

Bayesian Learning of MicroRNA Targets from Sequence and Expression Data

Jim C. Huang¹, Quaid D. Morris², Brendan J. Frey^{1,2}

¹ Probabilistic and Statistical Inference Group, University of Toronto
10 King's College Road, Toronto, ON, M5S 3G4, Canada

² Banting and Best Department of Medical Research, University of Toronto
160 College St, Toronto, ON, M5S 3E1, Canada
jim@psi.toronto.edu

Abstract. MicroRNAs (miRNAs) regulate a large proportion of mammalian genes by hybridizing to targeted messenger RNAs (mRNAs) and down-regulating their translation into protein. Although much work has been done in the genome-wide computational prediction of miRNA genes and their target mRNAs, an open question is how to efficiently obtain functional miRNA targets from a large number of candidate miRNA targets predicted by existing computational algorithms. In this paper, we propose a novel Bayesian model and learning algorithm, GenMiR++ (**Generative model for miRNA regulation**), that accounts for patterns of gene expression using miRNA expression data and a set of candidate miRNA targets. A set of high-confidence functional miRNA targets are then obtained from the data using a Bayesian learning algorithm. Our model scores 467 high-confidence miRNA targets out of 1,770 targets obtained from TargetScanS in mouse at a false detection rate of 2.5%: several confirmed miRNA targets appear in our high-confidence set, such as the interactions between miR-92 and the signal transduction gene MAP2K4, as well as the relationship between miR-16 and BCL2, an anti-apoptotic gene which has been implicated in chronic lymphocytic leukemia. We present results on the robustness of our model showing that our learning algorithm is not sensitive to various perturbations of the data. Our high-confidence targets represent a significant increase in the number of miRNA targets and represent a starting point for a global understanding of gene regulation.

1 Introduction

One of the main goals in genomics is to understand how genes are regulated, both transcriptionally and post-transcriptionally. A landmark advancement in understanding the full scope of post-transcriptional gene regulation is the discovery of microRNAs. These short, 23 nt-long RNAs suppress protein synthesis from specific transcripts that contain anti-sense target sequences to which the miRNAs can hybridize with complete or partial complementarity [1, 5, 15]. MicroRNAs have emerged as an important component in the regulatory circuitry of the cell, with current estimates suggesting that the human genome contains

but this has the problem of becoming intractable when the regulatory action of many miRNAs across multiple tissues must be taken into account.

While computational sequence analysis methods for finding targets and expression profiling methods have their own respective limitations, we can combine the two methods in order to learn miRNA targets from both sequence and expression data. Given the thousands of miRNA targets being output by target-finding programs [14, 16] and given the ability to profile the expression of thousands of mRNAs and miRNAs using microarray technology [12, 20], we motivate GenMiR++ (**G**enerative model for **mi**RNA regulation), a Bayesian model for miRNA regulation and learning algorithm which will allow us to obtain a set of functional miRNA targets from both sequence and expression data. The pipeline for finding functional miRNA targets is shown in Fig. 1: a set of candidate miRNA targets is first generated using a sequence-based target-finding program. Given this set of candidate targets, our Bayesian model for miRNA regulation accounts for mRNA expression using miRNA expression data while also modeling the combinatorial nature of miRNA regulation. We then apply a Bayesian learning algorithm to the data in order to find a set of functional miRNA targets.

In this paper, motivated by evidence which suggests that miRNAs can down-regulate protein expression by causing mRNA target degradation [4, 9, 18], we will formulate a Bayesian model for miRNA regulation. Under this model, the expression of a targeted mRNA transcript can be explained through the regulatory action of multiple miRNAs. We show that GenMiR++ allows us to accurately identify miRNA targets from both sequence and expression data. We also show that we can recover a significant number of experimentally verified targets, many of which provide insight into miRNA regulation. We will show that GenMiR++ also achieves significantly improved performance over our previous model and learning algorithm, GenMiR [10, 11]. Finally, we will present results on the robustness of GenMiR++ to sub-sampling of the data and to adding fake targets to our list of candidate miRNA targets.

2 Exploring miRNA targets using microarray data

In this section, we will motivate the use of both mRNA and miRNA expression data to detect miRNA-target regulatory relationships. Recently, it has been shown that mRNA expression data can be used to detect direct miRNA-mRNA target interactions [18] by looking for transcripts that appear to be down-regulated after over-expression of a tissue-specific miRNA. Given the availability of data sets that profile the expression of miRNAs as well as mRNAs across many tissues, we expected to find miRNA-target relationships in two separate expression data sets profiling both miRNAs and mRNAs. To explore putative relationships between mRNAs and miRNAs, we used mRNA expression data from [25] profiling 41,699 mouse mRNAs in tandem with data profiling 78 mouse miRNAs [3] across 17 mouse tissues. Both data sets consisted of arcsinh-normalized intensity values in the same range with negative miRNA intensities thresholded to 0. We used a set of human miRNA targets predicted by the target-finding

program TargetScanS [16, 27]: these consisted of a total of 12,839 target predictions in human genes. These were identified based on conservation in 3'UTR regions across five mammalian species (human, mouse, rat, dog and chicken) in addition to miRNA-target sequence complementarity. After mapping these predictions to the mouse mRNAs and miRNAs in the two expression data sets using Ensembl and BLAT [13], we were left with 1,770 target predictions involving 788 unique mRNA transcripts and 22 unique miRNAs. Given this information, we looked for examples of miRNA target down-regulation in which predicted target mRNA transcript expression was low in a given tissue and the predicted targeting miRNA was highly expressed in that same tissue.

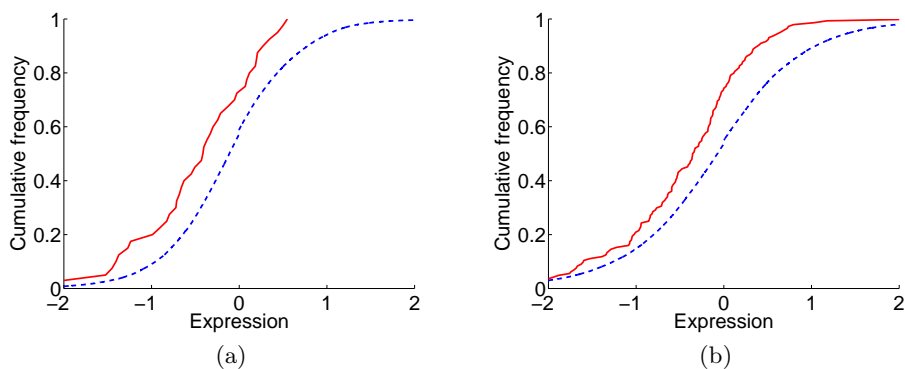


Fig. 2. Effect of miRNA negative regulation on mRNA transcript expression: shown are cumulative distributions for (a) Expression in embryonic tissue for mRNA transcripts targeted by miR-205 (b) Expression in spleen tissue for mRNA transcripts targeted by miR-16. A shift in the curve corresponds to down-regulation of genes targeted by miRNAs. Targets of miR-205 and miR-16 (solid) show a negative shift in expression with respect to the background distribution (dashed) in tissues where miR-205 and miR-16 are highly-expressed ($p < 10^{-7}$ and $p < 0.0015$).

Among miRNAs in the data from [3], miR-16 and miR-205 are two that are highly expressed in spleen and embryonic tissue respectively. The cumulative distribution of expression of the TargetScanS-predicted targets for the miRNAs in these same two tissues is shown in Fig. 2. The plots show that the expression of targeted mRNAs is negatively shifted with respect to the background distribution of expression in these two tissues ($p < 10^{-7}$ and $p < 0.0015$ using a one-tailed Wilcoxon-Mann-Whitney test, Bonferroni-corrected at $\alpha = 0.05/22$ for the 22 miRNAs in our data set). This result suggests that regulatory interactions predicted on the basis of genomic sequence can be observed in microarray data in the form of high miRNA/low targeted transcript expression relationships. While it is feasible to find such relationships for a single miRNA using an expression profiling method [18], to exhaustively test for all regulatory patterns due to multiple miRNAs would require a number of microarray experiments that grows exponentially in the number of miRNAs being tested: this would be

prohibitive for even a modest number of miRNAs and tissues. There would also be additional uncertainty due to miRNAs that are expressed in many tissues. An alternative is to use data which profiles the expression of mRNAs and miRNAs across many tissues and formulate a probabilistic model which links the two using a set of candidate miRNA targets. A sensible model would account for negative shifts in tissue expression for targeted mRNA transcripts given that the corresponding miRNA was also highly expressed in the same tissue. By accounting for the fact that miRNA regulation is combinatorial in nature [5, 14], we will construct such a model which will capture the basic mechanism of miRNA regulation. The model will use a set of candidate miRNA targets and expression data sets profiling both mRNA transcripts and miRNAs to account for examples of down-regulation. We can then learn our model on both sequence and expression data in order to obtain a set of functional miRNA targets.

3 A Bayesian model for miRNA regulation

In this section, we describe our Bayesian model for miRNA regulation. Our model takes into account the down-regulation of target mRNAs subject to the action of multiple miRNAs. Each miRNA can reduce the expression of its targets by some fixed amount: given high expression of one or many miRNAs, the expression of a targeted transcript is negatively shifted with respect to a background expression level which is to be estimated. For any given transcript, the particular miRNAs that can regulate it will be selected using a set of unobserved binary indicator variables. Thus, the problem of finding functional miRNA targets consists of inferring which indicator variables are turned on and which are turned off, given the data.

Consider two separate expression data sets profiling G messenger RNA transcripts and K microRNAs across T tissues. Let indices $g = 1, \dots, G$ and $k = 1, \dots, K$ denote particular messenger RNA transcripts and microRNAs in our data sets. Let $\mathbf{x}_g = (x_{g1} \cdots x_{gT})^\top$ and $\mathbf{z}_k = (z_{k1} \cdots z_{kT})^\top$ be the expression profiles over the T tissues for mRNA transcript g and miRNA k such that x_{gt} is the expression of the g^{th} transcript in the t^{th} tissue and z_{kt} is the expression of the k^{th} miRNA in the same tissue. Now suppose we are given a set of candidate miRNA-target interactions in the form of a binary matrix \mathbf{C} where $c_{gk} = 1$ if miRNA k putatively targets transcript g and $c_{gk} = 0$ otherwise. The matrix \mathbf{C} therefore contains an initial set of candidate target mRNAs for different miRNAs: these are putative miRNA-mRNA regulatory relationships within which we will search for high-confidence cases.

Because of noise in the sequence and expression data as well as the limited accuracy of sequence-based methods for identifying miRNA targets, there is uncertainty as to which miRNA targets are in fact functional. This uncertainty can be represented using a set of unobserved binary random variables indicating which of the candidate miRNA-target interactions are supported by the observed patterns of expression for miRNAs and their putative target mRNAs. We will assign an unobserved random variable s_{gk} to each candidate interaction so that $s_{gk} = 1$ if miRNA k genuinely targets mRNA transcript g . Then, the problem

of finding functional miRNA-target interactions can be formulated in terms of finding a subset of pairs (g, k) so that both $c_{gk} = 1$ and $s_{gk} = 1$. These will correspond to interactions which are supported by the observed expression data and which are *bona fide*.

Having established some notation, we can describe a relationship between the expression of a targeted mRNA transcript, the background of expression and a set of targeting miRNAs in tissue t :

$$E[x_{gt} | \{s_{gk}\}, \{z_{kt}\}, \mathbf{\Lambda}, \mu_t, \gamma_t] = \mu_t - \gamma_t \sum_k \lambda_k s_{gk} z_{kt}, \quad \lambda_k > 0 \quad (1)$$

where λ_k is some positive regulatory weight that determines the relative amount of down-regulation incurred by miRNA k , γ_t is a positive tissue scaling parameter which accounts for differences in hybridization conditions and normalization between the miRNA and mRNA expression data, and μ_t is a background expression parameter. Thus, given the expression of a set of targeting miRNAs, the expression of a targeted transcript is negatively shifted with respect to the background level of expression. This down-regulation can be caused by multiple miRNAs which cooperate in tuning the expression of a targeted transcript. Note that we will not allow miRNAs to directly increase the expression of their target transcripts, in accordance with current evidence about their functions [1, 18].

Given the above, we can present our model for miRNA regulation: we will do so using a probabilistic graphical model in which variables are represented as nodes in a graph and edges represent dependencies between variables. Furthermore, we will adopt a Bayesian modeling framework in which we average over all possible settings of model parameters according to prior probability distributions which encode our *a priori* beliefs about plausible parameter settings. Thus, denote the base targeting probability as $p(s_{gk} = 1 | c_{gk} = 1) = \pi$. Let \mathbf{S} be the set of s_{gk} variables and let $\mathbf{\Lambda}$ be the set of regulatory weights λ_k . Let $\mathbf{\Gamma} = \text{diag}(\gamma_1, \dots, \gamma_T)$ be the $T \times T$ diagonal matrix of tissue-scaling parameters which allow us to account for the fact that the mRNA and miRNA data sets correspond to two experiments with different hybridization conditions and the fact that they were normalized differently from one another. We will assign prior distributions $p(\mathbf{\Lambda})$ and $p(\mathbf{\Gamma})$ for the parameters $\mathbf{\Lambda}, \mathbf{\Gamma}$: the priors will be used to average over all possible settings of $\mathbf{\Lambda}, \mathbf{\Gamma}$. Having defined the above parameters and variables, we can write the probabilities in our model, conditioned on the expression of miRNAs and a set of candidate miRNA targets, as

$$p(\mathbf{x}_g | \mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \Theta) = \mathcal{N}(\mathbf{x}_g; \boldsymbol{\mu} - \sum_k \lambda_k s_{gk} \mathbf{\Gamma} \mathbf{z}_k, \boldsymbol{\Sigma}) \quad (2)$$

$$p(\mathbf{S} | \mathbf{C}, \Theta) = \prod_{(g,k)} p(s_{gk} | \mathbf{C}, \Theta) = \prod_{(g,k) | c_{gk}=0} [s_{gk} = 0] \prod_{(g,k) | c_{gk}=1} \pi^{s_{gk}} (1 - \pi)^{1-s_{gk}} \quad (3)$$

$$p(\mathbf{\Gamma}|\Theta) = \prod_{t=1}^T p(\gamma_t|\Theta) = \prod_{t=1}^T \mathcal{N}(\gamma_t; 1, s^2) \quad (4)$$

$$p(\mathbf{\Lambda}|\Theta) = \prod_{k=1}^K p(\lambda_k|\Theta) = \prod_{k=1}^K \frac{1}{\alpha} \exp\left(-\frac{\lambda_k}{\alpha}\right) \quad (5)$$

$$p(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{Z}, \Theta) = p(\mathbf{S}|\mathbf{C}, \Theta)p(\mathbf{\Gamma}|\Theta)p(\mathbf{\Lambda}|\Theta) \prod_g p(\mathbf{x}_g|\mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \Theta) \quad (6)$$

$$p(\mathbf{X}|\mathbf{C}, \mathbf{Z}, \Theta) = \sum_{\mathbf{S}} \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} p(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{Z}, \Theta) d\mathbf{\Lambda} d\mathbf{\Gamma} \quad (7)$$

where \mathbb{I} is an indicator function, \mathbf{X} and \mathbf{Z} are the sets of expression profiles for mRNAs and miRNAs and \mathbf{C} is the set of candidate miRNA targets. Note that in the above model, we do not average over all of the model parameters: we have defined the parameter set $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi, \alpha\}$ as the set of regular parameters whose settings are estimated in a point-wise fashion.

The above model links the expression profiles of mRNA transcripts and miRNAs given a set of candidate miRNA targets: within these, we search for functional relationships by inferring the settings for the unobserved s_{gk} indicator variables. Fig. 3 shows the Bayesian network for our model of miRNA regulation. The boxes in Fig. 3, or "plates", indicate that the structures contained within them are replicated a number of times indicated at the top of each plate. This replication allows for multiple miRNAs to regulate any given target mRNA and also allows for different parameters for each tissue and miRNA in our data. Under our model, each transcript in the network is assigned a set of indicator variables which select for miRNAs that are likely to regulate its expression given the model parameters and the data.

Having presented the above model for miRNA post-transcriptional regulation, we would now like to infer the settings for the s_{gk} variables, or computing the posterior probability $p(\mathbf{S}|\mathbf{X}, \mathbf{Z}, \mathbf{C}, \Theta) \propto \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} p(\mathbf{X}, \mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{Z}, \mathbf{C}, \Theta) d\mathbf{\Lambda} d\mathbf{\Gamma}$. Exact inference would require integrating over the parameters $\mathbf{\Lambda}, \mathbf{\Gamma}$ in addition to summing over an exponential number of combinations of miRNA interactions per mRNA transcript, which generally will be difficult to compute. Thus, we will turn instead to an approximate method for learning which will make the problem tractable.

4 Variational Bayesian learning of miRNA targets

For variational Bayesian learning [2, 21] in a graphical model with unobserved variables \mathbf{u} , observed variables \mathbf{v} and model parameters η , the exact posterior over the unobserved variables and model parameters $p(\mathbf{u}, \eta|\mathbf{v})$ is approximated by a surrogate distribution $q(\mathbf{u}, \eta) = q(\mathbf{u})q(\eta)$ which is factorized. The problem of variational learning consists of an optimization problem in which we optimize

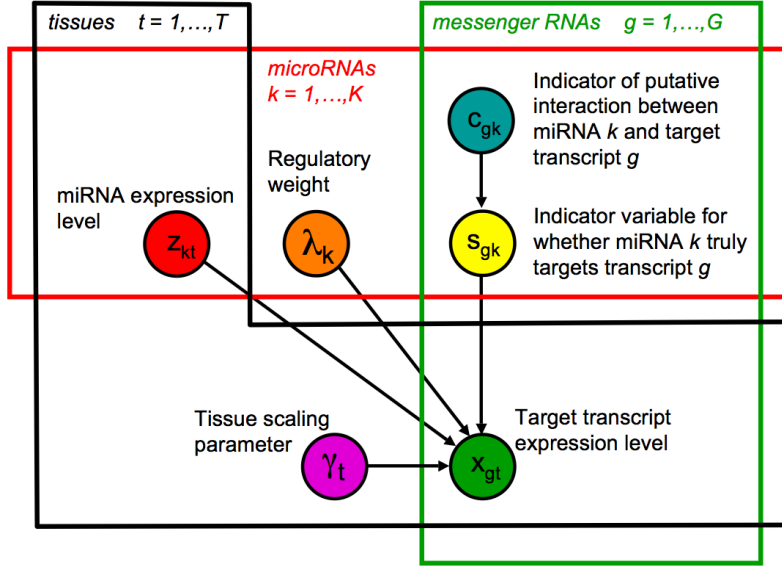


Fig. 3. Bayesian network used for finding functional miRNA targets. Nodes correspond to observed and unobserved variables as well as model parameters, with directed edges between nodes representing conditional dependencies encoded by our probability model. The boxes, or "plates", indicate that the structures contained within them are replicated a number of times indicated at the top of each plate. This replication allows for multiple miRNAs to regulate any given target mRNA and also allows for different parameters for each tissue and miRNA in our data. Each mRNA transcript in our network is assigned a set of indicator variables which select for miRNAs that are likely to regulate it given the model parameters and the data.

the fit between the surrogate distribution $q(\mathbf{u}, \eta)$ and the exact posterior distribution $p(\mathbf{u}, \eta | \mathbf{v})$. This fit is measured by the *KL-divergence* $D(q||p)$, which can be written as

$$\begin{aligned}
 D(q||p) &= \int_{\mathbf{u}, \eta} q(\mathbf{u})q(\eta) \log \frac{q(\mathbf{u})q(\eta)}{p(\mathbf{u}, \mathbf{v}, \eta)} d\mathbf{u} d\eta \\
 &= \int_{\mathbf{u}, \eta} q(\mathbf{u})q(\eta) \log \frac{q(\mathbf{u})q(\eta)}{p(\mathbf{u}, \eta | \mathbf{v})} d\mathbf{u} d\eta - \log p(\mathbf{v}) \\
 &\geq -\log p(\mathbf{v})
 \end{aligned} \tag{8}$$

We see that by optimizing the fit between the surrogate distribution $q(\mathbf{u}, \eta)$ and the exact posterior distribution $p(\mathbf{u}, \eta | \mathbf{v})$, we also minimize the negative log-likelihood $-\log p(\mathbf{v})$ of the observed data and we hence optimize the fit of our model to the data.

For the GenMiR++ model, \mathbf{X} , \mathbf{Z} and \mathbf{C} correspond to the observed variables \mathbf{v} , \mathbf{S} correspond to the unobserved variables \mathbf{u} , and $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ correspond to the

model parameters η . Thus, substituting these into Eqn. 8, we can write out the KL-divergence for our model as

$$D(q||p) = \sum_{\mathbf{S}} \int_{\Gamma} \int_{\Lambda} q(\mathbf{S}, \Lambda, \Gamma | \mathbf{C}) \log \frac{q(\mathbf{S}, \Lambda, \Gamma | \mathbf{C})}{p(\mathbf{X}, \mathbf{S}, \Gamma, \Lambda | \mathbf{C}, \mathbf{Z}, \Theta)} d\Lambda d\Gamma \quad (9)$$

To allow for tractable learning, we will simplify the q-distribution via a mean-field factorization with $q(\mathbf{S}, \Lambda, \Gamma | \mathbf{C}) = q(\mathbf{S} | \mathbf{C})q(\Lambda)q(\Gamma)$ with

$$\begin{aligned} q(\mathbf{S} | \mathbf{C}) &= \prod_{g,k} q(s_{gk} | \mathbf{C}) = \prod_{(g,k) | c_{gk}=1} \beta_{gk}^{s_{gk}} (1 - \beta_{gk})^{1-s_{gk}} \\ q(\Lambda) &= \prod_k q(\lambda_k) = \prod_k \frac{1}{\nu_k} \exp\left(\frac{-\lambda_k}{\nu_k}\right) \\ q(\Gamma) &= \prod_t q(\gamma_t) = \prod_t \mathcal{N}(\gamma_t; \omega_t, \phi_t^2) \end{aligned} \quad (10)$$

so that the targeting indicator variables s_{gk} , the regulatory weights λ_k and the tissue-scaling parameters γ_t are all independent of one another given the data. We thus approximate the true posterior distributions over these unobserved variables and model parameters using the above q-distributions, which we parameterize using variational parameters $\beta_{gk}, \nu_k, \omega_t, \phi_t^2$. Thus, β_{gk} represents the probability that miRNA k targets mRNA g given the data, ν_k represents the expected values of the regulatory weights and ω_t, ϕ_t^2 represent the means and variances of the tissue scaling parameters. The problem of inference in this setting is then to fit these variational parameters to the data by minimizing $D(q||p)$ given the regular model parameters and the data.

If we define the expected sufficient statistics $\Omega, \Phi, \mathbf{y}_g, \mathbf{U}, \mathbf{V}$ and \mathbf{W} as

$$\begin{aligned} \Omega &= \text{diag}(\omega_1, \dots, \omega_T), & \Phi &= \text{diag}(\phi_1^2, \dots, \phi_T^2) \\ \mathbf{y}_g &= \sum_{k:c_{gk}=1} \nu_k \beta_{gk} \mathbf{z}_k \\ \mathbf{U} &= \frac{1}{G} \sum_g \left(\mathbf{x}_g - (\boldsymbol{\mu} - \Omega \mathbf{y}_g) \right) \left(\mathbf{x}_g - (\boldsymbol{\mu} - \Omega \mathbf{y}_g) \right)^\top \\ \mathbf{V} &= \frac{1}{G} \sum_g \sum_{k:c_{gk}=1} \nu_k^2 (2\beta_{gk} - \beta_{gk}^2) (\Omega^2 + \Phi) \mathbf{z}_k \mathbf{z}_k^\top \\ \mathbf{W} &= \frac{1}{G} \Phi \sum_g \mathbf{y}_g \mathbf{y}_g^\top \end{aligned} \quad (11)$$

then the KL-divergence $D(q||p)$ can be written compactly as

$$\begin{aligned}
D(q||p) = & \sum_{(g,k)|c_{gk}=1} \left(\beta_{gk} \log \frac{\beta_{gk}}{\pi} + (1 - \beta_{gk}) \log \frac{1 - \beta_{gk}}{1 - \pi} \right) \\
& + \frac{G}{2} \log |\Sigma| + \frac{G}{2} \text{tr} \left(\Sigma^{-1} (\mathbf{U} + \mathbf{V} + \mathbf{W}) \right) \\
& + \frac{1}{2} \left[T \log s^2 + \frac{1}{s^2} \text{tr} ((\mathbf{\Omega} - \mathbf{I})^2 + \mathbf{\Phi}) - \log |\mathbf{\Phi}| \right] \\
& + \sum_k \left(\frac{\nu_k}{\alpha} - \log \frac{\nu_k}{\alpha} \right) + \text{const.} \tag{12}
\end{aligned}$$

We are now ready to learn our model from the data. Having defined the above objective function for learning and parameter estimation, we require an efficient method for optimization. We will use the variational Bayes algorithm which iteratively minimizes $D(q||p)$ with respect to the distribution over unobserved variables $q(\mathbf{S})$ (variational Bayes E-step), the distribution over model parameters $q(\mathbf{\Gamma}, \mathbf{\Lambda})$ (variational Bayes M-step) and with respect to the regular parameters (regular parameter optimization step) until convergence to a minimum of $D(q||p)$. Thus, the steps to the variational Bayes algorithm are given below:

Variational Bayes E-step:

$$\beta_{gk} = \arg \min_{\beta_{gk}} D(q||p) \tag{13}$$

Variational Bayes M-step:

$$\mathbf{\Phi} = \arg \min_{\mathbf{\Phi}} D(q||p)$$

$$\mathbf{\Omega} = \arg \min_{\mathbf{\Omega}} D(q||p)$$

$$\nu_k = \arg \min_{\nu_k} D(q||p)$$

(14)

Regular Parameter Optimization step:

$$\begin{aligned}
\boldsymbol{\mu} &= \arg \min_{\boldsymbol{\mu}} D(q||p) = \frac{1}{G} \sum_g (\mathbf{x}_g + \boldsymbol{\Omega} \mathbf{y}_g) \\
\boldsymbol{\Sigma} &= \arg \min_{\boldsymbol{\Sigma}} D(q||p) = \text{diag}(\mathbf{U} + \mathbf{V} + \mathbf{W}) \\
\pi &= \arg \min_{\pi} D(q||p) = \frac{\sum_{(g,k)|c_{gk}=1} \beta_{gk}}{\sum_{(g,k)|c_{gk}=1} 1} \\
\alpha &= \arg \min_{\alpha} D(q||p) = \frac{\sum_k \nu_k}{K}
\end{aligned}
\tag{15}$$

We iterate between the above three steps until we arrive at a minimum of $D(q||p)$, at which point we examine the β_{gk} targeting probabilities computed from the data. With the above variational Bayes algorithm and a set of candidate miRNA targets, we can now learn our model to obtain a set of high-confidence, functional miRNA targets. Once we have done so, we can analyze the performance of GenMiR++ to that of GenMiR [10, 11], as well as gauge the robustness of the model to various perturbations of the data.

5 Performance of GenMiR++

To compare GenMiR++ with our previous GenMiR model and learning algorithm, we used the above set of 1,770 TargetScanS 3'-UTR targets as well as the microarray data sets from [25] and [3]. Both GenMiR and GenMiR++ were applied to the above data by running their respective learning algorithms until convergence. Both learning algorithms were initialized with $\beta_{gk} = \pi = 0.5$ for all (g, k) pairs with $c_{gk} = 1$, $\alpha = \lambda_k = 5 \times 10^{-2} \forall k = 1, \dots, K$ and $s^2 = 10^{-2}$. The parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ were initialized to the sample mean and covariance of the mRNA expression data. Once both of the learning algorithms have converged, we assign a score to each of the candidate miRNA-target interactions according to

$$\text{Score}(g, k) = \log_{10} \left(\frac{\beta_{gk}}{1 - \beta_{gk}} \right)
\tag{16}$$

where each β_{gk} parameter, which corresponds to the probability that miRNA k targets transcript g , is computed from the data by each of the two methods. Thus a miRNA-mRNA pair is awarded a higher score if it was assigned a higher probability β_{gk} of being *bona-fide* given the observed expression profiles of mRNAs and miRNAs.

To assess how good of a representation of miRNA regulation has been learned by each method, we performed a series of permutation tests. Using the null hypothesis that there are no regulatory interactions between mRNAs and miRNAs,

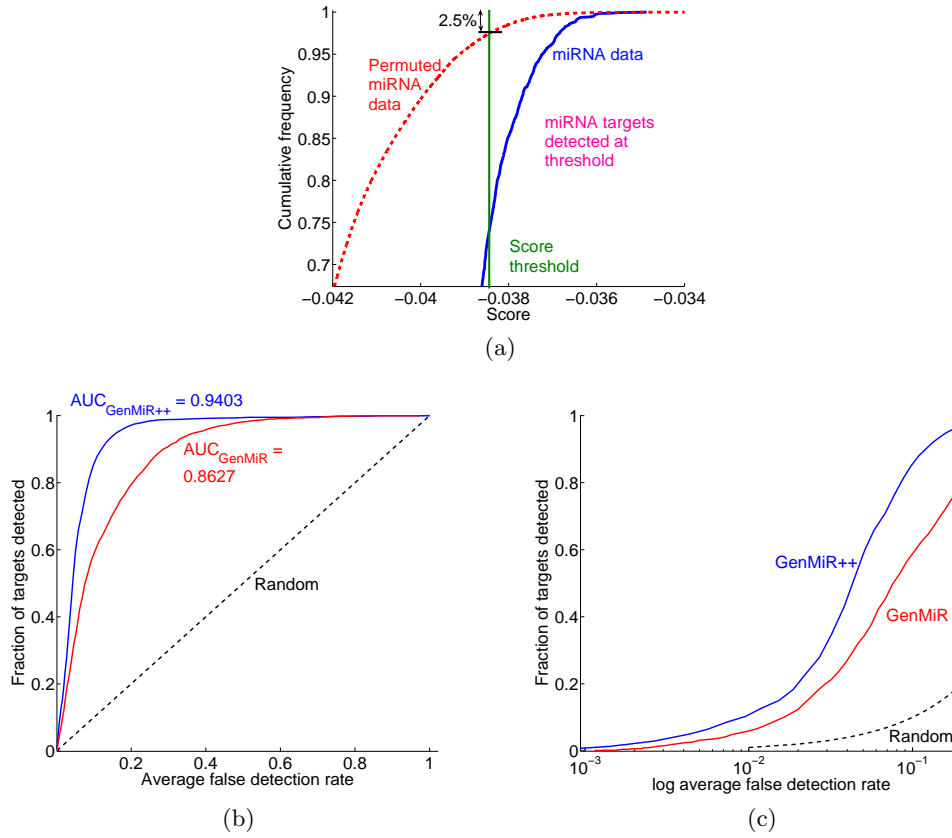


Fig. 4. (a) Empirical cumulative distribution of scores computed using GenMiR++ for the permuted data (red dashed) and unpermuted data for the set of 1,770 miRNA targets (blue solid); all scores above the threshold score correspond to high-confidence interactions (b) Fraction of candidate interactions detected ($\{\# \text{ of candidate interactions detected}\} / \{\# \text{ of candidate interactions}\}$) VS. average false detection rate ($\{\text{Average } \# \text{ of permuted interactions detected}\} / \{\# \text{ of candidate interactions}\}$) using GenMiR++ (blue) and GenMiR (red), computed over the set of 1,770 miRNA targets. GenMiR++ significantly outperforms the GenMiR model and learning algorithm from [11]. (c) Fraction of candidate interactions detected VS. log average false detection rate using both GenMiR++ and GenMiR.

we independently generated 100 different data sets corresponding to the original data with the gene labels permuted. By permuting the gene labels, we are destroying any information about relationships between miRNAs and mRNAs. We then applied both methods separately to each permuted data set and then computed the above score for each candidate miRNA target in the data set: the complete set of scores obtained from these 100 permuted data sets then provides us with an estimate of the null distribution over scores for each method. The resulting empirical cumulative distributions over scores obtained from GenMiR++

for both the permuted and unpermuted data are shown in Fig. 4(a). The plot shows that the set of scores learned from the unpermuted data (blue solid) are significantly higher than those learned from permuted data (red dashed), indicating that many of the candidate miRNA targets can be used to explain the expression levels of mRNA targets.

To find functional miRNA targets, we can threshold the scores computed from the original data: for different values of this threshold, we will get a certain number of false detections over the 100 permuted data sets. We can estimate the sensitivity and specificity for each threshold value by comparing the number of miRNA targets with score above the threshold to the average number of targets from the permuted data sets which also have a score above the threshold. The curve shown in Fig. 4(b) relates the fraction of candidate interactions detected ($\{\# \text{ of candidate targets detected}\} / \{\# \text{ of candidate targets}\}$) to the average false detection rate ($\{\text{Average } \# \text{ of permuted targets detected}\} / \{\# \text{ of candidate targets}\}$) for different threshold values, where the average false detection rate is computed for each threshold value using the average fraction of permuted miRNA-target interactions that score above that threshold. Figs. 4(b),4(c) show that GenMiR++ (blue), by averaging over a wide range of model parameter settings and by taking into account differences in normalization and hybridization between the mRNA and miRNA expression data sets, reduces overfitting and hence allows us to obtain a significantly larger fraction of functional miRNA targets in our set of 1,770 candidates for a given false detection rate as compared to GenMiR (red). Having shown that we can find functional miRNA targets with a small number of false detections, we would now like to assess what known miRNA-target interactions are recovered by GenMiR++: we address this in the next section.

6 Functional GenMiR++ miRNA targets

We used GenMiR++ to score each miRNA target in the above set of 1,770 TargetScanS candidates. Setting the threshold score to -0.0384 to control for a false detection rate of 2.5%, we have a set of 467 high-confidence miRNA targets, or 26.4% of the 1,770 candidates. Within our set of high-confidence targets, we manage to recover several confirmed miRNA-target interactions (Fig. 5). We have recovered 3 confirmed targets obtained from a survey of [5, 8]: we have recovered the interactions between miR-101 and the mouse homolog of MYCN [5], between miR-92 and MAP2K4 [5], a signal transduction gene, as well as the relationship between miR-16 and BCL2, a gene which has been implicated in chronic lymphocytic leukemia [8]. We also recover six out of 16 mouse homologs (SLC15A4, NEK9, PGRMC2, ITGB1, SFRS12 and HIPK1) of human transcripts which are predicted by TargetScanS as being targeted by miR-124a and miR-1 and were experimentally shown to be downregulated [9, 18, 27] by these two miRNAs. Given that we can recover the above nine out of 19 known miRNA targets amongst our 467 high-confidence targets (hypergeometric p-value $p < 0.0391$), we believe that the remainder of our high-confidence miRNA targets in fact represent a

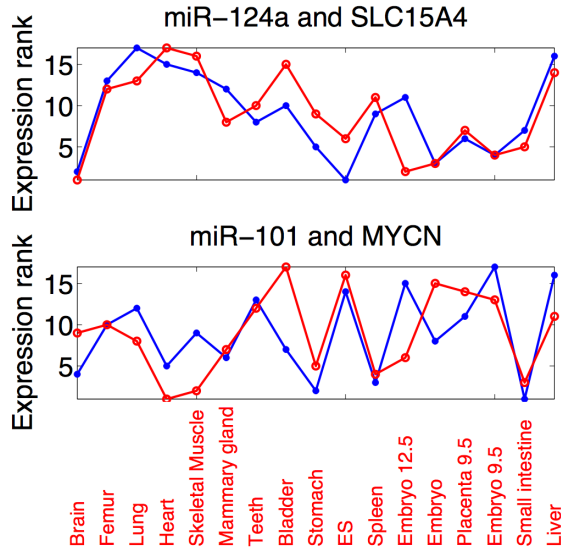


Fig. 5. Rank expression profiles of two of the nine high-scoring experimentally validated miRNA targets across 17 mouse tissues. For targeted mRNA transcripts (blue star), a rank of 17 denotes that the expression in that tissue was the highest amongst all tissues in the profile whereas a rank of 1 denotes that expression in that tissue was lowest amongst all tissues. Targeting miRNA expression profiles (red circle) are shown using a reverse scale, with a rank of 17 denoting that the expression was lowest and a rank of 1 denotes that expression in that tissue was highest.

significant increase in the number of known miRNA targets and merit further experimental investigation.

We have so far shown that in addition to being able to recover many known miRNA targets, GenMiR++ also affords us a significant increase in performance with respect to our previous GenMiR model and learning algorithm. Some readers may be curious as to how our set of high-confidence targets would change if we were to delete miRNAs or tissues from our data set. We will now assess the robustness of GenMiR++ to sub-sampling of the data and we will outline conditions in which we can learn meaningful miRNA targets from the data.

7 Robustness of GenMiR++

In this section, we examine the dependence of the performance of GenMiR++ on the amount of data used for learning miRNA targets. We will first outline the effect of removing miRNAs or tissues from our data set and then learning from the sub-sampled data. We will then examine how our model's discriminative ability is affected by introducing fake targets into our set of candidate miRNA targets.

7.1 Robustness to sub-sampling of miRNAs

In this part, we generated ten distinct data sets for each miRNA-target pair (g, k) with $c_{gk} = 1$ and for each miRNA sub-sample size $K_0 \in \{2, \dots, K - 2\}$. We then applied GenMiR++ to each of these data sets and scored all targets. For each putative miRNA-target interaction (g, k) with $c_{gk} = 1$, we computed the mean targeting probability $\langle \beta_{gk} \rangle$ and the standard deviation of the targeting probability β_{gk} over the ten test data sets for a given sub-sample size. Fig. 6(a) shows a scatter plot of the coefficient of variation (defined as the ratio of the standard deviation to the mean) versus the deviation from the mean $\beta_{gk} - \langle \beta_{gk} \rangle$ for each (g, k) pair. As can be seen, our method is quite robust to sub-sampling of the data set by miRNAs, as both the coefficient of variation and deviation from the mean remain very small for larger sub-sample sizes and thus the targeting probabilities computed using our method should have low uncertainty. Note, however, that in the limit of having only two miRNAs to learn from, both the coefficient of variation and the deviation from the mean become larger, indicating higher uncertainty in the estimated β_{gk} probabilities. Given that the set of points corresponding to learning from a sub-sample of 20 miRNAs has low variance and given the fact that our full data set contains 22 miRNAs, we infer that the set of miRNA targets we have learned using GenMiR++ should not vary significantly if we were to add miRNAs to our data set.

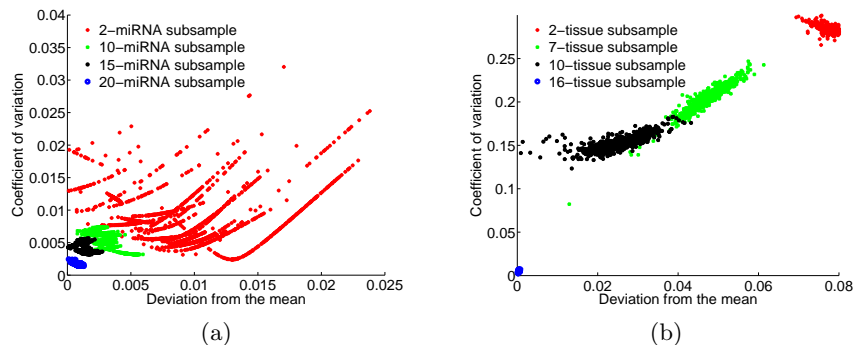


Fig. 6. Coefficient of variation versus deviation from the mean $\beta_{gk} - \langle \beta_{gk} \rangle$ when sub-sampling by (a) miRNA and (b) by tissue for various sub-sample sizes. Each point corresponds to a (g, k) pair for which $c_{gk} = 1$. In both tests, the uncertainty for any given (g, k) pair increases as sub-sample size decreases.

7.2 Robustness to sub-sampling of tissues

We repeated the above analysis for tissues by generating ten distinct data sets for tissue sub-sample sizes $T_0 \in \{2, \dots, T - 2\}$ and learning our model on each of these. Fig. 6(b) shows a scatter plot of the coefficient of variation versus the deviation from the mean for all pairs (g, k) with $c_{gk} = 1$. Given that the set of points for a sub-sample size of 16 tissues has low variance, we can conclude

that the targeting probabilities which are computed from the full data set with 17 tissues have low uncertainty and thus our set of functional miRNA targets should not vary considerably if we delete a small number of tissues from our data set.

However, from comparing Fig. 6(a) to Fig. 6(b), we see that our model is not as robust to sub-sampling by tissues as compared to sub-sampling by miRNA. Indeed, Fig. 6(b) shows that the uncertainty in the estimated β_{gk} probabilities increases significantly as we remove tissues from our data set, whereas we do not observe as dramatic of an increase in uncertainty of the targeting probabilities from sub-sampling by miRNA. Our results suggest that GenMiR++ is more robust to sub-sampling by miRNA than to sub-sampling by tissue. From this, we infer that any increase in accuracy in our learning procedure is more likely to come from adding tissues to our data set rather than from adding miRNAs and their corresponding targets.

7.3 Robustness to fake targets in data

An interesting question is how our model is affected by introducing large numbers of fake targets into a list of candidates. To address this, we introduced increasing numbers of fake targets into the set of 1,770 candidates by permuting a random subset of the candidate miRNA-mRNA interactions and then applying GenMiR++ to this modified set of targets. Thus, we apply our method to a data set which contains both fake and non-fake targets. For a given number of fake targets, we assign a score to each of the targets, including both fake and non-fake targets. Given the two sets of scores for fake and non-fake targets, we can then assess whether our method can discriminate between the two types: Table 1 shows the p-values obtained from a one-tailed Wilcoxon-Mann-Whitney test of the two sets of scores for a given number of fakes in the list of candidate miRNA targets. As can be seen, our method is able to discriminate between fake and non-fake targets by assigning lower scores to the fake targets in our list and higher scores to the non-fake targets. Note that our method is able to discriminate between fakes and non-fakes even when as many as half of the candidate miRNA targets are known to be fake. This result, in combination with the permutation testing above, suggests that our method can learn an accurate representation of miRNA regulation from a set of candidate miRNA targets and expression data.

8 Discussion

In this paper, we have presented GenMiR++, a Bayesian model and learning algorithm which combines candidate miRNA targets with mRNA and miRNA expression data to find a set of functional miRNA targets. Our model accounts for both mRNA and miRNA patterns of expression given a set of candidate miRNA targets: we then learn our model from this to obtain a set of functional miRNA targets. We have shown how to learn the model from sequence and

Fraction of targets which are fakes	p-value
10%	0.0230
25%	0.0100
50%	$< 10^{-9}$

Table 1. One-tailed Wilcoxon-Mann-Whitney test p-values for testing differences in scores for fake and non-fake targets in the set of TargetScanS candidates. Fake targets are assigned lower scores under our model than non-fake targets, indicating that our model is capable of discriminating between the two types.

expression data: GenMiR++ was shown to significantly outperform the original GenMiR method for learning miRNA targets. Our results suggest that our method is robust to sub-sampling of the data both by miRNA and by tissue, as well as to inserting additional fake miRNA targets into our data set. We’ve shown that our method learns an accurate representation of miRNA regulation in the form of a set of functional miRNA targets, even in the presence of many fake candidate targets. Using our method, we’ve obtained a set of 467 high-confidence miRNA targets out of a total of 1,770 TargetScanS miRNA targets at a false detection rate of 2.5%: given that we recover several confirmed miRNA targets and given the above results, we believe most of these remaining targets constitute a significant increase in the number of known miRNA targets and merit further experimental investigation.

Our model is the first to explicitly use patterns of miRNA and gene expression and the combinatorial aspect of miRNA regulation to learn a set of *bona fide* miRNA targets from sequence and expression data. Our model extends previous work [26], which has focused on the *de novo* finding of targets based on sequence and then associating miRNAs to their activity conditions through mRNA expression data alone. Our model, which also uses patterns of mRNA expression, improves on the above by taking into account patterns of miRNA expression to learn miRNA targets. More recent work in [23] has focused on identifying regulatory relationships by examining mRNA transcripts which are downregulated in select tissues due to the over-expression of tissue-specific miRNAs: they subsequently link the expression of the downregulated mRNAs with the number of occurrences of miRNA target sites in the 3’UTR regions of these targeted mRNAs. Our model extends this work by leveraging information about patterns of expression for both miRNAs and mRNAs across multiple tissues and across multiple miRNAs under a Bayesian framework for miRNA regulation which allows us to find a set of functional miRNA targets.

One issue we have not discussed in this paper is the fact that under our model, miRNAs putatively regulate their mRNA target transcripts by post-transcriptionally degrading them [4, 9, 18]. Current evidence suggests, however, that miRNAs can downregulate protein expression either by causing mRNA target degradation or by inhibiting translation without causing transcript degradation [19, 22]. It has been suggested that translational repression by miRNAs may be accompanied by the degradation of mRNA targets [4, 19], so that the effect

of down-regulation in miRNA-target interactions which are better accounted for by the TR model can nevertheless be observed at the mRNA level alone in the form of high miRNA expression/low targeted mRNA expression. Given this and the results presented in this paper, we believe that the GenMiR++ model allows us to infer whether a putative miRNA-target relationship is functional for many putative miRNA-target relationships, whether they be regulated via translational repression, by post-transcriptional degradation, or both. However, as there might be a significant fraction of putative miRNA-target relationships which cannot be explained using miRNA and mRNA expression data alone, we are actively pursuing the idea of extending the above model for miRNA regulation to account for both possible mechanisms for miRNA regulation. This would allow us to increase the accuracy with which we can identify functional miRNA targets from biological data. Establishing the extent to which either of these regulatory mechanisms would provide an important modeling framework for miRNA activity on global scale.

9 Acknowledgments

We would like to thank Tomas Babak for many insightful discussions and anonymous referees for recommending means for testing our model and learning algorithm. JCH was supported by a NSERC Postgraduate Scholarship and a CIHR Net grant. QDM was supported by an NSERC Discovery Grant. BJB was supported by a Premier's Research Excellence Award and a gift from Microsoft Corporation.

References

1. Ambros V (2004) The functions of animal microRNAs. *Nature* **431**, 350-355.
2. Attias H (1999) Inferring parameters and structure of unobserved variable models by variational Bayes. *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, Morgan Kaufmann Publishers, 21-30.
3. Babak T, Zhang W, Morris Q, Blencowe BJ and Hughes TR (2004) Probing microRNAs with microarrays: Tissue specificity and functional inference. *RNA* **10**, 1813-1819.
4. Bagga S *et al* (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**, 553-63.
5. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116**, pp.281-297.
6. Bentwich I *et al* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics* **37**, 766-770.
7. Berezikov E *et al* (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120(1)**, 21-24.
8. Cimmino A *et al* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *PNAS*, **102**, 13944-13949.
9. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310**, 1817-1821.

10. Huang JC, Morris QD and Frey BJ (2005) A computational high-throughput method for detecting miRNA targets. *University of Toronto Technical Report*, PSI TR 2005-026, August 15 2005.
11. Huang JC, Morris QD and Frey BJ (2006) Detecting microRNA targets by linking sequence, microRNA and gene expression data. *Proceedings of the Tenth Annual Conference on Research in Computational Molecular Biology*.
12. Hughes TR *et al* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19**, 342-347.
13. Kent WJ (2002) BLAT – The BLAST-Like Alignment Tool. *Genome Research* **4**, 656-664.
14. Krek A *et al* (2005) Combinatorial microRNA target predictions. *Nature Genetics* **37**, 495-500.
15. Lee RC, Feinbaum RL, and Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843-854.
16. Lewis BP, Burge CB and Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20.
17. Lim LP *et al* (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991-1008.
18. Lim LP *et al* (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-773.
19. Liu J, Valencia-Sanchez MA, Hannon GJ, and Parker R. (2006) MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol* **7**, 719-723.
20. Lockhart M *et al* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675-1680.
21. Neal, RM and Hinton, GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, 355-368.
22. Sen GL, Wehrman TS, and Blau HM. (2006) mRNA translation is not a prerequisite for small interfering RNA-mediated mRNA cleavage. *Nat Cell Biol* **73**, 287-293.
23. Sood P *et al* (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proceedings of the National Academy of Sciences (PNAS)* **103**, 2746-2751.
24. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123**, 1133-1146.
25. Zhang W, Morris Q *et al* (2004) The functional landscape of mouse gene expression. *J Biol* **3**, 21-43.
26. Zilberstein CBZ, Ziv-Ukelson M, Pinter RY and Yakhini Z (2005) A high-throughput approach for associating microRNAs with their activity conditions. *Proceedings of the Ninth Annual Conference on Research in Computational Molecular Biology*.
27. Supplemental Data for Lewis *et al*. *Cell* **120**, pp. 15-20.
<http://web.wi.mit.edu/bartel/pub/Supplemental%20Material/Lewis%20et%20al%202005%20Supp/>