Disease-specific module discovery is an important problem to understand the disease behavior. A successful method to address this problem is the integration of gene expression data with the protein-protein interaction (PPI) network. Many tools have been developed to efficiently perform this integration. However, these tools focus only on the genes existing in the PPI network; totally neglecting other genes that we do not yet have information regarding their interaction. In addition, they only make use of the gene expression data which does not give the true picture about the actual protein expression levels. In fact, the cell uses different mechanisms, such as microRNAs, to post-transcriptionally regulate the proteins without affecting the corresponding genes expressions. The unprecedented amount of publicly available disease-related data encourages the development of new methodologies for a further understanding the disease behavior.

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 network information to successfully return disease-specific gene modules. The novelty of the workflow lies in many directions, including the adjustment of miRNA expression with microRNA to better highlight indirect dependencies between the different 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 returned MICA gene modules unravel new and interesting dependencies between the different genes and miRNAs.

1. INTRODUCTION

In complex diseases, genes do not act in isolation, rather, they interact together in pathways and modules to perform the designated function [11]. In addition, their interaction patterns are changed based on the type of the cell and the condition [8]. A well-structured characterization and analysis of such modules have always been intriguing for the researchers, especially for extremely heterogeneous diseases. Cancer is such a disease: the derivative tissue differs for many cancer types. Besides, each cancer type can have many subtypes. Identifying a biologically correct and valid module is important for each cancer type and subtype since the treatment options and their success rates can significantly differ [2].

One way to find such modules is to look for clusters of genes with certain properties, e.g., dense cluster, in different biological networks, such as the PPI network or the gene co-expression network. A more efficient method is the integration of different biological data to better highlight these gene modules [40]. Following this idea, various techniques that integrate gene-expression values or p-values with biological networks to extract such gene modules have been proposed, e.g., [29, 16, 53]. Such extracted modules are called active modules since the gene expression data, which is dynamically changing, is integrated with the PPI network, which is static. Hence, the word active comes from the notion that these modules are active in certain cells or conditions. Following this track, many other tools have been developed to better make use of the network structure and other types of data as well, such as genotypic data. An excellent review and categorization of these tools was recently provided [40].

Although the gene expression signature-based tools and algorithms have proven to be flexible in practice, they do not provide a be-all and end-all solution for the active modules discovery problem. Today, we have various data types that can be used to increase the accuracy, but many of the existing tools and workflows do not exploit such heterogeneity. Besides, these tools are usually restricted to the proteins/genes in the networks they use and ignore the other genes in the gene expression data that we do not yet have any information regarding their interaction patterns.

MicroRNAs (miRNAs) are small non-coding RNAs that are used by the cell to post-transcriptionally regulate gene expression levels [18]. miRNAs 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 [30]. 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]. Hence, using miRNAs for the active module discovery is a promising technique to increase the accuracy and success rate of the cancer treatments.

Most of the works that integrate miRNA and mRNA data assumes that the miRNA effect on the mRNA is distinguishable from the gene expression levels [26, 58]. However, the protein expression level can be significantly affected by the miRNA without having any apparent effect on the gene ex-
pression level [1]. [13] suggested another method to integrate miRNA and mRNA by integrating the PPI network and miRNA-target gene network into one heterogeneous network. They focused on prioritizing the genes using the suggested network. Indeed, such integration would work around the miRNA-mRNA integration problem. However, by focusing only in prioritizing genes through the PPI network, they cannot detect connected modules of genes with indirect dependencies, e.g., through other genes not in the PPI network or through other genes with no change in expression at mRNA level.

Even though the techniques using gene expression levels provide valuable information, they cannot show the whole picture. Here, we try to exploit another miRNA and mRNA interaction pattern, which is the inhibition of protein translation rather than mRNA degradation. We believe that if the gene expression levels are adjusted based on the expression levels of the corresponding miRNAs, novel and interesting gene-gene dependencies can be unraveled.

In this work, we propose a workflow Mica which employs heterogeneous data sources and adopts independent component analysis [28] to extract active modules. To unravel new types of gene-gene dependencies, we provide a novel data integration technique that adjusts the expression level of the genes based on the expression level of the corresponding miRNA. These dependencies are then mapped back to the PPI network to extract the connected modules. Compared to existing active module discovery tools, Mica is less dependent on the given biological network it uses hence does not need to ignore the information for the entities which are not in the network.

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 [51]. Here, we will focus on the complementary and the additive effects.

The rest of the paper is organized as follows: In Section 2, we provide a background on the techniques we used in this work. Our methods and experimental results are presented in Section 3 and Section 4, respectively. Section 5 concludes the paper.

2. BACKGROUND

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 tries to distinguish the sources by minimizing the statistical dependencies between them [28]. In the context of gene expression, ICA decomposes an input expression into its possible expression modes [38]. For an \( n \times m \) input gene expression matrix \( \mathbf{X} \), where rows correspond to genes and columns correspond to samples, ICA decomposes \( \mathbf{X} \) into:

\[
\mathbf{X}^T = \mathbf{A} \times \mathbf{S}
\]

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 independent components. The columns of \( \mathbf{S} \) correspond to the genes and the entry \( \mathbf{S}_{g,s} \) shows the contribution of a gene \( g \) 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 contribution of each component \( c \) for a sample \( s \). Many approximation algorithms have been proposed to find \( \mathbf{A} \) and \( \mathbf{S} \) in an efficient way, e.g., fastICA [27], JADE [6], and InfoMax [3]. fastICA tries to identify non-Gaussian components under the assumption that Gaussian components represent the noise. This algorithm can stick in a local minima, hence multiple iterations, thus multiple estimates can be necessary [21, 10].

ICA has been used extensively to cluster different genes together or for sample classification [38, 33, 19, 49, 45, 17, 44, 54]. All of these studies have shown the efficiency of ICA in producing biologically relevant results.

3. METHODS

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

3.1 Data integration

The miRNA and gene expression data are usually integrated by using correlation-based methods with the assumption that the effect of miRNA on mRNA should be apparent on the gene expression level. Rather than the suppression of the gene expression, one can also exploit another type of miRNA effect on mRNA; the inhibition of the protein translation. Traditional approaches cannot exploit such an effect since it will not be apparent on the gene expression level. Our novel integration step is based on this fact. We use miRNA expression level to adjust the expression level of the genes. Therefore, if a gene is affected by an miRNA at the inhibition level, the proposed integration makes the effect visible on the expression level. For each sample \( s \), we first calculate the ratio:

\[
\beta_{g,s} = \frac{\sum_{\{r: r \text{ affects } g\}} Z_{r,s}}{\sum_{\{r: r \text{ affects } g\}} Z_{r,s}^+}
\]

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.
where \( Z^c_{g,s} \) (\( Z^-_{c,s} \)) is the positive (negative) z-score of miRNA \( r \) in sample \( s \) that is experimentally verified to affect gene \( g \). The z-score is calculated by

\[
Z_{r,s} = \frac{x_{r,s} - \mu_r}{\sigma_r}
\]

where \( x_{r,s} \) is the expression level of miRNA \( r \) in sample \( s \), and \( \mu_r \) and \( \sigma_r \) are the mean and standard deviation of \( r \)’s expression level across all the control samples. The z-score is divided into positive and negative groups since each group differently affect gene \( g \). In general, when a miRNA \( r \) is down-regulated, i.e., -ve z-score, then the expression of \( g \) will increase. On the other hand, when \( r \) is up-regulated, i.e., +ve z-score, then the expression of \( g \) will decrease. Accordingly, the final gene expression is calculated as follows:

\[
e'_{g,s} = \beta_{g,s} \times e_{g,s}
\]

where \( e_{g,s} \) and \( e'_{g,s} \) are the original and adjusted expression levels of gene \( g \).

For data integration, (4) is applied to each gene-sample pair. Only the absolute significant z-scores, i.e., the ones greater than a threshold \( t_R \), are taken into account. 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 \( e_{g,s} \). Such a constraint is meant to make sure that either the up-regulated group of miRNAs or the down-regulated group of miRNAs has a larger effect on \( g \).

As mentioned previously, a group of miRNAs can affect the same gene in a synergetic, complementary, or additive way. Our integration equation (4) is additive and partially complementary, i.e., the gene expression level will be affected more if several miRNAs affect it on a sample (additive). When only a single miRNA is active in the sample, it will still affect the expression level (complementary). At the end, our goal is to better highlight the dependency between different 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 the data integration step, the adjusted gene expression values are then fed to the ICA for which the R version of the fastICA algorithm is used [27]. To avoid local minima and unreliable independent component estimates, we follow the method in [10]: we run fastICA \( k \) times and obtain different independent component estimates at each run. Then, the Pearson correlation coefficients between the components from different estimates are computed to distinguish the most similar ones. We constructed a \( k \)-partite similarity graph \( G = (V, E) \) where \( V = V_1 \cup \cdots \cup V_k \) are the set of all components returned by ICA and \( V_i \) is the set of components obtained in the \( i \)th run. The edge set \( E \) contains an edge \((c, c')\) if the Pearson correlation coefficient between \( c \) and \( c' \) is at least 0.9 and they are not obtained in the same run, i.e., \( c \in V_i, c' \in V_j, i \neq j \). To obtain the final component set, we partition \( G \) to its maximally connected subgraphs. Then for each connected subgraph \( C \) of \( G \) with at least \( k \) vertices, we construct a final representative component by computing the average of the \( |C| \) rows corresponding to the vertices in \( C \).

An important parameter of ICA is the number of components \( \ell \) to be generated; when \( \ell \) is large ICA will probably return subcomponent-type structures which are not very interesting [37]. A naïve method is setting \( \ell = m \), the number of samples, which is not useful in our case since we have hundreds of them. We follow another approach [44] based on an earlier method proposed by [23]. We first apply Singular Value Decomposition (SVD) to the actual 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, 45]. Here, we used a variant of the correlation method suggest by [45]. Basically, instead of calculating the correlation between the component weight across the samples and the type (control/case) of the samples, the Wilcoxon signed-rank test is used to calculate a \( p \)-value for each component based on its weight distribution over the controls and cases. The Bonferroni correction method is then used to correct the \( p \)-value. We further compute \( \mu \) and \( \sigma \) for each component by using 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 determine the set of genes related to a component, we use the z-score threshold based method [46, 49] which was shown to be effective to return the most important genes for each component. We calculated the z-score of each gene in a component by using its weight, \( \mu \), and \( \sigma \) that are computed by using all the gene weights inside this component. Then for each component, the genes with a z-score at least \( t_C \) is considered to be a member of the component.

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 \( t_C \) used to pick the member genes for each component is relaxed to allow more connectivity. However, as the results will show, each component yield 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 [58] or highly enriched with signaling pathways [49]. Each component we found after the second step is expected to generate a connected modules. It is crucial to define a scoring function to determine which module is the most important one, i.e., containing important member genes. 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:

\[
scr(c) = \frac{\sum_{z \in c} Z_{z,g}}{\sqrt{|c|}}
\]
where \( |c| \) is the number of member genes in \( c \). We used \( \sqrt{|c|} \) instead of \( |c| \) since we want to give a higher score to larger modules. A gene \( g \) will have a high \( Z_{g} \) value if it is significant for \( c \). Therefore, if a connected module contains many important genes the module is considered to be important.

## 4. RESULTS

We implemented our proposed workflow Mica in R and used the available implementation of the \texttt{fastICA} algorithm. To demonstrate the effectiveness of the proposed workflow, that is, the added benefits of early integration of microRNA datasets, we compared the modules obtained by our workflow Mica against the ones obtained using ICA and DEGAS [53], using the original gene expression values. DEGAS is a set-cover based algorithm known for its efficiency 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 a measure provided by the tool based on how far the size of the module is from a randomly generated subnetwork of \( k \) genes. We set the maximum number of modules for DEGAS to 5. Still, it returned a single module in the experiments. In the rest of the text, DEGAS output modules are referred to as \texttt{degas}, ICA modules as \texttt{ica}, and Mica modules as \texttt{mica}.

We carried out the experiments on two datasets for two breast-cancer subtypes: invasive lobular carcinoma (ILC) and Invasive ductal carcinoma (IDC) datasets. Both datasets are from TCGA (https://tcga-data.nci.nih.gov/tcga/) and they both contain RNA-Seq and miRNA-Seq data. High throughput sequencing data was used in our experiments since it can provide a complete image about all the miRNAs and mRNAs in the cell without requiring any \textit{a-priori} information. The main aim of using two different subtypes of the same disease is 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, which also have miRNA expression information, are used in our experiments.

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 interactions between 18,170 proteins. The experimentally validated miRNA-target interactions used in data integration are obtained from mir-TarBase (rel. 4.5) [25].

The number of runs \( k \) for ICA is set to 100 while \( t_{R} \) threshold is set to 4 and \( t_{C} \) and \( t_{O} \) are set to 2. We set the threshold high since we only want to keep the values that would have a potential of being important.

The qualities of the output modules are verified using different methods, including, pathway enrichment analysis, GO enrichment analysis, disease ontology (DO) enrichment analysis, and finally using the evidence in the literature on the importance of the modules/genes. Enrichment analysis is performed using ReactomePA [56], FunDo [41], and cluster-Profiler [57].

### 4.1 Results on ILC data

The Mica modules are meaningfully different from ICA modules. Table 1 shows the number of samples they cover, the size of each component, the number of member genes in the PPI network, the size of the largest connected module, and the score. In general, for each of ICA and Mica components, there is a large connected module in the PPI network. Interestingly, Mica modules have higher scores than ICA modules in addition to being more common across the samples.

We also use DEGAS on the ILC dataset for comparison purposes. The \texttt{degas} module consists of 347 genes with 730 interactions between them and the number of DE genes in this module is 200. 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 module we obtained.

**Statistical analysis of the obtained components:** An important step is to first ensure that the obtained Mica components, hence the active modules, cannot be obtained from a random matrix. Therefore, we set our null hypothesis to be that the t-score calculated for each component from its weight across the case and control samples is 0. According to the random matrix by randomly permuting the modified gene expression values for each gene across the case and control samples. Afterwards, we applied Mica on the random matrices and calculated the t-score for the randomly generated components. For each 1000 run, we only kept the max/min t-score value. Finally, using the t-scores from the random runs, we generated the distribution for the random t-scores and compared our actual t-scores against them. The random t-score distribution and the components t-score values are shown in Figure 2. Clearly, the components cannot randomly gain such a high t-score (i.e., \( p \)-value = 0). Therefore, the null hypothesis is rejected.

**Classification using modified and original gene expression:** It is important to ensure that the modified gene expression data better differentiate between case and control samples. To this end, a comparison between the prediction accuracy using Mica modules on the modified gene expression data and ICA and DEGAS modules on the original data was carried out. Basically, for Mica modules, a Support Vector Machine (SVM) was trained on each module separately, with the genes in each module used as the input.

<table>
<thead>
<tr>
<th>#</th>
<th>S</th>
<th>c</th>
<th>c_mm</th>
<th>N</th>
<th>E</th>
<th>scr(c)</th>
<th>#</th>
<th>S</th>
<th>c</th>
<th>c_mm</th>
<th>N</th>
<th>E</th>
<th>scr(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>754</td>
<td>657</td>
<td>221</td>
<td>348</td>
<td>39.43</td>
<td>1</td>
<td>103</td>
<td>504</td>
<td>475</td>
<td>164</td>
<td>272</td>
<td>55.63</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>34</td>
<td>31</td>
<td>2</td>
<td>1</td>
<td>3.35</td>
<td>2</td>
<td>49</td>
<td>284</td>
<td>242</td>
<td>21</td>
<td>2</td>
<td>12.71</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>279</td>
<td>267</td>
<td>143</td>
<td>25.33</td>
<td>3</td>
<td>67</td>
<td>1007</td>
<td>879</td>
<td>330</td>
<td>585</td>
<td>49.51</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>703</td>
<td>641</td>
<td>274</td>
<td>510</td>
<td>50.70</td>
<td>4</td>
<td>30</td>
<td>455</td>
<td>446</td>
<td>283</td>
<td>506</td>
<td>52.41</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>542</td>
<td>448</td>
<td>116</td>
<td>141</td>
<td>28.80</td>
<td>5</td>
<td>68</td>
<td>931</td>
<td>876</td>
<td>541</td>
<td>153</td>
<td>66.91</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>349</td>
<td>320</td>
<td>116</td>
<td>337</td>
<td>26.68</td>
<td>6</td>
<td>9</td>
<td>889</td>
<td>752</td>
<td>253</td>
<td>34.6</td>
<td>46.04</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>204</td>
<td>176</td>
<td>30</td>
<td>29</td>
<td>12.81</td>
<td>7</td>
<td>3</td>
<td>790</td>
<td>738</td>
<td>410</td>
<td>1297</td>
<td>51.04</td>
</tr>
</tbody>
</table>
features. Afterwards, a voting was performed between the modules to determine the output classification. The same was applied on ICA but with the original data. For DEGAS, no voting was required since it only has one module. The results for a 10-fold cross validation is shown in Figure 2. In general, Mica and ICA obtain a better classification accuracy than DEGAS, with Mica being more stable across the different runs and obtaining an AUC value of 1 in almost all of the runs.

Active modules analysis: The next step is to see which genes exist in each active module, how the different active modules overlap, and the enrichment of each module with important GO annotations. Interestingly, there was not a large overlap between Mica, ICA, and DEGAS; degas overlaps with 12% of mica5 while ica4 overlaps with 17% of mica6. Nevertheless, there were some similarities in the top enriched GO annotations (i.e., with corrected p-value < 10⁻¹⁵). Among the top similar ones are: translational elongation between ica6 and mica7, and positive regulation of biological process between ica4 and mica6, cellular macromolecule metabolic process in mica1 and degas, and organelle organization between mica4 and degas. On the other hand, the top different ones included protein transport in ica1, cardiovascular system development and extra cellular matrix organization in ica5, response to endoplasmic reticulum stress in mica2, RNA processing and splicing in mica3, and cell cycle and cell cycle process in mica5.

Since we are working with active modules that are going to be further used to extract important pathways, we further performed pathway enrichment analysis to better evaluate the quality of the active modules. The results are shown in Table 2. Similar to GO annotations, some pathways are common between Mica, ICA, and DEGAS. For instance, both degas and mica5 were enriched with the cell cycle pathway, however, the p-value for degas was much smaller than the p-value in mica5. Remarkably, mica5 was enriched with more cell cycle-related pathways, such as, the cell cycle, mitotic, and check points pathways, with BRCA1 common among most of these pathways. Mutations in BRCA1 lead to genetic instability and deficiency in the different cell cycle phases [14]. 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. 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 gene is a growth factor receptor that is over-expressed in breast cancer and usually related to the aggressiveness of the tumor and the resistance to the chemotherapy [43]. RB1 gene is mutated in breast cancer [22] while the NFKB1 gene has a major role in invasive breast cancer [34]. MYC is a multifunctional protein that plays a role in cell cycle progression and cellular transformation. Amplification of MYC is found to be a frequent event in breast cancer that is often more associated with the metastatic version of the tumor [47]. The protein processing in endoplasmic reticulum pathway is another interesting pathway that is enriched in both mica2 and ica1. The endoplasmic reticulum (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 [50]. Interestingly, EDEM1 and SEL1L genes (mica2) are important parts of the ERAD component in addition to being de-regulated in cancer cells [50].

Since mica1, mica2, ica1, and ica5 contain interesting pathways, we further performed disease ontology enrichment analysis on these modules using FunDO [41]. The top diseases enriched in the modules, 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 mica3, cancer (3.64×10⁻¹⁰) in ica1, and cancer (1.11×10⁻⁴) in ica2.
in ica5, and cancer \((6.2 \times 10^{-5})\) and Melanoma \((1.1 \times 10^{-4})\) in ica1. Clearly, mica1 is the most enriched and related module to cancer in general and breast cancer, in specific.

### 4.2 Results on IDC data

Invasive Ductal Carcinoma is another famous breast cancer subtype. Previous works showed that IDC and ILC act differently and have different sets of DE genes \([59, 55]\). Nevertheless, we expect to find some common pathways between them, even though each pathway might include different sets of genes \([52]\).

Similar to ILC, we first used the dataset with ICA and Mica to see how different the output is when the miRNA data is added. As shown in Table 1, there is a significant difference between ICA and Mica modules. The Mica produced more highly scoring modules than ICA. In addition, Mica produced 66 modules while ICA produced 35 modules. We further analyzed the highest scoring modules from the two methods, namely, ica18, ica21, and ica30 from ICA and mica7, mica15, mica33, mica42, and mica63 from Mica. Those modules are the highest scoring modules with a score > 60. By comparing between the modules from ICA and Mica, we found that the most similar ones are mica42 and ica30; with 266 genes exist in both. The remaining Mica

<table>
<thead>
<tr>
<th>Database/Pathway</th>
<th>MICA pval</th>
<th>MICA Net</th>
<th>ICA pval</th>
<th>ICA Net</th>
<th>DEGAS pval</th>
<th>DEGAS Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reactome</td>
<td>23.81</td>
<td>6.78</td>
<td>10^{-05}</td>
<td>mica2</td>
<td>3.64</td>
<td>8.20</td>
</tr>
<tr>
<td>Processing of Capped Intron-Containing Pre-mRNA</td>
<td>5.60</td>
<td>4.21</td>
<td>10^{-03}</td>
<td>mica1</td>
<td>4.98</td>
<td>6.98</td>
</tr>
<tr>
<td>mRNA Splicing</td>
<td>5.30</td>
<td>4.21</td>
<td>10^{-03}</td>
<td>mica3</td>
<td>2.19</td>
<td>3.49</td>
</tr>
<tr>
<td>Cell Cycle, Mitotic</td>
<td>18.48</td>
<td>1.19</td>
<td>10^{-21}</td>
<td>mica5</td>
<td>11.53</td>
<td>7.79</td>
</tr>
<tr>
<td>Cell Cycle</td>
<td>19.96</td>
<td>7.30</td>
<td>10^{-15}</td>
<td>mica5</td>
<td>14.12</td>
<td>7.32</td>
</tr>
<tr>
<td>Mitotic M-M/G1 phases</td>
<td>13.31</td>
<td>3.75</td>
<td>10^{-18}</td>
<td>mica5</td>
<td>11.21</td>
<td>7.30</td>
</tr>
<tr>
<td>Elastic fibre formation</td>
<td>4.74</td>
<td>4.65</td>
<td>10^{-06}</td>
<td>mica6</td>
<td>22.41</td>
<td>8.20</td>
</tr>
<tr>
<td>Molecules associated with elastic fibres</td>
<td>3.95</td>
<td>2.81</td>
<td>10^{-06}</td>
<td>mica6</td>
<td>22.41</td>
<td>8.20</td>
</tr>
<tr>
<td>3'-UTR-mediated translational regulation</td>
<td>8.29</td>
<td>3.77</td>
<td>10^{-05}</td>
<td>mica7</td>
<td>22.41</td>
<td>8.20</td>
</tr>
<tr>
<td>L13a-mediated translational silencing of Ceruloplasmin expression</td>
<td>8.29</td>
<td>3.77</td>
<td>10^{-05}</td>
<td>mica7</td>
<td>22.41</td>
<td>8.20</td>
</tr>
<tr>
<td>Formation of a pool of free 40S subunits</td>
<td>7.80</td>
<td>3.98</td>
<td>10^{-06}</td>
<td>mica7</td>
<td>19.83</td>
<td>5.30</td>
</tr>
<tr>
<td>Eukaryotic Translation Initiation</td>
<td>8.29</td>
<td>3.98</td>
<td>10^{-06}</td>
<td>mica7</td>
<td>18.97</td>
<td>3.03</td>
</tr>
<tr>
<td>Antigen Presentation: Folding, assembly and peptide loading of class I MHC</td>
<td>1.73</td>
<td>7.32</td>
<td>10^{-03}</td>
<td>ic4</td>
<td>1.72</td>
<td>7.32</td>
</tr>
<tr>
<td>Extracellular matrix organization</td>
<td>21.55</td>
<td>5.25</td>
<td>10^{-15}</td>
<td>ic5</td>
<td>3.75</td>
<td>7.32</td>
</tr>
<tr>
<td>Molecules associated with elastic fibres</td>
<td>9.48</td>
<td>3.27</td>
<td>10^{-09}</td>
<td>ic5</td>
<td>11.21</td>
<td>2.02</td>
</tr>
<tr>
<td>Degradation of collagen</td>
<td>8.62</td>
<td>5.17</td>
<td>10^{-06}</td>
<td>ic5</td>
<td>10.24</td>
<td>4.25</td>
</tr>
<tr>
<td>Translation</td>
<td>24.13</td>
<td>8.66</td>
<td>10^{-14}</td>
<td>ic6</td>
<td>22.41</td>
<td>8.66</td>
</tr>
<tr>
<td>Cap-dependent Translation Initiation</td>
<td>22.41</td>
<td>8.66</td>
<td>10^{-14}</td>
<td>ic6</td>
<td>22.41</td>
<td>8.66</td>
</tr>
<tr>
<td>Eukaryotic Translation Initiation</td>
<td>22.41</td>
<td>8.66</td>
<td>10^{-14}</td>
<td>ic6</td>
<td>22.41</td>
<td>8.66</td>
</tr>
<tr>
<td>RTP hydrolysis and joining of the 60S ribosomal subunit</td>
<td>21.55</td>
<td>2.74</td>
<td>10^{-13}</td>
<td>ic6</td>
<td>21.55</td>
<td>2.74</td>
</tr>
<tr>
<td>GTP hydrolysis and joining of the 60S ribosomal subunit</td>
<td>21.55</td>
<td>2.74</td>
<td>10^{-13}</td>
<td>ic6</td>
<td>21.55</td>
<td>2.74</td>
</tr>
<tr>
<td>Peptide chain elongation</td>
<td>18.10</td>
<td>9.89</td>
<td>10^{-11}</td>
<td>ic6</td>
<td>18.10</td>
<td>9.89</td>
</tr>
<tr>
<td>Nonsense Mediated Decay Independent of the Exon Junction Complex</td>
<td>18.10</td>
<td>9.89</td>
<td>10^{-11}</td>
<td>ic6</td>
<td>18.10</td>
<td>9.89</td>
</tr>
<tr>
<td>Repair synthesis for gap-filling by DNA polymerase in TC-NER</td>
<td>1.73</td>
<td>7.32</td>
<td>10^{-03}</td>
<td>ic4</td>
<td>1.72</td>
<td>7.32</td>
</tr>
<tr>
<td>Removal of the Flap Intermediate from the C-strand</td>
<td>3.75</td>
<td>7.32</td>
<td>10^{-03}</td>
<td>ic4</td>
<td>3.75</td>
<td>7.32</td>
</tr>
<tr>
<td>Telomere Maintenance</td>
<td>3.75</td>
<td>7.32</td>
<td>10^{-03}</td>
<td>ic4</td>
<td>3.75</td>
<td>7.32</td>
</tr>
</tbody>
</table>

### Table 2: Pathway enrichment analysis for Mica, ICA, and DEGAS modules on the ILC dataset.
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, and \( N \) and \( E \) are the number of nodes and edges, respectively, for the largest connected module in the PPI, and \( \text{sc}(c) \) is the score of the largest connected module.

| #  | \( S \) | \( | c | \) | \( | c |_{ppi} \) | \( N \) | \( E \) | \( \text{sc}(c) \) |
|----|------|-------|-------|-----|-----|-------|
| 1  | 418 | 533 | 477 | 114 | 140 | 42.29 |
| 2  | 130 | 643 | 556 | 95 | 105 | 24.5 |
| 3  | 201 | 507 | 441 | 130 | 182 | 45.78 |
| 4  | 199 | 660 | 488 | 72 | 92 | 22.36 |
| 5  | 15 | 638 | 542 | 102 | 124 | 30.26 |
| 6  | 278 | 385 | 333 | 69 | 122 | 20.86 |
| 7  | 28 | 388 | 341 | 118 | 179 | 52.08 |
| 8  | 11 | 53 | 49 | 4 | 3 | 3.41 |
| 9  | 0 | 45 | 37 | 2 | 1 | 3.14 |
| 10 | 400 | 370 | 311 | 50 | 53 | 17.72 |
| 11 | 88 | 187 | 169 | 7 | 6 | 6.18 |
| 12 | 130 | 129 | 109 | 4 | 3 | 4.37 |
| 13 | 184 | 492 | 419 | 55 | 69 | 33.4 |
| 14 | 693 | 812 | 659 | 185 | 248 | 40.82 |
| 15 | 64 | 752 | 622 | 117 | 206 | 31.5 |
| 16 | 200 | 119 | 107 | 4 | 3 | 4.91 |
| 17 | 246 | 500 | 450 | 97 | 108 | 41.98 |
| 18 | 87 | 897 | 849 | 391 | 775 | 61.05 |
| 19 | 145 | 263 | 231 | 25 | 25 | 11.13 |
| 20 | 316 | 171 | 158 | 33 | 71 | 14.19 |
| 21 | 123 | 234 | 749 | 303 | 522 | 61.13 |
| 22 | 164 | 315 | 266 | 9 | 8 | 7.49 |
| 23 | 136 | 386 | 343 | 77 | 109 | 46.12 |
| 24 | 201 | 503 | 447 | 32 | 22 | 10 | 10.93 |
| 25 | 253 | 423 | 376 | 110 | 153 | 49.2 |
| 26 | 173 | 690 | 601 | 197 | 346 | 45.53 |
| 27 | 29 | 3 | 2 | 2 | 0 | 3.4 |
| 28 | 216 | 145 | 122 | 5 | 4 | 5.1 |
| 29 | 679 | 708 | 419 | 85 | 128 | 34.55 |
| 30 | 172 | 36 | 125 | 314 | 403 | 0.05 |
| 31 | 14 | 75 | 69 | 78 | 7 | 5.3 |
| 32 | 38 | 50 | 47 | 4 | 3 | 5.26 |
| 33 | 38 | 50 | 47 | 4 | 3 | 5.26 |
| 34 | 16 | 749 | 588 | 176 | 220 | 45.63 |
| 35 | 554 | 501 | 457 | 84 | 95 | 45.25 |

(1) http://cancer.sanger.ac.uk/cancergenome/projects/census/
down-regulated in cancer. Additionally, they are considered as potential therapeutic targets in cancer [12].

The DO enrichment analysis using FunDO is showed in Table 4. In general, Mica and MICA modules are significantly enriched with cancer and breast cancer genes than DEGAS, with Mica better enriched with breast cancer and cancer than ICA. Additionally, mica15 is enriched with metastatic to brain disease genes with APOA1 among those genes.

5. CONCLUSIONS

The unprecedented amount of publicly available disease-related data encourages the development of new methodologies and algorithms for a better analysis and further understanding the disease behavior. 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.

The experimental results show that the modules found by Mica are more disease-related while unraveling new dependencies between the genes which were hidden via previous techniques. Albeit the simplicity of the proposed workflow, Mica successfully includes many novel ideas, including how we adjust the gene expression levels with the miRNA expression to mimic the protein expression level and how we work on the genes first to get the related ones and map them to the PPI network rather than working only on the genes existing in the PPI. To the best of our knowledge, this is the first study that integrates miRNA, mRNA, and PPI network information for active module extraction. 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.

The results obtained from IDC and ILC datasets show the ability of Mica to generate disease specific modules. 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 would add more value and more understanding for the results. For instance, it would be more beneficial to extract a smaller module of 10 or 20 genes from each module that can be further used as a module 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. Pathways extraction can also benefit from adding directionality information to the PPI network. We are planning to tackle all such improvements in our future work.

6. ACKNOWLEDGMENTS

Funding: This work was partially supported by the NHI/NCI grant R01CA141090.

7. REFERENCES


Cell Cycle Checkpoints

- 9.48 \times 10^{-13} \text{ mica42} 8.15 \times 10^{-12} \text{ mica30} 8.03 \times 10^{-9}
- 4.02 \times 10^{-13} \text{ mica42} 3.52 \times 10^{-14} \text{ mica30}
- 3.45 \times 10^{-13} \text{ mica42} 2.64 \times 10^{-11} \text{ mica30}
- 3.45 \times 10^{-13} \text{ mica42} 2.64 \times 10^{-11} \text{ mica30}
- 8.62 \times 10^{-12} \text{ mica42} 8.15 \times 10^{-11} \text{ mica30}
- 6.90 \times 10^{-11} \text{ mica42} 5.51 \times 10^{-10} \text{ mica30} 4.4 \times 10^{-7}
- 6.90 \times 10^{-11} \text{ mica42} 5.51 \times 10^{-10} \text{ mica30} 4.4 \times 10^{-7}
- 3.16 \times 10^{-10} \text{ mica42} 8.66 \times 10^{-12} \text{ mica30}
- 3.16 \times 10^{-10} \text{ mica42} 8.66 \times 10^{-12} \text{ mica30}
- 7.4 \times 10^{-9} \text{ mica42} 3.08 \times 10^{-10} \text{ mica30}
- 2.87 \times 10^{-9} \text{ mica42} 2.64 \times 10^{-11} \text{ mica30}
- 2.87 \times 10^{-9} \text{ mica42} 2.64 \times 10^{-11} \text{ mica30}
- 2.59 \times 10^{-9} \text{ mica42} 1.76 \times 10^{-6} \text{ mica30} 1.55 \times 10^{-4}
- 2.87 \times 10^{-8} \text{ mica42} 2.64 \times 10^{-11} \text{ mica30}

DNA Replication Pre-Initiation

- 1.72 \times 10^{-7} \text{ mica42} 1.32 \times 10^{-9} \text{ mica30}
- 1.72 \times 10^{-7} \text{ mica42} 1.32 \times 10^{-9} \text{ mica30}
- 5.17 \times 10^{-7} \text{ mica42} 4.41 \times 10^{-7} \text{ mica30}

Fanconi Anemia pathway

- 3.16 \times 10^{-7} \text{ mica42} 2.86 \times 10^{-7} \text{ mica30}

Global Genomic NER (GG-NER)

- 3.45 \times 10^{-6} \text{ mica42} 3.08 \times 10^{-6} \text{ mica30}

Regulation of DNA replication

- 4.6 \times 10^{-5} \text{ mica42} 3.3 \times 10^{-4} \text{ mica30} 4.15 \times 10^{-2}

Repair of damaging factors from origins

- 4.6 \times 10^{-5} \text{ mica42} 3.3 \times 10^{-4} \text{ mica30} 4.15 \times 10^{-2}

Nuclosome assembly

- 4.31 \times 10^{-5} \text{ mica42} 3.52 \times 10^{-5} \text{ mica30} 4.4 \times 10^{-4}

Deposition of New CENPA-containing Nuclosomes at the Centromere

- 1.1 \times 10^{-5} \text{ mica42} 1.1 \times 10^{-5} \text{ mica30} 1.3 \times 2 \times 10^{-5}

Chk1/Cdc2 expression

- 1.15 \times 10^{-6} \text{ mica42} 1.1 \times 10^{-5} \text{ mica30} 1.04 \times 8 \times 10^{-6}

APC/C-mediated degradation of cell cycle proteins

- 4.31 \times 10^{-4} \text{ mica42} 3.74 \times 10^{-4} \text{ mica30} 4.66 \times 8 \times 10^{-5}

Regulation of mitotic cell cycle

- 4.31 \times 10^{-4} \text{ mica42} 3.74 \times 10^{-4} \text{ mica30} 4.66 \times 8 \times 10^{-5}

E2F-enabled inhibition of pre-replication complex formation

- 1.44 \times 10^{-5} \text{ mica42} 1.32 \times 10^{-5} \text{ mica30}

Homologous Recombination Repair

- 1.72 \times 10^{-4} \text{ mica42} 1.98 \times 10^{-4} \text{ mica30}

Homologous recombination repair of replication-independent double-strand breaks

- 1.72 \times 10^{-4} \text{ mica42} 1.98 \times 10^{-4} \text{ mica30}

Processive synthesis on the C-strand of the telomere

- 1.44 \times 10^{-4} \text{ mica42} 1.54 \times 10^{-5} \text{ mica30}

Double-Strand Break Repair

- 2.01 \times 10^{-4} \text{ mica42} 1.98 \times 10^{-4} \text{ mica30}

Activation of NIMA Kinases NEK9, NEK6, NEK7

- 1.15 \times 10^{-3} \text{ mica42} 0.88 \times 10^{-3} \text{ mica30}

G2/M DNA damage checkpoint

- 1.44 \times 10^{-4} \text{ mica42} 1.32 \times 10^{-4} \text{ mica30}

Kinesins

- 2.59 \times 10^{-3} \text{ mica42} 2.64 \times 10^{-3} \text{ mica30} 2.59 \times 10^{-3}

Base Excision Repair

- 1.72 \times 10^{-3} \text{ mica42} 1.32 \times 10^{-3} \text{ mica30}

Resolution of Abasic Sites (AP sites)

- 1.72 \times 10^{-3} \text{ mica42} 1.32 \times 10^{-3} \text{ mica30}

CDC6 association with the ORC:origin complex

- 1.15 \times 10^{-3} \text{ mica42}

G2/M DNA replication checkpoint

- 0.86 \times 10^{-3} \text{ mica42} 0.88 \times 10^{-4} \text{ mica30} 0.78 \times 10^{-3}

Removal of the Flap Intermediate from the C-strand

- 1.15 \times 10^{-3} \text{ mica42} 1.32 \times 10^{-4} \text{ mica30}

G2 Phase

- 0.86 \times 10^{-3} \text{ mica42}

Removal of DNA patch containing abasic residue

- 1.44 \times 10^{-3} \text{ mica42}

Resolution of AP sites via the multiple-nucleotide patch re-placement pathway

- 1.44 \times 10^{-3} \text{ mica42}

Regulation of APC/C activators between G1/S and early anaphase

- 3.45 \times 10^{-3} \text{ mica42} 4.15 \times 4 \times 10^{-4}

Post-transcriptional Silencing By Small RNAs

- 1.79 \times 10^{-6} \text{ mica42}

Pre-NOTCH Transcription and Translation

- 2.05 \times 10^{-5} \text{ mica42}

Cohesin Loading onto Chromatin

- 1.53 \times 10^{-3} \text{ mica42}

Pre-NOTCH Expression and Processing

- 2.05 \times 10^{-3} \text{ mica42}

Small Interfering RNA (siRNA) Biogenesis

- 1.28 \times 10^{-3} \text{ mica42}

Mitotic Telophase/Cytokinesis

- 1.53 \times 10^{-3} \text{ mica42}

DNA Polymerase II Transcription Termination

- 3.3 \times 10^{-3} \text{ mica21}

Cleavage of Growing Transcript in the Termination Region

- 3.3 \times 10^{-3} \text{ mica21}

Post-elongation Processing of the Transcript

- 3.3 \times 10^{-3} \text{ mica21}

DNA Polymerase II Transcription

- 5.28 \times 10^{-3} \text{ mica21}

Mitotic G2-G2/M phases

- 6.83 \times 10^{-11} \text{ mica30}

G2/M Transition

- 6.39 \times 10^{-10} \text{ mica30}
Table 5: Pathway enrichment analysis for ICA, DEGAS, and Mica.