Briefings in Bioinformatics Advance Access published June 12, 2012

doi:10.1093/bib/bbs028

Joint analysis of miRNA and mRNA expression data

Ander Muniategui, Jon Pey, Francisco Planes and Angel Rubio

Submitted: 21st February 2012; Received (in revised form): 6th May 2012

Abstract

miRNAs are small RNA molecules (′22 nt) that interact with their target mRNAs inhibiting translation or/and cleavaging the target mRNA. This interaction is guided by sequence complentarity and results in the reduction of mRNA and/or protein levels. miRNAs are involved in key biological processes and different diseases. Therefore, deciphering miRNA targets is crucial for diagnostics and therapeutics. However, miRNA regulatory mechanisms are complex and there is still no high-throughput and low-cost miRNA target screening technique. In recent years, several computational methods based on sequence complementarity of the miRNA and the mRNAs have been developed. However, the predicted interactions using these computational methods are inconsistent and the expected false positive rates are still large. Recently, it has been proposed to use the expression values of miRNAs and mRNAs (and/or proteins) to refine the results of sequence-based putative targets for a particular experiment. These methods have shown to be effective identifying the most prominent interactions from the databases of putative targets. Here, we review these methods that combine both expression and sequence-based putative targets to predict miRNA targets.

Keywords: *miRNA; miRNA target prediction; miRNA mRNA integration; miRNA mRNA expression*

INTRODUCTION

miRNAs are small RNA molecules (′22 nt) that have been shown to be one of the key regulators of the expression of their mRNA targets in many metazoans and plants. Currently, the number of miRNA sequences annotated in miRBase—*de facto* database for miRNAs—is over 17 000 in 140 species, including 1400 human miRNA sequences (release 17, miRBase) [1]. microRNAs are processed from double-stranded hairpin precursors by Drosha protein in the nucleus and by Dicer protein in the cytoplasm [2, 3]. The final single-stranded mature microRNA hybridizes with the RNA-induced silencing complex (RISC) to undergo gene inhibition. Although there are exceptions of miRNAs that show up-regulation effects on mRNA expression [4], the general statement is that miRNAs act to repress the expression of their targets [5–7]. Gene regulation by RISC complex is guided by sequence complementarity between the 'seed region' (nucleotides 2–7 and 8) of the microRNA and the 3′-UTR of the mRNA [8]. In plants, perfect base pairing of miRNA and mRNA leads to mRNA degradation and the subsequent reduction of both mRNA and

Corresponding author. Angel Rubio, Group of Bioinformatics, CEIT and TECNUN, University of Navarra, San Sebastian, Spain. Tel: +34 943 21 28 00; Fax: +34 943 21 30 76; E-mail: arubio@ceit.es

Ander Muniategui obtained his BS in Industrial and Materials Engineering in 2007 and 2008, respectively, and his Master's degree in Biomedical Engineering in 2010. He is currently completing his doctoral studies at CEIT in the area of Bioinformatics.

Jon Pey received his MSc in Engineering in 2009. In 2009, he joined CEIT where he is doing his PhD studies. His research interest field is focused on the Elementary Flux Modes and their application in the flux calculation, coupling this methodology with other omics (proteomics, gene expression, carbon labeling...).

Francis Planes received his PhD in Bioinformatics at Brunel University, UK, under the supervision of Professor John E. Beasley in 2008. He was awarded the Vice-Chancellor's Prize for Doctoral Research at Brunel University. In 2009, he joined the Bioinformatics group at CEIT. His research is devoted to the analysis of topology, structure, dynamics and regulation of metabolic networks and related applications.

Angel Rubio is a researcher at CEIT since 1995. He is Head of the Bioinformatics Group. His interest field is focused on alternative splicing, analysis of copy number alterations, and joint analysis of disparate sources of data (miRNA and mRNA expression, copy number and expression, etc.). His expertise includes genomic and proteomic data analysis, statistical treatment of data and systems modeling.

protein levels. On the contrary, in animals, base pairing is not perfect and miRNAs act to degrade and/or translationally inhibit the mRNA. Although determining the predominant mechanism of miRNA regulation has undergone extensive debate [9, 10], it has been recently shown that animal miRNAs mainly act to degrade mRNA targets [11]. miRNA regulation is only one of the many regulatory mechanisms of mRNA expression. Other mechanisms such as alternative splicing, polyadenylation and regulation of transcription [12] do also take place. miRNAs are generally assumed to be fine-tuners of protein expression.

Each miRNA is potentially able to regulate around 100 or more mRNA targets and 30% of all human genes are supposed to be regulated by miRNAs [6, 13]. miRNAs are involved in key biological processes, such as development, differentiation, apoptosis and proliferation [14, 15]. Furthermore, alterations in their regulatory pathways can cause different diseases such as cancer, neurodegenerative (Alzheimer, schizophrenia), cardiovascular diseases and metabolic disorders [16–20]. Therefore, identification and validation of miRNA–mRNA targets is essential since unveiling their regulation network may lead to new therapeutic targets [19, 21, 22]. Unfortunately, the validation of putative miRNA–mRNA interactions is not straightforward (see Figure 1a–d brief summaries of the respective following paragraphs are showed).

Experimental methods

The most extended experimental technique for determining miRNA targets (Figure 1a) is the transfection of mimic miRNAs or miRNA inhibitors (i.e. anti-miRs [23], antagomiRs [24] and miRNA sponges [25]). The effects on the expression levels of the mRNAs and proteins are measured by using transcriptomic and proteomic tools (i.e. qRT-PCR, microarrays, RNA-seq, western blot, SILAC, 2D-DIGE). However, with this technique it is not possible to distinguish indirect and direct interactions. Adding reporters or labels to miRNAs or the 3′-UTR of transcripts of interest during transfection focus the experiment on direct interactions as done in LAMP or luciferase report assays. Other direct methods for miRNA target prediction are based on the immunoprecipitation of RISC complexes such us Argonaute bound miRNA–mRNA molecules (i.e. HITS-CLIP [26] and PAR-CLIP [27]). There are other alternative experimental techniques, such as the detection of the 5′-UTR of the degraded mRNAs with 5′ RLM RACE, the detection of mRNAs undergoing translation by using polysome profiling or the detection of mRNAs by biotin labeled miRNAs. Each experimental technique has its own reliability (i.e. direct measurement of miRNA binding sites by HITS-CLIPS and PAR-CLIP makes the results more reliable than an indirect measurement from the analysis of expression patterns). Due to this, combining different experimental tools is a good method to ensure the authenticity of a miRNA target [28]. For a review of experimental methods for miRNA target prediction, see Refs [29–32]. The outputs of these experiments are a set of experimentally validated interactions (see Figure 1c, further explained in a following section).

Sequence-based methods

Despite the wide range of experimental tools for miRNA target validation available, the lack of high-throughput and low-cost methods has enforced the development of computational techniques (Figure 1b). These are based on experimentally determined rules of miRNA targeting: (i) sequence complementarity between the 3′-UTR of the mRNAs and the 'seed region' of the miRNA (nucleotides 2–7), (ii) possible functional target sites along the coding sequence and 5′-UTR of the mRNA, (iii) conservation of some of the miRNA target sites between related species and (iv) the target site accessibility due to the RNA secondary structure (i.e. free energy costs to unfold the mRNA secondary structure surrounding the target site and free energy of the miRNA-target pairing) [33–35]. Although the methods that use these rules are far from perfect, the putative lists of targets generated by computational methods entangle a considerable reduction of experimental work as they significantly reduce the number of interactions that must undergo validation.

Databases of interactions

The lists of miRNA targets predicted by experimental and computational tools have been included in several databases (Figure 1c). The most cited computationally-based databases are miRanda [36], miRBase [37], TargetScan [38–40], DIANAmicroT [41] and PicTar [42]. Computational methods predict hundreds of thousands target mRNAs per miRNA [43]. According to several studies, the

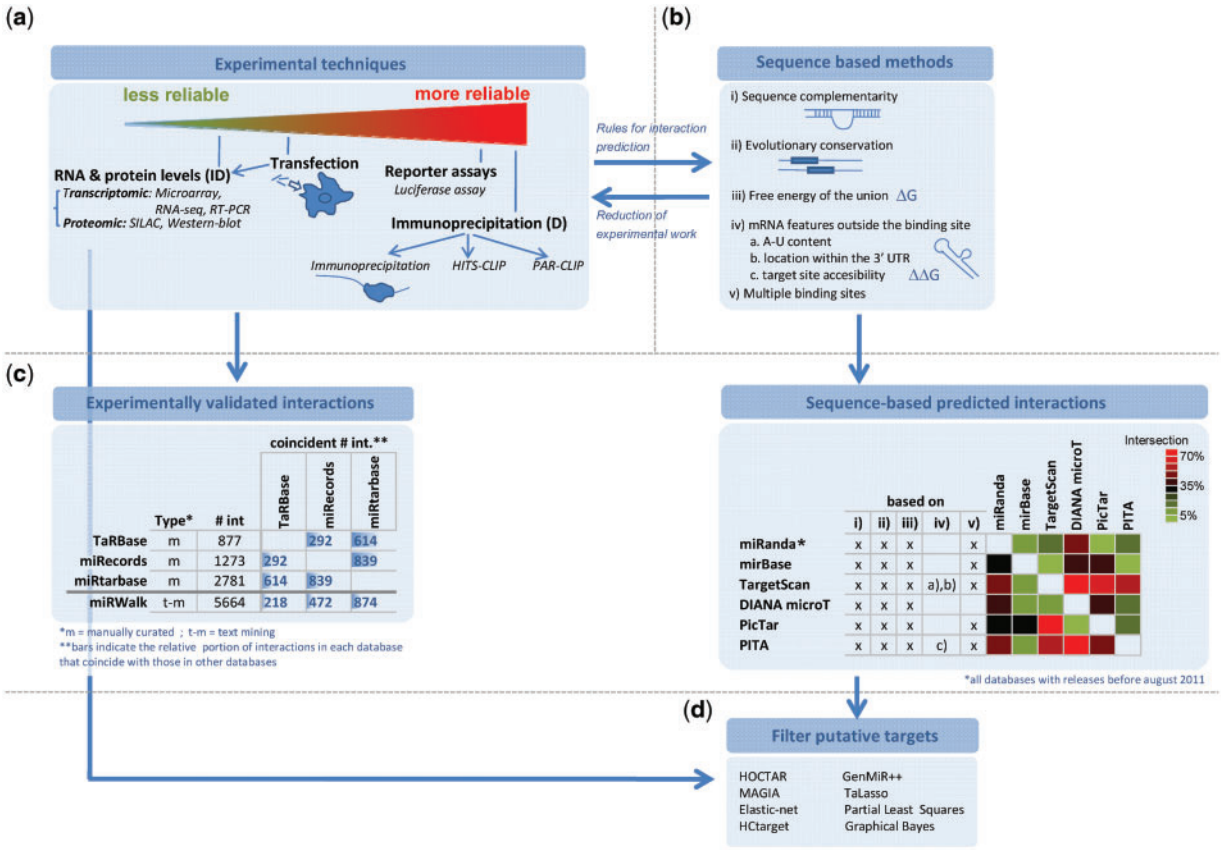

Figure 1: Combining experimental and computational tools for deciphering miRNA functions and targets. Several computational and experimental tools for the identification and validation of miRNA targets have emerged in the last years. Although the wide range of experimental tools for miRNA target validation available (**a**), the lack of high-throughput and low-cost methods enforces the use of computational techniques (**b**). These are based in experimentally determined rules of miRNA targeting. Putative lists of targets generated by computational methods entangle a considerable reduction of work since the number of experiments to carry out is greatly diminished (**c**). Further reduction of the number of putative miRNA targets is achieved combining experimental data and sequence-based predictions (**d**).

estimated false positive rate of these predictions ranges from 24% to 70% [44, 45]. A consequence of this is the lack of concordance between the predictions of different databases (as shown in Figure 1). A comparison of these databases and different combinations of them can be found in Ref. [44]. On the other hand, the most important databases that include experimentally validated targets are TarBase [46], miRecords [47], miRtarbase [48], miRWalk [49] and miRNAMAP [50]. Among these, TarBase and miRecords include manually curated experimental interactions, while miRWalk and miRNAMAP use text mining tools. miRtarbase database uses text mining techniques and manual curation and it includes most of the interactions on TarBase and MiRecords. In TaRBase, miRecords and miRtarbase databases the experimental techniques used for the validation of each miRNA target is also included. As mentioned before, each experimental technique has its reliability and thus, the addition of this information gives a confidence level to each interaction. Compared to the hundreds of thousands of putative targets predicted by sequence-based algorithms, the number of experimentally validated targets is very low. For instance, TarBase includes 1300 experimentally validated targets in humans.

Combination of experimental data and sequence-based predictions

The reliable prediction of miRNA targets is still a challenge. One appealing possibility to accomplish

this task is to combine high-throughput experimental data together with sequence-based putative predictions to improve the reliability of the predictions in a particular experiment [51–59] (Figure 1d). In this review, we focus on computational methods that combine sequence-based interactions and miRNA and mRNA expression data so as to filter putative lists of miRNA targets.

TARGET PREDICTION BASED ON EXPRESSION DATA AND SEQUENCE-BASED PUTATIVE INTERACTIONS

Computational models able to predict the most outstanding miRNA–mRNA interactions combine information from three different sources as depicted in Figure 2: mRNA (or protein) expression, miRNA expression and putative interactions. The mathematical tools used by these models move from simple Correlation analysis to more complex Bayesian inference methods (Table 1). Most of these models assume that mRNAs are repressed by miRNAs. Nevertheless, there are some studies, based on correlation analysis, that also consider possible enhancement effects from miRNAs.

Henceforth, matrices $\mathbf{X_{JxT}} = [x_{jt}]$ and $\mathbf{Z_{KxT}} = [z_{kt}]$ denote the expression values of mRNAs j ($j = 1, \ldots, J$) and miRNAs k ($k = 1, \ldots, K$) in sample t ($t = 1, \ldots, T$), respectively. The set of putative targets is represented by a binary matrix $\mathbf{C_{JxK}} = [c_{jk}]$, where c_{jk} is 1 if the pair mRNA j–miRNA k has been putatively predicted from a sequence-based method, 0 otherwise. For the sake of simplicity, we also denote the expression across samples as the row vectors $\mathbf{x_j} = [x_{j1}, \ldots, x_{jT}]$ and $\mathbf{z_k} = [z_{k1}, \ldots, z_{kT}]$. Observe that mRNA and miRNA expression data as assumed to be sample matched. The output of any method described here is a scored version of the matrix C of putative targets (C′ in Figure 2).

Table 1 summarizes methods based on **X**, **Z** and **C** matrices previously introduced. We describe below each of these methods in detail. In this review, we have not studied the methods that using **X** and **C** predict **Z**, the activity of miRNAs. Most of them use enrichment analysis methods borrowed from GO enrichment analysis (i.e. Genecodis [60, 61], GSEA [62, 63], MMIA [64], etc.).

Correlation and mutual information

A straightforward method to analyze the relationship between miRNAs and mRNAs is the Pearson correlation. It is a measure of linear-dependency and is mathematically represented as a scalar product, as shown in Equation (1). The notation $\mu_j^x = 0$ and $\sigma_j^x = 1$ indicates that the expression data for mRNA j is standardized. Similarly, $\mu_j^z = 0$ and $\sigma_j^z = 1$ indicates that expression of miRNA k is standardized.

$$\rho_{jk} = (\mathbf{x_j})^T_{\mu_j=0|\sigma_j=1} \cdot (\mathbf{z_k})_{\mu_k=0|\sigma_k=1} = \left(\frac{\mathbf{x_j} - \mu_j^x}{\sigma_j^x}\right)^T \cdot \left(\frac{\mathbf{z_k} - \mu_k^z}{\sigma_k^z}\right). \quad (1)$$

Due to its simplicity and intuitive interpretation, Pearson correlation is widely used [65–67]. Only those putative miRNA–mRNA pairs that show a statistically significant correlation values are

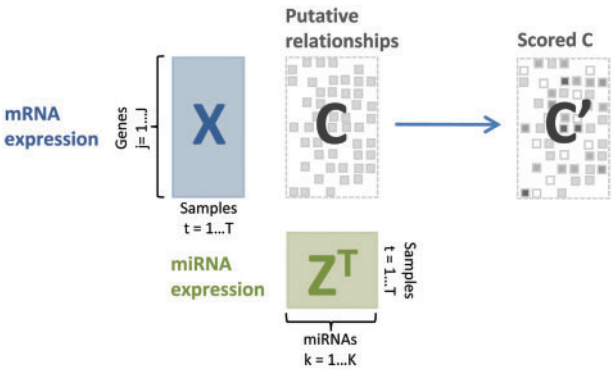

Figure 2: Scheme representing how sequence-based putative targets are scored by using mRNA and miRNA expression data. Matrices X, Z and C represent mRNA and miRNA expression data and putative interaction matrices, respectively. C is a binary matrix, which takes 1 is the corresponding pair is included in the considered database and 0 otherwise. A stringent set of targets can be used by combining C matrices of different databases. The output is C′, a matrix that ranks the putative interactions using the information from X and Z.

Table 1: Different methods in the literature to filter sequence-based interactions by using expression data

Method	References	Description	Input data					Result
Correlation								
Pearson correlation	[47,48,55]	The correlation coeficient is calculated for each of the miRNA-mRNA pairs. Only putative relationships are further considered.	X	Z	C	<0		$\omega_{jk} = (x_j - \mu_j)\cdot(z_k - \mu_k)^T/(\sigma_j \cdot \sigma_k)$
Spearman correlation	[48]	Non-parametric measure of correlation. It is computed from the ranks of the values.	X	Z	C	<0		
Mutual information	[48]	A generalized measure of dependency that extends Correlation from linear dependencies to any type of functional relationship. It cannot distinguish between positive and negative regulations.	X	Z	C			
Linear regression								
MLR	[56]	A multiple linear regression model. Only valid in case the number of putative miRNA regulators of a mRNA is lower than that of samples.	X	Z				$\omega_j = x_j \cdot Z^T \cdot (Z \cdot Z^T)^{-1}$
MLR + R^2 method	[57]	The model solves a multiple linear regression between each mRNA and its miRNA regulators and assigns a confidence value by using R^2 statistics.	X	Z	C			
Partial LS	[58]	Alternative method to linear regression used if the number of predictors is larger than the observations or in case of collinearity of the data. Results of combining X and Z are validated with C. Therefore C is used only for validation.	X	Z				
Regularized LS								
Lasso regression (TaLasso)	[59,60]	Least Squares with norm-1 regularization. Provides a sparse solution, i.e. a relatively small set of outstanding interactions.	X	Z	C	<0		$\omega_j = x_j \cdot Z^T \cdot (Z \cdot Z^T + \nu \cdot I)^{-1}$
Ridge regression[a]		Least Squares with norm-2 regularization. It has an explicit solution.	X	Z	C			
Elastic net	[51]	It uses both norm-1 and norm-2 regularizations. It can be reduced to Lasso and Ridge regressions by making the regularization parameters of norm-2 or norm-1 equal to zero.	X	Z	C			
Bayesian inference	[46]	(i) GenMiR++ is based on the expected inverse relationship among the expressions of the miRNAs and their targets.	X	Z	C	<0		
	[61]	(ii) GenMiR3 is a newer release of GenMiR++ that accounts for sequence-based information.	X	Z	C	R[b]	<0	
	[50]	(iii) a Bayesian Inference model with miRNA and protein expression levels.	X	Z		P	<0	
	[62]	(iv) HCtarget is a variation of GenMiR++ with modified priors over some of the parameters.	X	Z	C		<0	
	[63]	(v) A Bayesian Graphical method. Here authors restrict miRNA interaction search to down-regulation effects, allows adding weights (representing the confidence) to the sequence based putative predictions and forces the sparsity of the solution.	X	Z	C	R	<0	

Matrices X, Z, P and C represent expression matrices of mRNA, miRNA and proteins and the matrix of putative miRNA–mRNA interactions, respectively. The <0 indicates that the model includes a restriction so that only down-regulation effects from miRNAs are considered. This can be done: (i) by selecting the negative results or (ii) by imposing to the model to search only negative regressors (i.e. adding non-positive constraints). [a]To our knowledge, there is no computational method that uses Ridge regression for miRNA target prediction. However, as shown in the next section, Ridge regression and GenMiR++ are strongly related and thus we have included this method in the table. [b]GenMiR3 does not account for the scores of the sequence-based putative interactions. Alternatively, it considers other sequence-based information: total hybridization energy, context score and PhastCons score [61]. R = reliability of putative targets: scores from sequence-based databases or sequence features.

considered for further experimental validation. However, the number of significant miRNA–mRNA pairs can be too high to undergo posterior experimental validation. For this reason, approaches using correlation analysis usually include other constraints: differential expression [58, 68], sequence-based complementarity (i.e. putative targets) [53, 67, 69] and other biological information (i.e. conservation of target sites) [67]. Note here that these methods typically consider only negative correlation values. However, studies do exist considering positive correlations since some miRNAs may act as transcription factors [67, 70, 71].

There are several web-based tools and databases that use correlation for miRNA–mRNA target research [52, 53, 72, 73]. One outstanding example of these methods is HOCTAR [52]. HOCTAR was developed to determine mRNA targets of intragenic miRNAs [52, 74]. The expression values of host genes and intragenic miRNAs are strongly related, thus, the expression of host genes can be used to estimate the expression of intragenic miRNAs. Since the number of samples with available mRNA expression is large (this study includes 3445 arrays), the correlation can be highly significant even for small values.

In some situations, particularly when the underlying relationship is not linear or in the presence of outliers, Spearman correlation outperforms the Pearson correlation. Several web tools provide it as an alternative to correlation [53].

A different measure of independence of variables is mutual information (MI). While correlation values can distinguish the sign of the miRNA–mRNA relationship, MI only indicates whether (or not) two given variables are independent. These three measures (Pearson, Spearman and MI) were integrated in the web-based tool called MAGIA [53].

Multiple linear regression

Multiple linear regression (MLR) evaluates the relationship between the complete set of miRNA regulators and a target mRNA at the same time, in contrast to correlation techniques, which focuses on particular interactions. Some authors have used MLR for miRNA target prediction: by only considering expression data [75] and by combining both expression and sequence-based data [76]. In Ref. [76], the *R*-squared statistics is used for measuring the goodness of fit of the data.

An ordinary MLR model for mRNA j and its K miRNA putative regulators can be formulated as follows:

$$\mathbf{x_j} = \sum_{k=0}^{K^j} \omega_{jk} \cdot \mathbf{z_k} + \varepsilon_\mathbf{j} = \boldsymbol{\omega}_\mathbf{j} \cdot \mathbf{Z^j} + \varepsilon_\mathbf{j} \quad (2)$$

where $\boldsymbol{\omega}_\mathbf{j} = \{\omega_{jk}\} = [\omega_{j0}, \omega_{j1}, \ldots, \omega_{jK}]$ is the vector of regulatory weights and $\varepsilon_\mathbf{j}$ is an error term. For simplicity of notation, here and in the following, K^j will represent the number of putative miRNA regulators of mRNA j (those with $c_{jk} = 1$). In the matricial form, $\mathbf{Z}$ is a matrix of size $(K^j + 1) \times T$ of miRNA expression, in where, $\boldsymbol{z_0} = [1, 1, \ldots, 1]^T$ has been added to $\mathbf{Z}$ to account for the intercept, ω_{j0}.

The explicit solution of a MLR after applying least squares is given by:

$$\omega_\mathbf{j} = \mathbf{x_j} \cdot \left(\mathbf{Z^j}\right)^\mathbf{T} \cdot \left[\left(\mathbf{Z^j}\right)^\mathbf{T} \cdot \left(\mathbf{Z^j}\right)\right]^{-1} \quad (3)$$

If data is highly correlated—collinearity—or the number of samples is smaller than that of miRNA regulators, the matrix $(\mathbf{Z^j})^\mathbf{T} \cdot \mathbf{Z^j}$ can be singular and provokes instabilities in the solution. Since the number of putative miRNA regulators of an mRNA is usually larger than that of samples, plain MLR cannot be applied as a general purpose method to find miRNA–mRNA relationships and other alternatives must be considered.

Partial least squares

If the number of samples with available expression data is smaller than the number of covariates (miRNAs) the linear model is undetermined, the computed solution is not unique and therefore plain MLR cannot be applied. An extension of MLR suited for these cases and for possible collinear covariates are the partial least squares (PLS). A PLS model extracts the main miRNAs that explain the maximum variance in the mRNA expression by ensuring a good fit of the underlying model. This method is applied for miRNA target prediction in Ref. [77]. There, the authors determine putative miRNA–mRNA interactions by only considering expression data from both miRNAs and mRNAs. In their method, they reduce the number of possible miRNA–mRNA pairs by selecting differentially expressed miRNAs and mRNAs, do bootstrap for statistical analysis of the results and determine the validity of the method by comparing their predicted interactions with those included in sequence-based databases (TargetScan and miRanda). Although

putative interactions are difficult to integrate with standard PLS, there are some adaptations that can include this information and use the information in sequence-based databases.

Regularized least squares

An alternative method to deal with undetermined linear systems is regularization. In regularized least squares, aside from minimizing the error, there is an extra term that forces the coefficients of the solution to be somehow small. This approach can be formulated as the optimization problem,

$$\min\left\{ \left\| \mathbf{x_j} - \omega_j \cdot \mathbf{Z}^j \right\|_2^2 + \nu \cdot R(\omega_j) \right\}, \quad (4)$$

where $R(\boldsymbol{\omega_j})$ indicates the regularization of $\boldsymbol{\omega_j}$ and ν is a tuning parameter that controls the degree of the regularization. The most common regularizations are the norm-1 ($|\boldsymbol{\omega_j}|_1$, LASSO regression), the norm-2 [$|\boldsymbol{\omega_j}|_2$, Ridge regression (RR)] and a combination of both ($\nu_1 \cdot |\boldsymbol{\omega_j}|_1 + \nu_2 \cdot |\boldsymbol{\omega_j}|_2$, elastic-net). Norm-1 regularization fosters the number of non-zero covariates mRNA–miRNAs interactions to be small. On the other hand, norm-2 provides a solution in which all the coefficients are small but not null. Norm-1 and elastic-net regularizations have been used for miRNA target predictions in Refs [78] and [56], respectively. Since miRNA usually down-regulates the target mRNA, it is sensible to add the restriction of negative relationships to the optimization shown in Equation (4). In Ref. [79], we showed that adding non-positivity constraints to LASSO regression provides more experimentally validated interactions and better biological interpretation.

Bayesian inference

Bayesian inference refers to statistical models that use a priori information to estimate parameters and predict values in a probability framework. We found four methods based on Bayesian techniques for miRNA target prediction [51, 55, 80, 81]. Three of them use mRNA and miRNA expression data [51, 80, 81] while the other uses miRNA and protein expression values [55]. In this review, we will focus on the first three. A schematic representation of the first three methods is shown in Figure 3.

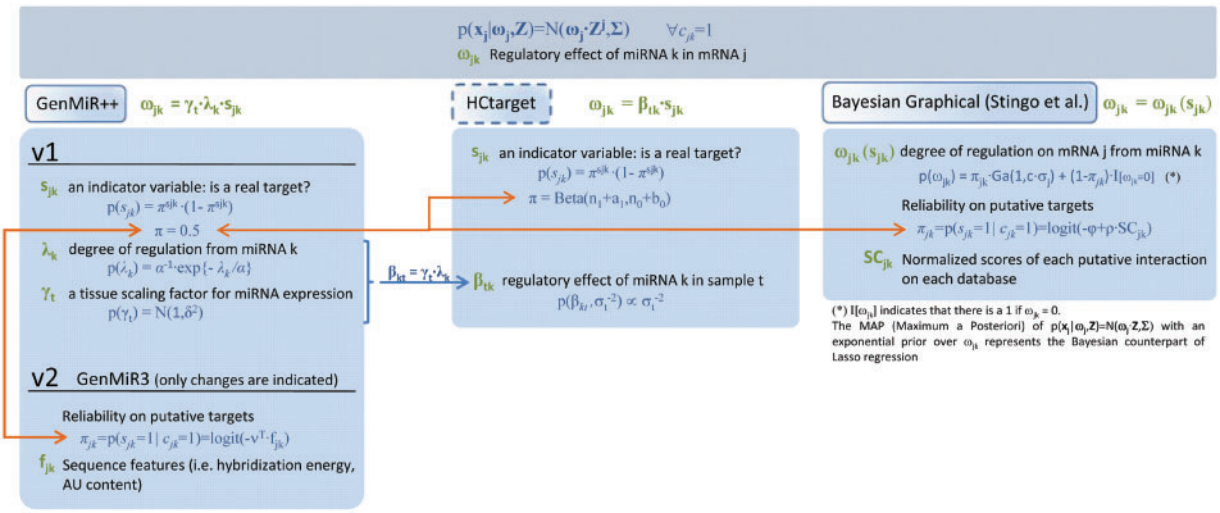

Figure 3: Bayesian inference methods for scoring putative miRNA–mRNA targets based on miRNA and mRNA expression data: relationships between them. Bayesian inference methods use a priori information to estimate parameters in a probability framework. The models of the figure assume normal distribution over $\mathbf{x_j}$ with a mean value dependent on the regressors and miRNA expression. The models divide the regulatory effect into different factors: (i) in GenMiR++ the regressors are divided onto the tissue scaling of miRNA expression, the degree of regulation form miRNAs and the probability of a putative target of being real or not; (ii) HCtarget groups the tissue effect and the degree of regulation of GenMiR++ and accounts for the same indicator variable as GenMiR++ and (iii) the Bayesian graphical method only accounts for the degree of regulation from miRNA, that depend on the indicator variables for putative targets (i.e. the model uses a mixed prior for the regressors). Each of the models assumes a different prior probability over π (orange arrows), the probability of a putative target of being real. The dashed lines on HCtarget indicate that the model is not feasible when the number of regressors is much larger than that of samples. VI and V2 refer to version 1 and version 2, respectively.

GenMiR++ was the first developed method for miRNA–mRNA target prediction based on expression data [51]. In the model, the expression of mRNA *j*, $\mathbf{x_j}$, is assumed to be normally distributed around its mean expression value and the regulation effects from miRNAs, ω_{jk}. These regressors are divided into three factors: γ_t a tissue scaling factor of miRNA expression, λ_k the degree of regularization of miRNA *k* over all its putative mRNA targets and $\cdot s_{jk}$ an indicative variable determining whether or not a putative target is a real target. The aim of the model is to infer the probabilities of putative targets of being real, $p(s_{jk}=1|c_{jk}=1)$ by approximating the posterior with variational inference techniques. In a second version [82], GenMiR3, the model was extended to also account for the reliability of putative interactions. The authors evaluated different sequence features (i.e. AU content, hybridization energy) by adding logit priors to the GenMiR++ model.

A drawback of GenMiR++ is that the expectation maximization (EM) algorithm obtained from variational inference is computationally expensive. In HCtarget [81], a variation of GenMiR++, the computing time is reduced by redefining some of the priors and by solving the full posterior by MCMC (Markov Chain Monte Carlo) techniques. This method considers a different regulation effect from each miRNA *k* for all its putative mRNA targets on each sample (i.e. variables γ_t and λ_k are grouped into β_{kt}). A major drawback of HCtarget is that, as with MLR, its posterior is not suitable for data where the number of regressors is higher than the number of samples.

In GenMiR++ and HCtarget, each miRNA is assumed to regulate to 'the same degree' all its putative mRNA targets. However, since the regulatory effect is governed by the sequence complementarity, it seems natural to expect a different degree of regulation from each miRNA *k* on each putative target mRNA *j*. In a recently published work, the Bayesian graphical approach [80], a different regulation effect for each putative mRNA *j*—miRNA *k* pair is determined. Contrarily to GenMiR++, this model groups all the effects into a single variable, ω_{jk} that accounts for the regulatory effects from miRNAs. By assuming a gamma prior over ω_{jk}, only down-regulation effects from miRNAs are considered and the sparsity of the solution is enforced [83]. This procedure can be seen as a Bayesian alternative of Lasso [84, 85]. Furthermore, authors use a full MCMC procedure for the direct quantification of the posterior. As in GenMiR3, this method also considers the reliability of sequence-based putative interactions by adding logit priors over sequence features. For this method, the sequence features are the scores (i.e. expected quality), or combination of scores, of the putative interactions from different databases.

RELATIONSHIPS BETWEEN DIFFERENT MODELS

We have classified the methods into three main groups. First of all, we include methods that perform pairwise comparisons between mRNA and miRNA, namely Pearson correlation, Spearman correlation and MI (Figure 4a). The second group includes MLR and regularized least squares (LASSO regression, RR and the elastic-net) (Figure 4b). Finally, the different variants of Bayesian methods (GenMiR++, HCtarget and the Bayesian graphical model) are grouped (Figure 4c). As indicated before, the aim of all these methods is to score the putative lists of mRNA–miRNA interactions. Thus, in practice, methods that provide similar ranking of their scores can be considered to be related. According to this, there are relevant relationships among and within these groups to be pinpointed (arrows in Figure 4).

Concerning the first group, under the normality assumption, MI can be derived from Pearson correlation values ρ_{jk}. MI is the quantity $-1/2\cdot\log(1-\rho_{jk}^2)$ [86]. In this case, the ranking of MI and the absolute value of the correlation is identical. This equivalence clarifies that MI does not take into account the sign of the interaction. Regarding the linear models, it is well known that the elastic-net is an extension to LASSO regression and RR). The elastic net is converted into any of them by adjusting either of the tuning parameters (ν_1 or ν_2) to zero. Of course, if both tuning parameters of the elastic-net are set to zero, no regularization is performed and is converted into a MLR. Contrarily, if the tuning parameter of the RR is assumed to be large, the matrix to be inverted, in the explicit solution shown in Table 1 is almost diagonal and RR can be viewed as a scalar product of vectors.

Apart from these relationships, there is one non-obvious connection between GenMiR++ and RR. GenMiR++ can be shown to be equivalent to solve *J* independent RRs, one per each mRNA (see Supplementary Data), in where gene

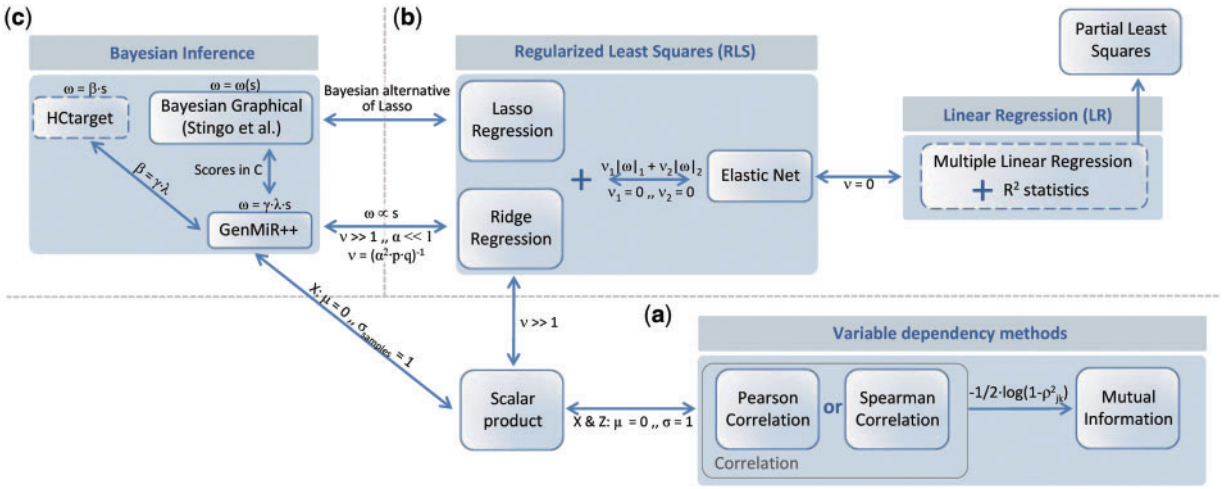

Figure 4: Relationships between different models: a) variable dependency methods, b) algorithms based on Regularized Least Squares and c) Bayesian Inference methods. Dashed boxes refer to algorithms that can only be applied if the number of regressors is larger than the number of samples. In general, the number of samples for which mRNA and miRNA expression is available is limited and the number of regressors tends to be higher than that of samples.

expression data is normalized to have sample variance equal to one (see Supplementary Data). This equivalence is accurate for low values of α, the hyperparameter of the degree of down-regulation from miRNAs and this is the case in the GenMiR++ model. Due to the relationship between GenMiR++ and RR (see Figure 4), these RR models have a large penalization parameter identical for all mRNAs, i.e. GenMiR++ can be further simplified to a scalar product of vectors. Therefore, the output of GenMiR++, the probabilities of each putative mRNA j—miRNA k interaction of being real $p(s_{jk}=1|c_{jk}=1)$, are proportional to the simple scalar product,

$$p(s_{jk}=1|c_{jk}=1) \propto \left(\mathbf{x_j}\right)_{\mu_j=0|\sigma^2_{samples}=1} \cdot \mathbf{z_k^T}, \quad (5)$$

where $\left(\mathbf{x_j}\right)_{\mu_j=0|\sigma^2_{samples}=1}$ is the standardized expression of mRNAs and z_k^T is the expression of the putative miRNA regulators (see Supplementary Data). Notice that once more, gene expression is standardized.

The suggested approximation in Equation (5) is quite accurate. The Spearman correlation is a good measure to compare the ranking of the interactions using both methods. This correlation ranges from 0.9913 to 0.9989 for different data sets. We have compared the expression data used in GenMiR++ and it is available on http://www.psi.toronto.edu/genmir/, acute lymphoblastic leukemia data from Ref. [67], multiple myeloma Refs [87] and [88], and NCI-60 data set from Ref. [71] (see Supplementary Data for a brief description of the data sets). For these data sets, the rankings provided by Equation (5) or GenMiR++ are almost identical (see in Supplementary Data). However, Equation (5) is several orders of magnitude less expensive in computational time (0.025 versus 300 s). Since the ranking is almost identical, in the following sections we have used the approximation given by Equation (5) instead of using the EM method proposed by GenMiR++ authors. The difference between GenMiR++ and the approximation given by Equation (5) is negligible in all used data sets.

Although Equations (1), corresponding to Pearson correlation and Equation (5), the approximation of GenMiR++, are both scalar products of mRNA and miRNA expressions, the solutions of GenMiR++ were shown by the authors to outperform those of Pearson correlation (and we have also checked that this is indeed the case, see next section). An interesting point to note here is that the only difference between them is data normalization.

SIDE BY SIDE COMPARISON

In this section, we compare the performance of the different methods described in this article. However, we could not run some of the methods due to several reasons: (i) PLS method does not have the code available and (ii) HCtarget is not suitable for data

where the number of predictors is larger than that of samples. The graphical Bayesian method is the only method that considers the scores of the predictions of putative interactions. These scores are difficult to combine across different databases. Furthermore, in the case where the scores from databases are not considered, the graphical Bayesian method is supposed to work as an alternative to Lasso regression. For this reason, we decided not to include the graphical Bayesian method in the analysis. In Ref. [56], the elastic net is used to solve the proposed problem. In this model, the effect of transcription factors is also included. Preliminary results in our analysis showed that elastic-net method for some of the data sets behaved similarly to TaLasso (see Supplementary Data). For the sake of simplicity, we have not included these results in the main manuscript since they only reproduce part of the model in the cited reference. MI was determined using the equivalence between correlation and MI indicated in Figure 4. Due to this, we have determined two values for the MI: one obtained from the Pearson correlation and the other from the Spearman correlation. We refer to them as |Pearson Cor.| and |Spearman Cor.|, respectively.

We have used a simple and intuitive score to compare the described algorithms: the number of experimentally validated interactions among the top-ranked predictions for each algorithm. A good algorithm will expectedly provide a large number of interactions that have already been experimentally validated. The number of checked interactions has been arbitrarily set to 1000, i.e. we score the quality of the algorithms according to the number of validated interactions within the top-1000. Other numbers of interactions provide similar results.

In addition to this, since genes are expected to behave coordinately, we have also computed the enrichment on KEGG pathway categories within the genes in the top-ranked interactions. It will be shown that although these two measurements are completely different, algorithms that perform well according to the enrichment on experimentally validated interactions also perform well according to KEGG pathway enrichment. The predicted interactions are more likely to be true if the computed KEGG pathways are related with the biology behind the data set used. KEGG pathway enrichment analysis was performed by using GeneCodis 2.0 [60, 61] for those 200 mRNAs with miRNA regulators within the top predicted interactions (for more information on enrichment analysis, refer to Ref. [79]).

We used five different data sets to analyze the relevance of the results: acute lymphoblastic leukemia [67], multiple cancer class data [89, 90], multiple myeloma [87] and [88], and NCI-60 cancer cell panel [71] (downloaded from Ref. [91]) (see Supplementary Data for a brief description of the data sets). Results are shown in Figure 5 that shows the experimentally validated interactions using TarBase [46] and miRecords [47] and Figure 6 that shows the enriched KEGG pathways (no enriched KEGG pathway was found for MCC results).

Our results show that GenMiR++ has a slightly better performance than Correlation in terms of the number of validated interactions. This improvement is more apparent in the KEGG pathways enrichment results. Correlation and MI obtain few or no pathways statistically significant. These results allow us to conclude that data normalization has key importance for miRNA target prediction.

Note in Figure 6 that MI results are outperformed by plain correlation. One important drawback of MI is that it does not take into account the sign of the relationship. In the same way, the addition of non-positive constraints to the LASSO improves KEGG enrichment. This result was also observed in Ref. [79]. It seems that considering only down-regulatory effects is relevant to find the relationships between miRNA and their targets.

DISCUSSION AND CONCLUSION

miRNAs are small RNA molecules (′22 nt) that interact with their target mRNAs inhibiting translation or/and cleavaging the target mRNA. This interaction is guided by sequence complementarity through RISC compound (multiprotein complex that incorporates mature miRNA) and results in both the mRNA and protein levels reduction. Deciphering miRNA targets is crucial for diagnosis and therapeutics since miRNAs are involved in key biological processes and different diseases. However, miRNA regulatory mechanisms are complex and the lack of a high-throughput and low-cost miRNA target screening technique, make miRNA target prediction laborious. Although, in the last years several computational methods based on sequence complementarity of the miRNA and the mRNAs have been developed, their predictions are inconsistent and their expected false positive rates are large.

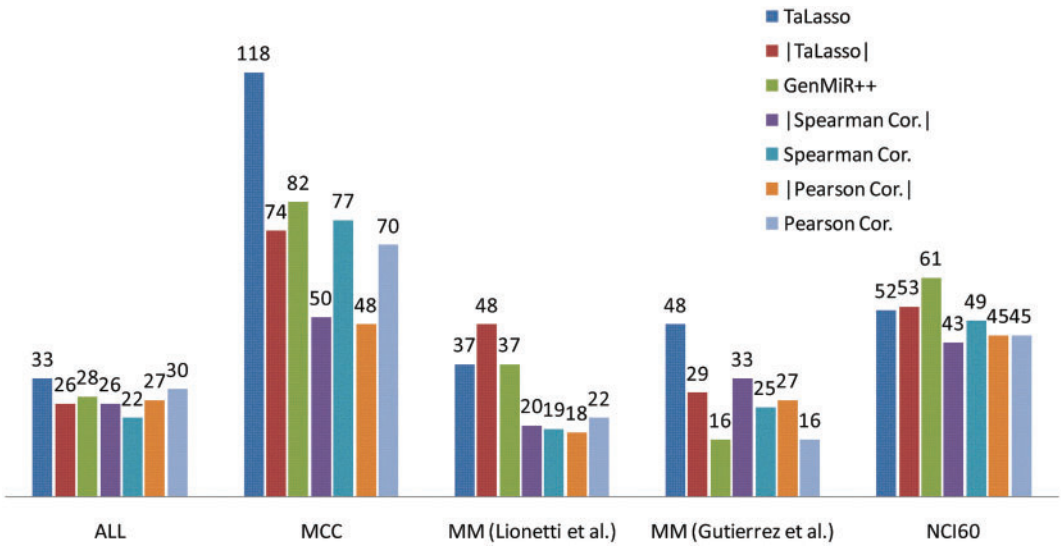

Figure 5: Enrichment values on experimentally validated targets for different data sets and methods. Number of experimentally validated targets predicted within the top-500 retrieved interactions. The names within an absolute value indicate that they do not differentiate the sign of the relationships. In the case of TaLasso, the absolute value indicates that non-positive restrictions are not considered. ALL = acute lymphoblastic leukemia, MM = multiple myeloma, MCC = multiple class cancer.

Recently, new methods based on the joint analysis of miRNA and mRNA expression for the filtering of sequence-based putative targets have been proposed. In brief, by assuming the expected inverse (or direct) relationship between the expression of a miRNA and its mRNA targets (or proteins), these methods determine whether a putative target is real for a particular set of experimental data. These methods have shown to be effective identifying the most prominent interactions from the databases of putative targets (see Supplementary Data for a brief analysis of the added value of using expression data for target prediction by using sequence-based predictions as initial set of putative interactions).

Filtering of putative miRNA–mRNA interactions involves three matrices: X, Z and C that represent mRNA and miRNA expressions and putative lists of targets, respectively. The models described in this review use this information in different mathematical models that have been grouped onto: dependency analysis, linear regression (multiple and regularized) and Bayesian methods. Among the different relationships between the models described here, special interest is on the one concerning GenMiR++ and RR. We have shown that owing to the low degree of miRNA regulation considered in GenMiR++, the ranking of its results can be obtained by applying RR with a high regularization term. Furthermore, we have shown that GenMiR++, in the same way as Pearson correlation, can be expressed as a scalar product of miRNA and mRNA expressions. The only difference between both scalar products is data normalization—the first standardizes only the expression of mRNAs while the latter standardizes both, the expressions of mRNAs and miRNAs. We have shown that this difference is crucial for miRNA target prediction. Our KEGG enrichment results have shown that GenMiR++ recovers more reliable interactions than Pearson correlation.

Most of the described methods account for down-regulatory effects from miRNAs. This is done indirectly—determining all the regressors and restricting afterward their interest to the negative ones—or directly—adding particular priors in Bayesian methods or adding constraints to the linear models to force them to only take into account negative regressors. Few studies do also consider enhancement effects from miRNAs (i.e. analyzing positive correlation values or using MI). However, our results have shown that taking into account the sign of the regulation improves the results, i.e. the results are more enriched in experimental interactions and are biologically sound.

In general, matrix C is considered to be an indicator matrix, with values equal to 1 (putative target) or 0 (not a putative target). This matrix is generally created by assigning a 1 to those interactions of a database that surpass a threshold and a 0 otherwise. These scores indicate the reliability of the predicted putative interaction on that database. Combining

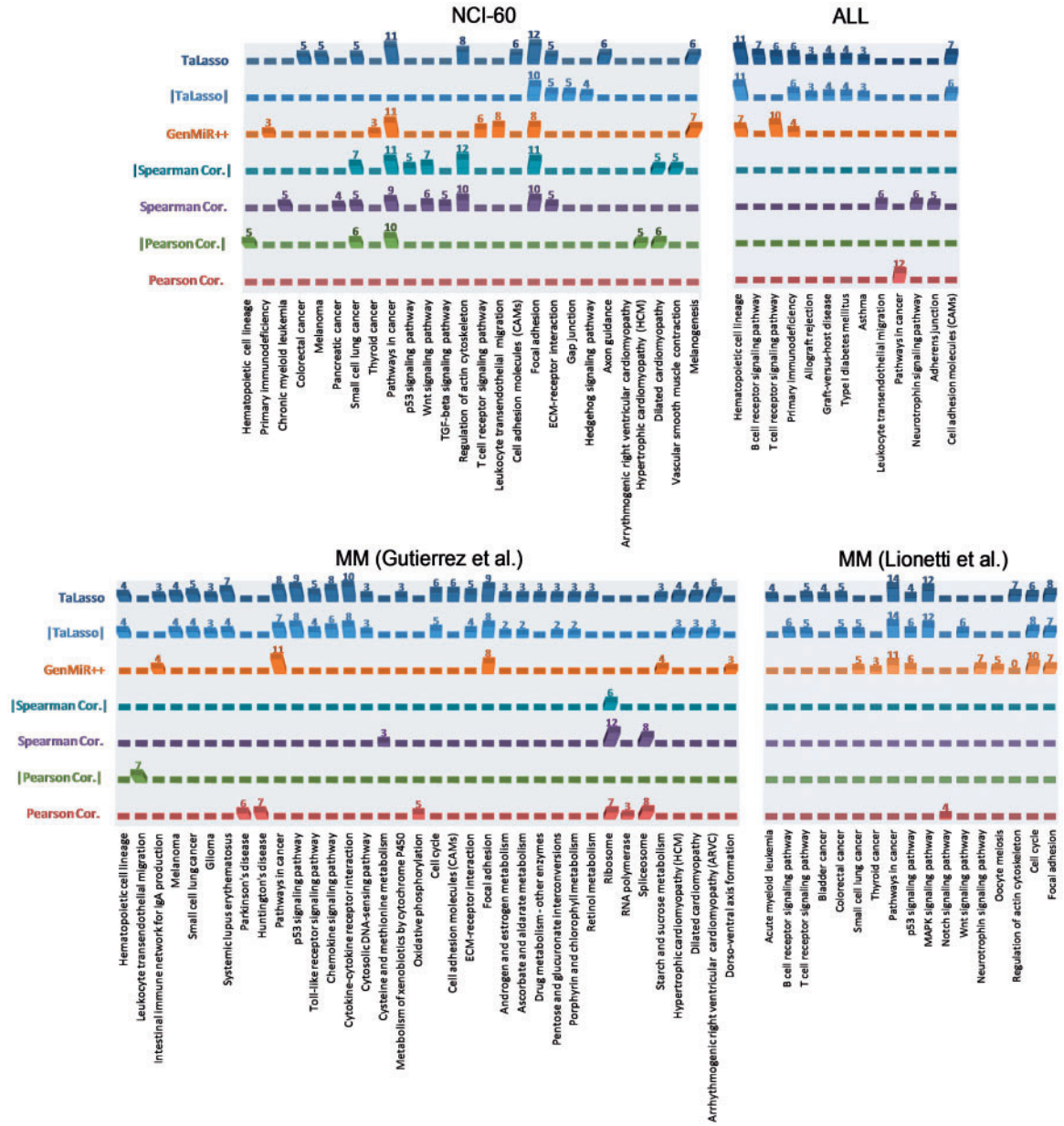

Figure 6: KEGG pathway enrichment results for different methods and data sets. The names within an absolute value indicate that they do not differentiate the sign of the relationships. In the case of TaLasso, the absolute value indicates that non-positive restrictions are not considered. ALL = acute lymphoblastic leukemia, MM = multiple myeloma.

putative targets from different databases—by taking their unions and/or intersections—allows their reliability to increase. Some of the methods include the scores of each interaction from a particular database (or combinations of scores from different databases) in their mathematical models. Among the methods described in this review, only the Bayesian models [80, 82, 92] account for these scores. GenMiR3 [82] was the first model that included logit priors to consider sequence features (i.e. hybridization energy, AU content). Alternatively, the Bayesian graphical method of Stingo *et al.* [80] uses similar logit functions to include the scores of a database or a combination of databases. The use of prior knowledge over the reliability of each putative interaction can be extended to the majority of the rest of the methods of this review. For example, correlation can be multiplied—or modified in some way—by the reliability of the interaction. The ranking will change accordingly. In addition to this, in regularized least squares methods the regularizing parameter can be weighted according to each individual interaction. To this end,

it would be necessary to combine the scores of different sequence-based databases which limit the use of these scores for miRNA target prediction. In this respect, ExprTarget [92] combines the scores of different databases using logistic regression to provide an overall score for each of the putative miRNA–mRNA interactions. The probability of each miRNA–mRNA interaction to be real from expression data is modeled via a logit function. In the end, the weights of each putative interaction are determined also by using a logistic regression.

A possible limitation of these methods occurs if samples are obtained from heterogenous experimental conditions. Since mRNA regulation is not only driven by miRNAs, other regulators—such as transcription factors—have larger effects on mRNA expression and target prediction become less reliable. In these cases, it would be interesting to combine TF activity—and other regulators—with miRNA activity. An example of this approach is Ref. [56]. Since heterogenous sources of information can blur miRNA effects on mRNA regulation, it would also be interesting to develop mathematical methods for data classification (i.e. a mixture of cancer samples could be divided into different classes of cancers with common groups of regulatory pathways). In this respect, some authors have developed tools directed to the search of miRNA regulatory modules that are able to discern regulation between different samples [93–98]. Clustering-based method could also be used.

Although huge advances have been made in miRNA target prediction, there is still much work to do. Until high-throughput experimental techniques reach the market, computational methods will continue to be of high importance. Combination of expression data with sequence-based prediction have shown to be feasible. Although, the number of predicted targets is still high, these methods have marked new future working lines. In this respect, models that combine more heterogeneous experimental data (i.e. TF, protein, time-course data, miRNA transfection effects on mRNA and proteins) could be more reliable on the predicted miRNA–mRNA interactions.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- miRNAs down-regulate their target mRNAs. This effect has shown to play a key role in different biological processes.
- There is still no high-throughput experimental technique for miRNA target prediction and thus, several computational methods have emerged. Nevertheless, their expected false positive rates are still large and predictions of different methods do not match at all. Some of these methods combine both expression data with sequence analysis.
- The integration of miRNA and mRNA expression data have shown to be a good method for filtering sequence-based putative predictions. The algorithms to develop this integration can be categorized into three groups: dependence-based methods (Pearson and Spearman correlation and MI), MLR and regularized least squares (MLR, Lasso, Ridge and Elastic-net), and Bayesian inference methods (GenMiR++, HCtarget and a Bayesian graphical method).
- Their comparison shows: (i) that restricting the search of miRNA regulation to down-regulation improves the reliability of the results and (ii) that normalization of mRNA and miRNA expressions is crucial for miRNA target prediction.

FUNDING

The work of A.M. and J.P. was funded by the grant of the Basque Government (Programas de Formación y Perfeccionamiento de Personal Investigador, http://www.hezkuntza.ejgv.euskadi.net/r43-5552/es/contenidos/informacion/dib4/es_2035/bfi_c.html).

References

1. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 2011;**39**:D152–D157.
2. Rana TM. Illuminating the silence: understanding the structure and function of small RNAs. *Nat Rev Mol Cell Biol* 2007;**8**:23–36.
3. Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 2009;**10**:126–139.
4. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science* 2007;**318**:1931–1934.
5. Hobert O. Gene regulation by transcription factors and microRNAs. *Science* 2008;**319**:1785–6.
6. Flynt AS, Lai EC. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat Rev Genet* 2008;**9**:831–842.
7. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 2010;**79**:351–379.
8. Mattick JS. RNA regulation: a new genetics? *Nat Rev Genet* 2004;**5**:316–323.
9. Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 2011;**12**:99–110.

10. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 2008; **9**:102–14.
11. Guo H, Ingolia NT, Weissman JS, *et al*. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 2010;**466**:835–840.
12. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 2007;**8**: 93–103.
13. Nilsen TW. Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends Genet* 2007;**23**: 243–249.
14. Lynam-Lennon N, Maher SG, Reynolds JV. The roles of microRNA in cancer and apoptosis. *Biol Rev Camb Philos Soc* 2009;**84**:55–71.
15. Inoue K. MicroRNA function in animal development. *Tanpakushitsu Kakusan Koso* 2007;**52**:197–204.
16. Alvarez-Garcia I, Miska EA. MicroRNA functions in animal development and human disease. *Development* 2005;**132**:4653–4662.
17. Taft RJ, Pang KC, Mercer TR, *et al*. Non-coding RNAs: regulators of disease. *J Pathol* 2010;**220**:126–139.
18. Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nat Rev Cancer* 2006;**6**:857–866.
19. Pfeifer A, Lehmann H. Pharmacological potential of RNAi–focus on miRNA. *Pharmacol Therap* 2010;**126**: 217–227.
20. Huang Y, Shen XJ, Zou Q, *et al*. Biological functions of microRNAs: a review. *J Physiol Biochem* 2011;**67**:129–139.
21. Gentner B, Visigalli I, Hiramatsu H, *et al*. Identification of hematopoietic stem cell-specific miRNAs enables gene therapy of globoid cell leukodystrophy. *Sci Trans Med* 2010;**2**:58ra84.
22. Brown BD, Naldini L. Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. *Nat. Rev. Genet.* 2009;**10**(8):578–585.
23. Weiler J, Hunziker J, Hall J. Anti-miRNA oligonucleotides (AMOs): ammunition to target miRNAs implicated in human disease? *Gene Ther* 2006;**13**:496–502.
24. Krützfeldt J, Rajewsky N, Braich R, *et al*. Silencing of microRNAs in vivo with ''antagomirs''. *Nature* 2005;**438**: 685–689.
25. Ebert MS, Neilson JR, Sharp PA. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods* 2007;**4**:721–726.
26. Chi SW, Zang JB, Mele A, *et al*. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 2009; **460**:479–486.
27. Hafner M, Landthaler M, Burger L, *et al*. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 2010;**141**:129–141.
28. Kuhn DE, Martin MM, Feldman DS, *et al*. Experimental validation of miRNA targets. *Methods San Diego Calif* 2008; **44**:47–54.
29. Jin H, Tuo W, Lian H, *et al*. Strategies to identify microRNA targets: new advances. *New Biotechnol* 2010;**27**: 734–738.
30. Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Res* 2011;**61**:1–9.
31. Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA target prediction. *Curr Mol Med* 2011;**11**: 93–109.
32. Ørom UA, Lund AH. Experimental identification of microRNA targets. *Gene* 2010;**451**:1–5.
33. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. *Nat Struct Mol Biol* 2010;**17**:1169–1174.
34. Saito T, Saetrom P. MicroRNAs–targeting and target prediction. *New Biotechnol* 2010;**27**:243–249.
35. Mazière P, Enright AJ. Prediction of microRNA targets. *Drug Discov Today* 2007;**12**:452–458.
36. Betel D, Wilson M, Gabow A, *et al*. The microRNA.org resource: targets and expression. *Nucleic Acids Res* 2008;**36**: D149–D153.
37. Griffiths-Jones S, Saini HK, Van Dongen S, *et al*. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008;**36**: D154–D158.
38. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;**120**:15–20.
39. Grimson A, Farh KK-H, Johnston WK, *et al*. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 2007;**27**:91–105.
40. Friedman RC, Farh KK-H, Burge CB, *et al*. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009;**19**:92–105.
41. Maragkakis M, Reczko M, Simossis VA, *et al*. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res* 2009;**37**: W273–W276.
42. Krek A, Grün D, Poy MN, *et al*. Combinatorial microRNA target predictions. *Nat Genet* 2005;**37**:495–500.
43. Alexiou P, Maragkakis M, Papadopoulos GL, *et al*. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics* 2009;**25**: 3049–3055.
44. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods* 2006; **3**:881–886.
45. Bentwich I. Prediction and validation of microRNAs and their targets. *FEBS Lett* 2005;**579**:5904–5910.
46. Papadopoulos GL, Reczko M, Simossis VA, *et al*. The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* 2009;**37**:D155–D158.
47. Xiao F, Zuo Z, Cai G, *et al*. miRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Res* 2009;**37**:D105–D110.
48. Hsu S-D, Lin F-M, Wu W-Y, *et al*. miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res* 2011;**39**:D163–D169.
49. Dweep H, Sticht C, Pandey P, *et al*. miRWalk - database: prediction of possible miRNA binding sites by ''walking'' the genes of three genomes. *J Biomed Inform* 2011;**44**:1–9.
50. Hsu PWC, Huang H-D, Hsu S-D, *et al*. miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Res* 2006;**34**: D135–D139.
51. Huang JC, Babak T, Corson TW, *et al*. Using expression profiling data to identify human microRNA targets. *Nat Methods* 2007;**4**:1045–1049.

52. Gennarino VA, Sardiello M, Mutarelli M, *et al*. HOCTAR database: a unique resource for microRNA target prediction. *Gene* 2011;**480**:51–8.
53. Sales G, Coppe A, Bisognin A, *et al*. MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Res* 2010;**38**:W352–W359.
54. Elkan-Miller T, Ulitsky I, Hertzano R, *et al*. Integration of transcriptomics, proteomics, and MicroRNA analyses reveals novel MicroRNA regulation of targets in the mammalian inner ear. *PLoS One* 2011;**6**:12.
55. Li J, Min R, Bonner A, *et al*. A probabilistic framework to improve microrna target prediction by incorporating proteomics data. *J Bioinform Comput Biol* 2009;**7**:955–972.
56. Beck D, Ayers S, Wen J, *et al*. Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in Myelodysplastic syndromes. *BMC Med Genomics* 2011;**4**:19.
57. Ritchie W, Rajasekhar M, Flamant S, *et al*. Conserved expression patterns predict microRNA targets. *PLoS Computat Biol* 2009;**5**:8.
58. Jayaswal V, Lutherborrow M, Ma DDF, *et al*. Identification of microRNAs with regulatory potential using a matched microRNA-mRNA time-course data. *Nucleic Acids Res* 2009;**37**:e60.
59. Ragan C, Zuker M, Ragan MA. Quantitative Prediction of miRNA-mRNA Interaction Based on Equilibrium Concentrations. *PLoS Computat Biol* 2011;**7**:11.
60. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, *et al*. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res* 2009;**37**:W317–W322.
61. Carmona-Saez P, Chagoyen M, Tirado F, *et al*. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol* 2007;**8**: R3.
62. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–15550.
63. Mootha VK, Lindgren CM, Eriksson K-F, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**:267–73.
64. Nam S, Li M, Choi K, *et al*. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res* 2009; **37**:W356–W362.
65. Liu H, Brannon AR, Reddy AR, *et al*. Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. *BMC Syst Biol* 2010;**4**:51.
66. Van Der Auwera I, Limame R, Van Dam P, *et al*. Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype. *Br J Cancer* 2010;**103**:532–541.
67. Fulci V, Colombo T, Chiaretti S, *et al*. Characterization of B- and T-lineage acute lymphoblastic leukemia by integrated analysis of MicroRNA and mRNA expression profiles. *Genes Chromosomes Cancer* 2009;**48**:1069–1082.
68. Guimbellot JS, Erickson SW, Mehta T, *et al*. Correlation of microRNA levels during hypoxia with predicted target mRNAs through genome-wide microarray analysis. *BMC Med Genomics* 2009;**2**:15.
69. Zhu M, Yi M, Kim C-H, *et al*. Integrated miRNA and mRNA expression profiling of mouse mammary tumor models identifies miRNA signatures associated with mammary tumor lineage. *Genome Biol* 2011;**12**:R77.
70. Nunez-Iglesias J, Liu C-C, Morgan TE, *et al*. Joint Genome-Wide Profiling of miRNA and mRNA Expression in Alzheimer's Disease Cortex Reveals Altered miRNA Regulation. *PLoS One* 2010;**5**:9.
71. Wang Y-P, Li K-B. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics* 2009;**10**:218.
72. Ritchie W, Flamant S, Rasko JEJ. mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics* 2010;**26**:223–227.
73. Cho S, Jun Y, Lee S, *et al*. miRGator v2.0: an integrated system for functional investigation of microRNAs. *Nucleic Acids Res* 2011;**39**:D158–D162.
74. Gennarino VA, Sardiello M, Avellino R, *et al*. MicroRNA target prediction by expression analysis of host genes. *Genome Res* 2009;**19**:481–490.
75. Kim S, Choi M, Cho K-H. Identifying the target mRNAs of microRNAs in colorectal cancer. *Computat Biol Chem* 2009;**33**:94–99.
76. Wang H, Li W-H. Increasing MicroRNA target prediction confidence by the relative R(2) method. *J Theor Biol* 2009; **259**:793–798.
77. Li X, Gill R, Cooper NGF, *et al*. Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. *BMC Med Genomics* 2011;**4**:44.
78. Lu Y, Zhou Y, Qu W, *et al*. A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics* 2011;**27**:1–8.
79. Muniategui A, Nogales-Cadenas R, Vázquez L, *et al*. Quantification of miRNA-mRNA interactions. *PLoS One* 2012;**7**(2):e30766.
80. Stingo FC, Chen YA, Vannucci M, *et al*. A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann Appl Stat* 2011;**4**:2024–2048.
81. Su N, Wang Y, Qian M, *et al*. Predicting MicroRNA targets by integrating sequence and expression data in cancer. *IEEE Int Conf Syst Biol* 2011.
82. Huang JC, Frey BJ, Morris QD. Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pacif Symp Biocomput* 2008;**63**:52–63.
83. Figueiredo MAT. Adaptive sparseness for supervised learning. *IEEE Trans Pattern Anal Machine Intell* 2003;**25**: 1150–1159.
84. Kyung M, Gill J, Ghosh M, *et al*. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 2010;**5**: 369–412.
85. Park T, Casella G. The Bayesian Lasso. *J Am Stat Assoc* 2008; **103**:681–686.
86. Cover TM, Thomas JA. *Elements of Information Theory,* 2nd Revised edn, John Wiley & Sons Inc, 2006, 776.
87. Lionetti M, Biasiolo M, Agnelli L, *et al*. Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. *Blood* 2009;**114**:e20–e26.
88. Gutiérrez NC, Sarasquete ME, Misiewicz-Krzeminska I, *et al*. Deregulation of microRNA expression in the different

genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia* 2010;**24**:629–637.

89. Lu J, Getz G, Miska EA, *et al*. MicroRNA expression profiles classify human cancers. *Nature* 2005;**435**:834–838.
90. Ramaswamy S, Tamayo P, Rifkin R, *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* 2001;**98**:15149–15154.
91. Shankavaram UT, Varma S, Kane D, *et al*. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* 2009;**10**:277.
92. Gamazon ER, Im H-K, Duan S, *et al*. ExprTarget: an integrative approach to predicting human MicroRNA targets. *PLoS One* 2010;**5**:8.
93. Liu B, Liu L, Tsykin A, *et al*. Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics* 2010;**26**:3105–3111.
94. Yoon S, De Micheli G. Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics* 2005;**21(Suppl 2)**:ii93–i100.
95. Savage RS, Ghahramani Z, Griffin JE, *et al*. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* 2010;**26**:i158–i167.
96. Jayaswal V, Lutherborrow M, Ma DD, *et al*. Identification of microRNA-mRNA modules using microarray data. *BMC Genomics* 2011;**12**:138.
97. Joung J-G, Hwang K-B, Nam J-W, *et al*. Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics* 2007;**23**:1141–1147.
98. Liu B, Li J, Tsykin A. Discovery of functional miRNA-mRNA regulatory modules with computational methods. *J Biomed Inform* 2009;**42**:685–691.