A BAYESIAN GRAPHICAL MODELING APPROACH TO MICRORNA REGULATORY NETWORK INFERENCE

BY FRANCESCO C. STINGO, YIAN A. CHEN, MARINA VANNUCCI[1],
MARIANNE BARRIER AND PHILIP E. MIRKES

*University of Florence and Rice University, Moffitt Cancer Center, Rice
University, Texas A&M University and Texas A&M University*

It has been estimated that about 30% of the genes in the human genome are regulated by microRNAs (miRNAs). These are short RNA sequences that can down-regulate the levels of mRNAs or proteins in animals and plants. Genes regulated by miRNAs are called targets. Typically, methods for target prediction are based solely on sequence data and on the structure information. In this paper we propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating expression levels of miRNAs with their potential mRNA targets and, via the prior probability model, with their sequence/structure information. We use a directed graphical model with a particular structure adapted to our data based on biological considerations. We then achieve network inference using stochastic search methods for variable selection that allow us to explore the huge model space via MCMC. A time-dependent coefficients model is also implemented. We consider experimental data from a study on a very well-known developmental toxicant causing neural tube defects, hyperthermia. Some of the pairs of target gene and miRNA we identify seem very plausible and warrant future investigation. Our proposed method is general and can be easily applied to other types of network inference by integrating multiple data sources.

1. Introduction. One of the major tasks and challenges in the post-genomics era is to decipher how genes and their products (proteins) are regulated. Regulation can happen at transcriptional, post-transcriptional, translational and post-translational level. Transcription is the process of synthesizing a stretch of ribonucleic acids (RNA) based on a specific DNA sequence. Transcriptional regulation can affect whether or not a specific RNA is transcribed as well as the amount of RNA produced. RNA can be regulated post-transcriptionally through degradation or modification of the RNA strand, which can affect its function. A segment of RNA can interact with other genes or proteins or can encode a protein. Translation, the process of forming a protein based on an RNA sequence, can also be positively or negatively regulated. Proteins often undergo post-translational modifications, which can affect their function. An abundant class of short ($\sim$22 nucleotide) RNAs, known as microRNAs (miRNAs), plays crucial regulatory roles

in animals and plants [Farh et al. (2005)]. It has been estimated that at least 30% of the genes in human genomes are regulated by miRNAs; see Lewis, Burge and Bartel (2005) and also Rajewsky (2006). Genes regulated by miRNAs are generally called "targets." The actual mechanism of miRNA regulation is still an active area of research and the complete picture of the regulatory mechanism is still to be understood [Thermann and Hentze (2007)]. According to current knowledge, it is believed that miRNAs regulate their targets either by degrading mRNA post-transcriptionally [Bagga et al. (2005)], or by suppressing initiation of protein synthesis [Pillai et al. (2005)], and/or by inhibiting translation elongation after initiation of protein synthesis [Petersen et al. (2006)]. The biosynthesis and maturation of miRNAs is composed of distinctive events. Briefly, the miRNA precursor is processed by Dicer to produce the mature, single-stranded molecule that is incorporated into the RNA-induced silencing complex (RISC) and possesses a 6–8 bases of "seed" sequence that mainly targets complementary sequences within the 3'-untranslated regions (UTR) of mRNA transcripts.

Many algorithms have been developed to predict potential target sequences for miRNAs based on their specific sequence and structure characteristics. These algorithms mainly use sequence information, hybridization energy for structure prediction and cross-species comparisons [Rajewsky (2006)]. Target prediction algorithms generally take into account different factors that influence miRNA/target interactions, such as seed match complementarity, 3'-UTR seed match context, the conservation, favorability of free energy binding and binding site accessibility. Some of the more widely used algorithms include the following: TargetScan of Lewis, Burge and Bartel (2005), PicTar of Krek et al. (2005), miRanda of John et al. (2004), PITA of Kertesz et al. (2007) and DIANA-microT of Kiriakidou et al. (2004). A comprehensive review of these and other methods can be found in Yoon and Micheli (2006). Typically, a large amount (e.g., hundreds to thousands) of potential targets are predicted by these algorithms, and it can be overwhelming for researchers to search through the candidate targets for those which play critical regulatory roles under particular experimental or clinical conditions.

Our goal is to develop a statistical approach to identify a small set of potential targets with high confidence, making future experimental validation feasible. Since miRNAs down-regulate the expression of their targets, expression profile of miRNAs and their potential targets can be used to infer their regulatory relationships. We propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating these two types of expression levels. We use a directed graphical model with a particular structure adapted to our data based on biological considerations. We take into account current knowledge on the down-regulation effect of mRNA expression (by miRNA) by imposing constraints on the sign of the regression coefficients of our proposed model. The model also integrates the sequence/structure information, as generated by widely used target prediction algorithms, via the prior probability model. We then achieve network inference using stochastic search methods for variable selection.

We consider experimental data from a study on a very well-known developmental toxicant causing neural tube defects, hyperthermia. We have available 23 mouse miRNAs and a total of 1297 potential targets. We infer their regulatory network under two different treatment conditions and also investigate time-dependent regulatory associations. Some of the pairs of target gene and miRNA we identify seem very plausible and warrant future investigation.

Huang, Morris and Frey (2007), Huang, Frey and Morris (2008) have proposed a Bayesian model for the regulatory process of targets and miRNAs which is similar to the one we propose here. However, in their model formulation the authors consider regression coefficients that are constant with respect to the mRNAs, while our formulation allows a more efficient way of selecting gene-miRNA pairs. Also, in order to achieve posterior inference, we implement a full MCMC procedure while Huang, Morris and Frey (2007) adopt a variational method that only approximates the posterior distribution. More importantly, Huang, Morris and Frey (2007) restrict their search algorithm to a preselected subset of possible gene-miRNA relations, which they select based on the available sequence information, therefore excluding a priori a large number of associations that could instead occur in specific experimental conditions, such as hyperthermia.

This paper is organized as follows. Section 2 introduces the experimental study and describes the available data, that is, the expression data of miRNAs and their potential mRNA targets, and the corresponding association scores. Section 3 illustrates the proposed modeling approach via a Bayesian graphical model and describes the prior model and the variable selection scheme. Section 4 describes how to perform posterior inference and Section 5 provides a detailed analysis of the miRNA regulatory network reconstruction based on the available data. Section 6 concludes the paper.

2. Neural tube defects. Neural tube defects (NTDs) are some of the most common congenital defects, with approximately 12 per day in the United States [Finnell et al. (2000)]. NTDs are generally related to failure of embryonic neural folds to fuse properly along the neuroaxis during development. Studies in both humans and animals suggest a complex genetic component to NTDs, likely involving multiple loci, together with environmental factors. MicroRNAs are believed to play important regulatory roles in mouse development and human disease [see, for example, Conrad, Barrier and Ford (2006)], although detailed regulatory mechanisms are still unknown.

In this paper we consider experimental data from a study on a very well-known developmental toxicant causing neural tube defects, hyperthermia. In the study mice are used as the animal model to study NTDs. Time-mated female C57Bl/6 mice were exposed in vivo to a 10 minute hyperthermia or control treatment on gestational day 8.5, when the neural folds are fusing to form the neural tube. Four litters were collected for each treatment at 5, 10 and 24 hours after exposure. Each litter was treated as a single biological sample. MiRNAs and mRNAs were extracted from each sample for expression analysis.

2.1. *miRNA expression levels.* As the regulatory network can be very complex, we focus on a small sets of mRNA targets with high confidence. With a limited budget available, a pilot study was performed to screen the expression profiles of most of the known ($\sim$240) mouse microRNAs based on one set of samples, for both heat shock and control at 4 different time points, and using TaqMan miRNA RTPCR assays available at the time (Applied Biosystems, Foster City, CA; provided in collaboration with Ambion, Austin, TX). Of the 240 miRNA evaluated, 50 had none or very low expression at all time points, while 86 had a 2-fold or greater change in expression in response to hyperthermia exposure at one or more time point. From this set of 86 miRNA, we chose a subset of 23 miRNA whose patterns of expression were interesting enough for further analysis and obtained replicate sample sets. The complete experiment was therefore carried out using only this set of 23 miRNAs. Results from the analysis of these data, and their biological interpretation, are clearly limited by this initial choice.

MicroRNA was extracted from each sample at each time point under each experimental condition. Two technical replicates were prepared for RTPCR quantification to confirm the technical reproducibility. In RTPCR experiments, fluorescence techniques are used to detect the amplification of miRNAs to assess their abundance. A fluorescence threshold is determined for an experiment, and the cycle number, which reaches the predetermined threshold level of log2-based fluorescence, is defined as the Ct number. An inverse linear relationship exists between Ct number and the logarithm of input quantity of the gene when the amplification efficiency is perfect [Pfaffl (2001)]. The Ct numbers of the miRNA technical replicates were averaged across the two technical replicates.

2.2. *Target prediction via sequence and structure information.* Four of the most widely used algorithms, miRanda, TargetScan, PITA and PicTar, were used in our study to retrieve the sequence and structure information for target prediction. The algorithm miRanda [see John et al. (2004) and Enright et al. (2003)] computes optimal sequence complementarity between mature microRNAs and a mRNA using a weighted dynamic programming algorithm with weights that are position-dependent and that reflect the relative importance of the $5'$ and $3'$ regions. Its alignment score is a weighted sum of match and mismatch scores for base pairs and gap penalties. The free-energy of the formation of the microRNA:mRNA duplex is used by miRanda as a filtering step. PITA of Kertesz et al. (2007) focuses on the overall effect of all potential binding sites combined together on the given UTR. Pictar [see Krek et al. (2005)] utilizes genome-wide alignment among species to take conservation into consideration. Finally, TargetScan of Lewis, Burge and Bartel (2005) utilizes two orthogonal scores, one is the total context score, and the other independent score is the probability of conserved targeting. More details on each algorithm can be found in the original papers.

The prediction scores using the four algorithms, miRanda, TargetScan, PITA and PicTar, can be obtained from the respective websites. Predictions by PicTar

were obtained from the PicTar site[2]. A zero or absent PicTar score indicates that the raw score did not exceed a prespecified threshold, that is, the algorithm suggests no indication of a regulatory association. The current release (September 2009) comprising 1,209,841 predicted microRNA target sites in 26,697 mouse gene isoforms for 491 mouse miRNAs, generated by the miRanda algorithm of John et al. (2004), was downloaded from MICRORNA.ORG; see Betel et al. (2008). PITA Catalog version 6 (31-Aug-08) was downloaded from Segal lab's website[3]. The target scores from all predictions were used in our study. TargetscanMouse 5.1 (released April 2009) was downloaded from the internet[4]. Scores for preferential conservation of the sites (Aggregate PCT) and the context of the sites within the UTR (total context score) were parsed and used in our study. Matlab scripts were written to retrieve the prediction scores from each algorithm using the RefSeq Ids of all potential targets downloaded for the 23 mouse miRNAs of interest.

2.3. *Target mRNA expression levels.*   RNA was extracted from each sample at each time point and hybridized to GE Codelink Mouse Whole Genome Microarrays (GE Healthcare Life Sciences, Piscataway, NJ). The slides were scanned and mRNA expression levels were quantified. One biological sample was not prepared properly at hour 10 in the control group, and therefore discarded.

The RefSeq Ids of the probes spotted on the Codelink microarrays were linked to the retrieved potential targets of the 23 miRNAs previously identified. The mRNAs were included in the analysis only if they were among the potential targets predicted by the four algorithms described above. Genes with missing or negative values were excluded from the analysis. The expression levels of the remaining mRNAs were then log2 transformed so that both miRNA and mRNA expression were on the log2 scale. A total of 1297 potential targets were included in the final analysis. The transformed expressions across the 3 time points were centered by subtracting their means.

3. Model.   We have available expression levels on a set of miRNAs and their potential targets. For each target we are interested in identifying a small number of regulatory associations with high confidence. We have also available sequence information for target prediction in the form of scores of regulatory associations. We propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating the expression data and, via the prior probability model, the sequence/structure information. An important aspect of our methodology is the concept of sparsity, that is, we believe that most genes are regulated by a small number of miRNAs.

---

[2]http://pictar.mdc-berlin.de/.

[3]http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html.

[4]http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=mmu_50.
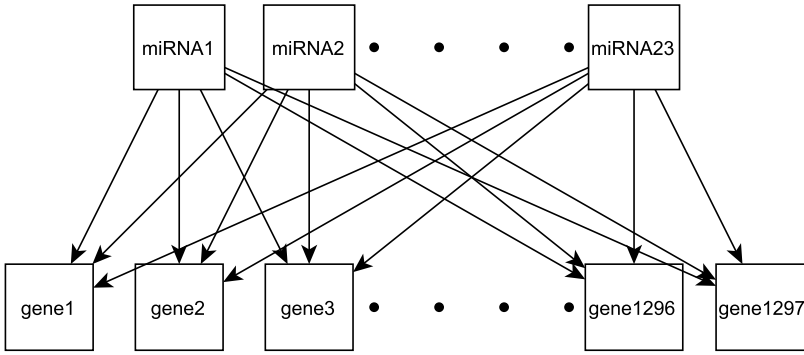
FIG. 1.   *Graphical representation of the miRNA regulatory network.*

3.1. *A Bayesian network for gene & miRNA expression.*   We use a directed graphical model (Bayesian Network) with a particular structure adapted to our data that uses a predetermined ordering of the nodes based on biological considerations. This model is able to answer to the baseline question of "*which miRNAs regulate which targets*" and, in addition, allows us to build a fast computational procedure required in such a high-dimensional framework. A graphical representation of the full miRNA network is given in Figure 1. Our task is to find a significant subset of edges.

Graphical models are graphs in which nodes represent random variables and the lack of arcs represents conditional independence assumptions; see, for example, Cowell et al. (1999). Graphical models provide a compact representation of joint probability distributions. Here we work with a multivariate normal distribution, and therefore with a Graphical Gaussian model (GGM). A graph $\mathcal{G}$ and the covariance matrix Ω entirely define a GGM $\mathcal{M}$, $\mathcal{M} \equiv (\mathcal{G}, \Omega)$. Arcs can be undirected, indicating symmetric dependencies, or directed, when there is a direction of the dependence. These dependencies can come from prior knowledge or from data analysis. Undirected graphical models have a simple definition of independence, for example, two nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C. Directed graphical models need a specific ordering of the variables. Graphs that do not allow the presence of cycles are called directed acyclic graphs (DAG). Conditional independencies in a DAG depend on the ordering of the variables.

We work with a DAG and impose an ordering of the variables such that each target can be affected only by the miRNAs and that the miRNAs can affect only the targets. Let $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G, \mathbf{X}_1, \ldots, \mathbf{X}_M)$ with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$ the matrix representing the targets and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ the miRNAs . Specifically, y_{ng} indicates the normalized averaged $\log_2$ gene expression of gene $g = 1, \ldots, G$ in sample $n = 1, \ldots, N$. These expression values are biological replicates obtained by averaging two technical replicates. Similarly, x_{nm} indicates the expression of

the mth miRNA in sample n, with $m = 1, \ldots, M$. We have $G = 1297$ and $M = 23$. In addition, we have $N = 11$ i.i.d. observations under the control status and $N = 12$ i.i.d. observations under hyperthermia. We infer the miRNA regulatory network separately under the two conditions.

Our assumptions are that $\mathbf{Z}$ is a matrix-variate normal variable with zero mean and a variance matrix Ω for its generic row, that is, following Dawid (1981) notation,

$$\mathbf{Z} - \mathbf{0} \sim \mathcal{N}(I_N, \Omega).$$

In addition, we assume that the target genes are independent conditionally upon the miRNAs, that is, $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \mathbf{X}_1, \ldots, \mathbf{X}_M$ and, without loss of generality, that the miRNAs are independent, that is, $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$. Note that the marginal distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ does not affect the regulatory network. In a Bayesian Network framework these assumptions imply an ordering of the nodes and, consequently, a likelihood factorization of the type:

$$(1) \qquad f(\mathbf{Z}) = \prod_{g=1}^{G} f(\mathbf{Y}_g | \mathbf{X}) \prod_{m=1}^{M} f(\mathbf{X}_m),$$

where $f(\mathbf{Y}_g | \mathbf{X}) \sim N(\mathbf{X}\beta_g, \sigma_g I_N)$ and $f(\mathbf{X}_m) \sim N(0, \sigma_m I_N)$, with $\beta_g = \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}_g}$ and $\sigma_g = \omega_{gg} - \Omega_{\mathbf{XY}_g}^{T} \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}_g}$. Here ω_{gg} indicates the gth diagonal element of Ω and $\Omega_{\mathbf{XX}}, \Omega_{\mathbf{XY}}$ are the blocks of the covariance matrix according to the following partition:

$$\Omega = \begin{pmatrix} \Omega_{\mathbf{YY}} & \Omega_{\mathbf{YX}} \\ \Omega_{\mathbf{XY}} & \Omega_{\mathbf{XX}} \end{pmatrix}.$$

For $m = 1, \ldots, M$ we have $\sigma_m = \omega_{mm}$.

According to current knowledge, miRNAs down-regulate gene expression. It therefore seems appropriate to include this information into our statistical model. This is achieved by specifying negative regression coefficients via the prior model. First, we note that our model is equivalent to the following system of equations:

$$(2) \qquad \begin{cases} \mathbf{Y}_1 = -\mathbf{X}\beta_1 + \varepsilon_{\sigma_1}, \\ \vdots \\ \mathbf{Y}_G = -\mathbf{X}\beta_G + \varepsilon_{\sigma_G}, \end{cases}$$

where ε_{σ_g} is distributed as a multivariate normal with zero mean and covariance matrix $\sigma_g I_N$. Then, we complete the model specification by specifying prior distributions on the regressions coefficients and the error variances. We impose our biological constraints by using Gamma distribution priors for the positive regressions coefficients, $(\beta_{gm} | \sigma_g) \sim Ga(1, c\sigma_g)$, and Inverse-Gamma distributions for the error variances, $\sigma_g^{-1} \sim Ga((\delta + M)/2, d/2)$. Figure 2 shows a graphical representation of our model. Circles indicate parameters and squares observed random variables. The parameters $\mathbf{R}$ and τ are involved in the variable selection and are introduced in the section below.
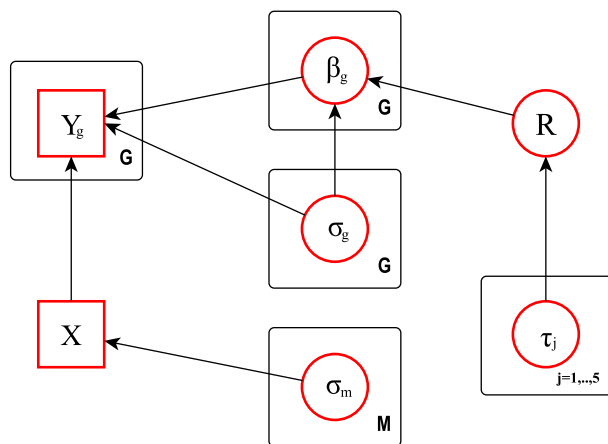
FIG. 2. *Structure of the graphical model.*

3.2. *Prior model for variable selection.* The goal of the analysis is to find, for each target, a small subset of miRNAs that regulate that target with high probability. This can be framed into a variable selection problem. Specifically, we can introduce a $(G \times M)$ matrix $\mathbf{R}$ with elements $r_{gm} = 1$ if the mth miRNA is included in the regression of the gth target and $r_{gm} = 0$ otherwise. Conditioned upon $\mathbf{R}$, expression (2) is equivalent to a system of linear equations where the included regressors are only those miRNAs corresponding to $r_{gm} = 1$. To emphasize the variable selection nature of our model, we write it as follows:

$$(3) \qquad \mathbf{Y}_g = -\mathbf{X}_{(R)}\beta_{g(R)} + \varepsilon_{\sigma_g},$$

where $\beta_{g(R)}$ is the vector that is formed by taking only the nonzero elements of β_g and $\mathbf{X}_{(R)}$ is the matrix that is formed by taking only the corresponding columns of $\mathbf{X}$. The goal of our modeling is to infer which elements of the vectors β_g's are non-zero, indicating a relationship between the corresponding genes and miRNAs. This underlying regulatory network is encoded by the association matrix $\mathbf{R} = \{r_{gm}\}$. The elements of the vectors β_g's are then stochastically independent, given the regulatory network $\mathbf{R}$, and have the following mixture prior distribution:

$$(4) \qquad \pi(\beta_{gm}|\sigma_g, r_{gm}) = r_{gm}Ga(1, c\sigma_g) + (1 - r_{gm})I_{[\beta_{gm}=0]}.$$

In addition, taking into account the regulatory network, we obtain that $\sigma_g^{-1}|\mathbf{R} \sim Ga((\delta + k_g)/2, d/2)$, where k_g is the number of significative miRNAs in the regression of the gth target.

Mixture priors have been used extensively for variable selection in linear regression settings; see George and McCulloch (1993) for univariate regression and Brown, Vannucci and Fearn (1998) and Sha et al. (2004) for multivariate models. According to prior (4), when $r_{gm} = 0$, then β_{gm} is estimated by 0 and the corresponding column of $\mathbf{X}$ is excluded from the gth equation in model (2). Notice that

the dimensions of the matrix $\mathbf{X}$ are such that there are many more columns than rows. In the domain of classical regression, this results in insufficient degrees of freedom to fit the model unless constraints are placed on the regression coefficients β_g's. Conversely, this problem is readily addressed in the Bayesian paradigm and is known as the "small n, large p" framework. The variable selection formulation we adopt here overcomes the somehow rigid structure of the model in Brown, Vannucci and Fearn (1998), which does not allow to select different predictors for different responses. See also Monni and Tadesse (2009) for an approach based on partition models.

3.3. *Using association scores in the prior model.* Scores of possible associations between gene-miRNA pairs obtained from sequence/structure information were used to estimate prior probabilities of miRNAs binding to their target genes. Let s_{gm} denote a generic score for gene g and miRNA m, obtained, for example, by the PicTar algorithm. As previously described, s_{gm} is either positive or, in the case of a regulatory association that is believed to be absent, equal to zero. Also, the PicTar algorithm shrinks small values to zero, setting $s_{gm} = 0$ if $s_{gm} < \xi$ where ξ is a prespecified threshold used by the algorithm. In our model the Bernoulli random variable r_{gm} indicates whether there is a relationship between gene g and miRNA m. We choose to model the success probability of r_{gm} as a function of the s_{gm} score as follows:

$$(5) \qquad P(r_{gm} = 1|\tau) = \frac{\exp[\eta + \tau s_{gm}]}{1 + \exp[\eta + \tau s_{gm}]},$$

where τ is an unknown parameter. We then assume that the elements of $\mathbf{R}$ are stochastically independent given τ. Notice that for $s_{gm} = 0$, we have that $P(r_{gm} = 1) = \exp[\eta]/(1 + \exp[\eta])$, which gives a 0.5 prior probability when $\eta = 0$. Thus, the inverse logit transformation of η can be interpreted as the false negative rate associated with the PicTar thresholding scheme. For a score $s_{gm} > 0$ we have $P(r_{gm} = 1) > \eta$, with higher scores yielding higher prior probabilities of association. We further specify a hyperprior on τ as a gamma distribution $\tau \sim Ga(a_\tau, b_\tau)$, ensuring the positivity of the parameter.

Since we have available multiple prior sources of information, from different sequence/structure algorithms, it makes sense to combine them all by incorporating all scores into the prior distribution via additional τ parameters. For example, in the application of Section 5 we combine five different scores as

$$P(r_{gm} = 1|\tau) = \frac{\exp[\eta + \tau_1 s_{gm}^1 + \tau_2 s_{gm}^2 + \tau_3 s_{gm}^3 + \tau_4 s_{gm}^4 + \tau_5 s_{gm}^5]}{1 + \exp[\eta + \tau_1 s_{gm}^1 + \tau_2 s_{gm}^2 + \tau_3 s_{gm}^3 + \tau_4 s_{gm}^4 + \tau_5 s_{gm}^5]},$$

with $\tau = (\tau_1, \ldots, \tau_5)$ and where the s_{gm}^j's, with $j = 1, \ldots, 5$, denote the PicTar, miRanda, aggregate Target Scan, total Target Scan and PITA scores, respectively. Scores should be normalized to obtain positive values that lie in the same range, with bigger values corresponding to stronger prior connections.

3.4. *Time-dependent coefficients model.* The previous model implies that the relation between gene g and miRNA m is constant over time. In the experimental study for which we developed our model there is no dependence between the measurements at different time points, since these observations come from independent units. However, one may still wish to incorporate into the model the fact that relations may possibly change with time. This can be done by allowing different regression coefficients at different time points, as follows:

(6)
$$\begin{cases} \mathbf{Y}_1 = -\mathbf{X}\beta_1 - \mathbf{X}_2^*\beta_1' - \mathbf{X}_3^*\beta_1'' + \varepsilon_{\sigma_1}, \\ \vdots \\ \mathbf{Y}_G = -\mathbf{X}\beta_G - \mathbf{X}_2^*\beta_G' - \mathbf{X}_3^*\beta_G'' + \varepsilon_{\sigma_G}, \end{cases}$$

where the $\mathbf{Y}_g$'s are $N \times 1$ vectors and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{pmatrix}, \qquad \mathbf{X}_2^* = \begin{pmatrix} \mathbf{0} \\ \mathbf{X}_2 \\ \mathbf{0} \end{pmatrix}, \qquad \mathbf{X}_3^* = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{X}_3 \end{pmatrix},$$

are the $N \times M$ matrices of the observed values, with $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ the miRNA expressions collected at the first, the second and the third time point, respectively. The element $\beta_{gm} \in \beta_g$ represents the relation between gene g and miRNA m at the first time point, $\beta_{gm} + \beta_{gm}'$, with $\beta_{gm}' \in \beta_g'$, represents the relation at the second time point and $\beta_{gm} + \beta_{gm}''$, with $\beta_{gm}'' \in \beta_g''$, at the third time point.

In order to do variable selection on the elements of β_g' and β_g'', we introduce two additional binary matrices $\mathbf{R}'$ and $\mathbf{R}''$, with a similar role to $\mathbf{R}$ in the time-invariant model (3). We consider the elements of $\mathbf{R}'$ and $\mathbf{R}''$ independently distributed and following a Bernoulli distribution with parameter $P(r_{gm}' = 1) = \eta_b = P(r_{gm}'' = 1)$. Because of the way we implement the MCMC (see Section 4), we do not need to impose the sequence information on the prior on $\mathbf{R}'$ and $\mathbf{R}''$.

As for the elements of the β_g's vectors, we assume that the elements of the β_g''s and β_g'''s vectors are stochastically independent given the regulatory networks $\mathbf{R}'$ and $\mathbf{R}''$, respectively, and that they have the following prior distributions:

$$\pi(\beta_{gm}' | \sigma_g, r_{gm}') = r_{gm}' N(0, c^{-1}\sigma_g\zeta) + (1 - r_{gm}')I_{[\beta_{gm}'=0]},$$

$$\pi(\beta_{gm}'' | \sigma_g, r_{gm}'') = r_{gm}'' N(0, c^{-1}\sigma_g\zeta) + (1 - r_{gm}'')I_{[\beta_{gm}''=0]},$$

where the hyperparameter ζ, usually ≤ 1, reflects the prior information on the magnitude of the β_g''s and β_g'''s.

We can reframe the time-dependent coefficients model in the same way we have framed model (3), that is,

$$\mathbf{Y}_g = -\mathbf{X}_{(R)}\beta_{g(R)} - \mathbf{X}_{2(R')}^*\beta_{g(R')}' - \mathbf{X}_{3(R'')}^*\beta_{g(R'')}'' + \varepsilon_{\sigma_g},$$

where the columns of $\mathbf{X}_2^*$ are selected if the corresponding elements of $\mathbf{R}'$ are equal to 1 and the columns of $\mathbf{X}_3^*$ are selected if the corresponding elements of $\mathbf{R}''$ are equal to 1, for each equation.

4. Posterior inference.   For posterior inference the primary interest is in estimating the association matrix $\mathbf{R}$. Here we show that $\mathbf{R}$ can be estimated by designing a simple extension of the stochastic search procedures used for variable selection; see George and McCulloch (1993) and Sha et al. (2004), among many others.

We use a Metropolis–Hastings within Gibbs to explore the huge model space and find the most influential predictors. Our model has 23 regressors for each of 1297 equations, that is a total of 29,831 regression coefficients for the time invariant model (3) and 89,493 for the time dependent model (6). Clearly, exploring such a huge posterior space is challenging. Here we exploit the sparsity of our model, that is, the belief that most of the genes are well predicted by a small number of regressors, and resort to a Stochastic Search Variable Selection (SSVS) method. A stochastic search allows us to explore the posterior space in an effective way, quickly finding the most probable configurations, that is, those corresponding to the coefficients that have high marginal probability of $r_{gm} = 1$, while spending less time in regions with low posterior probability.

In order to design this MCMC search, we need to calculate the marginal posterior distribution of $\mathbf{R}$ by integrating out β_g from the posterior:

$$f(\mathbf{Y}_g|\mathbf{X}_{(R)}, \sigma_g, \mathbf{R}) \propto \frac{1}{(2\pi)^{(N-k_g)/2}\sigma_g^{N/2}c^{k_g}}|U_g|^{1/2}$$

$$\times \exp\left[\frac{1}{2\sigma_g}q_g\right]\Phi_{k_g}(0; -U_gC_g, \sigma_g U_g),$$

where $U_g = (\mathbf{X^T}_{(R)}\mathbf{X}_{(R)})^{-1}$, $C_g = \mathbf{Y}_g^T\mathbf{X}^T_{(R)} - (\sigma_g^{1/2}/c)\mathbf{1}_{k_g}$ and $q_g = \mathbf{Y}_g^T\mathbf{Y}_g - C_gU_gC_g^T$ and with k_g the number of selected regressors. Here $\Phi_{k_g}(0; -U_gC_g, \sigma_g U_g)$ indicates the cdf of a multivariate normal, with mean $-U_gC_g$ and covariance matrix $\sigma_g U_g$, calculated at the zero vector.

Our algorithm consists of three steps. The first step is based on the marginal posterior distribution conditioned upon $\tau_1, \ldots, \tau_5$ and σ_g and consists of either the addition or the deletion of one arrow in our graphical model or the swapping of two arrows. The second step generates new values of τ_j's from their posterior distribution. In the last step values of all the error variances σ_g are updated. The unnormalized full conditionals needed for the Gibbs sampler can be derived from the conditional independencies of our model, as given in Figure 2. We now describe the three steps of the algorithm:

1. We use one of two types of moves to update $\mathbf{R}$:

   • with probability ϕ, we add or delete an element by choosing at random one component in the current $\mathbf{R}$ and changing its value;
   • with probability $1 - \phi$, we swap two elements by choosing independently at random one 0 and one 1 in the current $\mathbf{R}$ and changing the value of both of them.

The proposed $\mathbf{R}^{\text{new}}$ is then accepted with a probability that is the ratio of the relative posterior probabilities of the new versus the current model:

$$(7) \qquad \min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R^{\text{new}})}, \mathbf{R}^{\text{new}}, \sigma_g)\pi(\mathbf{R}^{\text{new}}|\tau)}{f(\mathbf{Y}|\mathbf{X}_{(R^{\text{old}})}, \mathbf{R}^{\text{old}}, \sigma_g)\pi(\mathbf{R}^{\text{old}}|\tau)}, 1\right].$$

Because these moves are symmetric, the proposal distribution does not appear in the previous ratio.

2. In order to update the τ_j's, we employ Metropolis steps. The proposal is made via a truncated normal random walk kernel. The proposed τ_j^{new} is then accepted with probability

$$(8) \qquad \min\left[\frac{\pi(\mathbf{R}|\tau_j^{\text{new}})\pi(\tau_j^{\text{new}})q(\tau_j^{\text{old}}; \tau_j^{\text{new}})}{\pi(\mathbf{R}|\tau_j^{\text{old}})\pi(\tau_j^{\text{old}})q(\tau_j^{\text{new}}; \tau_j^{\text{old}})}, 1\right],$$

where $q(\tau_j^{\text{old}}; \tau_j^{\text{new}})$ is a truncated normal with mean τ_j^{new} and truncation at 0, given the constraint of positivity on τ_j. The variance of this distribution represents the tuning parameter and has to be set in such a way to explore the parameter space and have a good acceptance rate; see also Section 5.

3. For $g = 1, \ldots, G$, we update the error variance σ_g using a Metropolis step where the proposal distribution $q(\sigma_g^{\text{old}}; \sigma_g^{\text{new}})$ is a Gamma distribution with parameters a_σ and b_σ. The proposed new value is then accepted with probability

$$(9) \qquad \min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R)}, \mathbf{R}, \sigma_g^{\text{new}})\pi(\sigma_g^{\text{new}})q(\sigma_g^{\text{old}}; \sigma_g^{\text{new}})}{f(\mathbf{Y}|\mathbf{X}_{(R)}, \mathbf{R}, \sigma_g^{\text{old}})\pi(\sigma_g^{\text{old}})q(\sigma_g^{\text{new}}; \sigma_g^{\text{old}})}, 1\right].$$

To obtain an efficient exploration of the parameter space we set $a_\sigma = \sigma_g^{\text{old}}/b_\sigma$ and $b_\sigma = e/\sigma_g^{\text{old}}$, where e represents the variance of the proposal distribution and can be set to obtain a suitable acceptance ratio.

Posterior inference can then be performed based on the MCMC output using the marginal probabilities of the singles r_{gm}'s.

The MCMC algorithm for the time-dependent coefficient model (6) is pretty similar to the procedure described above, the main difference being that at the first step we update either $\mathbf{R}$, $\mathbf{R}'$ or $\mathbf{R}''$. We then derive the marginal posterior distribution $f(\mathbf{Y}_g|\mathbf{X}_{(R)}, \mathbf{R})$ for the time dependent model obtaining

$$f(\mathbf{Y}_g|\mathbf{X}_{(R)}, \mathbf{X}_{2(R')}^*, \mathbf{X}_{3(R'')}^*, \mathbf{R}, \mathbf{R}', \mathbf{R}'', \sigma_g)$$
$$= (2\pi)^{-(n-k_g)/2}\sigma_g^{-n/2}c^{-k_g-(k_{2g}+k_{3g})/2}\zeta^{-(k_{2g}+k_{3g})/2}|A_g|^{-1/2}|C_g|^{-1/2}$$
$$\times |E_g|^{-1/2}\exp\left[\frac{1}{2\sigma_g}q_g\right]\Phi_{k_g}(0; -E_g^{-1}F_g, \sigma_g E_g^{-1}),$$

with

$$q_g = \mathbf{Y}_g^T\mathbf{Y}_g - \mathbf{Y}_{2g}^T\mathbf{X}_{2(R')}A_g^{-1}\mathbf{X}_{2(R')}^T\mathbf{Y}_{2g} - \mathbf{Y}_{3g}^T\mathbf{X}_{3(R'')}C_g^{-1}\mathbf{X}_{3(R'')}^T\mathbf{Y}_{3g}$$
$$- F_g^T E_g^{-1} F_g,$$

$$F_g = -\mathbf{X}_{(R)}^T \mathbf{Y}_g + \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')} C_g^{-1} \mathbf{X}_{3(R'')}^T \mathbf{Y}_{3g} + \mathbf{X}_{2(R)}^T \mathbf{X}_{2(R')} A_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{Y}_{2g}$$
$$\qquad - \sigma_g^{1/2} c^{-1} \mathbf{1}_{k_g},$$
$$E_g = \mathbf{X}\mathbf{X}^T - \mathbf{X}_{2(R)}^T \mathbf{X}_{2(R')} A_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)} - \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')} C_g^{-1} \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R)},$$
$$A_g = \big(\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')} + (c\zeta)^{-1} I_{k_{2g}}\big),$$
$$C_g = \big(\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')} + (c\zeta)^{-1} I_{k_{3g}}\big)$$

and $\mathbf{Y}_g^T = (\mathbf{Y}_{1g}^T, \mathbf{Y}_{2g}^T, \mathbf{Y}_{3g}^T)$; k_{2g} and k_{3g} are the number of selected β_{gm}' and β_{gm}''. We can now write the first step of the MCMC as follows:

1′. We first select which of the three matrices to update. We choose $\mathbf{R}$ with probability λ and $\mathbf{R}'$ (or $\mathbf{R}''$) with probability $(1 - \lambda)/2$. We then use the same add/delete or swap scheme described above and we accept the proposed $\mathbf{R}^{\text{new}}$ (or $\mathbf{R}'^{\text{new}}$ or $\mathbf{R}''^{\text{new}}$). For $\mathbf{R}$ the acceptance probability is

$$\min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R^{\text{new}})}, \mathbf{X}_{2(R'^{\text{old}})}^*, \mathbf{X}_{3(R''^{\text{old}})}^*, \mathbf{R}^{\text{new}}, \mathbf{R}'^{\text{old}}, \mathbf{R}''^{\text{old}}) \pi(\mathbf{R}^{\text{new}}|\tau)}{f(\mathbf{Y}|\mathbf{X}_{(R^{\text{old}})}, \mathbf{X}_{2(R'^{\text{old}})}^*, \mathbf{X}_{3(R''^{\text{old}})}^*, \mathbf{R}^{\text{old}}, \mathbf{R}'^{\text{old}}, \mathbf{R}''^{\text{old}}) \pi(\mathbf{R}^{\text{old}}|\tau)}, 1 \right]$$

and similarly if $\mathbf{R}'$ or $\mathbf{R}''$ is selected. Note that to perform this step we need to use only the prior distribution of the selected matrix.

This algorithm can be run either without any constraint on the moves relative to $\mathbf{R}$, $\mathbf{R}'$ and $\mathbf{R}''$ or with the constraint that the elements of $\mathbf{R}'$ (or $\mathbf{R}''$) can be selected only when the corresponding element of $\mathbf{R}$ is already selected and that the elements of $\mathbf{R}$ can be eliminated only when the corresponding element of $\mathbf{R}'$ and $\mathbf{R}''$ are not selected. For our application we adopted the constraint strategy. To implement this, we do not need to add the ratio of the proposal distributions into (7), since we use symmetric moves. This choice, jointly with some empirical results (not reported here), led us to not use association scores into the prior distribution of $\mathbf{R}'$ and $\mathbf{R}''$, because the selecting constraints imply that the prior probability of selecting the generic element r_{gm}' (or r_{gm}'') already depends on the association scores information through the prior probability on the corresponding element r_{gm}. This also implies a faster computational procedure in comparison with the option of including the external information into the prior of $\mathbf{R}'$ and $\mathbf{R}''$.

5. Neural tube defects application. We now apply our model to analyze the data described in Section 2, combining miRNA and mRNA expression levels with sequence information. Our model allows us to identify significant miRNAs for each target, possibly along the time.

5.1. *Parameter settings.* We first need to set the values of the hyperparameters of the model. The parameter c of the prior distribution of the regression coefficients

β_{gm} can be interpreted as a correction factor. Since truncating at zero a zero mean normal distribution with variance σ^2 results in a half normal distribution with variance $\approx 0.7\sigma^2$, we decided to set $c = 0.7$. Also, we specified a vague prior on σ_g by setting $\delta = 3$, the minimum value such that the expectation of Ω exists, and chose $c = 0.2$, setting the expected value of the variance parameter σ_g comparable in size to a small percentage of the expected error variances of the standardized $\mathbf{Y}$ given $\mathbf{X}$.

In our variable selection framework, the parameter η of the Bernoulli distribution (5) reflects the prior belief about the percentage of significant coefficients in the model. In this application, having 23 regressors for each of the 1297 equations, we set $\eta = -3$ to obtain a prior expected number of regressors approximately equal to 1. This setting also corresponds to a 5% prior probability of selection. In Section 5.2 we show that, even though η affects the number of selected coefficients, no sensitivity to the posterior selection of the most influential arcs is observed. For the more computationally expensive time dependent model, we set $\eta = -3$ and $\eta_b = 0.05$, to avoid memory problems. We also set the hyperparameters $a_\tau = 1.5$ and $b_\tau = 0.2$ to obtain a Gamma distribution that gives high probability to a broad set of values of $\tau_1, \ldots, \tau_5$, taking into account the scale of values that come from PicTar and the other algorithms. However, the posterior distributions we obtained, in all the different chains we ran, showed that the parameter setting of the Gamma distribution is not strongly informative. When running MCMCs we have set the variance of the truncated normal proposal distribution of τ_j equal to 0.01 to obtain an acceptance rate close to 25%.

We ran two different chains for each of the four possible combinations, the time invariant model for the control and the hyperthermia group and the time dependent model for the control and the hyperthermia group. We used either adding/deleting or swapping moves with equal probability at each step of the chain; we assigned a probability of $\lambda = 0.5$ to the move that updates $\mathbf{R}$ and then probability 0.25 to each of the moves that update $\mathbf{R}'$ and $\mathbf{R}''$. In all cases, after the initial burn-in, both chains mostly explored the same region of the parameter space corresponding to configuration of $\mathbf{R}$ with high posterior probability. In general, we found good agreement between the two chains, which were run from different starting points. To be more precise, correlations between the posterior probabilities of the two chains ranged from 0.89 to 0.91.

Figure 3 gives the summary trace plots for the number of selected coefficients β_{gm} and corresponding log-posterior probabilities for the time invariant model on the hyperthermia group. In this case the chain was run for three million iterations, from a starting randomly chosen set of 1000 arrows, and mostly visited models with 1000–1200 edges, that is, on average roughly 1 edge per gene, a number not too far from the prior specification.

5.2. *Results.* The huge number of potential coefficients in the model implies that the weight of a single coefficient toward the posterior probability of the entire
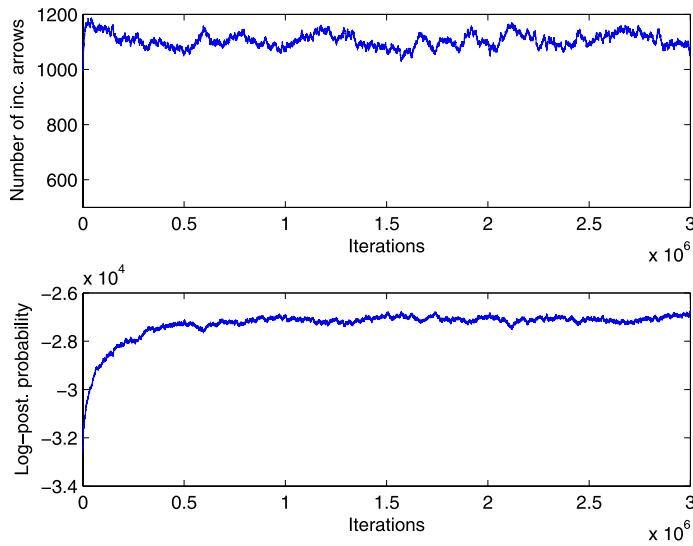
FIG. 3. *Trace plot for number of selected arrows and for the log-posterior probability for the time invariant model.*

model can be potentially very small. Also, due to sparsity, there may be many models with almost the same (small) posterior probability. Because of this, it is good practice to perform posterior inference based on the marginal posterior probability of the single coefficients, rather than on their joint distribution. These posterior probabilities of the presence of single interactions, that is, $P(r_{gm} = 1|\mathbf{Y}, \mathbf{X})$, can be estimated directly from the MCMC samples by taking the proportion of MCMC iterations for which $r_{gm} = 1$.

The small sample size of our experimental groups does not allow us to create a validation set and, therefore, all the samples are used to fit the model. Selected models are then evaluated based on the R^2 statistic, calculated using the posterior mean of regression coefficients. As expected, when more covariates are included into the model, based on their posterior probabilities, the statistic R^2 increases. We observed that coefficients with highest posterior probability explain most of the variability, while the increment in R^2 becomes marginal when adding coefficients with relatively low posterior probability. We take this as an indication of the fact that the ordering created by the posterior probabilities correctly maps the significant variables. For the time invariant model a threshold of 0.15, corresponding to 1224 included edges, gave an R^2 of 0.36, for the control group, and of 0.44 for the hyperthermia group, with 1473 included edges. Identical behavior was observed for the additional coefficients of the time dependent model, that is, when the number of included β'''s and β''''s increases, then the quality of the fitting improves; with a threshold of 0.15 for β's and a threshold of 0.1 for β'''s and β''''s, we obtain a $R^2 = 0.45$ for the control group, including 1826 β's, 366 β'''s and 222

β'''s, and a $R^2 = 0.47$ for the hyperthermia group, including 2173 β's, 439 β'''s and 296 β''''s.

In an effort to assess whether our model correctly selects miRNAs that under-regulate target genes, we also calculated the ordinary least square estimates of the regression coefficients and checked how many of them were negative; see the Appendix for the calculation of the OLS estimates. Notice that this approach does not impose the negative constraint on β's. By including all coefficients with posterior probability greater than 0.2 (0.15), we found that, for the control and hyperthermia group, respectively, 96.4 (93.5)% and 95.0 (90.4)% of the estimated coefficients were correctly negative.

Important pairs of target genes and miRNAs can be selected as those corresponding to arrows with highest posterior probabilities. For example, by exploring the regulatory network as a function of this posterior probability of the arrows, we found that, for the time invariant model on the control group, a posterior probability cutoff of 0.8 selected 88 arrows between 88 target genes and 7 miRNAs. These correspond to an expected rate of false detection (Bayesian FDR) of 7.5%, that we calculated, following Newton et al. (2004), as

$$FDR = C(\kappa)/|J_\kappa|,$$

where $C(\kappa) = \sum_{g,m} \psi_{gm} I_{[\psi_{gm} \leq \kappa]}$ and $\psi_{gm} = 1 - P(r_{gm} = 1|\mathbf{Y}, \mathbf{X})$, with $|J_\kappa|$ the size of the list ($|J_\kappa| = \sum_{g,m} I_{[\psi_{gm} \leq \kappa]}$). We set $\kappa = 1 - k$ with k the chosen threshold (i.e., 0.8). For the same cutoff the time-dependent analysis on the control group showed 11 miRNAs with at least one target gene for a total of 107 gene targets. There were 7 miRNA and 75 gene targets in common between the time-independent and time-dependent analyses of the control data. For the hyperthermia-treated group, the time-invariant model with a 0.8 cutoff led to 93 selected arrows, between 91 target genes and 11 miRNAs, corresponding to a Bayesian FDR of 9.0%. The time-dependent analysis showed 12 miRNAs with at least one target gene for a total of 120 gene targets. There were 10 miRNA and 77 gene targets in common between the time-independent and time-dependent analysis of the hyperthermia-treated data.

Figure 4, produced using GraphExplore of Wang, Dobra and West (2004), displays the selected network for the hyperthermia group and a threshold of 0.8 on the posterior probability under the time invariant model. A close look at the pairs of target genes and miRNAs with high posterior probabilities reveals that some of the regulatory relationships seem plausible and warrant future investigation. For example, links between miR-367 and target Egr2 and Mob1, selected with posterior probability of 0.97 and 1, are particularly interesting, as described below.

Increasing the cutoff value on the posterior probabilities clearly reduces the number of selected arrows and results in lower Bayesian FDRs. For example, a cutoff value of 0.9 identified 60 arrows between 60 target genes and 6 miRNAs, corresponding to a Bayesian FDR of 4.1%, for the control group, and 50 selected
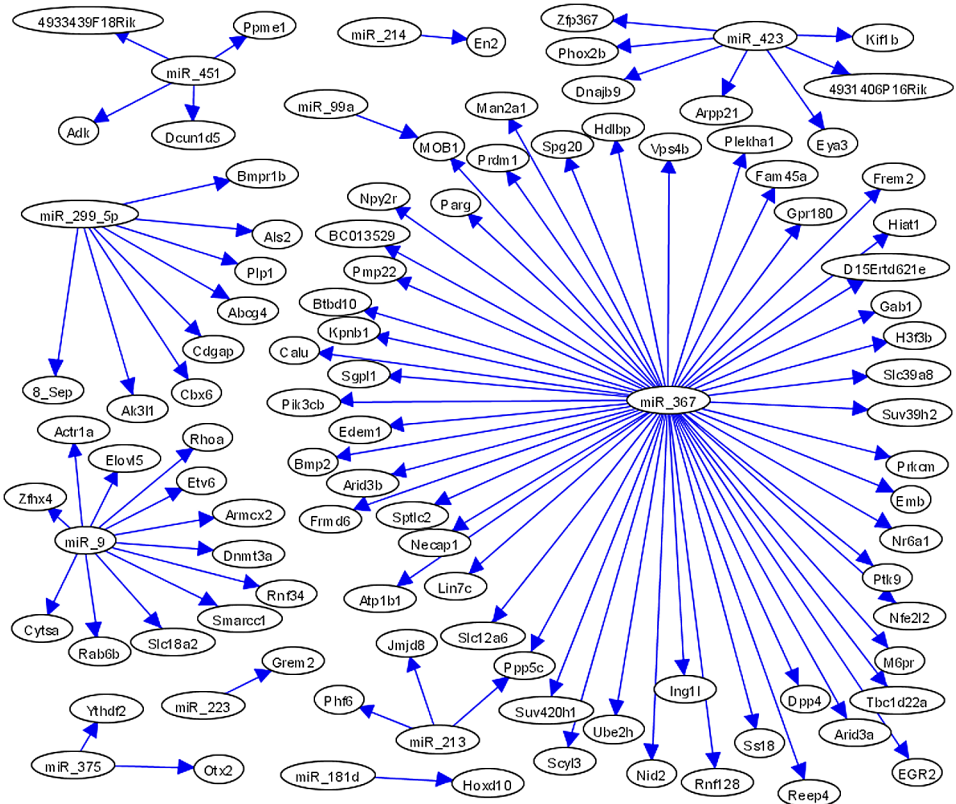
FIG. 4. *Selected network for the hyperthermia group using a threshold of* 0.8 *on the posterior probability.*

arrows, between 50 target genes and 9 miRNAs, corresponding to a Bayesian FDR of 3.9%, for the hyperthermia-treated group.

Overall, there were 252 unique miRNA-gene target associations identified with a posterior probability of at least 0.8, including 15 miRNAs and 221 gene targets. Of the 252 miRNA-gene target associations identified, 35 were predicted by miRanda only, 26 by PicTar only, zero by total Target Scan only, 14 by aggregate Target Scan only, 8 by PITA only, and 45 by at least one of the five algorithms considered. 108 of the gene targets identified were associated with miR-367, a pluripotency-specific marker in human and mouse ES cells [Li et al. (2009)], while 27 of the gene targets were associated with miR-423, which has previously been shown to be expressed in the adult and/or developing brain [Zhang and Pan (2009)]. Expression of both miR-367 and 423 decreased over time in control and hyperthermia-treated embryos, which is consistent for a marker of pluripotency in a differentiating embryo. While decreasing, the expression levels of these miRNAs were higher after hyperthermia exposure when compared to controls, which may

indicate a delay in the differentiation program. In addition, 62 of the gene targets were associated with miR-299-5p, which has been shown to regulate de novo expression of osteopontin, a protein that plays a role in enhancing proliferation and tumorigenicity [Shevde et al. (2010)].

Among the target genes identified, 12 genes associated with brain development or expressed in brain/whole embryo (including Egr2, Hnf1b, and Mob1) were associated with miR-367 with a posterior probability in 0.82–1.0 in the time-dependent analysis of hyperthermia-treatment data at the 5-hour time point. 11 of these gene target associations also had a posterior probability in 0.68–1.0 with the time-independent analysis of hyperthermia treatment data. At the 5-hour time point after hyperthermia treatment, miR-367 expression increased 1.7-fold, while expression of these associated gene targets decreased 1.1–2 fold when compared to control-treatment. This pattern of expression might be indicative of down-regulation of these gene targets by increased expression of miR-367 in response to hyperthermia treatment. Such gene-miRNA associations, identified by our methods as possibly related to brain and embryo development, are interesting to pursue in follow-up NTD studies.

It is also interesting to look at the inference on the regression coefficients. Figure 5 shows the estimates of the significant β_{gm}'s for the time invariant model under hyperthermia condition. Each bar in the plot represents the 1297 regression coefficients for one of the 23 miRNAs. Nonzero values correspond to the posterior mean estimates of the best β_{gm}'s with posterior inclusion probability above 0.15 (all other β's are estimated by zero). Notice, for example, how miRNAs miR-423, corresponding to the 20th bar, and miR-367, corresponding to the 18th bar, play an important role in the down-regulatory mechanism.

Let us now comment on the inference on $\tau_1, \ldots, \tau_5$. These parameters measure the influence of the prior information on the posterior inference. In general, we noticed that posterior inference on these parameters showed some sensitivity to the value assigned to η. When selecting edges the hyperparameter η represents the weight assigned to the data and, consequently, $\tau_1, \ldots, \tau_5$ play the role of the weight of the prior sequence information derived from the five used algorithms. The bigger the value of η, the more the posterior distribution of τ_j will be concentrated around small values. Besides this general rule, inference on the τ_j's generally depends on the concordance between data and prior information, the number of observations and the number of parameters in the model. As an example, the behavior of the posterior distribution of τ_1 (the parameter associated to PicTar), for different values of η, is summarized in Figure 6. The scale of the estimates compensates the very large values we observe for some of the PicTar scores. We can clearly see how the posterior distribution concentrates on bigger values when η decreases. To evaluate the influence of the different scores, we calculated how the prior probabilities (5) increase, on average, for a set of pairs target-miRNA with high scores. With $\eta = -3.5$ the prior probability of $r_{gm} = 1$ increase by 35.9%, while when setting $\eta = -2.5$ this increment is equal only to 8.7%.
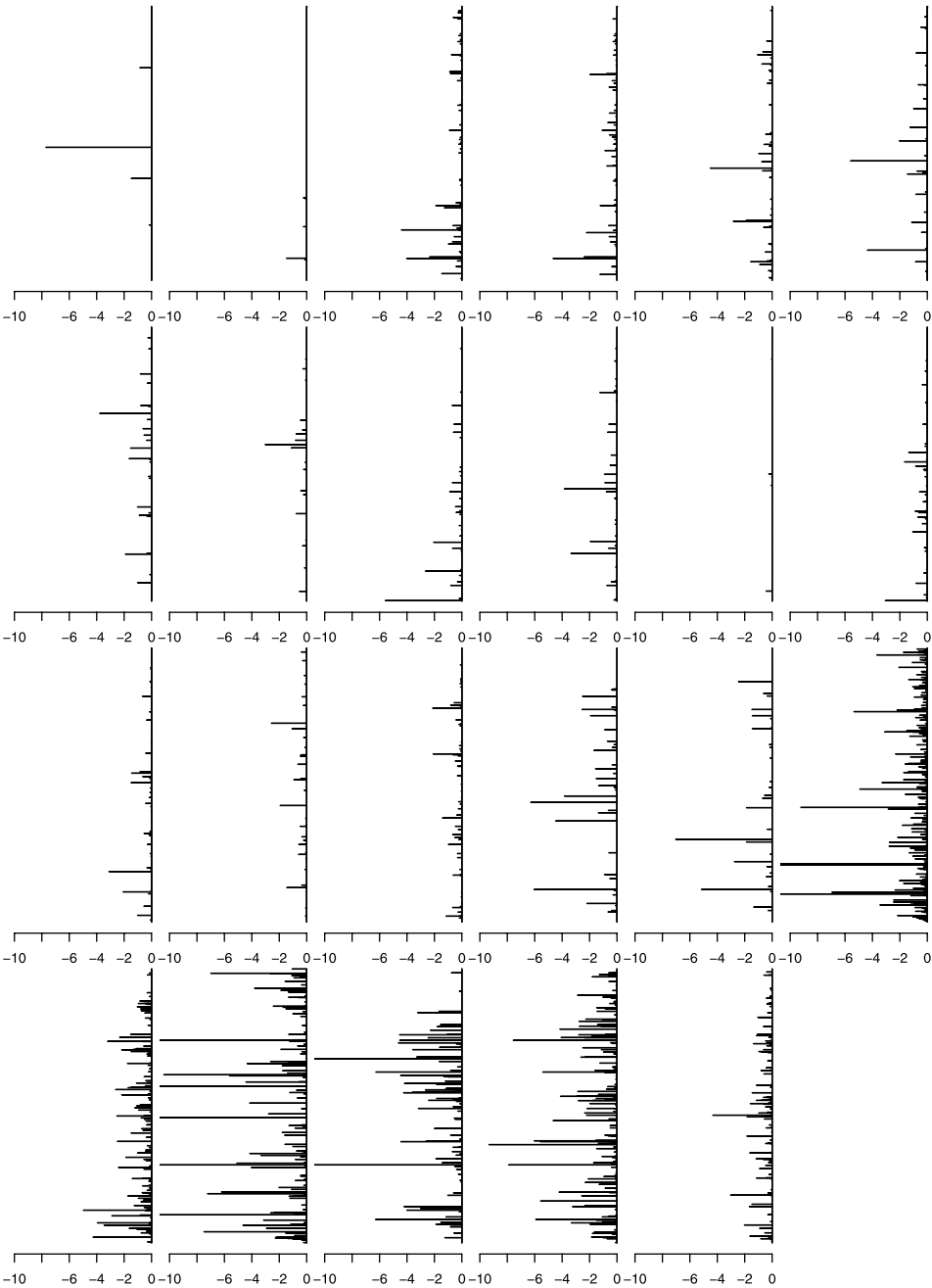
FIG. 5.   *Estimation of the significant β_{gm}'s for the time invariant model under hyperthermia condition. The y-axis indicates the* 1297 *targets, listed in the same order in all* 23 *plots.*
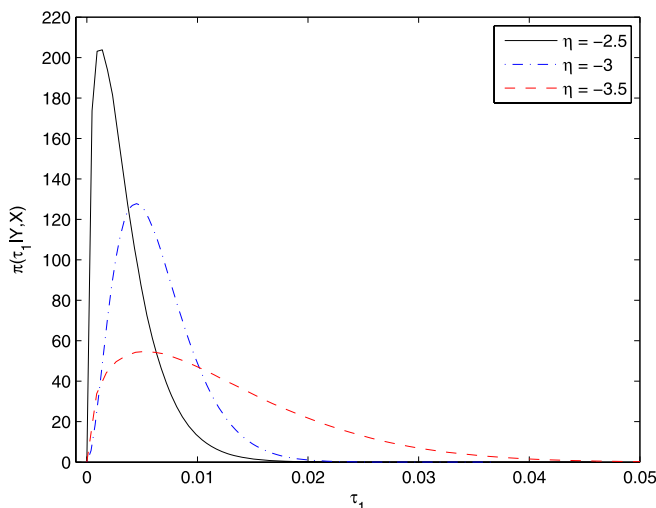
FIG. 6.   *Density Kernel estimate using the time independent model for the control group.*

Our results suggested that information extracted from PicTar is the most influential on the posterior inference. MiRanda and Target Scan aggregate also contribute somehow to the selection process, while Target Scan total and PITA do not affect the posterior inference. Figure 7 shows the posterior inference for the three most influential algorithms when $\eta = -3$. The other two algorithms resulted in posterior densities that were very concentrated around zero (plot not shown). Notice that, while the importance order does not change, the 3 algorithms are generally more influential in the case of the time dependent model for the hyperthermia group (right panel). Also, Target Scan aggregate seems to have an increased effect for this case, compared to the model with time invariant coefficients on the control group (left panel). In general, we found that with $\eta = -3$ the posterior distributions of $\tau_1, \ldots, \tau_5$, for the control group, are concentrated around values that imply a 16.4% increase on the prior probability of $r_{gm} = 1$, for high scores. For the hyperthermia group the corresponding percentage, as consequence of the increased influence of PicTar and Target Scan aggregate, increases to 37.8%. When using the time dependent model the prior probability of $r_{gm} = 1$ increases by 32.6% for the control group and, due mostly to the increased effect of PicTar and Target Scan aggregate, by 118.2% for the hyperthermia group.

Because different sequence/structure based methods do not result in exactly the same set of miRNA target predictions, we decided to perform a systematic analysis to understand how different prior target predictions can influence the final inference. Accordingly, we selected the control group and the model with time invariant coefficients and repeated our analysis by integrating in the prior formulation Pic-Tar, miRanda, Target Scan aggregate, Target Scan total and PITA scores, one set at the time. We then compared the selected arrows obtained integrating single sets
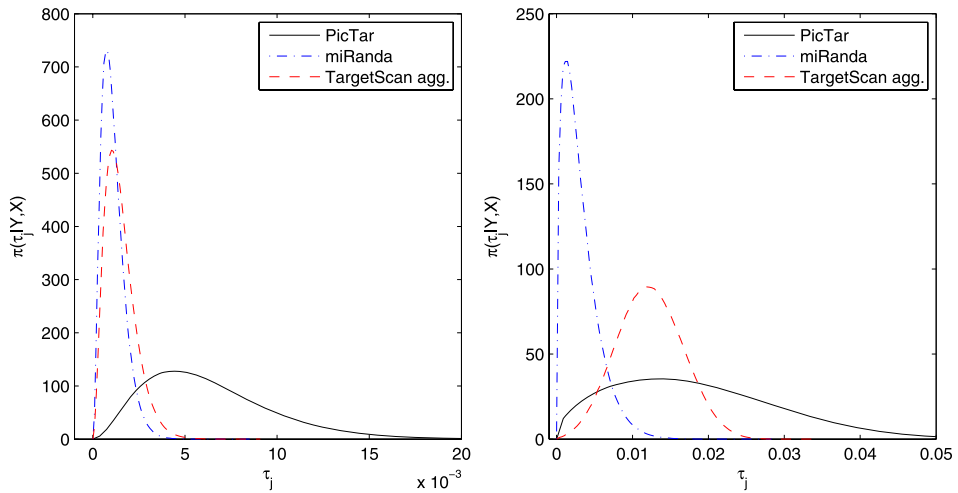
FIG. 7. *Density Kernel estimate using the time independent model for the control group* (*left panel*) *and the time dependent model for the hyperthermia group* (*right panel*).

to those selected by integrating all five available sets of scores. Using $\eta = -3$, we found that the network selection is only partially affected by the different data integrations. In particular, by considering the first 200 arrows, selected according to the posterior probability, we found that the set of selected ones using PicTar overlaps at the 73.5% with the set selected using all 5 databases; this percentage is equal to 66.0% for miRanda, to 75.0% for Target Scan aggregate, to 70.0% for Target Scan total and 72.5% for PITA. If we instead consider the first 1000 arrows, then these percentages get lower (66.4% for PicTar, 60.0% for miRanda, 68.8% for Target Scan aggregate, 65.8% for Target Scan total and 68.2% for PITA). These results indicate that although the selection is mostly data driven, especially when restricted to the most important arrows, it is also partially affected by the integrated sets.

6. Conclusions. We have proposed a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating expression levels of miR-NAs with their potential mRNA targets and, via the prior probability model, with their sequence/structure information. Our model is able to incorporate multiple data sources directly into the prior distribution, avoiding arbitrary prior data synthesis. We have used stochastic search variable selection methods to infer the miRNA regulatory network. We have considered experimental data from a study on a very well-known developmental toxicant causing neural tube defects, hyperthermia. The analysis has involved 23 mouse miRNAs and a total of 1297 potential targets. Our goal was to identify a small set of potential targets with high confidence. Some of the pairs of target gene and miRNA selected by our

model seem promising candidates for future investigation. In addition, the time-dependent model has achieved significant improvement in the percentage of explained variance, only slightly increasing the size of the selected model. Our proposed modeling strategy is general and can easily be applied to other types of network inference by integrating multiple data sources.

An interesting feature of our inference is that there is only a partial overlap between the connections selected by our model and those indicated by the sequence/structure algorithms. This phenomenon has been observed by other authors in models for data integration. Wei and Li (2008), for example, attribute this to the fact that our knowledge of biological processes is not complete and can potentially include errors and therefore induce misspecified edges on the networks. They also suggest to first check the consistency of the prior information with the available data. In our case, if the correlation between a miRNA and a target gene is very small, we may want to remove the edge from the network. On the other hand, given the limited number of observations typical of experimental studies in genomics, it would seem important to retain as much, possibly accurate, prior information as possible. This important aspect of models for data integration certainly deserves future investigation.

Extensions and generalizations of our model are possible. One future avenue we intent to pursue is trying to relax the assumption on the conditional independence of the targets given the miRNAs. This assumption is necessary in order to integrate out the covariance matrix, as in Brown, Vannucci and Fearn (1998), and still allow the selection of individual relations between a gene and a miRNA. Looking at this as a computational issue, it may be possible to still sample the values of this huge covariance matrix in the MCMC, perhaps by reducing the number of nonzero elements via the prior information on the gene network.

APPENDIX A: POSTERIOR INFERENCE ON REGRESSION COEFFICIENTS

If inference on regression coefficients is desirable, these can estimate either via the posterior distributions or the least squares estimates. For model (1) we have the following posterior distribution:

$$\pi\left(\beta_g | \mathbf{Y}, \mathbf{X}_{(R)}, \omega^2\right) \sim HN^+(U_g C_g, \sigma_g U_g), \tag{10}$$

where HN^+ indicates a k_g-variate half-normal distribution that gives positive probability only to vectors formed by elements bigger than zero.

For the more general time-dependent model we have the following posterior distributions:

$$\begin{cases} \pi\left(\beta_g | \mathbf{Y}, \mathbf{X}_{(R)}, \omega^2\right) \sim HN^+(E_g^{-1} F_g, \sigma_g E_g^{-1}), \\ \pi\left(\beta_g'' | \mathbf{Y}, \mathbf{X}_{(R)}, \omega^2\right) \sim N(J_g^{-1} H_g, \sigma_g J_g^{-1}), \end{cases} \tag{11}$$

with

$$J_g = \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')} - \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R)} L_g^{-1} \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')} + (\zeta c)^{-1} I_{k_{3g}},$$

$$H_g = \mathbf{Y}_{3g}^T \mathbf{X}_{3(R'')}$$
$$+ \left(\mathbf{Y}_g^T \mathbf{X}_{(R)} - \mathbf{Y}_{2g}^T \mathbf{X}_{2(R')} D_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)} + \sigma_g^{1/2} c^{-1} \mathbf{1}_{k_g} \right) L_g^{-1} \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')},$$

$$D_g = \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')} + (\zeta c)^{-1} I_{k_{2g}},$$

$$L_g = \mathbf{X}_{(R)}^T \mathbf{X}_{(R)} - \mathbf{X}_{2(R)}^T \mathbf{X}_{2(R')} D_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)}.$$

The posterior distribution of β' has the same form as the posterior distribution of β''. Using the least squares approach, instead, we obtain the following equations for β, β' and β'':

$$\begin{cases} \hat{\beta}_{gLS} = (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (Y_g - \mathbf{X}_{2(R')}^* \beta_g' + \mathbf{X}_{3(R'')}^* \beta_g''), \\ \hat{\beta}'_{gLS} = (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T (Y_{2g} - \mathbf{X}_{2(R)} \beta_g), \\ \hat{\beta}''_{gLS} = (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T (Y_{3g} - \mathbf{X}_{3(R)} \beta_g), \end{cases}$$

and then

$$\hat{\beta}_{gLS} = K_g^{-1} [\hat{\beta}_{gOLS} - (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (\mathbf{X}_{2(R')}^* (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T Y_{2g}$$
$$+ \mathbf{X}_{3(R'')}^* (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T Y_{3g})],$$

with

$$K_g = I_{k_g} - (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (\mathbf{X}_{2(R')}^* (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)}$$
$$+ \mathbf{X}_{3(R'')}^* (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R)}).$$

REFERENCES

BAGGA, S., BRACHT, J., HUNTER, S., MASSIRER, K., HOLTZ, J., EACHUS, R. and PASQUINELLI, A. E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122** 553–563.

BETEL, D., WILSON, M., GABOW, A., MARKS, D. S. and SANDER, C. (2008). The microRNA.org resource: Targets and expression. *Nucleic Acids Research* **36** D149–D153.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B* **60** 627–641. MR1626005

CONRAD, R., BARRIER, M. and FORD, L. P. (2006). Role of miRNA and miRNA processing factors in development and disease. *Birth Defects Research (Part C)* **78** 107–117.

COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York. MR1697175

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

ENRIGHT, A. J., JOHN, B., GAUL, U., TUSCHL, T., SANDER C. and MARKS, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biology* **5** R1.

FARH, K. K. H., GRIMSON, A., JAN, C., LEWIS, B. P., JOHNSTON, W. K., LIM, L. P., BURGE, C. B. and BARTEL, D. P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolutions. *Science* **310** 1817–1821.

FINNELL, R. H., WAES, J. G. V., BENNETT, G. D., BARBER, R. C., WLODARCZYK, B., SHAW, G. M., LAMMER, E. J., PIEDRAHITA, J. A. and EBERWINE, J. H. (2000). Genetic basis of susceptibility to environmentally induced neural tube defects. *Ann. New York Acad. Sci.* **919** 261–277.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable Selection via Gibbs sampling. *J. Amer.* **85** 398–409.

HUANG, J. C., FREY, B. J. and MORRIS, Q. D. (2008). Comparing sequence and expression for predicting microRna targets using GenMiR3. *Pacific Symposium for Biocomputing (PSB)* **13** 52–63.

HUANG, J. C., MORRIS, Q. D. and FREY, B. J. (2007). Bayesian inference of microRNA targets from sequence and expression data. *J. Comput. Biol.* **14** 550–563. MR2344257

JOHN, B., ENRIGHT, A. J., ARAVIN, A., TUSCHL, T., SANDER, C. and MARKS, D. S. (2004). Human MicroRNA targets. *PLoS Biol.* **2** e363.

KERTESZ, M., IOVINO, N., UNNERSTALL, U., GAUL, U. and SEGAL, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics* **39** 1278–1284.

KIRIAKIDOU, M., NELSON, P. T., KOURANOV, A., FITZIEV, P., BOUYIOUKOS, C., MOURELATOS, Z. and HATZIGEORGIOU, A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes and Development* **18** 1165–1178.

KREK, A., GRN, D., POY, M. N., WOLF, R., ROSENBERG, L., EPSTEIN, E. J., MACMENAMIN, P., PIEDADE, I. D., GUNSALUS, K. C., STOFFEL, M. and RAJEWSKY, N. (2005). Combinatorial microRNA target predictions. *Nature Genetics* **37** 495–500.

LEWIS, B. P., BURGE, B. and BARTEL, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120** 15–20.

LI, S. S., YU, S., KAO, L., TSAI, Z. Y., SINGH, S., CHEN B. Z., HO, B., LIU, Y. and YANG, P. (2009). Target identification of microRNAs expressed highly in human embryonic stem cells. *Journal of Cellular Biochemistry* **106** 1020–1030.

MONNI, S. and TADESSE, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Anal.* **4** 413–436.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PETERSEN, C. P., BORDELEAU, M. E., PELLETIER, J. and SHARP, P. A. (2006). Short RNAs repress translation after initiation in mammalian cells. *Molecular Cell* **4** 533–542.

PFAFFL, M. W. (2001). A new mathematical model for relative quantification real-time RT-PCR. *Nucleic Acid Research* **29** 20027.

PILLAI, R. S., BHATTACHARYA, S. N., ARTUS, C. G., ZOLLER, T., COUGOT, N., BASYUK, E., BERTRAND, E. and FILIPOWICZ, W. (2005). Inhibition of translational initiation by let-7 microRNA in human cells. *Science* **209** 1573–1576.

RAJEWSKY, N. (2006). MicroRNA target predictions in animals. *Nature Genetics* **38** S8-13.

SHA, N., VANNUCCI, M., TADESSE, M. G., BROWN, P. J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, N., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian

variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60** 812–819. MR2089459

SHEVDE, L. A., METGE, B. J., MITRA, A., XI, Y., JU, J., KING, J. A. and SAMANT, R. S. (2010). Spheroid-forming sub-population of breast cancer cells demonstrates vasculogenic mimicry via hsa-miR-299-5p regulated de novo expression of osteopontin(p). *J. Cell. Mol. Med.* **14** 1693–1706.

THERMANN, R. and HENTZE, M. W. (2007). Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. *Nature* **447** 875–879.

WANG, Q., DOBRA, A. and WEST, M. (2004). GraphExplore: A software tool for graph visualization. *ISDS Discussion Paper* **04-22**.

WEI, Z. and LI, H. (2008). A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. App. Statist.* **2** 408–429. MR2415609

YOON, S. and MICHELI, G. D. (2006). Computational identification of microRNAs and their targets. *Birth Defects Research (Part C)* **78** 118–128.

ZHANG, B. and PAN, X. (2009). RDX induces aberrant expression of microRNAs in mouse brain and liver. *Environ. Health Perspect.* **117** 213–240.

F. C. STINGO
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORENCE
50134 FLORENCE
ITALY
E-MAIL: stingo@ds.unifi.it

M. VANNUCCI
DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TEXAS 77005
USA
E-MAIL: marina@rice.edu

Y. A. CHEN
DEPARTMENT OF BIOSTATISTICS
MOFFITT CANCER CENTER
TAMPA, FLORIDA 33612
USA
E-MAIL: Ann.Chen@moffitt.org

M. BARRIER
US EPA, ORD, NHEERL
ISTD, SBB (MD-72)
RESEARCH TRIANGLE PARK,
   NORTH CAROLINA 27711
USA
E-MAIL: barrier.marianne@epa.gov

P. E. MIRKES
DEPARTMENT OF VETERINARY,
   PHYSIOLOGY & PHARMACOLOGY
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77845
USA
E-MAIL: pmirkes@gmail.com