BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i89–i97
doi:10.1093/bioinformatics/btt231

Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation

Hai-Son Le[1] and Ziv Bar-Joseph[1,2,*]

[1]Machine Learning Department and [2]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally. MiRNAs were shown to play an important role in development and disease, and accurately determining the networks regulated by these miRNAs in a specific condition is of great interest. Early work on miRNA target prediction has focused on using static sequence information. More recently, researchers have combined sequence and expression data to identify such targets in various conditions.

Results: We developed the Protein Interaction-based MicroRNA Modules (PIMiM), a regression-based probabilistic method that integrates sequence, expression and interaction data to identify modules of mRNAs controlled by small sets of miRNAs. We formulate an optimization problem and develop a learning framework to determine the module regulation and membership. Applying PIMiM to cancer data, we show that by adding protein interaction data and modeling cooperative regulation of mRNAs by a small number of miRNAs, PIMiM can accurately identify both miRNA and their targets improving on previous methods. We next used PIMiM to jointly analyze a number of different types of cancers and identified both common and cancer-type-specific miRNA regulators.

Contact: zivbj@cs.cmu.edu

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

1 BACKGROUND

MicroRNAs (miRNAs) are a family of small non-coding RNA molecules that regulate gene expression post-transcriptionally. These single-stranded RNAs, 19–25 nt long, are initially transcribed as longer independent genes or from introns of protein-coding genes. MiRNAs are now known to play a major role in development (Bartel, 2009), various brain functions (Shao *et al.*, 2010) and diseases (Meola *et al.*, 2009). Since their discovery, several 100 miRNAs were identified in each of several different species, including mammals, worms, flies and plants (He and Hannon, 2004). Initial discovery of large sets of miRNAs relied heavily on sequence and conservation analysis (Bartel, 2009), although recent advances in sequencing capacity are now allowing researchers to validate and identify additional miRNAs experimentally (Motameny, 2010). Most miRNAs target the genes they regulate by binding to the 3′-untranslated region of the target mRNAs (using complementary base pairing) and recruiting additional machinery to either degrade these mRNAs or prevent them from being translated. The miRNA regulation is ubiquitous, and a single miRNA can target 100s and even 1000s of genes. As the effect of each miRNA on any single target is often limited, they often work cooperatively with multiple miRNAs targeting the same mRNA in a specific condition (Krek *et al.*, 2005; Krol *et al.*, 2010).

Although the set of active miRNAs can often be determined experimentally (by measuring their expression levels), identifying their targets is much more challenging. Determining such target set is important for fully understanding the role of various miRNAs and to model the networks they regulate in a condition of interest. Initially, computational methods developed to predict such targets primarily relied on sequence information, in some cases, also using conservation information and/or secondary structure predictions. These methods search for base pair complementarity between the mature miRNA and 3′-untranslated regions of all mRNAs, allowing for some mismatches (the penalty for mismatches differs between the methods). Popular methods include TargetScan (Lewis *et al.*, 2005), miRBase (now called MicroCosm) (Griffiths-Jones *et al.*, 2006), miRanda (John *et al.*, 2004) and PicTar (Krek *et al.*, 2005).

Although these predictions are useful, because of the short length of miRNAs, they lead to many false positives and some false negatives (Betel *et al.*, 2010). Conservation analysis has proven especially problematic in this domain, as several real targets are not well conserved and would be ignored if conservation is a requirement (Barakat *et al.*, 2007). In addition, sequence data are static and do not change in different conditions or at different times. Thus, based on sequence data alone, it is impossible to map the set of targets for specific miRNA in a condition of interest (as most genes are not expressed in any specific condition or tissue). Finally, miRNAs often work cooperatively in small groups. As miRNA activation is condition specific, using this cooperative regulation property requires the use of condition-specific data, which of course cannot be inferred from sequence information alone.

Transcription factors (TFs) also play a major role in regulating gene expression, and they have been shown to work combinatorially with miRNAs (Sun *et al.*, 2012). However, a pre-requisite for such combinatorial analysis is a list of targets for individual miRNAs. Unlike TFs, which can serve as activators or repressors and are often post-transcriptionally regulated, miRNAs are only transcriptionally regulated and inhibit their direct targets. This has led to several studies that isolated the miRNA target prediction task by integrating sequence, mRNA and miRNA expression data (Cheng and Li, 2008; Huang *et al.*, 2007a; Joung *et al.*, 2007; Ooi *et al.*, 2011). Unlike sequence data, expression data are dynamic and condition-specific and thus provide useful clues about the set of active miRNAs and mRNAs. A number of methods, mostly based on (anti) correlation or regression analysis using the expression levels of miRNAs and predicted

*To whom correspondence should be addressed.

mRNA targets, were suggested for this task (Huang *et al.*, 2011; Wang and Li, 2009). A representative example for this group is GenMiR++ (Huang *et al.*, 2007a), one of the first methods to integrate miRNA and mRNA expression profiles in a unified probabilistic model. Given an expression dataset for both miRNAs and mRNAs and a set of putative miRNA–mRNA interactions (inferred from sequence data), GenMiR++ uses a generative probabilistic regression model to assign targets to miRNAs. It was successfully applied to identify targets of let-7b in retinoblastoma. Another approach is to project mRNA expression data on pathway databases and compute the correlation between miRNAs and average pathway expression levels to identify likely regulators of signaling pathways (Ooi *et al.*, 2011). Although this method does not identify specific targets, it can be used to infer the function of specific miRNAs based on the pathways they regulate. A number of other methods for integrating miRNA and mRNA expression data have been proposed, see (Muniategui *et al.* 2012) for a recent review.

Finally, there is growing evidence that interacting proteins are more likely to be co-regulated by the same miRNAs (Hsu *et al.*, 2008; Liang and Li, 2007). It has also been shown that some miRNAs coordinately target protein complexes (Sass *et al.*, 2011). Although such complementary information may be important, few previous works have taken advantage of it to predict condition-specific interactions. An exception is a recent work by Zhang *et al.* (2011), which developed SNMNMF to integrate protein interactions with miRNA and mRNA expression data. The method is based on a non-negative matrix factorization analysis, which factorizes the two expression data matrices such that the two share one common factor, which is assumed to be the module basis matrix **W**. Note, however, that although this method was successfully applied to analyze Ovarian cancer data, it does not use a regression model to explain mRNA expression levels, or requires that miRNAs and mRNAs in the same module be anti-correlated; therefore, the resulting modules do not fully use current knowledge regarding the inhibitory role of miRNAs, which may lead to missing important interactions.

The methods discussed earlier in the text successfully integrated expression and sequence data. However, a major point that is often ignored by these prediction methods is the combinatorial aspect of miRNA regulation. Several studies have shown that individual miRNAs have only limited impact on their targets (Malumbres, 2012) and multiple (different) miRNAs are needed to drastically reduce transcription levels of targets. To allow the use of such group- or module-based regulatory model, we have recently developed GroupMiR (Le and Bar-Joseph, 2011), which uses a non-parametric Bayesian prior based on the Indian Buffet Process (IBP; Griffiths and Ghahramani, 2006) to identify modules of co-regulated miRNAs and their target mRNAs. As we have shown, by using a module-based approach, we can improve on methods that treat miRNAs or mRNAs individually improving the set of correctly recovered miRNA–mRNA interactions (Le and Bar-Joseph, 2011).

Here, we present the Protein Interaction based MicroRNA Modules (PIMiM) method, which extends the regression framework of GroupMiR by using an additional type of data: protein interactions (Fig. 1). As we show, by defining a new target

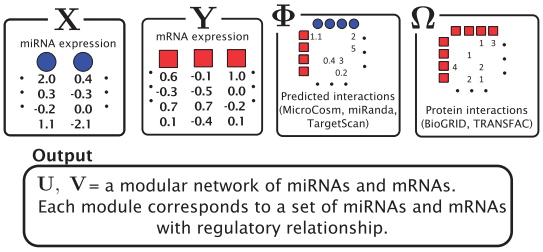

Fig. 1. Data used as input for PIMiM. In addition the miRNA and mRNA expression data, PIMiM uses sequence-based predictions of miRNA–mRNA interactions and protein–protein interactions. These datasets are integrated as discussed in Section 2

function that encourages interacting proteins to belong to the same module, we can use such data and integrate it with expression and sequence-based data in a probabilistic model. We develop an iterative learning procedure to learn the parameters of our model and show that it converges to a local minima. Comparison of PIMiM with previous methods indicates that by combining a module-based approach with protein interaction data, we can improve on both methods that only rely on modules (GroupMiR) and methods that rely on protein interaction (SNMNMF). We used PIMiM to study miRNA in several types of cancers, allowing us to identify novel regulators that either span multiple cancer types or are unique to specific cancers.

2 METHODS

2.1 Overview

We developed PIMiM, a module-based method that predicts targets for miRNAs by assigning them, together with the mRNAs they regulate, to one of K modules. Modules may contain several miRNAs and many mRNAs, and both miRNA and mRNAs can be assigned to 0, 1 or multiple modules, and thus modules may overlap.

The input to PIMiM is condition-specific miRNA and mRNA expression data (usually multiple measurements from patients or different time points). In addition, we use sequence-based predictions of miRNA–mRNA interactions (any probabilistic predictions can be used) and static protein interaction data. Using these datasets we learn a regularized probabilistic regression model in which mRNA data are regressed to the expression data of miRNAs assigned to modules regulating it. The down-regulation effect of an miRNA on the expression of its target mRNA is aggregated across all modules, allowing information to be shared between modules in the learning process. Our probabilistic model rewards the assignments of predicted miRNA–mRNA pairs to the same module and also rewards assignment of mRNAs of interacting proteins to the same module. Combined, the modules explain the observed mRNA expression data as a function of their regulating miRNAs and the set of proteins they interact with.

2.2 Notations

We use the following notation in the rest of the article. We assume there are M miRNAs and N mRNAs in each sample. We denote expression profiles of miRNAs by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)^{\mathrm{T}}$ and of mRNAs by $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^{\mathrm{T}}$, where $\mathbf{x}_i$ and $\mathbf{y}_j$ are vectors with the expression levels of miRNA i and mRNA j, respectively, in all samples. Both matrices have P columns corresponding to the P-matched samples. In addition, let Ω (sparse $N \times N$ matrix) be the weighted adjacency matrix of the

protein interactions [obtained from databases, such as BioGRID (Stark *et al.*, 2011) or TRANSFAC (Wingender *et al.*, 2000)] and Φ (sparse $M \times N$ matrix) be the list of predicted interactions of miRNAs and mRNAs from sequence data (obtained from prediction databases, such as MicroCosm; Griffiths-Jones *et al.*, 2006). We also define $\mathbf{I}_\Phi$ and $\mathbf{I}_\Omega$ as binary matrices indicating whether an entry of Φ and Ω, respectively, is non-zero.

For learning K modules, our goal is to determine (learn) the values of the membership parameters u_{ik} and v_{jk}, which represent the propensity that miRNA i or mRNA j belong to module k. Naturally, we restrict these parameters to be non-negative: $u_{ik} \geq 0$ and $v_{jk} \geq 0$, where we interpret that an miRNA or an mRNA is not assigned to a module if the corresponding parameter is zero. We use matrices $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_M)^{\mathrm{T}}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)^{\mathrm{T}}$ to represent this complete set of membership parameters. Finally, we use the following subscript such as $\mathbf{u}_{,k}$ or $\mathbf{v}_{,k}$ to denote the kth column of the matrices.

$\mathbf{U}, \mathbf{V}$: miRNA and mRNA module membership
K : number of modules
$\mathbf{u}_i, \mathbf{v}_j$: ith or jth rows of the matrices
$\mathbf{u}_{,k}, \mathbf{v}_{,k}$: kth columns of the matrices
$\mathbf{I}_\Phi, \mathbf{I}_\Omega$: binary indicators of Φ, Ω

2.3 Probabilistic regression model

Following previous works (Huang *et al.*, 2007b; Le and Bar-Joseph, 2011), we use a regression-based method to link the expression profiles of miRNAs and mRNAs. Expression values of mRNAs are assumed to be downregulated from a baseline expression level by a linear combination of expression profiles of all their predicted miRNA regulators. For example, mRNA j's expression values are distributed as: $\mathbf{y}_j \sim \mathcal{N}\left(\mu - \sum_{i \in \mathcal{S}_j} w_{ij}\mathbf{x}_i, \Sigma\right)$, where μ is the baseline expression level, $\mathbf{w}_i$ are weights associated with miRNAs (which previous methods learn individually for each mRNA) and $\mathcal{S}_j$ is the set of predicted miRNA regulators of mRNA j.

We depart from these previous models in how we specify miRNA regulators and how we learn the weights $\mathbf{w}_i$. First, each mRNA is assumed to be a target of all miRNAs assigned to the modules it belongs to as long as they are predicted to regulate it ($\phi_{ij} \neq 0$). Formally, mRNA j is the target of the set of miRNAs $\mathcal{S}_j = \{i : \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j > 0 \text{ and } \phi_{ij} \neq 0\}$. Second, the downregulation weights are aggregated across all modules, such as $w_{ij} = \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$.

Given these assumptions, the likelihood of the observed expression values is

$$\begin{aligned} p(\mathbf{Y}|\mathbf{U}, \mathbf{V}, \mathbf{X}, \mu, \Sigma) &= \prod_j \mathcal{N}\left(\mathbf{y}_j|\mu - \sum_{i \in \mathcal{S}_j} \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j\mathbf{x}_i, \Sigma\right) \\ &= \prod_j \mathcal{N}\left(\mathbf{y}_j|\mu - \mathbf{X}^{\mathrm{T}}((\mathbf{I}_\Phi)_{,j} \circ (\mathbf{U}\mathbf{v}_j)), \Sigma\right) \end{aligned} \tag{1}$$

where $\Sigma = \mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_P^2)$ is the per-sample variance terms.

2.4 Using protein interactions

So far PIMiM only uses expression values in a regression setting (although we constrain the regulators to come from the sequence-based predicted set, the regression model itself does not directly encourage the assignment of miRNA and predicted mRNA targets to the same module).

To incorporate the input interaction data (predicted miRNA–mRNA pairs Φ and protein interactions Ω), we use a function that rewards assignments to the same module based on the strength of the predicted edge as follows:

$$\begin{aligned} p(I_{\phi_{ij}} = 1|\mathbf{U}, \mathbf{V}) &= \frac{1}{1 + \exp(-\alpha\ > \phi_{ij}\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j)} = \sigma(\alpha\ > \phi_{ij}\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j) \\ p(I_{\phi_{ij}} = 0|\mathbf{U}, \mathbf{V}) &= 1 - \sigma(\alpha\ > \mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j) \\ p(I_{\omega_{jj'}} = 1|\mathbf{V}) &= \sigma(\beta\ > \omega_{jj'}\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_{j'}) \end{aligned} \tag{2}$$

Where α and β are positive tuning parameters that are used to adjust the contributions of the two types of interaction data in our model and $\sigma(.)$ is the logistic-sigmoid function. If available (as is the case for the miRNA–mRNA interaction data), we use probabilities for Φ and Ω derived directly from the prediction or experimental databases (see Section 4). We deliberately do not include penalty terms for zero entries of Ω because this interaction matrix is extremely sparse (the number of known protein–protein interactions is small compared with the total number of possible interactions). Penalizing zero entries when using such a sparse matrix would lead to small modules and may be less biologically accurate, as not all co-targets of a miRNA interact.

These terms indicates that the higher the probability of interaction (both miRNA–mRNA and protein–protein) the more likely it is that the interacting entities would be assigned to the same set of modules. This is done globally across all modules. For instance, if ϕ_{ij} is positive, we have previous knowledge that miRNA i and mRNA j interact. To maximize the likelihood $p(\phi_{ij}|\mathbf{U}, \mathbf{V})$, we would need to learn parameters that lead to large values of $\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$, which means that the method is more likely to place them in the same module.

2.5 Overall log-likelihood

To summarize, our target is to minimize the following negative log-likelihood:

$$\begin{aligned} \mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{\Phi}, \mathbf{\Omega}) &= -\log\ p(\mathbf{Y}|\mathbf{U}, \mathbf{V}, \mathbf{X}, \boldsymbol{\mu}, \Sigma) \\ &- \sum_{i,j} \log p(I_{\phi_{ij}}|\mathbf{U}, \mathbf{V}) - \sum_{j \neq j'} \log p(I_{\omega_{jj'}} = 1|\mathbf{V}) \end{aligned} \tag{3}$$

The first term evaluates how well the miRNA expression explains the observed mRNA expression, whereas the second and third terms are rewards for assigning predicted miRNA–mRNA pairs and protein interaction pairs to the same module, respectively. This function is non-convex and thus can have multiple local minima solutions. To constrain the set of solutions, we add a number of regularization terms. First, we add two sets of ℓ_1 norm constraints for the vectors $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_j\}$. ℓ_1 norm contraints encourage sparsity leading to smaller and tighter modules. As our goal is to reduce false positives, such constraints are useful, as they reduce the set of predicted miRNA–mRNA pairs. Specifically, we require that

$$\begin{aligned} ||\mathbf{u}_i||_1 &\leq C_1, \quad i = 1, \ldots, M \\ ||\mathbf{v}_j||_1 &\leq C_2, \quad j = 1, \ldots, N \end{aligned}$$

We are using two different regularization parameters C_1 and C_2. This is because the number of miRNAs and mRNAs are different; therefore, a single number does not yield good solutions. Moreover, we choose to use these constraints explicitly instead of adding them to the objective function (using Lagrangian multipliers), as this formulation is simpler to solve in our optimization procedure.

Together, our learning phase solves the following optimization:

$$\begin{aligned} \min_{\mathbf{U} \geq 0, \mathbf{V} \geq 0, \mu, \Sigma} \quad & \mathcal{F} = \mathcal{L}(\mathbf{Y}, \mathbf{X}, \Phi, \Omega) \\ \text{s.t.} \quad & ||\mathbf{u}_i||_1 \leq C_1, i = 1, \ldots, M \\ & ||\mathbf{v}_j||_1 \leq C_2, j = 1, \ldots, N \end{aligned} \tag{4}$$

2.6 Learning the parameters of our model

In this section, we discuss how to solve the optimization problem from (4) to determine values for the parameters of our model. As aforementioned, this problem is non-convex, and we cannot analytically compute general solutions. However, we notice that by holding $\mathbf{U}$ and $\mathbf{V}$ fixed, we can solve for μ and Σ in a closed form using standard linear regression:

$$\hat{\mu}_p = \frac{1}{N}\sum_{j=1}^{N}(z_{jp} + y_{jp}) \qquad \hat{\sigma}_p^2 = \frac{1}{N}\sum_{j=1}^{N}(\hat{\mu}_p - y_{jp} - z_{jp})^2$$

where $z_{jp} = \mathbf{x}_{\cdot,p}^{\mathrm{T}}((\mathbf{I}_\Phi)_{\cdot,j} \circ (\mathbf{U}\mathbf{v}_j))$ for $j = 1, \ldots, N$ and $p = 1, \ldots, P$.

To solve for $\mathbf{U}$ and $\mathbf{V}$ for given values of μ and Σ, we use a projected quasi-Newton (PQN) method (Schmidt *et al.*, 2009). Quasi-Newton methods construct an approximation to the Hessian by using the observed gradients at successive iterations. We use the MATLAB implementation min_PQN (http://www.di.ens.fr/mschmidt/Software/PQN.html). There are several reasons why we chose this method instead of directly working with the Hessian. First, our set of constraints is convex, and the projection on this set can be done analytically. Second, although we can compute both the gradients and Hessian of $\mathcal{F}$, the memory required to store the Hessian is often too large given the dimensions of the expression data ($O((M+N)^2K^2)$). Moreover, because of interactions between miRNAs and mRNAs, the Hessian is not necessary sparse even if both Φ and Ω are. During the projection step, to speed-up the convergence of the algorithm, we set the entries of $\mathbf{U}$, which do not have predicted interactions to zero.

Using the updated values for $\mathbf{U}$ and $\mathbf{V}$, we once again solve for μ and Σ and so on. These two steps lead to an iterative procedure to solve (4) along the lines of coordinate-descent methods. This procedure converges to the local minima because of the fact that the objective function is bounded below, and the sequence of function values is monotonically decreasing, and the gradients at the convergence are zeros. As the problem is non-convex, we perform the learning process several times, randomly initializing the parameters each time. After repeating this process several times (10 iterations in our experiments), we select the parameters from the result that leads to the lowest value for our objective function.

Finally, the regularization and data-type–weighting parameters α, β, C_1 and C_2 are chosen based on an external evaluation discussed in Section 4.

3 CONSTRAINT MODULE LEARNING FOR MULTIPLE CONDITION ANALYSIS

So far we have discussed our approach for identifying miRNA-regulated modules using a condition-specific expression dataset. Although the optimization problem in Equation (4) can be used with expression data from multiple conditions (e.g. different types of cancer), the output is one set of modules for all conditions. In some cases, directly identifying similar and divergent modules across conditions is an important goal. Consider, for example, joint analysis of multiple types of cancers. Although some researchers may be interested in regulatory modules that are activated in all different cancer types, others may be interested in unique aspects, or modules, of a specific cancer type when compared with other types of cancer.

In our problem, we would like to learn a set of modules for T different conditions. The interaction input matrices Φ and Ω are fixed, whereas for each condition t, we have a set of expression measurements $\mathbf{X}_t$ and $\mathbf{Y}_t$. Given this input, we jointly learn T sets of modules $\{\mathbf{U}^t, \mathbf{V}^t\}_{t=1,\ldots,T}$. The number of modules is also fixed for all conditions.

This type of learning is called multi-task learning (Caruana, 1997) in the machine-learning community, where many related models are learned simultaneously using the same internal representation. Such learning allows different models (or cancer types) to share some parameters, which improves learning while at the same time it can also identify unique parameters for specific types. In several cases, such framework was shown to lead to better solutions (Caruana, 1997). Many existing methods proposed for multi-task learning focus on multi-output regression problems, where it is often desirable to obtain sparse solutions by performing covariate selection. They rely on regularization technique to jointly select a set of covariates that are relevant to many tasks. One can apply ℓ_1/ℓ_2 penalty of group lasso to select covariates relevant to all tasks (Obozinski *et al.*, 2010).

Here, we adopt the ℓ_1/ℓ_2 penalty of group lasso to regularize the modules over T conditions with the following penalty:

$$\lambda\left(\sum_{i,k}\sqrt{\sum_t (u_{ik}^t)^2} + \sum_{j,k}\sqrt{\sum_t (v_{jk}^t)^2}\right)$$

This penalty encourages entries $\{u_{ik}^t\}_{t=1,\ldots,T}$ and $\{v_{jk}^t\}_{t=1,\ldots,T}$ to be selected together, which means that miRNAs and mRNAs are assigned to the same modules across conditions. As the penalty is not differentiable at 0, we reformulate the optimization problem by moving the non-differentiable part to the constraints as suggested in (Liu *et al.*, 2009):

$$\begin{aligned}
&\min_{\mathbf{U}\geq 0, \mathbf{V}\geq 0, \mu, \Sigma, \{a_{ik}, b_{jk}\}} \quad \mathcal{F} + \lambda\left(\sum_{i,k} a_{ik} + \sum_{j,k} b_{jk}\right) \\
&\text{s.t.} \sqrt{\sum_t (u_{ik}^t)^2} \leq a_{ik}; \quad \sqrt{\sum_t (v_{ik}^t)^2} \leq b_{ik} \qquad (5)\\
&||\mathbf{u}_i||_1 \leq C_1; \quad ||\mathbf{v}_j||_1 \leq C_2 \\
&i = 1, \ldots, M; j = 1, \ldots, N; k = 1, \ldots K
\end{aligned}$$

Here, we have introduced new variables $\{a_{ik}\}$ and $\{b_{jk}\}$ into the problem. We update the projection step in Section 2.6 with the projection on the new ℓ_2 norm balls in the constraint set as shown in Liu *et al.* (2009) (Theorem 4).

4 RESULTS

4.1 MiRNA regulation in ovarian cancer

To test PIMiM and to compare it with previous methods for determining condition-specific miRNA regulation (SNMNMF and GroupMiR), we use the ovarian cancer dataset from Zhang *et al.* (2011). This dataset contains 385 samples from cancer patients, each measuring the expression of 559 miRNAs and 12 456 mRNAs and was downloaded from the Cancer Genome Atlas data portal (TCGA) (https://tcga-data.nci.nih.gov/tcga/). In addition to expression data, the sequence-based prediction of miRNA–mRNA interactions was downloaded from MicroCosm (Griffiths-Jones *et al.*, 2006), and protein interaction data were downloaded from TRANSFAC (Wingender *et al.*, 2000). We only use MicroCosm here to allow a fair comparison with SNMNMF, which only uses these data. In subsequent analysis, we use other sequence-based prediction methods as well. To evaluate the accuracy of each method, we used a set

of 115 cancer miRNAs that were determined to participate in ovarian cancer in a recent review article (Koturbash *et al.*, 2011; Tables 1 and 2). Using this set we compute the precision, recall and F1 score (the harmonic mean of precision and recall) of the set of miRNAs identified by each method.

The number of modules K was set to 50 for the non-negative matrix factorization method (SNMNMF) as suggested in Zhang *et al.* (2011). PIMiM also requires setting regularization and weight parameters α, β, C_1 and C_2. To set these, we performed an iterative line search (holding three of the four parameters fixed and adjusting the value of the fourth until convergence) to determine the values of these parameters using the F1 score as the target function to optimize. Based on this analysis, we selected $K = 40$ for PIMiM (see Supplementary Fig. S3 for details). SNMNMF was also run with the optimized set of parameters and input data described in Zhang *et al.* (2011). Unlike PIMiM and SNMNMF, GroupMiR uses a non-parametric Bayesian prior for the number of modules; therefore, this number cannot be fixed in advance. Thus, for GroupMiR, we report modules and interactions with posterior probability at least 0.3 to get a set of comparable size with other methods. Previously, GroupMiR was shown to outperform several other methods (Le and Bar-Joseph, 2011) including GenMiR++ (Huang *et al.*, 2007b); therefore, we omitted comparison with these methods here. Figure 2 presents a graphical view of the modules identified by PIMiM and SNMNMF. We color interaction edges between genes using different colors for each module. The modules identified by PIMiM are more dense and, hence, are in better agreement with previous findings regarding the regulation of interacting proteins by miRNAs.

4.1.1 Evaluation: identifying cancer miRNAs We first looked at the set of miRNAs identified by each method (those belonging to the modules returned by each of the methods). The results in Table 1 demonstrate that using the protein interaction data greatly increases precision, recall and the F1 score. Both methods that use these data (PIMiM and SNMMNF) clearly outperform GroupMiR on this set. In addition, using a regression model also helps as indicated by the increase in F1 score PIMiM obtains over SNMNMF.

4.1.2 Expression coherence In addition to analyzing the set of identified miRNAs, we also computed the average anti-correlation between miRNAs and mRNAs in the modules identified by each of the methods (Table 1). In this analysis, GroupMiR achieves the highest anti-correlation between miRNAs and the mRNAs they regulate in a module. This is the result of a much smaller module size identified by GroupMiR. As protein interactions are not used, mRNAs in these modules are selected because they are strongly anti-correlated with the miRNAs predicted to regulate the modules. This requirement leads to smaller modules and a better (anti) correlation between miRNAs and mRNAs. Still, PIMiM improves on SNMNMF in identifying anti-correlated miRNA–mRNA pairs. SNMNMF's objective function does not explicitly include a component for expression anti-correlation between miRNAs and mRNAs, which may explain why it does not capture the inhibitory role of miRNAs. Thus, PIMiM provides a useful compromise between relying strongly on protein interactions, which improves accuracy and using the observed expression values in a regression setting.

Table 1. Evaluation of all methods on the ovarian cancer dataset

F1 score	Cancer miRNAs			Expression correlation	Number of genes/module
	F1	Precision	Recall		
PIMiM	**0.3768**	**0.3230**	**0.4522**	−0.0131	67.80
SNMNMF	0.3588	0.3197	0.4087	0.0745	79.26
GroupMiR	0.1227	0.2083	0.0870	**−0.0408**	54.82

Note: The expression correlation values and number of genes are averaged across modules. Expression correlation: the correlation of expression values of miRNAs and mRNAs. Bold values are the best values for the column (highest or lowest depending on the context).

Table 2. miRNAs specifically identified for a cancer type

MiRNAs	Predicted type	BRCA	GBM	AML
hsa-miR-663	BRCA	Khoshnaw *et al.* (2009)	–	–
hsa-miR-433	GBM	–	Hua *et al.* (2012)	–
hsa-miR-99b	AML	–	–	Garzon *et al.* (2007)

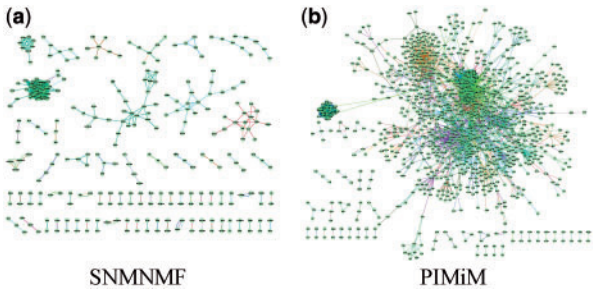

Fig. 2. Interactions between genes of the modules. We show an edge between two genes if they are members of a module and their interaction exists in the database. Each color corresponds to one module. Genes with no edges are omitted to improve visualization

4.1.3 MSigDB enrichment analysis To test the biological function of the modules, we used 880 gene sets of canonical pathways (C2-CP, v.3.0) from MSigDB (Subramanian *et al.*, 2005). We used the hypergeometric distribution to compute enrichment *P*-values for each of the modules with each of the MSigDB gene sets. To correct for the multiple hypothesis testings, we used the Benjamini–Hochberg procedure implemented in the R function p.adjust, which computes a q-value for each intersection. The results are presented in Figure 3, which depicts the number of modules with at least one enriched set in the

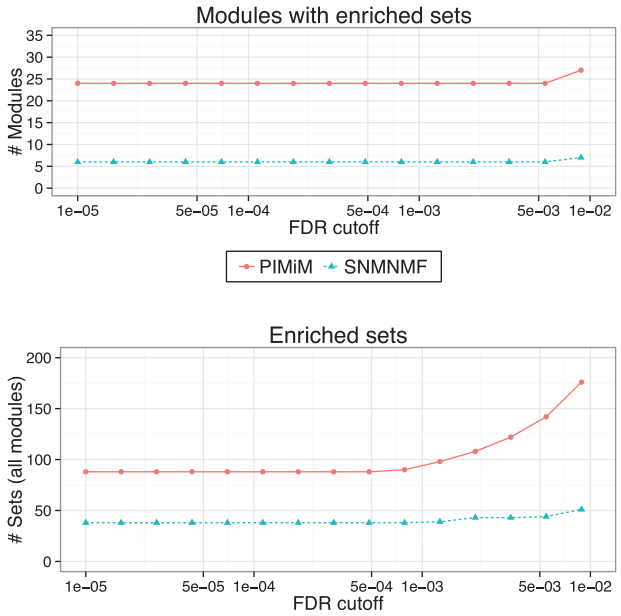

Fig. 3. MSigDB enrichment analysis: pathway enrichment analysis was done using 880 gene sets of canonical pathways (C2-CP) from MSigDB (Subramanian *et al.*, 2005). *P*-values were computed using hypergeometric test (with 10 000 random permutations) on the intersection of the set of genes in each module with MSigDB gene sets. Benjamini–Hochberg procedure was used to control the false discovery rate. Top: Number of modules significantly enriched for at least one MSigDB category for different significance cut-offs. Bottom: Number of MSigDB categories identified as in enriched in at least one of the modules for different significance cut-off

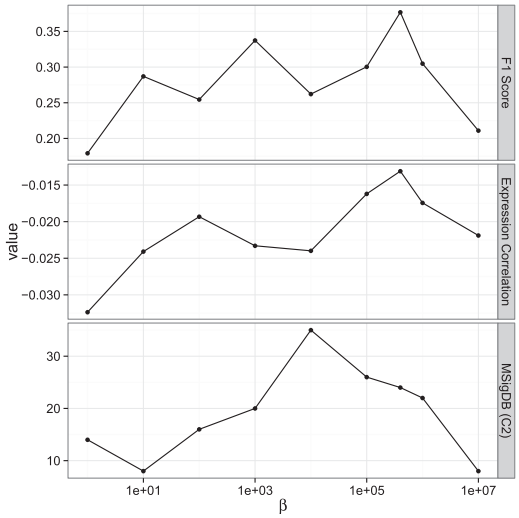

Fig. 4. The effect of protein interaction data to the result. We varied the value of β and tested the different metrics discussed in Section 4. As can be seen, both high and low values lead to reduced performance

MSigDB enrichment analysis and the total number of unique enriched gene sets. PIMiM outperforms SNMNMF, achieving both better enrichment for individual modules and better coverage of different MSigDB sets. MSigDB pathways are biased toward cancer pathways and so may be more relevant for the data we are analyzing here than Gene Ontology analysis. In addition to cancer hits, top hits for MSigDB include signatures for β cells that have been linked to cancer (Pelengaris and Khan, 2001) and several translation-related categories.

4.1.4 The effect of β on the performance of PIMiM To test the effects of using the protein interaction data in PIMiM, we re-run PIMiM with different β values. The results are presented in Figure 4. As the figure shows, when decreasing the value of β, the performance of PIMiM on all evaluation metrics decreases indicating the protein protein interactions (PPI) data are useful for identifying coherent modules. On the other hand, increasing β too much leads to high weight for PPI data at the expense of the expression information, which also negatively affects the performance of PIMiM. Thus, balancing the two data types, which is done by setting an intermediate value for β is key to the success of PIMiM.

4.2 Integrating data from multiple types of cancers

To further investigate miRNA control of different cancers, we applied PIMiM to a dataset of three cancer types using the multi-task learning framework described in Section 3. We learn three sets of modules for three types of cancer: breast invasive carcinoma (BRCA), Glioblastoma multiforme (GBM) and acute myeloid leukemia (AML). The miRNA and gene expression profiles of 89 BRCA, 498 GBM and 173 AML patients were downloaded from the TCGA. This set has 285 miRNAs and 10 922 mRNAs in common. Here, we combine the miRNA–mRNA predicted interactions from three public databases [MicroCosm (Griffiths-Jones *et al.*, 2006), miRanda (John *et al.*, 2004) and TargetScan (Lewis *et al.*, 2005)] and protein interaction data from TRANSFAC (Wingender *et al.*, 2000). For each cancer type, PIMiM learns 1 set of 50 modules. The parameters were set by optimizing for the F1 score of identifying miRNAs relevant to this dataset based on the set of cancer-related miRNAs from Koturbash *et al.* (2011). Figure 5 displays the miRNA-regulating modules in all three cancer types.

4.2.1 Analysis of identified miRNAs Several of the modules identified by PIMiM are regulated by known cancer miRNAs. The overall F1 score for cancer miRNAs for the joint analysis was high for all three cancer types: BRCA (0.6167), GBM (0.5789) and AML (0.6111). Well-known cancer miRNAs reported by PIMiM include the let-7b/c/d/e (active in BRCA: Yu *et al.*, 2007, GBM: Lee *et al.*, 2011 and AML: Jongen-Lavrencic *et al.*, 2008), mirR-302a/b/c/d cluster [suppression of the CDK2 and CDK4/6 cell cycle pathways (Lin *et al.*, 2010)] and miR-96 (active in BRCA: Guttilla and White, 2009, AML: Zhao *et al.*, 2010, miR-34a (active in BRCA: O'Day and Lal, 2010, GBM: Li *et al.*, 2009, AML: Zenz *et al.*, 2009, miR-15a/b (active in AML: Calin *et al.*, 2008). Some members of the miR-17-92 cluster (miR-18b, miR-19a, miR-20a/b and miR-93) are also identified by PIMiM (active in BRCA: Mendell, 2008, GBM: Ernst *et al.*, 2010, AML: Mi *et al.*, 2010). Note that some well-known cancer miRNAs, including miR-17 and miR-92, are missing from the modules because their expression is not available for enough of the samples. Several other subsets of miRNAs were assigned to

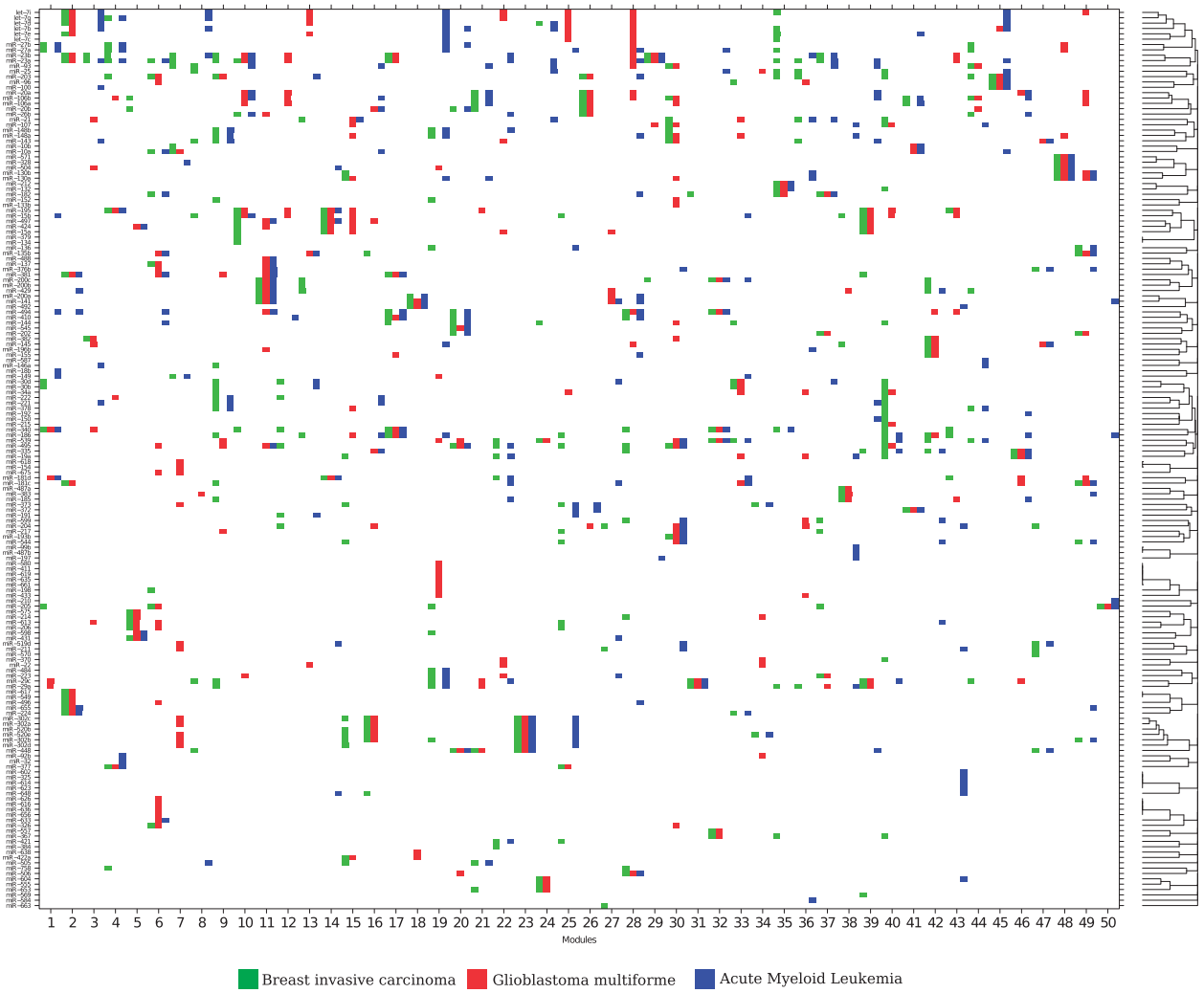

Fig. 5. Inferred miRNA modules of the three cancer types (BRCA, GBM and AML). The *x*-axis shows the 50 × 3 modules learned for the three cancer types (each *x*-axis bar is subdivided into three with the color corresponding to the cancer type). The *y*-axis shows miRNAs ordered by hierarchical clustering of their module membership vector. In several cases, the same miRNAs are predicted for all or two of the three cancer types

cooperatively regulate modules in multiple types of cancer as shown in Figure 5.

4.2.2 Cancer-specific miRNAs In addition to finding common cancer regulators, PIMiM can be used to identify cancer-type–specific regulators. These can either be used as biomarkers for a sub-type or can be studied to determine the unique properties of each cancer type. Although it is hard to obtain negative information (i.e. an article that mentions that a certain miRNA does not regulate a specific cancer type) several of the predictions made by PIMiM agree with current literature that, at least so far, only mentions their role in the cancer they were assigned to by PIMiM. Table 2 lists a few of these miRNAs and the cancer type they were predicted to regulate.

4.2.3 Analyzing the miRNAs and mRNAs in an identified module In addition to identifying important miRNAs for this particular study, PIMiM returns a set of modules providing predictions of cooperative regulation of miRNAs and their mRNAs targets. To demonstrate the informative power of this modular structure, we analyze in more details one of these modules (see also Supplementary Results for detailed discussion of other modules). Figure 6 depicts a network of miRNAs and mRNAs identified as part of Module 11. Across all cancer types, PIMiM identified a set of 14 strongly connected proteins. MiR-200a/b/c, miR-141 and miR-429 are predicted to regulate this set of mRNAs in all types of cancer. These miRNAs have previously been reported to play a role in cancer and cell proliferation (Korpal *et al.*, 2008; Peter, 2009). Interestingly, the miR-200 family is located in two chromosomal regions on 1p36.33 (200b, 200a and 429) and 12p13.31 (200c and 141), respectively (Uhlmann *et al.*, 2010), which may support our prediction of their cooperative regulation. Applying Gene Ontology analysis [using FuncAssociate (Berriz *et al.*, 2009)] and MSigDB enrichment analysis to the set of 14 mRNAs in this module indicates

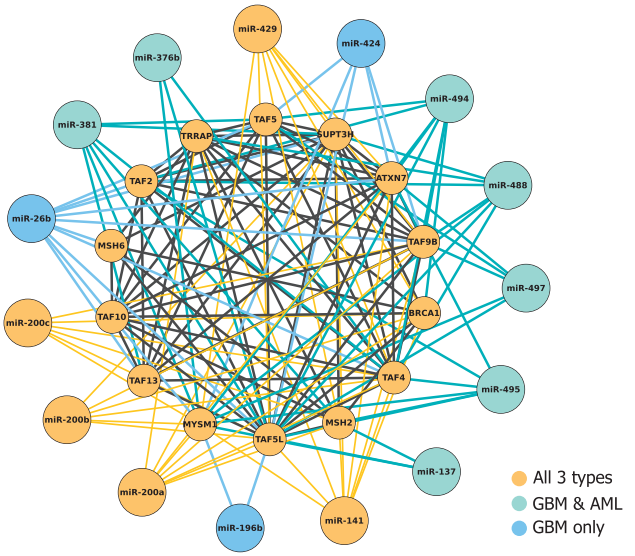

Fig. 6. miRNAs and mRNAs assigned to Module 11 in all three cancer types. Color indicates the specific cancer type for which the mRNA or miRNA was selected as part of the module

that this set is enriched with members of transcription factor TFTC/STAGA and TFFIID complexes. Recent findings support the link between these complexes and cancer (Kurabe *et al.*, 2007). This module also includes a tumor suppressor gene MSH2 (Wada-Hiraike *et al.*, 2005) and a famous breast cancer susceptibility gene BRCA1 (Miki *et al.*, 1994).

5 CONCLUSIONS

We presented PIMiM, a new method for inferring condition-specific regulation of miRNAs and for identifying their targets. PIMiM combines sequence, expression and interaction data to discover miRNA-regulated modules of mRNAs. We use a probabilistic model that combines regression with network information to discover these modules. We developed an iterative learning procedure to learn the parameters of our model and a multi-task learning method for combining data from multiple conditions.

We tested PIMiM on ovarian cancer expression data and have shown that it correctly identifies miRNAs regulating this cancer type, and that it is able to group relevant genes together. Comparison with other methods indicates that by using protein interaction data, we can improve accuracy while at the same time PIMiM also maintains expression coherence among mRNAs and anti-correlation between miRNAs and the mRNAs they are predicted to regulate improving on previous methods that have also used protein interaction data. Application of the method to compare and contrast three types of cancer identified both common and unique regulators, which can allow researchers to determine the core cancer regulatory network and the differences in regulation among the various cancers we studied.

Although we believe PIMiM can already be of use to researchers that collect mRNA and miRNA expression data, there are a number of extensions that can further improve it. As aforementioned, we follow several other articles in isolating the miRNA target prediction task from the combinatorial analysis of miRNA–TF regulation. Although such an approach leads to good results as discussed earlier in the text, our longer term goal is to develop a method that can incorporate both types of regulation in a single-modeling framework. For this, we would need to determine the role a specific TF plays (activator or repressor) and its activity level [either based on its expression levels or on the set of its targets (Shi *et al.*, 2009)]. With this information, we can incorporate TFs into our regression model to account for their part in regulating expression, which will hopefully lead to better results regarding the role played by specific miRNAs. The regression component that we considered in PIMiM uses a simple linear model to explain the regulation effect of multiple miRNAs. We could also extend this to incorporate other complex combinatorial analysis (for example, AND, OR logic). However, this requires an extension of the methods derived in the article. We thus leave this for future work. In addition, we would like to incorporate additional types of high-throughput data, for example, epigenetic data to our analysis framework.

Funding: National Institutes of Health (U01 HL108642) and National Science Foundation (DBI-0965316) award (to Z.B.J.).

Conflict of Interest: none declared.

REFERENCES

Barakat,A. *et al.* (2007) Conservation and divergence of microRNAs in populus. *BMC Genomics*, **8**, 481.

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

Berriz,G.F. *et al.* (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043–3044.

Betel,D. *et al.* (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.

Calin,G.A. *et al.* (2008) miR-15a and miR-16-1 cluster functions in human leukemia. *Proc. Natl Acad. Sci. USA*, **105**, 5166–5171.

Caruana,R. (1997) Multitask learning. *Mach. Learn.*, **28**, 41–75.

Cheng,C. and Li,L.M. (2008) Inferring microRNA activities by combining gene expression with microRNA target prediction. *PLoS One*, **3**, 1989.

Ernst,A. *et al.* (2010) De-repression of CTGF via the miR-17-92 cluster upon differentiation of human glioblastoma spheroid cultures. *Oncogene*, **29**, 3411–3422.

Garzon,R. *et al.* (2007) MicroRNA gene expression during retinoic acid-induced differentiation of human acute promyelocytic leukemia. *Oncogene*, **26**, 4148–4157.

Griffiths,T. and Ghahramani,Z. (2006) Infinite latent feature models and the Indian buffet process. *Adv. Neural Inform. Process. Syst.*, **18**, 475.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34(Suppl. 1)**, D140–D144.

Guttilla,I.K. and White,B.A. (2009) Coordinate regulation of foxo1 by miR-27a, miR-96, and miR-182 in breast cancer cells. *J. Biol. Chem.*, **284**, 23204–23216.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.

Hsu,C.W. *et al.* (2008) Characterization of microRNA-regulated protein-protein interaction network. *Proteomics*, **8**, 1975–1979.

Hua,D. *et al.* (2012) A catalogue of glioblastoma and brain microRNAs identified by deep sequencing. *OMICS*, **16**, 690–699.

Huang,J.C. *et al.* (2007a) Using expression profiling data to identify human microRNA targets. *Nat. Methods*, **4**, 1045–1049.

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

Huang,G.T. *et al.* (2011) mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res.*, **39(Suppl. 2)**, W416–W423.

John,B. *et al.* (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
Jongen-Lavrencic,M. *et al.* (2008) MicroRNA expression profiling in relation to the genetic heterogeneity of acute myeloid leukemia. *Blood*, **111**, 5078–5085.
Joung,J.G. *et al.* (2007) Discovery of microRNA–mRNA modules via population-based probabilistic learning. *Bioinformatics*, **23**, 1141.
Khoshnaw,S.M. *et al.* (2009) MicroRNA involvement in the pathogenesis and management of breast cancer. *J. Clin. Pathol.*, **62**, 422–428.
Korpal,M. *et al.* (2008) The mir-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. *J. Biol. Chem.*, **283**, 14910–14914.
Koturbash,I. *et al.* (2011) Small molecules with big effects: the role of the microRNAome in cancer and carcinogenesis. *Mutat. Res.*, **722**, 94–105.
Krek,A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
Krol,J. *et al.* (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
Kurabe,N. *et al.* (2007) Deregulated expression of a novel component of TFTC/STAGA histone acetyltransferase complexes, rat SGF29, in hepatocellular carcinoma: possible implication for the oncogenic potential of c-Myc. *Oncogene*, **26**, 5626–5634.
Le,H.-S.P. 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 24*, Curran Associates, Inc. pp. 235–243.
Lee,S.T. *et al.* (2011) Let-7 microRNA inhibits the proliferation of human glioblastoma cells. *J. Neurooncol.*, **102**, 19–24.
Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
Li,Y. *et al.* (2009) Microrna-34a inhibits glioblastoma growth by targeting multiple oncogenes. *Cancer Res.*, **69**, 7569–7576.
Liang,H. and Li,W.H. (2007) MicroRNA regulation of human protein–protein interaction network. *RNA*, **13**, 1402.
Lin,S.L. *et al.* (2010) MicroRNA miR-302 inhibits the tumorigenecity of human pluripotent stem cells by coordinate suppression of the CDK2 and CDK4/6 cell cycle pathways. *Cancer Res.*, **70**, 9473–9482.
Liu,J. *et al.* (2009) Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 339–348.
Malumbres,M. (2012) miRNAs versus oncogenes: the power of social networking. *Mol. Syst. Biol.*, **8**, 569.
Mendell,J.T. (2008) miRiad roles for the mir-17-92 cluster in development and disease. *Cell*, **133**, 217–222.
Meola,N. *et al.* (2009) microRNAs and genetic diseases. *Pathogenetics*, **2**, 7.
Mi,S. *et al.* (2010) Aberrant overexpression and function of the miR-17-92 cluster in MLL-rearranged acute leukemia. *Proc. Natl Acad. Sci. USA*, **107**, 3710–3715.
Miki,Y. *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 7.
Motameny,S. *et al.* (2010) Next generation sequencing of miRNAs–strategies, resources and methods. *Genes*, **1**, 70–84.
Muniategui,A. *et al.* (2012) Joint analysis of miRNA and mRNA expression data. *Brief. Bioinformatics*, [Epub ahead of print, doi: 10.1093/bib/bbs028, June 12, 2012].
Obozinski,G. *et al.* (2010) Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.*, **20**, 231–252.
O'Day,E. and Lal,A. (2010) MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Res.*, **12**, 201.
Ooi,C.H. *et al.* (2011) A densely interconnected genome-wide network of microRNAs and oncogenic pathways revealed using gene expression signatures. *PLoS Genet.*, **7**, e1002415.
Pelengaris,S. and Khan,M. (2001) Oncogenic co-operation in beta-cell tumorigenesis. *Endocr. Relat. Cancer*, **8**, 307–314.
Peter,M.E. (2009) Let-7 and mir-200 microRNAs: guardians against pluripotency and cancer progression. *Cell Cycle*, **8**, 843–852.
Sass,S. *et al.* (2011) MicroRNAs coordinately regulate protein complexes. *BMC Syst. Biol.*, **5**, 136.
Schmidt,M. *et al.* (2009) Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, MIT press, pp. 456–463.
Shao,N.Y. *et al.* (2010) Comprehensive survey of human brain microRNA by deep sequencing. *BMC Genomics*, **11**, 409.
Shi,Y. *et al.* (2009) A combined expression-interaction model for inferring the temporal activity of transcription factors. *J. Comput. Biol.*, **16**, 1035–1049.
Stark,C. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39(Suppl. 1)**, D698–D704.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Sun,J. *et al.* (2012) Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput. Biol.*, **8**, e1002488.
Uhlmann,S. *et al.* (2010) miR-200bc/429 cluster targets plcγ1 and differentially regulates proliferation and EGF-driven invasion than miR-200a/141 in breast cancer. *Oncogene*, **29**, 4297–4306.
Wada-Hiraike,O. *et al.* (2005) The DNA mismatch repair gene hmsh2 is a potent coactivator of oestrogen receptor α. *Br. J. Cancer*, **92**, 2286–2291.
Wang,H. and Li,W.H. (2009) Increasing MicroRNA target prediction confidence by the relative R2 method. *J. Theor. Biol.*, **259**, 793–798.
Wingender,E. *et al.* (2000) Transfac: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
Yu,F. *et al.* (2007) let-7 regulates self renewal and tumorigenicity of breast cancer cells. *Cell*, **131**, 1109–1123.
Zenz,T. *et al.* (2009) miR-34a as part of the resistance network in chronic lymphocytic leukemia. *Blood*, **113**, 3801–3808.
Zhang,S. *et al.* (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**, i401.
Zhao,H. *et al.* (2010) MicroRNA and leukemia: tiny molecule, great function. *Crit. Rev. Oncol Hematol.*, **74**, 149–155.