", "ZEB2", "NEUROD1", "BCL2L11", "ERBB4". Its p-value is 1.16×10^{-2}. The co-regulation is related to GO:0005515 (protein binding) domain. hsa-miR-101, hsa-miR-130 and hsa-miR-19 have been validated to co-regulate neurodegenerative diseases related gene "ATXN1".[26] In Lee *et al.* paper,[26] three miRNAs hsa-miR-130/19/101 are confirmed to be co-regulating with gene ATXN1, but our approach found the 4th miRNA hsa-miR-144 co-regulating hsa-miR-130/19/101. It suggests that hsa-miR-144 is another co-regulating miRNA for gene ATXN1. It is notable that if all co-regulating miRNAs are co-expressed, the target genes will be suppressed heavily. Otherwise, the target genes cannot be inhibited effectively.

Figure 8 shows the histogram of p-values for 4-term co-regulating miRNA groups. The two miRNA groups mentioned have the smallest p-values in all 4-term miRNA groups.

5. Conclusions

In this work, a statistic approach has been proposed to reduce the number of miRNAs that are regulating a specific gene. Some identified co-regulating miRNAs are supported by biological literature. The main reason may be that Gene Ontology has been used for evaluation, because miRNAs usually inhibit genes in protein level; Gene Ontology annotation is based on UniProt (Universal Protein Resource) Knowledgebase.[16]

From the predicted result by TargetScanS, we found highly probable co-regulating miRNA groups in terms of the number of miRNAs: 65 (1-term); 193

(2-term); 97 (3-term) and 12 (4^+-term). For given genes, we can find a very small candidate set of co-regulating miRNAs. It significantly reduces labor and financial cost to inhibit the genes of interest.

Acknowledgments

We thank Professor Limsoon Wong from National University of Singapore for making essential contribution to this paper. We also thank the research associations below: JA was supported by a grant to the National Centre for Adult Stem Cell Research from the Australian Department of Health and Ageing. KPC was supported by an NUS ARF grant R-155-000-051-112. CAW was supported by NHMRC CDA 481945. PYC was supported by Australian Research Council grants DP0559251 and LX0560616.

References

1. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, Human microRNA targets, *LOS Biol* **2**(11):e363, 2004.
2. Lewis B, Shih I, Jones-Rhoades M, Bartel D, Burge C, Prediction of mammalian microRNA targets, *Cell* **115**:787–798, 2003.
3. Lewis B, Burge C, Bartel D, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell* **120**:15–20, 2005.
4. Kiriakidou M, Nelson P, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A, A combined computational-experimental approach predicts human microRNA targets, *Genes Dev* **18**(10):1165–1178, 2004.
5. Lall S, Grun D, Krek A, Chen K, Wang Y, Dewey C, Sood P, Colombo T, Bray N, Macmenamin P, Kao H, Gunsalus K, Pachter L, Piano F, Rajewsky N, A genome-wide map of conserved microRNA targets in *C. elegans*, *Curr Biol* **16**:460–471, 2006.
6. Sethupathy P, Corda B, Hatzigeorgiou A, TarBase: A comprehensive database of experimentally supported animal microRNA targets, *RNA* **12**:192–197, 2006.
7. Thadani R, Tammi M, MicroTar: Predicting microRNA targets from RNA duplexes, *BMC Bioinformatics* **7**(Suppl 5):S20, 2006.
8. Rajewsky N, MicroRNA target predictions in animals, *Nat Genet* **38**:S8–S13, 2006.
9. Hua Z, Lv Q, Ye W, Wong CK, Cai G, Gu D, Ji Y, Zhao C, Wang J, Yang B, Zhang Y, MiRNA-directed regulation of VEGF and other angiogenic factors under hypoxia, *PLoS ONE* **1**:e116, 2006.
10. Krek A, Gruen D, Poy M, Wolf R, Rosenberg L, Epstein E, MacMenamin P, da Piedade I, Gunsalus K, Stoffel M, , Rajewsky N, Combinatorial microRNA target predictions, *Nat Genet* **37**:495–500, 2005.
11. Lai E, Predicting and validating microRNA targets, *Genome Biol* **5**(9):115, 2004.
12. Yoon S, Micheli G, Prediction of regulatory modules comprising microRNAs and target genes, *Bioinformatics* **21**(Suppl 2):ii93–ii100, 2005.
13. Joung JG, Hwang K, Nam J, Kim S, Zhang B, Discovery of microRNA-mRNA modules via population-based probabilistic learning, *Bioinformatics* **23**(9):1141–1147, 2007.
14. Blanchette M, Kent W, Riemer C, Elnitski L, Smit A, Roskin K, Baertsch R, Rosenbloom K *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner, *Cancer Res* **14**:708–715, 2004.

15. Wu J, Kasif S, DeLisi C, Identification of functional links between genes using phylogenetic profiles, *Bioinformatics* **19**(12):1524–1530, 2003.
16. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K *et al.*, Gene ontology: Tool for the unification of biology, *Nat Genet* **25**:25–29, 2000.
17. Alexa A, Rahnenfuhrer J, Lengauer T, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, *Bioinformatics* **22**(13):1600–1607, 2006.
18. Grossmann S, Bauer S, Robinson P, Vingron M, Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis, *Bioinformatics* **23**:3024–3031, 2007.
19. Shalgi R, Lieber D, Oren M, Pilpel Y, Global and local architecture of the mammalian microRNAtranscription factor regulatory network, *PLoS Comput Biol* **3**(7):e131, 2007.
20. Calin G, Croce C, MicroRNA-cancer connection: The beginning of a new tale, *Cancer Res* **66**:7390–7394, 2006.
21. An J, Chen YPP, Finding rule groups to classify high dimensional gene expression datasets, *Comput Biol Chem* **33**(1):108–113, 2009.
22. An J, Chen YPP, Finding edging genes from microarray data, *J Biotechnol* **135**(3):233–240, 2008.
23. Ein-Dor L, Kela I, Getz G, Givol D, Domany E, Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **21**:171–178, 2005.
24. Caldas C, Brenton J, Sizing up miRNAs as cancer genes, *Nat Med* **11**:712–714, 2005.
25. Liang H, Li W, MicroRNA regulation of human protein-protein interaction network, *RNA* **13**:1402–1408, 2007.
26. Lee Y, Samaco R, Gatchel J, Thaller C, Orr H, Zoghbi H, miR-19, miR-101 and miR-130 co-regulate ATXN1 levels to potentially modulate SCA1 pathogenesis, *Nat Neurosci* **11**(10):1137–1139, 2008.

Jiyuan An received his Bachelor's degree from Hefei University of Technology, China, in 1986 and his Master's degree from Kyushu Institute of Technology, Fukuoka, Japan, in 1998. From 1998 to 2000, he worked for the Hitachi Government & Public Corporation System Engineering, Ltd., Tokyo, Japan. In 2003, he obtained his Ph.D. degree in Engineering (Information Sciences and Electronics) from the University of Tsukuba, Ibaraki, Japan. In April 2003, he moved to Australia to start his academic career. He has worked as a research fellow in biological data mining at the Queensland University of Technology, Deakin University and Griffith University. His research interests include identification of microRNA targets and transcript factors. He is also working on finding correlated genes based on microarray and sequencing data.

Kwok-Pui Choi received his B.Sc. (1st Class) degree from the University of Hong Kong, and M.Sc. and Ph.D. degree from the University of Illinois at Urbana-Champaign. He is now an Associate Professor at the Department of Statistics and Applied Probability at the National University of Singapore (NUS). He has a joint appointment with the Department of Mathematics at NUS. His research interests include probability and computational biology.

Christine Wells' research interests are in the field of genome biology, and its application to innate immunity. Her primary research goal is the characterization of a new class of inflammatory regulator. She was recently awarded an NHMRC Career Development Award fellowship to develop new paradigms in innate immune signalling. Recently, she has applied these technologies to human adult stem cell biology, as a member of the collaborative team at the National Centre for adult Stem Cell Research, Griffith University. Christine is a member of the FANTOM and Genome Network transcriptome consortia, as well as the Functional Glycomics consortium.

Yi-Ping Phoebe Chen received the B.Inf.Tech.(First Honor) and Ph.D. degrees in Computer Science from the University of Queensland, Brisbane, Qld., Australia. She is currently an Associate Professor at Deakin University, Melbourne, Vic., Australia, where she is the Director of the Bioinformatics Group, and the Chief Investigator of the ARC Centre of Excellence in Bioinformatics. Her current research interests include bioinformatics, multimedia databases and technology, Web information systems, machine learning, and data mining. She is the recipient of 23 research grants (including 12 prestigious Australia Research Council grants). She has authored or co-authored more than 130 refereed publications, in journals including the *Nucleic Acids Research*, *BMC Genomics*, *BMC Bioinformatics*, *Data Mining and Knowledge Discovery*, *ACM Transactions on Multimedia Computing*, *Communications and Applications*, *IEEE Transaction on Knowledge and Data Mining*. She is an Associate Editor for a number of journals. Dr. Chen is the Chair of the Steering Committee of Asia Pacific Bioinformatics Conference (founder) and Multimedia Modeling.