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

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

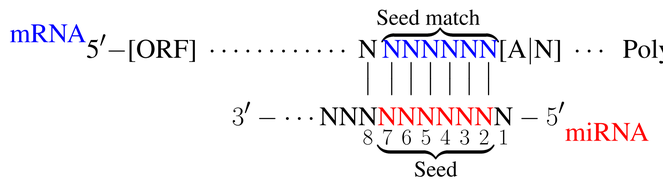
[1]*Department of Computer Science,* [2]*The Donnelly Centre,* [3]*Banting and Best Department of Medical Research,* [4]*Genetics and Genome Biology, SickKids Research Institute, University of Toronto*

Abstract

Systematic identification of miRNA targets remains a challenge. 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. 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 GEO 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 down-regulation and are more significantly enriched for Gene Ontology terms. Using targetScore, we explored oncomir-oncogenes network and predicted several potential cancer-related miRNA-mRNA interactions. *TargetScore* is available at Bioconductor.

Introduction:

MicroRNAs (miRNAs) repress protein production in animal species by forming Watson-Crick (WC) base-pairing to the $3'$ UTR regions of the target mRNAs (Friedman et al., 2009). The binding primarily occurs at the 2-7 nucleotide (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 are implicated in many developmental and pathogenic processes.



Current sequence-based algorithms achieve less than 50% specificity and having poor agreement among them (Alexiou et al., 2009). Overexpression of miRNA coupled with expression profiling of mRNA by either microarray or RNA-seq has proved to be a promising approach (Lim et al., 2005; Arvey et al., 2010). To take advantage of the increasingly available data of such kind, we developed a probabilistic approach termed *TargetScore* to infer miRNA targets by integrating expression change due to miRNA overexpression and sequence information such as context score and other orthogonal features such as conservation into a probabilistic score.

Methods:

Bayesian Gaussian 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). The latent variables $\mathbf{z}$ are sampled at probabilities $\boldsymbol{\pi}$ (mixing coefficient), that follow a Dirichlet prior $Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}_0)$ with hyperparameters $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{0,1}, \ldots, \boldsymbol{\alpha}_{0,K})$. To account for the relative frequency of targets and non-targets for any miRNA, we set the $\boldsymbol{\alpha}_{0,1}$ (associated with the target component) to aN and other $\boldsymbol{\alpha}_{0,k} = (1-a) \times N/(K-1)$, where $a = 0.01$ (by default). Assuming $\mathbf{x}$ follows a Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}$ (precision matrix) is the inverse covariance matrix, $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ together follow a Gaussian-Wishart prior $\prod_k^K \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda})^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0)$, where the hyperparameters $\{\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0\} = \{\hat{\boldsymbol{\mu}}, 1, \mathbf{I}_{D\times D}, D+1\}$.

Variational Bayesian Expectation Maximization

Let $\boldsymbol{\theta} = \{\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$. 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):

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

where $q(\boldsymbol{\theta})$ is a proposed distribution for $p(\boldsymbol{\theta}|\mathbf{x})$, 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 *w.r.t* other variables, $\mathbb{E}_{i\neq j}[\ln p(\mathbf{x}, \boldsymbol{\theta})]$. In particular, we define $q(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{z})q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The expectations for the three terms (at log scale), namely $\ln q^*(\mathbf{z}), \ln q^*(\boldsymbol{\pi}), \ln q^*(\boldsymbol{\mu})$, have the same forms as the initial distributions due to the conjugacy of the priors. However, they require evaluation of the parameters $\{\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, which in turn all depend on the expectations of $\mathbf{z}$ or the posterior of interest:

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

(Methods continued)

where $\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k,\boldsymbol{\Lambda}_k}[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)]$. The inter-dependence between the expectations and model parameters calls for EM framework, namely VB-EM. We first initialize the model parameters based on priors and randomly sample K data points $\boldsymbol{\mu}$. 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 iterates until $\mathcal{L}(q)$ converges. Please refer to Bishop (2006) p485 for more details.

TargetScore

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

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

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

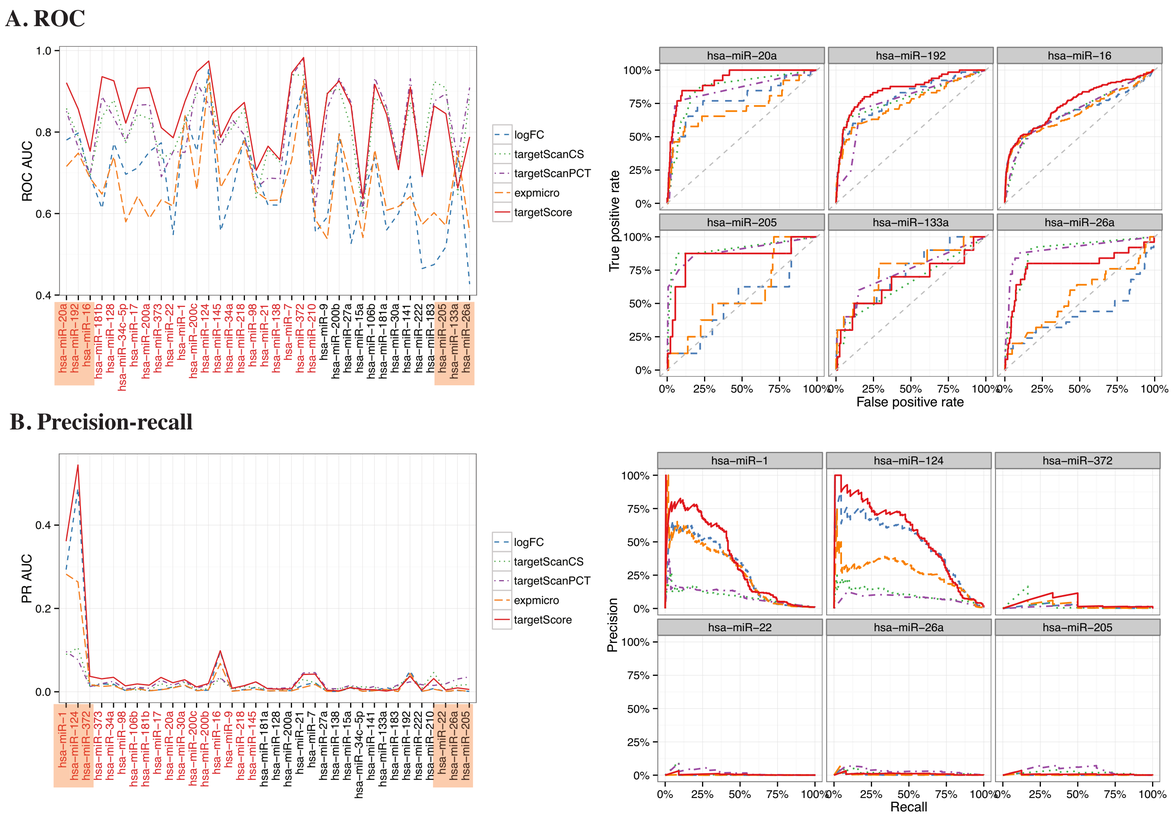
Test Data

We collected miRNA-overexpression data corresponding to 84 GEO series, 6 platforms, 77 human cells or tissues, and 113 distinct miRNAs. Validated miRNA-target interactions from MirTarBase 3.5 were used, which contains 3565 validated interactions between 432 human miRNAs and 1959 target genes.

Discussion and Future work:

We introduce TargetScore, a Bayesian probabilistic scoring method taking into account both the fold-change (FC) due to miRNA overexpresion and sequence information. The success of our method underlines the importance of integrating independent informative predictors into a unified framework. A few issues remain to be addressed in future work. Several studies have shown that expression changes also depends on the initial abundance and the natural decay of the corresponding mRNA species (Larsson et al., 2010). Presumably, an algorithm will benefit from the use of this information in a cell-specific context.TargetScore assumes that the mRNA targets are independent of each other. Arvey et al. (2010) have shown that miRNAs that have a higher number of available target transcripts will down-regulate each individual target gene to a lesser extent than those with a lower number of targets. Thus, it would be more realistic to consider the expressed target sites.

Results:



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.1. Evaluation on validated targets from 35 miRNAs. (**A**) and precision-recall (**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 indicates that TargetScore is best among all. The specific ROC/PR curves for the top and bottom 3 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.

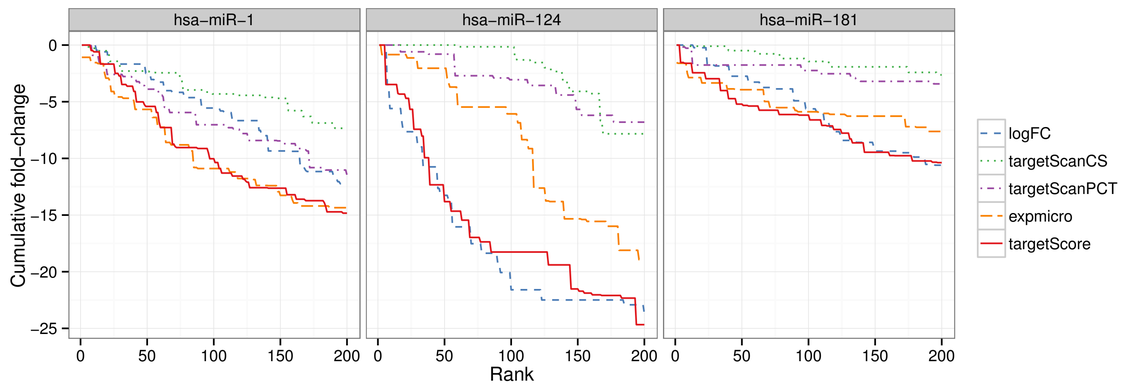


Fig.2. Cumulative sum of protein down-regulation. Cumulative sum of protein down-regulation as a function of the top 200 rankings of target predictions for hsa-miR-1, 124, and 181. For each comparison method, cumulative sum of log2 fold-change of protein outputs measured in miRNA transfection experiments (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.

Results (continued):

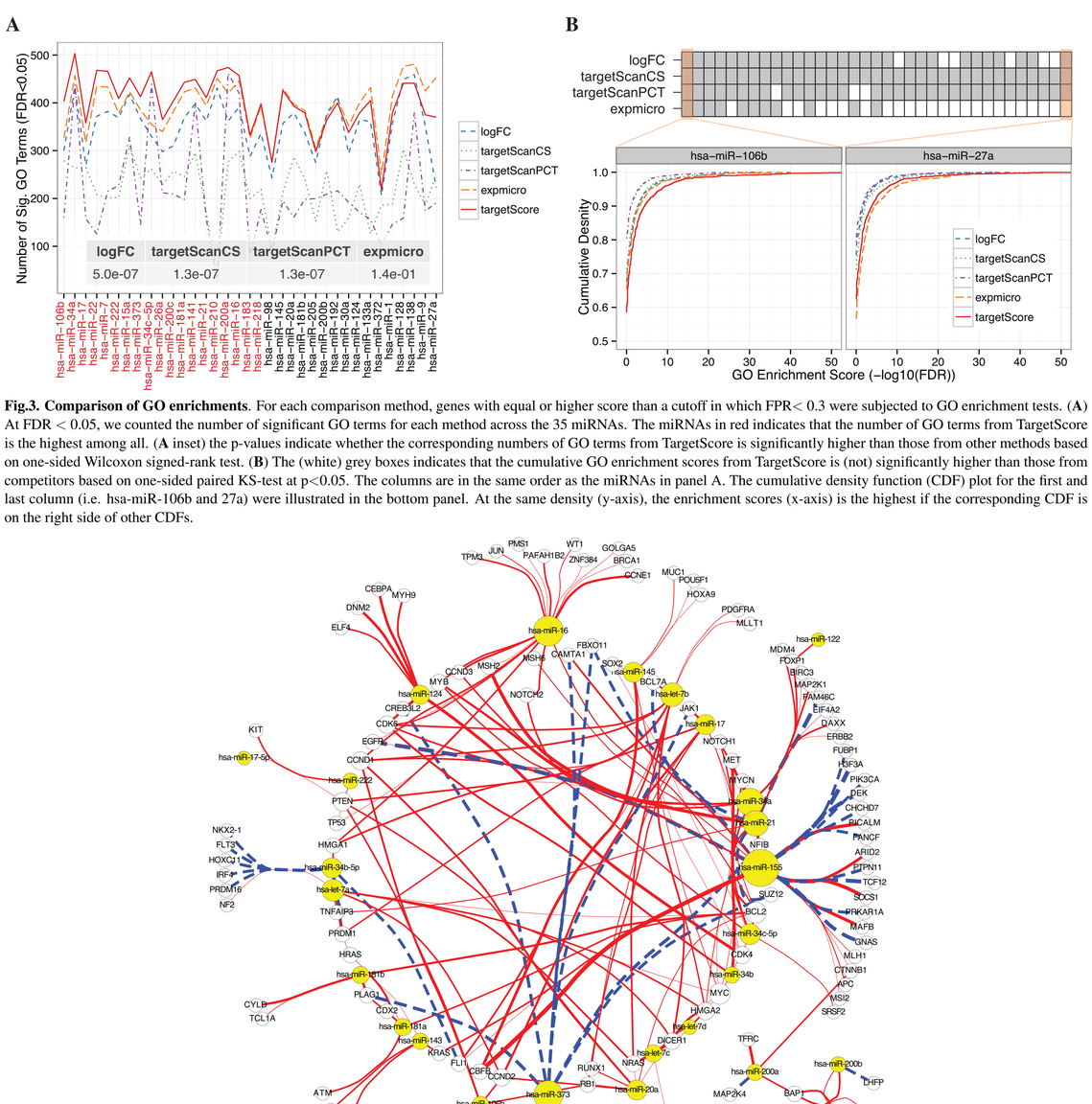


Fig.3. Comparison of GO enrichments. For each comparison method, genes with equal or higher score than a cutoff in which FPR$<$ 0.3 were subjected to GO enrichment tests. (**A**) At FDR $<$ 0.05, we counted the number of significant GO terms for each method across the 35 miRNAs. The miRNAs in red indicates that the number of GO terms from TargetScore is the highest among all. (**A** inset) the p-values indicate whether the corresponding numbers of GO terms from TargetScore is significantly higher than those from other methods based on one-sided Wilcoxon signed-rank test. (**B**) The (white) grey boxes indicates that the cumulative GO enrichment scores from TargetScore is (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) were illustrated in the bottom panel. At the same density (y-axis), the enrichment scores (x-axis) is the highest if the corresponding CDF is on the right side of other CDFs.

Fig.4. Oncomir-oncogene network based on confidence targetScore. The network drawn by Cytoscape 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 proportion to the connection degree. Edge widths are prop. to the targetScore.

Reference:

Alexiou, P., et al. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**(23), 3049-3055.

Arvey, A., et al. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular systems biology*, **6**, 1-7.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer, Information Science and Statistics. NY, USA.

Friedman, R.C., et al. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, **19**(1), 92-105.

Larsson, E., et al. (2010). mRNA turnover rate limits siRNA and microRNA efficacy. *Molecular systems biology*, **6**, 1-9.

Lewis, B. P., et al. (2003). Prediction of mammalian microRNA targets. *Cell*, **115**(7), 787-798.

Lim, L. P., et al. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**(7027), 769-773.

Li, Y., Goldenberg, A., Wong, K.-C., & Zhang, Z. A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information. *Bioinformatics*, first published online October 17, 2013.