MICA: MicroRNA Integration for Active Module Discovery

Ayat Hatem
The Biomed. Informatics Dept.
The Elec. & Comp. Eng. Dept.
dayat@bmi.osu.edu

Kamer Kaya
The Biomed. Informatics Dept.
kamer@bmi.osu.edu

Jeffrey Parvin
The Biomed. Informatics Dept.
Jeffrey.Parvin@osumc.edu

Kun Huang
The Biomed. Informatics Dept.
kun.huang@osumc.edu

Ümit V. Çatalyürek
The Biomed. Informatics Dept.
The Elec. & Comp. Eng. Dept.
umit@bmi.osu.edu

ABSTRACT

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 mRNA 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

n 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:

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

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 independent components. The columns of $\mathbf{S}$ correspond to the genes and the entry $\mathbf{S}_{cg}$ 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-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 [21, 10].

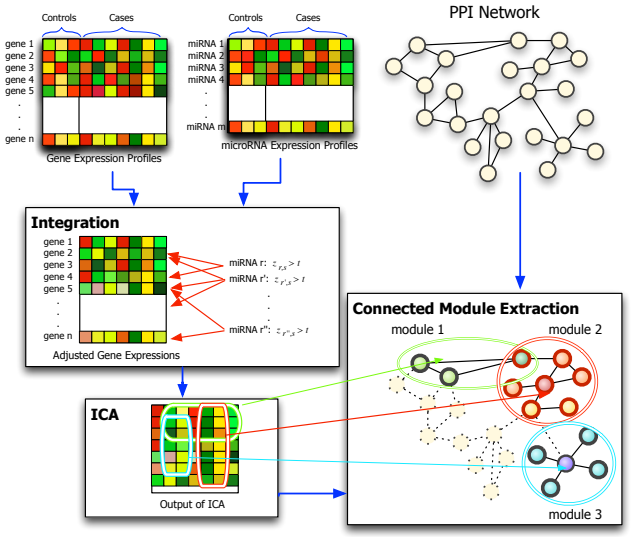


Figure 1: MICA: The workflow starts with integrating miRNA and mRNA data by adjusting the mRNA data using the miRNA data. Then, ICA is applied on the resulting new gene-expression matrix. Finally, for each independent component obtained by ICA, the largest connected module from the PPI network is extracted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

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:

$$\beta_{g,s} = \frac{\mid \sum_{\{r:\ r \text{ affects } g\}} Z^{-}_{r,s} \mid}{\sum_{\{r:\ r \text{ affects } g\}} Z^{+}_{r,s}} \tag{2}$$

where $Z^+_{r,s}$ ($Z^-_{r,s}$) 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

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

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. 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:

$$e'_{g,s} = \beta_{g,s} \times e_{g,s} \quad (4)$$

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 minimas and unreliable independent component estimates, we follow the method in [10]: we run `fastICA` κ 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 $V = V_1 \cup \cdots \cup V_\kappa$ are 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') 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 κ 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 ℓ to be generated; when ℓ 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 $\mathbf{S}$ and $\mathbf{A}$ 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 μ and σ 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, μ, and σ 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_G 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_G 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:

$$scr(c) = \frac{\sum_{g \in c} Z_{cg}}{\sqrt{|c|}} \quad (5)$$

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_{cg} 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 `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 `degas`, ICA modules as `ica`, and MICA modules as `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 *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 miRTarBase (rel. 4.5) [25].

The number of runs κ for ICA is set to 100 while t_R threshold is set to 4 and t_C and t_G 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 clusterProfiler [57].

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 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	$scr(c)$
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

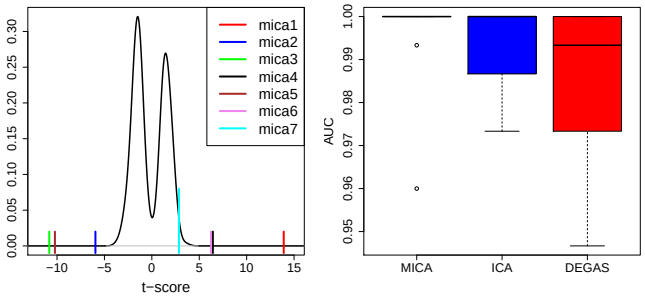
4.1 Results on ILC data

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 `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 in the A matrix can be obtained if we have a random input matrix. Accordingly, we generated 1000 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. 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 predication 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



(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 a random run. b) MICA modules prediction performance after a 10-fold cross validation in comparison to `ICA` and `DEGAS`.

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`

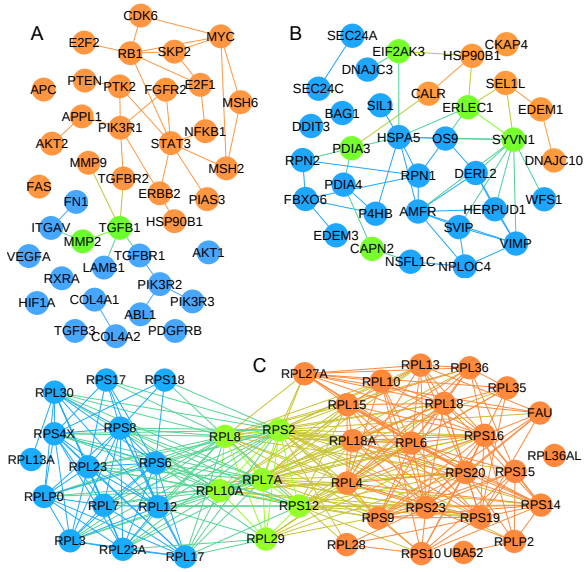


Figure 3: Overlap between Important pathways enriched in both MICA and `ICA` modules. Orange is for MICA, blue is for `ICA`, and green for genes in both. A) Pathways in cancer (`mica1` and `ica5`, B) Protein processing in endoplasmic reticulum (`mica2` and `ica1`, C) Ribosome (`mica7` and `ica6`).

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 overexpressed 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 rule 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})

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	Unfolded Protein Response	23.81	6.78×10^{-05}	mica2	3.64	8.20×10^{-03}	ica4		
	Processing of Capped Intron-Containing Pre-mRNA	5.60	4.21×10^{-03}	mica1					
	mRNA Splicing	5.30	4.21×10^{-03}	mica3					
	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					
	Elastic fibre formation	4.74	4.05×10^{-05}	mica6	11.21	7.30×10^{-11}	ica5		
	Molecules associated with elastic fibres	3.95	2.81×10^{-04}	mica6					
	3' -UTR-mediated translational regulation	8.29	3.77×10^{-05}	mica7	22.41	8.20×10^{-14}	ica6		
	L13a-mediated translational silencing of Ceruloplasmin expression	8.29	3.77×10^{-05}	mica7	22.41	8.20×10^{-14}	ica6		
	Formation of a pool of free 40S subunits	7.80	3.98×10^{-05}	mica7	19.83	5.30×10^{-12}	ica6		
	Eukaryotic Translation Initiation	8.29	3.98×10^{-05}	mica7	18.97	3.03×10^{-11}	ica6		
	Antigen Presentation: Folding, assembly and peptide loading of class I MHC				4.52	1.62×10^{-06}	ica1		
	Interferon alpha/beta signaling				5.88	7.99×10^{-05}	ica1		
	Golgi Cisternae Pericentriolar Stack Reorganization				2.71	5.10×10^{-04}	ica1		
	ER-Phagosome pathway				4.98	6.98×10^{-04}	ica1		
	PERK regulated gene expression				2.19	3.49×10^{-03}	ica4		
	Toll Like Receptor 4 (TLR4) Cascade				5.47	4.25×10^{-03}	ica4		
	Cytokine Signaling in Immune system				10.21	4.25×10^{-03}	ica4		
	Antigen Presentation: Folding, assembly and peptide loading of class I MHC				2.55	6.00×10^{-03}	ica4		
	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		
	Degradation of collagen				8.62	5.17×10^{-06}	ica5		
	Translation				24.13	8.66×10^{-14}	ica5		
	Cap-dependent Translation Initiation				22.41	8.66×10^{-14}	ica6		
	Eukaryotic Translation Initiation				22.41	8.66×10^{-14}	ica6		
	GTP hydrolysis and joining of the 60S ribosomal subunit				21.55	2.74×10^{-13}	ica6		
	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		
	Repair synthesis for gap-filling by DNA polymerase in TC-NER							1.73	7.32×10^{-3}
	Removal of the Flap Intermediate from the C-strand							1.72	7.32×10^{-3}
	Telomere Maintenance							3.75	7.32×10^{-3}
KEGG	Pancreatic cancer	6.70	1.05×10^{-04}	mica1	6.03	4.15×10^{-03}	ica5		
	Pathways in cancer	15.24	1.05×10^{-04}	mica1	14.66	2.59×10^{-03}	ica5		
	Small cell lung cancer	7.31	1.05×10^{-04}	mica1	7.75	7.07×10^{-04}	ica5		
	Chronic myeloid leukemia	6.09	7.01×10^{-04}	mica1	6.89	1.26×10^{-03}	ica5		
	Colorectal cancer	5.49	8.10×10^{-04}	mica1	5.17	9.87×10^{-03}	ica5		
	Bladder cancer	4.27	2.18×10^{-03}	mica1					
	Prostate cancer	6.09	2.24×10^{-03}	mica1					
	Non-small cell lung cancer	74.27	8.10×10^{-03}	mica1					
	Protein processing in endoplasmic reticulum	52.38	4.65×10^{-11}	mica2	12.22	1.10×10^{-08}	ica1		
	Spliceosome	6.19	1.24×10^{-03}	mica3					
	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		
	Focal adhesion				16.28	3.83×10^{-07}	ica6		
	TGF-beta signaling pathway				7.76	7.07×10^{-04}	ica6		
	Renal cell carcinoma				6.03	4.15×10^{-03}	ica6		

in `ica5`, and cancer (6.2×10^{-5}) and Melanoma (1.1×10^{-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 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	$scr(c)$
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	30.08
6	278	385	333	69	122	20.86
7	28	388	341	118	179	52.08
8	11	53	49	4	3	4.31
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	131	34.5
16	200	119	107	4	3	4.91
17	246	500	450	97	108	41.98
18	**87**	**897**	**849**	**391**	**775**	**61.95**
19	145	263	231	25	25	11.15
20	316	171	158	33	71	14.19
21	**123**	**744**	**669**	**303**	**522**	**61.43**
22	164	315	266	9	8	7.49
23	136	386	343	77	109	46.12
24	201	503	447	112	137	26.47
25	253	423	376	110	153	49.62
26	173	690	601	197	316	44.53
27	29	3	2	2	0	3.4
28	216	145	122	5	4	5.1
29	6	708	612	186	234	34.55
30	**513**	**675**	**649**	**454**	**1851**	**83.63**
31	42	540	457	171	252	33.83
32	38	603	502	111	140	27.59
33	5	228	201	7	6	6.65
34	16	749	588	176	220	45.63
35	554	501	457	84	95	45.25

(b) MICA

#	S	$\|c\|$	$\|c\|_{ppi}$	N	E	$scr(c)$
1	324	595	538	154	182	45.82
2	76	571	526	212	329	37.71
3	523	535	473	68	78	35.5
4	308	289	245	22	23	11.28
5	319	679	604	169	249	37.61
6	134	412	376	52	57	18.07
7	**296**	**400**	**374**	**147**	**234**	**61.78**
8	174	655	592	188	266	36.24
9	296	380	329	50	57	19.55
10	294	483	413	99	137	25.97
11	414	661	583	136	176	34.89
12	254	83	68	5	5	4.85
13	516	323	279	34	35	14.67
14	284	55	48	2	1	3.27
15	**336**	**317**	**267**	**42**	**47**	**59.76**
16	255	733	670	299	458	47.19
17	216	542	425	67	86	36.61
18	260	335	296	55	70	19.59
19	319	159	145	58	98	18.6
20	325	623	510	62	66	25.49
21	436	272	258	101	208	58.79
22	20	565	473	54	58	28.33
23	208	543	473	91	113	33.63
24	262	570	512	167	275	34.7
25	309	532	483	184	244	57.42
26	328	403	377	152	243	54.86
27	278	455	389	80	88	31.39
28	262	655	579	162	214	36.99
29	237	341	303	13	13	9.18
30	196	420	369	122	148	28.42
31	257	682	602	202	726	50.76
32	3	212	173	11	10	9.83
33	**245**	**289**	**280**	**138**	**297**	**79.69**
34	362	174	153	6	5	6
35	380	495	433	106	135	31.81
36	160	768	662	286	909	54.72
37	169	534	471	135	199	30.98
38	166	700	619	178	218	36.7
39	132	665	607	197	298	36.41
40	466	378	332	69	78	21.61
41	246	156	153	8	8	6.83
42	**544**	**682**	**633**	**348**	**1063**	**66.97**
43	51	473	397	19	35	12.81
44	209	8	7	7	0	5.84
45	185	634	565	156	202	32.8
46	32	379	317	39	43	15.52
47	214	444	379	53	66	29.8
48	450	248	217	18	17	9.9
49	278	290	247	38	42	15.14
50	170	110	96	6	5	5.78
51	300	4	4	4	0	4.22
52	363	2	2	2	0	3.38
53	179	5	5	5	0	5.05
54	314	581	453	12	14	23.74
55	0	731	618	161	192	48.34
56	361	110	92	4	3	5.26
57	374	289	255	17	20	11.91
58	29	764	594	67	85	28.48
59	496	1	1	1	0	2.46
60	432	1	0	0	0	NA
61	99	535	433	78	104	40.37
62	306	457	425	60	68	20.36
63	**242**	**243**	**230**	**101**	**188**	**66.43**
64	186	565	506	163	222	49.83
65	1	494	444	159	246	58.05

and ICA modules did not have any significant overlap.

By further examining the genes in mica42 and ica30, we found that both contain BRCA1, BRCA2, BRIP1, BLM, RAD51, UBE2C, and CKS2. BLM and RAD51 have a tumorigenic significance [15], UBE2C and CKS2 are among the genes that are DE in IDC [39], and BRCA1, BRIP1, and BRCA2 are known breast cancer mutated [1]. On the other hand mica42 only contains TOP3A, HMG20B, RAD51C, CDC6, and U2AF1 genes. HMG20B gene interacts directly with BRCA2. The inhibition, of the interaction between HMG20B and BRCA2 lead to progression of tumor [32]. TOP3A and BLM genes interact with RMI1 gene forming a complex that is very important in genome stability [7]. The mutations in this complex increase the risk of breast cancer in addition to other types of cancer [5]. RAD51C gene was also found to be mutated in breast cancer [35]. The deregulation of CDC6 poses a serious risk of carcinogenesis [36] while U2AF1 is a splicing factor protein that is mutated in cancer in general [20].

The degas module on IDC data contains 386 genes with $1,056$ interactions and 190 DE genes. Based on the quality measure, the module has a p-value of 0, i.e., it cannot be randomly obtained. There are 105 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 exclusively exist in mica42 exist in degas.

In addition to examining the different obtained modules, we performed classification analysis using the different modules and datasets to ensure that the adjusted gene expression data better correlate with the disease behavior. Similar to the ILC dataset, a SVM was trained on the active modules obtained from each tool separately. Then, a 10-fold cross validation was performed using the original data for ICA and DEGAS and modified gene expression data for MICA. The three tools almost performed the same 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 further performed pathway enrichment analysis, as shown in Table 5. There are a lot of pathways common between mica42, mica30, and degas such as Cell cycle, Tolemere maintenance, and DNA strand elongation. However, mica42 alone was enriched with the p53 signaling pathway. Interestingly, there are many important pathways enriched in mica15 which were not enriched in any other tools, including the complement and coagulation cascades, platelet degranulation, and Hemostasis pathways. All of these pathways are part of the hemostatic system of the cell. Hemostatic elements are considered important in facilitating the metastatic potential of breast cancer [31]. Additionally, A proteomic based study has shown the complement and coagulation pathway to be DE in IDC([48] . Figure 4 shows the genes in mica15 module. Among the nodes in this network and also in the Hemostasis pathway is the APOA1 gene. APOA1 gene was found DE in IDC samples vs control samples in a proteomic study [42]. In addition, mutations in this gene lead to poor outcome for post-surgery breast cancer patients [24]. Other interesting genes in mica15 are GADD45A, GADD45B, and GADD45G genes. GADD45 genes are stress sensor genes that are activated in respond to cell stress and DNA damage. GADD45 genes were found

[1] http://cancer.sanger.ac.uk/cancergenome/projects/census/

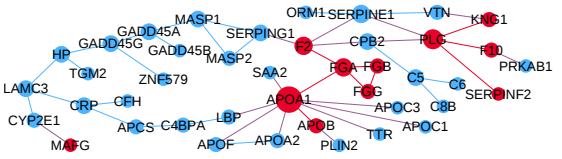


Figure 4: `mica15` module. The red nodes are for the nodes in the Hemostasis pathway.

Table 4: 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, metastatic to brain	4.67×10^{-9}, 1.16×10^{-8}, 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}

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

[1] D. Baek, J. Villén, C. Shin, et al. The impact of microrna 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 Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 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] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka. Network biology approach to complex diseases. *PLoS comp biol*, 8(12):e1002820, 2012.

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

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

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

[15] 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.

[16] 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.

[17] J. M. Engreitz, B. J. Daigle, J. J. Marshall, and R. B. Altman. Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of biomed. info.*, 43(6):932–944, 2010.

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

[19] A. Frigyesi, S. Veerla, D. Lindgren, and M. Höglund. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC bioinf.*, 7(1):290, 2006.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

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

[31] I. Lal, K. Dittus, and C. E. Holmes. Platelets, coagulation and fibrinolysis in breast cancer progression. *Breast Cancer Research*, 15(4):1–11, 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] S.-I. Lee, S. Batzoglou, et al. Application of independent component analysis to microarrays. *Genome biology*, 4(11):R76–R76, 2003.

[34] 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.

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

[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, S. M. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila, G. Feng, and R. L. Chisholm. 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] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] R. A. Verdugo, T. Zeller, M. Rotival, et al. Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers. *PloS one*, 8(1):e50888, 2013.

[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, 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.

[59] 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.

Database	Pathway	MICA			ICA			DEGAS	
		%	pval	Name	%	pval	Name	%	pval
KEGG	Complement and coagulation cascades	42.86	1.17×10^{-23}	mica15					
	Staphylococcus aureus infection	14.29	2.97×10^{-5}	mica15					
	DNA replication	6.32	6.68×10^{-17}	mica42	5.51	1.13×10^{-18}	ica30		
	Cell cycle	11.21	6.68×10^{-17}	mica42	10.13	1.13×10^{-18}	mica30	6.22	3.04×10^{-4}
	Mismatch repair	3.16	5.53×10^{-07}	mica42	3.30	1.11×10^{-10}	ica30		
	Nucleotide excision repair	3.45	1.62×10^{-4}	mica42	3.3	9.57×10^{-6}	mica30		
	Homologous recombination	2.59	3.57×10^{-04}	mica42	2.64	6.97×10^{-06}	ica30		
	Base excision repair	2.59	1.65×10^{-03}	mica42	2.64	5.46×10^{-05}	ica30		
	p53 signaling pathway	3.45	7.86×10^{-03}	mica42					
	Spliceosome				6.60	8.20×10^{-04}	ica21		
	Oocyte meiosis				4.63	1.43×10^{-03}	ica30		
Reactome	Formation of Fibrin Clot (Clotting Cascade)	16.67	2.68×10^{-8}	mica15					
	Complement cascade	16.67	3.60×10^{-8}	mica15					
	Platelet degranulation	21.43	6.66×10^{-08}	mica15					
	Common Pathway	11.90	6.66×10^{-08}	mica15					
	Response to elevated platelet cytosolic Ca2+	21.43	7.88×10^{-8}	mica15					
	Chylomicron-mediated lipid transport	9.52	7.98×10^{-06}	mica15					
	Platelet activation, signaling and aggregation	23.81	9.16×10^{-06}	mica15					
	Intrinsic Pathway	9.52	2.81×10^{-05}	mica15					
	Retinoid metabolism and transport	11.90	4.16×10^{-05}	mica15					
	Terminal pathway of complement	7.14	5.70×10^{-5}	mica15					
	Lipoprotein metabolism	9.52	6.14×10^{-5}	mica15					
	Hemostasis	30.95	6.80×10^{-05}	mica15					
	Visual phototransduction	11.9	1.29×10^{-4}	mica15					
	Dissolution of Fibrin Clot	7.14	1.29×10^{-4}	mica15					
	Diseases associated with visual transduction	11.90	1.29×10^{-04}	mica15					
	Platelet Aggregation (Plug Formation)	9.52	3.18×10^{-04}	mica15					
	p130Cas linkage to MAPK signaling for integrins	7.14	3.97×10^{-04}	mica15					
	GRB2:SOS provides linkage to MAPK signaling for Intergrins	7.14	3.97×10^{-04}	mica15					
	Lectin pathway of complement activation	4.76	4.20×10^{-4}	mica15					
	Lipid digestion, mobilization, and transport	9.52	4.27×10^{-4}	mica15					
	Integrin alphaIIb beta3 signaling	7.14	2.07×10^{-3}	mica15					
	Transport of gamma-carboxylated protein precursors from the endoplasmic reticulum to the Golgi apparatus	4.76	4.06×10^{-3}	mica15					
	Creation of C4 and C2 activators	4.76	4.06×10^{-3}	mica15					
	Removal of aminoterminal propeptides from gamma-carboxylated proteins	4.76	4.65×10^{-3}	mica15					
	Gamma-carboxylation of protein precursors	4.76	4.65×10^{-3}	mica15					
	Amyloids	9.52	5.11×10^{-3}	mica15					
	Integrin cell surface interactions	9.52	5.14×10^{-3}	mica15					
	Gamma-carboxylation, transport, and amino-terminal cleavage of proteins	4.76	6.03×10^{-3}	mica15					
	HDL-mediated lipid transport	4.76	6.86×10^{-3}	mica15					
	Binding and Uptake of Ligands by Scavenger Receptors	4.76	8.30×10^{-3}	mica15					
	Scavenging of Heme from Plasma	4.76	8.30×10^{-3}	mica15					
	Regulation of Complement cascade	4.76	8.30×10^{-3}	mica15					
	mRNA Splicing	9.42	1.52×10^{-04}	mica33	6.60	7.65×10^{-05}	ica21		
	mRNA Splicing - Major Pathway	9.42	1.52×10^{-4}	mica33	6.6	7.65×10^{-5}	mica21		
	Processing of Capped Intron-Containing Pre-mRNA	9.42	1.52×10^{-4}	mica33	6.6	9.06×10^{-5}	mica21		
	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}
	Cell Cycle	35.06	4.67×10^{-49}	mica42	33.26	7.77×10^{-59}	mica30	21.76	3.46×10^{-16}
	Mitotic M-M/G1 phases	23.28	3.79×10^{-39}	mica42	20.48	9.30×10^{-41}	mica30	13.73	4.74×10^{-13}
	Mitotic Prometaphase	13.79	3.81×10^{-27}	mica42	12.56	5.37×10^{-31}	mica30	7.51	1.88×10^{-8}
	DNA strand elongation	7.18	3.92×10^{-25}	mica42	5.95	6.16×10^{-26}	ica30	2.59	1.45×10^{-04}
	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}
	M Phase	16.67	1.69×10^{-22}	mica42	15.2	7.73×10^{-25}	mica30	11.4	5.22×10^{-10}
	DNA Replication	10.92	3.10×10^{-21}	mica42	8.59	2.57×10^{-18}	mica30	5.7	3.70×10^{-6}
	Activation of the pre-replicative complex	6.32	1.20×10^{-19}	mica42	5.07	9.03×10^{-19}	mica30	2.33	1.01×10^{-3}
	G2/M Checkpoints	7.47	1.66×10^{-19}	mica42	6.39	2.02×10^{-20}	mica30	4.4	7.13×10^{-8}
	S Phase	11.21	3.69×10^{-19}	mica42	10.57	4.78×10^{-23}	mica30	6.74	4.03×10^{-7}
	Mitotic G1-G1/S phases	11.78	1.30×10^{-18}	mica42	10.13	1.68×10^{-18}	mica30	6.74	4.94×10^{-6}
	Synthesis of DNA	9.77	1.80×10^{-18}	mica42	7.71	7.13×10^{-16}	mica30	5.44	4.36×10^{-6}
	Mitotic Metaphase and Anaphase	13.51	2.96×10^{-17}	mica42	13	3.69×10^{-21}	mica30	9.33	8.39×10^{-8}
	Mitotic Anaphase	13.22	1.35×10^{-16}	mica42	12.78	1.56×10^{-20}	mica30	9.33	8.17×10^{-8}
	Activation of ATR in response to replication stress	6.32	1.64×10^{-16}	mica42	5.29	1.56×10^{-16}	mica30	3.63	1.94×10^{-6}
	G1/S Transition	9.77	6.96×10^{-16}	mica42	8.37	9.68×10^{-16}	mica30	6.48	3.82×10^{-7}
	Separation of Sister Chromatids	12.36	1.84×10^{-15}	mica42	12.11	1.03×10^{-19}	mica30	8.55	4.86×10^{-7}
	Telomere C-strand (Lagging Strand) Synthesis	4.6	4.39×10^{-15}	mica42	3.96	1.78×10^{-16}	mica30		
	Chromosome Maintenance	9.77	1.40×10^{-14}	mica42	8.81	7.82×10^{-16}	mica30	5.96	2.29×10^{-5}
	Extension of Telomeres	4.6	3.55×10^{-14}	mica42	3.96	2.32×10^{-15}	mica30		
	Unwinding of DNA	3.16	1.78×10^{-13}	mica42	2.42	3.05×10^{-12}	mica30	1.55	2.68×10^{-4}
	E2F mediated regulation of DNA replication	5.17	3.40×10^{-13}	mica42	3.96	2.55×10^{-11}	mica30	2.07	8.80×10^{-3}

Cell Cycle Checkpoints	9.48	4.77×10^{-13}	mica42	8.15	1.15×10^{-12}	mica30	8.03	1.94×10^{-9}
Lagging Strand Synthesis	4.02	6.05×10^{-13}	mica42	3.52	1.96×10^{-14}	mica30		
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		
Polymerase switching on the C-strand of the telomere	3.45	6.05×10^{-13}	mica42	2.64	1.40×10^{-11}	mica30		
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}
Gap-filling DNA repair synthesis and ligation in TC-NER	3.16	4.05×10^{-10}	mica42	2.86	6.15×10^{-12}	mica30		
Gap-filling DNA repair synthesis and ligation in GG-NER	3.16	4.05×10^{-10}	mica42	2.86	6.15×10^{-12}	mica30		
G0 and Early G1	3.74	3.16×10^{-9}	mica42	3.08	5.41×10^{-9}	mica30		
Repair synthesis for gap-filling by DNA polymerase in TC-NER	2.87	4.86×10^{-9}	mica42	2.64	5.77×10^{-11}	mica30		
Repair synthesis of patch 27-30 bases long by DNA polymerase	2.87	4.86×10^{-9}	mica42	2.64	5.77×10^{-11}	mica30		
Condensation of Prometaphase Chromosomes	2.59	7.26×10^{-9}	mica42	1.76	1.63×10^{-6}	mica30	1.55	4.81×10^{-4}
G1/S-Specific Transcription	2.87	1.17×10^{-8}	mica42	2.2	1.36×10^{-7}	mica30	1.55	2.64×10^{-3}
Processive synthesis on the lagging strand	2.59	1.35×10^{-7}	mica42	2.42	2.11×10^{-9}	mica30		
DNA replication initiation	1.72	2.31×10^{-7}	mica42	1.32	9.12×10^{-7}	mica30		
Telomere C-strand synthesis initiation	1.72	2.31×10^{-07}	mica42	1.32	9.12×10^{-07}	ica30		
Telomere Maintenance	5.17	2.68×10^{-07}	mica42	4.41	4.87×10^{-07}	ica30	3.63	1.01×10^{-03}
Fanconi Anemia pathway	3.16	8.48×10^{-07}	mica42	2.86	1.23×10^{-07}	ica30		
Removal of the Flap Intermediate	2.30	1.37×10^{-06}	mica42	2.20	2.13×10^{-08}	ica30		
Global Genomic NER (GG-NER)	3.45	1.75×10^{-06}	mica42	3.08	4.60×10^{-07}	ica30		
Regulation of DNA replication	4.6	6.64×10^{-6}	mica42	3.3	4.84×10^{-4}	mica30	4.15	7.20×10^{-5}
Removal of licensing factors from origins	4.6	6.64×10^{-6}	mica42	3.3	4.84×10^{-4}	mica30	4.15	7.20×10^{-5}
Nucleosome assembly	4.31	1.03×10^{-5}	mica42	3.52	4.33×10^{-5}	mica30	4.4	4.74×10^{-6}
Deposition of New CENPA-containing Nucleosomes at the Centromere	4.31	1.03×10^{-5}	mica42	3.52	4.33×10^{-5}	mica30	4.4	4.74×10^{-6}
Phosphorylation of Emi1	1.44	2.04×10^{-05}	mica42	1.10	5.57×10^{-05}	ica30		
Cyclin A/B1 associated events during G2/M transition	2.01	2.40×10^{-5}	mica42	1.98	4.72×10^{-7}	mica30	2.07	9.42×10^{-6}
Nucleotide Excision Repair	3.74	2.40×10^{-05}	mica42	3.30	1.33×10^{-05}	ica30		
Transcription-coupled NER (TC-NER)	3.45	3.60×10^{-05}	mica42	3.08	1.46×10^{-05}	ica30		
Orc1 removal from chromatin	4.02	1.03×10^{-4}	mica42	2.86	3.86×10^{-3}	mica30	3.89	1.82×10^{-4}
Switching of origins to a post-replicative state	4.02	1.03×10^{-4}	mica42	2.86	3.86×10^{-3}	mica30	3.89	1.82×10^{-4}
Nuclear Envelope Breakdown	2.01	1.09×10^{-4}	mica42	1.54	4.41×10^{-4}	mica30	1.55	3.78×10^{-3}
Assembly of the pre-replicative complex	3.74	1.66×10^{-4}	mica42	2.64	5.40×10^{-3}	mica30	3.63	2.66×10^{-4}
Inhibition of replication initiation of damaged DNA by RB1/E2F1	1.72	2.02×10^{-4}	mica42	1.32	6.39×10^{-4}	mica30		
Cyclin B2 mediated events	1.15	2.65×10^{-4}	mica42	1.1	1.10×10^{-5}	mica30	1.3	2.00×10^{-5}
Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex	1.15	2.65×10^{-4}	mica42	1.1	1.10×10^{-5}	mica30	1.04	8.63×10^{-4}
APC/C-mediated degradation of cell cycle proteins	4.31	4.54×10^{-4}	mica42	3.74	6.24×10^{-4}	mica30	4.66	7.87×10^{-5}
Regulation of mitotic cell cycle	4.31	4.54×10^{-4}	mica42	3.74	6.24×10^{-4}	mica30	4.66	7.87×10^{-5}
E2F-enabled inhibition of pre-replication complex formation	1.44	5.67×10^{-4}	mica42	1.32	1.07×10^{-4}	mica30		
Homologous Recombination Repair	1.72	7.34×10^{-4}	mica42	1.98	2.04×10^{-6}	mica30		
Homologous recombination repair of replication-independent double-strand breaks	1.72	7.34×10^{-4}	mica42	1.98	2.04×10^{-6}	mica30		
Processive synthesis on the C-strand of the telomere	1.44	9.48×10^{-4}	mica42	1.54	1.33×10^{-5}	mica30		
Double-Strand Break Repair	2.01	1.17×10^{-3}	mica42	2.42	1.54×10^{-6}	mica30		
Activation of NIMA Kinases NEK9, NEK6, NEK7	1.15	1.49×10^{-3}	mica42	0.88	3.18×10^{-3}	mica30		
G2/M DNA damage checkpoint	1.44	1.49×10^{-3}	mica42	1.32	3.94×10^{-4}	mica30		
Kinesins	2.59	1.51×10^{-3}	mica42	2.64	8.91×10^{-5}	mica30	2.59	1.48×10^{-3}
Base Excision Repair	1.72	1.95×10^{-3}	mica42	1.32	6.00×10^{-3}	mica30		
Resolution of Abasic Sites (AP sites)	1.72	1.95×10^{-3}	mica42	1.32	6.00×10^{-3}	mica30		
CDC6 association with the ORC:origin complex	1.15	2.68×10^{-3}	mica42					
G2/M DNA replication checkpoint	0.86	3.22×10^{-3}	mica42	0.88	1.31×10^{-4}	mica30	0.78	8.80×10^{-3}
Removal of the Flap Intermediate from the C-strand	1.15	7.17×10^{-3}	mica42	1.32	1.07×10^{-4}	mica30		
G2 Phase	0.86	7.52×10^{-3}	mica42					
Removal of DNA patch containing abasic residue	1.44	8.20×10^{-3}	mica42					
Resolution of AP sites via the multiple-nucleotide patch replacement pathway	1.44	8.20×10^{-3}	mica42					
Regulation of APC/C activators between G1/S and early anaphase	3.45	8.57×10^{-3}	mica42				4.15	4.81×10^{-4}
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		
Cohesin Loading onto Chromatin				1.53	1.41×10^{-03}	ica18		
Pre-NOTCH Expression and Processing				2.05	3.32×10^{-3}	ica18		
Small Interfering RNA (siRNA) Biogenesis				1.28	8.16×10^{-03}	ica18		
Mitotic Telophase/Cytokinesis				1.53	8.16×10^{-03}	ica18		
RNA Polymerase II Transcription Termination				3.3	2.18×10^{-3}	mica21		
Cleavage of Growing Transcript in the Termination Region				3.3	2.18×10^{-3}	mica21		
Post-Elongation Processing of the Transcript				3.3	2.18×10^{-3}	mica21		
RNA Polymerase II Transcription				5.28	2.18×10^{-3}	mica21		
Mitotic G2-G2/M phases				6.83	2.74×10^{-11}	mica30		
G2/M Transition				6.39	3.70×10^{-10}	mica30		

Centrosome maturation	5.07	1.26×10^{-7}	mica30		
Recruitment of mitotic centrosome proteins and complexes	5.07	1.26×10^{-7}	mica30		
Loss of Nlp from mitotic centrosomes	3.96	2.04×10^{-6}	mica30		
Loss of proteins required for interphase microtubule organizationåÊfrom the centrosome	3.96	2.04×10^{-6}	mica30		
Establishment of Sister Chromatid Cohesion	1.54	1.33×10^{-5}	mica30		
Interactions of Rev with host cellular proteins	2.42	9.41×10^{-5}	mica30		
Recruitment of NuMA to mitotic centrosomes	1.98	2.24×10^{-4}	mica30		
Rev-mediated nuclear export of HIV-1 RNA	2.2	2.26×10^{-4}	mica30		
Homologous DNA pairing and strand exchange	0.88	1.51×10^{-3}	mica30		
Presynaptic phase of homologous DNA pairing and strand exchange	0.88	1.51×10^{-3}	mica30		
Nuclear import of Rev protein	1.98	1.52×10^{-3}	mica30		
mRNA 3'-end processing	1.98	3.01×10^{-3}	mica30		
Post-Elongation Processing of Intron-Containing pre-mRNA	1.98	3.01×10^{-3}	mica30		
Transport of Mature Transcript to Cytoplasm	1.76	3.68×10^{-3}	mica30		
Polo-like kinase mediated events	0.66	5.53×10^{-3}	mica30		
Transport of Mature mRNA derived from an Intron-Containing Transcript	1.54	6.81×10^{-3}	mica30		
Recruitment of repair and signaling proteins to double-strand breaks	0.88	9.29×10^{-3}	mica30		
Interactions of Vpr with host cellular proteins	1.76	9.87×10^{-3}	mica30		
APC/C:Cdc20 mediated degradation of mitotic proteins				3.89	5.18×10^{-4}
Cdc20:Phospho-APC/C mediated degradation of Cyclin A				3.89	5.18×10^{-4}
Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins				3.89	5.93×10^{-4}
Meiotic Recombination				3.89	1.48×10^{-3}
Cyclin A:Cdk2-associated events at S phase entry				3.37	1.48×10^{-3}
Cyclin E associated events during G1/S transition				3.11	4.76×10^{-3}
Packaging Of Telomere Ends				2.59	5.17×10^{-3}
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1				3.37	5.75×10^{-3}
Meiosis				4.4	6.10×10^{-3}
p53-Independent G1/S DNA damage checkpoint				2.59	8.80×10^{-03}
p53-Independent DNA Damage Response				2.59	8.80×10^{-3}
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A				2.59	8.80×10^{-3}
G1/S DNA Damage Checkpoints				2.85	9.81×10^{-3}

Table 5: Pathway enrichment analysis for `ICA`, `DEGAS`, and MICA.