BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 621–628
doi:10.1093/bioinformatics/btt599

Genome analysis

Advance Access publication October 17, 2013

A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information

Yue Li[1,2,*], Anna Goldenberg[1,3], Ka-Chun Wong[1,2] and Zhaolei Zhang[1,2,4,*]

[1]Department of Computer Science, [2]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, [3]Genetics and Genome Biology, SickKids Research Institute, Toronto, Ontario M5G 1L7 and [4]Banting and Best Department of Medical Research and Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 3E1, Canada

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Systematic identification of microRNA (miRNA) targets remains a challenge. The miRNA overexpression coupled with genome-wide expression profiling is a promising new approach and calls for a new method that integrates expression and sequence information.

Results: We developed a probabilistic scoring method called *targetScore*. TargetScore infers miRNA targets as the transformed fold-changes weighted by the Bayesian posteriors given observed target features. To this end, we compiled 84 datasets from Gene Expression Omnibus corresponding to 77 human tissue or cells and 113 distinct transfected miRNAs. Comparing with other methods, targetScore achieves significantly higher accuracy in identifying known targets in most tests. Moreover, the confidence targets from targetScore exhibit comparable protein downregulation and are more significantly enriched for Gene Ontology terms. Using targetScore, we explored oncomir–oncogenes network and predicted several potential cancer-related miRNA–messenger RNA interactions.

Availability and implementation: *TargetScore* is available at Bioconductor: http://www.bioconductor.org/packages/devel/bioc/html/TargetScore.html.

Contact: yueli@cs.toronto.edu or zhaolei.zhang@utoronto.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 18, 2013; revised on October 6, 2013; accepted on October 15, 2013

1 INTRODUCTION

MicroRNAs (miRNAs) repress protein production in animal species by forming Watson–Crick base pairing to the 3′-untranslated regions of the target messenger RNAs (mRNAs) (Friedman *et al.*, 2009). The binding primarily occurs at the 2–7 nt positions from the 5′ end of the miRNA, which is termed as the 'seed' and the binding as the 'seed match' (Lewis *et al.*, 2003). MiRNA regulations have been implicated in numerous developmental and pathogenic processes (Bartel, 2009). Functional characterization of miRNAs depends on precise identification of their targets. However, it has proved difficult to experimentally identify miRNA–mRNA interactions.

To date, according to mirTarBase, only 3565 interactions between 432 miRNAs and 1959 target genes in human have been confirmed using high-confidence low-throughput assays such as Western blot (Hsu *et al.*, 2011). On the other hand, computational prediction provides a rapid alternative method to identify putative miRNA targets that can be subsequently validated. However, accurate prediction of miRNA targets remains a challenge with the current state-of-the-art algorithms achieving <50% specificity and having poor agreement among them (Alexiou *et al.*, 2009). Most of these prediction programs are based on sequence complementarity, evolutionary conservation (Friedman *et al.*, 2009; Lewis *et al.*, 2003), free energy (Enright *et al.*, 2004; Krek *et al.*, 2005; Lewis *et al.*, 2003) and/or target site accessibility (Kertesz *et al.*, 2007). Although it is shown that evolutionary conservation can improve signal-to-noise ratios, the conservation approach is limited, as not all of the functional target sites are conserved (e.g. lineage- and species-specific target sites) and vice versa. In particular, the performance of conservation-based methods drops drastically when restricted to only mammalian genomes because of short evolutionary time (Friedman *et al.*, 2009). Similarly, the energy- or accessibility-based methods rely on secondary structure prediction tools such as RNAfold (Lorenz *et al.*, 2011), which itself has room for improvements. These limitations also underscore the historical lack of genome-wide functional data that measure the *in vivo* impact of miRNA regulation, which was alleviated by the recent developments of transcriptomic and proteomic profiling methods (Baek *et al.*, 2008; Lim *et al.*, 2005; Selbach *et al.*, 2008).

Particularly, overexpression of miRNA coupled with expression profiling of mRNA by either microarray or RNA-seq has proved to be a promising approach (Arvey *et al.*, 2010; Lim *et al.*, 2005). Consequently, genome-wide comparison of differential gene expression holds a new promise to elucidate the global impact of a specific miRNA regulation without solely relying on evolutionary conservation. To improve the prediction of relative repression of mRNA, several studies have proposed a series of additional determinants including seed match type (6mer seed, 7mer-tA1, 7mer-m8 and 8mer), number of target sites, site relative location on the 3′-untranslated region, local AU content, 3′-supplementary paring, seed-pairing stability and target-site abundance (Arvey *et al.*, 2010; Garcia *et al.*, 2011; Grimson *et al.*, 2007), which are combined together in a single value termed as the 'context score' made available from targetScan Web site (Garcia *et al.*, 2011). Accordingly, we hypothesize that target prediction can be improved by integrating expression change and sequence information such as context score and

*To whom correspondence should be addressed.

other orthogonal sequence-based features such as conservation (Friedman *et al*., 2009) into a probabilistic score.

The proposed model differs from most of the previous expression-based target prediction methods in three important aspects. First, our model is specifically designed for miRNA-overexpression experiments to interrogate targets of a particular miRNA in a specific cell condition. To our knowledge, only a few methods are suitable for such task (Liu *et al*., 2010a). Most existing methods such as GenMiR++ (Huang *et al*., 2007) and GroupMiR (Le and Bar-Joseph, 2011) are based on global miRNA-target expression correlation, which requires a large set of expression profiles of both mRNA and miRNA measured across various distinct conditions and may miss targets that are specific to only a subset of the input samples. Second, the proposed model is unsupervised such that it infers miRNA-targets solely based on their distinct high-dimensional patterns of expression fold-changes and sequence features. However, the existing methods are mostly regression-based frameworks by treating gene expression as response and miRNA expression as input variables (Huang *et al*., 2007; Le and Bar-Joseph, 2011) or supervised learning by explicitly operating on a training dataset of confidence positive and negative targets (Liu *et al*., 2010b; Sumazin *et al*., 2011), which are incomplete and difficult to obtain. Third, our method operates on the entire gene set to more closely model the overall likelihood rather than only on a subset of genes prefiltered by targetScan score (Huang *et al*., 2007) or sample variance (Le and Bar-Joseph, 2011).

2 METHODS

2.1 Overview

We describe a novel probabilistic method for miRNA target prediction problem by integrating miRNA-overexpression data and sequence-based scores from other prediction methods. Briefly, each score feature is considered an independent observed variable as input to a variational Bayesian–Gaussian mixture model (VB-GMM). We chose a Bayesian over a maximum likelihood approach to avoid overfitting. Specifically, given expression fold-change (due to miRNA transfection), we use a three-component VB-GMM to infer downregulated targets accounting for genes with little or positive fold-change [due to off-target effects (Khan *et al*., 2009)]. Otherwise, two-component VB-GMM is applied to unsigned sequence scores. The parameters for the VB-GMM are optimized using variational Bayesian expectation maximization algorithm. The mixture component with the largest absolute means of observed negative fold-change or sequence score is associated with miRNA targets and denoted as 'target component'. The other components correspond to the 'background component'. It follows that inferring miRNA–mRNA interactions is equivalent to inferring the posterior distribution of the target component given the observed variables. The targetScore is computed as the sigmoid-transformed fold-change weighted by the averaged posteriors of target components over all of the features (3).

2.2 Bayesian mixture model

Assuming there are N genes, we denote $\mathbf{x} = (x_1, \ldots, x_N)^T$ as the log expression fold-change ($\mathbf{x}_f$) or sequence scores ($\mathbf{x}_l$). Thus, for L sets of sequence scores, $\mathbf{x} \in \{\mathbf{x}_f, \mathbf{x}_1, \ldots, \mathbf{x}_L\}$. To simplify the following equations, we use $\mathbf{x}$ to represent one of the independent variables without loss of generality. To infer target genes for a miRNA given $\mathbf{x}$, we need to obtain the posterior distribution $p(\mathbf{z}|\mathbf{x})$ of the latent variable $\mathbf{z} \in \{z_1, \ldots, z_K\}$, where $K=3$ ($K=2$) for modeling signed (unsigned) scores such as logarithmic fold-changes (sequence scores).

We follow the standard Bayesian GMM based on Bishop (2006, pp. 474–482) with only minor modifications. Although univariate GMM ($D = 1$) is applied to each variable separately, we implemented and describe the following formalism as a more general multivariate GMM, allowing modeling the covariance matrices. Briefly, the latent variables $\mathbf{z}$ are sampled at probabilities π (mixing coefficient), that follow a Dirichlet prior $Dir(\pi|\alpha_0)$ with hyperparameters $\alpha_0 = (\alpha_{0,1}, \ldots, \alpha_{0,K})$. To account for the relative frequency of targets and non-targets for any miRNA, we set the $\alpha_{0,1}$ (associated with the target component) to aN and other $\alpha_{0,k} = (1-a) \times N/(K-1)$, where $a = 0.01$ (by default). Assuming $\mathbf{x}$ follows a Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda}^{-1})$, where $\mathbf{\Lambda}$ (precision matrix) is the inverse covariance matrix, $p(\mu, \mathbf{\Lambda})$ together follow a Gaussian–Wishart prior $\prod_k^K \mathcal{N}(\mu_k|\mathbf{m}_0, (\beta_0\mathbf{\Lambda})^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0)$, where the hyperparameters $\{\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0\} = \{\hat{\mu}, 1, \mathbf{I}_{D\times D}, D+1\}$.

2.3 Variational Bayesian expectation maximization

Let $\theta = \{\mathbf{z}, \pi, \mu, \mathbf{\Lambda}\}$. The marginal log likelihood can be written in terms of lower bound $\mathcal{L}(q)$ (first term) and Kullback–Leibler divergence $\mathcal{KL}(q||p)$ (second term):

$$\ln p(\mathbf{x}) = \int q(\theta) \ln \frac{p(\mathbf{x}, \theta)}{q(\theta)} + \int q(\theta) \ln \frac{q(\theta)}{p(\theta|\mathbf{x})} \quad (1)$$

where $q(\theta)$ is a proposed distribution for $p(\theta|\mathbf{x})$, which does not have a closed form distribution. Because $\ln p(\mathbf{x})$ is a constant, maximizing $\mathcal{L}(q)$ implies minimizing $\mathcal{KL}(q||p)$. The general optimal solution $\ln q_j^*(\theta_j)$ is the expectation of variable j with respect to other variables, $\mathbb{E}_{i\neq j}[\ln p(\mathbf{x}, \theta)]$. In particular, we define $q(\mathbf{z}, \pi, \mu, \mathbf{\Lambda}) = q(\mathbf{z})q(\pi)q(\mu, \mathbf{\Lambda})$. The expectations for the three terms (at log scale), namely, $\ln q^*(\mathbf{z})$, $\ln q^*(\pi)$, $\ln q^*(\mu)$, have the same forms as the initial distributions due to the conjugacy of the priors. However, they require evaluation of the parameters $\{\mathbf{z}, \pi, \mu, \mathbf{\Lambda}\}$, which in turn all depend on the expectations of $\mathbf{z}$ or the posterior of interest:

$$p(z_{nk}|\mathbf{x}_n, \theta) \equiv \mathbb{E}[z_{nk}] = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}} \quad (2)$$

where
$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\mathbf{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k, \mathbf{\Lambda}_k}[(\mathbf{x}_n - \mu_k)^T\mathbf{\Lambda}_k(\mathbf{x}_n - \mu_k)]$. The inter-dependence between the expectations and model parameters falls naturally into an expectation-maximization (EM) framework, namely variational Bayesian expectation maximization. Briefly, we first initialize the model parameters based on priors and randomly sample K data points μ. At the i^{th} iteration, we evaluate (2) using the model parameters (VB-E step) and update the model parameters using (2) (VB-M step). The EM iteration terminates when $\mathcal{L}(q)$ improves by $<10^{-20}$ (default). Please refer to Bishop (2006) for more details.

2.4 TargetScore

We define the targetScore as an integrative probabilistic score of a gene being the target t of a miRNA:

$$\text{targetScore} = \sigma(-\log FC)\left(\frac{1}{K+1}\sum_{\mathbf{x}\in\{\mathbf{x}_f, \mathbf{x}_1, \ldots, \mathbf{x}_L\}} p(t|\mathbf{x})\right) \quad (3)$$

where $\sigma(-\log FC) = \frac{1}{1+\exp(\log FC)}$, $p(t|\mathbf{x})$ is the posterior in (2).

2.5 miRNA-overexpression data collection

We collected miRNA-overexpression data corresponding to 84 Gene Expression Omnibus (GEO) series, 6 platforms, 77 human cells or tissues and 113 distinct miRNAs (Supplementary Table S1). To our knowledge, this is by far the largest compendium of miRNA-overexpression data. To automate data downloading and processing, we developed a pipeline written in R, making use of the function `getGEO` from *GEOquery* R/Bioconductor package (Davis and Meltzer, 2007). For each dataset, the pipeline downloads and quantile normalizes (if necessary) the expression data and calculates (when necessary) the log fold-change (logFC) in

treatment (miRNA transfected) versus (mock) control. For mRNAs interrogated by multiple probes in a single experiment, we took the average of the fold-changes. Each of the 286 data vectors contains logFC values for $N \leq 18560$ RefSeq mRNAs (i.e. with prefix NM for known protein coding mRNA obtained from University of California, Santa Cruz) due to transfection of one of the 113 miRNAs. For two logFC vectors associated with the same miRNA transfection (conducted in different studies), we removed mRNAs for which neither of the vectors contains a logFC value and filled the remaining missing values in one vector with the non-missing values in the other. For more than two logFC vectors for the same miRNA transfection, we removed mRNAs absent in all of those vectors and performed imputation for the remaining missing values using `impute.knn` from *impute* R package (Troyanskaya *et al.*, 2001). Finally, we picked one representative logFC vector per miRNA, which has the highest Pearson correlation with the binary vector of the validated targets (Hsu *et al.*, 2011) or averaged them if no validated target was available. As a result, we have 113 logFC data vectors for 113 distinct miRNA transfections.

2.6 Comparison with other prediction methods

We compared targetScore with seven published methods (Table 1). Among these methods, Expmicro is the only method that also uses miRNA-overexpression data. The other six methods are sequence-based, and their predictions are fine-tuned by the authors and the results provided on their Web site are the most accurate ones. Thus, we only ran Expmicro locally on the same test data and directly downloaded the target predictions by the other six programs from their corresponding Web sites. Specifically, predictions on all target sites (including conserved and non-conserved sites) from targetScan 6.1 were obtained for both context + scores (CS) and probability of conserved targeting (PCT). The latest PicTar2 target predictions were obtained from doRiNA database (database of RNA interactions in post-transcriptional regulation) (Anders *et al.*, 2011). The miRanda predictions for sites with both good and non-good mirSVR scores were obtained at http://www.microrna.org/. For Probability of Interaction by Target Accessibility (PITA) predictions, the accessibility energy $\Delta\Delta G$ on target sites flanked by 3/15 nt upstream/downstream ('3/15 flank') was chosen as recommended by the authors (Kertesz *et al.*, 2007). SVMicro predictions were obtained from http://compgenomics.utsa.edu/svmicro.html. Expmicro was run on the SVMicro scores and expression fold-changes (Liu *et al.*, 2010a).

2.7 Evaluation

2.7.1 Gold standard Experimentally validated targets were downloaded from mirTarBase (Hsu *et al.*, 2011) (http://mirtarbase.mbc.nctu.edu.tw). MirTarBase is one of the most comprehensive databases that support bulk download and agree with other databases such as TarBase (Vergoulis *et al.*, 2012) and miRecords (Xiao *et al.*, 2009). MirTarBase 3.5 contains 3565 validated interactions between 432 human miRNAs and 1959 target genes. The average (median) number of targets per miRNA is 8.2 (3).

2.7.2 Methods of comparison We first compared the performances of using logFC and the six methods that use sequence information alone (Table 1) in discriminating validated targets from the rest. To have a reliable estimate, we selected the test data corresponding to miRNAs that have at least three validated targets from mirTarBase and at least 1 predicted target from each of the predictors. Based on the results, we then picked two overall best sequence-based methods. The scores from those two methods were then integrated along with logFC into the proposed targetScore Equation (3). Next, we modified our testing set by including miRNAs that have at least one target predicted from each of the two best-performing sequence-based methods and at least 10 validated targets from mirTarBase. Finally, we evaluated the performances of logFC and sequence-based methods alone in comparison with Expmicro (Liu *et al.*, 2010a) and the proposed targetScore.

2.7.3 Systematic evaluation For each comparison, the sensitivity and specificity of the methods were systematically assessed using receiver operating characteristic (ROC) curve and summarized by the area under the curve (AUC) (Fig. 2). For a given score cutoff, the true- and false-positive rates (TPR and FPR) are estimated as the respective ratios of TPR = TP/P and FPR = FP/N, where TP and FP are the numbers of true and false positives, and P and N are the total numbers of positive and negative miRNA targets *within the test data*. Predicted targets outside of the test data were not counted. ROC is plotted by iteratively evaluating TPR (y-axis) and FPR (x-axis) while relaxing the scoring cutoff. In addition, we constructed precision-recall (PR) curve (PRC) and assessed the precision TP/(TP + FP) of each method at the same recall (or TPR). Comparing to ROC, PR is more informative in evaluating the performance at the top predictions. Both ROC and PR statistics were obtained using *ROCR* package (Sing *et al.*, 2005).

2.7.4 Evaluation using proteomic data To further evaluate the performance of each method, we used the available protein output data

Table 1. Comparison between miRNA target prediction methods

Method	Seed match	Energy	Conservation	Context score[a]	SVMicro score	P_{CT}[b]	logFC[c]	References
TargetScanPCT	✓		✓			✓		Friedman et al. (2009)
TargetScanCS	✓		✓	✓				Garcia *et al.* (2011)
PicTar2	✓	✓	✓					Anders et al. (2011)
miRanda		✓	✓					Enright et al. (2004)
PITA	✓	✓						Kertesz *et al.* (2007)
SVMicro					✓			Liu *et al.* (2010b)
Expmicro					✓		✓	Liu *et al.* (2010a)
TargetScore				✓		✓	✓	Proposed

Note: Check marks indicate the type of information used by each method.
[a]Context score (CS) is a sequence-based score for individual target sites calculated by targetScan (Grimson *et al.*, 2007; Garcia *et al.*, 2011).
[b]P_{CT} is the probability of conserved targeting for individual target sites and also available from targetScan Web site (Friedman *et al.*, 2009);
[c]logFC: logarithmic fold-change due to miRNA transfection.

following overexpression of hsa-miR-1, 124 and 181 (Baek *et al.*, 2008; Liu *et al.*, 2010a). Despite higher cost, protein abundance is a more accurate indicator of miRNA regulation than mRNA level, as miRNA is known to cause not only mRNA degradation but also translational repression. Presumably, a better prediction algorithm should rank the top targets with larger negative protein fold-changes. Thus, we selected the top 200 target predictions from selected prediction methods and plotted the cumulative sum of the protein log2 fold-change as a function of their rankings (Liu *et al.*, 2010a). At the same ranking, a superior method is expected to reach greater protein downregulation, indicated by a steeper curve (Fig. 3).

2.7.5 GO enrichment analysis As an additional metric previously used by Huang *et al.* (2007), we examined whether the miRNA targets predicted by each method are biologically meaningful via Gene ontology (GO) enrichment analysis. Specifically, GO terms in biological processes (BP) (GO-BP) were downloaded using `getBM` function from R package *biomaRt* (Durinck *et al.*, 2009), where GO terms with fewer than five genes or with evidence codes equal to Inferred from Electronic Annotation (IEA), Non-traceable Author Statement (NAS) or No biological Data available (ND) were discarded, giving 1717 GO-BP terms and 10 222 unique genes. Based on ROC analysis (Section 2.7.3), we chose for each method a specific scoring cutoff that resulted in FPR <0.3 and used such cutoff to select genes with equal or higher scores. The resulting list of Ensembl gene IDs for each method was then subjected to hypergeometric enrichment test for each GO-BP term using R built-in function `phyper`. The corrected *P*-values or false discovery rate (FDR) with R function `p.adjust` was converted to enrichment scores as $-\log_{10}$(FDR). Cumulative density plot was constructed as a function of the enrichment scores (Fig. 4B).

3 RESULTS

3.1 Constructing sequence-based predictors

Among the 113 overexpressed miRNAs, 11 of them have at least 3 validated and 1 predicted target from the 6 sequence-based methods (Table 1). Using these test data, we established that the best-overall sequence-based methods are targetScanCS and targetScanPCT in terms of the AUCs of ROC and PR (Supplementary Fig. S1). Accordingly, we constructed the proposed targetScore by integrating logFC and sequence scores from targetScanCS and targetScanPCT into Equation (3). Modifying the test cases by including all of the overexpressed miRNAs that have at least one predicted targets in targetScanCS, targetScanPCT and SVMicro (required as input to Expmicro), we obtained 35 miRNAs for subsequent comparisons.

3.2 TargetScore

Figure 1 illustrates the intuition behind targetScore. The top 10 interactions in terms of logFC, targetScanCS or PCT from the real data were plotted for the validated and unvalidated targets. Rectangle size corresponds to the magnitude of the scores rescaled to [0, ± 1]. Intuitively, we observe higher targetScores as the three features scores (logFC, targetScanCS and PCT) become more negative. In particular, the superior performance of targetScore is more attributed to the logFC feature than the other two *in silico* predictors; this is consistent with our hypothesis that miRNA-overexpression data is valuable in miRNA target prediction. Among the validated targets (left portion of the plot), for instance, the top 10 logFC attribute to higher targetScore than the top 10 scores from targetScanCS/PCT. On the other hand, targetScanCS and PCT complement logFC, as some (un)validated targets have (high) low logFC but (low) high targetScanCS/PCT scores. Thus, targetScore is a more robust predictor by integrating the three scores.

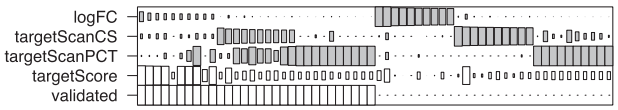


Fig. 1. Hinton plot (Rumelhart and Hintont, 1986) of target feature scores. Top 10 interactions in terms of logFC, targetScanCS or PCT for validated and unvalidated targets are displayed. Gray/white indicates +/− sign. Rectangle size is proportional to the feature size rescaled to [0, ±1]

3.3 Comparisons of target prediction methods

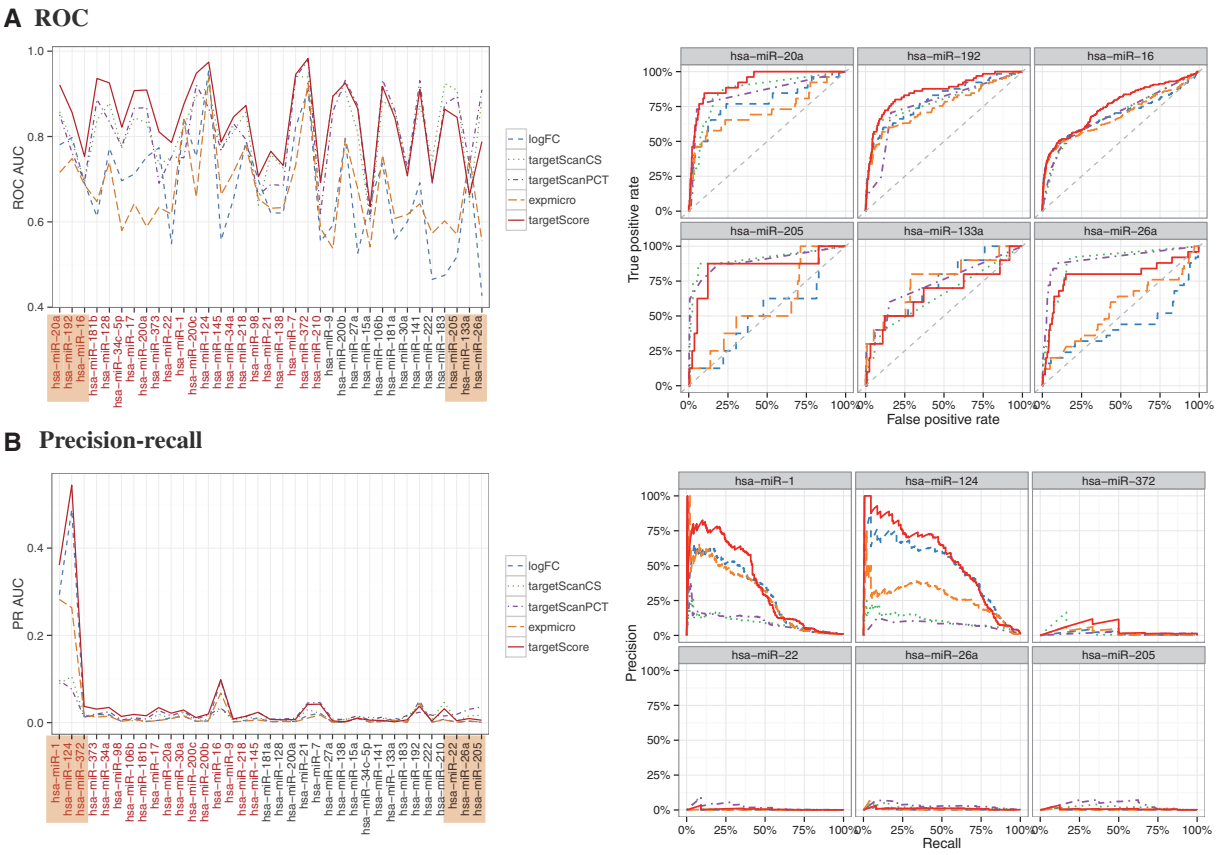
Figure 2 and Supplementary Figures S2 and S3 show that our targetScore method outperforms other existing methods. Specifically, targetScore achieves significantly higher AUC of both ROC and PR than those from other methods (except for the PR-AUC from targetScanPCT) ($P<0.01$; one-sided Wilcoxon signed-rank test; Fig. 2C). Additionally, targetScore dominates the largest number of tests among the 35 miRNAs: it has the best ROC and PRC for 22 and 17 miRNAs, respectively (Fig. 2D). It is also worth noting that our method had a large leading margin ahead of the second best method in several tests (e.g. hsa-miR-20a/192/16 in ROC and hsa-miR-1/124 in PRC; right panels from Fig. 2A and B). Thus, targetScore is able to outperform the best individual method by integrating the complementary information generated from each method.

3.4 Protein downregulation

Because of the potential translational repression by miRNA, negative logFC of protein outputs due to transfection of hsa-miR-1/124/181 is a more direct indicator of the true miRNA targets than mRNA expression. Targets ranked by targetScore exhibit comparable negative cumulative logFC comparing with the best among other methods (Fig. 3). Thus, the protein outputs are mostly consistent with the above ROC/PR statistics, indicating that the confidence target predictions from targetScore at the mRNA level translate well to the miRNA effects at the protein level.

3.5 GO enrichment of predicted targets

The confidence targets filtered by targetScore (at FPR <0.3) are enriched for more significant GO terms at FDR <0.05 than other methods in 19 of the 35 tests (Fig. 4A and Supplementary Table S2). Overall, targetScore recovers significantly higher number of GO terms than other methods ($P<$ 1e-6) except for Expmicro ($P<0.14$; one-sided Wilcoxon signed-rank test; Fig. 4A inset table). Moreover, GO terms identified from targetScore targets (targetScore-GO) exhibit significantly higher cumulative enrichment scores than those from Expmicro in nine tests and those from logFC, targetScanCS/PCT in majority of the tests as indicated by the gray boxes in Figure 4B ($P<0.05$, one-sided paired Kolmogorov-Smirnov (KS) test; see Supplementary Fig. S4 for details). In contrast, Expmicro-GO is significantly more enriched than targetScore-GO in only four



C Paired test on AUC of TargetScore vs other method

	logFC	targetScanCS	targetScanPCT	expmicro
ROC	1.3e-09	8.7e-04	1.3e-03	5.5e-10
PRC	2.5e-07	3.3e-03	1.8e-01	1.9e-08

D Tests for which method X has the best performance

	logFC	targetScanCS	targetScanPCT	expmicro	targetScore
ROC	1	5	6	1	22
PRC	2	3	13	0	17

Fig. 2. Evaluation of target prediction methods on 35 miRNAs. Based on the experimentally validated targets (Hsu *et al.*, 2011), ROC (**A**) and PR (**B**) curves were constructed using scores from logFC, targetScanCS, targetScanPCT, expmicro and targetScore. The left panels display the AUC of ROC/PR for each method across the 35 miRNAs, which were ordered by the decreasing differences between targetScore-AUC and AUC from the best-performing competitor. The miRNAs in red indicate that targetScore is best among all. The specific ROC/PR curves for the top and bottom three miRNAs as highlighted in the orange boxes were illustrated on the right panels. (**C**) *P*-values from one-sided Wilcoxon signed-rank test by comparing the AUCs for ROC/PR from targetScore with the AUCs from other methods. (**D**) The number of miRNAs for which a particular prediction method has the best performance in ROC/PR

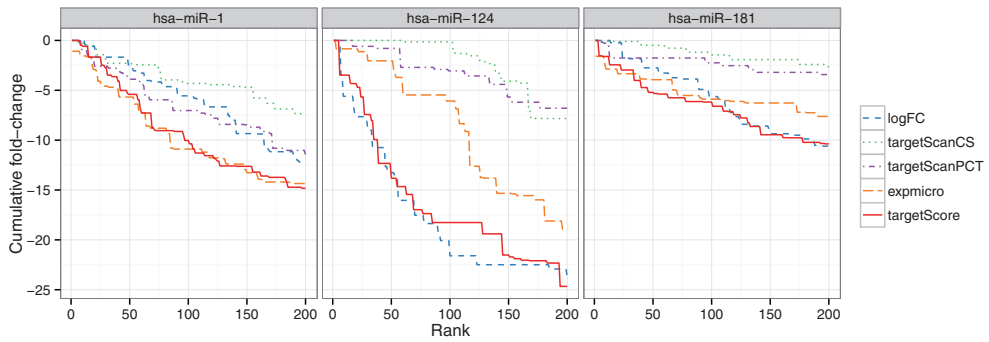


Fig. 3. Cumulative sum of protein downregulation as a function of the top 200 rankings of target predictions for hsa-miR-1, 124 and 181. For each comparison method, cumulative sum of log2 fold-change of protein outputs measured in miRNA transfection experiments (Baek *et al.*, 2008) is plotted as a function of the top 200 rankings. At the same rank index, a superior method is expected to reach greater cumulative sum of protein down-fold, indicated by the steeper curve

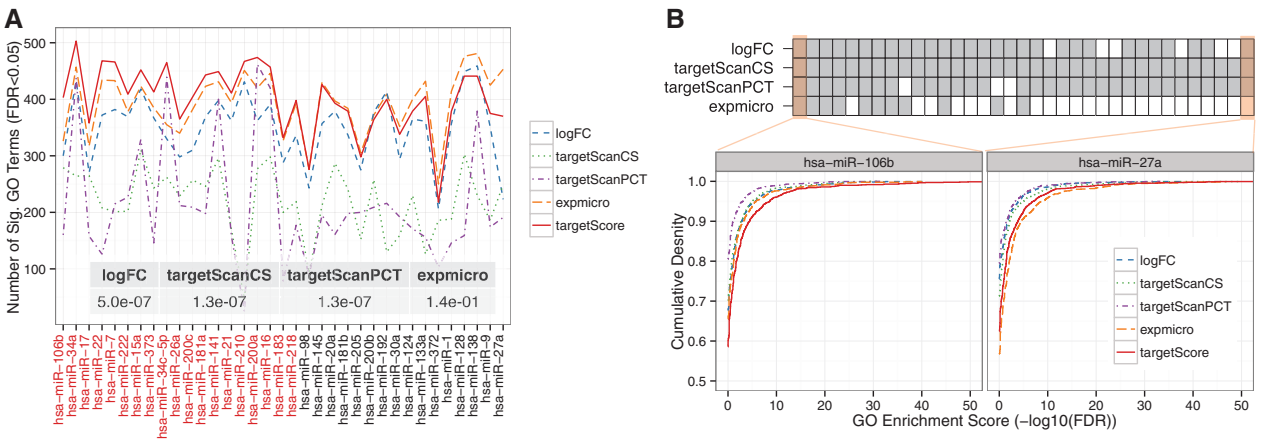


Fig. 4. Comparison of GO enrichments. For each comparison method, genes with equal or higher score than a cutoff in which FPR < 0.3 were subjected to GO enrichment tests. (**A**) At FDR < 0.05, we counted the number of significant GO terms for each method across the 35 miRNAs. The miRNAs in red indicate that the number of GO terms from targetScore is the highest among all. (Inset) the *P*-values indicate whether the corresponding numbers of GO terms from targetScore are significantly higher than those from other methods based on one-sided Wilcoxon signed-rank test. (**B**) The (white) gray boxes indicate that the cumulative GO enrichment scores from targetScore are (not) significantly higher than those from competitors based on one-sided paired KS-test at $P < 0.05$. The columns are in the same order as the miRNAs in panel A. The cumulative density function (CDF) plot for the first and last column (i.e. hsa-miR-106b and 27a) was illustrated in the bottom panel. At the same density (*y*-axis), the enrichment scores (*x*-axis) are the highest if the corresponding CDF is on the right side of other CDFs. All of the CDF plots and detailed *P*-values are presented in Supplementary Figure S4

tests. Thus, targetScore-predicted targets are more likely enriched for meaningful BP.

3.6 Oncomir and oncogene targets network

Encouraged by the aforementioned results, we analyzed the target relationships between oncomirs (Croce, 2009; Spizzo *et al.*, 2009) and oncogenes downloaded from COSMIC (Forbes *et al.*, 2011) (Supplementary Table S3). Specifically, we identified 207 oncomir–oncogene interactions, involving 26 oncomirs (yellow node) and 113 oncogenes (white node) (Fig. 5). Of these interactions, 166 are validated (red solid edges) and the remaining 41 are predicted with targetScore ≥ 0.6 (blue dash edges). The scoring cutoff was chosen based on the differential targetScore distributions for the validated and non-validated targets (Supplementary Fig. S5). The constructed oncomir–oncogene network is highly connected. In particular, the top three oncomirs with the highest out-degree, namely, hsa-miR-155, 16 and 373 target 27, 19 and 18 oncogenes, respectively. Interestingly, 12 and all of the 19 targets of hsa-miR-155 and 16 are, respectively, validated; in contrast, 17 of the 19 targets of hsa-miR-373 are predicted with high confidence, yet call for experimental validation.

4 DISCUSSION

Most of recently developed miRNA target prediction methods exhibit a paradigm shift from rule-based binary classification to a more context-dependent, quantitative and probabilistic approach (Huang *et al.*, 2007; Le and Bar-Joseph, 2011). The momentum of this shift is largely facilitated by the increasing amount of expression profiling data of mRNAs (and miRNAs) across various experimental conditions. However, most of these expression-based methods are based on correlation and thus require a large set of expression profiles of mRNAs and miRNAs across various tissues, cell lines or patients, which limit their applications to only a general survey of the robust miRNA targets rather than condition-specific miRNA targets. On the other hand, expression profiling following specific miRNA transfection (knockin) provides the most direct clue to identify *in vivo* functional miRNA targets. However, expression data measured by either microarrays or RNA-seq are noisy, and it is known that changes in expression can be caused by indirect regulatory effect by miRNAs (Khan *et al.*, 2009), which is not easily distinguishable from direct effects without the aid of sequence-based information. To our knowledge, few programs are specifically developed for transfection-based miRNA target prediction. Among them, Sylamer is the earliest work developed to identify enriched k-mer motifs, which are the seed regions among the top ranked targets based on *P*-values obtained from *t*-test (van Dongen *et al.*, 2008). Thus, the method does not model the distribution of fold-changes or sequence features. Instead, Sylamer is mainly designed to visually inspect the enrichment of k-mer patterns rather than predicting specific targets.

In this article, we introduce targetScore, a Bayesian probabilistic scoring method taking into account both the fold-change due to miRNA overexpression and sequence-based information. The proposed method is similar to Expmicro (Liu *et al.*, 2010a). However, Expmicro requires training data to calculate the sequence-based scores using SVMicro and used only a two-component GMM to model the fold-changes. In contrast, targetScore models the distributions of multiple sets of precomputed or user-supplied sequence-based scores and fold-changes using respective two- and three-component VB-GMM.

We complied (to our knowledge) the largest set of overexpression data compendium and demonstrated the utilities of targetScore in extensive tests. TargetScore achieved superior ROC and PR performances (Fig. 2 and Supplementary Figs S2

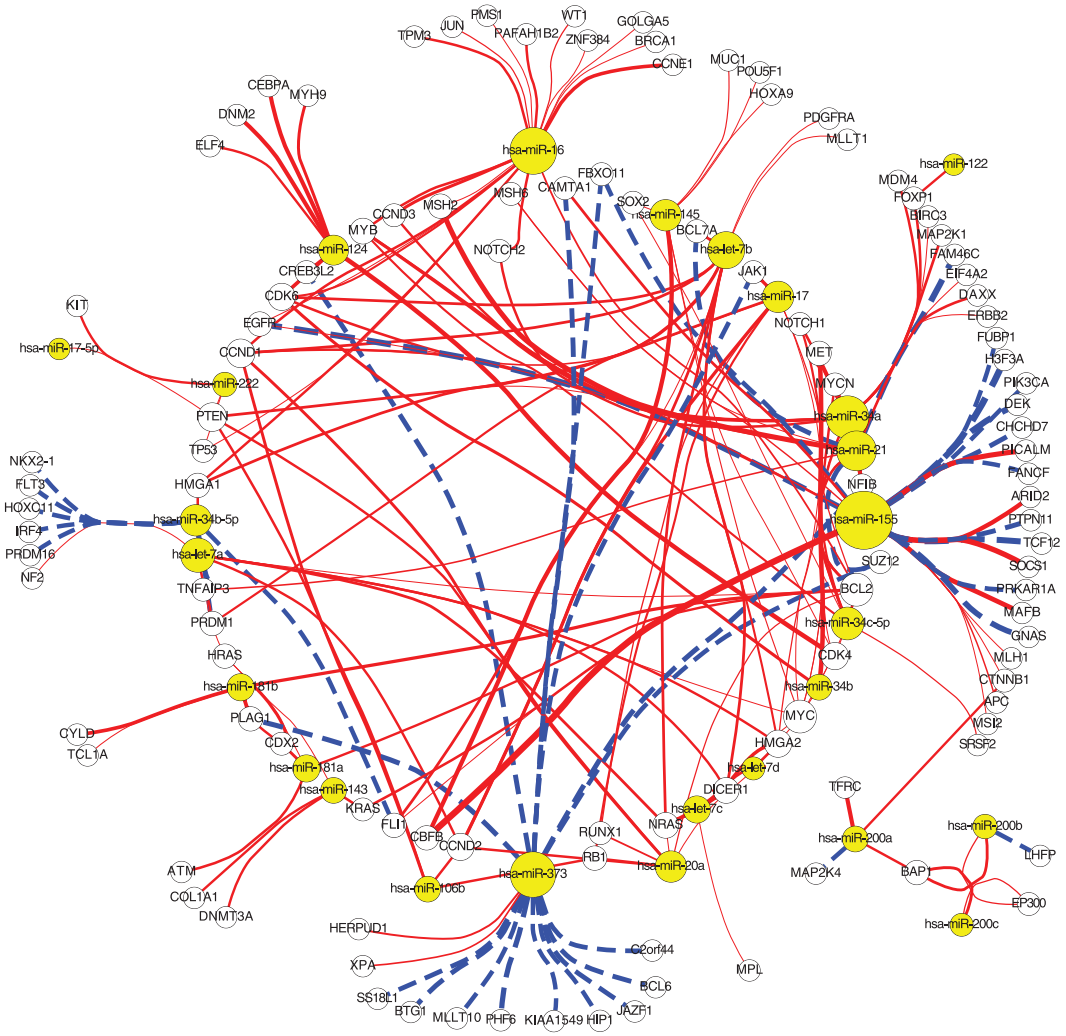


Fig. 5. Oncomir–oncogene network. The network drawn by Cytoscape 3 (Shannon *et al*., 2003) comprises 207 oncomir–oncogene interactions, where 166 (41) are validated (predicted with targetScore $\geq$ 0.6), involving 26 oncomirs (yellow) and 113 oncogenes (white). The red solid and blue dash edges are the validated and predicted interactions, respectively. The size of the node is proportional to the connection degree. Edge widths are proportional to the targetScore

and S3). The cumulative protein outputs for miR-1/124/181 as a function of the top 200 rankings from targetScore reveals the (second) deepest protein downregulation comparing with the top targets from other methods (Fig. 3). Thus, the relative overall trends at the protein level are consistent with the predictions at the mRNA level, where targetScore dominates most of the tests. Moreover, targetScore targets are more enriched for meaningful BP when compared with targets predicted by other methods (Fig. 4 and Supplementary Fig. S4). Using targetScore, we constructed an oncomir–oncogene regulatory network (Fig. 5) and hypothesized that our predicted oncomir–oncogene interactions may provide further insights into cancer network.

The success of our method underlines the importance of integrating (preferably independent) informative predictors into a unified framework. On the other hand, the absolute PR performances are generally poor among all tested methods perhaps because of the limited number of validated targets and/or the limitations of the methods. Although targetScore compares favorably with the published methods, there are a few issues that remain to be addressed in future work. Several studies have shown that the expression fold-change also depends on the initial abundance and the natural decay or turnover rate of the corresponding mRNA species (Larsson *et al*., 2010). Presumably, a future target prediction algorithm will benefit from the use of this information in a cell-specific context. Additionally, targetScore assumes that the mRNA targets are independent of each other. However, Arvey *et al*. (2010) have shown that miRNAs that have a higher number of available target transcripts will downregulate each individual target gene to a lesser extent than those with a lower number of targets. Although the context score from targetScanCS has incorporated the *static* target abundance information as the total number of target sites that can be recognized by the same miRNA, it would be more realistic to consider the underlying expression level of the co-targeted mRNA to infer the actual *available* target sites.

ACKNOWLEDGEMENTS

The authors thank Quaid Morris for useful discussion, and Yufei Huang and Hui Liu for the helps on running Expmicro. This paper is dedicated to the memory of SZ.

Funding: Natural Sciences and Engineering Research Council (NSERC) Canada Graduate Scholarship (to Y.L.) and Ontario Research Fund-Global Leader (Round 2) and an NSERC grant (to Z.Z.).

Conflict of Interest: None declared.

REFERENCES

Alexiou,P. *et al.* (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.

Anders,G. *et al.* (2011) doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **40**, D180–D186.

Arvey,A. *et al.* (2010) Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.*, **6**, 1–7.

Baek,D. *et al.* (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, New York.

Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, **10**, 704–714.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

Durinck,S. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.

Enright,A.J. *et al.* (2004) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1–R14.

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Garcia,D.M. *et al.* (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.

Grimson,A. *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hsu,S.-D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.

Huang,J.C. *et al.* (2007) Bayesian inference of microRNA targets from sequence and expression data. *J. Comput. Biol.*, **14**, 550–563.

Kertesz,M. *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Khan,A.A. *et al.* (2009) Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat. Biotechnol.*, **27**, 549–555.

Krek,A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Larsson,E. *et al.* (2010) mRNA turnover rate limits siRNA and microRNA efficacy. *Mol. Syst. Biol.*, **6**, 1–9.

Le,H.S. and Bar-Joseph,Z. (2011) Inferring interaction networks using the IBP applied to microRNA target prediction. In: Shawe-Taylor,J. *et al.* (eds) *Advances in Neural Information Processing Systems*. Vol. 24, Sierra Nevada, Spain, pp. 235–243.

Lewis,B.P. *et al.* (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Lim,L.P. *et al.* (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Liu,H. *et al.* (2010a) A Bayesian approach for identifying miRNA targets by combining sequence prediction and gene expression profiling. *BMC Genomics*, **11** (**Suppl. 3**), S12.

Liu,H. *et al.* (2010b) Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, **11**, 476.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Rumelhart,D. and Hintont,G. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Selbach,M. *et al.* (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Spizzo,R. *et al.* (2009) SnapShot: microRNAs in cancer. *Cell*, **137**, 586–586.e1.

Sumazin,P. *et al.* (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, **147**, 370–381.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

van Dongen,S. *et al.* (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, **5**, 1023–1025.

Vergoulis,T. *et al.* (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.*, **40**, D222–D229.

Xiao,F. *et al.* (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.