MICRORNA-AUGMENTED PATHWAYS (mirAP) AND THEIR APPLICATIONS TO PATHWAY ANALYSIS AND DISEASE SUBTYPING

DIANA DIAZ[1], MICHELE DONATO[3], TIN NGUYEN[1], SORIN DRAGHICI[1,2]

[1] *Department of Computer Science, Wayne State University,*
Detroit, MI 48202, U.S.A.
[2] *Department of Obstetrics and Gynecology, Wayne State University,*
Detroit, MI 48202, U.S.A.
[3] *Institute for Immunity, Transplantation and Infection, Stanford University Medical Center,*
Stanford, CA 94305, U.S.A.
E-mail: sorin@wayne.edu

MicroRNAs play important roles in the development of many complex diseases. Because of their importance, the analysis of signaling pathways including miRNA interactions holds the potential for unveiling the mechanisms underlying such diseases. However, current signaling pathway databases are limited to interactions between genes and ignore miRNAs. Here, we use the information on miRNA targets to build a database of miRNA-augmented pathways (mirAP), and we show its application in the contexts of integrative pathway analysis and disease subtyping. Our miRNA-mRNA integrative pathway analysis pipeline incorporates a topology-aware approach that we previously implemented. Our integrative disease subtyping pipeline takes into account survival data, gene and miRNA expression, and knowledge of the interactions among genes. We demonstrate the advantages of our approach by analyzing nine sample-matched datasets that provide both miRNA and mRNA expression. We show that integrating miRNAs into pathway analysis results in greater statistical power, and provides a more comprehensive view of the underlying phenomena. We also compare our disease subtyping method with the state-of-the-art integrative analysis by analyzing a colorectal cancer database from TCGA. The colorectal cancer subtypes identified by our approach are significantly different in terms of their survival expectation. These miRNA-augmented pathways offer a more comprehensive view and a deeper understanding of biological pathways. A better understanding of the molecular processes associated with patients' survival can help to a better prognosis and an appropriate treatment for each subtype.

1. Introduction

The identification of biological processes underlying conditions is crucial for disease prognosis and treatment programs. As gene signaling pathways are capable of representing complex interactions between genes, pathway databases have become essential for several gene expression analyses. Signaling pathway databases are remarkably important because they allow researchers to analyze high-throughput data in a functional context, reducing complexity and increasing the explanatory power. However, there are other molecules that play important roles in gene regulation, such as microRNAs, which are not included into current pathway databases. MicroRNAs (miRNAs) are small RNA molecules capable of suppressing protein production by binding to gene transcripts. In fact, more than 30% of the protein-coding genes in humans are miRNA-regulated. Additionally, miRNAs have been shown to play an important role in diagnosis and prognosis for different types of diseases[1].

The integration of miRNA into signaling pathways have multiple applications, such as pathway analysis and disease subtyping. Pathway analysis techniques and methods aim to analyze high-throughput data with the goal of identifying pathways that are significantly impacted by a given condition. The typical input of pathway analysis includes gene expression data from two different phenotypes (e.g., condition vs. control) and a set of signaling pathways. Although current pathway analysis methods using gene expression (mRNA) have achieved excellent results[2–4], mRNA expression alone is unable to capture the complete picture of biological processes, as other entities also play important roles. Relevant work has been done to elucidate the important interplay between miRNAs and biological pathways[5–9]. The state-of-the-art approach for miRNA-mRNA pathway analysis is microGraphite[8] which uses an empirical gene set approach. microGraphite's main goal is the identification of signal transduction paths correlated with the condition under study[10].

A second crucial process in the understanding of complex diseases is disease subtyping. Identifying clinically meaningful subtypes in complex diseases is crucial for improving prognosis, treatment, and precision medicine[11]. A typical input of disease subtyping consists of various clinical variables and gene expression data from patients affected by a particular disease. The expected output consists of well-identified groups of patients that highly correlate with one or more variables, such as observed survival (e.g., long-term vs. short-term survival patients). Disease subtyping is typically expressed as a clustering problem with the goal of partition patients in groups based on their genetic similarities with the additional complexity that the number of clusters is unknown. Several methods for disease subtyping using gene expression data have been developed[11–15]. Integrative analysis using clinical data, multi-'omics' data, and prior biological knowledge can leverage current disease subtyping methods.

In this paper, we present a tool for integrating miRNA into signaling pathways (mirIntegrator), a publicly available miRNA-augmented pathway database (mirAP), and we show the applications of such augmentation to pathway analysis and disease subtyping. We have used mirIntegrator previously as a part of our orthogonal meta-analysis approach[16].

Our pathway analysis pipeline uses mirAP and Impact Analysis[3,4], a topology-aware pathway analysis method previously developed by our group. To demonstrate the advantage of our method, we analyze 9 datasets studying 7 different diseases with mRNA and miRNA expression. We show that the proposed approach is able to identify the pathways that describe the underlying diseases as significant. The p-values and rankings of these pathways are significantly smaller than those obtained without data integration as well as when using microGraphite[8].

Our disease subtyping pipeline uses miRNA and mRNA expression data, available clinical variables, and prior biological knowledge. This method includes a feature selection approach based on mirAP to reduce the effective dimensionality of the unsupervised clustering problem. We analyze colorectal cancer miRNA, gene expression data, and clinical records downloaded from the Cancer Genome Atlas (TCGA) with our pipeline and SNF[15], a recently proposed integrative disease subtyping method. The colorectal cancer-relevant pathways and subgroups identified with our approach are significantly different in terms of their survival expectation, outperforming the approach that does not use miRNA, and providing information on biological mechanisms relevant to the difference in survival.

2. Methods

In this section, we propose an algorithm for integrating miRNA into signaling pathways. We also describe two pipelines using miRNA-augmented pathways (mirAP). The first pipeline is for pathway analysis (PA) and the second one is for disease subtyping (DS). The scenarios for these analyses are different. PA is used in biological studies comparing genetic samples from two different phenotypes (e.g., disease vs. control samples), and DS is used in studies with samples of patients undergoing the same disease for which the clinical subtypes are unknown. Our PA pipeline is able to integrate miRNA and mRNA expression data and identify pathways that are related to the disease under study. Our DS pipeline is able of incorporate biological pathways to partition patients into groups with very different survival patterns.

2.1. *Pathway augmentation*

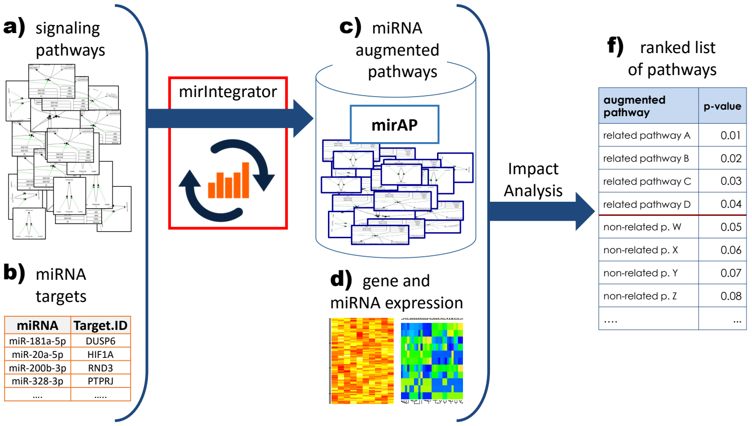
This method augments the graphical representation of original signaling pathways with interactions between miRNAs and their target genes. The input of this method includes a set of signaling pathways and known miRNA-mRNA interactions (Fig. 1a,b). The output is a set of augmented pathways that consists of the original genes, the miRNAs that target those genes and their interactions. Let $P = (V, E)$ denote the graphical representation of the original gene-gene pathway, and $T : M \to V$ a function that identifies the target genes of miRNAs in M. An edge $e \in E$ can be represented as a 3-tuple $e = (g_1, g_2, \mathit{interaction})$. We augment the nodes and edges of the original pathway as follows:

$$\bar{V} = V \cup \{m \in M | T(m) \cap V \neq \varnothing\}$$
$$\bar{E} = E \cup \{(m, g, \mathit{inhibition}) | m \in V \cap M \wedge g \in T(m)\}$$

We implemented this algorithm in R and published it as the Bioconductor package named mirIntegrator (`http://bit.ly/mirIntegrator`). mirIntegrator is flexible and allows users to integrate user-specific pathway databases with user-specific miRNA-mRNA target databases. Additionally, it generates graphical representations of the augmented pathways (see Fig. 5). We integrated pathways from Kyoto Encyclopedia of Genes and Genomes[17] (KEGG) (version 73) with miRNA targets from miRTarBase[18] (version 4.5) to generate mirAP, a database of miRNA-augmented pathways (`http://www.cs.wayne.edu/dmd/mirAP`).

2.2. *Integrative pathway analysis*

Our pathway analysis pipeline consists of two main steps. In the first step, we augment the signaling pathways with interactions between miRNAs and their targets. Once this is done, the data integration problem is mapped to the original pathway analysis problem for which existing methods can be applied. The difference is that here both miRNA and mRNA expression can be taken into consideration. In the second step, we apply any pathway analysis that uses fold change and p-value as input, e.g., Over-representation analysis[19] (ORA) and Impact Analysis[3,4]. ORA and Impact Analysis are well-known methods developed by our group to identify signaling pathways that are impacted by the effects of diseases. Fig. 1 displays the overall pipeline of our approach.


a) signaling pathways
b) miRNA targets
mirIntegrator
c) miRNA augmented pathways
mirAP
d) gene and miRNA expression
Impact Analysis
f) ranked list of pathways

miRNA	Target.ID
miR-181a-5p	DUSP6
miR-20a-5p	HIF1A
miR-200b-3p	RND3
miR-328-3p	PTPRJ
....	.....

augmented pathway	p-value
related pathway A	0.01
related pathway B	0.02
related pathway C	0.03
related pathway D	0.04
non-related p. W	0.05
non-related p. X	0.06
non-related p. Y	0.07
non-related p. Z	0.08
....	...

Fig. 1. Workflow of pathway analysis using augmented pathways.

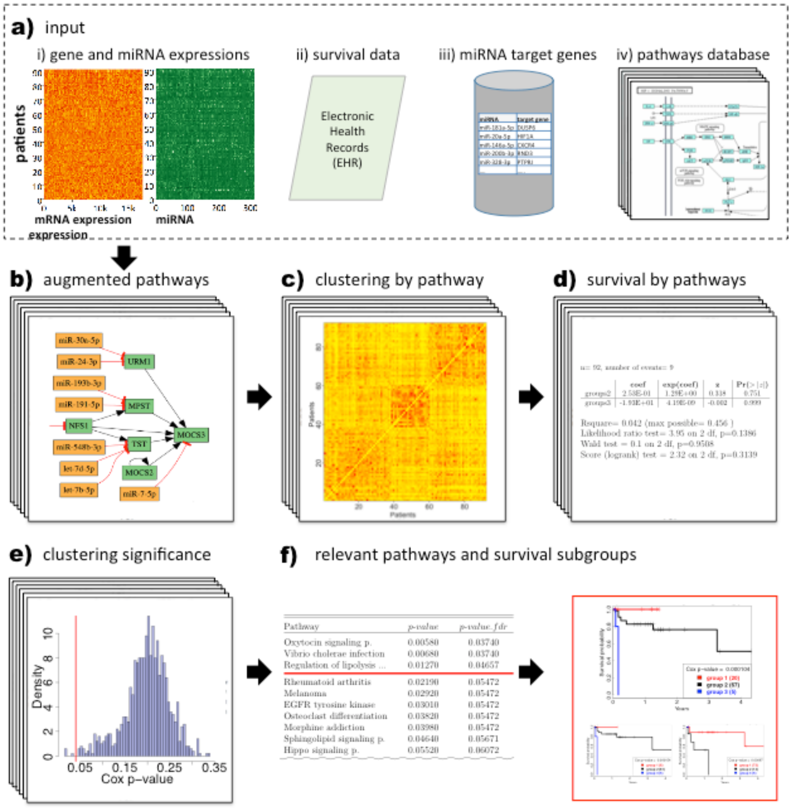
Impact Analysis[3,4] is a widely used topology-aware method that combines two types of evidence: i) the over-representation (ORA) of differentially expressed (DE) genes in a pathway[19], and ii) the perturbation (PERT) of such a pathway, as measured by propagating expression changes through the pathway topology. These two types of evidence are captured by two independent p-values[4]: p_{ORA} and p_{PERT}. These p-values are combined using Fisher's method to obtain a global p-value per pathway. Each global p-value represents the probability of having the observed number of DE genes, as well as the observed amount of impact just by chance (i.e. when the null hypothesis is true)[4]. To calculate p_{ORA} on mirAP, we assumed that the number of DE entities (genes and miRNAs) on the given pathway follows a hypergeometric distribution. The following information is needed to compute p_{ORA}: i) the total number of measured entities, ii) the number of entities belonging to the given augmented pathway, iii) the total number of DE entities, and iv) the number of DE entities in the given augmented pathway. To calculate p_{PERT} on mirAP, we perform a bootstrap procedure using the following input: i) the log-fold change of DE entities, and ii) the given augmented pathway.

2.3. *Integrative disease subtyping*

Our disease subtyping pipeline is presented on Fig. 2. The input includes: i) mRNA and miRNA sample-matched expression data, ii) survival records, iii) a database of miRNA-target gene interactions, and iv) a database of signaling pathways (see Fig. 2a). The output is a set of selected pathways (Fig. 2f) yielding to subtypes with significantly distinct survival patterns.

First, we obtain the miRNA-augmented pathways from mirAP (Fig. 2b). Second, we partition the patients using the genes and miRNAs provided by each augmented pathway (Fig. 2c). e.g., let us say that we want to analyze gene and miRNA expression from $\mathcal{N}$ number of patients and we obtained $\mathcal{P}$ number of augmented pathways from mirAP. Taking one pathway at the time, we filter the gene expression data by selecting only genes that belong to the pathway. Similarly, we filter the miRNA expression data by selecting only miRNAs that belong to the pathway. Now, we need to combine the filtered gene expression and miRNA data and then perform clustering on the combined data. So, we use Similarity Network Fusion method[15] (SNF) in conjunction with spectral clustering[20] for this purpose. We repeat this process with each pathway to obtain $\mathcal{P}$ different pathway-based clusterings, one per each pathway.

Third, we perform survival analysis on each of the pathway-based clusterings (Fig. 2d). In order to do this, we compute the log-rank test p-value (Cp) of Cox proportional hazards regression analysis by using the input survival information.This p-value represents how significant the difference between the survival curves is. For instance, a Cox log-rank test p-value close to zero may indicate that these groups have well-differentiated survival patterns. Now the question is whether we could obtain the same clustering just by chance[21]. To answer this question we use the random sampling technique. For example, if the pathway has G number of genes and m number of miRNAs, we randomly select G genes and m miRNAs from the measured values. Then, we partition the patients using this randomly selected set of entities and then compute its Cox p-value (rCp). We repeat this random selection a large number of times (e.g., 2,000 times) to construct an empirical distribution of Cox p-values (Fig. 2d). Next, we compare the observed Cox p-value Cp with the distribution of rCp, calculated from randomly selected genes and miRNAs. We estimate the probability of obtaining this Cp by computing the proportion of resampling p-values less than or equal to the observed Cp (e.g., In Fig. 2d the vertical red line indicates the observed Cp). For each pathway, we estimate this probability in order to quantify how likely it is to observe by chance a Cox p-value less than or equal to the one observed with the actual genes and miRNAs in the pathway.



```
n= 92, number of events= 9

         | coef      | exp(coef) | z      | Pr(>|z|)
groups2  | 2.53E-01  | 1.29E+00  | 0.338  | 0.751
groups3  | -1.93E+01 | 4.19E-09  | -0.002 | 0.999

Rsquare= 0.042 (max possible= 0.456 )
Likelihood ratio test= 3.95 on 2 df, p=0.1386
Wald test = 0.1 on 2 df, p=0.9508
Score (logrank) test = 2.32 on 2 df, p=0.3139
```

Pathway	p-value	p-value.fdr
Oxytocin signaling p.	0.00580	0.03740
Vibrio cholerae infection	0.00680	0.03740
Regulation of lipolysis ...	0.01270	0.04657
Rheumatoid arthritis	0.02190	0.05472
Melanoma	0.02920	0.05472
EGFR tyrosine kinase	0.03010	0.05472
Osteoclast differentiation	0.03820	0.05472
Morphine addiction	0.03980	0.05472
Sphingolipid signaling p.	0.04640	0.05671
Hippo signaling p.	0.05520	0.06072

Fig. 2. The proposed pipeline for disease subtyping.

The final step is to select the pathways that are relevant to survival, i.e., pathways yielding to significantly distinct survival curves. To do this, we adjust the p_i p-values for multiple comparisons using False Discovery Rate (FDR). We then rank the pathways by FDR.p-value and select those less than or equal to the significance threshold of 5% as *relevant pathways*. We note that this pipeline can be used in conjunction with other integrative clustering methods.

3. Results

In this section, we present the results of our pathway analysis and disease subtyping pipelines using the miRNA-augmented pathways (mirAP). First, we perform pathway analysis of 9 mRNA/miRNA sample-matched datasets using two different methods (Impact Analysis and ORA) and show that mirAP offers a significant improvement over analyzing mRNA data alone. We also compare the obtained results with the state-of-the-art method (microGraphite)[8]. Second, we perform disease subtyping of a colorectal cancer dataset from TCGA using our subtyping pipeline and compare with the traditional pipeline for subtyping.

3.1. *Validation of our pathway analysis pipeline*

We analyze nine sample-matched datasets from seven different diseases: GSE43592 (multiple sclerosis, 10 controls, 10 cases), GSE35389 (melanoma, 4 controls, 4 cases), GSE35982 (colorectal cancer, 8 controls, 8 cases), GSE26168 (type II diabetes, 8 controls, 9 cases), GSE62699 (alcoholism, 18 controls, 18 cases), GSE35834 (colorectal cancer, 23 controls, 55 cases), GSE43797 (pancreatic cancer, 5 controls, 7 cases), GSE29250 (non-small cell lung cancer, 6 controls, 6 cases), and GSE32688 (pancreatic cancer, 7 controls, 25 cases). For each of these datasets, we used the normalized expression values as found in GEO.[22] The microarray probes were annotated according to their corresponding platform's metadata using GEOquery.[23] Next, we estimated log-fold-change between disease and control groups by fitting to a gene-wise linear model using the R package limma[24]. We use the following two criteria to identify differentially expressed (DE) genes: i) genes with adjusted p-value lower than 5%, and ii) among the genes that satisfy the first criterion, we choose the genes with the highest log-fold change, up to 10% of measured genes. We use the same criteria to identify DE miRNAs.

The nine datasets were selected due to two important reasons. First, these datasets have both mRNA and miRNA measurements for the same set of patients. Second, for each of the underlying diseases, there is a KEGG pathway, henceforth *target pathway*, that was created to describe the underlying mechanisms of the disease. To demonstrate the advantage of the miRNA data integration, we compare the use of the original KEGG pathways with the use of our miRNA augmented pathways (mirAP) by performing two pathway analysis methods that use p-value and fold-change: Impact Analysis (IA)[4] and over-representation analysis (ORA)[19]. The input for IA and ORA using KEGG is mRNA expression data. The input for IA and ORA using mirAP includes both mRNA and miRNA expression data. The output of each method is a list of p-values – one per pathway. These p-values are adjusted for multiple comparisons using False Discovery Rate (FDR)[25].

We also analyze the nine GEO datasets using microGraphite[8] after quantile normalization to compare with our pipeline. The main goal of microGraphite is the identification of signal transduction paths correlated with the condition under study. It is implemented in a four-steps recursive procedure as follows: (i) selection of pathways, (ii) best path identification, (iii) metapathway construction, and (iv) metapathway analysis. Here we only consider the first step of the approach, which is the selection of significant pathways. This selection is based on the significance levels obtained from the test on the mean of the pathways (alpha-mean). The input is the mRNA and miRNA expression data and it does not take in account fold-changes nor differentially expressed entities.

For each dataset, we expect a good method to identify the target pathway as significant, as well as to rank it on top. For instance, in the colorectal cancer dataset which compares colorectal cancer tissue vs. normal, the *Colorectal cancer pathway* must be shown as significant and should be as close to the top of the ranking as possible since this is the pathway that describes the phenomena involved in colorectal cancer. Based on this, we compare the rank and p-value of the target pathway in each disease using the five methods: i) mRNA expression alone using standard KEGG pathways with ORA and ii) IA, iii) mRNA and miRNA expression data using the augmented pathways (mirAP) with iii) ORA and iv) IA, and v) mRNA and miRNA expression data analyzed with microGraphite.

Table 1. Results of target pathway identification using traditional ORA (column 3), traditional IA (col. 4), ORA on mirAP (col. 5), IA on mirAP (col. 6), microGraphite (col. 7)

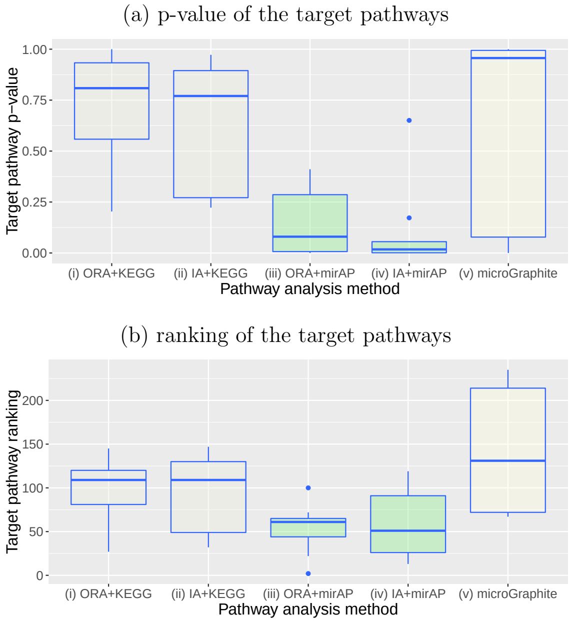
GEO ID	Target pathway	ORA	IA	ORAmir	IAmir	microGraphite
GSE26168	Type II diabetes mellitus	no	no	no	no	yes
GSE29250	Non-small cell lung cancer	no	no	yes	no	no
GSE35982	Colorectal cancer	no	no	no	no	no
GSE32688	Pancreatic cancer	no	no	yes	yes	no
GSE35389	Melanoma	no	no	yes	yes	no
GSE35834	Colorectal cancer	no	no	yes	yes	no
GSE43592	Amyotrophic lateral scle.	no	no	no	yes	no
GSE43797	Pancreatic cancer	no	no	yes	yes	yes
GSE62699	Alcoholism	no	no	no	yes	no

Table 1 shows the target pathways and their significance for the 9 datasets. The first and second columns display the datasets and their corresponding target pathways while the other five columns indicate whether the target pathways are identified as significant using the five methods: ORA of mRNA expression on KEGG pathways (ORA+KEGG), IA of mRNA expression on KEGG (IA+KEGG), ORA of miRNA and mRNA expression data on mirRNA-augmented pathways (ORA+mirAP), our approach IA of miRNA and mRNA expression on mirAP (IA+mirAP), and miRNA and mRNA expression analysis using microGraphite, respectively. The significance threshold is 5% for FDR p-values. IA and ORA fail to identify any target pathway as significant when using only mRNA whereas our approach (IA+mirAP) correctly identify the target in 6 out of 9 datasets (GSE32688, GSE35389, GSE35834, GSE43592, GSE43797, GSE62699) and ORA+mirAP correctly identify the target pathway as significant in 5 out of 9 datasets (GSE29250, GSE32688, GSE35389, GSE35834, GSE43797). microGraphite correctly identifies the target pathway as significant in only 2 out of 9 datasets (GSE26168, GSE43797). The results demonstrate that our integration of mRNA and miRNA lifts the statistical power for both pathway analysis techniques (ORA and IA) and outperforms microGraphite in target pathway identification.

Fig. 3 shows the p-values and rankings of the target pathways using the five methods. The panel (a) shows the FDR corrected p-values of the target pathways. We compare the lists of p-values using Wilcoxon test. The FDR p-values produced by IA+mirAP are significantly smaller than by IA+KEGG (p=0.007), ORA+KEGG (p=0.005), and microGraphite (p=0.009).

The panel (b) shows the rankings of the target pathways. Again, the rankings produced by IA+mirAP are significantly smaller than those of IA+KEGG (p=0.03 using t-test, and p=0.04 using Wilcoxon test), ORA+KEGG (p=0.03 using t-test and p=0.04 using Wilcoxon test), and microGraphite (p=0.0051 using t-test and p=0.0058 using Wilcoxon test). This confirms that our augmented pathways, mirAP, improve the performance of traditional Impact Analysis and ORA. Also, the results show that the proposed integrative pathway analysis also outperforms microGraphite in terms of both p-values and rankings for target pathway identification.

Furthermore, our pathway database (mirAP) is generated with validated miRNA-mRNA interactions, while microGraphite uses predicted interactions, which increases the number of false positive miRNA-target interactions. Another drawback of microGraphite is it execution time. A typical analysis with microGraphite takes approximately 22 hours while our approach takes only a few minutes. We ran these experiments on a typical desktop workstation with a 2.6 GHz Intel Core i5, 8GB of RAM, on a single thread, and the OS X 10.11 operative system.


(a) p-value of the target pathways

(b) ranking of the target pathways

Fig. 3. Corrected p-values and rankings of the target pathways using different methods.

3.2. *Validation of our disease subtyping pipeline*

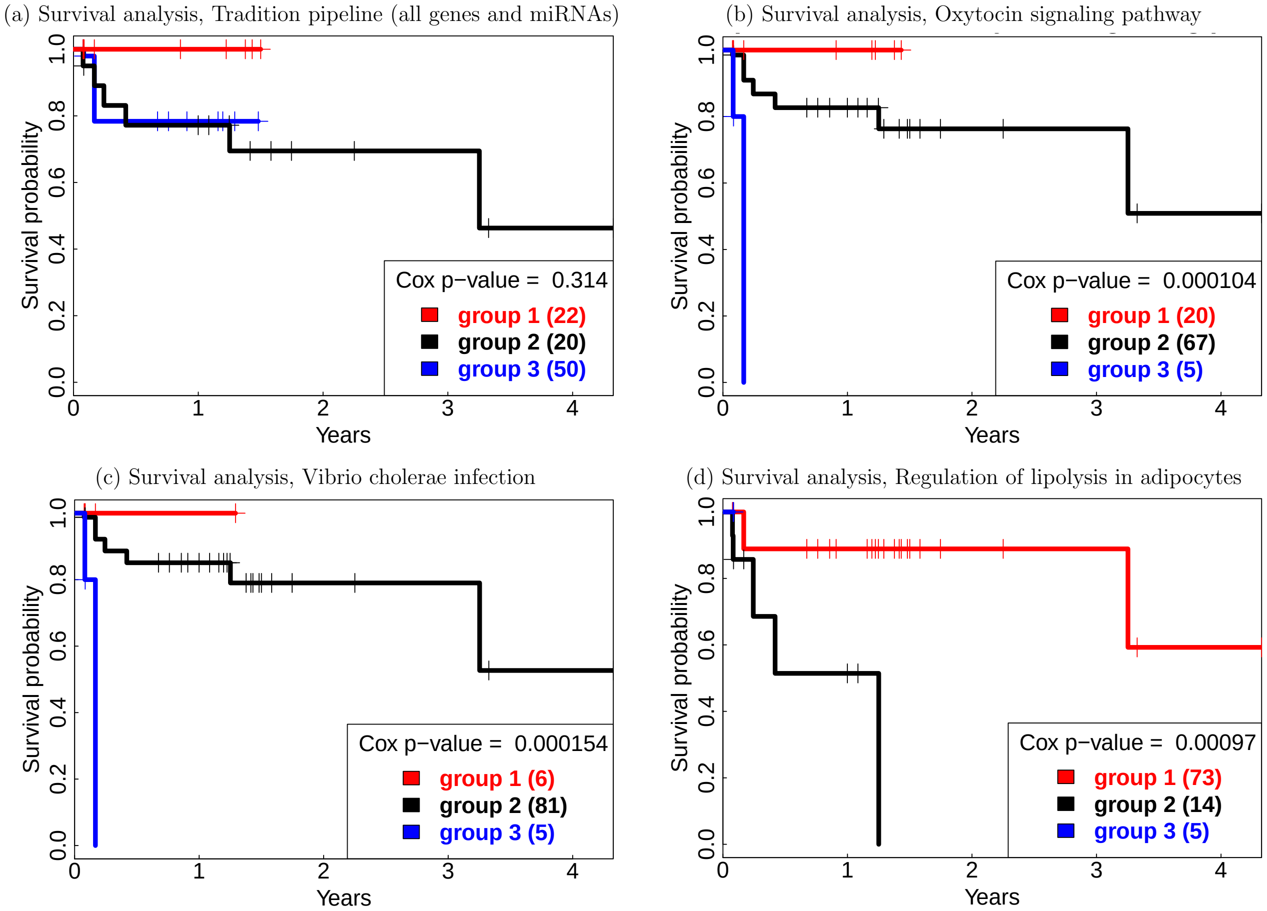
To assess our disease subtyping pipeline we use matched-sample gene and miRNA expression data (level 3 from platforms Agilent G4502A-07 and Illumina GASeq miRNASeq, respectively) of colorectal cancer patients (COAD) downloaded from the Cancer Genome Atlas (TCGA) (`cancergenome.nih.gov`). We selected the largest set of patients with miRNA-mRNA matched samples and available survival records, as were selected in SNF[15]. The number of patients is $M = 92$, the number of genes is $N_g = 17,814$, and the number of miRNAs is $N_m = 705$. We performed unsupervised clustering with the number of clusters set as $k = 3$ according to prior knowledge of the number of subtypes of COAD[15]. We use SNF[15] in conjunction with spectral clustering[20] as integrative clustering method. To perform SNF clustering, we used the SNFtool package with the suggested parameters.

For each miRNA-augmented pathway, our method partitions the patients using the genes and miRNAs in the pathway as clustering features, resulting in a total of 184 clusterings. Then for each pathway-based clustering, we construct the empirical distribution and then estimated the *p-value* of how likely the pathway helps to improve disease subtyping. The *p-values* of the relevant pathways are shown in Table 2. We select the pathways with a FDR-corrected *p-value* ≤ 0.05 as *relevant pathways*. The horizontal red line represents the significance cutoff at 5%. For TCGA-COAD, we identify three relevant pathways: *Oxytocin signaling pathway*, *Vibrio cholerae infection*, and *Regulation of lipolysis in adipocytes*.

We also cluster the 92 patients using SNF with the traditional pipeline, i.e., using all the measured genes and miRNAs. We compare these partitions with those obtained by our pipeline. To assess the correlation between the obtained groups and survival patterns (e.g., long-term vs. short-term survival), we performed survival analysis for all the cases using Kaplan-Meier analysis.

Table 2. List of relevant pathways for colorectal subtyping.

Pathway	*p-value*	*p-value.fdr*
Oxytocin signaling pathway	0.00580	0.0374
Vibrio cholerae infection	0.00680	0.0374
Regulation of lipolysis in adipocytes	0.01270	0.0466
Rheumatoid arthritis	0.02190	0.0547
...	...	...



(a) Survival analysis, Tradition pipeline (all genes and miRNAs)

(b) Survival analysis, Oxytocin signaling pathway

(c) Survival analysis, Vibrio cholerae infection

(d) Survival analysis, Regulation of lipolysis in adipocytes

Fig. 4. Kaplan-Meier survival analysis of the obtained COAD subtypes. a) Survival curves using all genes and miRNAs. b), c), and d) Survival curves using relevant pathways.

Fig. 4 shows the Kaplan-Meier plots, each one represents the association of the obtained groups with the observed patient survival. Fig. 4a shows the subtypes obtained with the traditional pipeline using all $17,814$ genes and 705 miRNAs. In a Cox proportional hazards regression analysis, we find that there is no statistically significant difference between survival groups obtained with the traditional pipeline (log rank test p-value $= 0.314$). Fig. 4b, c, and d. shows the resultant clustering on the relevant pathways identified with our approach (Table 2). Clustering based on *Oxytocin signaling pathway* entities gives a log rank test p-value of 0.000104, which indicates a significant difference between the survival curves (Fig. 4b). Similarly, clusterings based on *Vibrio cholerae infection* and *Regulation of lipolysis in adipocytes* augmented pathways indicate significant differences between the survival curves with p-values of $p = 0.000154$ and $p = 0.00097$, respectively (Fig. 4c and d). As we can see, integrative clustering based on relevant mirAP pathways produce subtypes significantly more related to survival data than the traditional subtyping pipeline (approximately 1000 times lower p-values).

Given that our approach requires resampling for computing the pathways' significance (*p-values*), our pipeline is more time consuming than the traditional pipeline. For the computational experiments presented here, we generated $2,000$ random clusterings per each pathway. Our pipeline took some hours to subtype the set of patients (approximately 4 hours) while running SNF alone takes only some minutes (less than 3 minutes).

3.2.1. *Biological Significance of relevant Signaling Pathways*

Our pipeline identifies the *Oxytocin signaling pathway* to be related to the survival subtyping of colorectal cancer patients ($p = 0.000104$). Oxytocin (OXT) is a hormone with a well-known effect on uterine smooth muscles and myoepithelial cells. Additionally, it has been shown that oxytocin is expressed along the entire human gastrointestinal (GI) tract, including colon, and it contributes to the control of the GI motility[26]. Moreover, studies have shown that exposure to OXT leads to a significant decrease in cell proliferation for some epithelial cancer cells (e.g., breast and prostate cancer)[27]. In contrast, OXT has a growth-stimulating effect in other types of cancer cells (e.g., small-cell lung cancer, endothelial cancer, and Kaposiâs sarcoma)[28,29]. We think that the evidence of OXT expression on colon and the dual role that OXT has in some cancer cells (as inhibitor and promoter of cancer cells proliferation) may indicate that OXT could also play an important role in differentiating short and long-term survival COAD patients. In addition, OXT is also known to be capable of mitigating symptoms caused by stress, OXT levels increase in acute(short-lived) stress and decrease during chronic stress. Also, it is well-known that chronic stress has an outstanding role in cancer growth and metastasis.[30] From this, we also hypothesize that patients in the short term survival group (Fig. 4b, gr. 3) may have been in a metastatic stage with chronic stress and different OXT expression than patients in the other groups (Fig. 4b,1-2).

Similarly, we identify *Vibrio cholerae infection* pathway as relevant. This pathway describes the colonization of the intestine by Vibrio cholerae bacteria (VC). The main factor involved in this process is Cholera toxin (CTX). Several studies have exhibit relations between gastrointestinal tract bacteria and colon cancer progression. In particular, it has been shown that CTX suppresses carcinogenesis of inflammation-driven sporadic colon cancer[31].

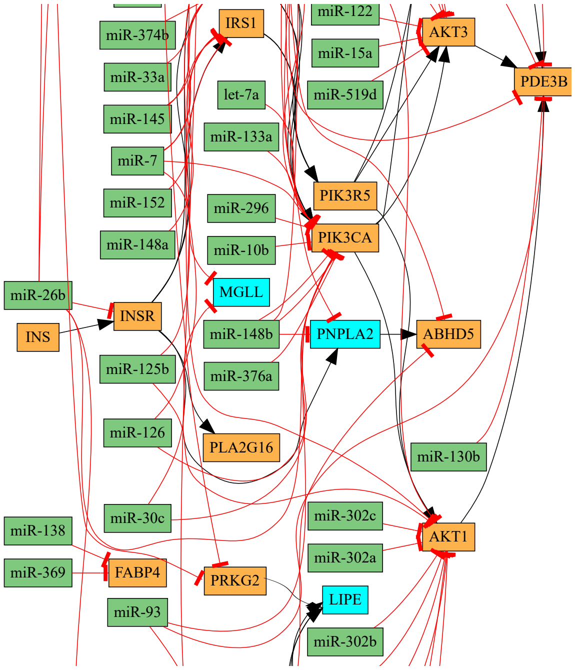


Fig. 5. Portion of the miRNA-augmented *Regulation of lipolysis in adipocytes* pathway.

Ultimately, the *Regulation of lipolysis in adipocytes* pathway describes a unique function of white adipose tissue in which triacylglycerols (TAGs) are broken down into fatty acids and glycerol. Fatty acid (FA) pathways play an important role in cancer[32]. In particular, increased gene expression of AGPAT9(PNPLA2), MAGL(MGLL), and HSL(LIPE), FA metabolism regulators, is associated with increased cancer cells proliferation in colorectal cancer[32] (see blue boxes in Fig. 5). By instance, MAGL pharmacological inhibition attenuated aggressiveness of colorectal cancer cells. On the other hand, decreased gene expression of CD36/FAT regulator has been implicated in contributing to colorectal cancer progression, a higher metastasis grade, and low relapse-free survival[33].

Fig. 5 shows a portion of the *Regulation of lipolysis in adipocytes* augmented pathway obtained from our database (see the complete pathway at `http://bit.ly/hsa04923`).The green boxes show the protein coding genes while the orange boxes display the miRNAs. The black arrows denote activation and the red bar-headed arrows denote repression.

4. Discussion

In this article, we present a method to augment signaling pathways with miRNA-target interactions. The miRNA-augmented pathways (mirAP) offer a more comprehensive view and a deeper understanding of complex diseases. We also present two pipelines that use mirAP to integrate miRNA and mRNA expression data for the purpose of pathway analysis and disease subtyping. As miRNA expression data are becoming freely accessible, miRNA-mRNA integrative analyses are likely to become a routine.

Our pathway analysis pipeline augments gene-gene signaling pathways with miRNA-target interactions. Then we perform a topology-based pathway analysis that takes into consideration both types of molecular data. We analyze 9 sample-matched datasets that were assayed in independent labs. Our pipeline outperforms traditional methods in identifying target pathways (smaller p-values and rankings of the target pathways). We plan to explore methods for augmenting the pathways using only the process(es) described by each given pathway.

Our disease subtyping pipeline combines gene and miRNA expression data, clinical records, and mirAP. The contribution of our disease subtyping pipeline is two-folds. First, this framework introduces a way to exploit the additional information available in biological databases and integrates clinical data, miRNA and gene expression data for disease subtyping. Second, it identifies pathways associated with survival differentiated subgroups of diseases, which bring us closer to the identification of causal pathways associated with survival. We analyze a colorectal cancer data downloaded from TCGA. Our framework provides pathways relevant to survival patterns and subtypes significantly difference between the survival curves. It greatly improves the former approach with p-values $1,000$ times lower than the former. This pipeline is limited by the availability of datasets containing survival records, miRNA, and mRNA expression matched-samples. We plan to extend this study by investigating more diseases and larger datasets.

Acknowledgments

This work was supported by the National Institutes of Health [R01 DK089167, R42 GM087013]; National Science Foundation [DBI-0965741]; and the Robert J. Sokol Endowment in Systems Biology. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Y. S. Lee and A. Dutta, *Annual Review of Pathology* **4** (2009).
2. P. Khatri, M. Sirota and A. J. Butte, *PLoS Computational Biology* **8**, p. e1002375 (2012).
3. S. Drăghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu and R. Romero, *Genome Research* **17**, 1537 (2007).
4. A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, *Bioinformatics* **25**, 75 (2009).

5. C. Backes, E. Meese, H.-P. Lenhof and A. Keller, *Nucleic Acids Research* **38**, 4476 (July 2010).
6. J. B.-K. Hsu, C.-M. Chiu, S.-D. Hsu, W.-Y. Huang, C.-H. Chien, T.-Y. Lee and H.-D. Huang, *BMC Bioinformatics* **12**, p. 300 (July 2011).
7. I. S. Vlachos, N. Kostoulas, T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas and A. G. Hatzigeorgiou, *Nucleic Acids Research* **40**, W498 (July 2012).
8. E. Calura, P. Martini, G. Sales, L. Beltrame, G. Chiorino, M. D'Incalci, S. Marchini and C. Romualdi, *Nucleic Acids Research* **42**, p. e96 (2014).
9. S. Nam, M. Li, K. Choi, C. Balch, S. Kim and K. P. Nephew, *Nucleic Acids Research* **37**, W356 (May 2009).
10. P. Martini, G. Sales, M. S. Massa, M. Chiogna and C. Romualdi, *Nucleic Acids Research* **41**, e19 (2013).
11. S. Saria and A. Goldenberg, *IEEE Intelligent Systems* **30**, 70 (2015).
12. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science* **286**, 531 (October 1999).
13. T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler *et al.*, *Proceedings of the National Academy of Sciences* **100**, 8418 (2003).
14. P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag *et al.*, *Breast Cancer Research* **10**, p. R65 (2008).
15. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nature Methods* **11**, 333 (2014).
16. T. Nguyen, D. Diaz, R. Tagett and S. Draghici, *Nature Scientific Reports* **6**, p. 29251 (2016).
17. M. Kanehisa and S. Goto, *Nucleic acids research* **28**, 27 (2000).
18. S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho *et al.*, *Nucleic Acids Research* **42**, D78 (January 2014).
19. S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier and S. A. Krawetz, *Genomics* **81**, 98 (2003).
20. U. Von Luxburg, *Statistics and Computing* **17**, 395 (2007).
21. E. Czwan, B. Brors and D. Kipling, *BMC Bioinformatics* **11**, p. 19 (2010).
22. T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi and R. Edgar, *Nucleic Acids Research* **33**, D562 (2005).
23. S. Davis and P. Meltzer, *Bioinformatics* **14**, 1846 (2007).
24. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, *Nucleic Acids Research* **43**, e47 (April 2015).
25. Y. Benjamini and D. Yekutieli, *Annals of Statistics* **29**, 1165 (August 2001).
26. B. Ohlsson, M. Truedsson, P. Djerf and F. Sundler, *Regulatory Peptides* **135**, 7 (July 2006).
27. A. Reversi, V. Rimoldi, T. Marrocco, P. Cassoni, G. Bussolati, M. Parenti and B. Chini, *Journal of Biological Chemistry* **280**, 16311 (April 2005).
28. P. Cassoni, T. Marrocco, S. Deaglio, A. Sapino and G. Bussolati, *Annals of Oncology* **12**, S37 (January 2001).
29. C. Pqueux, B. P. Keegan, M.-T. Hagelstein, V. Geenen, J.-J. Legros and W. G. North, *Endocrine-Related Cancer* **11**, 871 (December 2004).
30. M. Moreno-Smith, S. K. Lutgendorf and A. K. Sood, *Future Oncology* **6**, 1863 (December 2010).
31. M. Doulberis, K. Angelopoulou, E. Kaldrymidou, A. Tsingotjidou, Z. Abas, S. E. Erdman and T. Poutahidis, *Carcinogenesis* **36**, p. bgu325 (December 2014).
32. S. Balaban, L. S. Lee, M. Schreuder and A. J. Hoy, *BioMed Research International* **2015**, p. 274585 (2015).
33. S. M. Rachidi, T. Qin, S. Sun, W. J. Zheng and Z. Li, *PLOS ONE* **8**, p. e57911 (March 2013).