Published in final edited form as:

Chin Sci Bull. 2010 November ; 55(31): 3576–3589. doi:10.1007/s11434-010-4051-1.

Identification of common microRNA-mRNA regulatory biomodules in human epithelial cancers

Xinan Yang^{1,2}, Younghee Lee², Hong Fan³, Xiao Sun^{1,*}, and Yves A Lussier^{2,4,5,*}

¹State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China

²Center for Biomedical Informatics and Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL 60637 USA

³MOE Key Laboratory of Developmental Genes & Human Diseases, Southeast University, Nanjing 210009, China

⁴The University of Chicago Cancer Research Center, and The Ludwig Center for Metastasis Research, The University of Chicago, Chicago, IL 60637 USA

⁵The Institute for Genomics and Systems Biology, and the Computational Institute, Argonne National Laboratories and The University of Chicago, Chicago, IL 60637 USA

Abstract

The complex regulatory network between microRNAs and gene expression remains unclear domain of active research. We proposed to address in part this complex regulation with a novel approach for the genome-wide identification of biomodules derived from paired microRNA and mRNA profiles, which could reveal correlations associated with a complex network of de-regulation in human cancer. Two published expression datasets for 68 samples with 11 distinct types of epithelial cancers and 21 samples of normal tissues were used, containing microRNA expression (Lu et al. *Nature Letters* 2005) and gene expression (Ramaswamy et al. *PNAS* 2001) profiles, respectively. As results, the microRNA expression used jointly with mRNA expression can provide better classifiers of epithelial cancers against normal epithelial tissue than either dataset alone ($p=1\times 10^{-10}$, F-Test). We identified a combination of six microRNA-mRNA biomodules that optimally classified epithelial cancers from normal epithelial tissue (total accuracy = 93.3%; 95% confidence intervals: 86% - 97%), using penalized logistic regression (PLR) algorithm and three-fold cross-validation. Three of these biomodules are individually sufficient to cluster epithelial cancers from normal tissue using mutual information distance. The biomodules contain 10 distinct microRNAs and 98 distinct genes, including well known tumor markers such as *miR-15a*, *miR-30e*, *IRAK1*, *TGFBR2*, *DUSP16*, *CDC25B* and *PDCD2*. In addition, there is a significant enrichment (Fisher's exact test $p=3\times 10^{-10}$) between putative microRNA-target gene pairs reported in five microRNA target databases and the inversely correlated micro-RNA-mRNA pairs in the biomodules. Further, microRNAs and genes in the biomodules were found in abstracts mentioning epithelial cancers (Fisher Exact Test, unadjusted $p<0.05$). Taken together, these results strongly suggest that the discovered microRNA-mRNA biomodules correspond to regulatory mechanisms common to human epithelial cancer samples. In conclusion, we developed and evaluated a novel comprehensive method to systematically identify, on a genome scale, microRNA-mRNA expression biomodules common to distinct cancers of the same tissue. These biomodules also comprise novel microRNA and genes as well as an imputed

*Corresponding authors.

Supplementary materials are available at (<http://www.lussierlab.org/publication/biomodule>).

regulatory network, which may accelerate the work of cancer biologists as large regulatory maps of cancers can be drawn efficiently for hypothesis generation.

Keywords

biomodule; microRNA expression; gene expression; cancer; diagnosis

Mounting evidence shows that common gene expression signatures across many types of cancer are useful markers for medical diagnostics[1-3]. And recent work also reveals the universal diagnostic role of microRNAs for human tumors[4] which extends the potential application of microRNA profiling from specific cells within a tissue[5-10] to a more functional understanding of cancer development[11,12]. For example, Volinia *et al.* identified a set of universal microRNA signatures for solid cancers by a large-scale analysis on patients including lung, breast, stomach, prostate, colon, and pancreatic tumors[4]. Lu *et al.* found that with lower expression levels in tumors than in normals, microRNAs are unexpectedly rich in information about cancer[13]. In addition, microRNA and messenger RNA (mRNA) interactions remain incompletely understood in most cancers. Among solutions conducted to clarify these interactions, hundreds of putative microRNA targets have generally been calculated from nucleic acid sequence similarities and are provided in datasets (miRBase[14,15], miRanda[12], PicTar[16,17], TarBase[18] and TargetScan[19], reviewed by Bartel[20]). However, sequence similarity alone, taking out of cellular context, leads to about 40% accuracy in microRNA target validations[21,22]. Further, traditional biological characterization of microRNA targets is time consuming. Since expression arrays of microRNA and mRNA can be conducted over the same samples, there is an opportunity to reverse engineer regulatory mechanisms and to provide microRNA-target hypotheses. Though previous studies also identified regulatory bio-modules of specific human cancer by combining microRNA and mRNA expression profiles[23-26], their algorithms focus exclusively on linear inverse correlations[25,26] and/or on previously predicted microRNA targets[24-27] between two expression profiles that may not reveal other complex patterns of regulations that may appear paradoxical (such as co-expression). However, co-expression has been observed between intragenic microRNAs and a target gene, when this target happens to be the microRNA host gene[28]. Further, one can hypothesize downstream signaling that may paradoxically lead to co-expression of microRNAs with genes that aren't their immediate target. Of note, certain microRNAs have already been discovered to be either positively co-expressed with their target genes[29], for example, *miR-17-5p* and its target *E2F1* are positively coregulated by the proto-oncogene *c-Myc*[30]. Therefore, in this manuscript, we propose a novel unbiased computational strategy to derive microRNA-mRNA interactions biomodules that represent fundamental regulatory mechanism common to several cancers inclusive of inverse correlation and correlation patterns between microRNAs and mRNAs.

In bioinformatics, supervised machine learning techniques have been successfully used for classification and cancer diagnosis. So far, many supervised prediction metrics have been built to find the relation between gene expression and clinical conditions. The Support Vector Machine (SVM) is among the methods that have successful applications to the problems of cancer diagnosis[31]. The model of Partial least squares (PLS) builds weighted linear combinations of genes that have maximal covariance with conditions of interest, but usually result in thousands of genes on microarray expression data. Assuming that only a few gene markers determine the classification of a genetic problem, the penalized logistic regression (PLR) is a technique that performs comparably to the SVM[32] and provides an estimation of the underlying class probabilities. PLR has been considered as a stand-alone for classification of microarray gene expression data with "small sample size, larger

variances”[32,33] and has been shown strongly predictive for clinical response in cancer as compared with the other above mentioned methods[33]. However to our knowledge, PLR has never been used over microRNA data before, nor for reverse engineering regulatory networks.

We hypothesized that we could identify in high throughput microRNA/mRNA biomodules common to different types of epithelial cancers that would also be associated to some fundamental biological mechanisms. We address the problem by re-analyzing the public data which includes 89 human epithelial samples including cancers and controls that have both mRNA and microRNA expression profiles[13]. Using PLR, we identify common microRNA-mRNA biomodules across epithelial cancers that best distinguishes them from normal epithelial tissues. Specifically, to reverse engineer the regulatory network consisting of microRNA and mRNA interactions, we assumed that microRNAs might act in concert with other regulatory processes to regulate gene’s expression[34] without using any prior knowledge of microRNA target genes or any assumption on the inverse correlation or co-regulation.

The remainder of the paper is organized as follows. In the **Result section**, we first describe (i) a preliminary validation study comparing PLR usage in mRNA expression data alone, microRNA expression data alone and the combined results, followed by (ii) identification and the quantitative description of the discovered microRNA-mRNA biomodules and their interaction networks. Subsequently, we describe qualitatively the biomodules by (iii) showing how the expression patterns of the microRNA-mRNA biomodules classify epithelial cancers and normal samples across all 11 types of tissues, and by (iv) conducting the Gene Ontology enrichment of the genes associated with these biomodules to provide an unbiased description of their biological mechanisms. Finally, we evaluate the identified genes and microRNAs in the biomodules in two ways: (v) we calculate the enrichment of gene targets of microRNAs among the negative co-expression patterns of microRNA and mRNA, (vi) we conduct a systematic review of literature to identify the relevance of the gene patterns observed in the biomodules. Specifically, we calculate the enrichment of co-occurrence of one “genes or microRNA” and one epithelial cancer in the abstracts indexed in PubMed. The **Discusses section** discusses the results, limitation and future work of the study and is followed by the **Conclusion**. The detailed **Methods section** follows with in depth details on our approach and evaluations.

1 Results

1.1 Preliminary Validation Study Comparing PLR usage in mRNA Expression Data alone, microRNA Expression Data alone and the Combined Results

In order to demonstrate the increased statistical power of classification using joint microRNA and mRNA expressions, we compared the classification accuracy for microRNA, mRNA, and the expression profiles of both kinds of microarrays on the same 89 epithelial samples, representing 11 types of human tumor (colon, pancreas, kidney, bladder, prostate, ovary, uterus, lung, mesothelioma, melanoma, and breast cancer)[13].

To estimate the total accuracy of expression profiles, we repeatedly ($n=100$) performed 3-fold[35,36] stratified cross-validation, using standard PLR as the classifier to predict cancer from normal samples (Details are given in section Methods). The meta-analysis of both mRNA and microRNA expression profiles resulted in a higher accuracy with smaller variance (Figure 1). The differences between the predictions from individual mRNA or microRNA and the predictions from integrated data are significant (F-test = 56.27, $p = 1 \times 10^{-10}$). Our results suggest that there are groups of non-coding and coding genes that work together with respect to the classification (total accuracy = 93.3%; 95% confidence

intervals: 86% - 97%) of cancer tissue versus normal tissue for the 11 types of epithelial tissues.

To identify the contribution of each specific epithelial tissue to the accuracy, we further counted the classification accuracy in each type of tissue (Table 1). The results indicate that, with the exception of colon cancer, no specific tissue type alters the total accuracy of classification of the combined expression profile, implying that there are common expression patterns among distinct kinds of epithelial cancers in the context of expression levels of both mRNA and microRNA. The lower accuracy of cancer-vs-normal diagnosis of colon sample in this dataset suggests a tissue-specific expression pattern among these colon samples because the dataset contains eleven different anatomical locations of epithelial tissues. Interestingly, for each tissue type, the analysis of the combined microRNA-mRNA expression resulted in a relatively higher accuracy than either the microRNA or the mRNA expression profiles conducted alone did.

1.2 Identification and the Quantitative Description of the Discovered Biomodules and their Interaction Networks

We previously confirmed that combining the expression profiles of mRNAs and microRNAs can lead to higher classification power. In order to discover the common group of mRNA and microRNAs interactions that best classified epithelial cancers separately from normal tissue, regardless of the specific cancer or anatomical location, we employed PLR on the whole combined microRNA and mRNA expression profiles to generate ten prioritized groups (biomodule candidates). To answer the question of how many biomodules should be optimal to best classify, we conducted a prediction-based statistical inference on the datasets[39]. Six microRNA-mRNA biomodules contributed to the lowest 95% quintile of the error-rate on the combined data (Figure 2). In particular, the error-rate of microRNA data stably decreased when $n < 3$, while the combined data showed an increased tendency of error-rate when $n > 6$. Our results confirm the six microRNA-mRNA biomodules that perform commonly in several types of human epithelial tumor. Thus our further investigation and discussion will focus on the leading six groups which are referred as “expression **microRNA-mRNA biomodule**”. The biomodules contain 10 distinct microRNAs and 98 distinct genes, (3 microRNAs and 24 genes, 3 microRNAs and 22 genes, 4 microRNAs and 21 genes, 5 microRNAs and 24 genes, 1 microRNAs and 6 genes, 1 microRNAs and 19 genes, respectively).

Two microRNAs and 11 mRNAs were involved in multiple microRNA-mRNA biomodules, such as *miR-30e*, *miR-15a*, *ABO*, *METTL7A*, *CFD*, *IRAK1*, *PTMS*, *MED6*, etc (Suppl. Table 1). A two-dimensional projection of tumor (red 1) and normal (black 0) samples into the space of the expression level of the first biomodule (x axis) and second biomodule (y axis) separates the samples of tumor and the normal tissue (Figure 3). Moreover, the combined analysis of microRNA and mRNA (leftmost projection of Figure 3) yields larger between-cluster distance than within-cluster distances.

1.3 Expression Patterns of the MicroRNA-mRNA Biomodules that Classify Epithelial Cancers and Normal Samples Across 11 Types of Tissues

The genes and/or microRNAs within a biomodule should have the biggest within-module correlation and biggest central distance between cancer and normal sample groups. The Mutual information (MI) has been used as a measurement between two genes related to their degree of independence[40]. The hypothesis here is that a higher measurement of mutual information observed between the expression of two genes and/or microRNAs indicates a closer biological relationship. Therefore, we used MI to validate the expression pattern of the combined six biomodules identified by PLR algorithm as best predictors (lowest

prediction error using cross-validation in Figure 1). Figure 4 shows that the jointed six modules can distinguish epithelial cancer from normal tissue samples using MI measurements. We also found high mutual information between the microRNAs and/or genes in the six biomodules (data not shown).

We further observed the mutual information within each of the six leading biomodules. We found that the mutual information was higher in the leading three biomodules, which included the known cancer suppressor microRNA *miR-30e* and *miR-15a* (Suppl. Figure 1). For instance, previous research proposed *miR-30e* as tumor suppressor because the region encoding *miR-30e* is lost in 79% of neuroblastoma[41] whereas *miR-15a* was found as prostate cancer suppressor[42]. Our result suggests that *miR-30e* and *miR-15a* play as universal suppressor for more than one type of epithelial cancers.

1.4 Gene Ontology enrichment of the genes associated with these biomodules

To further understand the biological mechanism underling the identified merged microRNA-mRNA biomodule, we conducted a Gene Ontology enrichment of the biological processes and also reviewed the literature for the genes involved in the merged microRNA-mRNA biomodule. Table 2 gives 7 Biological Process terms defined by Gene Ontology (Built: 15-Mar-2006), which are significantly over-represented within the 98 genes based on running hypergeometric tests for each GO term of two Affymetrix chips (hu35ksuba and hu6800). Many of these biological processes are known to be involved in oncogenesis (*e.g.* DNA replication, growth, nucleic acid metabolism, etc.). The hypergeometric tests evaluate the likelihood that the corresponding number of annotations is occurring in a random list of genes of the same size using R *CompdiagTools* package. *PTMS*[43], *MCMC7*[44,45], *TK1*, and *NFIB* are the genes important for DNA replication (hypergeometric p-value is 0.007). Higher expression levels of *PTMS*[46], *TK1*[47], and *MCM7*[45] in epithelial carcinoma cells than in normal cells has been reported. The over-expression of *IRAK1* and under-expression of *TGFBR2* in tumor cells have been previously reported as well[48,49]. The genes enriched in protein amino acid dephosphorylation are all famous tumor suppressors or oncogenes. Both *DUSP16*[50], positively, and *TGFBR2*[51,52], negatively, are involved in the *MAPK* signaling pathway. *CDC25B* phosphatase, as a kind of cyclin-dependent kinase activator, plays a key role in cell-cycle progression and the *MAPK* signaling pathway with positive expression[53]. In contrast, each biomodule individually did not have sufficient genes to achieve statistic power or to identify meaningful non-trivial processes (Suppl. Table 2).

1.5 Evaluation 1 -- Enrichment of MicroRNA targets among inversely correlated microRNA-mRNA pairs

The PLR identified six microRNA-mRNA biomodules comprising 10 microRNAs and 98 mRNAs. The average expressions of every one of the 10 microRNAs were down-regulated in the epithelial cancer group as compared with the normal tissue group, while 59 out of 98 mRNAs were up-regulated. There were another 39 mRNAs in this biomodule that were down-regulated in cancer. Five databases of microRNA target predictions based on sequences were used: *miRBase* v5[14,15], *miRanda* version on July 2007[12], *PicTar* server version 4.0.24[16,17], *TarBase* v4.0[18] and *TargetScan* v3.1[19], and they were used to ascribe relationships between genes and microRNAs within each biomodule clusters to generate the network (shown in Figure 5) comprising 35 microRNA-targets relationships. Among the biomodules, 208 distinct pairs of anti-correlated microRNA-mRNA were found, of which 29 were also found in the five putative microRNA target databases based on sequence similarities that contain 121,755 distinct pairs of microRNA-target. All together, 217 microRNAs and 16,063 genes were studied, with a potential for 3,485,671 microRNA-mRNA interactions. Therefore, the Fisher's exact test showed a significant enrichment

($p=3\times 10^{-10}$, odds ratio=4.5) of observed microRNA-mRNA inverse correlations pairs in the six biomodules, corroborating the validity of the proposed approach.

In some cases, the microRNA target was only up-regulated in some subset of epithelial cancers (tissue-specific inverse correlation with the microRNA). To address this question, we further measured the tissue-specific log-transformed group distance of expression for every microRNA and gene in the six biomodules (Suppl. Table 3). Eight putative target genes were down-regulated in the epithelial cancers group as compared with the normal tissue group and up-regulated in a specific cancer tissue as compared with the expression in the comparable normal tissue (Suppl. Table 3). And three microRNAs (*has-miR-143*, *-193* and *has-let-7b*) were down-regulated in every cancers except in colon cancer; while *has-miR-1* was down-regulated in all cancers except in pancreas cancer. Besides, there are four down-regulated putative target genes (*SPCS1*, *FABP4*, *PDK4* and *IHPK2*) of one of the 10 microRNAs in the combined biomodules. Note that our method didn't make any assumptions on the mechanism of correlation or inverse correlation between microRNAs and genes expressions in the same biomodule; therefore it identified co-expression as well inversely correlated patterns. As described in the background, correlation of microRNA and mRNA expression may be attributed to mechanisms such as genes hosting an intragenic microRNA correlated with the expression of their intragenic microRNA[54], or target genes regulated by other mechanism such as transcriptional factor[29, 34, 55], and other complex feed-back controls.

1.6 Evaluation 2 -- Enrichment of Co-occurrence of One "Genes or MicroRNA" and One Epithelial Cancer in the Abstracts Indexed in PubMed

To further evaluate the functional role of the 10 microRNAs and 98 genes in our merged epithelial cancer biomodule, we conducted a Pubmed literature search to estimate the enrichment of co-occurrence of the symbol of (i) [the microRNA or gene] with that of (ii) any of the 11 types of epithelial cancers (Fisher's exact test). A true positive result was assumed to be unadjusted p-value < 0.05. As of April 15th, 2009, there were 18,789,608 papers indexed in *Pubmed* database. The number of abstracts containing each symbol, cancer type and both were recorded. There are 78% (69 out of 89) of the gene and microRNA symbols taken from the six biomodules that have *Pubmed* that co-occurred significantly ($p<0.05$) with at least one term associating to epithelial cancers (Suppl. Table 4). This result strongly suggest that the proposed method can enrich known facts about the regulation for epithelial cancers, that it likely predicts new fact, and that it is likely scalable to other biological problems. Moreover, Among the results, 4 out of 9 identified microRNAs and 33 out of these 80 genes that have official Symbols (89 total) met the threshold of $p<0.05$ for the co-occurrences at least two epithelial cancers in PubMed, there were *miR-30e*, *miR-15a*, *let-7b* and *miR-143*, *ABO*, *MYO9B*, *CFD*, *RAD1*, *IL1RN*, *S100A2*, *HBA2*, *SH3BP2*, *BMP2*, *MT1E*, *CEBPD*, *C1R*, *ADRM1*, *NUAK1*, *DNAJB6*, *CCNDBP1*, *MMP1*, *NDE1*, *MCM7*, *GPX3*, *BFAR*, *ARMCX1*, *ZBTB7A*, *PKP1*, *ITIH1*, *GRB14*, *FPGS*, *SLC31A1*, *TK1*, *TGFBR2*, *SFRP1*, *CDC25B* and *TUBB3* (Suppl. Table 4). Of note, there were 12 gene symbols for which we did not find a single *PubMed* result, which may be novel predictions. They were *METTL7A*, *KRTCAP2*, *tcag7.1314*, *SPCS1*, *psiTPTE22*, *CCDC103*, *RASL12*, *TRIM69*, *KIAA2026*, *OSBPL10*, *C1orf142* and *hCG_1731871*.

All the identified microRNAs have already been reported associated with human cancer, except *miR-33*. The known tumor-suppressive microRNAs are *miR-30e* in breast, head and neck, and lung cancer[56], *miR-193a* and *miR-338* in oral squamous cell carcinoma[57], *miR-15a* in prostate cancer[42], *miR-130a* in the drug-resistant cell lines of ovarian cancer[58], *miR-136* in leukemia[59], *miR-143* in cervical cancer[60], colorectal cancer[61,62] and nasopharyngeal carcinoma[63], *miR-1* in lung cancer[64] and hepatocellular cancer[65], *let-7b* in the poor-prognostic ovarian carcinoma[66], colon

cancer[67], lung cancer[68] and melanoma [69]. Our result suggests that the new candidate *miR-33* may also function as growth suppressors in human epithelial cancer cells.

We found that apoptosis related microRNAs (*miR-338*, *miR-193a*) and genes (*BCL2L13*, *PDCD2*, *IRAK1*, *IHPK2*, *INHBA*, *BFAR* and *MCM7*), as well as genes involved in cell division cycle pathway (*RAD1*, *PTMS*, *INHBA* and *CDC25B*) and DNA replication (*NFIB*, *PTMS*, *MCM7* and *TK1*), are promising markers for diagnosis of human epithelial cancer. Other known tumor markers that we found include *IRAK1*, *TGFB2*, *DUSP16*, and *CDC25B*. Some tumor suppressor genes such as *PDCD2* also showed consistent down-regulation in all human tumors derived from epithelial tissues. Our most interesting findings are that many genes in the group of diagnostic markers are the expected targets of the identified microRNAs. Although computational prediction of microRNA targets requires experimental validation, this observation further reveals the complicated relationship between microRNA and genes in tumors[70]. We suggest that combining expression profiling of microRNA with mRNA is a promising strategy to predict the risk of epithelial tumors.

We also summarized and did basic statistics on the genes and microRNAs in these microRNA-mRNA biomodules for further possible investigation. All 10 microRNAs and 75% identified genes show strongly significant changes ($q\text{-value}[71] < 0.05$), when we did fold-change equivalent measures on the log-transformed expression levels[72] (Suppl. Table 1). Note that since the biological events are not independent, we cannot assess the performance by individual genes. Individually, some of the identified diagnostic signatures have mediocre identification power but if joined together, they may have good classification power.

2 Discussion

In this paper, we described an integrated analysis of expression levels of protein-coding transcripts and non-coding transcripts. We used modern classification and evaluation methods based on penalized logistic regression (PLR) and found that the combining analysis of mRNA and microRNA expression profiles resulted in the lowest rate of misclassification. Furthermore, these biomodules found by PLR are sorted according to their variances in descending order. We then identified and investigated a group of six leading microRNA-mRNA biomodules whose collective expression was strongly associated with epithelial cancers from different originals as compare with their corresponding normal tissue. The first, second and third biomodules could individually distinguish cancers from controls, suggesting the diversity of distinct biomodules involved in distinct mechanisms associated with epithelial cancers.

Tumorigenesis is a multi-step process during which cancer cells acquire characteristics such as self-sufficiency in growth, evasion of apoptosis, insensitivity to growth inhibitory signals, limitless replicative potential, and so on. Early detection of cancer is the key for the application of successful therapies. MicroRNAs have recently been identified as a new class of genes with tumor suppressor and oncogenic functions. More researchers are considering the advances of microRNA expression profiling and discussing their potential in cancer diagnostics and prevention. It is known that most microRNAs are controlled by developmental or tissue-specific signaling and that there are tissue special regulatory microRNA and gene expression patterns. However, our study shows that some microRNAs are consistently under-expressed in 11 different types of epithelial tumors compared with normal tissues, strongly suggesting that there are common microRNA regulatory patterns occurring in distinct types of epithelial cancers from distinct anatomical origins. The regulatory network of the identified microRNA-gene biomodules include the co-expressed

and inversely co-expressed patterns within a biomodule, implying the complexity of regulatory mechanisms beyond the direct microRNA target, such as a host microRNA gene co-expressed with its intragenic microRNA, more complex interactions mediated via signaling and transcriptional activity [30,34,54], or a microRNA and gene activated by the same regulator[30].

In contrast to the algorithm of decision fusion by Wang *et al.* using the same datasets, our study is superior in two points. Wang's "high level decision fusion" consists of analyzing separately microRNAs and mRNAs. **1)** Our proposed "low-level data fusion" method is more suitable for mining the oncogenic microRNAs and their regulation of gene expression. We combined and scaled the two sources of expression profiles to produce a new meta-data. Moreover, the null hypothesis of the method adopted by Wang et al. was that the expression levels for the variables were the same for normal tissues and tumors[73], however a large fraction of the world of known microRNAs were down-expressed in cancer[6,74]. **2)** Our method is good for feature selection because it is based on the PLR algorithm which is designed for searching small functional groups of genes/microRNAs that act together and whose collective expression is strongly associated with response conditions. PLR is also superior to classification with state-of-the-art methods based on single genes[33]. Therefore, our method can identify the signature from complete mRNA and microRNA expression changes which might be ignored in an individual data source.

Compared with other studies that calculate the biomodules of microRNA-mRNA by combining both expression profiles, our study is different in two ways: **1)** The proposed unbiased and genome-wide method identifies microRNA-mRNA biomodules inclusive of both correlations and inverse correlations between microRNA-mRNA, while previous methods only addressed the latter (biased) and generally also used previous knowledge from microRNA target databases to reduce the number of comparisons (not genome-wide) [24-27]. **2)** The proposed method is an alternative to bi-clustering which results in biomodules made of de-regulated microRNA and mRNAs whose collective expression are strongly associated with cancer conditions, because the algorithm balances between the within group co-expression and between group distance. In contrast previous methods based on negative linear correlation between microRNA and mRNA expression profiles did not distinguish cancer and normal conditions[24,26] or distinguished them depending on arbitrary threshold of differential expression either[25].

Future studies of microRNA-mRNA biomodules common to multiple cancers will be: **1)** the development of robust biomodules by generating a subset of robust genes and microRNA within biomodules using cross-validations or other studies; and **2)** conducting biological validations on some uncharacterized mechanisms that we have predicted. **3)** In addition, the open question in our identified microRNA-mRNA biomodules is that some microRNAs are co-expressed with their putative target genes in this dataset. We should add additional databases, such as *transfac*, or other approaches to further characterize the putative mechanism involved in this otherwise paradoxical result, particularly when this correlation is observed between a microRNA and its putative target gene. It either challenges the disease specific roles of microRNA on its target genes[75], or challenges the prediction precision of putative target genes in those databases. In fact, neuronal-enriched microRNAs have been discovered to be either positively or negatively co-expressed with their target genes[29]. Therefore, besides the repressive effect on target expression, microRNAs might act in concert with other regulatory processes[54] to regulatory target gene expression[29,34,55]. Therefore the combination of genome-wide microRNA and mRNA expression provides increasing information to further understanding of the multiple levels of de-regulatory network in human cancer by integrating microRNA into regulatory networks[76].

3 Conclusions

Current approaches to derive microRNA-mRNA interactions from their respective expression profiles over paired tissue samples are generally focused on statistical associations of their expressions and in many cases, assume an inverse correlation to the interaction. Different from others, we focused on coordinating mRNA and microRNA expressions together by rescaling their normalized values and treating them as probe-sets of a common pool. We then developed and evaluated a novel comprehensive method to systematically identify, on a genome scale, pan-expression biomodules commonly to distinct cancers of the same tissues. These pan-expression biomodules are enriched in genes and microRNAs that have previously been identified in epithelial cancers, using literature review. Since microRNA can sometimes induce mRNA degradation with their complementary mRNA molecules, we confirmed that the biomodules were significantly enriched in microRNAs associated with their inversely correlated putative gene targets. Additionally, we have demonstrated that the pan-expression biomodules are robust classifiers of cancer vs. normal conditions. Taken together, these observations support the internal validity and biological consistency of the pan-expression biomodules in epithelial cancers. Further, the regulatory network of the identified microRNAs and corresponding genes in these biomodules may contribute to the understanding of the complexity of microRNA-mRNA interactions by providing unbiased synthetic models and hypothesis generation to cancer biologists.

4 Methods

4.1 Datasets

Expression profiles of 217 microRNA and 16,063 mRNA for the same 89 epithelial samples were published by Lu et al[13] (GSE2564) and Ramaswamy et al[31], respectively. They were among the very first paired samples of microRNA and mRNA expression, including 21 normals and 68 tumors. To generate the regulatory network, five microRNA gene target databases were downloaded and parsed to set up the comprehensive tables for human microRNA target genes. The five databases are *miRBase* v5[14,15], *miRanda* version on July 2007[12], *PicTar* server version 4.0.24[16,17], *TarBase* v4.0[18] and *TargetScan* v3.1[19], which contain all together 1,112 human microRNAs and 22,084 predictive putative microRNA gene-targets. Additionally, Gene Ontology and PubMed databases were used to do evaluation in this study, using Bioconductor softwares[77].

For the expression profiles, we did preprocessing and filtering on the downloaded data in two steps: **(A) Preprocessing and preliminary filtering suggested by the author**[13]: log₂-transformation, keeping only those probes exceeding a measurement of 7.25 (on log₂ scale) in one or more samples. This preprocessing and pre-filtering resulted in 195 microRNAs and 14,546 mRNAs[13]. **(B) Additional preprocessing steps** to the microRNA and mRNA expression profiles to concentrate on probe-sets that vary among different conditions. This variation filter[13,31] involves the following steps: a) setting a threshold using a ceiling of 1,600 units and a floor of 20 units of expression measurement; b) the maximum expression measurement was at least 5 fold greater than the minimum measurement, and c) the maximum measurement was more than 500 greater than the minimum measurement. The filtering step resulted in 130 microRNAs and 6,621 mRNAs. Finally, we performed “low-level fusion”, that is, we combined expression profiles from two sources of microarrays before step-down analysis. This is different from “high-level fusion (decision fusion)”[73] in which microRNA-mRNA analysis are done in parallel rather than together. To allow expression levels to be comparable, we first scaled the combined data. Then we performed further supervised grouping.

4.2 Algorithm for Sample Classification and Predictor-variable Grouping

Penalized logical regression (PLR) algorithm has been proposed as a stand-alone for classification of microarray gene expression data with “small sample size, larger variances” [32,33] and has been shown strongly predictive for clinical response in cancer as compared with the other above mentioned methods[33]. Let $G = (1, g_1, \dots, g_q)$ be a group of q probe-sets, θ be a vector of q parameters (each associated with one probe-set) that are trained to optimize (for example, to minimize the $S(\theta)$ in Equation 3) the PLR model by a penalized maximum likelihood principle[78]. The classical logical model of PLR is defined as[79]:

$$\log\left(\frac{p_{\theta}(G)}{1 - p_{\theta}(G)}\right) = \sum_{k=0}^q \theta_k g_k \quad (\text{Equation 1})$$

Two implementations of PLR algorithms in this study were employed. **1)** In the preliminary validation study comparing PLR usage in mRNA expression data alone, microRNA expression data alone and the combined profiles (step **i** in Figure 6), we applied three-fold cross validation using Bioconductor package MCRestimate and penalized maximum likelihood algorithm to estimate the classical PLR model, using the R package *Design*[80]. The PLR model was estimated from the data of training samples in cross-validation (CV), then in the subsequent tests within the CV, we performed a *predict* function for classification test data using the fitted model. The input data were the mRNA and/or microRNA expression together with classification annotation of test samples and the trained PLR model. The corresponding outputs here were predicted class labels of each test sample. **2)** To identify the merged microRNA-mRNA biomodule that best classified normal epithelial samples from the combined set of epithelial cancers, we performed Supervised Grouping of Predictor Variables (*pelora*) method developed by Dettling *et al.*[78,81], using the R package *Supercluster*.

In summary, the second implementation of PLR finds prioritized groups of variables (probes in a microarray) that can best explain the sample classes (cancer and normal in this study), employing the penalized log-likelihood function that is based on estimations of conditional class probabilities from PLR. As a result, the algorithm estimates conditional probabilities while focusing on similarities and intersections among predictor-variables in a supervised way[78,81]. The assumption is as follows: let's assume there exist two groups of probe-sets, set_A and set_B , and two conditions of samples, e.g. tumor and normal. If it is indicative of cancer that the centroid of set_A (C_A) is high while the centroid of set_B (C_B) is low, then two such probe-sets and their contributions to the centroids can be understood as molecular signatures to gain insights into molecular regulation in cancer[81]. For details, let x be the observed measurement of expression of one probe-set, the centroid of a group of probe-sets G is given as:

$$C_G = \frac{1}{|G|} \sum_{g \in G} \alpha_g \theta_g x_g, \quad (\text{Equation 2})$$

using a discrete parameter $\alpha_g \in \{-1, 1\}$ that allows up- or down-regulated genes to contribute in the same group. In this way, this centroid approach can derive biomodules consisting of both up- and down-expressed probe-sets. This characteristic allows for unbiased observation of both the negative expression correlation between microRNA and target gene within the group[82], and positive expression correlation between the microRNA and their host gene in the combined profiles[54]. Let θ be a vector of parameters, λ be a variable parameter for penalization control, and P be a penalty matrix, the *pelora*[78,81]

measures the strength of a probe-set group G for distinguishing the n tumor/normal phenotypes $(y_1, y_2, \dots, y_n)$ as:

$$S(\theta) = - \sum_{i=1}^n [y_i \cdot \log p_{\theta}(C_G) + (1 - y_i) \log(1 - p_{\theta}(C_G))] + n \frac{\lambda}{2} \theta^T P_{\theta}, \quad (\text{Equation 3})$$

where $p_{\theta}(C_G) = p_{\theta}[y = 1 | C_G]$ is the estimated conditional class probability from penalized logical regression analysis. The *pelora* automatically scratches the starting probe-set for each group of probe-sets, then it incrementally increases the number of probe-sets in a group G_i by adding or pruning one probe-set after the other, and recalculating after each change until it optimizes the sample classification in the training set. Once a group of probe-set and trained parameters are optimized as best, then *pleora* identifies the second best one. Therefore, by inputting a matrix of mRNA and/or microRNA expression with classification annotation of each samples, and setting the number of biomodules that should be searched to 10, while keeping other default parameter, we obtained 10 biomodules that PLR searched, each associated with their respective “rescaled penalty parameter θ ”, and the fitted values of each biomodule.

Step i: Comparing PLR Usage in mRNA Expression Data, MicroRNA expression data and the combined profiles

We performed a three-fold stratified cross-validation[35,36] as well as a re-randomization algorithms to assess and compare the predictive power of mRNA, microRNA and their combined expression profiles, where “three-fold” means random choosing the training-set to be two thirds of the samples, and “stratified” means the proportion of tumor tissues in the training set more or less the same as in the whole data set. Then, for the samples in the training-set, we employed penalized logistic regression (PLR) to train the classifier by modeling the probability of class membership. The resulting classifier model was then used to predict the samples in the remaining one-third test-set. Thus each run of the cross-validation yielded predictive class labels for the samples in the test-set. We iterated the above process 100 times and calculated the frequency of each sample correctly predicted in the test sets. The three-fold cross validation has been shown at least as conservative as the leave-one out method to provide robust estimates that are not over trained when using machine learning methods[83,84,85].

After completing the above-mentioned cross-validations, we plotted the distribution of resulting predictions for the microRNA expression profiles alone, the gene expression profiles alone and also the combined microRNA-mRNA expression profiles. Then we further compared such three distributions of the validation results using F-test. The total accuracy, $(TP+TN)/(TP+FP+TN+FN)$, of cross-validation was calculated, where TP (True positive) is the correctly classified cancer samples, FP (false positive) is the wrongly classified cancer samples, TN (True negative) is the correctly classified normal samples, and FN (false negative) is the wrongly classified normal samples. The 95% central confidence interval[37,38] of the total accuracy for the population (89 samples) in this data was estimated using an online Bayesian calculator (http://www.causascientia.org/math_stat/ProportionCI.html).

Step ii: Identification and the Quantitative Description of the Discovered MicroRNA-mRNA Biomodules and Their Interaction Networks

The two kinds of filtered expression data were finally combined (6,621 mRNAs and 130 microRNAs), log 10 transformed, and standardized. We used penalized regression

model[86] to search prioritized microRNA-mRNA biomodules with good predictive potential based on this combined data, using Bioconductor package *supclust*[78,81].

First, we searched $N = 1 \dots 10$ leading biomodules from the combined microRNAs and mRNAs data. Second, we estimated the number of including biomodules that yielded the best prediction accuracy on the validated samples by cross-validation. We iteratively ran a three-fold cross validation 100 times, using Bioconductor package *MCRestimate*. A PLR model of the learning set was obtained for each number of including group $n \in N$, and then applied to the test set while the predicted labels were compared with the true labels. The fraction of misclassified individuals was estimated for each number of including biomodules. For comparison, we also did the same estimation for mRNA and microRNA data, respectively. The optimal number of included biomodules is the n^* that achieved the lowest misclassification. We then merged the n^* leading biomodules into a merged epithelial cancer biomodule for further validation.

We summarized the genes and microRNAs involved in the merged epithelial cancer biomodule as comprehensive tables of human microRNA target genes. To describe the regulatory network, we made use of shapes and colors: circle (gene), box (microRNA), pink node (averagely up-regulated gene and/or microRNA in cancer), blue node (averagely down-regulated gene and/or microRNA in cancer), and orange node (tissue-specific expressed gene target of down-expressed microRNAs).

Step iii: Expression Patterns of the MicroRNA-mRNA Biomodules Classify Epithelial Cancers and Normal Samples Across all 11 Types of Tissues

The 10 microRNAs and 98 genes identified in the microRNA-mRNA biomodule and all samples were hierarchically clustered using complete Mutual Information distance of the expression levels, (Bioconductor package *Biodist*). We also observed the expression patterns of microRNAs and genes in each prioritized biomodule, to see whether individual biomodules could classify epithelial cancers and normal samples across 11 types of tissue.

Step iv: Gene Ontology Enrichment

Validation of Gene Ontology enrichment of the biological processes was conducted in the genes of the epithelial cancer biomodule in order to biologically characterize the biomodule. We used the Bioconductor package *GOstats*[87] to estimate the Biological Processes (BP) enriched in genes of six microRNA-mRNA biomodules (version 2.6.0 based on *Hu6800.db* v2.2.0, *hu35ksuba.db* v2.2.0 and *GO.db* v2.2.0, parameters $p\text{-value} \leq 0.01$, $gene\ count > 2$). This package *GOstats* reports a standard hypergeometric p -value to access whether the number of selected genes among the genes in an annotated array associated with a GO term is larger than expected[87], and it is a standard tool to evaluate the overrepresentation of GO terms among a given gene set and has been successfully implemented[88]. As each GO term is considered independently in *GOstat*, the review of the enrichment was conducted with the understanding that if related or possibly redundant GO terms were found to be enriched, the nature of this relationship could depend on the intrinsic annotations rather than the structure of GO (e.g. redundant annotations of genes to related GO terms). Additionally, because the raw data contains two different mRNA platforms, we also used an old version of Bioconductor package *Compdiagtools* (version in April 2007) to access the BP from whole human genome (*Gene Ontology* Built: 15-Mar-2006) enriched in these 98 genes identified in biomodules (parameters: $p \leq 0.05$, $gene\ count \geq 2$).

Step v: Evaluation of Enrichment of inversely correlated microRNA-mRNA pairs in the Biomodules computed by expression among microRNA-target pairs predicted from nucleotide sequences microRNA target

Enrichment of microRNA targets in the network was evaluated using Fisher exact test to estimate possibility of drawing the intersection of biomodule observed inversely correlated microRNA-mRNA pairs from all putative microRNA-target pairs based on sequence similarities that contained in five databases (*miRBase* v5[14,15], *miRanda* version on July 2007[12], *PicTar* server version 4.0.24[16,17], *TarBase* v4.0[18] and *TargetScan* v3.1[19]). This study was performed on the expression profiles of 217 microRNAs and 16,063 genes, about which there were 121,755 putative microRNA-target pairs among all possible 3,485,671 pairs. We then counted the six biomodules identified inversely correlated microRNA-mRNA pairs and the intersection number of identified inversely correlated pairs among putative target pairs to calculate the *p*-value.

We looked into the significance of differential expression for each gene and microRNA that act together to best classify epithelial tumors versus normal tissues. To estimate the individually false positives rate when expression changes were significant[72], we calculated the *q*-value[72] of fold-change equivalent measures on the log-transformed expression levels using *Bioconductor* package *twilight*[71]. The *Entrez Gene IDs* for mRNAs were mapped from probe sets using package *hu6800* version 2.2.0, and package *hu35ksuba* version 2.2.0 in R.

Step vi: Systematic Review of Literature to Identify the Relevance of the Gene Patterns Observed in the Biomodules

We assessed the significance of the co-occurrence for every pair of two terms in which one was a microRNA or gene symbol and the other was a cancer type, to see the association between cancer and gene or microRNAs in the identified biomodule. Similarly, using the online literature mining tool *PubMatrix*[89], we searched the *Pubmed* indexed publications and counted the number of abstracts containing the symbol of the identified microRNA or gene and of the epithelial cancer type, also the number of co-occurrence of pair of terms. The total number of indexed publication was derived from the PubMed online review tool. Subsequently, Fisher's exact test was used to estimate the significance to co-occurrence. Let *N* be the totally records in PubMed, *n*₂ be the number of abstracts containing a symbol of microRNA or gene, *n*₃ be the number of abstracts about cancer, and *n*₁ is the number of abstract mentioning both, we got the contingency table (Table 3) for each symbol and cancer. Then one-side Fisher's exact test was applied to evaluate the significance of co-occurrence. A threshold of unadjusted *p* < 0.05 was set for to indicate a true positive result in this evaluation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by the National Natural Science Foundation of China (Grant Nos. 60971099, 60671018 and 60771024), Center for Multilevel Analyses of Genomic and Cellular Networks (Grant No. 1U54CA121852-01A1), and the Cancer Research Foundation. We thank WALTERS Adrienne for her contribution to editing. We also thank XIE Jianming for contributive discussion on the biological impact.

References

1. Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* 2004;101(25):9309–9314. [PubMed: 15184677]
2. Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 2004;36(10):1090–1098. [PubMed: 15448693]
3. Yang X, Bentink S, Spang R. Detecting common gene expression patterns in multiple cancer outcome entities. *Biomed Microdevices* 2005;7(3):247–251. [PubMed: 16133813]
4. Volinia S, Calin GA, Liu CG, et al. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* 2006;103(7):2257–2261. [PubMed: 16461460]
5. Calin GA, Dumitru CD, Shimizu M, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* 2002;99(24):15524–15529. [PubMed: 12434020]
6. Calin GA, Ferracin M, Cimmino A, et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* 2005;353(17):1793–1801. [PubMed: 16251535]
7. Iorio MV, Ferracin M, Liu CG, et al. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res* 2005;65(16):7065–7070. [PubMed: 16103053]
8. Yanaihara N, Caplen N, et al. Bowman E, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* 2006;9(3):189–198. [PubMed: 16530703]
9. Ruike Y, Ichimura A, Tsuchiya S, et al. Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. *J Hum Genet* 2008;53(6):515–523. [PubMed: 18465083]
10. Ambs S, Prueitt RL, Yi M, et al. Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. *Cancer Res* 2008;68(15):6162–6170. [PubMed: 18676839]
11. Pasquinelli AE, Reinhart BJ, Slack F, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 2000;408(6808):86–89. [PubMed: 11081512]
12. John B, Enright AJ, Aravin A, et al. Human MicroRNA targets. *PLoS Biol* 2004;2(11):e363. [PubMed: 15502875]
13. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature* 2005;435(7043):834–838. [PubMed: 15944708]
14. Griffiths-Jones S. miRBase: the microRNA sequence database. *Methods Mol Biol* 2006;342:129–138. [PubMed: 16957372]
15. Griffiths-Jones S, Saini HK, van Dongen S, et al. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008;36(Database issue):D154–158. [PubMed: 17991681]
16. Yang Y, Wang YP, Li KB. MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics* 2008;9(Suppl 12):S4. [PubMed: 19091027]
17. Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* 2006;38(12):1452–1456. [PubMed: 17072316]
18. Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* 2006;12(2):192–197. [PubMed: 16373484]
19. Grimson A, Farh KK, Johnston WK, et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 2007;27(1):91–105. [PubMed: 17612493]
20. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;136(2):215–233. [PubMed: 19167326]
21. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods* 2006;3(11):881–886. [PubMed: 17060911]
22. Lewis BP, Shih IH, Jones-Rhoades MW, et al. Prediction of mammalian microRNA targets. *Cell* 2003;115(7):787–798. [PubMed: 14697198]

23. Lanza G, Ferracin M, Gafa R, et al. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Mol Cancer* 2007;6:54. [PubMed: 17716371]
24. Joung JG, Hwang KB, Nam JW, et al. Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics* 2007;23(9):1141–1147. [PubMed: 17350973]
25. Liu B, Li J, Tsykin A. Discovery of functional miRNA-mRNA regulatory modules with computational methods. *Journal of Biomedical Informatics* 2009;42(4):685–691. [PubMed: 19535005]
26. Tran DH, Satou K, Ho TB. Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics* 2008;9(Suppl 12):S5. [PubMed: 19091028]
27. Xin F, Li M, Balch C, et al. Computational analysis of microRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. *Bioinformatics* 2009;25(4):430–434. [PubMed: 19091772]
28. Gennarino VA, Sardiello M, Avellino R, et al. MicroRNA target prediction by expression analysis of host genes. *Genome Res* 2009;19(3):481–490. [PubMed: 19088304]
29. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell* 2007;26(5):753–767. [PubMed: 17560377]
30. O'Donnell KA, Wentzel EA, Zeller KI, et al. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* 2005;435(7043):839–843. [PubMed: 15944709]
31. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 2001;98(26):15149–15154. [PubMed: 11742071]
32. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics* 2004;5(3):427–443. [PubMed: 15208204]
33. Shen L, Tan EC. Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Trans Comput Biol Bioinform* 2005;2(2):166–175. [PubMed: 17044181]
34. Yu X, Lin J, Zack DJ, et al. Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. *Nucleic Acids Res* 2008;36(20):6494–6503. [PubMed: 18927108]
35. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002;97(457):77–87.
36. Fort G, Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression. *Bioinformatics* 2005;21(7):1104–1111. [PubMed: 15531609]
37. Nicholson BJ. On the F-Distribution for Calculating Bayes Credible Intervals for Fraction Nonconforming. *IEEE Transactions on Reliability* 1985:227–228.
38. Harper WL, Hooker CA, Harper WL, et al. Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science. 1976:175.
39. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002;3(7):RESEARCH0036. [PubMed: 12184810]
40. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;7(Suppl 1):S7. [PubMed: 16723010]
41. Stallings RL, Nair P, Maris JM, et al. High-resolution analysis of chromosomal breakpoints and genomic instability identifies PTPRD as a candidate tumor suppressor gene in neuroblastoma. *Cancer Res* 2006;66(7):3673–3680. [PubMed: 16585193]
42. Bonci D, Coppola V, Musumeci M, et al. The miR-15a/miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. *Nat Med* 2008;14(11):1271–1277. [PubMed: 18931683]
43. Clinton M, Graeve L, Dorry H, et al. Evidence for nuclear targeting of prothymosin and parathymosin synthesized in situ. *Proc Natl Acad Sci U S A* 1991;88(15):6608–6612. [PubMed: 1862085]
44. Vareli K, Frangou-Lazaridis M, van der Kraan I, et al. Nuclear distribution of prothymosin alpha and parathymosin: evidence that prothymosin alpha is associated with RNA synthesis processing and parathymosin with early DNA replication. *Exp Cell Res* 2000;257(1):152–161. [PubMed: 10854063]

45. Lei M. The MCM complex: its role in DNA replication and implications for cancer therapy. *Curr Cancer Drug Targets* 2005;5(5):365–380. [PubMed: 16101384]
46. Tsitsilonis OE, Bekris E, Voutsas IF, et al. The prognostic value of alpha-thymosins in breast cancer. *Anticancer Res* 1998;18(3A):1501–1508. [PubMed: 9673361]
47. Fujiwaki R, Hata K, Nakayama K, et al. Thymidine kinase in epithelial ovarian cancer: relationship with the other pyrimidine pathway enzymes. *Int J Cancer* 2002;99(3):328–335. [PubMed: 11992400]
48. Holleman A, den Boer ML, de Menezes RX, et al. The expression of 70 apoptosis genes in relation to lineage, genetic subtype, cellular drug resistance, and outcome in childhood acute lymphoblastic leukemia. *Blood* 2006;107(2):769–776. [PubMed: 16189266]
49. Biswas S, Chytil A, Washington K, et al. Transforming growth factor beta receptor type II inactivation promotes the establishment and progression of colon cancer. *Cancer Res* 2004;64(14):4687–4692. [PubMed: 15256431]
50. Hoornaert I, Marynen P, Goris J, et al. MAPK phosphatase DUSP16/MKP-7, a candidate tumor suppressor for chromosome region 12p12-13, reduces BCR-ABL-induced transformation. *Oncogene* 2003;22(49):7728–7736. [PubMed: 14586399]
51. Ku JL, Yoon KA, Kim WH, et al. Establishment and characterization of four human pancreatic carcinoma cell lines Genetic alterations in the TGFBR2 gene but not in the MADH4 gene. *Cell Tissue Res* 2002;308(2):205–214. [PubMed: 12037578]
52. Grady WM, Markowitz SD. Genetic and epigenetic alterations in colon cancer. *Annu Rev Genomics Hum Genet* 2002;3:101–128. [PubMed: 12142355]
53. Guo J, Kleeff J, Li J, et al. Expression and functional significance of CDC25B in human pancreatic ductal adenocarcinoma. *Oncogene* 2004;23(1):71–81. [PubMed: 14712212]
54. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 2005;11(3):241–247. [PubMed: 15701730]
55. Shalgi R, Lieber D, Oren M, et al. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput Biol* 2007;3(7):e131. [PubMed: 17630826]
56. Wu F, Zhu S, Ding Y, et al. MicroRNA-mediated regulation of Ubc9 expression in cancer cells. *Clin Cancer Res* 2009;15(5):1550–1557. [PubMed: 19223510]
57. Kozaki K, Imoto I, Mogi S, et al. Exploration of tumor-suppressive microRNAs silenced by DNA hypermethylation in oral cancer. *Cancer Res* 2008;68(7):2094–2105. [PubMed: 18381414]
58. Sorrentino A, Liu CG, Addario A, et al. Role of microRNAs in drug-resistant ovarian cancer cells. *Gynecol Oncol* 2008;111(3):478–486. [PubMed: 18823650]
59. Yu J, Wang F, Yang GH, et al. Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. *Biochem Biophys Res Commun* 2006;349(1):59–68. [PubMed: 16934749]
60. Lui WO, Pourmand N, Patterson BK, et al. Patterns of known and novel small RNAs in human cervical cancer. *Cancer Res* 2007;67(13):6031–6043. [PubMed: 17616659]
61. Chen X, Guo X, Zhang H, et al. Role of miR-143 targeting KRAS in colorectal tumorigenesis. *Oncogene* 2009;28(10):1385–1392. [PubMed: 19137007]
62. Michael MZ, SM OC, van Holst Pellekaan NG, et al. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol Cancer Res* 2003;1(12):882–891. [PubMed: 14573789]
63. Chen HC, Chen GH, Chen YH, et al. MicroRNA deregulation and pathway alterations in nasopharyngeal carcinoma. *Br J Cancer* 2009;100(6):1002–1011. [PubMed: 19293812]
64. Nasser MW, Datta J, Nuovo G, et al. Down-regulation of microRNA-1 (miR-1) in lung cancer. Suppression of tumorigenic property of lung cancer cells and their sensitization to doxorubicin-induced apoptosis by miR-1. *J Biol Chem* 2008;283(48):33394–33405. [PubMed: 18818206]
65. Datta J, Kutay H, Nasser MW, et al. Methylation mediated silencing of MicroRNA-1 gene and its role in hepatocellular carcinogenesis. *Cancer Res* 2008;68(13):5049–5058. [PubMed: 18593903]
66. Karwowska S, Zolla-Pazner S. Passive immunization for the treatment and prevention of HIV infection. *Biotechnol Ther* 1991;2(1-2):31–48. [PubMed: 1845123]
67. Akao Y, Nakagawa Y, Naoe T. let-7 microRNA functions as a potential growth suppressor in human colon cancer cells. *Biol Pharm Bull* 2006;29(5):903–906. [PubMed: 16651716]

68. Takamizawa J, Konishi H, Yanagisawa K, et al. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* 2004;64(11):3753–3756. [PubMed: 15172979]
69. Schultz J, Lorenz P, Gross G, et al. MicroRNA let-7b targets important cell cycle molecules in malignant melanoma cells and interferes with anchorage-independent growth. *Cell Res* 2008;18(5):549–557. [PubMed: 18379589]
70. Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res* 2006;34(5):1646–1652. [PubMed: 16549876]
71. Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics* 2005;21(12):2921–2922. [PubMed: 15817688]
72. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003;100(16):9440–9445. [PubMed: 12883005]
73. Wang, Y.; Dunham, MH.; Waddle, JA., et al. Classification for poorly differentiated tumor classification using both messenger rna and microrna expression profiles. 2006 Computational Systems Bioinformatics Conference (CSB 2006); Stanford, California. 2006.
74. Thomson JM, Newman M, Parker JS, et al. Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev* 2006;20(16):2202–2207. [PubMed: 16882971]
75. Kuhn DE, Martin MM, Feldman DS, et al. Experimental validation of miRNA targets. *Methods* 2008;44(1):47–54. [PubMed: 18158132]
76. Kanellopoulou C, Monticelli S. A role for microRNAs in the development of the immune system and in the pathogenesis of cancer. *Semin Cancer Biol* 2008;18(2):79–88. [PubMed: 18291671]
77. Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol* 2006;411:119–134. [PubMed: 16939789]
78. Dettling M, Buhlmann P. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 2004;90(1):106–131.
79. Cessie SL, Houwelingen JV. Ridge estimators in logistic regression. *Applied Statistics* 1990;41:191–201.
80. Alzola C, Harrell F. An Introduction to S and the Hmisc and Design Libraries.
81. Dettling M, Buhlmann P. Supervised clustering of genes. *Genome Biol* 2002;3(12):RESEARCH0069. [PubMed: 12537558]
82. Huang JC, Morris QD, Frey BJ. Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol* 2007;14(5):550–563. [PubMed: 17683260]
83. Feng CJ, Yu AG, Unnati K, et al. Threefold vs fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. *Journal of manufacturing systems* 2005;24(2):93–105.
84. Campbell GP, Curran JM, Miskelly GM, et al. Compositional data analysis for elemental data in forensic science. *Forensic Sci Int* 2009;188(1-3):81–90. [PubMed: 19411149]
85. Marchese A, Beischlag TV, Nguyen T, et al. Two gene duplication events in the human and primate dopamine D5 receptor gene family. *Gene* 1995;154(2):153–158. [PubMed: 7890157]
86. Dettling M, Buhlmann P. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 2004;90(1):106–131.
87. Falcon S, Gentleman R. Using GStats to test gene lists for GO term association. *Bioinformatics* 2007;23(2):257–258. [PubMed: 17098774]
88. Martin-Subero JI, Kreuz M, Bibikova M, et al. New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* 2009;113(11):2488–2497. [PubMed: 19075189]
89. Becker KG, Hosack DA, Dennis G, et al. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* 2003;4:61. [PubMed: 14667255]

Figure 1. Standard box—and-whisker plots of the error rate of repeated cross-validation on mRNA profiles, microRNA profiles and on the combined profiles

Each box presents the smallest observation, lower quartile, median, upper quartile and the largest observation, the whiskers are the lines that extend to a maximum of 1.5 times of the inter-quartile range (the range of the middle two quartiles) excluding outliers, and the circles represent the outliers. The y-axis gives the probability of samples to be wrongly predicted. The dashed line is the normal (lower) prevalence of the samples.

Figure 2. Box and whisker plot showing the variation of the misclassification rates (y axis) of different number of microRNA-mRNA biomodules (x axis), conducted in three expression datasets

mRNA alone (left), microRNA alone (middle), and the combined microRNA-mRNA expression (right). The optimal area is the lowest error rate as measured by the median (black horizontal line) and the direction of the box plot. Optimal results are observed in the leftmost group at X=6. The upper edge of the boxes indicates the 75th percentile of the data set, and the lower hinges indicate the 25th percentile. The “whiskers” are the lines that extend to a maximum of 1.5 times of the inter-quartile range (the range of the middle two quartiles) excluding outliers, and the circles represent the outliers. The dashed line in each plot is the normal (lower) prevalence of the samples.

Figure 3. Two-dimensional projection of supervised cluster of the microarray samples based on the mRNA (left), microRNA (middle), and the combined (right) profiling data

In each plot, the expression level of its first biomodule for discrimination of tumor (labeled as 1 in the plot) versus normal (labeled as 0 in the plot) is given on the x axis and of its second biomodule on the y axis.

Figure 4. The heatmap of log-transformed expression levels of microRNAs-genes biomodules across 89 epithelial samples

The 10 microRNAs (symbols in red) and 98 genes and all samples are ordered by a hierarchical cluster agglomerated on complete Mutual Information distance of the expression levels, using Bioconductor package Biodist. The two black vertical lines range the normal samples that were clustered together. The annotation of the 89 sample are “tumor_T” or “tumor_N” (e.g. normal kidney tissue is Kid-N); Kid=kidney, BLDR=bladder, PAN=pancreas, BRST=breast, BLDR=bladder, UT=uterus, MELA= melanoma, MESO=mesothelioma, PROST=prostate.

Figure 5. Regulatory network consisting of microRNAs and their putative gene targets in the six combined microRNA-mRNA biomodules

Circle (gene), box (microRNA), pink node (up-regulated gene or microRNA in epithelial cancer), blue node (down-regulated gene or RNA in epithelial cancer), and orange node (tissue-specific up-regulation of the gene target associated with the down-expressed microRNAs as described in Suppl. Table 3). It includes 9 out of 10 down-regulated microRNAs and their predicted target genes (n=26) summarized from five public databases observed among the 98 genes of the six combined biomodules.

Figure 6. Data and six steps of analysis

Data) data collection, profile preprocessing and filtering; The **analysis** was performed as: **i)** a preliminary validation study comparing PLR usage in mRNA expression data alone, microRNA expression data alone and the combined results; **ii)** identification and the quantitative description of the discovered microRNA-mRNA biomodules and their interaction networks; **iii)** expression patterns of the microRNA-mRNA biomodules classify epithelial cancers and normal samples across all 11 types of tissues; **iv)** conducting the Gene Ontology enrichment of the genes associated with these biomodules to provide an unbiased description of their biological mechanisms. Finally, we **evaluated** the identified genes and microRNAs in the biomodules in two ways: **(v)** calculating the enrichment of gene targets of microRNAs among the negative co-expression patterns of microRNA and mRNA; **(vi)** systematic review of literature to identify the relevance of the gene patterns observed in the biomodules. **Legend:** GO: Gene Ontology, PPI: protein-protein interaction, PLR: penalized logistic regression, CV: Cross-Validation.

Table 1
The population size and classification total accuracy in percentage (%) from 3-fold cross-validation for each type of cancer and as a whole in this study

The last row of this table lists the sample populations for each tissue and as a whole. The last three columns list the total accuracy for all types of cancer as a whole and the corresponding confidence intervals (CI). CI: 95% confidence interval calculated by a Bayesian calculator[37,38]; TP: true positive; TN: true negative.

	colon	prostate	uterus	kidney	ovary	breast	bladder	melanoma	mesothelioma	lung	pancreas	Total Accuracy	CI -Lower Limit	CI -Upper limit
microRNA/mRNA (%)	63.6	91.7	90.9	100	100	100	100	100	100	100	100	93	86	97
MicroRNA (%)	45.5	75	81.8	85.7	80	100	100	100	100	100	100	85	77	91
mRNA (%)	45.5	91.7	72.7	71.4	100	88.9	71.4	0	100	100	100	80	70	87
Sample Population	11	12	11	7	5	9	7	3	8	7	9	89	-	-

Table 2
Biological Processes Enriched in 98 Genes of merged MicroRNA-mRNA Biomodule

The listed GO terms meet the standard of significance ($p \leq 0.05$, gene count ≥ 2 , CompdiagTools software, see **Methods**)

GO ID	Biological Process	p	Genes
0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	0.003	IRAK1, TGFBR2
0051216	cartilage development	0.003	BMP2, ZBTB7A
0006957	complement activation, alternative pathway	0.005	CFD, CFB
0006260	DNA replication	0.007	NFIB, PTMS, MCM7, TK1
0040007	growth	0.008	BMP2, INHBA
0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.05	FPGS, TK1, FPGS
0006470	protein amino acid dephosphorylation	0.05	DUSP16, TGFBR2, CDC25B

Table 3**Contingency table**

Ref= individual reference in PubMed

	#ref. including gene or microRNA symbol	#ref. not including the symbol	Σ
#ref. Including epithelial cancers	n_1	$n_3 - n_1$	n_3
#ref. not including epithelial cancers	$n_2 - n_1$	$N - n_2 - n_3 + n_1$	$N - n_3$
Σ	n_2	$N - n_2$	N