RESEARCH

DeMoS: Dense Module-based Gene Signature Detection through Quasi-Clique: An Application for Cervical Cancer Prognosis

Suparna Saha[1,†], Saurav Mallik[2,†] and Sanghamitra Bandyopadhyay[1*]

*Correspondence:
sanghami@isical.ac.in
[1] Machine Intelligence Unit, Indian Statistical Institute, 700108 Kolkata, India
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Cervical cancer is a major cause of death among women. MicroRNA (miRNA) and its associated gene play a vital role in human cancer evolution. Identifying the gene signature is a crucial problem in bioinformatics. The gene signature represents the molecular alteration in a disease at specified phenotypic conditions. It is often used to differentiate samples into various groups for an improved research perspective as well as a clinical treatment. Although many methodologies and applications have been suggested in recent literature, efficient techniques that can consider the complex gene expression profile and be able to identify the most relevant signatures are required.

Methods: In this article, we present a new framework to identify dense module-based gene signatures (DeMoS) and their targeting miRNAs through the quasi-clique detection algorithm and discuss their application in a prognosis survival study. We used a cervical cancer data repository with prognosis clinical data to conduct our experiment. First, we performed the empirical Bayes test by using the linear model for microarray method to identify dysregulated genes or miRNAs. MiRNA-mediated dysregulated target genes were extracted from the dysregulated miRNAs. Thereafter, we detected dense co-expressed modules by using a quasi-clique identification technique. The average correlation coefficient was computed for each resultant module, and the module containing the highest correlation was formulated as the resultant gene signature. We then applied three well-known classifiers: support vector machine (SVM), prediction analysis for microarrays (PAM), and random forest (RF) using 10-fold cross-validation, and obtained the area under the curve (AUC). Finally, we conducted a prognosis survival study for the resultant gene signature.

Results: The resultant signature consisted of 10 genes: FGF9, FGF18, PPP1R9A, ERBB4, DCDC2, TOX3, ARMC3, DNALI1, RGL3, and ENPP3. Additionally, we identified eight dysregulated miRNAs that targeted the aforementioned gene signature. Hsa-mir-34c was strongly associated with the genes signature because its out-degree centrality score was the highest. On the other hand, the p-value of the Cox regression in the prognosis study for the resultant gene signature was significant (=4.2e-02). Finally, DeMoS evaluated the highest AUC values (0.95 for SVM, 0.955 for RF, and 0.955 for PAM) for the resultant gene signature compared with the other state-of-the-art techniques.

Conclusions: Our framework estimated the most promising gene signature that could classify multiple groups/subtypes of samples with higher AUCs as well as had a statistically significant p-value in regression-based prognosis analysis. Our method is useful for determining the signature for any microarray or RNA-Seq profile. The code is available at https://github.com/sahasuparna/DeMoS.

Keywords: Gene signature; Quasi-clique; co-expression; Limma; Disease classification; Prognosis survival study

Background

Identifying gene signatures from a genomic profile has become a growing concern in biomedical research in the last couple of decades. A gene signature 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 [1]. The gene signature offers numerous benefits in various cancers and their prognosis. According to Chanrion et al. [2], 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 for their relationship with overall and metastasis-free survival in patients with breast or other types of cancer [3]. 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 [4]. 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 [5]. These miRNAs are abnormally expressed in various malignancies as tumor suppressors or oncogenes [6]. They regulate various processes in carcinogenesis such as metastasis [7] and cell proliferation [8]. 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 [9]. Dysregulation of miRNAs is responsible for the formation and progression of tumors [10], [11]. 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 [12]. 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 [13], [14]. In graph theory, a clique is a complete graph in which each vertex is adjacent to each other. In general, a clique is used for selecting pairwise relationships in the gene regulatory network and protein–protein interaction. There are several types of cliques. Biclique [15] is a special type of bipartite graph where every vertex of the first set is connected to every vertex of the second set. Biclique has been applied in numerous contexts such as optimization problems for identifying the maximal biclique of a graph [16], [17] in covering the problem domain [18]. A quasi-clique, which is a generalized version of a clique, is preferable to a clique in dense subgraph detection and a robust connectivity-finding problem. Let us consider a graph $G = (V, E)$, where V denotes the set of all vertices,

while E is the set of all edges in G. A quasi clique say Q_c, is a subset of V, such that a subgraph induced by Qc consists of at least $\left\lfloor \frac{n(n-1)}{2} \right\rfloor$ edges, where n=$|Qc|$. In other words, finding quasi-cliques is a better method of mining co-functional genes in a large and scale-free gene co-expression network (GCN). 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 [19], and protein-protein interaction network [20].

In our paper, we introduce a new framework to determine dense module-based gene signatures (DeMoS) and their targeting miRNAs by identifying quasi-cliques and discuss 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 gene signature by using quasi-clique. 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, we applied the local maximal quasi-clique merger (lmQCM) [21] algorithm to the GCN comprising of the extracted common genes. The module with the highest average correlation is termed as the resultant gene signature. In addition, we conducted a comparative study between the resultant gene signatures obtained using our method and those obtained using other existing methods. Because our resultant gene signature produced a significant p-value for the prognosis survival analysis, it could be called a clinically promising signature. Moreover, our proposed method is highly effective in identifying the molecular signature from any microarray or RNA-Seq profile.

Results

The gene expression profile consisted of 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 discarded 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 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. 3(a) and Figure. 3(b), respectively. The volcano plots in Figure. 3(c) and Figure. 3(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 the volcano plot. The boxplot on both the gene expression data before and after Voom transformation were shown in Figure 4(a) and Figure 4(b), respectively.

We then identified the significant target genes that also existed in significantly expressed genes in the underlying dataset. Notably, the total number of over-expressed genes in SCC targeted by the over-expressed miRNAs in SCC was 43, while the total number of under-expressed genes in SCC (also termed as over-expressed genes in ADENO) targeted by the over-expressed miRNAs in SCC was 41. On the other hand, the total number of over-expressed genes in SCC targeted by the under-expressed miRNAs in SCC (i.e., over-expressed miRNAs in ADENO) was 22, whereas the total number of under-expressed genes in SCC (also called as over-expressed genes in ADENO) targeted by the under-expressed miRNAs in SCC (also termed as over-expressed miRNAs in ADENO) was 37. Finally, a total of 143 predicted target genes (=43+41+22+37) were determined for 18 significantly expressed miRNAs. Thereafter, we obtained two dense gene modules through LmQcm technique. From these two modules, one module was found to be significant (called as signature) which consisted of 10 genes. Intuitively, we observed the network in Figure 2 to identify those miRNAs that were associated with those genes belonging to the signature. Those miRNAs were hsa-mir-34b (p-value=2.96e-04), hsa-mir-34c (p-value=1.35e-04), hsa-mir-615 (p-value=1.55e-04), hsa-mir-137 (p-value=2.93e-12), hsa-mir-1910 (p-value=4.87e-06), hsa-mir-375 (p-value=4.51e-25), hsa-mir-577 (p-value=1.25e-13), hsa-mir-215 (p-value=4.42e-18) and hsa-mir-548j (p-value=5.45e-10).

Intuitively, we obtained a gene signature consisting of ten genes with the p-values in the differential analysis were: FGF9 (p-value=1.42e-08), FGF18 (p-value=1.73e-14), PPP1R9A (p-value=7.66e-69), ERBB4 (p-value=9.20e-45), DCDC2 (p-value=5.68e-46), TOX3 (p-value=1.09e-49), ARMC3 (p-value=2.17e-36), DNALI1 (p-value=1.00e-58), RGL3 (p-value=1.18e-47) and ENPP3 (p-value=2.51e-14). We also identified eight miRNAs those were associated with the gene signature. Among them, hsa-mir-34c was strongly associated with the signature as its out-degree is the highest in the network as mentioned in the Figure 2. The miRNA hsa-mir-34c has differential effects in migration and cell proliferation in Cervical Cancer was already proved in the literature [22].

Furthermore, we conducted KEGG pathway and Gene Ontology (GO) analyses using the DAVID database. Here, we observed that the genes belonging to the signature followed several significant biological processes, as mentioned in Table. 1; e.g., the genes, ERBB4 and FGF9 were involved in positive regulation of vascular endothelial growth factor receptor signalling pathway (p-value= 3.140e-03). ERBB4 and FGF9 were associated with the positive regulation of cardiac muscle cell proliferation (p-value= 5.88e-03). FGF18 and FGF9 followed the fibroblast growth factor receptor signalling pathway (p-value=1.29e-02), FGF18 and

FGF9 were linked in angiogenesis (p-value=3.907e-02). In contrast, the genes FGF18 and ERBB4 were associated with positive regulation of ERK1 and ERK2 cascade (p-value=5.081e-02). Of note, the correlation of regulation of vascular endothelial growth factor receptor with cancer cells was already well established for a diverse of malignant tumors, inclusive of CESC [23]. Therefore, the significant enrichment of the biological process `positive regulation of vascular endothelial growth factor receptor signalling pathway` had a meticulous connection with cancers originating from the cervix. The other notably enriched biological processes like `Positive regulation of cardiac muscle cell proliferation` was also known to be involved in the development of different cancer, including CESC [24]. Moreover, the association between CESC and `Fibroblast growth factor receptor signalling pathway` was already in the literature [25]. The expression of members of the Fibroblast growth factor receptor family implied prognostic concernment in early-stage cervical cancer patients. The complex biological process `Angiogenesis` plays a vital role in the development of cancer. There was evidence of association of CESC with Angiogenesis [26] in the presence of human papillomavirus (HPV). Another biological process of ERK1 and ERK2 activation are essential for the development and progression of cancer. Moreover, cascade (ERK1 and ERK2) had a specific association with HPV and CESC [27].

In addition, we explored the survival prognosis analysis for our gene signature. For survival analysis, we used Cox proportional hazards regression model to predict the survival time of the underlying patients (dead or alive) concerning our gene signature. For the living and deceased patients, we extracted the days from the last follow-up time and overall survival time, respectively. We obtained Cox regression p-value for each gene belonging to the signature. Among them, we obtained nine significant p-values (<0.05), while one was insignificant. E.g., for the gene FGF9, the p-value was 1.8e-02 (significant). Similarly, in the case of the gene FGF18, the p-value was 1.0e-02 (significant), whereas, for the gene PPP1R9A, it was 4.7e-04 (significant). For the cases of remaining genes (ERBB4, DCDC2, TOX3, ARMC3, DNALI1, RGL3 and ENPP3), p-values were 2.0e-01 (insignificant), 5.2e-03 (significant), 2.5e-02 (significant), 8.0e-03 (significant), 1.6e-03 (significant), 9.0e-03 (significant) and 1.3e-02 (significant), respectively. For these individual gene-wise survival analyses, the corresponding plots were illustrated in Figure 6(a)-(j). Additionally, we performed survival prognosis study and corresponding Cox regression for all those ten genes together belonging to the signature and then obtained significant p-value (p-value=4.2e-02). The survival plot for this signature (integrated case) was provided in Figure 6(k). Overall, since we found the significant p-value for the integrated case as well as the most of individual survival cases (nine out of ten cases), it implies that our resultant gene signature was powerful enough to say clinically promising.

In addition, we carried out the performance analysis of our resultant gene signature by estimating the prediction accuracy of the cervical cancer subtypes (Adenocarcinoma and Squamous cell carcinoma). There were several gene signatures of cervical cancer available in the literature. Hence, we provided here a comparative study of the AUC score of the gene signature with the related previous signature found in the literature. For two-class classification, we used Support Vector Machine (SVM), Random Forest (RF) and Prediction Analysis for Microarrays (PAM)

classifiers. In contrast, the Area under the Curve (AUC) was used as a performance metric. We obtained the highest AUC values (viz., 0.95 for SVM, 0.955 for RF, 0.955 for PAM) for our resultant gene signature estimated by our proposed method, DeMoS. On the other hand, for the gene signature obtained by Li et al. (2017) [6], the corresponding AUC values across these three classifiers were 0.535 for SVM, 0.5625 for RF and 0.585 for PAM, whereas for the gene signature found by Li et al. (2018) [28], those values were 0.535 for SVM, 0.5625 for RF and 0.585 for PAM. Another gene signature estimated by Huang et al. (2011) [29], the AUC values for SVM, RF and PAM classifiers were 0.83, 0.66 and 0.75, respectively. Figure 5 represented the comparative analysis. Finally, our proposed method DeMoS produced the best AUC score to classify the cervical cancer subtypes among the state-of-the-art methods. Moreover, our method is useful and powerful enough to identify a molecular signature from RNA-seq or similar data.

Discussion

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 dense module based signature, DeMoS used gene co-expression network to determine the local maximal quasi clique and finally extracted the significant dense gene module. The resultant gene module was termed as gene signature. Overall, DeMoS 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 DeMoS are the highest across the three classifiers presented in Fig. 5. (iii) The resultant gene signature was found clinically validated since it produced significant p-value in cox regression-based survival analysis for integrated study along with most of the individual gene-based survival studies.

Conclusion

In this article, we developed a new framework to extract dense module-based gene signature and their targeting miRNAs through Quasi-Clique 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 co-expressed modules using Quasi-Clique identification method. We then computed the average correlation for each resultant module. The module that contained the highest average correlation was considered as the resultant gene signature. The signature consisted of ten genes, FGF9, FGF18, PPP1R9A, ERBB4, DCDC2, TOX3, ARMC3,

DNALI1, RGL3 and ENPP3. A total of eight dysregulated miRNAs that targeted the genes belonging to the signature was also identified. DeMoS produced the best AUC values (>=0.95 for all classifiers) for our resultant gene signature in compared to the other state-of-the-art algorithms. In addition, the Cox regression analysis in the prognosis study for the resultant gene signature was found to be p-value significant (=4.2e-02). 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, 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 [30], [31]. In addition, We found the comparative study of various DMR finding methods in [32] that might motivate us to extend our future work.

Methods

Literature review

Since the last two decades, gene signatures are widely used in omics data analysis. In this article, we propose a framework that can identify DeMoS and their targeting miRNAs through a quasi-clique method 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 [30], [31]. Mallik et al. [32] 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 [33], leukemia [34], and breast cancer [35]. 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 [36]. A predictive gene signature predicts the outcome of therapeutic intervention and does not depend on the prognosis [37]. Hence, these signatures contain crucial information.

Several highly efficient biological networks can be used to predict the new functionality of genes [38]. 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 [39]. 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) [40] 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 [41]. 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 [42]. 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 [43].

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 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) [44] is computed on the basis of the power of the PCC values. The gene modules are identified on the basis of the dissimilarity score, that is, (=1-TOM score). The TOM score between two nodes in an adjacency matrix is defined as follows.

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

By observing formula (1), the generalization of TOM can be expressed as follows.

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

Here, Y denotes the adjacency matrix of the nodes containing the first nearest neighbours of each nodes, '$|.|$' symbol indicates the cardinality of the set, $nbd_1(j)$ signifies the set of neighbours of j excluding itself (i.e., j), and $nbd_1(k)$ signifies the set of neighbours of k excluding itself (i.e., k), whereas $|nbd_1(j) \cap nbd_1(k)|$ 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 dense module-finding algorithm by merging the local maximal quasi-clique 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 lmQCM algorithm was then applied to find the dense modules, and the average correlation coefficient was calculated for the resultant modules. The best module can be considered as the gene signature. We used three consecutive classifiers to classify the class labels by using all the features belonging to the signature. Compared with the other methods, DeMoS produced the highest accuracy across all the classifiers. Additionally, we conducted a gene enrichment analysis [Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway] to identify disease-related pathways as well as GO terms for the participating genes belonging to the signature. We also computed the degree centrality of the network and identified 10 miRNAs to be associated with the gene signature. Our framework may prove useful for extracting gene signatures for other microarrays/RNA-Seq datasets for cancer or any other disease.

Proposed Gene Signature Discovery Technique

In this article, we introduce a novel framework for detecting DeMoS and their targeting miRNAs through a quasi-clique detection methodology and discuss their application in a prognosis survival study. We performed an integrative analysis of the mRNA and miRNA from the TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma 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, [14] 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.

$$\widetilde{t_g} = (\sqrt{(D_0 + D_g)/D_g}) * (\widetilde{\beta_g}/\sqrt{S^2_{*,g}\, V_g}) \tag{3}$$

The degree of freedom is D_0+D_g, $\widetilde{\beta_g}$ is the contrast estimator, and $\widetilde{S^2_{*,g}}$ 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., $FC \geq 2$), whereas for the downregulated gene probes (DG) and downregulated miRNAs ($DMIR$), the threshold value of FC is 0.5 (i.e., $FC \leq 0.5$). In this manner, the UG), UMIR, DG, and DMIR were selected on the basis of p and FC values. The overview of the workflow of DeMoS is presented in Figure 1.

Prediction of target genes

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 [45] 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.

Detection of dense modules and gene signature

We attempted to extract the dense gene module from the given gene expression profiles containing a set of samples and with the resultant target genes. A GCN

can be established by considering each gene as a node, and the correlation of the expression score between any gene pair annotating the edge between them. lmQCM generated the dense gene modules in the network by merging the quasi-cliques existing on the WGCN. With an undirected WGCN, G1 ={V,E,W} where V is the set of vertices, i.e., V = $\{v_1,v_2,...,v_n\}$, where n is the number of vertices in the weighted network, and W = $[w_{ij}]$ with $w_{ij} \geq 0$, while $w_{ij} = 0$ states the non-negative weights of the edges e_{ij} (i.e., no self-loop is allowed). The density of the graph Den(G1) is defined as follows.

$$Den(G1) = \frac{\sum_{j=i+1}^{n} \sum_{i=1}^{n} w_{ij}}{\frac{n(n-1)}{2}}. \quad (4)$$

The algorithm depends on four input parameters: σ, μ, xi, and η. The symbols are denoted as follows:

- σ is used for controlling the thresholds at the initial stage of each new module,
- μ and xi represent the adaptive thresholds of module density to confirm the appropriate stopping criterion, and
- η is used as the threshold for measuring the overlapping ratio while merging the two modules, that is, if M1 and M2 are two overlapping modules; they will be merged based on the threshold value , that is, M1 and M2 will be overlapped if $\frac{|M1 \cap M2|}{min\{|M1|,|M2|\}} > \eta$.

We obtained some dense modules by using lmQCM. Thereafter, we computed the average correlation of the individual modules and identified the module with the highest average correlation, which was treated as the potential gene signature. 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 [45] 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.

Classification model of gene signature

We extracted the data for all the evolved genes (features) belonging to the gene signature, 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 gene signatures obtained using our proposed method as well as other existing methods to compare their performance.

Gene set enrichment analysis

In addition, we conducted the KEGG 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.

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 [46] 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 (DeMoS). 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 KaplanMeier 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 signature, and the frequency of a significant p-value was obtained. This resultant signature was termed the clinically verified potential gene signature.

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. 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.

Abbreviations

ADENO	Adeno Carcinoma
CESC	Cervical cancer
CV	Cross Validation
DEG	Significantly expressed genes
DG	Down-regulated gene probes
DLPCA	Double-Label Propagation Clustering Algorithm
DMR	Differentially Methylated Regions
FC	Fold Change
FGMD	Functional Gene Module Detection
GCN	Gene Co-expression Network
GRNs	Gene regulatory network
HPV	human papillomavirus
IGS	Invasiveness genesignature
miRNAs	MicroRNAs
MLPA	Multilabel Clustering Algorithm
PAM	Prediction Analysis for Microarrays
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction
RF	Random Forest
ROC	Receiver-Operating Characteristic Curves
SCC	Squamous Cell Carcinoma
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
TOM	Topological Overlap Matrix
UG	Up-regulated gene probes
WGCNA	Weighted Correlation Network Analysis

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SM and SS devised the study and coordinated this research. SM and SS designed the experiments. SS carried out the experiments and data analysis. SM, SS and SB drafted the manuscript. All the authors read and approved the final manuscript.

Acknowledgments

We would like to thank all the lab members of Indian Statistical Institute, Kolkata, India.

Funding

The work is partially supported by the SyMec project (grant number BT/Med-II/NIBMG/SyMeC/2014/Vol.II), funded by the Department of Biotechnology, Government of India.

availability of data and materials

The codes are made freely available through Github (https://github.com/sahasuparna/DeMoS)

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author details

[1] Machine Intelligence Unit, Indian Statistical Institute, 700108 Kolkata, India. [2] Center of Precision Health, Department of School of Biomedical Informatics, University of Texas Health Science Center , Houston,Texas, USA.

References

1. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., *et al.*: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113 (2013)
2. Chanrion, M., Negre, V., Fontaine, H., Salvetat, N., Bibeau, F., Mac Grogan, G., Mauriac, L., Katsaros, D., Molina, F., Theillet, C., *et al.*: A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. Clinical Cancer Research **14**(6), 1744–1752 (2008)
3. Liu, R., Wang, X., Chen, G.Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., Clarke, M.F.: The prognostic role of a gene signature from tumorigenic breast-cancer cells. New England Journal of Medicine **356**(3), 217–226 (2007)
4. Arranz, E.E., Vara, J.Á.F., Gámez-Pozo, A., Zamora, P.: Gene signatures in breast cancer: current and future uses. Translational oncology **5**(6), 398–403 (2012)
5. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., *et al.*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences **98**(19), 10869–10874 (2001)

6. Hou, L.-K., Ma, Y.-S., Han, Y., Lu, G.-X., Luo, P., Chang, Z.-Y., Xie, R.-T., Yang, H.-Q., Chai, L., Cai, M.-X., *et al.*: Association of microrna-33a molecular signature with non-small cell lung cancer diagnosis and prognosis after chemotherapy. PloS one **12**(1), 0170431 (2017)
7. Munding, J.B., Adai, A.T., Maghnouj, A., Urbanik, A., Zöllner, H., Liffers, S.T., Chromik, A.M., Uhl, W., Szafranska-Schwarzbach, A.E., Tannapfel, A., *et al.*: Global microrna expression profiling of microdissected tissues identifies mir-135b as a novel biomarker for pancreatic ductal adenocarcinoma. International journal of cancer **131**(2), 86–95 (2012)
8. Shenoy, A., Blelloch, R.H.: Regulation of microrna function in somatic stem cell proliferation and differentiation. Nature Reviews Molecular Cell Biology **15**(9), 565 (2014)
9. Bartel, D.P.: Micrornas: target recognition and regulatory functions. cell **136**(2), 215–233 (2009)
10. Croce, C.M.: Causes and consequences of microrna dysregulation in cancer. Nature reviews genetics **10**(10), 704 (2009)
11. Dou, C., Wang, Y., Li, C., Liu, Z., Jia, Y., Li, Q., Yang, W., Yao, Y., Liu, Q., Tu, K.: Microrna-212 suppresses tumor growth of human hepatocellular carcinoma by targeting foxa1. Oncotarget **6**(15), 13216 (2015)
12. 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 **11**(1), 95–115 (2014)
13. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research **43**(7), 47–47 (2015)
14. Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology **3**(1), 1–25 (2004)
15. Groshaus, M., Szwarcfiter, J.L.: Biclique graphs and biclique matrices. Journal of Graph Theory **63**(1), 1–16 (2010)
16. Peeters, R.: The maximum edge biclique problem is np-complete. Discrete Applied Mathematics **131**(3), 651–654 (2003)
17. Yannakakis, M.: Node-and edge-deletion np-complete problems. In: Proceedings of the Tenth Annual ACM Symposium on Theory of Computing, pp. 253–264 (1978). ACM
18. Amilhastre, J., Vilarem, M.-C., Janssen, P.: Complexity of minimum biclique cover and minimum biclique decomposition for bipartite domino-free graphs. Discrete applied mathematics **86**(2-3), 125–144 (1998)
19. Xiang, Y., Zhang, C.-Q., Huang, K.: Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on tcga data. In: BMC Bioinformatics, vol. 13, p. 12 (2012). BioMed Central
20. Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., *et al.*: An empirical framework for binary interactome mapping. Nature methods **6**(1), 83 (2008)
21. Zhang, J., Huang, K.: Normalized lmqcm: an algorithm for detecting weak quasi-clique modules in weighted graph with application in functional gene cluster discovery in cancer. Cancer Inform **1**(1) (2016)
22. Córdova-Rivas, S., Fraire-Soto, I., Mercado-Casas Torres, A., Servín-González, L.S., Granados-López, A.J., López-Hernández, Y., Reyes-Estrada, C.A., Gutiérrez-Hernández, R., Castañeda-Delgado, J.E., Ramírez-Hernández, L., *et al.*: 5p and 3p strands of mir-34 family members have differential effects in cell proliferation, migration, and invasion in cervical cancer cells. International journal of molecular sciences **20**(3), 545 (2019)
23. He, M., Cheng, Y., Li, W., Liu, Q., Liu, J., Huang, J., Fu, X.: Vascular endothelial growth factor c promotes cervical cancer metastasis via up-regulation and activation of rhoa/rock-2/moesin cascade. BMC cancer **10**(1), 170 (2010)
24. Mitchelson, K.R., Qin, W.-Y.: Roles of the canonical myomirs mir-1,-133 and-206 in cell development and disease. World journal of biological chemistry **6**(3), 162 (2015)
25. Choi, C.H., Chung, J.-Y., Kim, J.-H., Kim, B.-G., Hewitt, S.M.: Expression of fibroblast growth factor receptor family members is associated with prognosis in early stage cervical cancer patients. Journal of translational medicine **14**(1), 124 (2016)
26. Tomao, F., Papa, A., Rossi, L., Zaccarelli, E., Caruso, D., Zoratto, F., Panici, P.B., Tomao, S.: Angiogenesis and antiangiogenic agents in cervical cancer. OncoTargets and therapy **7**, 2237 (2014)
27. Branca, M., Ciotti, M., Santini, D., Di Bonito, L., Benedetto, A., Giorgi, C., Paba, P., Favalli, C., Costa, S., Agarossi, A., *et al.*: Activation of the erk/map kinase pathway in cervical intraepithelial neoplasia is related to grade of the lesion but not to high-risk human papillomavirus, virus clearance, or prognosis in cervical cancer. American journal of clinical pathology **122**(6), 902–911 (2004)
28. Li, X., Tian, R., Gao, H., Yan, F., Ying, L., Yang, Y., Yang, P., Gao, Y.: Identification of significant gene signatures and prognostic biomarkers for patients with cervical cancer by integrated bioinformatic methods. Technology in cancer research & treatment **17**, 1533033818767455 (2018)
29. Huang, L., Zheng, M., Zhou, Q.-M., Zhang, M.-Y., Jia, W.-H., Yun, J.-P., Wang, H.-Y.: Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. Cancer **117**(15), 3363–3373 (2011)
30. Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S.: Epigenome-wide association studies for common human diseases. Nature Reviews Genetics **12**(8), 529 (2011)
31. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., *et al.*: Alzheimer's disease: early alterations in brain dna methylation at ank1, bin1, rhbdf2 and other loci. Nature neuroscience **17**(9), 1156 (2014)
32. Mallik, S., Odom, G.J., Gao, Z., Gomez, L., Chen, X., Wang, L.: An evaluation of supervised methods for identifying differentially methylated regions in illumina methylation arrays. Briefings in bioinformatics (2018)
33. Hoshida, Y., Villanueva, A., Sangiovanni, A., Sole, M., Hur, C., Andersson, K.L., Chung, R.T., Gould, J., Kojima, K., Gupta, S., *et al.*: Prognostic gene expression signature for patients with hepatitis c–related early-stage cirrhosis. Gastroenterology **144**(5), 1024–1030 (2013)
34. Verhaak, R.G., Goudswaard, C.S., van Putten, W., Bijl, M.A., Sanders, M.A., Hugens, W., Uitterlinden, A.G.,

Erpelinck, C.A., Delwel, R., Löwenberg, B., *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 **106**(12), 3747–3754 (2005)
35. Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D., Barry, G., Dowidar, N., Maysuria, M., Storhoff, J.: 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 **14**(1), 177 (2014)
36. Nguyen, H.G., Welty, C.J., Cooperberg, M.R.: Diagnostic associations of gene expression signatures in prostate cancer tissue. Current opinion in urology **25**(1), 65–70 (2015)
37. Baker, S.G., Kramer, B.S.: Evaluating surrogate endpoints, prognostic markers, and predictive markers: some simple themes. Clinical Trials **12**(4), 299–308 (2015)
38. Zhang, J., Lu, K., Xiang, Y., Islam, M., Kotian, S., Kais, Z., Lee, C., Arora, M., Liu, H.-w., Parvin, J.D., *et al.*: Weighted frequent gene co-expression network mining to identify genes involved in genome stability. PLoS computational biology **8**(8), 1002656 (2012)
39. Mallik, 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, 1–1 (2018). doi:10.1109/TCBB.2018.2868348
40. Jin, D., Lee, H.: Fgmd: A novel approach for functional gene module detection in cancer. PloS one **12**(12), 0188900 (2017)
41. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics **24**(5), 719–720 (2007)
42. Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics **9**(1), 559 (2008)
43. 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 **12**(5), 0178006 (2017)
44. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks. science **297**(5586), 1551–1555 (2002)
45. Cava, C., Colaprico, A., Bertoli, G., Graudenzi, A., Silva, T.C., Olsen, C., Noushmehr, H., Bontempi, G., Mauri, G., Castiglioni, I.: Spidermir: an r/bioconductor package for integrative analysis with mirna data. International journal of molecular sciences **18**(2), 274 (2017)
46. Therneau, T.M., Lumley, T.: Package 'survival'. R Top Doc **128** (2015)

Figures

Figure 1 Flowchart of DeMoS for identifying gene signature.

Figure 2 A miRNA-Gene interaction network of significantly expressed miRNAs and their predicted target genes through DeMoS.

Figure 3 (a) Voom: Mean-Variance trend plot during the extraction of significantly expressed genes in Adenocarcinoma vs Squamous cell carcinoma, (b) Voom: Mean-Variance trend plot during the extraction of significantly expressed miRNA in Adenocarcinoma vs Squamous cell carcinoma, (c) Volcano plot for finding significantly expressed genes in Adenocarcinoma vs Squamous cell carcinoma, (d) Volcano plot for finding significantly expressed miRNA in Adenocarcinoma vs Squamous cell carcinoma.

Figure 4 a) Performing the Box plots before and after Voom normalization while extracting the significantly expressed genes. b) Performing the Box plots before and after Voom normalization while extracting the significantly expressed microRNAs.

Figure 5 Comparative study of area under curve (AUC) scores through the use of multiple classifiers (viz., SVM, RF and PAM).

Tables

Figure 6 From (a) - (j) Survival analysis of each participating gene in the Signature: Kaplan-Meier plots and Log-Rank Test p-values comparing overall survival times. (k) The integrated Kaplan-Meier plots and Log-Rank Test p-values for all participating genes in the signature.

Table 1 Enriched Gene Ontology (GO) terms for the participating significantly expressed genes belonging to the gene signature (named as DE_{sig}) for our proposed method, where "BP" denotes Biological Process.

GO ID	**GO Name**	**Category**	$\#DE_{sig}$	**Names of** DE_{sig}	**Enrichment p-value**
GO:0030949	Positive regulation of vascular endothelial growth factor receptor signaling pathway	GO:BP	2	FGF18, FGF9	3.140e-03
GO:0060045	Positive regulation of cardiac muscle cell proliferation	GO:BP	2	ERBB4, FGF9	5.88e-03
GO:0008543	Fibroblast growth factor receptor signaling pathway	GO:BP	2	FGF18, FGF9	1.29e-02
GO:0001525	Angiogenesis	GO:BP	2	FGF18, FGF9	3.907e-02
GO:0070374	Positive regulation of ERK1 and ERK2 cascade	GO:BP	2	FGF18, ERBB4	5.081e-02