33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Analyzing mRNA and microRNA co-expression profiles to identify pathways and their potential regulators in ER+ and ER- breast tumors

Lei Huang*, Damian Roqueiro*, *Member, IEEE,* Yang Dai*, Member, IEEE*

Abstract— **Transcription factors and microRNAs are both considered pivotal regulators of gene expression. Numerous computational methods have been developed to predict their targets. These methods, although powerful, provide a static snapshot of how genes may be regulated by transcription factors and microRNAs. We propose a method that combines these prediction data with co-expression analysis and a supervised learning algorithm to determine the main regulators in different pathways of ER+ and ER- tumors.**

I. Introduction

TRANSCRIPTION factors (TFs) and microRNAs are well-known regulators of gene expression. The former bind directly to regulatory regions of DNA whereas the latter regulate the expression of genes at a post-transcriptional level. Although they have different mechanisms of regulation, there is evidence [1] suggesting that transcription factors and microRNAs regulate target genes in a coordinated way. In order to facilitate the elucidation of these regulation mechanisms, we propose an integrative approach to analyze mRNA and microRNA expression data together with functional target predictions of TFs and microRNAs.

Traditional microarray expression profiling often leads to the identification of hundreds, even thousands, of differentially expressed genes in a study. Subsequent functional annotation analysis (e.g., gene ontology) can classify these genes into different biological functional groups. However, this does not always reflect collaborative activities among the genes in a pathway with less significant expression changes but still critical for signaling transduction and transcriptional regulation. To help with this, differential co-expression analysis can be used to identify genes whose expressions are highly correlated in the phenotype of interest. As a variant of this approach, gene set co-expression analysis takes advantage of predefined gene sets with known biological functions and identifies subtle co-expression differences between different phenotypes [2].

Manuscript received April 15, 2011. Revised June 20, 2011. This work was supported in part by the Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust.

L. Huang, D. Roqueiro and Y. Dai are with the Department of Bioengineering, University of Illinois at Chicago (MC063), Chicago, IL 60607, USA (phone: 312-413-1819; fax: 312-413-2018; e-mails: {lhuang7, droque1, yangdai}@uic.edu)

* Both authors contributed equally

In regards to prediction, computational methods that predict the gene targets for different TFs and microRNAs are a key to postulating TF-gene and microRNA-mRNA regulatory modules. There are numerous prediction algorithms based solely on sequence analysis, for TFs [3][4], and on sequence and structure data for microRNAs [5]. Successful attempts to create a framework that integrates both types of predictions have been implemented [6]. However, none of these predictive or integrative approaches take into consideration expression data. They simply provide a static picture of how the TFs and microRNAs might regulate their target genes. Our method combines prediction information with mRNA and microRNA expression data. By applying machine learning methods to the target genes in differentially co-expressed genes sets (pathways and targets of TFs) and expression data, we attempt to provide a ranking list of TFs and microRNAs for a specific pathway.

We describe our approach using the mRNA and microRNA expression data generated from a breast tumor study [7]. Estrogen receptor (ER) plays a pivotal role in breast cancer development and progression. As a ligand-dependent TF, ER exerts both genomic and non-genomic effects that are involved in breast cancer cell differentiation, proliferation, survival, invasion and angiogenesis by interacting with other TFs and signaling pathway genes. Understanding the molecular mechanisms of ER action in tumors with different ER status (ER+ and ER–) will provide insight into the potential novel targets for breast cancer treatments [8]. Here we demonstrate the usefulness of our integrative method by analyzing microarray data sets from ER+/ER– breast tumors and uncovering the relationships among TFs, microRNAs and pathway genes that are associated with these tumors.

II. Methods

Our integrative approach started by obtaining the differentially expressed mRNAs and microRNAs from ER+ and ER– tumor microarray gene profiles. We then identified the differentially co-expressed (DC) pathway gene sets and TF target gene sets. MicroRNA target genes were predicted using an online tool. Gene set overlap test was performed to associate the TF and microRNA target sets to the DC pathway genes. Finally for a DC pathway with multiple associated TFs and microRNAs, we used a random forest classifier to determine the importance of these putative

regulators. Figure 1 shows a flowchart of the entire analysis process.

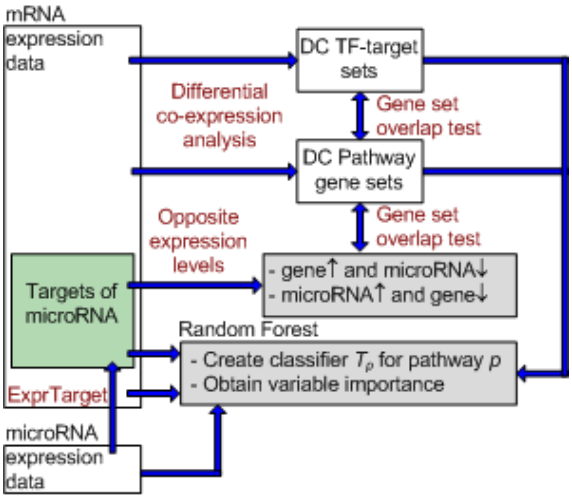

Fig. 1. Data analysis flowchart. The two gray modules provide the final prediction of the microRNAs and TFs that regulate a pathway.

A. Expression data for microRNA and mRNA

The recently published concurrent mRNA and microRNA expression profiles on ER+/ER– breast tumors were used in this work [7]. The profiling was performed using duplicate hybridizations for 99 tumor samples on Agilent "Human microRNA Microarray Kit (V2)" and human genome 4×44K one-color oligo array. Among these tumors, 60 are ER+ and 35 are ER–. We downloaded the raw microRNA data from the Gene Expression Omnibus (GEO) with accession number GSE19536 and the normalized mRNA expression data with accession number GSE19783.

B. Differential expression analysis of microRNA and mRNA

Expression analysis was carried out using packages in Bioconductor. The raw microRNA data were normalized with the Robust Multichip Averaging (RMA) method using the AgiMicroRna package [9][10]. Those microRNAs that were detected in less than 10% of the samples were filtered out. This preprocessing step yielded 498 microRNAs that were subsequently analyzed for differential expression between the ER+ and ER– tumors. We used the limma package with the Benjamini–Hochberg correction for multiple tests. The adjusted p-value threshold was set to 0.05.

Since the mRNA data were already normalized, our first preprocessing step was to identify the least variable mRNA probes. Probes with a coefficient of variability of less than 50% were filtered out. This left a total of 8,589 probes for further analysis, all of them with a unique Entrez Gene ID. The differential expression analysis was performed using a similar procedure as the one described for microRNA data.

C. Identification of differentially co-expressed gene sets

Gene set co-expression analysis (GSCA) [2] is a statistical method to identify DC gene sets. It can identify whether genes in a given pathway have distinct co-expression patterns between two different biological conditions. The R package GSCA [2] was used for the analysis of the following two main corpora of data:

1. Pathways: We considered two types: a) *canonical pathways*: obtained from the Molecular Signatures Database (MSigDB) [11]. It consists of gene sets for 880 pathways collected from Reactome, KEGG and BioCarta; and b) *signaling transduction pathways*: obtained all genes that are part of the 25 pathways listed by NETPATH [12]. Each pathway gene set has experimentally confirmed genes known to play an active role in the pathway
2. Transcription factor target sets: we downloaded information about 615 TFs from MSigDB [11]. These TFs are defined in TRANSFAC ver.7.4 [13]. For each TF, we obtained a TF target gene set that lists all the genes for which the TF is predicted to bind in their promoter region (defined as +2Kb, –2Kb from the transcription start site)

Following the procedure of GSCA, we computed Spearman's correlation coefficients for all gene pairs within the gene sets described above and a dispersion index to quantify the extent of differential co-expression of each individual set (See [2] for details). Significant DC gene sets were identified through sample permutation between the ER+ and ER– tumors. The permutation was performed 10,000 times to produce gene set specific p-values. The gene sets with a p-value < 0.05 were considered to be differentially co-expressed.

D. microRNA target prediction

ExprTarget [14] is an online human microRNA target prediction database which integrates several microRNA target prediction algorithms (such as PicTar [5], and others). ExprTarget has shown to greatly improve microRNA target prediction, compared to individual prediction algorithms, in terms of sensitivity and specificity based on the evaluation of a gold standard dataset formed from the experimentally supported targets in TarBase [15]. We downloaded all microRNA-gene predictions and filtered out those with a score less than 3. We further eliminated from our analysis the microRNA-gene pairs when the microRNA was not differentially expressed based on our differential expression analysis. We finally obtained a total of unique 3,959 predicted target genes.

E. Gene set overlap test

In order to sort out potential TFs and microRNAs that may regulate genes on each DC pathway, we selected DC TF target sets and differentially expressed microRNAs target sets that significantly overlapped with the genes in the DC pathways. For each DC pathway, the significance of overlap of genes in the pathway between a DC TF target gene set and between a microRNA target gene set, was tested using the hypergeometric distribution. Before the test, the gene sets with the same TF but different TRANSFAC matrix IDs were combined into a unique transcription factor name. This yielded a smaller number of DC TF target sets. In addition,

before testing the overlap between pathway genes and a microRNA target gene set, we first reduced the number of microRNA-gene pairs by excluding the genes which did not have opposite expression profiles as the microRNAs that are predicted to regulate them (i.e., the gene is over-expressed when the microRNA is under-expressed, and vice versa)

F. Random forest and variable importance

To determine the level of confidence for microRNAs and TFs as being putative regulators of a pathway, we applied the random forest (RF) classification algorithm [16]. RF was used as a supervised classification method on each pathway to predict the genes' expression level. We used as predictor variables the information of the microRNAs' expression levels and the TFs that are predicted to target those genes. Our ultimate goal was not to find a classifier to predict the expression level of genes but to use RF to measure the importance of each predictor variable, and thus obtain a group of TFs and microRNAs that can help differentiate the expression level of the genes in the pathway. These, in turn, will be the putative regulators.

We limited our analysis to the differentially expressed genes and microRNAs. For each gene, we used the information obtained about the TFs predicted to regulate that gene [11] and the microRNAs predicted to target the gene [14], as described before. For the microRNAs, we did not impose the conditions described in section E (i.e., opposite expression levels). The union of all TFs and microRNAs that are linked to a pathway were used as variables for a supervised learning predictor:

$T_p = (y_i, \boldsymbol{x}_i)$ with $i = 1$ to n_p, where n_p is the number of differentially expressed genes in pathway p. The response vector y contains the n_p expression levels for the genes: "up" or "down" if the gene is over- or under-expressed in ER+. Figure 2 illustrates the layout of the data. Each vector $\boldsymbol{x}_i$ has $m = k + r$ predictor variables:

x_{ij} for $j = 1$ to k contains information related to the k microRNAs in the pathway. Each x_{ij} is coded as: "0" if microRNA j does not target gene i; "+1" or "–1" if microRNA j targets gene i and is over- or under-expressed in ER+ respectively.

x_{ij} for $j = k+1$ to $k+r$ contains the information of the r TFs in the pathway. The variables are coded with a "0" if TF k is not present in the promoter region of gene i; or with a "1" if present.

		microRNAs					Transcription factors				
	expression	1	2	....	k		1	2	....		r
$gene_1$	up		-1		+1			0		1	
$gene_2$	down		0		+1			1		1	
	⋮		⋮		⋮			⋮		⋮	
$gene_m$	up		-1		0			1		0	

Fig. 2. Data layout for random forest classification.

For each pathway, an ensemble of 200 trees was created. One third of the variables were randomly chosen at each tree level and one third of the samples were left as out of bag. Variable importance was determined after performing permutations on the trees to assess the change in their predicting power. Each variable was assigned a mean decrease of accuracy score and the ranking of variables for the pathway was based on this score. The analysis was implemented with the R package randomForest.

III. Experimental Results and Discussion

For the microRNA and mRNA microarray expression profiles described in section A, the number of differentially expressed microRNAs is 60, of which 24 are over-expressed and 36 are under-expressed in ER+ tumors. The number of differentially expressed genes is 2,770 which includes 1,333 over-expressed and 1,437 under-expressed in ER+ tumors.

We applied the GSCA approach to the mRNA expression profiles. For canonical pathways (880) and signaling transduction pathway gene sets (25), we identified 253 pathway gene sets as DC between ER+ and ER– tumors (p-value < 0.05). For TF target gene sets (615), 404 of them were identified as DC (p-value < 0.05). We further took 97 DC pathways with smaller p-values and 115 DC TF target gene sets to test for gene set overlap. 66 pathways were identified having statistically significant gene overlap with various TF target sets at a p-value of 0.05. A selected list of pathways can be found in Table I.

TABLE I
List of pathways with significantly overlapped genes over TF target sets and their biological functions

Pathway name	Biological function
KEGG_CELL_CYCLE	Cell cycle
KEGG_ECM_RECEPTOR_INTERACTION	Cell adhesion, proliferation
KEGG_ERBB_SIGNALING_PATHWAY	Signal transduction
KEGG_MAPK_SIGNALING_PATHWAY	Signal transduction
KEGG_MOTOR_SIGNALING_PATHWAY	Signal transduction
NETPATH_Alpha6_Beta4_Integrin	Cell invasion, differentiation
NETPATH_AR	Signal transduction
NETPATH_EGFR1	Signal transduction
NETPATH_leptin	Signal transduction
NETPATH_TGF-beta	Signal transduction

The gene set overlap test procedure output 2 pathways (MAPK signaling pathway and NETPATH AR) whereas the RF analysis provided a ranking of TFs and microRNAs for each pathway. Focusing on the MAPK signaling pathway, the overlap test yielded 8 under-expressed microRNAs (hsa-miR-19a, hsa-miR-135b, hsa-miR-203, hsa-miR-223, hsa-miR-23a, hsa-miR-24) and 7 over-expressed target genes (CACNA1D, DUSP5, DUSP16, IKBKB, MAP3K12, RASA1, RPS6KA5) in ER+ tumors (Figure 3). The figure was constructed, based on the known pathway structure, by taking the above mentioned genes and microRNAs and the 8 top-ranked TFs (with their gene targets) reported by RF (Figure 4). In Figure 3, the circles denote the TFs; the genes (rectangles) and microRNAs (ovals) are grey if under-expressed and white if over-expressed in ER+ tumors. Dotted arrows between genes indicate precedence through intermediate genes in the pathway.

Relating our analysis to known facts about this pathway, we note that the epidermal growth factor receptor (EGFR) is significantly over-expressed in ER– tumors. Activation of

the EGFR/ERBB2 pathway initiates a kinase signaling cascade that has a variety of effects on tumor cells, such as stimulation of cell proliferation, enhanced invasion and cell motility as well as inhibition of apoptosis. The hyperactivity of GFR signaling has been associated to the dynamic nature of ER status and resistance to endocrine therapy [17]. EGFR is targeted by the TF CEBPA, highly ranked in our RF analysis for this pathway and known to cause growth arrest by inhibiting different kinases. The rest of the TFs have been reported to regulate genes involved in the MAPK signaling pathway [18].

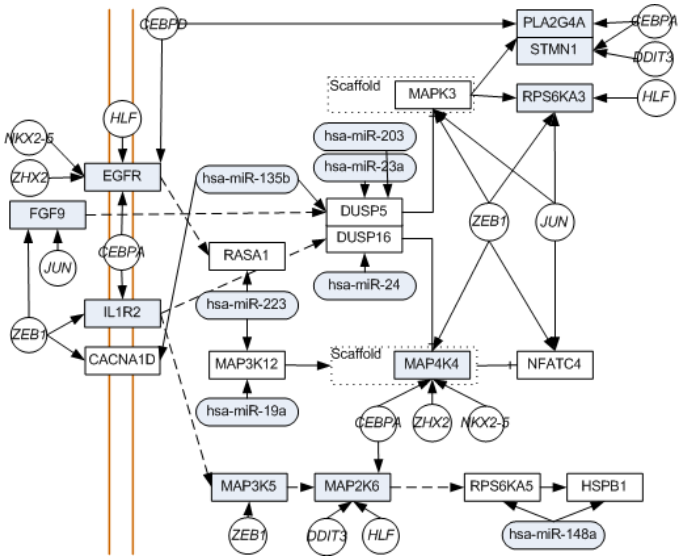

Fig. 3. Identified microRNAs and TFs as potential regulators in the MAPK signaling pathway.

Traditional gene functional enrichment analysis may miss some biological relationships between differentially expressed genes and their common regulators such as TFs and microRNAs. The approach adopted in our work may help reveal those underlying relationships in the specific biological context. Our study demonstrates that this integrative approach can uncover important biological pathways and identify important players involved in the regulation of those pathways such as key pathway genes as well as TFs and microRNAs.

IV. Conclusion

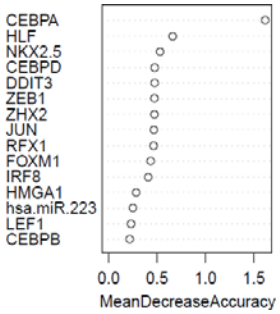

Fig. 4. Ranking obtained from RF for the MAPK signaling pathway.

We proposed an integrative bioinformatics approach to combine microRNA and mRNA expression profiling to discover important biological pathways involved in ER+ and ER– breast tumors. Our approach based on statistically significant gene set overlap gives us a coherent set of genes, framed in the context of a pathway, and their possible regulators. Additionally, by applying the RF algorithm we were able to enhance our initial list of possible regulators to include interactions between TFs and microRNAs.

References

[1] S. Bandyopadhyay and M. Bhattacharyya, "Analyzing miRNA co-expression networks to explore TF-miRNA regulation," *BMC Bioinformatics*, vol. 10, 2009, p. 163.

[2] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinformatics*, vol. 25, Nov. 2009, pp. 2780 -2786.

[3] T.L. Bailey, N. Williams, C. Misleh, and W.W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Research*, vol. 34, Jul. 2006, p. W369 -W373.

[4] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner, "Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data," *Nucleic Acids Research*, vol. 23, Jan. 1995, pp. 4878 -4884.

[5] A. Krek, D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, et al., "Combinatorial microRNA target predictions," *Nat Genet*, vol. 37, May. 2005, pp. 495-500.

[6] A. Le Bechec, E. Portales-Casamar, G. Vetter, M. Moes, P.-J. Zindy, A. Saumet, et al., "MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model," *BMC Bioinformatics*, vol. 12, 2011, p. 67.

[7] E. Enerly, I. Steinfeld, K. Kleivi, S.-K. Leivonen, M.R. Aure, H.G. Russnes, et al., "miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors," *PLoS ONE*, vol. 6, Feb. 2011, p. e16915.

[8] C.K. Osborne and R. Schiff, "Mechanisms of Endocrine Resistance in Breast Cancer," *Annu. Rev. Med.*, vol. 62, Apr. 2011, pp. 233-247.

[9] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, Jan. 2003, pp. 185 -193.

[10] P. Lopez-Romero, "Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library," *BMC Genomics*, vol. 12, 2011, p. 64.

[11] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, Oct. 2005, pp. 15545 -15550.

[12] K. Kandasamy, S.S. Mohan, R. Raju, S. Keerthikumar, G. Kumar, A. Venugopal, et al., "NetPath: a public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, 2010, p. R3.

[13] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, Jan. 2003, pp. 374 -378.

[14] E.R. Gamazon, H.-K. Im, S. Duan, Y.A. Lussier, N.J. Cox, M.E. Dolan, and W. Zhang, "ExprTarget: An Integrative Approach to Predicting Human MicroRNA Targets," *PLoS ONE*, vol. 5, Oct. 2010, p. e13534.

[15] P. Sethupathy, B. Corda, and A.G. Hatzigeorgiou, "TarBase: A comprehensive database of experimentally supported animal microRNA targets," *RNA*, vol. 12, Feb. 2006, pp. 192 -197.

[16] L. Breiman, "Random Forests," *Mach. Learn.*, vol 45, 2001, pp 5-32.

[17] S. Lopez-Tarruella and R. Schiff, "The Dynamics of Estrogen Receptor Status in Breast Cancer: Re-shaping the Paradigm," *Clinical Cancer Research*, vol. 13, Dec. 2007, pp. 6921 -6925.

[18] S. Maijgren, I. Sur, M. Nilsson, and R. Toftgård, "Involvement of RFX proteins in transcriptional activation from a Ras-responsive enhancer element," *Archives of Dermatological Research*, vol. 295, Apr. 2004, pp. 482-489.