BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2627–2635
doi:10.1093/bioinformatics/btu373

Systems biology

Advance Access publication June 3, 2014

Mirsynergy: detecting synergistic miRNA regulatory modules by overlapping neighbourhood expansion

Yue Li[1,2,†,*], Cheng Liang[3,†], Ka-Chun Wong[1,2], Jiawei Luo[3] and Zhaolei Zhang[1,2,4,*]

[1]Department of Computer Science, [2]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada, [3]College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China and [4]Banting and Best Department of Medical Research and Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 3E1, Canada

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Identification of microRNA regulatory modules (MiRMs) will aid deciphering aberrant transcriptional regulatory network in cancer but is computationally challenging. Existing methods are stochastic or require a fixed number of regulatory modules.

Results: We propose Mirsynergy, an efficient deterministic overlapping clustering algorithm adapted from a recently developed framework. Mirsynergy operates in two stages: it first forms MiRMs based on co-occurring microRNA (miRNA) targets and then expands each MiRM by greedily including (excluding) mRNAs into (from) the MiRM to maximize the synergy score, which is a function of miRNA–mRNA and gene–gene interactions. Using expression data for ovarian, breast and thyroid cancer from The Cancer Genome Atlas, we compared Mirsynergy with internal controls and existing methods. Mirsynergy-MiRMs exhibit significantly higher functional enrichment and more coherent miRNA–mRNA expression anti-correlation. Based on Kaplan–Meier survival analysis, we proposed several prognostically promising MiRMs and envisioned their utility in cancer research.

Availability and implementation: Mirsynergy is implemented/available as an R/Bioconductor package at www.cs.utoronto.ca/~yueli/Mirsynergy.html

Contact: yueli@cs.toronto.edu; zhaolei.zhang@utoronto.ca

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

Received on February 14, 2014; revised on May 28, 2014; accepted on May 29, 2014

1 INTRODUCTION

MicroRNAs (miRNAs) are ~22 nt small non-coding RNA that base pair with mRNAs primarily at the 3′ untranslated region (UTR) to cause mRNA degradation or translational repression (Bartel, 2009). Aberrant miRNA expression is implicated in tumourigenesis (Spizzo *et al.*, 2009). Based on miRBase (release 20) (Kozomara and Griffiths-Jones, 2014), the human genome encodes 1872 (2578) precursor (mature) miRNAs, which potentially target majority of the human genes (Friedman *et al.*, 2009). Although targets of individual miRNAs are significantly enriched for certain biological processes (Papadopoulos *et al.*, 2009; Tsang *et al.*, 2010), it is also likely that multiple miRNAs are coordinated together to synergistically regulate one or more pathways (Boross *et al.*, 2009; Krek *et al.*, 2005; Xu *et al.*, 2011). Despite their limited number, miRNAs may be in charge of more evolutionarily robust and potent regulatory effects through coordinated collective actions. The hypothesis of miRNA synergism is also parsimonious or biologically plausible because the number of possible combinations of the 2578 human miRNAs is extremely large to potentially react to virtually countless environmental changes. Intuitively, if a group of (miRNA) workers perform similar tasks together, then removing a single worker will not be as detrimental as assigning each worker a unique task (Boross *et al.*, 2009).

Several related methods have been developed to study miRNA synergism. Some early methods were based on pairwise overlaps (Shalgi *et al.*, 2007) or score-specific correlation (Xu *et al.*, 2011) between predicted target sites of any given two (co-expressed) miRNAs. For instance, Shalgi *et al.* (2007) devised an overlapping scoring scheme to account for differential 3′UTR lengths of the miRNA targets, which may otherwise bias the results if standard hypergeometric test was used. Methods beyond pairwise overlaps have also been described. These methods consider not only the sequence-based miRNA-target site information but also the respective miRNA–mRNA expression correlation (MiMEC) across various conditions to detect miRNA regulatory modules (MiRMs).

Briefly, Joung *et al.* (2007) developed a probabilistic search procedure to separately sample from the mRNA and miRNA pools candidate module members with probabilities proportional to their overall frequency of being chosen as the 'fittest', which is determined by their target sites and MiMEC relative to the counterparts. The algorithm finds only a single best MiRM, which varies depending on the initial mRNA and miRNA set. Peng *et al.* (2009) used a maximal bi-clique enumeration (MBE) technique to discover complete bipartite subgraphs, where every miRNA is connected with every mRNA. The approach operates on unweighted edges only, which requires discretizing MiMEC. Moreover, maximal bi-clique does not necessarily imply functional MiRMs and vice versa. Le and Bar-Joseph (2011) developed GroupMiR to form MiRMs by assigning each miRNA or mRNA (m/miRNA) one or more memberships

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

based on a non-parametric Bayesian prior called Indian Buffet Process. Although GroupMiR is statistically sound, it has time complexity cubic to the number of observations and converges slow on large datasets (Griffiths and Ghahramani, 2005).

More recently, Zhang *et al.* (2011) described a framework of sparse network-regularized multiple non-negative matrix factorization (SNMNMF) to identify miRNA–mRNA modules based on the factorized coefficient matrices. Importantly, SNMNMF uses not only the expression and target-site information but also the gene–gene interaction (GGI) derived from protein–protein interaction (PPI) and transcription factor binding sites (TFBS). To define the reduced dimensionality of the factorized matrices, however, the SNMNMF approach requires a predefined number of modules, which may be data-dependent and difficult to determine beforehand. Additionally, the solution to NMF is often not unique, and the identified modules do not necessarily include both miRNAs and mRNAs, which makes reproducing and interpreting the results difficult. Similarly, the regression-based model called PIMiM (Protein Interaction-based MicroRNA Modules) developed by Le and Bar-Joseph (2013) also requires a fixed number of unique memberships, and the non-convexity of the algorithm leads to different outcomes with random initializations. Additionally, both algorithms incur a time complexity quadratic to both the number of miRNA and mRNA multiplied by the (squared) total number of modules per iteration. We herein hypothesize that an intuitively simple and efficient deterministic framework may serve as an attractive alternative.

In this article, we propose a novel model called *Mirsynergy* that integrates m/miRNA expression profiles, target site information and GGI to form MiRMs, where an m/miRNA may participate in multiple MiRMs, and the module number is systematically determined given the predefined model parameters. The clustering algorithm of Mirsynergy adapts from ClusterONE (Nepusz *et al.*, 2012), which was intended to identify protein complex from PPI data. The ultimate goal here, however, is to construct a priori the MiRMs and exploit them to better explain clinical outcomes such as patient survival rate. We apply Mirsynergy to ovarian (OV), breast (BRCA) and thyroid cancer (THCA) datasets from TCGA (Cancer Genome Atlas Research Network, 2008). Comparing with alternative formalisms, we find that the Mirsynergy-based modules are more functionally enriched and exhibit more negative MiMEC. Through Kaplan–Meier (KM) survival analysis, we propose several prognostically promising cancer-specific MiRMs as biomarkers.

2 METHODS

2.1 Mirsynergy model overview

We formulate the construction of synergistic MiRMs as an overlapping clustering problem with two main stages. Figure 1 depicts our model schema. Before the two clustering stages, we first inferred miRNA–mRNA interaction weights (MMIW) (**W**) using m/miRNA expression data and target site information (Section 2.2). At stage 1, we only cluster miRNAs to greedily maximize miRNA–miRNA synergy, which is proportional to the correlation between miRNAs in terms of their MMIW (See Section 2.2). At stage 2, we fix the MiRM assignments and greedily add (remove) genes to (from) each MiRM to maximize the synergy score, which is defined as a function of the MMIW matrix and the gene–gene interaction weight (GGIW) matrix (**H**). The underlying clustering algorithm at both stages (Section 2.3) adapts from ClusterONE, which was originally developed for PPI network (Nepusz *et al.*, 2012).

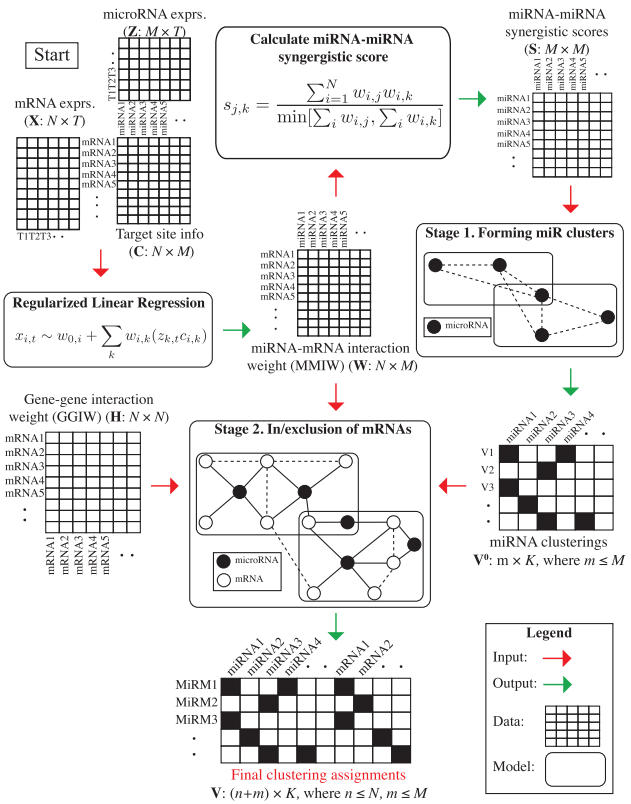

Fig. 1. Mirsynergy workflow. Given the inputs of m/miRNA expression profiles and sequence-based target site information, we first derive an expression-based MMIW matrix **W** using L1-norm regularized linear regression model (i.e. LASSO). We then compute miRNA–miRNA synergistic scores **S** based on the MMIW matrix, which is subsequently provided as input to form miRNA clusters $\mathbf{V^0}$ at Stage 1 by an overlapping neighbourhood expansion clustering algorithm. Next, we fix the MiRM assignments $\mathbf{V^0}$ and greedily add (remove) mRNA to (from) each module that maximizes the synergy score, where the edge weights are defined by the MMIW matrix and GGIW matrix involving known TFBS and PPI. The end results are overlapping clustering assignments of m/miRNAs (i.e., MiRMs)

2.2 Evaluation of expression-based prediction models

To compute MMIW as the edge-weight matrix for downstream clustering, we first evaluated four prominent expression-based miRNA target prediction algorithms, namely Pearson correlation coefficient (PCC) (Zhang *et al.*, 2011), GenMiR++ (Huang *et al.*, 2007) and two formulations of LASSO (Lu *et al.*, 2011). Specifically, PCC between each pair of m/miRNA across T samples was computed using R built-in function `cor`. Targets with negative correlation were ranked at the top. GenMiR++ was ran in Matlab with default setting using as input the binary target site matrix and the expression profiles. We implemented two versions of LASSO both using *glmnet* with $\alpha = 1$ except that the best λ was chosen using cross-validation function `cv.glmnet` (Friedman *et al.*, 2010). For the first LASSO, we used expression of miRNAs $k \in \{1, \ldots, M\}$ that have non-zero target sites to model the corresponding

mRNA expression $i \in \{1 \ldots, N\}$ across samples $t \in \{1 \ldots T\}$: $x_{i,t} \sim w_{i,0} + \sum_k w_{i,k}(z_{k,t}c_{i,k})$, where $w_{i,0}$ is the bias, $c_{i,k}$ is the number of target sites in mRNA i for miRNA k, $w_{i,k} \in \mathbf{W}_{N\times M}$ is the fitted linear coefficients used for the MMIW. For the second LASSO, namely LASSO_RISC (RNA-induced silencing complex), we used target site counts, miRNA, and Ago 1–4 mRNA expression as the input variable: $x_{i,t} \sim w_{i,0} + \sum_k w_{i,k}Ago2 z_{k,t}c_{i,k} + \sum_k w'_{i,k}Ago_{1,3,4}z_{k,t}$. The rationale behind LASSO_RISC is that Ago2 is the only protein component known to catalyse target mRNA degradation whereas the non-catalytic Ago 1,3,4 proteins may compete with Ago2 to bind to miRNAs; thus, the expression changes in the Ago 1, 3, 4 proteins will affect the ability of Ago2 to bind to miRNAs (Lu *et al.*, 2011). For LASSO and LASSO_RISC, interactions associated with negative linear coefficients $w_{i,k}$ (i.e. MMIW) were ranked at the top as the candidate interactions.

Because of the scarcity of the validated miRNA targets, conventional power analysis such as ROC (receiver operating characteristic) cannot distinguish the method performances. Instead, we assessed each method by the number of validated interactions they identified among their top ranked 1000–5000 targets (1000-interval) based on miRTarBase (Hsu *et al.*, 2011). Among the four methods, LASSO achieved the best overall performance on our test data (Supplementary Fig. S1). Accordingly, we adopted the linear coefficient matrix $w_{i,k} \in \mathbf{W}$ from LASSO as the $N \times M$ MMIW matrix and set all of its positive coefficients to zero, assuming only negative miRNA regulation.

2.3 Two-stage clustering

Let $\mathbf{W}$ denote the expression-based $N \times M$ MMIW matrix obtained from LASSO, determined as the best performing target prediction model in Section 2.2, where $w_{i,k}$ is the scoring weight for miRNA k targeting mRNA i. Similar to the 'Meet/Min' score defined by Shalgi *et al.* (2007) for binary interactions of co-occurring targets of miRNA pairs, we define an $M \times M$ scoring matrix denoted as $\mathbf{S}$, indicating miRNA–miRNA synergistic scores between miRNA j and k ($j \neq k$):

$$s_{j,k} = \frac{\sum_{i=1}^{N} w_{i,j}w_{i,k}}{\min\left[\sum_i w_{i,j}, \sum_i w_{i,k}\right]} \tag{1}$$

Notably, if $\mathbf{W}$ were a binary matrix, Equation 1 became the ratio of number of targets shared between miRNA j and k over the minimum number of targets possessed by j or k, which is essentially the original 'Meet/Min' score.

Similar to the cohesiveness defined by Nepusz *et al.* (2012), we define *synergy* score $s(V_c)$ for any given MiRM V_c as follows. Let $w^{in}(V_c)$ denote the total weights of the internal edges within the miRNA cluster, $w^{bound}(V_c)$ the total weights of the boundary edges connecting the miRNAs within V_c to the miRNAs outside V_c, and $\alpha(V_c)$ the penalty scores for forming cluster V_c. The synergy of V_c (i.e. the objective function) is as follows:

$$s(V_c) = \frac{w^{in}(V_c)}{w^{in}(V_c) + w^{bound}(V_c) + \alpha(V_c)} \tag{2}$$

where $\alpha(V_c)$ reflects our limited knowledge on potential unknown targets of the added miRNAs as well as the false-positive targets within the cluster. Presumably, these unknown factors will affect our decision on whether miRNA k belong to cluster V_c. For instance, miRNAs may target non-coding RNAs and seedless targets, which are the mRNAs with no perfect seed-match (Helwak *et al.*, 2013). We currently consider only mRNA targets with seed-match to minimize the number of false-positive results. By default, we set $\alpha(V_c) = 2|V_c|$, where $|V_c|$ is the cardinality of V_c. Additionally, we define two scoring functions to assess the overlap $\omega(V_c, V_{c'})$ between V_c and $V_{c'}$ for $c \neq c'$ and the density $d_1(V_c)$ of any given V_c:

$$\omega(V_c, V_{c'}) = \frac{|V_c \cap V_{c'}|^2}{|V_c||V_{c'}|} \tag{3}$$

$$d_1(V_c) = \frac{2w^{in}(V_c)}{m(m-1)} \tag{4}$$

where $|V_c \cap V_{c'}|$ is the total number of common elements in V_c and $V_{c'}$, and m is the number of miRNAs in V_c.

The general solution for solving an overlapping clustering problem is Non-deterministic Polynomial-time hard (NP-hard) (Barthélemy and Brucker, 2001). Thus, we adapt a greedy-based approach (Nepusz *et al.*, 2012). The algorithm can be divided into two major steps. In step 1, we select as an initial *seed* miRNA k with the highest total weights. We then grow an MiRM V_t from seed k by iteratively including boundary or excluding internal miRNAs to maximize the synergy $s(V_t)$ (Equation 2) until no more node can be added or removed to improve $s(V_t)$. We then pick another miRNA that has neither been considered as seed nor included in any previously expanded V_t to form V_{t+1}. The entire process terminates when all of the miRNAs are considered. In step 2, we treat the clusters as a graph with V_c as nodes and $\omega(V_c, V_{c'}) \geq \tau$ as edges. Here τ is a free parameter. Empirically, we observed that most MiRMs are quite distinct from one another in terms of $\omega(V_c, V_{c'})$ (before the merging) (Supplementary Fig. S2A). Accordingly, we set τ to 0.8 to ensure merging only similar MiRMs, which avoids producing large MiRMs (when τ is too small). We then perform a breath-first search to find all of the weakly connected components (CC), each containing clusters that can reach directly/indirectly to one another within the CC. We merge all of the clusters in the same CC and update the synergy score accordingly. Algorithm 1 outlines the pseudocode.

After forming MiRMs at stage 1, we perform a similar clustering procedure as in Algorithm 1 by adding (removing) *only the mRNAs* to (from) each MiRM. Different from stage 1, however, we grow each existing MiRM separately with no prioritized seed selection or cluster merging, which allows us to implement a parallel computation by taking advantage of the multicore processors in the modern computers. In growing/contracting each MiRM, we maximize the same synergy function (Equation 2) but changing the edge weight matrix from $\mathbf{S}$ to a $(N + M) \times (N + M)$ matrix by combining $\mathbf{W}$ (the $N \times M$ MMIW matrix) and $\mathbf{H}$ (the $N \times N$ GGIW matrix). Notably, here we assume miRNA–miRNA edges to be zero. Additionally, we do not add/remove miRNAs to/from the MiRM at each greedy step. Finally, we define a new density function because of the connectivity change at stage 2:

$$d_2(V_c) = \frac{w^{in}(V_c)}{n(m+n-1)} \tag{5}$$

where n (m) are the number of mRNAs (miRNAs) in V_c. By default, we filter out MiRMs with $d_1(V_j) < 1e{-}2$ and $d_2(V_j) < 5e{-}3$ at stage 1 and 2, respectively. Both density thresholds were chosen based on our empirical analyses (Supplementary Fig. S2). For some datasets, in particular, we found that our greedy approach tends to produce a large cluster involving several hundred miRNAs or several thousand mRNAs at Stage 1 or 2, respectively, which are unlikely to be biologically meaningful. Despite the ever increasing synergy (by definition), however, the anomaly modules all have low density scores, which allows us to filter them out using the above chosen thresholds.

The time complexity of the algorithm is $O(M(N + M))$: stage 1 takes $O(M^2)$ in the worst case scenario by checking every miRNA in forming an MiRM using every miRNA as a seed; given that the maximum number of MiRMs is M, stage 2 takes $O(NM)$. In our actual implementation, the total weights of each node is pre-computed so that the synergy update is $O(1)$. Moreover, we maintain a sorted list using numbers to represent m/miRNA nodes for efficient binary search of neighbour or removable nodes.

Algorithm 1 Mirsynergy stage 1 clustering procedure

▷ Step1: Grow MiRMs by overlapping neighbourhood expansion
$\mathbf{V} \leftarrow \varnothing$; $K \leftarrow \{1, \ldots, M\}$ for M miRNAs
while $K \neq \varnothing$ **do**
  $k \leftarrow \arg\max_{k' \in K} \sum_j s_{k',j}$ ▷ k: miRNA with max total synergistic scores
  $V_t^* \leftarrow \{k\}$; $V_t \leftarrow \varnothing$
  while $V_t^* \neq V_t$ **do** ▷ Stop if cur. cluster is same as prev. cluster
    $V_t \leftarrow V_t^*$
    ▷ Choose the best neighbour miRNA to add
    $j' \leftarrow \arg\max_{j' \notin V_t} [s(V_t \cup j')]$, where $w_{j',k} \neq 0$ for $k \in V_t$
    ▷ Choose the best miRNA in the cluster to remove
    $j \leftarrow \arg\max_{j \in V_t} [s(V_t \setminus j)]$
    ▷ Grow, contract, or do nothing, whichever produces the highest synergy
    if $s(V_t \cup j') > \max[s(V_t \setminus j), s(V_t)]$ **then**
    $V_t^* \leftarrow V_t \cup \{j\}$
    else if $s(V_t \setminus j) > \max[s(V_t \cup j'), s(V_t)]$ **then**
      $V_t^* \leftarrow V_t \setminus j$
      end if
  end while
  $\mathbf{V} \leftarrow \mathbf{V} \cup \{V_t\}$
  $K \leftarrow K \setminus \{k \in V_t\}$
end while
▷ Step2: Merging MiRMs by breath-first search
for all $V_c \in \mathbf{V}$ **do**
  $\mathbf{C} \leftarrow \{V_c\}$; $\mathbf{V} \leftarrow \mathbf{V} \setminus \{V_c\}$
  ▷ Compare each cluster V_c with every other cluster $V_{c'}$ $(c \neq c')$
  for all $V_{c'} \subset \mathbf{V}$ **do**
    if $\omega(V_c, V_{c'}) \geq \tau$ **then**
      $\mathbf{C} \leftarrow \mathbf{C} \cup \{V_{c'}\}$; $\mathbf{V} \leftarrow \mathbf{V} \setminus \{V_{c'}\}$
      end if
  end for
  $V_t^* \leftarrow \varnothing$
  ▷ Merge all of the clusters that are in $\mathbf{C}$
  for all $V_k \subset \mathbf{C}$ **do**
    $V_t^* \leftarrow V_t^* \cup V_k$
  end for
  $\mathbf{V}^* \leftarrow \mathbf{V}^* \cup \{V_t^*\}$
end for
$\mathbf{V} \leftarrow \mathbf{V}^*$
return $\mathbf{V}$

2.4 Method comparisons

To compare the performance of Mirsynergy, we applied SNMNMF (Zhang *et al.*, 2011) to the same testing data. SNMNMF is one of the most cited works in the recent literature and has been shown to outperform the MBE approach introduced earlier by Peng *et al.* (2009). SNMNMF was implemented in Matlab with source code available. As suggested by the authors, we set the required module number to 50 and used the default settings throughout the tests. Additionally, we also compared our results (using the same test dataset) with the published results from another recent algorithm called PIMiM developed by Le and Bar-Joseph (2013), which was shown to outperform SNMNMF, but the software is not publicly available.

2.5 Functional enrichment comparison

We examined whether the target genes in each MiRM are involved in biologically meaningful Gene Ontology (GO) or pathways via functional enrichment analysis. Specifically, GO terms in biological processes (BPs) (GO-BPs) were downloaded using `getBM` function from R package *biomaRt*, where GO terms with fewer than five genes or with evidence codes equal to Electronic Annotation, Non-traceable Author Statement or No biological Data available were discarded, giving 2007 GO-BP terms and 10 315 unique genes. The canonical pathways were downloaded from MSigDB (c2.cp.v4.0.symbols.gmt) (Subramanian *et al.*, 2005). The list of Ensembl gene IDs for each MiRM was then subjected to hypergeometric enrichment test for each GO-BP term or pathways using R built-in function `phyper`. The *P*-values were corrected for multiple testing by R function `p.adjust` to produce false discovery rates (FDR).

2.6 Data collection and pre-processing

For the first test dataset, we obtained the same ovarian cancer dataset processed by the SNMNMF authors (Zhang *et al.*, 2011). The expression data from this dataset were originally downloaded from TCGA along with the target site information from MicroCosm (v5) and GGI from TRANSFAC (Wingender *et al.*, 2000). The expression dataset contains 385 samples, each measuring 559 miRNAs and 12 456 mRNAs. For the second and third test datasets, we downloaded the expression data for BRCA and THCA from TCGA data portal. The BRCA and THCA data contain 331 and 543 samples, respectively, each measuring 710 miRNAs and 13 306 mRNAs. For both datasets, processed (level 3) RNA-seq(V2) and miRNA-seq data were used, which record the RPKM (read per kilobase of exon per million mapped reads) values for mRNAs and RPM (reads per million miRNAs mapped) for miRNAs. The data were further log2-transformed and mean-centred. Clinical information for OV, BRCA and THCA were also collected from TCGA.

For the BRCA and THCA, we used target site information from TargetScanHuman 6.2 database (Friedman *et al.*, 2009). For each mRNA–miRNA pair that have measured expression in TCGA data, we extracted the corresponding conserved target sites. For multiple transcripts of the same gene, we used transcripts with the longest 3′UTR. GGI data matrix **H** including TFBS and PPI data were processed from TRANSFAC (Wingender *et al.*, 2000) and BioGrid (Stark *et al.*, 2011) databases, respectively. Cancer-related miRNAs as oncomirs or tumour suppressors (i.e. miRNAs highly or lowly expressed in cancers) were downloaded from Spizzo *et al.* (2009) and Koturbash *et al.* (2011) Tables 1 and 2, where 97 oncomirs have measured expression in the TCGA data.

3 RESULTS

3.1 Comparison of module sizes and connectivities

Mirsynergy identified 84, 53, 50 MiRMs in OV, BRCA and THCA datasets, respectively (Table 1, Supplementary Table S1). Intriguingly, the overall synergy landscapes are distinct for each cancer type at both clustering stages (Supplementary Fig. S3). Comparing with the MiRMs from SNMNMF, however, the modules identified by Mirsynergy are clearly more densely connected within and between MiRMs (Fig. 2, Supplementary Figs S4, S5), which are comparable with the module density associated with PIMiM-MiRMs (on the same OV dataset) (Le and Bar-Joseph, 2013) and perhaps more consistent with the intricacy of the underlying biological network. Moreover, Mirsynergy identified more modules (84, 53, 50) than SNMNMF did (49, 39, 39) for OV, BRCA and THCA, respectively (Table 1). Notably, although we set $K = 50$ for SNMNMF, some modules were empty and thus discarded by the algorithm. For OV, BRCA and THCA datasets, the respective averaged number of miRNAs (genes) per module are 4.76, 5.77, 7.60 (7.57, 24.15, 32.26) for Mirsynergy and 4.12, 2.62, 2.23 (81.37, 71.56, 74.82) for SNMNMF (Table 1). For the BRCA and THCA datasets, 8/39 modules from SNMNMF have only one miRNA (i.e. 'star-shape' basic network structure) and 20/39 modules have less than three miRNAs. In contrast, all of the modules from Mirsynergy have more than one miRNA, and most modules have at least

Table 1. Performance summary of Mirsynergy, SNMNMF and PIMiM

Cancer	Method	M#	$\overline{miR}$	$\overline{mR}$	GOES	MiMEC	Time
OV	Mirsynergy	**84**	**4.76**	7.57	**15.64**	**–0.05**	**1**
	SNMNMF	49	4.12	81.37	7.51	0.07	24+
	PIMiM†	40	4.7	67.80	NA	–0.013	NA
BRCA	Mirsynergy	**53**	**5.77**	24.15	**8.74**	**–0.08**	**1.5**
	SNMNMF	39	2.62	71.56	5.56	–0.04	24+
THCA	Mirsynergy	**50**	**7.60**	32.26	**8.04**	**–0.08**	**2**
	SNMNMF	39	2.23	74.82	6.73	–0.04	24+

Note: M: module number; $\overline{miR}$ and $\overline{mR}$: average miRNA and mRNA per module; GOES: GO enrichment score ($\frac{M}{G}\sum_{g}^{G} -\log FDR$); MiMEC: miRNA–mRNA expression correlation; Time: number of hours took to run. Notably, LASSO alone took about an hour, which is included in the above run time of Mirsynergy. †The results for PIMiM were directly taken from the original paper (Le and Bar-Joseph, 2013), where GOES and Time are not available (NA). The comparison here is fair because all of three methods were applied to the same OV dataset used in Zhang *et al.* (2011).

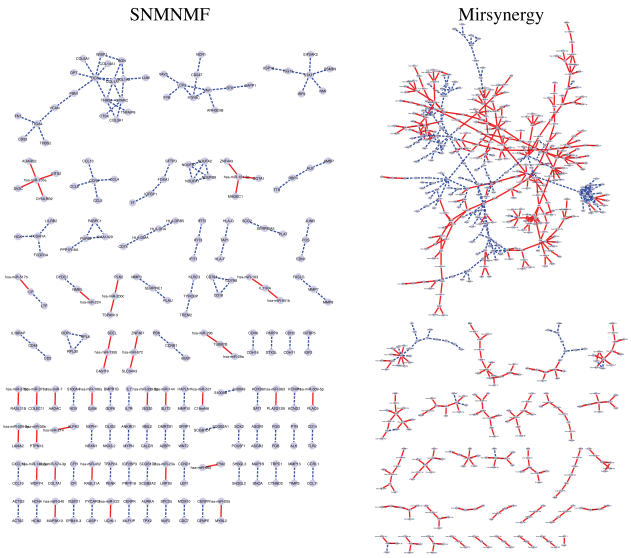

Fig. 2. Network overview using ovarian dataset. The networks from SNMNMF (left) and Mirsynergy (right) were rendered by Cytoscape 3 (Shannon *et al.*, 2003). Edges represent either miRNA–mRNA (solid red line) or gene–gene interactions (dash blue line)

three miRNAs (Supplementary Table S1). Moreover, the maximum number of miRNAs in a single Mirsynergy-MiRM are 12, 20 and 19 for OV, BRCA and THCA, respectively. In contrast, PIMiM obtained at maximum eight miRNAs per module for the OV dataset, and the highest miRNA counts for the SNMNMF modules are 9, 5 and 7 for OV, BRCA and THCA, respectively. Notably, there are considerably fewer genes in the Mirsynergy-MiRMs than in modules identified by SNMNMF and PIMiM. This perhaps implies that each MiRM identified by Mirsynergy is responsible for a more specific biological process or pathway, which is supported by the functional enrichment analysis described below. Thus, Mirsynergy predictions are in better agreement with the notion of miRNA synergistic regulation, where a relatively large number of miRNAs can engage in regulating a specialized cohort of interconnected genes.

3.2 Evaluating MiRMs by functional enrichments

As an internal control, we first compared whether the nodes included in each MiRM (i.e. internal nodes) tend to be more functionally enriched than the boundary nodes, which have non-zero connections to the internal nodes but were either disregarded or removed from the modules during the neighbourhood expansion. As illustrated in Figure 3 (Supplementary Fig. S6), the number of significant GO-BP terms at FDR < 0.05 (upper panel) and the total enrichment scores across all terms (lower panel) are significantly higher ($P < 1e\text{-}3$; one-sided Wilcoxon signed rank test) for the internal nodes comparing with boundary nodes. Interestingly, we also observed modestly positive Pearson correlations ($\rho_{OV} = 0.14$; $\rho_{BRCA} = 0.28$; $\rho_{THCA} = 0.39$) between the synergy and enrichment scores (Supplementary Fig. S7), further supporting the biological function of the identified modules.

To compare functional enrichments of the predicted modules from Mirsynergy and SNMNMF, we counted as a function of enrichment score the number of significant modules having at least one significant canonical pathways (CPs) or GO-BPs and the total number of significant CPs or GO-BPs across all modules. As depicted in Figure 4 (Supplementary Fig. S8), comparing with SNMNMF, Mirsynergy produced significantly higher number of functionally meaningful MiRMs (upper panel) ($P < 0.00781$, 0.00707, respectively; one-sided Wilcoxon signed rank test), which together contribute significantly higher number of distinct CPs or GO-BPs ($P < 0.00781$, 0.00195, respectively). Moreover, Mirsynergy also compares favourably with PIMiM, which used the same ovarian dataset, target site information and GGI data based on the published results from Le and Bar-Joseph (2013) (Fig. 4A). Notably, the averaged enrichment scores over all of the Mirsynergy-MiRMs are also higher than the scores from SNMNMF in all three test datasets (Table 1).

Examining enriched GO terms exclusive to Mirsynergy-MiRMs (Supplementary Table S2) revealed several interesting cancer-related processes, e.g. 'DNA damage response, signal transduction by p53 class mediator' (GO:0030330; FDR < 0.016; OV-MiRM-4), which involves oncogene BCL3 (Forbes *et al.*, 2011) and OV-related oncomir miR-20a (Koturbash *et al.*, 2011); 'ovarian follicle development'

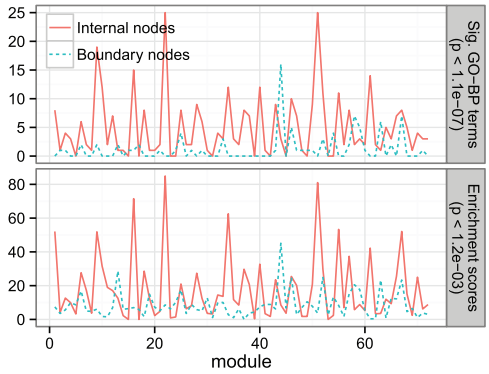

Fig. 3. Comparison of internal and boundary nodes of each MiRM in ovarian cancer. For each MiRM, we calculated the number of significant GO-BP terms (FDR < 0.05; upper panel) and GO enrichment scores (lower panel) using the internal nodes (solid) (i.e. nodes included in the MiRMs) and boundary nodes (dash) (i.e. nodes connected to internal nodes but disregarded or excluded from the MiRM to maximize synergy score). *P*-values from one-sided Wilcoxon signed rank test of internal nodes versus boundary nodes across all MiRMs are indicated in the right margins

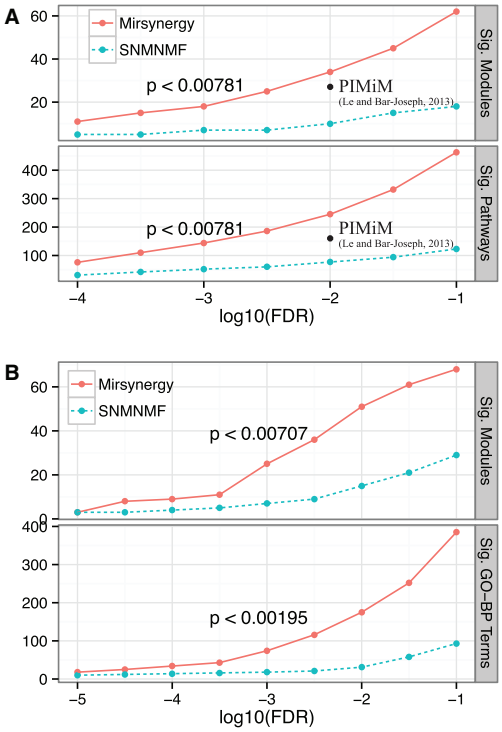

Fig. 4. Functional enrichment comparison in ovarian cancer. **(A)** The number of significant modules (upper panel) and CPs (lower panel) were plotted as a function of log10(FDR), where a significant module was defined as a module containing at least one significant CPs. We compared the distributions from Mirsynergy (solid) and SNMNMF (dash) by one-sided Wilcoxon signed rank test to obtain the *P*-values displayed above. Also, we plotted the performance of PIMiM on the same dataset based on the published results (Le and Bar-Joseph, 2013). **(B)** Same as **A** but for GO-BP terms enrichment comparison

(GO:0001541; FDR < 0.022; OV-MiRM-58), involving an OV-related oncomir miR-221; and 'negative regulation of epithelial to mesenchymal transition' (GO:0010719; FDR < 0.037; BRCA-MiRM-1; also see Supplementary Fig. S9), which is known to be associated with breast cancer (Davis *et al.*, 2013). It is also interesting to note that the miR-30 family members hsa-miR-30a (chr6: 72113254-324) and miR-30e (chr1: 41220027-118) (Kozomara and Griffiths-Jones, 2014) were both identified by Mirsynergy to be involved in BRCA-MiRM-1. Moreover, miR-30 has been predicted to regulate a number of breast cancer-associated genes in patients, who are non-carriers of BRCA1/2 mutations, lending the possibility that familial breast cancer may be caused by variation in these miRNAs (Kozomara and Griffiths-Jones, 2014). Additionally, module 11 and 23 from THCA data are respectively enriched for 'thymus development' (GO:0048538; FDR < 0.012; Supplementary Figs S10, S11) and 'positive regulation of natural killer cell mediated cytotoxicity directed against tumour cell target' (GO:0002860; FDR < 0.00012), both of which are exclusive to Mirsynergy-MiRMs. Moreover, module 11 and 23 involve THCA-related oncomirs miR-17 and 197, respectively (Supplementary Table S2) (Koturbash *et al.*, 2011).

3.3 Co-expression within MiRMs

We next examined the MiMEC involved in each predicted MiRM from Mirsynergy comparing with the boundary nodes (i.e. neighbour nodes of each MiRM) as the internal control and MiRMs identified by SNMNMF. Because miRNAs are known to repress mRNAs, the negativity of MiMEC is a reasonable indicator of the MiRM quality. In particular, we used OV, BRCA and THCA expression data from TCGA to compute for each module (or control) the averaged pairwise MiMEC. We then compared the cumulative density functions of the MiMEC from each method. We first compared the MiMEC between internal nodes (m/miRNAs within the modules) and boundary nodes (neighbour m/miRNAs with non-zero connection to the internal nodes) from each Mirsynergy-MiRMs. Indeed, miRNAs involved in the MiRMs exhibit significantly more negative correlation with the included mRNA targets for all three datasets (Fig. 5 left panels) [$P < 9.81e\text{-}10$, 3.35e-12, 0 for OV, BRCA, THCA, respectively; one-sided Kolmogorov–Smirnov (KS) test]. More impressively, Mirsynergy-MiRMs exhibit significantly higher MiMEC than SNMNMF-MiRMs ($P < 5.35e\text{-}12$, 0.0041, 0.013 for OV, BRCA and THCA, respectively) (Fig. 5 right panels). As summarized in Table 1, the MiMEC averaged over all of the Mirsynergy modules (–0.08, –0.08) are also twice as more negative as those of SNMNMF (–0.04, –0.04) for BRCA and THCA, and almost four folds more negative than the published result from PIMiM (Le and Bar-Joseph, 2013) on the same OV dataset.

3.4 KM survival analyses of MiRMs

Based on the clinical data from TCGA, we characterized the level of associations of patient survival time from the date of diagnosis to death with the discovered MiRMs by the averaged miRNA expression levels. After filtering out samples with unavailable survival time or not recorded in the expression data, we retained 376/385, 331/331, and 259/543 samples for OV, BRCA

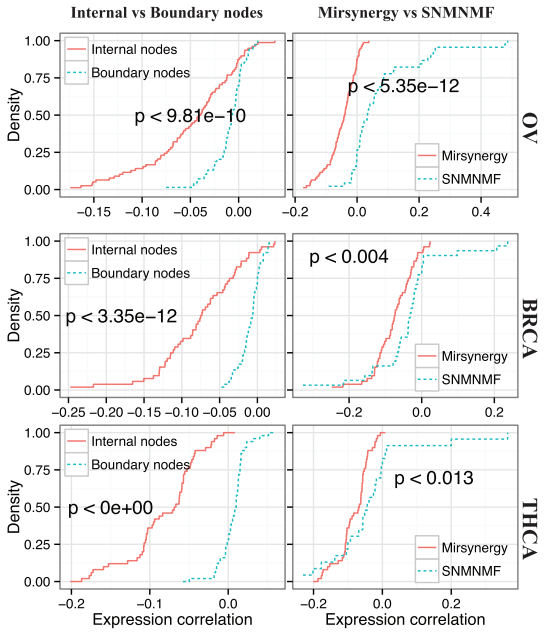

Fig. 5. Cumulative density function of expression correlation between miRNA and mRNA. Averaged Pearson correlation coefficients for each MiRM were plotted as a CDF for internal nodes/Mirsynergy (solid) and boundary nodes/SNMNMF outputs (dash). We observe a significantly more negative MiMEC distribution for the internal nodes or Mirsynergy-MiRMs, as their curves are clearly located on the left of the curves from the counterparts. *P*-values were computed by one-sided KS test

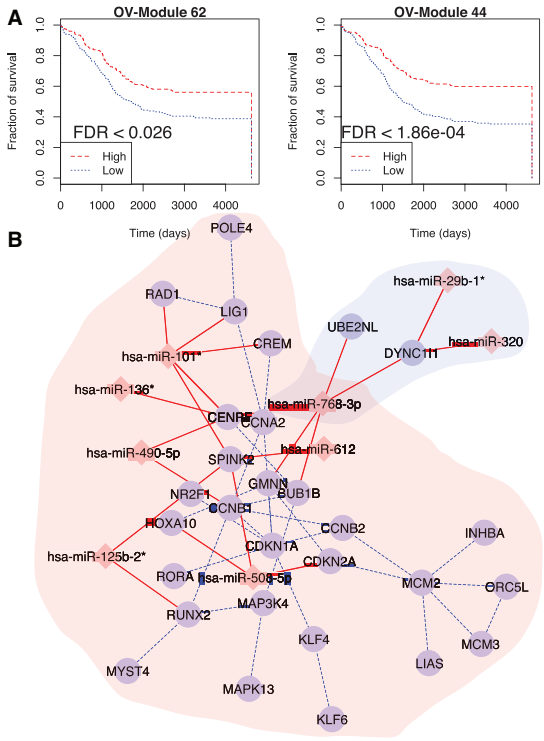

Fig. 6. KM survival analysis using MiRMs in ovarian cancer. **A.** Patients with miRNA expression higher and lower than the sample average within each MiRM were divided into 'High' (red dash) and 'Low' (blue dot) groups, respectively. Survival fractions as a function of time (days) between initial diagnosis and death were then plotted for the two groups, and the significant separation of the two curves were assessed by log-rank test. Module 62 and 44 with significant FDRs were displayed. **B.** The corresponding network for Module 62 (red) and 44 (blue). miRNAs are red diamonds and mRNAs are blue ellipses. Solid red and dash blue edges indicate the miRNA–mRNA and gene–gene interactions, respectively

and THCA, respectively (Supplementary Table S3). For each module, we divided the samples into two groups based on the lower and higher MiRM-averaged miRNA expression levels than the sample means. We then used KM method using R package *survival* to compare the survival characteristics of the two groups, with significance determined by the log-rank test. The resulting *P*-values were then corrected for multiple testing using the Benjamini–Hochberg method over all of the MiRMs (Benjamini and Hochberg, 1995). As a result, we found four, one and three MiRMs with FDR < 0.1 for OV, BRCA and THCA datasets, respectively (Supplementary Fig. S12). In contrast, when testing using individual miRNA expression profiles alone, none of the miRNAs survived the multiple testing correction, highlighting the statistical power gained by our network approach.

For BRCA, oncomir miR-100 is involved in the potential prognostic module 31. For THCA, we found oncomirs miR-29a (MiRM-8), miR-215 (MiRM-40) and miR-23a/191 (MiRM-45). Figure 6 illustrates the KM survival curves for modules 62 and 44 in ovarian cancer and the visualization of the two clusters generated from Cytoscape (Shannon *et al.*, 2003). Notably, OV-MiRM-62 involves three oncogenes, namely BUB1B, CDKN2A, MYST4 (Forbes *et al.*, 2011), and an oncomir miR-125a-3p (Koturbash *et al.*, 2011). Moreover, the two modules share a common miRNA miR-768-3p. Intriguingly, Jiang *et al.* (2013) showed that downregulation miR-768-3p is associated with MEK/ERK-mediated enhancement of protein synthesis in melanoma cells. Here we provided to our knowledge the first evidence of the potential prognostic value of miR-768-3p in the ovarian cancer network.

4 DISCUSSION

Recent works on miRNA dysregulation prove useful in cancer research (Koturbash *et al.*, 2011). To identify condition-specific networks, however, most existing methods exploited only the curated pathways or GO terms. Methods for *de novo* network reconstruction in some recent literatures operate only on interactions involving differentially expressed genes (DEG) or differentially expressed miRNA (Peng *et al.*, 2009) because the corresponding solutions quickly become intractable with increasing number of m/miRNAs. However, the DEG/M-driven approaches cannot capture subnetworks with accumulative changes of m/miRNA that have been filtered out either by individual hypothesis testings or by multiple testing corrections.

In this article, we introduce Mirsynergy as a reasonable alternative approach. Notably, standard clustering methods such as *k*-means or hierarchical clustering are not suitable for constructing MiRMs, as these methods assign each data point to a unique

cluster. A recently developed greedy-based clustering method ClusterONE is more realistic because it allows overlap between clusters (Nepusz *et al.*, 2012). However, ClusterONE was developed with physical PPI in mind. Mirsynergy extends from ClusterONE to detecting MiRMs. The novelty of our approach resides in a two-stage clustering strategy with each stage maximizing a synergy score as a function of either the miRNA-miRNA synergistic co-regulation or miRNA–mRNA/gene–gene interactions. Several methods have incorporated GGI as PPI/TFBS in predicting MiRMs (Le and Bar-Joseph, 2013; Zhang *et al.*, 2011), which proved to be a more accurate approach than using miRNA-mRNA alone. Comparing with recent methods such as SNMNMF (Zhang *et al.*, 2011) and PIMiM (Le and Bar-Joseph, 2013), however, an advantage of our deterministic formalism is the automatic determination of module number (given the predefined thresholds to merge and filter low quality clusters) and efficient computation with theoretical bound reduced from $O(K(T+N+M)^2)$ per iteration to only $O(M(N+M))$ for N (M) mRNAs (miRNAs) across T samples. Because N is usually much larger than M and T, our algorithm runs orders of faster. Based on our tests on a linux server, Mirsynergy took about 2 h including the run time for LASSO to compute OV ($N = 12\,456$; $M = 559$; $T = 385$), BRCA or THCA ($N = 13\,306$; $M = 710$; $T = 331$ or 543, respectively), whereas SNMNMF took more than a day for each dataset. More importantly, Mirsynergy-MiRMs are significantly more functionally enriched, coherently expressed and thus more likely to be biologically functional.

The success of our model is likely attributable to its ability to explicitly leverage two types of information at each clustering stage: (i) the miRNA–miRNA synergism based on the correlation of the inferred miRNA target score profiles from MMIW matrix; (ii) the combinatorial miRNA regulatory effects on existing genetic network, implicated in the combined MMIW and GGIW matrices. We also explored other model formulations such as clustering m/miRNAs in a single clustering stage or using different MMIW matrices other than the one produced from LASSO, which tends to produce MiRMs each containing only one or a few miRNAs or several large low-quality MiRMs, which were then filtered out by the density threshold in either clustering stage. Notably, an MiRM containing only a single miRNA can be directly derived from the MMIW without any clustering approach. Moreover, Mirsynergy considers only neighbour nodes with non-zero edges. Thus, our model works the best on a sparse MMIW matrix such as the outputs from LASSO, which is the best performing expression-based methods based on our comparison with other alternatives. Nonetheless, the performance of Mirsynergy is sensitive to the quality of MMIW and GGIW. In this regard, other MMIW or GGIW matrices (generated from improved methods) can be easily incorporated into Mirsynergy as the function parameters by the users of the Bioconductor package (please refer to the package vignette for more details). In conclusion, with large amount of m/miRNA expression data becoming available, we believe that Mirsynergy will serve as a powerful tool for analysing condition-specific miRNA regulatory networks.

Funding: Y.L. is funded by Natural Sciences and Engineering Research Council (NSERC) Canada Graduate Scholarship, and Z.Z. is supported by Ontario Research Fund - Global Leader (Round 2) and an NSERC grant. J.L. is supported by the National Natural Science Foundation of China (Grant NO. 61240046) and Hunan Provincial Natural Science Foundation of China (Grant NO.13JJ2017).

Conflicts of Interest: none declared.

REFERENCES

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

Barthélemy,J.P. and Brucker,F. (2001) Np-hard approximation problems in overlapping clustering. *J. Classif.*, **18**, 159–183.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Boross,G. *et al.* (2009) Human microRNAs co-silence in well-separated groups and have different predicted essentialities. *Bioinformatics (Oxford, England)*, **25**, 1063–1069.

Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Davis,F.M. *et al.* (2013) Induction of epithelial-mesenchymal transition (EMT) in breast cancer cells is calcium signal dependent. *Oncogene*, **33**, 2307–2316.

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

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

Griffiths,T. and Ghahramani,Z. (2005) Infinite latent feature models and the Indian buffet process. In: *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, Cambridge, Massachusetts, United States, pp. 475–482.

Helwak,A. *et al.* (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.

Hsu,S.D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D619.

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

Jiang,C.C. *et al.* (2013) Repression of microRNA-768-3p by MEK/ERK signalling contributes to enhanced mRNA translation in human melanoma. *Oncogene*, **33**, 2577–2588.

Joung,J.G. *et al.* (2007) Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics (Oxford, England)*, **23**, 1141–1147.

Koturbash,I. *et al.* (2011) Small molecules with big effects: the role of the microRNAome in cancer and carcinogenesis. *Mutat. Res.*, **722**, 94–105.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.

Krek,A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Le,H.S. and Bar-Joseph,Z. (2011) Inferring interaction networks using the ibp applied to microrna target prediction. In: *Advances in Neural Information Processing Systems, to appear*.

Le,H.S. and Bar-Joseph,Z. (2013) Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation. *Bioinformatics (Oxford, England)*, **29**, i89–i97.

Lu,Y. *et al.* (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics (Oxford, England)*, **27**, 2406–2413.

Nepusz,T. *et al.* (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**, 471–472.

Papadopoulos,G.L. *et al.* (2009) DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinformatics (Oxford, England)*, **25**, 1991–1993.

Peng,X. *et al.* (2009) Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, **10**, 373.

Shalgi,R. *et al.* (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput. Biol.*, **3**, e131.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Spizzo,R. *et al*. (2009) SnapShot: MicroRNAs in cancer. *Cell*, **137**, 586–586.e1.
Stark,C. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*., **39**, D698–D704.
Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc*. *Natl Acad. Sci. USA*, **102**, 15545–15550.
Tsang,J.S. *et al*. (2010) Genome-wide dissection of MicroRNA functionsand cotargeting networks using gene set signatures. *Mol. Cell*, **38**, 140–153.
Wingender,E. *et al*. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
Xu,J. *et al*. (2011) MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res*., **39**, 825–836.
Zhang,S. *et al*. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics (Oxford, England)*, **27**, i401–i409.