MICA: MicroRNA Integration for Active Module Discovery

Ayat Hatem
Dept. Computer Science
University of Massachusetts
Lowell
Lowell, MA
ayat_hatem@uml.edu

Kamer Kaya
Sabanci University
Faculty of Engineering and
Natural Sciences
Istanbul, Turkey
kaya@sabanciuniv.edu

Jeffrey Parvin
Dept. Biomedical Informatics
The Ohio State University
Columbus, OH
Jeffrey.Parvin@osumc.edu

Kun Huang
Dept. Biomedical Informatics
The Ohio State University
Columbus, OH
Kun.Huang@osumc.edu

Ümit V. Çatalyürek
Dept. Biomedical Informatics
Dept. Elect. and Comp. Eng.
The Ohio State University
Columbus, OH
umit@bmi.osu.edu

ABSTRACT

A successful method to address disease-specific module discovery is the integration of the gene expression data with the protein-protein interaction (PPI) network. Although many algorithms have been developed for this purpose, they focus only on the network genes (mostly on the well-connected ones); totally neglecting the genes whose interactions are partially or totally not known. In addition, they only make use of the gene expression data which does not give the complete picture about the actual protein expression levels. The cell uses different mechanisms, such as microRNAs, to post-transcriptionally regulate the proteins without affecting the corresponding genes' expressions. Due to this complexity, using a single data type is definitely not the correct way to find the correct module(s). Today, the unprecedented amount of publicly available disease-related heterogeneous data encourages the development of new methodologies to better understand complex diseases.

In this work, we propose a novel workflow MICA, which, to the best of our knowledge, is the first study integrating miRNA, mRNA, and PPI information to identify disease-specific gene modules. The novelty of the MICA lies in many directions, such as the early modification of mRNA expression with microRNA to better highlight the indirect dependencies between the genes. We applied MICA on microRNA-Seq and mRNA-Seq data sets of 699 invasive ductal carcinoma samples and 150 invasive lobular carcinoma samples from the Cancer Genome Atlas Project (TCGA). The MICA modules are shown to unravel new and interesting dependencies between the genes. Additionally, the modules accurately differentiate between the case and control samples while being highly enriched with disease-specific pathways and genes.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Algorithms, Experimentation

Keywords

microRNA, active modules, PPI network

1. INTRODUCTION

In complex diseases, genes interact via pathways and modules to function. Additionally, the interaction patterns in such diseases change based on the cell type and the cell surrounding conditions [8]. A well-structured characterization and analysis of gene modules have been intriguing, especially for extremely heterogeneous diseases. Cancer is such a disease: the derivative tissue differs for many cancer types which can have many subtypes. Identifying a biologically valid module is important for each cancer type and subtype since this information can be very useful to help the diagnosis and improve the treatment and its success rate significantly [2].

Arguably, one way to identify modules is to look for (denser) gene clusters in biological networks. The accuracy and benefits of this approach can be significantly improved by the integration of various data to better highlight these gene modules [40]. Following this idea, various module extraction techniques have been proposed, e.g., [26, 15, 54]. These dense modules are called *active modules* since the gene expression data, which is dynamically changing, is integrated with the static PPI network. Hence, these modules are *active* in certain cells or conditions. Many algorithms have been developed to better make use of the network and/or other types of data, e.g., genotypic data, as well [27, 29]. Although the algorithms based on gene expression signatures have proven to be flexible and useful for certain diseases, they do not provide a *be-all and end-all* solution. Today, we have heterogeneous data that can be used to boost the accuracy, but many of the existing algorithms cannot exploit heterogeneity. Besides, they are usually restricted only to the (possibly well-connected) proteins/genes in the networks while ignoring the genes whose interactions are not discovered yet.

MicroRNAs (miRNAs) are small non-coding RNAs used by the cell to post-transcriptionally regulate gene expression levels [16]. They inhibit protein synthesis by either stopping the protein translation or by performing mRNA degradation. miRNAs constitute an important inhibition technique that has been shown to be very important in different diseases, specifically, in cancer progression [28]. For instance, miRNAs were found to be differentially expressed in breast cancer in addition to successfully classifying estrogen and progesterone receptors, and HER2/neu status [4]. Additionally, many techniques have been proposed to extract miRNA-mRNA interactions that are specific to different cancer types [31, 58]. Hence, using miRNAs for active module discovery is a promising technique to increase the accuracy and success rate of the cancer treatments.

Many works that integrate miRNA and mRNA data assume that the miRNA effect on the mRNA can be detected by examining gene expression values [23, 59]. However, the protein expression levels can be significantly affected by the miRNA without having any apparent effect on the gene expression levels [1]. Cun and Frölich suggested another miRNA-mRNA integration method that overcomes the above problem [12]. Basically, they integrated the PPI and miRNA-target gene networks into one heterogenous network. The network is then further used to prioritize the different genes. However, by focusing only in prioritizing genes through the PPI network, it is hard to detect the active gene modules (having indirect dependencies) connected via genes that are not in the PPI network or have no change in their expression at mRNA level.

Even though the techniques using gene expression levels provide valuable information, they do not show the whole picture. In this work, we exploit another miRNA and mRNA interaction, the *inhibition of protein translation*, rather than mRNA degradation. We believe that if the gene expression levels are adjusted based on the corresponding miRNAs' expression levels more interesting gene-gene dependencies can be unraveled. We propose a workflow Mica which employs heterogeneous data sources and adopts independent component analysis [25] to extract active modules. Similar to Cun and Frölich, we propagate the effect of differentially expressed miRNA into its experimentally validated target-mRNA. Such a propagation is valid since if a miRNA is active, its target genes could be active too [35]. These dependencies are then mapped back to the PPI network to extract the connected modules.

The contribution of our work lies in many direction, including (1) the less dependence of Mica on the used PPI network, (2) the generation of different modules for different disease subtypes, (3) the enrichment of Mica modules with disease-specific pathways and genes; for invasive lobular carcinoma, they are enriched with breast cancer genes, such as BRCA1, and important pathways such as the *pathways in cancer* pathway with ERRB2, MYC, and RB1 genes, and (4) the accurate classification of case and control samples.

The rest of the paper is organized as follows: In the next section, we briefly explain how the miRNAs and genes interact with each other and describe the tools used in Mica. Section 3 described the proposed workflow and Section 4 describes our experimental setting and evaluates the results. Section 5 concludes the paper.

2. BACKGROUND

2.1 miRNA-mRNA interactions

There are three types of interactions between a group of miRNAs and a target gene; *synergetic*, *complementary*, and *additive*. A *synergetic* effect implies that all the miRNAs affecting the gene must be expressed together in order to have mRNA degradation or protein inhibition [9]. Rather, miRNAs can act *complementary* by requiring only one out of the miRNA set to be expressed [9]. In an *additive* interaction, each miRNA alone has an effect while the overall effect is increased if multiple miRNAs are expressed [52].

To propagate the effect of the differentially expressed miRNA into its experimentally validated target-mRNA, a mathematical model which can take *synergetic*, *complementary*, and *additive* behaviors into account at the same time is needed. Both *synergetic* and *complementary* behaviors imply that we must have a complete picture of the activity of all miRNAs affecting a certain gene. However, due to the lack of data, such a model would be harder to build. Therefore, in our paper, we build our mathematical model based on the *additive* type of interactions.

2.2 Independent Component Analysis

Independent Component Analysis (ICA) is a famous technique used to solve the *Blind Source Separation* problem: Given an input with multiple, linearly mixed sources, it distinguishes the sources by minimizing their statistical dependencies [25]. It has been used in the literature to cluster different genes together or for sample classification, e.g., [38, 50, 45, 44]. In the context of gene expression, ICA decomposes an input expression into its possible *expression modes* [38]. For an $n \times m$ gene expression matrix $\mathbf{X}$, where rows and columns correspond to genes and samples, respectively, ICA decomposes $\mathbf{X}$ into

$$\mathbf{X}^T = \mathbf{A} \times \mathbf{S} \tag{1}$$

such that $\mathbf{S}$ is a $\ell \times n$ matrix for $\ell \leq m$. The rows of $\mathbf{S}$ are (statistically) as independent as possible and correspond to the components. The columns correspond to the genes and the entry $\mathbf{S}_{cg}$ is the gene g's contribution to the component c. $\mathbf{A}$ is an $m \times \ell$ matrix where its rows correspond to samples. The entry $\mathbf{A}_{sc}$ shows the component c's contribution to sample s. Many approximation algorithms have been proposed to efficiently find $\mathbf{A}$ and $\mathbf{S}$, e.g., `fastICA` [24], `JADE` [6], and `InfoMax` [3]. `fastICA` tries to identify non-Guassian components under the assumption that Gaussian components represent the noise. This algorithm can stuck in a local minima, hence multiple iterations, thus multiple estimates can be necessary [18, 10].

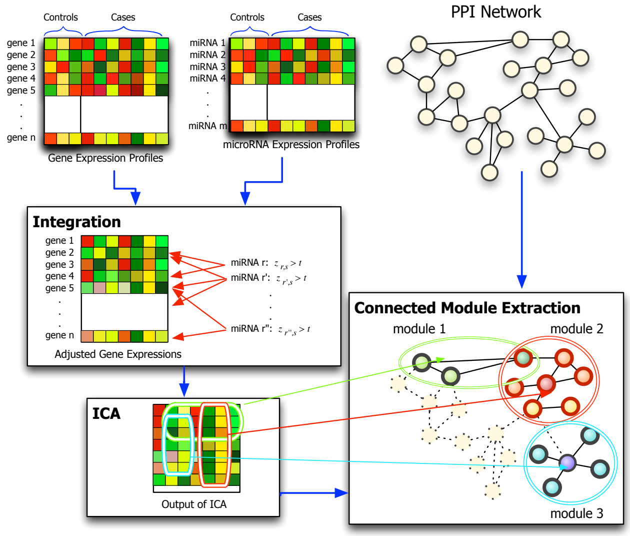


Figure 1: Mica: The workflow starts with integrating miRNA and mRNA data by adjusting the mRNA data using the miRNA data. Then, ICA is applied on the resulting new gene-expression matrix. Finally, for each independent component obtained by ICA, the largest connected module from the PPI network is extracted using the significant genes in the component.

3. THE MICA WORKFLOW

Mica consists of three main parts as shown in Figure 1:

3.1 Data integration

The miRNA and gene expression data are usually integrated using correlation-based methods with the assumption that the miRNA effect on mRNA should be apparent on the gene expression level. Rather than the suppression of the gene expression, *the inhibition of the protein translation* can also be used. Traditional approaches cannot exploit this effect. Our novel integration scheme uses miRNA expression levels to adjust the gene expression. Hence, if a gene is affected by an miRNA at the inhibition level, the proposed integration makes the effect visible on the expression level. To do this, for each sample s, we first compute

$$\beta_{g,s} = \frac{\left| \sum_{\{r:\ r \text{ affects } g,\ Z_{r,s}<0\}} Z_{r,s} \right|}{\sum_{\{r:\ r \text{ affects } g,\ Z_{r,s}>0\}} Z_{r,s}} \tag{2}$$

where $Z_{r,s}$ is the z-score of miRNA r in sample s that is experimentally verified to affect gene g. It is calculated by

$$Z_{r,s} = (x_{r,s} - \mu_r)/\sigma_r \tag{3}$$

where $x_{r,s}$ is the expression level of miRNA r in sample s, and μ_r and σ_r are the mean and standard deviation of r's expression level across all the control samples. In (2), the miRNAs are divided into two groups since they affect a gene differently. In general, when an miRNA r is down-regulated, i.e., has a negative z-score, then the expression of g will increase. On the other hand, when r is up-regulated then the expression of g will decrease. Accordingly, the final gene expression is calculated as

$$e'_{g,s} = \beta_{g,s} \times e_{g,s} \tag{4}$$

where $e_{g,s}$ and $e'_{g,s}$ are the original and adjusted expression levels of gene g in sample s.

For data integration, (4) is applied to each gene-sample pair. To avoid noise, only the miRNAs with an absolute z-score at least t_R in more than 10% of the samples are kept. Additionally, $\beta_{g,s}$ must be $> t_R$ or $< \frac{1}{t_R}$ in order to modify $e_{g,s}$, i.e., allowing only the group of significantly up-regulated or down-regulated miRNAs to change $e_{g,s}$.

A more accurate gene adjustment equation can be obtained by using miRNA, mRNA, and protein expression values for the same set of samples. However, to the best of our knowledge, with the current available technology, there is not enough miRNA, mRNA, and protein expression data for the same set of samples. Therefore, we believe that the above simple adjustment equation while including only significantly up(down)-regulated genes would be sufficient for our current study.

As mentioned above, miRNAs can affect the genes in a synergetic, complementary, or additive way. Our integration equation (4) is additive i.e., the gene expression level will be affected more if several miRNAs affect it (additive). Yet, when only a single miRNA is active, it will still affect the expression level. At the end, our goal is to better highlight the dependencies between the genes rather than finding exact protein expression values; there are many unknown factors affecting the actual protein expression.

3.2 ICA on gene expression values

After data integration, the adjusted gene expression values are fed to ICA for which the `R` version of `fastICA` is used [24]. To avoid local minimas and unreliable independent component estimates, we follow the method proposed by Chiappetta et al. [10]: we run `fastICA` κ times and store the estimates at each run. Then, the Pearson correlation coefficients between the components from different estimates are computed to distinguish the most similar ones. We construct a k-partite similarity graph $G = (V, E)$ where $V = V_1 \cup \cdots \cup V_\kappa$ is the set of all components returned by ICA and V_i is the set of components obtained in the ith run. The edge set E contains an edge (c, c') between two components from different runs if their Pearson correlation is at least 0.9. To obtain the final component set, we partition G to its maximally connected subgraphs. For each connected subgraph C of G with at least κ vertices, we construct a representative component by computing the average of the rows corresponding to the vertices in C.

An important ICA parameter is the number of components ℓ to be generated. A naïve method is setting $\ell = m$, the number of samples, which is not useful in our case, since when ℓ is large, ICA will probably return uninteresting, subcomponent-type structures [37]. We follow another approach [44] based on an earlier method [20]. We first apply Singular Value Decomposition (SVD) to the original gene expression matrix to reduce the dimensionality. We do the same for a randomly permuted version of the same matrix. The actual variance obtained from each SVD component is used to draw a curve of the information gain. A similar curve is also generated for the randomly permuted case. The optimal number of components would be the point of intersection of these two curves, i.e., when the information obtained from the random components is higher than the information obtained from the actual components.

The matrices **S** and **A** generated by ICA can be used to determine which genes are significant in each component and which components are significant in each sample, respectively. There are different options to pick the significant components, e.g., [46, 10]. Here we used a variant of the correlation method [45]; the Wilcoxon signed-rank test is used to calculate a p-value for each component based on its weight distribution over the case and control samples instead of computing the Pearson correlation. The Bonferroni correction method is then used to correct the p-value. The results of such a step are the components showing a significant change in weight distribution in the cases compared to control samples. We further classify which component weight per sample is causing such a high p-value by computing the component μ and σ values from its weights in the control samples. We then compute the z-score for each component-case sample pair. Hence, a component is *significant* for a case if the corresponding z-score is at least a threshold t_C.

To find the component-related genes, we use the z-score threshold based method [46, 50] which is very effective in returning the important genes for each component C. Basically, the z-score for each gene in C is computed from the μ and σ of all of the genes' weights inside C. Then for each C, the genes with a z-score at least t_G are considered to be a *member* of C.

3.3 Connected module extraction

The connected PPI modules are extracted by mapping the set of member genes in each component to the PPI network and extracting the largest connected module. If there is no connected module or if the largest one is not large enough the threshold t_G used to pick the member genes for each component is relaxed to allow more connectivity. However, as the results will show, each component yields a large connected module in PPI. In addition, recent studies also showed that the components generated by ICA (or similar techniques) are either highly enriched in the PPI network [59] or highly enriched with signaling pathways [50].

Each component we found after the second step is expected to generate a connected module. It is crucial to define a scoring function to determine which module is the most important one. Here, we define the importance of a module from the importance of the genes inside the module. In more details, a module containing many genes with high $Z_{c,g}$ values in a component c would be more important than another module containing many genes with low $Z_{c,g}$. Although a large module is preferable, we do not want the modules to be too large. Therefore, after determining the member genes in each component c, the following scoring function is used:

$$scr(c) = \sum_{g \in c} Z_{c,g} / \sqrt{|c|} \tag{5}$$

where $|c|$ is the number of member genes in c. We used $\sqrt{|c|}$ instead of $|c|$ since we still want to give a higher score to a larger module. A gene g will have a high $Z_{c,g}$ value if it is significant for c. Therefore, a module with many important genes is considered important. Albeit the simplicity of the scoring method, we believe that it would be sufficient to distinguish the modules with the most important genes.

4. EXPERIMENTAL RESULTS

We implemented our workflow Mica in `R` and used the available implementation of `fastICA`. To demonstrate the effectiveness of the proposed workflow, i.e., the added benefits of the early integration of microRNA datasets, we compared the modules obtained by Mica against the ones obtained using ICA and `DEGAS` [54], using the original gene expression values. However, we excluded Cun and Frölich [12] method from the comparison since it depends on using microarray datasets while we focus here on RNA-Seq datasets. `DEGAS` is a set-cover based algorithm efficient in detecting dysregulated pathways. It tries to detect a module with at least k differentially expressed (DE) genes shared between most of the samples. We tuned the `DEGAS` parameters to detect the best module according to the size measure provided by the tool, which is the probability of randomly obtaining a module with k genes. We set the maximum number of modules for `DEGAS` to 5, yet it returned a single module. In the following, `DEGAS`, `ICA`, and Mica output modules are referred to as `degas`, `ica`, and `mica`, respectively.

We used two datasets for two breast-cancer subtypes: invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC). Both datasets are from TCGA (`https://tcga-data.nci.nih.gov/tcga/`) and contain RNA-Seq and miRNA-Seq data. We used two different subtypes to understand how different techniques are able to detect modules specific to each subtype.

The ILC dataset has 106 control samples and 153 case samples. All of the 259 samples have gene expression information. Out of the 153 cases, only 150 contain miRNAs expression data as well. Therefore, only the 150 cases are used in our experiments. The IDC dataset shares the 106 control samples with the ILC. It also has 714 case samples with gene expression information, however, only 699 case samples having miRNA expression information are used.

The PPI network used for the module extraction was obtained from the BioGRID (`http://thebiogrid.org`) database (rel. 3.2.104). It contains 139,539 unique interactions between 18,170 proteins. The experimentally validated miRNA-target interactions are obtained from miRTarBase (rel. 4.5) [22]. The number of runs κ for ICA is set to 100 while t_R is set to 4, and t_C and t_G are set to 2 to keep only the potentially important values.

The module qualities are verified using pathway, GO, and disease ontology (DO) enrichment analyses and by also looking for evidences in the literature. Enrichment analyses are performed using ReactomePA [56], FunDo [41], and clusterProfiler [57]. A complete set of results and information about the components are available at `http://bmi.osu.edu/hpc/software/mica`. Here, we only show the highest scoring components and the most significant results.

4.1 Results on ILC data

Both Mica and `ICA` generated seven modules. However, the Mica modules are meaningfully different from `ICA` ones. Table 1 shows the number of covered samples, component size, the number of member genes in the PPI network, the size, and the score of the largest and top scored modules. In general, there is a large connected module in the PPI network. Mica modules have higher scores than `ICA` modules in addition to being more common.

We also used `DEGAS` on the ILC dataset for comparison purposes. The `degas` module consists of 347 genes with 730

Table 1: Size of the modules obtained using Mica and ICA. # is the component number, S is the number of samples a component covers, $|c|$ is the size of the component, $|c|_{ppi}$ is the number of genes that are both in the component and the PPI network, N, E, and $scr(c)$ are the number of nodes, number of edges, and scores, respectively, for the largest connected module in the PPI.

(a) ICA

#	S	$\|c\|$	$\|c\|_{ppi}$	N	E	$scr(c)$
1	55	754	657	221	348	39.43
2	18	34	31	2	1	3.35
3	54	279	267	103	143	25.33
4	28	703	641	274	510	50.70
5	4	542	448	116	141	28.80
6	7	349	320	116	337	26.68
7	2	204	176	30	29	12.81

(b) Mica

#	S	$\|c\|$	$\|c\|_{ppi}$	N	E	$scr(c)$
1	103	501	475	164	272	55.63
2	49	284	242	21	21	12.71
3	67	1007	879	339	585	49.51
4	30	455	446	283	506	52.41
5	68	931	876	541	1535	66.91
6	9	889	752	253	354	46.04
7	3	790	738	410	1297	51.04

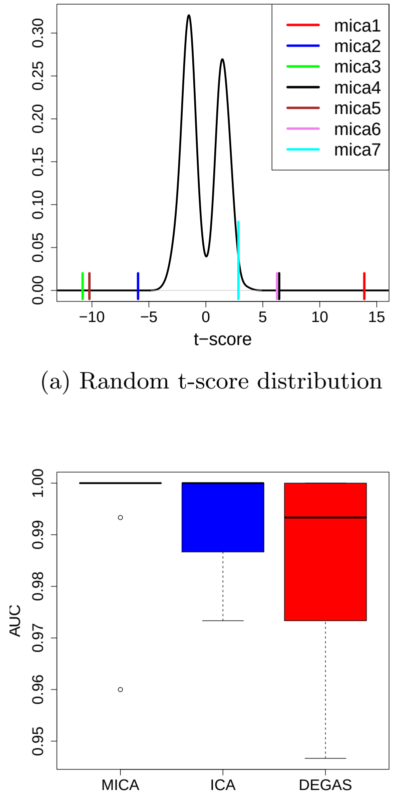
interactions and 200 DE genes. The quality, i.e., the module size p-value, is 0.19 which can be considered large. We tried different options for DEGAS to get a better module, however, this is the best we could get.

4.2 Statistical analysis of the components

The first step is to ensure that the Mica components, hence the active modules, are not random. Therefore, our null hypothesis is that the t-score computed for each component from its weight across the case and control samples in $\mathbf{A}$ can be random. We generated $1,000$ matrices by randomly permuting the modified gene expression values across the case and control samples. Afterwards, we applied Mica on these matrices and computed the t-score for the random components. For each $1,000$ runs, we only kept the max/min t-score value. Finally, we generated the random t-score distribution and compared our actual t-scores against it. The random t-score distribution and the components' t-scores are shown in Figure 2(a). Clearly, the components cannot randomly gain high t-scores (i.e., p-value $= 0$): the null hypothesis is rejected.

4.3 Classification using the modified and original gene expression data

To show that the modified gene expression data can better differentiate between case and control, we compared the predication accuracy of Mica modules on the modified data, and ICA and DEGAS modules on the original data. For Mica and ICA, a Support Vector Machine (SVM) is trained on each module separately, where the module genes are used as the input features. Afterwards, a uniform voting is performed between the modules to understand how the modules collectively affect the classification performance (since DEGAS has one module, no voting is required). A 10-fold cross validation is performed to better understand the prediction performance of the modules. The same 10-subsets were used for Mica, ICA, and DEGAS. The results are shown in Figure 2(b). In general, Mica and ICA obtain a better classification accuracy than DEGAS with Mica being more stable across different runs and almost obtaining an AUC value of 1.



(a) Random t-score distribution

(b) Prediction performance

Figure 2: Performance evaluation of Mica modules. a) Mica modules' t-scores in comparison to t-scores from random runs. b) Mica, ICA, and DEGAS prediction performance.

4.4 Active modules analysis

Here we analyze the genes in each module, the overlap between the modules, and the top enriched GO terms. Surprisingly, there is no large overlap among Mica, ICA, and DEGAS; degas overlaps with 12% of mica5, and ica4 overlaps with 17% of mica6. However, there are similarities in the top GO annotations (i.e., corrected p-value $< 10^{-15}$). Among them: *translational elongation* in ica6 and mica7, *positive regulation of biological process* in ica4 and mica6, *cellular macromolecule metabolic process* in mica1 and degas, and *organelle organization* in mica4 and degas. The top different ones include *protein transport* in ica1, *cardiovascular system development* and *extra cellular matrix organization* in ica5, *response to endoplasmic reticulum stress* in mica2, *RNA processing* in mica3, and *cell cycle* and *cell cycle process* in mica5.

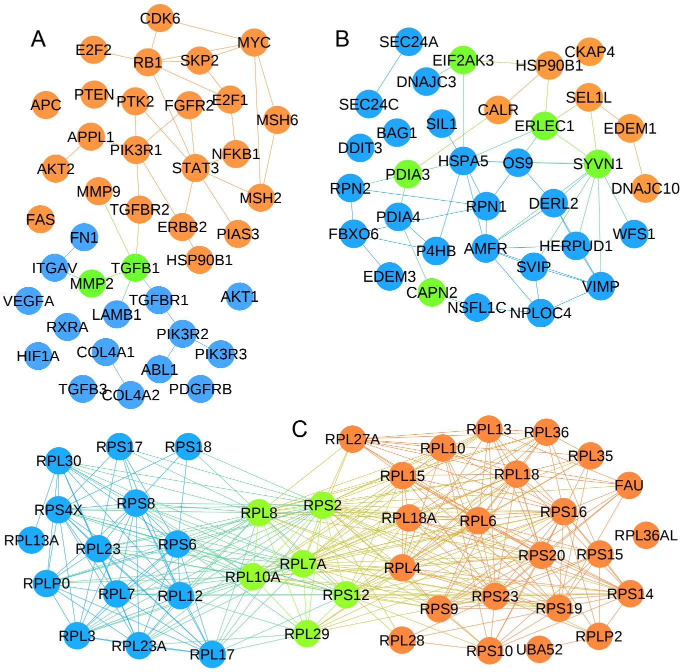


Figure 3: Overlap between important pathways enriched in both Mica and ICA modules. Orange is for Mica, blue is for ICA, and green for genes in both. A) Pathways in cancer (mica1 and ica5), B) Protein processing in ER (mica2 and ica1), C) Ribosome (mica7 and ica6).

We performed pathway enrichment to see how the active modules are enriched with important pathways. The results are shown in Table 2. Similar to GO annotations, common pathways among Mica, ICA, and DEGAS exist: both degas and mica5 are enriched with the *cell cycle* pathway, however, the p-value for degas is much smaller than the p-value in mica5. Remarkably, mica5 is enriched with more cell cycle-related pathways, e.g., the cell cycle, mitotic, and check points pathways, with BRCA1 common among most of them. BRCA1 mutations lead to genetic instability and deficiency in different cell cycle phases [13]. Additionally, its absence results in breast cancer formation.

Pathways that are highly enriched in both Mica and ICA modules include the *pathways in cancer*, *ribosome*, and *protein processing in endoplasmic reticulum* pathways. Figure 3 shows the overlap between Mica and ICA on those pathways. The *pathways in cancer* pathway is enriched in both mica1 and ica5. Remarkably, mica1 contains key breast cancer genes including ERBB2, MYC, RB1, and NFKB1. Additionally, mica1 is more common across the samples than ica5. ERBB2 is a growth factor receptor over-expressed in breast cancer and related to tumor aggressiveness and resistance to chemotherapy [43]. RB1 is mutated in breast cancer [19] while NFKB1 has a major rule in invasive breast cancer [33]. MYC is a multifunctional protein that plays a role in cell cycle progression and cellular transformation. MYC amplification is found to be frequent in breast cancer that is often more associated with the metastatic tumor version [47]. The *protein processing in endoplasmic reticulum (ER)* pathway is another interesting one that is enriched in both mica2 and ica1. The ER is an essential organelle involved in many important functions such as protein folding and secretion. In cancer cells, the *unfolded protein response (UPR)* and *ER-associated degradation (ERAD)* pathways, which are parts of the protein processing in ER pathway, are both activated to help in the survival and the metastasis of the cancer cells [51]. Surprisingly, EDEM1 and SEL1L in mica2 are important parts of the ERAD component in addition to being deregulated in cancer cells [51].

Table 3: The components obtained by ICA and Mica. # is the component number, S is the number of samples a component covers, $|c|$ is the size of the component, $|c|_{ppi}$ is the number of genes that are both in the component and the PPI network, N and E are the number of nodes and edges, respectively, for the largest connected module in the PPI, and $scr(c)$ is the score of the largest connected module.

(a) ICA

#	S	$\|c\|$	$\|c\|_{ppi}$	N	E	$scr(c)$
18	87	897	849	391	775	61.95
21	123	744	669	303	522	61.43
30	513	675	649	454	1851	83.63

(b) Mica

#	S	$\|c\|$	$\|c\|_{ppi}$	N	E	$scr(c)$
7	296	400	374	147	234	61.78
15	336	317	267	42	47	59.76
33	245	289	280	138	297	79.69
42	544	682	633	348	1063	66.97
63	242	243	230	101	188	66.43

Since mica1, mica2, ica1, and ica5 contain interesting pathways, we performed DO enrichment analysis on them using FunDO [41]. The top enriched diseases, after Bonferroni correction, are: cancer (2.11×10^{-21}) and breast cancer (1.11×10^{-4}) in mica1, cancer (1.15×10^{-3}) in mica2, cancer (2.34×10^{-12}) in ica5, and cancer (6.2×10^{-5}) and Melanoma (1.1×10^{-4}) in ica1. Clearly, mica1 is the most related module to cancer in general and breast cancer, in specific.

4.5 Results on IDC data

Invasive ductal carcinoma is a famous breast cancer subtype. In the literature, IDC and ILC act differently and have different sets of DE genes [60, 55]. Yet we expect to find common pathways between them, even though each pathway might include different sets of genes [53].

Similar to ILC, we used the dataset with ICA and Mica to see how different the output is when the miRNA data is added. As shown in Table 3, there is a significant difference between ICA and Mica modules. In addition, Mica produced 66 modules while ICA produced 35 modules (only the important modules are shown in the table). We analyzed the highest scoring modules for each method, i.e., with a score ≥ 60, namely, ica18, ica21, and ica30 from ICA and mica7, mica15, mica33, mica42, and mica63 from Mica. By comparing the ICA and Mica modules, we found that the most similar ones are mica42 and ica30; with 266 common genes existing in both. The remaining Mica and ICA modules do not have any large overlap.

A further examination of mica42 and ica30 shows that both contain BRCA1, BRCA2, BRIP1, BLM, RAD51, UBE2C,

Table 2: Pathway enrichment analysis for Mica, ICA, and DEGAS modules on the ILC dataset.

Database	Pathway	MICA			ICA			DEGAS	
		%	pval	Net	%	pval	Net	%	pval
Reactome	Cell Cycle, Mitotic	18.48	1.19×10^{-21}	mica5				11.53	7.79×10^{-3}
	Cell Cycle	19.96	7.30×10^{-19}	mica5				14.12	7.32×10^{-3}
	Mitotic M-M/G1 phases	13.31	3.75×10^{-18}	mica5					
	Extracellular matrix organization				21.55	5.25×10^{-15}	ica5		
	Molecules associated with elastic fibres				9.48	3.27×10^{-09}	ica5		
	Integrin cell surface interactions				11.21	2.02×10^{-07}	ica5		
	Translation				24.13	8.66×10^{-14}	ica5		
	GTP hydrolysis and joining of the 60S ribosomal subunit				21.55	2.74×10^{-13}	ica6		
	Peptide chain elongation				18.10	9.89×10^{-11}	ica6		
	Nonsense Mediated Decay Independent of the Exon Junction Complex				18.10	1.71×10^{-10}	ica6		
KEGG	Pathways in cancer	15.24	1.05×10^{-04}	mica1	14.66	2.59×10^{-03}	ica5		
	Protein processing in endoplasmic reticulum	52.38	4.65×10^{-11}	mica2	12.22	1.10×10^{-08}	ica1		
	Osteoclast differentiation	8.70	1.85×10^{-06}	mica6					
	Complement and coagulation cascades	4.74	1.62×10^{-03}	mica6					
	Ribosome	7.07	1.76×10^{-10}	mica7	17.24	3.34×10^{-14}	ica6		
	ECM-receptor interaction				11.21	3.83×10^{-07}	ica6		

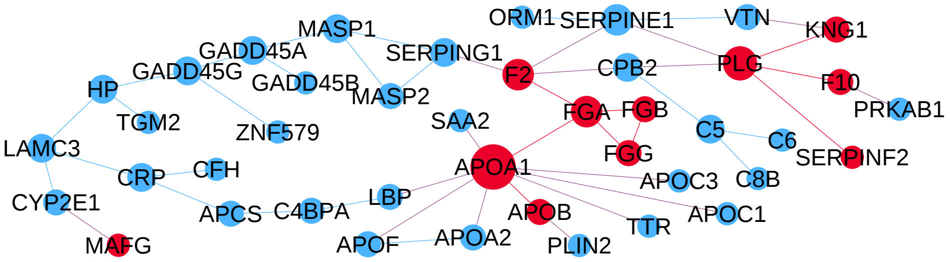


Figure 4: mica15 module. The red nodes are in the Hemostasis pathway.

and CKS2. BLM and RAD51 genes have a tumorigenic significance [14], UBE2C and CKS2 are DE genes in IDC [39], and BRCA1, BRIP1, and BRCA2 are known breast cancer mutated genes (http://cancer.sanger.ac.uk/cancergenome/projects/census/). On the other hand, mica42 only contains TOP3A, HMG20B, RAD51C, CDC6, and U2AF1. HMG20B interacts directly with BRCA2. The inhibition of the interaction between HMG20B and BRCA2 leads to tumor progression [32]. TOP3A and BLM interact with RMI1 forming a complex that is very important in genome stability [7]. The mutations in this complex increase the breast cancer risk [5]. RAD51C is also found to be mutated in breast cancer [34]. The de-regulation of CDC6 poses a serious risk of carcinogenesis [36] while U2AF1 is a splicing factor protein that is mutated in cancer [17].

The degas module on IDC data contains 386 genes with 1,056 interactions and 190 DE genes. Based on the quality, the module has a p-value of 0, i.e., it cannot be randomly generated. There are 105 common genes exist in degas, ica30, and mica42 including BRIP1, RAD51, BLM, UBE2C, and CKS2. However, degas did not contain other cancer related genes including BRCA1, BRCA2, XRCC1, XRCC2, and RRM2. Additionally, none of the genes that exclusively exist in mica42 also exist in degas.

In addition, we performed classification analysis on the modules and datasets to ensure that the adjusted gene expression data better correlate with the disease behavior. Similar to the ILC dataset, a SVM is trained on the top scoring modules obtained from each tool separately. Then a 10-fold cross validation is performed using the original data for ICA and DEGAS and modified gene expression data for Mica. The three tools had a similar performance with Mica having the least error of 0.0013. The error for ICA and DEGAS was 0.0038 and 0.0063, respectively.

To better evaluate ICA, DEGAS, and Mica modules, we performed pathway enrichment analysis as shown in Table 4. There are many common pathways among mica42, mica30, and degas such as *cell cycle*, *Tolemere maintenance*, and *DNA strand elongation*. However, mica42 alone is enriched with the *p53 signaling* pathway. Interestingly, there are many important pathways enriched in mica15 which are not enriched in any other modules, including the *complement and coagulation cascades*, *platelet degranulation*, and *Hemostasis* pathways. All of these pathways are part of the cell's hemostatic system which is important in facilitating the metastatic potential of breast cancer [30]. Additionally, a proteomic-based study has shown the complement and coagulation pathway to be DE in IDC [49]. Figure 4 shows mica15 genes. The APOA1 gene in mica15 is found DE in IDC samples versus control samples in a proteomic study [42]. In addition, mutations in this gene lead to poor outcome for post-surgery breast cancer patients [21]. Other interesting genes in mica15 are GADD45A, GADD45B, and GADD45G, which are found down-regulated in cancer [11]. They are stress sensor genes activated in response to cell stress and DNA damage. Interestingly, they are considered as potential therapeutic targets in cancer [11].

The results of the DO enrichment analysis are shown in Table 5. In general, Mica and ICA modules are significantly enriched with cancer and breast cancer than DEGAS, with

Table 4: Pathway enrichment analysis for ICA, DEGAS, and Mica.

Database	Pathway	MICA			ICA			DEGAS	
		%	pval	Name	%	pval	Name	%	pval
KEGG	Complement and coagulation cascades	42.86	1.17×10^{-23}	mica15					
	DNA replication	6.32	6.68×10^{-17}	mica42	5.51	1.13×10^{-18}	ica30		
	Mismatch repair	3.16	5.53×10^{-07}	mica42	3.30	1.11×10^{-10}	ica30		
	Homologous recombination	2.59	3.57×10^{-04}	mica42	2.64	6.97×10^{-06}	ica30		
	p53 signaling pathway	3.45	7.86×10^{-03}	mica42					
	Spliceosome				6.60	8.20×10^{-04}	ica21		
Reactome	Platelet degranulation	21.43	6.66×10^{-08}	mica15					
	Platelet activation, signaling and aggregation	23.81	9.16×10^{-06}	mica15					
	Hemostasis	30.95	6.80×10^{-05}	mica15					
	mRNA Processing	10.14	1.52×10^{-04}	mica33	6.93	2.45×10^{-04}	ica21		
	Cell Cycle, Mitotic	32.76	3.86×10^{-52}	mica42	31.28	4.26×10^{-64}	ica30	17.62	4.74×10^{-13}
	Resolution of Sister Chromatid Cohesion	12.07	1.57×10^{-22}	mica42	11.45	9.46×10^{-28}	ica30	6.74	2.29×10^{-07}
	Leading Strand Synthesis	3.45	6.05×10^{-13}	mica42	2.64	1.40×10^{-11}	ica30		
	Polymerase switching	3.45	6.05×10^{-13}	mica42	2.64	1.40×10^{-11}	ica30		
	DNA Repair	8.62	3.75×10^{-12}	mica42	8.15	2.16×10^{-14}	ica30		
	DNA Replication Pre-Initiation	6.90	2.29×10^{-11}	mica42	5.51	8.67×10^{-10}	ica30	4.4	7.74×10^{-05}
	M/G1 Transition	6.90	2.29×10^{-11}	mica42	5.51	8.67×10^{-10}	ica30	4.40	7.74×10^{-05}
	Telomere Maintenance	5.17	2.68×10^{-07}	mica42	4.41	4.87×10^{-07}	ica30	3.63	1.01×10^{-03}
	Post-transcriptional Silencing By Small RNAs				1.79	1.49×10^{-06}	ica18		
	Pre-NOTCH Transcription and Translation				2.05	1.77×10^{-05}	ica18		
	p53-Independent G1/S DNA damage checkpoint							2.59	8.80×10^{-03}

Table 5: DO enrichment analysis for ICA, DEGAS, and Mica.

name	DO	Corrected p-value
mica7	cancer	$5.38 \times 10-7$
mica15	liver cancer, systematic infection,	4.67×10^{-9}, 1.16×10^{-8},
	metastatic to brain	6.66×10^{-8}
mica33	cancer	5.2×10^{-5}
mica42	cancer, breast cancer	6.21×10^{-35}, 5.72×10^{-7}
mica63	cancer	2.30×10^{-4}
ica18	breast cancer, cancer	4.59×10^{-6}, 6.21×10^{-35}
ica21	cancer	1.36×10^{-5}
ica30	cancer, breast cancer	2.78×10^{-33},1.96×10^{-6}
degas	cancer, breast cancer	1.78×10^{-14}, 3.14×10^{-4}

Mica better enriched with them than ICA. Additionally, mica15 is enriched with metastatic to brain disease.

5. CONCLUSION

In this work, we proposed a new workflow, Mica, that successfully integrates miRNA data, mRNA data, and PPI network in a novel way to obtain active modules which can serve as powerful biomarkers. Experimental results show that Mica modules are more disease-related while unraveling new gene-gene dependencies which are hidden via existing techniques. Albeit the simplicity of the proposed workflow, Mica successfully includes many novel ideas, including the adjustment of the gene expression levels with the miRNA expression to mimic the protein expression levels, and the integration of non-network genes to include possible missing dependencies. To the best of our knowledge, this is the first study that integrates miRNA, mRNA, and PPI network for active module discovery. Furthermore, Mica provides information regarding which modules are active in which set of samples, hence, making it easier to understand the disease behavior for different patients.

Mica modules obtained from the IDC and ILC datasets are different, suggesting that Mica can be further used to generate disease specific modules and hence distinguish between the different diseases. Still, there are some pathways common between IDC and ILC, such as the cell cycle pathway with BRCA1 and BRCA2 retrieved with Mica in both datasets.

Further improvements for Mica can add more value and more understanding for the results. For instance, it may be more beneficial to extract smaller sub-modules of 10–20 genes that can be further used as an effective biomarker. Additionally, each module can be broken into smaller ones and each can be considered as a possible pathway. Hence, we can further understand how the different pathways interact together. A possible method to extract such smaller submodules would be by extracting the densest subgraph in each module. Pathway extraction can also benefit from adding directionality information to the PPI network.

In addition to the mentioned improvements, Mica currently depends on experimentally validated interactions which could generate biases in the results. However, with the advancement in the high throughput sequencing technology and with the increase in the reported protein expression data (e.g., CPTAC https://cptac-data-portal.georgetown.edu/cptacPublic/), Mica can be further improved to include non-experimentally validated microRNA-target genes in addition to building more accuate models representing the relation between the proteins and miRNAs [48]. As usual, the most important step is to carry wet lab experiments to validate the obtained results and to further transfer our study from being a theoretical work to actually be a part of the cancer treatment process.

6. REFERENCES

[1] D. Baek, J. Villén, C. Shin, et al. The impact of micrornas on protein output. *Nature*, 455(7209):64–71, 2008.

[2] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[3] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[4] C. Blenkiron, L. D. Goldstein, N. P. Thorne, et al. Microrna expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol*, 8(10):R214, 2007.

[5] K. Broberg, E. Huynh, K. S. Engström, et al. Association between polymorphisms in rmi1, top3a, and blm and risk of cancer, a case-control study. *BMC cancer*, 9(1):140, 2009.

[6] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. In *IEE Proc. F (Radar and Signal Processing)*, volume 140, pages 362–370, 1993.

[7] K.-L. Chan, P. S. North, and I. D. Hickson. Blm is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. *The EMBO journal*, 26(14):3397–3409, 2007.

[8] X. Chang, T. Xu, Y. Li, and K. Wang. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of date and party hubs. *Scientific reports*, 3, 2013.

[9] S. Chavali, S. Bruhn, K. Tiemann, et al. MicroRNAs act complementarily to regulate disease-related mRNA modules in human diseases. *RNA*, 19(11):1552–1562, 2013.

[10] P. Chiappetta, M.-C. Roubaud, and B. Torrésani. Blind source separation and the analysis of microarray data. *Journal of Comp Biol*, 11(6):1090–1109, 2004.

[11] A. Cretu, X. Sha, J. Tront, et al. Stress sensor gadd45 genes as therapeutic targets in cancer. *Cancer therapy*, 7(A):268, 2009.

[12] Y. Cun and H. Fröhlich. Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS one*, 8(9):e73074, 2013.

[13] C.-X. Deng. Brca1: cell cycle checkpoint, genetic instability, dna damage response and cancer evolution. *Nucleic acids research*, 34(5):1416–1426, 2006.

[14] S.-l. Ding, J.-C. Yu, S.-T. Chen, et al. Genetic variants of blm interact with rad51 to increase breast cancer susceptibility. *Carcinogenesis*, 30(1):43–49, 2009.

[15] M. T. Dittrich, G. W. Klau, A. Rosenwald, et al. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinf.*, 24(13):i223–i231, 2008.

[16] M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA translation and stability by microRNAs. *Ann. review of bioch.*, 79:351–379, 2010.

[17] A. R. Grosso, S. Martins, and M. Carmo-Fonseca. The emerging role of splicing factors in cancer. *EMBO reports*, 9(11):1087–1093, 2008.

[18] J. Himberg, A. Hyvärinen, and F. Esposito. Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, 22(3):1214–1222, 2004.

[19] A. Hollestelle, J. H. Nagel, M. Smid, et al. Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Br. Can. Res. and Treat.*, 121(1):53–64, 2010.

[20] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[21] M.-C. Hsu, K.-T. Lee, W.-C. Hsiao, et al. The dyslipidemia-associated snp on the apoa1/c3/a5 gene cluster predicts post-surgery poor outcome in taiwanese breast cancer patients: a 10-year follow-up study. *BMC cancer*, 13(1):330, 2013.

[22] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucl. Acids Res.*, 39(suppl 1):D163–D169, 2011.

[23] G. T. Huang, C. Athanassiou, and P. V. Benos. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucl. Acids Res.*, 39(suppl 2):W416–W423, 2011.

[24] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

[25] A. Hyvärinen. Independent component analysis: recent advances. *Philos. Trans. of the Royal Soc. A: Math., Phys. and Eng. Sci.*, 371(1984), 2013.

[26] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinf.*, 18(Suppl 1):S233–S240, 2002.

[27] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res.*, 18(4):644–652, 2008.

[28] M. V. Iorio and C. M. Croce. microRNA involvement in human cancer. *Carcinogenesis*, 33(6):1126–1133, 2012.

[29] M. Koyutürk. Algorithmic and analytical methods in network biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3):277–292, 2010.

[30] I. Lal, K. Dittus, and C. E. Holmes. Platelets, coagulation and fibrinolysis in breast cancer progression. *Breast Cancer Research*, 15(4):1–11, 2013.

[31] H.-S. Le and Z. Bar-Joseph. Integrating sequence, expression and interaction data to determine condition-specific mirna regulation. *Bioinformatics*, 29(13):i89–i97, 2013.

[32] M. Lee, M. Daniels, M. Garnett, and A. Venkitaraman. A mitotic function for the high-mobility group protein HMG20b regulated by its interaction with the brc repeats of the brca2 tumor suppressor. *Oncogene*, 30(30):3360–3369, 2011.

[33] F. Lerebours, S. Vacher, C. Andrieu, et al. Nf-kappa b genes have a major role in inflammatory breast cancer. *BMC cancer*, 8(1):41, 2008.

[34] E. Levy-Lahad. Fanconi anemia and breast cancer susceptibility meet again. *Nature genetics*, 42(5), 2010.

[35] L. Li, H.-Z. Chen, F.-F. Chen, et al. Global microrna expression profiling reveals differential expression of target genes in 6-hydroxydopamine injured mn9d cells. *Neuromolecular medicine*, 15(3):593–604, 2013.

[36] P. Li, Y. Lin, Y. Zhang, et al. SSX2IP promotes metastasis and chemotherapeutic resistance of hepatocellular carcinoma. *Jr. of Trans. Med.*, 2013.

[37] Y.-O. Li, T. Adalı, and V. D. Calhoun. Estimating the number of independent components for functional magnetic resonance imaging data. *Human brain mapping*, 28(11):1251–1266, 2007.

[38] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[39] X.-J. Ma, R. Salunga, J. T. Tuggle, et al. Gene

expression profiles of human breast cancer progression. *Proc. of the Nat. Acad. of Sci.*, 100(10):5974–5979, 2003.

[40] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Gen.*, 14(10):719–732, 2013.

[41] J. D. Osborne, J. Flatow, M. Holko, et al. Annotating the human genome with disease ontology. *BMC genomics*, 10(Suppl 1):S6, 2009.

[42] I. Pucci-Minafra, P. Cancemi, M. R. Marabeti, et al. Proteomic profiling of 13 paired ductal infiltrating breast carcinomas and non-tumoral adjacent counterparts. *PROT.-Clin. App.*, 1(1):118–129, 2007.

[43] F. Revillion, J. Bonneterre, and J. Peyrat. ERBB2 oncogene in human breast cancer and its clinical significance. *Euro. Jr. of Cancer*, 34(6):791–808, 1998.

[44] M. Rotival, T. Zeller, P. S. Wild, et al. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS gen.*, 7(12):e1002367, 2011.

[45] R. Schachtner, D. Lutter, P. Knollmüller, et al. Knowledge-based gene expression classification via matrix factorization. *Bioinf.*, 24(15):1688–1697, 2008.

[46] M. Scholz, S. Gatzek, A. Sterling, et al. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinf.*, 20(15):2447–2454, 2004.

[47] A. D. Singhi, A. Cimino-Mathews, R. B. Jenkins, et al. Myc gene amplification is often acquired in lethal distant breast cancer metastases of unamplified primary tumors. *Modern Path.*, 25(3):378–387, 2011.

[48] R. J. Slebos, X. Wang, X. Wang, et al. Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Scientific data*, 2, 2015.

[49] M.-N. Song, P.-G. Moon, J.-E. Lee, et al. Proteomic analysis of breast cancer tissues to identify biomarker candidates by gel-assisted digestion and label-free quantification methods using LC-MS/MS. *Arch. of Pharm. Res.*, 35(10):1839–1847, 2012.

[50] A. E. Teschendorff, M. Journée, P. A. Absil, et al. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comp. Biol.*, 3(8):e161, 2007.

[51] Y. C. Tsai and A. M. Weissman. The unfolded protein response, degradation from the endoplasmic reticulum, and cancer. *Genes & cancer*, 1(7):764–778, 2010.

[52] J. S. Tsang, M. S. Ebert, and A. van Oudenaarden. Genome-wide dissection of microrna functions and cotargeting networks using gene set signatures. *Molecular cell*, 38(1):140–153, 2010.

[53] G. Turashvili, J. Bouchal, K. Baumforth, et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer*, 7(1):55, 2007.

[54] I. Ulitsky, A. Krishnamurthy, R. M. Karp, and R. Shamir. DEGAS: de novo discovery of dysregulated pathways in human diseases. *PLoS One*, 5(10):e13367, 2010.

[55] N. Wasif, M. A. Maggard, C. Y. Ko, et al. Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. *Ann. of Surg. Oncol.*, 17(7):1862–1869, 2010.

[56] G. Yu. *ReactomePA: Reactome Pathway Analysis*, 2014. R package version 1.4.0.

[57] G. Yu, L. Wang, Y. Han, and Q. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Jr. of Int. Biol.*, 16(5):284–287, 2012.

[58] S. Zhang, Q. Li, J. Liu, and X. J. Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):i401–i409, 2011.

[59] S. Zhang, C.-C. Liu, W. Li, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucl. Acids Res.*, 40(19):9379–9391, 2012.

[60] H. Zhao, A. Langerød, Y. Ji, et al. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. of the Cell*, 15(6):2523–2536, 2004.