DOI: 10.14744/ejmo.2022.53376
EJMO 2022;6(2):156–171

Research Article

A Multi-Molecular Fusion to Detect Transcriptomic Signature in Tissue-Specific Cancer

Suparna Saha,[1] Saurav Mallik,[2] Sanghamitra Bandyopadhyay[3]

[1]SyMeC Data Center, Indian Statistical Institute, Kolkata, India
[2]Department of Environmental Epigenetics, Harvard T H Chan School of Public Health, Boston, USA
[3]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

Abstract

Objectives: Analysis of multi-molecular interactions and detection of combinatorial transcriptomic signatures are emerging as important research topics in disease analytics. Currently, a combination of gene and miRNA expression profiling in bioinformatic analysis enables us to comprehensively detect molecular changes in cancer and thereafter to identify integrated signatures and pathways that exist in the miRNA and gene interaction networks. Although many methodologies and applications have been suggested in recent literature, efficient techniques that can integrate the complex gene as well as miRNA expression profiles, and identify the most relevant signatures are required.

Methods: In this article, we presented a new framework of multi-molecular data integration to identify combinatorial transcriptomic signatures through the strategy of unsupervised learning and target detection. Later, we evaluated their utility in survival analysis through a multi-variate Cox regression study. We used a cervical cancer data repository to conduct our experiment. To construct the miRNA-mRNA interaction network, we selected the downregulated mRNAs that were negatively correlated with the upregulated miRNAs. Thereafter, we identified dense modules by using an unsupervised learning technique. The silhouette index value was computed for each cluster.

Results: By considering the network centrality of each molecule belonging to each cluster we identified top 3 combined signatures We also highlighted cluster-2 (hsa-mir-944, CFTR, GABRB2, HNF4G, TAC1, and C7orf57) for its high cohesiveness and contained a combined signature. We then applied three well-known classifiers (viz., SVM, KNN, and random forest) using 10-fold cross-validation, and obtained a high AUC score for cluster-2. Finally, we conducted a survival study with each molecule of the same cluster.

Conclusion: Finally, we conducted a survival study with each molecule of the same cluster. Our proposed combined signature detection strategy can determine the signature(s) for any microarray or RNA-Seq profile. The code is available at https://github.com/sahasuparna/DeMoS

Keywords: Co-expression, Limma, Combinatorial transcriptomic fused signature, Disease classification, miRNA target gene, survival study

Cite This Article: *Saha S, Mallik S, Bandyopadhyay S. A Multi-Molecular Fusion to Detect Transcriptomic Signature in Tissue-Specific Cancer. EJMO 2022;6(2):156–171.*

Identifying key molecular signatures[1-3] from a genomic profile has become a growing concern in biomedical research in the last couple of decades. A gene signature[4,5] is defined as a single gene or set of genes of a cell consisting of a distinctive gene expression pattern owing to modified pathogenic conditions or biological processes.[6] The gene signature offers numerous benefits in various cancers and their prognosis. According to Chanrion et al.[7], gene signatures can not only predict the relapse of primary breast cancers treated with tamoxifen but also help in the therapeutic management of estrogen receptor (ER)-positive cancers. Invasiveness gene signatures are significantly expressed genes that are assessed

Address for correspondence: Saurav Mallik, PhD. Department of Environmental Epigenetics, Harvard T H Chan School of Public Health, Boston, USA
Phone: (+1)617-432-0478 **E-mail:** sauravmtech2@gmail.com, smallik@hsph.harvard.edu

Submitted Date: March 26, 2022 **Accepted Date:** May 28, 2022 **Available Online Date:** June 06, 2022

for their relationship with overall and metastasis-free survival in patients with breast or other types of cancer.[8] Gene signatures are also identified as pathological factors that provide prognostic information in ER-positive breast cancers, and in the early stage of the disease, they help decide whether a patient will need supportive chemotherapy.[9] Gene signatures might be targeted by various miRNAs, which are non-coding RNAs (approximately 18{25 nucleotides long) that participate in the post-transcriptional regulation of gene expression.[10] These miRNAs are abnormally expressed in various malignancies as tumor suppressors or oncogenes.[11] They regulate various processes in carcinogenesis such as metastasis.[12] and cell proliferation.[13] Hence, miRNAs are promising markers in the diagnosis, prognosis, and therapy of cancer. MiRNAs regulate gene expression by binding to partially complementary sites in the target mRNAs.[14] Dysregulation of miRNAs is responsible for the formation and progression of tumors.[15,16] Cervical cancer, caused due to the alteration of cells in the cervix, is the leading cause of death in women, second only to breast cancer. This cancer, which is prone to metastasis, is difficult to diagnose in the early stages, and is tough to operate in the advanced stages when it is usually detected. Thus, gene signatures are useful for the early detection and treatment of cervical cancer by understanding the molecular level mechanisms underlying its progression. Errors in statistical analysis are one of the most complex issues in this decade. There are two types of hypothesis in statistical analysis, namely, the null hypothesis and alternative hypothesis. The null hypothesis denotes no significant difference between the mean of the diseased and control groups, whereas the alternative hypothesis signifies a significant difference between the mean of these two groups.[17] An appropriate statistical test is required to determine differentially expressed transcripts among samples. The linear model for microarray (Limma) package based on the empirical Bayes test is useful for all sizes and types of data distribution (normal or non-normal distribution) for RNA-Seq or similar type of data.[18-19] In this era of social networking and biomedical engineering, handling big data is challenging because of not only its size but also the heterogeneity with high dimensions, and other complicated relationships. Due to such challenges, network analysis is important because the network is a powerful way to represent complex relationships among a large number of objects. In biomedicine, a network is a convenient place such as the regulatory network, GCN[20], and protein-protein interaction network.[21]

In our paper, we introduce a new framework to determine dense module-based combinedsignatures. We identified the high cohesive cluster and discussed their application in a prognosis survival study. Most of the state-of-the-art approaches used hierarchical clustering that avoids overlapping modules. To overcome this limitation, we developed a novel framework to determine the potential dense module-based signatures by using network based model. First, we identified the predicted target genes corresponding to the differentially expressed miRNAs. After extracting the common genes between the significantly expressed genes and the predicted target genes. For the network construction we used the overexpressed miRNAs with their negatively correlated downregulated target genes. In addition, we conducted a comparative study between the high cohesive cluster obtained using our method and miRNA and mRNA signatures in separation, obtained using our method. As the aforementioned cluster produced a significant p-value for the prognosis survival analysis, it could produce a clinically promising signature. Moreover, our proposed method is highly effective in identifying the top 3 combined molecular signatures. Since the last two decades, gene signatures are widely used in omics data analysis. In this article, we propose a framework that can identify integrated modules and discuss their application in a prognosis study. In future research, we will consider the application of epigenetics (viz., methylation) to the existing framework. A recent study states that the epigenome contains contiguous regions denoted as differentially methylated regions (DMRs) that are significantly associated with numerous diseases[22-23]. Mallik et al.[24] conducted a comparative study of different DMR-finding methods; the results of this study might motivate our future work. Several types of gene signatures exist in the bioinformatics field, such as the prognostic gene signature, diagnostic gene signature, and predictive gene signature. The term `prognostic' signifies the prediction of the expected development (duration, description, and function) of the course of a disease. Hence, the prognostic gene signature is vital to the overall outcome of a disease, irrespective of therapeutic interference.

These prognostic signatures are useful in several tissue-Specific cancers such as hepatocellular carcinoma [25], leukemia [26], and breast cancer.[27] The diagnostic gene signature acts as a biomarker that differentiates the severity of phenotypes of analogous therapeutic conditions into the mild, moderate, or severe stage based on an inception point.[28]. A predictive gene signature predicts the outcome of therapeutic intervention and does not depend on the prognosis.[29] Hence, these signatures contain crucial information. Several highly efficient biological networks can be used to predict the new functionality of genes.[30] One of the most popular biological networks is GCN, where each node in the network denotes a gene. Based on the edge between the two genes (nodes in GCN) of the network, GCNs are of two types, namely, unweighted-GCNs (UGCNs) and weighted-GCNs (WGCNs). In a UGCN, a threshold value is applied to the correlation coefficient. If the correlation coefficient value is higher than the threshold value, an edge must exist between two genes; otherwise, no edge is made. In a WGCN, the result depends on the choice of the threshold of the correlation coefficient. Hence, the WGCN is a preferred network, where an edge exists between every pair of nodes, and the weights of the edges are determined by the correlation values between the corresponding nodes. Various techniques have been currently developed for multi-omics integration. Weighted connectivity measure integrating co-methylation, co-expression and protein-protein interactions (WeCoMXP), based on the weighted connectivity measure, is an approach for integrating multi-omics data from the weighted normalized gene regulatory network to detect gene modules.[31] WeCoMXP is the most promising integration technique in which tri-omics profiles (expression data, methylation data, and protein-protein interactions) and hierarchical clustering are used to identify gene modules. Functional gene module detection (FGMD)[32] is another method for determining gene modules. In a previous study on FGMD, single omics (gene expression data) data were used from two platforms, RNA-Seq and microarray, for the same gene pairs and paired samples. First, the log2 ratios between the tumor samples and the average of normal samples were calculated. To identify functional gene modules, gene expression data of the microarray and RNA-Seq platforms were compared using the Pearson correlation coefficient (PCC) for the same gene pairs. Then, the gene expression network was constructed using the PCC values, and the gene pairs having higher PCC values were extracted. After extraction, hierarchical clustering was applied to the selected gene pairs. The analogous modules based on the overlap ratio were combined using hierarchical clustering with the dynamic tree cut method.[33] Finally, the FGMD modules were identified. The double-label propagation clustering algorithm (DLPCA) is a new algorithm for determining disease-associated modules by using the gene expression dataset. In a previous study, DLPCA used the pathogenic records of genes as the properties of nodes from the GCN constructed using the WGCN analysis (WGCNA) tool.[34] Gene modules were constructed by applying a multilabel clustering algorithm, followed by the DLPCA. During the module detection phase, the DLPCA classified the corresponding modules into diseased and non-diseased samples.[35]

Several well-known algorithms are used to identify a gene module, and many of them are developed to extract network modules using a hierarchical clustering algorithm.[36,4] Scope exists for identifying stronger gene signatures that can exhibit higher classification performance of class labels and provide higher biological significance and validation. In the WGCNA method, hierarchical clustering and dynamic tree cut are used to discover the densely connected gene module. The PCC values are transformed by considering the recently evaluated power, and then, the topological overlap measure (TOM)[37] is computed on the basis of the power of the PCC values. The transcript modules are identified on the basis of the dissimilarity score, that is, (=1-TOM score). The TOM value between two vertices (e.g., transcripts) in an adjacency matrix is defined as follows.

$$\mathrm{TOM}(j,k)=\begin{cases}\dfrac{\sum\limits_{h\neq j,k} Y(j,v)Y(k,v)+Y(j,k)}{min\{\sum\limits_{h\neq l} Y(j,h),\sum\limits_{h\neq k} Y(k,h)\}-Y(j,k)+1}, & \text{if } j\neq k,\\ 1, & \text{if } j=\mathrm{k}.\end{cases} \tag{1}$$

By observing formula (3), the generalization of TOM can be expressed as follows.

$$\mathrm{GTOM1}(j,k)=\begin{cases}\dfrac{|nbd_1(j)\cap nbd_1(k)|+Y(j,k)}{min\{|nbd_1(j)|,|nbd_1(k)|\}-Y(j,k)+1}, & \text{if } j\neq k,\\ 1, & \text{if } j=\mathrm{k}.\end{cases} \tag{2}$$

Here, Y denotes the adjacency matrix of the nodes containing the first nearest neighbours of each nodes, '|.|' symbol indicates the cardinality of the set, $\mathrm{nbd}_1(j)$ signifies the set of neighbours of j excluding itself (i.e., *j*), and $\mathrm{nbd}_1(\mathrm{k})$ signifies the set of neighbours of k excluding itself (i.e., k), whereas $|\mathrm{nbd}_1(j)\cap\mathrm{nbd}_1(k)|$ refers to the number of common neighbours which are shared by the nodes j and k. A gene involved in more than one function might also exist in a different gene subnetwork and thus in different functions. However, with hierarchical clustering, the overlaps between different subnetworks are avoided. In this study, we used a graph based model by detecting high cohesive clusters to overcome this issue. We collected data from The Cancer Genome Atlas (TCGA), and then found significantly expressed genes using the Limma-voom tool. Thereafter, we identified the target genes for the significantly expressed miRNAs. We only considered these significantly expressed target genes for the analysis. The spectral clustering algorithm was then applied to find the dense modules, and the silhoette index was calculated for each cluster. For the purpose of emperical analysis we consider the cluster 2 as mentioned in Figure 4. We used three consecutive classifiers to classify the class labels by using all the features belonging to the signature. Compared with the other signatures, mirna-mrna signature produced the highest accuracy across all the classifiers. Additionally, we conducted a gene enrichment analysis [Gene Ontology (GO) and pathway] to identify disease-related pathways as well as GO terms for the participating genes belonging to the signature. Our framework may prove useful for extracting integrated signatures for other microarrays/ RNASeq datasets for cancer or any other disease. Since the last two decades, gene signatures are widely used in omics data analysis. In this article, we propose a framework that can identify integrated modules and discuss their application in a prognosis study. In future research, we will consider the application of epigenetics (viz., methylation) to the existing framework. A recent study states that the epigenome contains contiguous regions denoted as differentially methylated regions (DMRs) that are significantly associated with numerous diseases.[22,23] Mallik et al.[24] conducted a comparative study of different DMR-finding methods; the results of this study might motivate our future work. Several types of gene signatures exist in the bioinformatics field, such as the prognostic gene signature, diagnostic gene signature, and predictive gene signature. The term `prognostic' signifies the prediction of the expected development (duration, description, and function) of the course of a disease. Hence, the prognostic gene signature is vital to the overall outcome of a disease, irrespective of therapeutic interference. These prognostic signatures are useful in several tissue-Specific cancers such as hepatocellular carcinoma,[25] leukemia,[26] and breast cancer.[27] The diagnostic gene signature acts as a biomarker that differentiates the severity of phenotypes of analogous therapeutic conditions into the mild, moderate, or severe stage based on an inception point.[28] A predictive gene signature predicts the outcome of therapeutic intervention and does not depend on the prognosis.[29] Hence, these signatures contain crucial information. Several highly efficient biological networks can be used to predict the new functionality of genes.[30] One of the most popular biological networks is GCN, where each node in the network denotes a gene. Based on the edge between the two genes (nodes in GCN) of the network, GCNs are of two types, namely, unweighted-GCNs (UGCNs) and weighted-GCNs (WGCNs). In a UGCN, a threshold value is applied to the correlation coefficient. If the correlation coefficient value is higher than the threshold value, an edge must exist between two genes; otherwise, no edge is made. In a WGCN, the result depends on the choice of the threshold of the correlation coefficient. Hence, the WGCN is a preferred network, where an edge exists between every pair of nodes, and the weights of the edges are determined by the correlation values between the corresponding nodes. Various techniques have been currently developed for multi-omics integration. Weighted connectivity measure integrating co-methylation, co-expression and protein-protein interactions (WeCoMXP), based on the weighted connectivity measure, is an approach for integrating multi-omics data from the weighted normalized gene regulatory network to detect gene modules.[31] WeCoMXP is the most promising integration technique in which tri-omics profiles (expression data, methylation data, and protein-protein interactions) and hierarchical clustering are used to identify gene modules. Functional gene module detection (FGMD)[32] is another method for determining gene modules. In a previous study on FGMD, single omics (gene expression data) data were used from two platforms, RNA-Seq and microarray, for the same gene pairs and paired samples. First, the log2 ratios between the tumor samples and the average of normal samples were calculated. To identify functional gene modules, gene expression data of the microarray and RNA-Seq platforms were compared using the Pearson correlation coefficient (PCC) for the same gene pairs. Then, the gene expression network was constructed using the PCC values, and the gene pairs having higher PCC values were extracted. After extraction, hierarchical clustering was applied to the selected gene pairs. The analogous modules based on the overlap ratio were combined using hierarchical clustering with the dynamic tree cut method.[33] Finally, the FGMD modules were identified. The double-label propagation clustering algorithm (DLPCA) is a new algorithm for determining disease-associated modules by using the gene expression dataset. In a previous study, DLPCA used the pathogenic records of genes as the properties of nodes from the GCN constructed using the WGCN analysis (WGCNA) tool.[34] Gene modules were constructed by applying a multilabel clustering algorithm, followed by the DLPCA. During the module detection phase, the DLPCA classified the corresponding modules into diseased and non-diseased samples.[35]

Several well-known algorithms are used to identify a gene module, and many of them are developed to extract network modules using a hierarchical clustering algorithm. Scope exists for identifying stronger gene signatures that can exhibit higher classification performance of class labels and provide higher biological significance and validation. In theWGCNA method, hierarchical clustering and dynamic tree cut are used to discover the densely connected gene module. The PCC values are transformed by considering the recently evaluated power, and then, the topological overlap measure (TOM)[37] is computed on the basis of the power of the PCC values. The transcript modules are identified on the basis of the dissimilarity score, that is, (=1-TOM score). The TOM value between two vertices (e.g., transcripts) in an adjacency matrix is defined as follows.

$$\mathrm{TOM}(j,k)=\begin{cases}\frac{\sum\limits_{v\neq j,k} Y(j,v)Y(k,v)+Y(j,k)}{min\{\sum\limits_{h\neq l} Y(j,h),\sum\limits_{h\neq k} Y(k,h)\}-Y(j,k)+1}, & \text{if } j\neq k,\\ 1, & \text{if } j=\mathrm{k}.\end{cases}\tag{3}$$

By observing formula (3), the generalization of TOM can be expressed as follows.

$$\mathrm{GTOM1}(j,k)=\begin{cases}\frac{|nbd_1(j)\cap nbd_1(k)|+Y(j,k)}{min\{|nbd_1(j)|,|nbd_1(k)|\}-Y(j,k)+1}, & \text{if } j\neq k,\\ 1, & \text{if } j=\mathrm{k}.\end{cases}\tag{4}$$

Here, Y denotes the adjacency matrix of the nodes containing the first nearest neighbours of each nodes, '|.|' symbol indicates the cardinality of the set, $\mathrm{nbd}_1(j)$ signifies the set of neighbours of j excluding itself (i.e., *j*), and $\mathrm{nbd}_1(k)$ signifies the set of neighbours of k excluding itself (i.e., *k*), whereas $|\mathrm{nbd}_1(j)\cap\mathrm{nbd}_1(k)|$ refers to the number of common neighbours which are shared by the nodes j and k. In other words, in TOM/GTOM, final similarity score between two nodes has been computed based upon direct similarity between those two nodes and indirect similarity through the different levels of nearest neighbours, whereas for other general similarity measures like Pearson`s correlation, this indirect similarity through neighbourhood is not there.[38]

A gene involved in more than one function might also exist in a different gene subnetwork and thus in different functions. However, with hierarchical clustering, the overlaps between different sub-networks are avoided. In this study, we used a graph based model by detecting high cohesive clusters to overcome this issue. We collected data from The Cancer Genome Atlas (TCGA), and then found significantly expressed genes using the Limma-voom tool. Thereafter, we identified the target genes for the significantly expressed miRNAs. We only considered these significantly expressed target genes for the analysis. The spectral clustering algorithm was then applied to find the dense modules, and the silhoette index was calculated for each cluster. For the purpose of emperical analysis we consider the cluster 2 as mentioned in Figure 4. We used three consecutive classifiers to classify the class labels by using all the features belonging to the signature. Compared with the other signatures, mirna-mrna signature produced the highest accuracy across all the classifiers. Additionally, we conducted a gene enrichment analysis [Gene Ontology (GO) and pathway] to identify disease-related pathways as well as GO terms for the participating genes belonging to the signature. Our framework may prove useful for extracting integrated signatures for other microarrays/RNA-Seq datasets for cancer or any other disease.

Algorithm 1 Identification of combined molecular signatures from multi-omics data profiles

Input: Gene expression data matrix (GDM, row = genes, columns = unpaired samples).

miRNA datamatrix (MDM, row = miRNAs,columns = unpaired samples).

Output: Top ranked combined signatures, based on centrality score by analysing the multi-omics network.

Algorithm:

/*Extraction and filtering of features from GDM and MDM***/**

1: Selection of the common samples from GDM and MDM.

2: Identifying differentially expressed genes (DEG) and diferentially expressed miRNAs (DMR) w.r.t their fold change and p-values.

3: The target genes of DMR are identified.

4: The overlapping genes are obtained from DEG, and the target genes of DMR.

5: Selection of those down regulated genes that are negatively correlated with the upregulated miRNAs.

/*Construction of the multi molecular network with the upregulated miRNA-downregulated mRNAs***/**

6: Let N = (G;mR;E) is the network corresponding to multi-omics data profile. Each node in the network is denoted by G if the node is gene and denoted by mR if the node is a miRNA.

A member of G is linked by an edge if there exists a positive correlation between them. Similarly, a member of mR and a member of G is linked by an edge if there exists a negative correlation between them.

/*Finding the correlation between each gene pair for G ***/**

$$\mathrm{Gene}(N_i, N_j) = \begin{cases} 1, & \text{if } \mathrm{Cor}(N_i,\, N_j) > 0, \\ 0, & \text{Otherwise.} \end{cases} \tag{5}$$

/*Finding the negative correlation between each miRNA-gene pair ***/**

$$G_M(G_i, M_j) = \begin{cases} 1, & \text{if } \mathrm{Cor}(G_i,\, M_j) < 0, \\ 0, & \text{Otherwise.} \end{cases} \tag{6}$$

7: For the network N spectral clustering is applied.

8: For each gene and miRNA in each cluster degree centrality Dc is calculated,

9: Assigning rank to each miRNA in the network N based on the its Dc.

10: Among all the genes that are connected to the top ranked miRNA, we select those genes who have maximum degree centrality. Thus selecting the hub molecules from each clusters. By applying this approach we discovered top 3 ranked combined signatures.

11: End of the Algorithm

Proposed integrated Signature Discovery Technique

In this article, we introduced a novel framework for detecting dense modules and discussed their application in a prognosis survival study. We performed an integrative analysis of the mRNA and miRNA from the TCGA Cervical cancer datasets. The steps of the method are described as follows.

Identification of Significantly Expressed Transcripts

We first chose the common sample IDs from the multi-omics (mRNA and miRNA) datasets and then collected the subtypes of cervical cancer from the phenotype data for those sample IDs. Thereafter, the gene probes containing the missing values (i.e., NA values) and those with expression values of zero across all the samples in the dataset were eliminated to obtain the filtered dataset. We used the Limma method, a non-parametric test,[19] employing the empirical Bayes test to determine significantly expressed gene probes and miRNAs because it performs very well for any distribution (normal or non-normal distribution) and for all sample sizes. The empirical Bayes approach causes a reduction of the estimated sample variance toward a pooled estimate, resulting in a more stable inference. The use of moderated t-statistics is more advantageous than that of the posterior odds because the number of hyperparameters that need to estimate is reduced. To avoid a large number of item sets resulting from numerous genes, we considered only the top significantly expressed genes and miRNAs by using the Limma statistical test and generated a list of genes sorted according to their p -values from significant to insignificant. Thereafter, we assigned a weight to each gene and miRNA with respect to their p-values. The t-statistics in Limma is described as follows.

$$\tilde{t_g} = (\sqrt{(D_0 + D_g)/D_g}) * (\tilde{\beta_g}/\sqrt{S^2_{*;g} V_g}) \qquad (7)$$

The degree of freedom is D_0+D_g, βg is the contrast estimator, and S^2_* ;g the posterior variance. Gene probes and miRNAs whose p-values were less than 0.05, as obtained using the empirical Bayes test, were considered as significant gene probes and miRNAs, respectively. Simultaneously, we also considered fold change (FC) to identify significant gene probes and miRNAs. FC is a measure for quantifying the ratio of the mean score of the diseased samples to the mean score of the control samples. The FC value was used to analyze the changes in gene and miRNA expression between multiple normal and tumor samples. For upregulated gene probes (UG) and upregulated miRNAs (UMIR), the threshold value of FC is 2 (i.e., FC 2), whereas for the downregulated gene probes (DG) and downregulated miRNAs (DMIR), the threshold value of FC is 0.5 (i.e., FC 0.5). In this manner, the UG), UMIR, DG, and DMIR were selected on the basis of p and FC values. The overview of the work flow of the proposed method is presented in Figure 1.

MiRNA Target Detection Using SpirderMiR

Gene regulatory networks (GRNs) play a major role in various biological processes, such as cell cycle, cell differentiation, signal transduction, and metabolism, during the pathological process. The differences between GRNs in normal and pathological conditions may unveil the mechanisms underlying disease development. GRNs are split into some simple connections that describe how the network nodes interact. Users can integrate the miRNA{gene interaction into various network data, such as gene coexpression, genetic interactions, physical interactions, and pathways. We provided the significantly expressed miRNAs as inputs in the SpidermiR R tool[39] that consists of predicted miRNA{target gene interactions from eight external databases, namely, EIMMo, DIANA, miRanda, PITA, miRDB, MicroCosm, PicTar, and TargetScan, and validated miRNA-target gene interactions from miRecords, miRTarBase, and TarBase. Once the integrated network data were prepared, we conducted the downstream analysis.

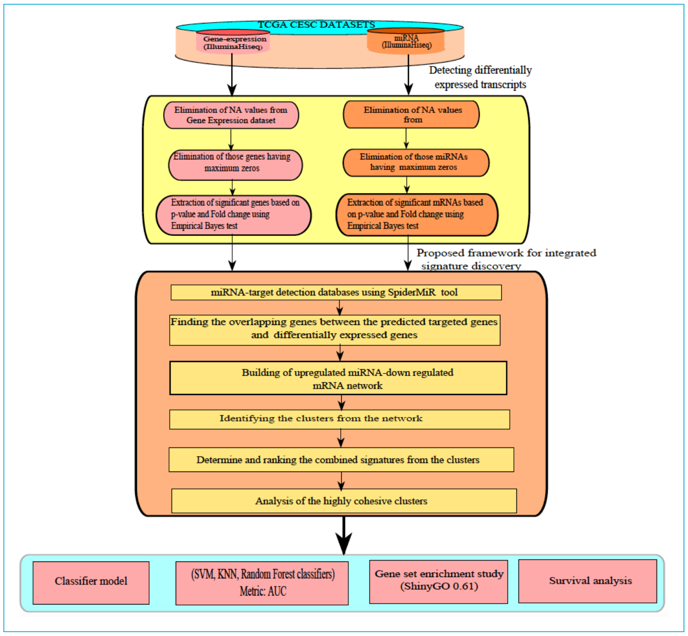

TCGA CESC DATASETS
Gene-expression (IlluminaHiseq)
miRNA (IlluminaHiseq)
Detecting differentially expressed transcripts
Elimination of NA values from Gene Expression dataset
Elimination of those genes having maximum zeros
Extraction of significant genes based on p-value and Fold change using Empirical Bayes test
Elimination of NA values from
Elimination of those miRNAs having maximum zeros
Extraction of significant mRNAs based on p-value and Fold change using Empirical Bayes test
Proposed framework for integrated signature discovery
miRNA-target detection databases using SpiderMiR tool
Finding the overlapping genes between the predicted targeted genes and differentially expressed genes
Building of upregulated miRNA-down regulated mRNA network
Identifying the clusters from the network
Determine and ranking the combined signatures from the clusters
Analysis of the highly cohesive clusters
Classifier model
(SVM, KNN, Random Forest classifiers) Metric: AUC
Gene set enrichment study (ShinyGO 0.61)
Survival analysis

Figure 1. Flowchart of proposed combined signature finding Strategy.

Detection of Integrative Signature Using Spectral Clustering

Gene expression is distinguished to be regulated by the association between transcription factors (TFs) and upstream regulatory components of target genes.[40] Moreover, miRNAs expression can be started or repressed by TFs, which therefore can act as miRNA regulators.[41] Thus a change in miRNA expression could result from an alteration in transcriptional activity. miR-17-5p and miR-20a which are the members of a cluster are dysregulated in cervical cancer.[42] Regulation of miRNAs by transcription factor has been studied only in rare studies and therefore it would be promising if important combined miRNAs and gene signature in cervical cancer are identified. We attempted to extract the modules with highly associated miRNAs and mRNAs. We built a network with upregulated miRNAs and its target genes. We considered those downregulated mRNAs that are negatively correlated with the upregulated miRNAs. As the network consist of two types of components, so there are two types of nodes representing miRNA and mRNA. There is an edge between two nodes (when two nodes are miRNA and mRNA) if there is a negative correlation between them, whereas an edge between two nodes (when two nodes are mRNAs) if there is a positive correlation between them. Spectral clustering is a method in graph theory, where the technique is used to identify communities of nodes in a connected network based on the edges connecting them. So in this study, we applied spectral clustering[43] on the entire miRNA-mRNA network. Once the integrated network data were prepared, we conducted the downstream analysis.

Combined Molecular Signature Discovery and Ranking

From the spectral clustering we found 4 clusters. Our aim was to detect those modules containing hub molecules. We applied network centrality measure on each molecule of the network and captured the high degree miRNAs with theirs associated high degree mRNAs.

The hub miRNA-mRNA modules are ranked and listed in Table 1. We ranked each combined module by looking at

Table 1. Degree-centrality based cluster Specific top combined signature for Cervical Cancer.

CI	SI	Cluster Elements with Degree centrality		Combined signature	Rank	Protein Class from Human Protein Atlas
1	0.85	hsa-mir-34c	9	fhsa-mir-34c-> DRD2, HNF4Ag	1	DRD2: FDA approved drug targets, G-protein coupled receptors, transporter HNF4A: Disease-related genes, Nuclear receptors, and transcription factors
		BMP3	1			
		DRD2	7			
		HNF4A	7			
		SIX3	1			
		TRIM10	1			
		CAPN9	1			
		CPLX2	1			
		FBXO16	1			
		PALM3	1			
1	0.85	hsa-mir-137	7	fhsa-mir-137-> TOX3, SLC39A5g	2	SLC39A5: Disease related genes, Potential drug targets transporters TOX3: Transcription factors
		ALPPL2	1			
		AQP2	1			
		FGF9	1			
		DLGAP1	1			
		TOX3	13			
		SLC39A5	9			
		SHISA9	1			
2	0.96	hsa-mir-944	6	fhsa-mir-944-> IYDg	3	IYD: Disease related genes, enzymes, and potential drug targets
		CFTR	1			
		GABRB2	1			
		HNF4G	1			
		TAC1	1			
		C7orf57	1			
		IYD	7			

SI represents Silhoutte Index; CI denotes Cluster Index.

the degree centrality of each miRNA. In the interaction in the real world, we usually recognize people with numerous connections to be central. Degree centrality conveys the same concept into a measure. The degree centrality of a node quantifies the ranks of the node with more connections more leading in terms of centrality. The degree centrality Dc for node Ni in an undirected graph is

$$D_c(N_i) = Deg_i \tag{8}$$

where Degi is the degree i.e the total number of connected edges with node Ni. For a directed graph the degree centrality, say DirC is considered either the in-degree or the out-degree or the fusion of both (in-degree and out-degree), i.e.,

$$Dir_c(N_i) = Deg_i^{in} \tag{9}$$

$$Dir_c(N_i) = Deg_i^{out} \tag{10}$$

$$Dir_c(N_i) = Deg_i^{in} + Deg_i^{out} \tag{11}$$

When we are utilizing in-degrees, degree centrality estimates how vital a node is and its value gives prominence or prestige, whereas out-degrees measure the gregariousness of a node. When we joining in-degrees and out-degrees, we are neglecting the edge directions, i.e., we are considering the undirected graph and the Equation.11 turns into Equation 8. Figure 2 is presenting a sample of an

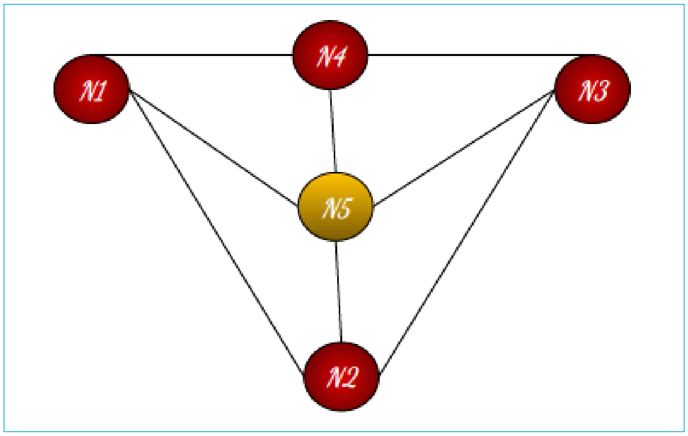


Figure 2. Sample graph for degree centrality.

undirected graph G_1. G_1 cosists of 5 nodes (N_1, N_2, N_3, N_4, and N_5). Where degree centrality Dc of each node is as follows $D_c(N_1) = 3; D_c(N_2) = 3; D_c(N_3) = 3; D_c(N_4) = 3$, and $D_c(N_5) = 4$. As a result, node N5 is most important in the graph as it is having highest degree centrality.

Implementation and Data Availability

Here we elaborated the information of such datasets that we utilized in this study, viz., gene expression (Illumina-HiSeq) dataset, miRNA expression dataset and clinical data containing phenotype information for Cervical cancer (CESC). Then we demonstrated the outcome of our experiment. We downloaded CESC gene expression data (Illumina-HiSeq), miRNA expression data (Illumina-HiSeq) and clinical data containing phenotype information from the TCGA database by using UCSC Xena browser (https://xenabrowser.net/datapages/). CESC gene expression and miRNA expression data sets contained a total of 275 samples. We extracted those sample IDs that are common in both data sets. From the clinical matrix, we obtained the phenotype information like patient survival time, survival status, cervical cancer subtypes, etc. As there exist few number of control samples in the TCGA CESC dataset, so we considered two subtypes of CESC as two groups, viz., Endocervical type of Adenocarcinoma (ADENO) and Cervical Squamous Cell Carcinoma (SCC) for our experiment. The total number of samples in ADENO and SCC are 22 and 253, respectively.

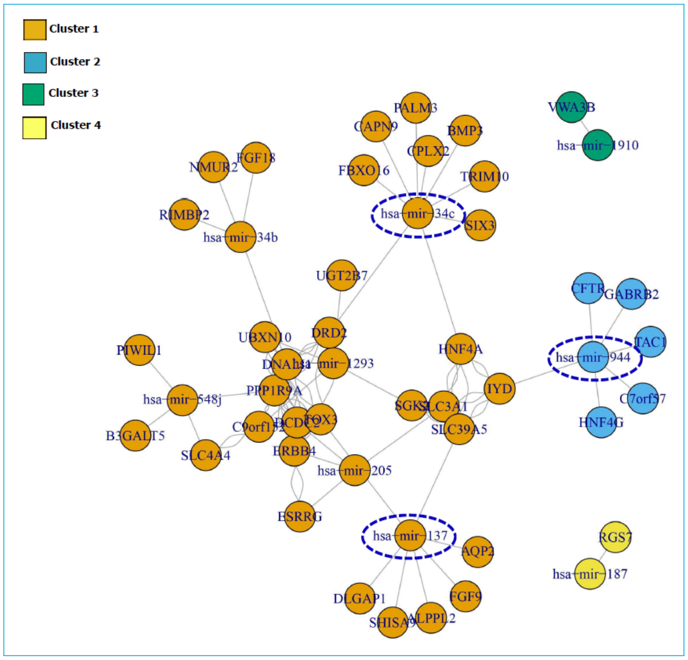


Figure 3. miRNA-mRNA integrated clusters through spectral clustering, where dotted line indicates hub nodes (hub miRNAs here.

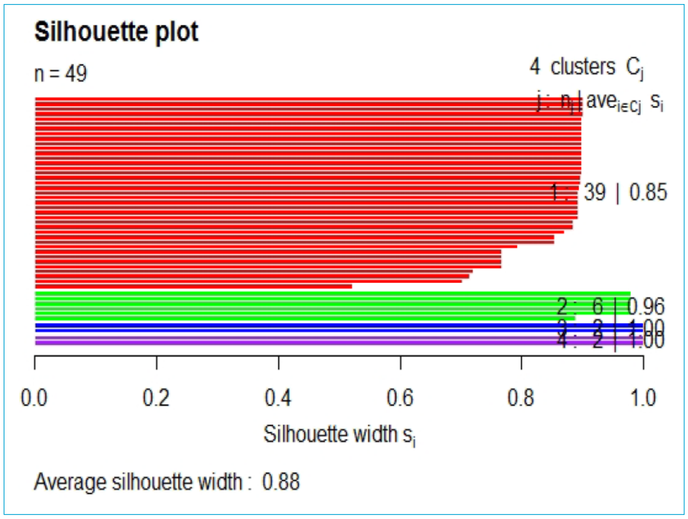
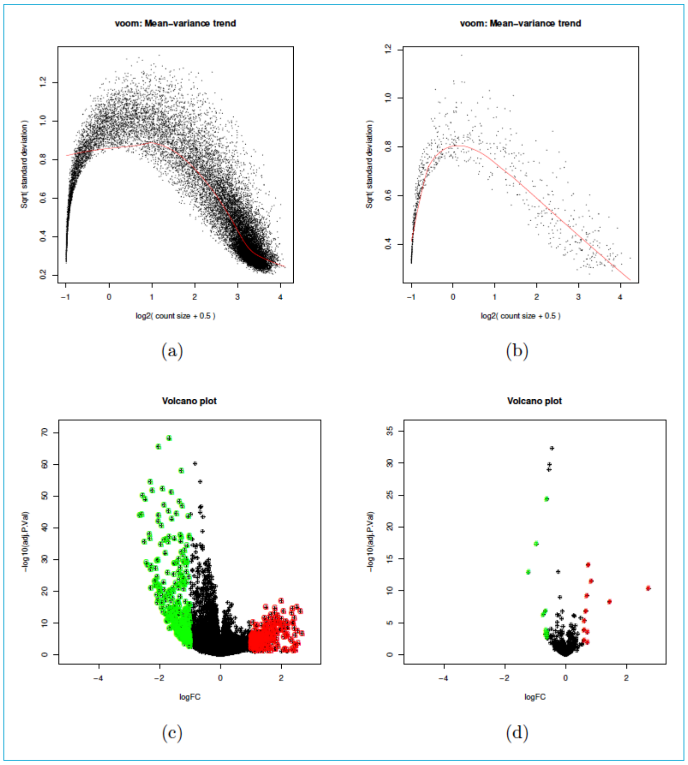


Figure 4. Silhouette plot for the 4 clusters.

Results

Initially, the gene expression profile contained a total of 20,530 gene probes, while the miRNA expression profile contained 1,046 miRNAs. There was a total of 313 samples in the phenotype data (the clinical matrix) among which 308 samples were common in both mRNA and miRNA expression profiles. We omiited the invalid features (genes). After filtering, we had a total of 20,501 genes and 275 samples of which the number of ADENO samples was 22, while the number of SCC samples was 253. Hence, a total of 275 samples were considered for the experiment, while after re-filtering, we obtained a total of 19,685 genes and 889 miRNAs. Next, we identified the significantly expressed genes by applying Limma statistical test as mentioned in section 2.1 on the dataset. Intuitively, the total number of significant genes in the dataset was found as 580. Of note, since no normal sample was available in the dataset, we considered the effect of one subtype versus others through the statistical test. While we tried to measure the effect of SCC over ADENO, we obtained 259 over-expressed genes in SCC. On the other hand, to measure the effect of ADENO over SCC, we identified 321 over-expressed genes in ADENO. In addition, we identified 20 significantly expressed miRNAs in the miRNA datasets of which 11 miRNAs were overexpressed in SCC, and the remaining 9 miRNAs were over-expressed in ADENO. The voom:mean-variance trend plot for the gene expression data and miRNA expression data were presented by Figure 5(a) and Figure 5(b), respectively.

The volcano plots in Figure 5(c) and Figure 5(d) represented the significantly expressed genes and significantly expressed miRNAs, respectively. The over-expressed genes/miRNAs in SCC were represented in red color while the over-expressed genes/miRNAs in ADENO were illustrated in green color in



(a) (b) (c) (d)

Figure 5. (a) Voom: Mean-Variance trend plot during the extraction of signicantly expressed genes in Adenocarcinoma vs Squamous cell carcinoma, **(b)** Voom: Mean-Variance trend plot during the extraction of signicantly expressed miRNA in Adenocarcinoma vs Squamous cell carcinoma, **(c)** Volcano plot for nding signicantly expressed genes in Adenocarcinoma vs Squamous cell carcinoma, **(d)** Volcano plot for nding signicantly expressed miRNA in Adenocarcinoma vs Squamous cell carcinoma.

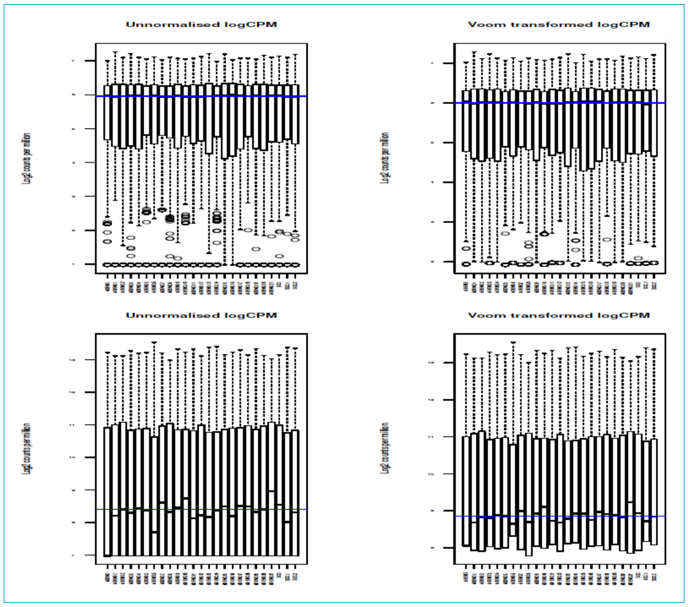


Figure 6. (a) Performing the Box plots before and after Voom normalization while extracting the signicantly expressed genes. **(b)** Performing the Box plots before and after Voom normalization while extracting the signicantly expressed microRNAs.

the volcano plot. The boxplot on both the gene expression data before and after Voom transformation were shown in Figure 6 (a) and Figure 6 (b), respectively.

In this study we considered two subtypes of CESC as there were few number of control samples. As a result, we obtained 9 upregulated miRNAs and 40 downregulated genes. The genes were also predicted target genes and negatively correlated with those overexpressed miRNAs (Table 4). For our experiment we constructed an integrated network with these 9 miRNAs and 40 mRNAs. After applying spectral clustering on the integrated network 4 integrated clusters were found (Fig 3). As we can see from the silhouette plot in Figure 4, the silhoette index of cluster 1,2,3, and 4 were 0.85, 0.96, 1, 1 respectively. It was noted from the network Figure 3 cluster 1 contains 2 and cluster 2 contains 1 hub miRNAs (based on its degree centrality) respectively demarcated with dotted oval. Table 1 representing the summary of network in Figure 3. From the degree centrality of miRNAs and its associated genes we ranked each signature. As the cohesiveness of cluster-2 was very high, and also contained a combined signature, we conducted geneset enrichment analysis and pathway analysis for cluster-2 separately.

Cluster Specific Combinatorial Fused Signature Analysis

After finding the cluster, we focused on extracting the combined molecular signature. In Table 1 the cluster index signies in which cluster the bio molecules belong to.The color of each cluster are depicted in miRNA-mRNA target network (Fig. 3). We performed a network analysis after applying the spectral clustering on the network. In the Orange colored cluster (cluster-1) the miRNAs hsa-mir-34c and hsa-mir-137 were demarcated with dotted oval. The highest ranked miRNA in cluster-1 is hsa-mir-34c as its degree centrality is the highest among the other miRNAs within the cluster. The associated target genes with highest in-degree are DRD2, and HNF4A. Intuitively, a combine biomolecular signature {hsa-mir-34c->DRD2,HNF4A} was formed in cluster-1. Similarly, the second highest ranked signature was {hsa-mir-137->TOX3, SLC39A5} in cluster-1. Another combined signature {hsa-mir-944->IYD} was selected in cluster-2 the blue colored cluster in Figure 3. We searched for annotation and protein class(Human Protein Atlas[1]) of every gene belonging to the top 3 ranked signatures. In {hsa-mir-34c->DRD2,HNF4A}, we found DRD2 is a FDA-approved

[1] https://www.proteinatlas.org/

drug target, and encodes the D2 subtype of the dopamine receptor. Also it functions as a G-protein coupled receptor. The protein encoded by HNF4A controls the expression of various genes, including hepatocyte nuclear factor 1 alpha, a transcription factor that regulates the expression of different hepatic genes. Diseases associated with HNF4A include Type 1, Maturity-Onset Diabetes. Also, it plays a vital role in the development of the liver, kidney, and intestines. Mutations in HNF4A are associated with monogenic autosomal dominant non-insulin-dependent type I diabetes mellitus. Whereas, miRNA hsa-mir-34c is associated with cervical cancer. The downregulated expression of hsa-miR-34c is associated with a small increase in cellular proliferation and a significant growth in cell migration.[44] As a result, the signature {hsa-mir-34c->DRD2,HNF4A} should have a vital role in cervical cancer. In {hsa-mir-137->TOX3, SLC39A5} we found the gene SLC39A5 is associated with Myopia, Autosomal Dominant. The associated pathways are Transport of glucose and bile salts, other sugars, metal ions and Metal ion SLC transporters. Where TOX3 is an important regulator of calcium-dependent transcription that employ its effect on CRE-mediated transcription with the CREB{CBP complex.[45]

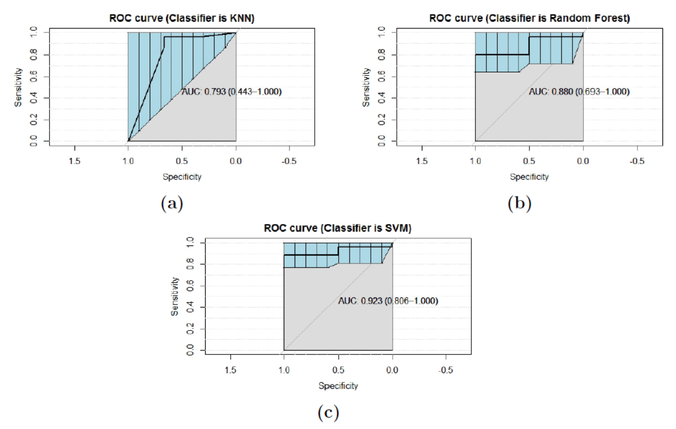
The upregulation of MiRNA hsa-miR-137 suppresses the tumor progression in Cervical Cancer by blocking the Transforming growth factor (TGF-) pathway.[46] Consequently, the combined signature {hsa-mir-137->TOX3, SLC39A5} might play an significant role in Cervical cancer. On the other hand in the rank-3 signature {hsa-mir-944->IYDg, IYD has been found as a FDA-approved drug target. Also, it encodes an enzyme that catalyzes the oxidative NADPH-dependent deiodination of mono as well as di-iodotyrosine, which are the halogenated by products of thyroid hormone production. Mutations in this gene can produce congenital hypothyroidism due to dyshormonogenesis type-4^2. In a study,[47] an association has been identified between high expression levels of hsa-mir-944 and expression levels of HPV in HPV-positive cervical cancer cells with TCGA dataset. Hence, the signature {hsa-mir-944->IYD} should have an influence in progress of cervical cancer.

Emperical Analysis and Classification Accuracy of Cluster-2

We considered the the evolved genes and miRNAs (features) belonging to the cluster 2, and then applied the cross-validation technique and several classifiers to classify the groups (diseased or control) toward the samples. We computed the area under the curve (AUC) for each classifier for the evolved integrated signatures obtained using our proposed method as well as we compared the performance by individually measuring the performance with miRNAs and mRNAs (Fig. 8). We plotted ROC curve (Fig. 7) (receiver operating characteristic curve) for showing the performance of the classication model.

Gene Set Enrichment Analysis of Cluster-2 Using ShinyGO

We conducted the pathway and GO analysis and selected pathways and GO terms having a significant enrichment score (p-value<0.05). We then highlighted pathways and GO terms that were associated with participating genes belonging to the signature. Intuitively, we obtained a combined signature consisting of genes and one miRNAs: CFTR, GABRB2, TAC1, c7orf57, HNF4G, and hsa-mir-944. Furthermore, we conducted pathway and Gene Ontology (GO) analyses using the Shiny GO http://bioinformatics.sdstate.edu/go/. Here, we observed that the genes belonging to the cluster-2 followed several significant biological processes, as mentioned in Table 2; e.g., the genes CFTR, TAC1, GABRB2, and HNF4G were involved in Response to



(a) (b) (c)

Figure 7. From **(a) - (c)** ROC curve through the use of multiple classiers (viz., SVM, RF and KNN) with the integrated signature.

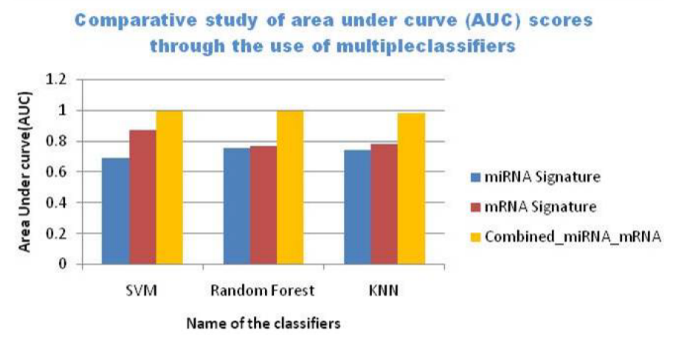


Figure 8. Comparative study of area under curve (AUC) scores through the use of multiple classifiers (viz., SVM, RF and KNN).

[2]https://www.ncbi.nlm.nih.gov/gene/389434

Table 2. Enriched GO-terms associated with the evolved genes in cluster 2.

Name of the High level GO category	p	Name of the Groups of GO	# Associated Genes	Genes
Response to organic cyclic compound	0.001462	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Regulation of membrane potential	0.001683	Biological process	3	GABRB2, CFTR, TAC1
Synaptic transmission, GABAergic	0.001683	Biological process	2	TAC1, GABRB2
Cellular response to endogenous stimulus	0.001683	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Chloride transmembrane transport	0.00225	Biological process	2	CFTR, GABRB2
Response to endogenous stimulus	0.00225	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Cellular response to organic cyclic compound	0.00225	Biological process	3	CFTR, GABRB2, HNF4G
Inorganic anion transmembrane transport	0.002735	Biological process	2	CFTR, GABRB2
Chloride transport	0.002735	Biological process	2	CFTR, GABRB2
Positive regulation of hormone secretion	0.003175	Biological process	2	CFTR, TAC1
Response to ammonium ion	0.003175	Biological process	2	TAC1, GABRB2
Monovalent inorganic cation homeostasis	0.003824	Biological process	2	CFTR, TAC1
Inorganic anion transport	0.004249	Biological process	2	CFTR, GABRB2
Response to organonitrogen compound	0.005033	Biological process	3	CFTR, TAC1, GABRB2
Response to lipid	0.005033	Biological process	3	CFTR, TAC1, HNF4G
Response to nitrogen compound	0.006322	Biological process	3	CFTR, TAC1, GABRB2
Organic hydroxy compound transport	0.006515	Biological process	2	CFTR, TAC1
Cellular response to organic substance	0.006515	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Anion transmembrane transport	0.007212	Biological process	2	CFTR, GABRB2
Regulation of hormone secretion	0.007212	Biological process	2	CFTR, TAC1
Lipid transport	0.009323	Biological process	2	CFTR, TAC1
Hormone transport	0.009323	Biological process	2	CFTR, TAC1
Response to organic substance	0.009323	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Hormone secretion	0.009323	Biological process	2	CFTR, TAC1
Cellular response to chemical stimulus	0.009323	Biological process	4	CFTR, TAC1, GABRB2, HNF4G
Lipid localization	0.010782	Biological process	2	CFTR, TAC1
Positive regulation of secretion by cell	0.012745	Biological process	2	CFTR, TAC1
Cell-cell signaling	0.013507	Biological process	3	GABRB2, CFTR, TAC1
Positive regulation of secretion	0.013507	Biological process	2	CFTR, TAC1
Signal release	0.01375	Biological process	2	CFTR, TAC1
Chloride channel complex	0.001464053	Cellular Component	2	CFTR, GABRB2
Ion channel complex	0.014161241	Cellular Component	2	CFTR, GABRB2
Transmembrane transporter complex	0.014161241	Cellular Component	2	CFTR, GABRB2
Transporter complex	0.014161241	Cellular Component	2	CFTR, GABRB2
Ligand-gated anion channel activity	0.000146823	Molecular function	2	GABRB2, CFTR
Anion channel activity	0.001149013	Molecular function	2	CFTR, GABRB2
Chloride channel activity	0.001149013	Molecular function	2	CFTR, GABRB2
Chloride transmembrane transporter activity	0.001149013	Molecular function	2	CFTR, GABRB2
Ligand-gated ion channel activity	0.001293524	Molecular function	2	GABRB2, CFTR
Ligand-gated channel activity	0.001293524	Molecular function	2	GABRB2, CFTR
Inorganic anion transmembrane transporter activity	0.001352648	Molecular function	2	CFTR, GABRB2

Table 2. CONT.

Name of the High level GO category	p	Name of the Groups of GO	# Associated Genes	Genes
Anion transmembrane transporter activity	0.005033167	Molecular function	2	CFTR, GABRB2
Gated channel activity	0.005033167	Molecular function	2	GABRB2, CFTR
Ion gated channel activity	0.005033167	Molecular function	2	GABRB2, CFTR
Ion channel activity	0.00634347	Molecular function	2	GABRB2, CFTR
Channel activity	0.00634347	Molecular function	2	GABRB2, CFTR
Passive transmembrane transporter activity	0.00634347	Molecular function	2	GABRB2, CFTR
Substratespecific channel activity	0.00634347	Molecular function	2	GABRB2, CFTR
Inorganic molecular entity transmembrane transporter activity	0.018378207	Molecular function	2	GABRB2, CFTR
Ion transmembrane transporter activity	0.020043027	Molecular function	2	GABRB2, CFTR
Transmembrane transporter activity	0.027244201	Molecular function	2	CFTR, GABRB2
Transporter activity	0.035929585	Molecular function	2	CFTR, GABRB2

endogenous stimulus (already identified in an earlier study [48]). TAC1, GABRB2, CFTR were associated with the Regulation of biological quality(earlier found in [49]). CFTR, TAC1 followed the Reproduction, Response to stress, Sexual reproduction, Reproductive process, Regulation of signaling, Regulation of localization, Macro-molecule localization, Multi-organism reproductive process, Multicellular organism reproduction, and Regulation of response to stimulus, and Multi-organism process were associated with cervical cancer, already known in literature.[50-58] GABRB2, and TAC1 were linked in Cellular component biogenesis. In contrast, the genes CFTR, and GABRB2 were associated with the cellular component Chloride channel complex, Ion channel complex, Transporter complex, and Transmembrane transporter complex. There were some important molecular functions in which genes CFTR, and GABRB2 were involved. The functions were Ligand-gated anion channel activity, Anion channel activity, Chloride channel activity, Chloride transmembrane transporter activity, Ligand-gated ion channel activity, Inorganic anion transmembrane transporter activity, Anion transmembrane transporter activity, Gated channel activity, Ion gated channel activity, Ion channel activity, Channel activity, Passive transmembrane transporter activity, Substrate-Specific channel activity, Inorganic molecular entity transmembrane transporter activity, Ion transmembrane transporter activity, Transmembrane transporter activity, and Transporter activity. Of note, gene HNF4G involved in the pathway Maturity onset diabetesof the young, Gene GABRB2 and TAC1 were associated with Neuroactiveligandreceptorinteraction, gene CFTR involved in Bile secretion, Pancreatic secretion and AMPK signaling pathway (Table 3).

Survival Analysis

Survival analysis is one of the key statistical methods for exploring data on time to the occurrence of an event of interest, such as death, or time to failure of a device. It can be applied to many aspects such as estimating the year of death, evaluating the reliability of a product, and measuring the capability of medical therapies. Survival analysis is difficult to perform in cases with undetectable or inexistent outcomes in the observation period. This type of event is called as censoring that can be dealt with the survival analysis strategy and is required to perceive how well the signature predicts the survival time for the patients in the respective clinical dataset.

In this study, we applied the Cox proportional-hazards regression (coxph R) package[59] to investigate the association between the survival time of the patients and one or more predictor variables, considering the gene expression profile of only the identified resultant module . We computed the Z-score for each gene to produce high- and low-risk patient groups. The difference in survival time between the two groups of patients was determined using the Kaplan-Meier estimator as well as the log-rank method. Genes in the modules were associated with the patient survival time in particular cancer. We predicted the patient survival time for each gene belonging to the resultant signature on the basis of gene expression and classified the patients into high- and low-risk groups, in whom the survival time was significantly different p-value<0.05. The same procedure was followed for all the genes belonging to the cluster 2, and the frequency of a significant p-value was obtained.

In addition, we explored the survival prognosis analysis for the genes belonging to cluster-2. For survival analysis, we used Cox proportional hazards regression model to predict the survival time of the underlying patients (dead or alive). For the living and deceased patients, we extracted the days from the last follow-up time and overall survival time, respectively. In Fig.9 we obtained Cox regression p-value for each gene and miRNA in cluster-2. Among them, we obtained 5 significant p-values (<0.05), while one was insignificant. E.g., for the gene TAC1, the p-value was 0.014 (significant). Similarly, in the case of the gene HNF4G, the p-value was 0.0071 (significant), whereas, for the miRNA hsa-mir-944 the p-value was 0.017(significant). For the gene CFTR and c7orf57 the p values were 0.028 and 0.038 respectively. For these individual gene and miRNA wise survival analyses, the corresponding plots were illustrated in Figure 9(a)-(f). Overall, since we found the significant p-values for the individual survival cases, it implies that our experimental cluster was powerful enough to say clinically promising. In addition, we carried out the performance analysis of the same cluster (i.e., cluster-2) by estimating the prediction accuracy of the cervical cancer subtypes (Adenocarcinoma and Squamous cell carcinoma). Hence, we provided here a comparative study of the AUC score of the elements in cluster-2 with mRNA and miRNA signatures individually, found by applying our method. For two-class classification, we used Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbors (KNN) classifiers. In contrast, the Area under the Curve (AUC) was used as a performance metric. We obtained the highest AUC values (>0.95) for the cluster-2 estimated by our proposed method. The classification accuary was performed using Caret R Package.[60] Figure 8 represented the comparative analysis. Finally, our proposed method produced the best AUC score to classify the cervical cancer subtypes among the other two types (with mRNA and miRNA) signatures. Moreover, our method is useful and powerful enough to identify a molecular signature from RNA-seq or similar data.

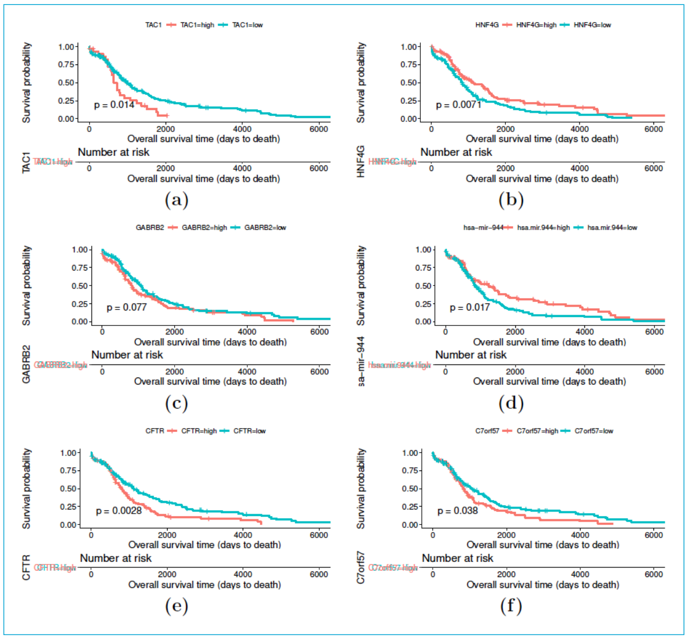


Figure 9. From **(a) - (f)** Survival analysis of each molecule of the Signature: Kaplan-Meier plots and Log-Rank Test p-values comparing overall survival times.

Discussion and Conclusion

Investigating the association between transcriptomic details is essential to understand the functionalities of the biological process. Recent innovations have made it conceivable to perform multi-omics profiling, including gene expression and miRNA expression. However, the integrative analysis of heterogeneous information provides biologically relevant information more precisely rather than the analysis with a single omic profile. Nowadays, most of the existing methods for integrating multi-omics profiles apply hierarchical clustering indicating the relationship the omics profiles. Since the hierarchical clustering does not consider the overlapping modules, there is a chance of losing important information. The combined signatures, used a network based model to determine the cluster Specific signatures. Overall, it possesses multiple unique advantages: (i) It provides a novel strategy for the integrative analysis of gene and miRNA expression data. (ii) It is progressively more potent than present-day techniques since the AUC scores of it are the highest across the three classifiers presented in Figure 8. (iii) The resultant signature was found clinically validated since most of the members in the signature produced significant p-value in cox regression-based survival analysis. In this article, we developed a new framework to extract dense modulebased integrative signature detection technique and their application in prognosis survival study. We used a cervical cancer data repository with clinical prognosis data to perform our experiment. At first, we applied Empirical Bayes test using Limma method to determine dysregulated genes (or, dysregulated miRNAs). MiRNA-mediated dysregulated target genes were identified from those dysregulated miRNAs. Next, we detected dense modules using spectral clustering technique. The cluster that contained the highest silhouette index (=0.96) was considered as the cluster for our analysis. MiRNA-mRNA signature produced the best AUC values (>=0.95 for all classifiers) for our resultant signature in compared to the individual signatures (as presented in Fig. 8). Our proposed method is efficient and useful to identify a molecular signature for any RNA-seq or similar profile.

The possible direction of our future work will lead to considering the apply this method in the study of epigenetics,

Table 3. Table shows the enriched pathways of the down regulated 40 genes, that targeted by 9 upregu- lated miRNA. Bold font gene (also demarcated with `*') indicates the evolved signature genes found in the corresponding pathway

Pathway	% of Associated Genes	Enrichment FDR of pathway	Names of Associated Genes
Maturity onset diabetes of the young	7.692%	0.015	HNF4A, HNF4G*
Neuroactive ligand-receptor interaction	1.187%	0.016	DRD2, GABRB2*, NMUR2, TAC1*
PI3K-Akt signaling pathway	1.133%	0.016	SGK2, ERBB4, FGF9, FGF18
Rap1 signaling pathway	1.456%	0.018	DRD2, FGF9, FGF18
Bile secretion	2.778%	0.018	CFTR*, SLC4A4
Melanoma	2.778%	0.018	FGF9, FGF18
Pancreatic secretion	2.062%	0.028	CFTR*, SLC4A4
MAPK signaling pathway	1.017%	0.028	ERBB4, FGF9, FGF18
AMPK signaling pathway	1.667%	0.032	CFTR*, HNF4A
Breast cancer	1.361%	0.039	FGF9, FGF18
Gastric cancer	1.351%	0.039	FGF9, FGF18

Table 4. miRNA and mRNA paires predicted by our proposed framework, also the pairs are negatively correlated

miRNA	Regulation of miRNAs	Number of target genes	Target genes
hsa-mir-205	Up-regulated	4	ERBB4, ESRRG, SLC3A1, DCDC2
hsa-mir-137	Up-regulated	7	ALPPL2, AQP2, FGF9, DLGAP1, TOX3, SLC39A5, SHISA9
hsa-mir-944	Up-regulated	6	CFTR, GABRB2, HNF4G, TAC1, C7orf57, IYD
hsa-mir-548j	Up-regulated	4	SLC4A4, PIWIL1, B3GALT5, PPP1R9A
hsa-mir-1293	Up-regulated	4	UGT2B7, SGK2, UBXN10, C9orf152
hsa-mir-1910	Up-regulated	1	VWA3B
hsa-mir-34c	Up-regulated	9	BMP3, DRD2, HNF4A, SIX3, TRIM10, CAPN9, CPLX2, FBXO16, PALM3
hsa-mir-34b	Up-regulated	4	DNALI1, FGF18, RIMBP2, NMUR2
hsa-mir-187	Up-regulated	1	RGS7

specially methylation. Interestingly, in a recent study, it has been observed that contiguous regions exist in the epigenome denoted as differentially methylated regions (DMRs) which are significantly associated with the various diseases.[22-23] In addition, We found the comparative study of various DMR finding methods in[24] that might motivate us to extend our future work.

Acknowledgments

SS and SB acknowledge the SyMeC Project (grant number BT/Med-II/NIBMG/SyMeC/2014/Vol.II) funded by the Department of Biotechnology, Government of India.

Disclosures

Ethics Committee Approval: The study was approved by the Local Ethics Committee.

Peer-review: Externally peer-reviewed.

Conflict of Interest: None declared.

Authorship Contributions: Concept - S.S., S.M., S.B.; Design - S.S., S.M.; Supervision - S.M., S.B.; Materials - S.S., S.M.; Data collection &/or processing - S.S., S.M.; Analysis and/or interpretation - S.S., S.M.; Literature search - S.S.; Writing - S.S., S.M.; Critical review - S.S., S.M.

References

1. Bhadra T, Mallik S, Sohel A, Zhao Z. Unsupervised feature selection using an integrated strategy of hierarchical clustering with singular value decomposition: an integrative biomarker discovery method with application to acute myeloid leukemia. IEEE/ACM Trans Comput Biol Bioinform. 2021 Sep 8;PP. doi: 10.1109/TCBB.2021.3110989. [Epub ahead of print].
2. Kandimalla R, Shimura T, Mallik S, Sonohara F, Tsai S, Evans DB, et al. Identification of serum miRNA signature and establish-

ment of a nomogram for risk stratification in patients with pancreatic ductal adenocarcinoma. Ann Surg 2022;275:229–37.

3. Mallik S, Zhao Z. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. Brief Bioinform 2020;21:368–94. [CrossRef]
4. Bhadra T, Mallik S, Hasan N, Zhao Z. Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. BMC Bioinformatics 2022;23:153. [CrossRef]
5. Maulik U, Sen S, Mallik S, Bandyopadhyay S. Detecting TF-miRNA-gene network based modules for 5hmC and 5mC brain samples: a intra- and inter-species case-study between human and rhesus. BMC Genet 2018;19:9. [CrossRef]
6. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20. [CrossRef]
7. Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, Mac Grogan G, et al. A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. Clin Cancer Res 2008;14:1744–52. [CrossRef]
8. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 2007;356:217–26.
9. Arranz EE, Vara JБ, Gбmez-Pozo A, Zamora P. Gene signatures in breast cancer: current and future uses. Transl Oncol 2012;5:398–403.
10. Sшrlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.
11. Hou LK, Ma YS, Han Y, Lu GX, Luo P, Chang ZY, et al. Association of microRNA-33a molecular signature with non-small cell lung cancer diagnosis and prognosis after chemotherapy. PLoS One 2017;12:e0170431. [CrossRef]
12. Munding JB, Adai AT, Maghnouj A, Urbanik A, Zцllner H, Liffers ST, et al. Global microRNA expression profiling of microdissected tissues identifies miR-135b as a novel biomarker for pancreatic ductal adenocarcinoma. Int J Cancer 2012;131:E86–95.
13. Shenoy A, Blelloch RH. Regulation of microRNA function in somatic stem cell proliferation and differentiation. Nat Rev Mol Cell Biol 2014;15:565–76. [CrossRef]
14. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009;136:215–33. [CrossRef]
15. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet 2009;10:704–14. [CrossRef]
16. Dou C, Wang Y, Li C, Liu Z, Jia Y, Li Q, et al. MicroRNA-212 suppresses tumor growth of human hepatocellular carcinoma by targeting FOXA1. Oncotarget 2015;6:13216–28. [CrossRef]
17. Bandyopadhyay S, Mallik S, Mukhopadhyay A. A survey and comparative study of statistical tests for identifying differential expression from microarray data. IEEE/ACM transactions on computational biology and bioinformatics 2014;11:95–115.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
19. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article3. [CrossRef]
20. Xiang Y, Zhang CQ, Huang K. Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. BMC Bioinformatics 2012;13:S12.
21. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods 2009;6:83–90. [CrossRef]
22. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet 2011;12:529–41. [CrossRef]
23. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 2014;17:1156–63. [CrossRef]
24. Mallik S, Odom GJ, Gao Z, Gomez L, Chen X, Wang L. An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. Brief Bioinform 2019;20:2224–35. [CrossRef]
25. Hoshida Y, Villanueva A, Sangiovanni A, Sole M, Hur C, Andersson KL, et al. Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. Gastroenterology 2013;144:1024–30.
26. Verhaak RG, Goudswaard CS, van Putten W, Bijl MA, Sanders MA, Hugens W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood 2005;106:3747–54. [CrossRef]
27. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer 2014;14:177. [CrossRef]
28. Nguyen HG, Welty CJ, Cooperberg MR. Diagnostic associations of gene expression signatures in prostate cancer tissue. Curr Opin Urol 2015;25:65–70.
29. Baker SG, Kramer BS. Evaluating surrogate endpoints, prognostic markers, and predictive markers: Some simple themes. Clin Trials 2015;12:299–308. [CrossRef]
30. Zhang J, Lu K, Xiang Y, Islam M, Kotian S, Kais Z, et al. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. PLoS Comput Biol

2012;8:e1002656.

31. Mallik S, S. Bandyopadhyay S. Wecomxp: Weighted connectivity measure integrating co-methylation, co-expression and protein-protein interactions for gene-module detection. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2018;17:690–703. [CrossRef]
32. Jin D, Lee H. FGMD: A novel approach for functional gene module detection in cancer. PLoS One 2017;12:e0188900.
33. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics 2007;24:719–20. [CrossRef]
34. Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
35. Jiang X, Zhang H, Quan X, Liu Z, Yin Y. Disease-related gene module detection based on a multi-label propagation clustering algorithm. PLoS One 2017;12:e0178006.
36. Seth S, Mallik S, Bhadra T, Zhao Z. Dimensionality reduction and louvain agglomerative hierarchical clustering for cluster-specified frequent biomarker discovery in single-cell sequencing data. Front Genet 2022;13:828479.
37. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–5. [CrossRef]
38. Mallik S, Zhao Z. ConGEMs: condensed gene co-expression module discovery through rule-based clustering and its application to carcinogenesis. Genes (Basel) 2017;9:7.
39. Cava C, Colaprico A, Bertoli G, Graudenzi A, Silva TC, Olsen C, et al. SpidermiR: An R/Bioconductor package for integrative analysis with miRNA data. Int J Mol Sci 2017;18:274. [CrossRef]
40. Srivastava P, Mangal M, Agarwal SM. Understanding the transcriptional regulation of cervix cancer using microarray gene expression data and promoter sequence analysis of a curated gene set. Gene 2014;535:233–8. [CrossRef]
41. Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res 2010;38:D119–22.
42. Wang X, Tang S, Le SY, Lu R, Rader JS, Meyers C, et al. Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth. PLoS One 2008;3:e2557.
43. Von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007;17:395–416. [CrossRef]
44. Sommerová L, Fraňková H, Anton M, Jandáková E, Vojtěšek B, Hrstka R. Expression and functional characterization of miR-34c in cervical cancer. Klin Onkol 2018;31:82–7.
45. Yuan SH, Qiu Z, Ghosh A. TOX3 regulates calcium-dependent transcription in neurons. Proc Natl Acad Sci U S A 2009;106:2909–14. [CrossRef]
46. Miao H, Wang N, Shi LX, Wang Z, Song WB. Overexpression of mircoRNA-137 inhibits cervical cancer cell invasion, migration and epithelial-mesenchymal transition by suppressing the TGF-β/smad pathway via binding to GREM1. Cancer Cell Int 2019;19:147.
47. Park S, Kim J, Eom K, Oh S, Kim S, Kim G, et al. microRNA-944 overexpression is a biomarker for poor prognosis of advanced cervical cancer. BMC Cancer 2019;19:419.
48. Pett MR, Herdman MT, Palmer RD, Yeo GS, Shivji MK, Stanley MA, et al. Selection of cervical keratinocytes containing integrated HPV16 associates with episome loss and an endogenous antiviral response. Proc Natl Acad Sci U S A 2006;103:3822–7. [CrossRef]
49. Wu X, Peng L, Zhang Y, Chen S, Lei Q, Li G, et al. Identification of key genes and pathways in cervical cancer by bioinformatics analysis. Int J Med Sci 2019;16:800–12.
50. International Collaboration of Epidemiological Studies of Cervical Cancer. Cervical carcinoma and reproductive factors: collaborative reanalysis of individual data on 16,563 women with cervical carcinoma and 33,542 women without cervical carcinoma from 25 epidemiological studies. Int J Cancer 2006;119:1108–24. [CrossRef]
51. Nelson EL, Wenzel LB, Osann K, Dogan-Ates A, Chantana N, Reina-Patton A, et al. Stress, immunity, and cervical cancer: biobehavioral outcomes of a randomized clinical trial [corrected]. Clin Cancer Res 2008;14:2111–8.
52. Brinton LA, Hamman RF, Huggins GR, Lehman HF, Levine RS, Mallin K, et al. Sexual and reproductive risk factors for invasive squamous cell cervical cancer. J Natl Cancer Inst 1987;79:23–30.
53. Manzo-Merino J, Contreras-Paredes A, Vázquez-Ulloa E, Rocha-Zavaleta L, Fuentes-Gonzalez AM, Lizano M. The role of signaling pathways in cervical cancer and molecular therapeutic targets. Arch Med Res 2014;45:525–39. [CrossRef]
54. Ewald PW, Swain Ewald HA. Infection and cancer in multicellular organisms. Phil Trans R Soc B 2015;370:20140224.
55. Singh S, Kumar PU, Thakur S, Kiran S, Sen B, Sharma S, et al. Expression/localization patterns of sirtuins (SIRT1, SIRT2, and SIRT7) during progression of cervical cancer and effects of sirtuin inhibitors on growth of cervical cancer cells. Tumour Biol 2015;36:6159–71.
56. Sun L, Fang J. Macromolecular crowding effect is critical for maintaining SIRT1's nuclear localization in cancer cells. Cell Cycle 2016;15:2647–55.
57. Carter J, Auchincloss S, Sonoda Y, Krychman M. Cervical cancer: issues of sexuality and fertility. Oncology (Williston Park) 2003;17:1229–34.
58. Lin M, Ye M, Zhou J, Wang ZP, Zhu X. Recent advances on the molecular mechanism of cervical carcinogenesis based on systems biology technologies. Comput Struct Biotechnol J 2019;17:241–50. [CrossRef]
59. Therneau TM, Lumley T. Package 'survival'. R Top Doc 2015;128:28–33.
60. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26. [CrossRef]