# Stage-Specific Co-expression Network Analysis for Cancer Biomarker Discovery

Raihanul Bari Tanvir
*School of Computing and Information Sciences*
*Florida International University*
Miami. Florida
rtanv003@fiu.edu

Ananda Mohan Mondal
*School of Computing and Information Sciences*
*Florida International University*
Miami. Florida
amondal@fiu.edu

***Abstract*— Identification of conserved gene network modules in different stages of cancer may lead to uncovering mechanisms behind cancer initiation and progression. This work is based on two hypotheses. *Hypothesis-1*: the network modules conserved in all cancer stages are potential biomarkers related to the trajectory of cancer development or progression of cancer from initiation to stage-to-stage to metastasis. *Hypothesis-2*: The network modules from a stage, which are not conserved in other stages, can be considered as the stage-specific biomarkers for diagnosis.**

**To test the hypotheses, gene expression and clinical data of Breast Invasive Carcinoma (BRCA) from The Cancer Genome Atlas (TCGA) were used for analysis. Gene expression data was divided into five groups - stage I to stage IV and normal tissue samples. First, the co-expression networks for each of the four stages and normal samples were generated. Second, the modules from each of the stage-specific networks were discovered using weighted gene co-expression network analysis (WGCNA). Third, survival analysis was performed to identify the prognostically significant modules. Fourth, module preservation analysis was performed to determine whether a module from one stage is preserved in other cancer stages as well as in normal stage. Finally, gene ontology and pathway enrichment analyses were performed for the prognostically significant and conserved modules.**

**The present study discovered several gene-network modules for breast cancer preserved in all cancer stages and are significant in overall survival; hence, they can be considered potential biomarkers for cancers, related to the trajectory of cancer development. The modules that were found not to be conserved in different stages can be considered as stage-specific biomarkers.**

***Keywords*— *Cancer Biomarker, Gene Co-expression, Module Preservation, Network Biomarker***

## I. INTRODUCTION

Gene expression profiles across cancer samples have been widely used to find cancer biomarkers [1]–[9], heterogeneity in cancer [10], signatures for cancer diagnosis [11]–[13], progression [14], prognosis [15], and therapeutic strategies [16]. Different methods and algorithms have been proposed to analyze gene expression profiles. One such method is the use of co-expression networks [1]–[3], [9], [17]–[19]. In a gene co-expression network, nodes represent genes and edges represent pairs of genes, denoting the strength of correlation larger than a threshold. The threshold can be soft or hard, leading to a weighted or a binary network [20]. A co-expression network is a gene network built based on "guilt by association" [21]. Several methods exist for inferring edges in a co-expression network. Among them, Pearson Correlation is the most widely used method to calculate the strength of the correlation between pairs of genes [1]–[3], [9], [20], [22]–[25]. Other methods include Spearman Rank Correlation [26], [27], Biweight Midcorrelation [28], etc.

A co-expression network has topological structures reflecting real gene interactions [29]. Highly connected groups of genes with a higher similarity among themselves can be considered biologically significant modules, which perform a task of interest [30]. Different methods have been proposed or applied to gene co-expression networks to find biomarker modules in various diseases. Concepts and algorithms from graph theory and network science such as clique-bipartite structure [1], [2], and community detection [5] were applied on co-expression networks of multiple cancers to find network biomarkers. Machine learning methods such as clustering [20] and bi-clustering [31] were also used to find biologically important clusters from gene co-expression networks. The WGCNA [25] is a widely used popular tool for co-expression analysis and has been used in gene expression data of breast [32], bone (osteosarcoma) [33], cervical [34], and esophageal [35] cancer.

The studies mentioned above used the whole cohort of cancer patients to do the co-expression analysis by proposing a novel method or using existing methods for biomarker module detection. A few works have been done based on cohorts divided into many groups based on some clinical traits, such as cancer pathological stages. Li *et al.* performed an analysis of structural changes between consecutive stage-specific co-expression networks of four cancers using Jacquard Similarity and Rank Analysis and found changes in the network topology across stages and types [36]. Sarathi and Palaniappan [7] discovered stage-specific differentially expressed genes in hepatocellular carcinoma using the linear model by fitting expression profiles of each gene at different stages, including normal. Huo *et al.* [37] performed a linear regression analysis on expression profiles of each gene for Colorectal Carcinoma (CRC) to find the gene expression trend with increasing cancer stages and discovered a total of 11 genes of interest. In another study, the evolutionary process was constructed, and the relevant evolutionary paths related to the stage-specific prognosis of kidney cancer were discovered by applying Bayesian Mutation Landscape [38] using stage-specific cross-sectional single nucleotide variants (SNV) [39]. In a study by Palaniappan *et al.*, the stage-specific protein-protein interaction networks were constructed using the driver genes corresponding to each stage of CRC and from which hub genes were extracted and were claimed to have the capability to pinpoint the progression of CRC [40].

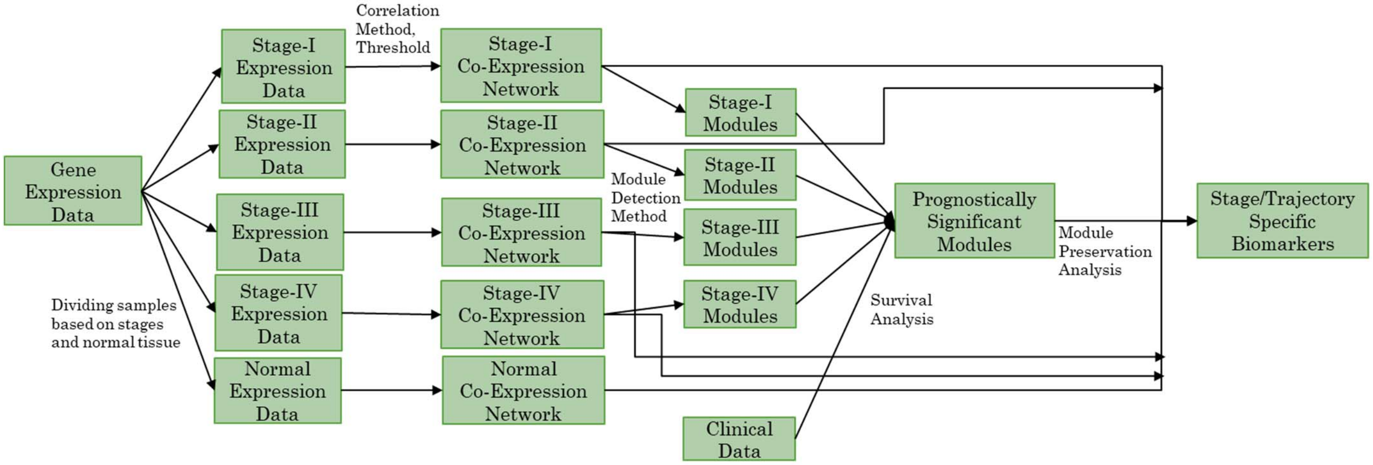
Gene Expression Data
Dividing samples based on stages and normal tissue
Stage-I Expression Data
Stage-II Expression Data
Stage-III Expression Data
Stage-IV Expression Data
Normal Expression Data
Correlation Method, Threshold
Stage-I Co-Expression Network
Stage-II Co-Expression Network
Stage-III Co-Expression Network
Stage-IV Co-Expression Network
Normal Co-Expression Network
Module Detection Method
Stage-I Modules
Stage-II Modules
Stage-III Modules
Stage-IV Modules
Clinical Data
Survival Analysis
Prognostically Significant Modules
Module Preservation Analysis
Stage/Trajectory Specific Biomarkers

**Fig 1.** Workflow of the study. Input: Gene expression data from TCGA. First: expression data was divided into five groups – stage I to IV and normal. Second: Co-expression networks were generated for each of the stages and normal samples. Third: Modules are identified from four stages of cancer. Fourth: Survival analysis was done for each module. Finally: Module preservation test was performed.

In Another study by Pradhan *et al.* [41], DNA methylation patterns were observed across different stages of Lung Adenocarcinoma (LUAD) by constructing a stage-specific subnetwork using a protein-protein interaction network, differentially expressed genes, and differentially DNA methylated genes. Then network analysis were performed on these stage-specific subnetworks to find essential hub genes, which were prognostic targets and conserved genes across all stages. In a study by Lalremmawia *et al.* [42], the stage-specific co-expression networks with a set of query genes were constructed for ovarian cancer (OV). The highly ranked genes co-expressed with query genes were analyzed further using support vector machine (SVM), resulting in 17 potential biomarker genes.

Amongst these stage-specific studies, there was no study that included the normal samples in their analysis. Additionally, there were not any studies that considers the stage-specific network modules that happens to be preserved in any other stages, which makes our work different from existing works. The flowchart of this study is shown in Figure 1. First, the co-expression networks for each tumor-stage samples and normal samples were created, and then the network modules were detected using WGCNA [25]. Second, survival analysis was done to find the prognostically significant modules. Third, module preservation analysis (MPA) was conducted on the prognostically significant modules or biomarkers to find the preserved ones. Prognostically significant modules which happens to be preserved in all the cancer stages are considered as trajectory-specific biomarker modules and not persevered in any other stages are considered as stage-specific biomarker modules. Then, the biological significance of these biomarker modules was identified using GO term and pathway enrichment analyses.

## II. Materials and Methods

### A. Dataset Preparation

The RNAseq gene expression data and clinical data for BRCA (Breast Invasive Carcinoma) are downloaded from UCSC Xena Browser [43]. The gene expression dataset contains expression profiles of 20,530 genes for 1,218 samples. The gene expression values are log2 transformed RSEM normalized count. The number of tumor, normal, and metastatic tissue samples are 1097, 114, and 7 respectively. The tumor samples are divided into four groups based on cancer stage information. The distribution of patients in four different stages and the number of genes with low expression in each stage are shown in Table I.

Table I. Summary of gene expression data preprocessing. Column 2: Stage-specific sample distribution. Columns 4 & 5: Finding insignificant genes applying threshold, Normalized Count ≤ 10 for 90% sample. Column 6: Remaining genes used for generating co-expression networks.

| | | **Number of Genes** | | | |
|---|---|---|---|---|---|
| | | | **Normalized≤10, 90% sample** | | |
| **Stage** | **Samples** | **Original** | **Each stage** | **Common** | **Remaining** |
| I | 182 | 20530 | 4460 | 3784 | 16746 |
| II | 621 | | 4654 | | |
| III | 250 | | 4565 | | |
| IV | 20 | | 4458 | | |
| Normal | 114 | | 4076 | | |

According to WGCNA package FAQ [44], genes with low expression values provide false correlations among themselves and contribute to unnecessary edges in the co-expression network. For each stage-specific gene expression data, the genes with normalized count ≤10 for 90% of the samples are selected, and different numbers of such genes are found in different stages, as shown in Table I. The common 3784 genes are removed from each stage-specific gene expression data. So, the final dataset contains gene expression profiles of 16746 genes.

### B. *Co-expression Network Construction and Module Detection*

WGCNA (Weighted Gene Co-expression Network Analysis) package [25] in R language was used to construct each stage's Co-expression network. Soft thresholding was used to create the co-expression networks. The optimal soft threshold (power), β, was identified by evaluating its effects on the scale-free topology model fit index $R^2$, by varying the value of soft threshold (power) ranging from 1 to 30. This threshold was selected to pick the lowest value of β while maintaining a high value of $R^2$, because the higher value of β leads to a sparser network with lower mean connectivity. So, the threshold would ensure the lower value of the β and higher value of $R^2$.

The adjacency matrix was first transformed into a Topological Overlap Matrix (TOM) based similarity matrix for module detection. Then it was converted into a TOM-based dissimilarity matrix by subtracting from one. This matrix was used as a distance metric to perform the average linkage hierarchical clustering algorithm, which outputs a dendrogram. Then Dynamic Tree Cut [45] method was performed for branch cutting to generate network modules. The minimum size of the modules was set at four genes.

### C. *Survival Analysis*

The workflow of survival analysis is shown in Figure 2. Module genes were used in survival analysis to check their validity as biomarkers.

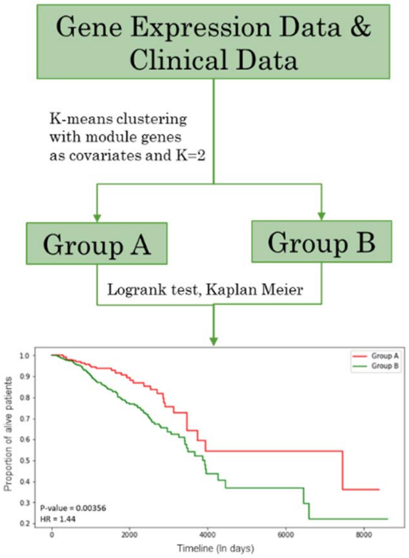

**Fig 2**. Flowchart of survival analysis. The patients are divided into two groups using K-means clustering and module genes as features. Then Logrank Test and Kaplan Meier test were performed.

For genes in the modules were taken as covariates, and the K-means clustering algorithm was used on cancer samples to create two distinguishing groups, meaning high-risk and low-risk groups. After dividing into two groups, the Log-rank test and Kaplan-Meier test, widely used in survival analysis, were applied to determine whether the modules found could differentiate between high-risk and low-risk groups. In our Survival Analysis method, we chose K-means clustering.

With K-means, we can take the whole sample into account to have the low-risk and high-risk groups. However, only after doing the Kaplan Meier test and generating the curve can we label the groups as high-risk or low-risk, not after the K-means clustering.

### D. *Module Preservation Analysis*

Modules obtained from each of the four stage-specific gene co-expression networks that passed the survival analysis test were then used for module preservation analysis to check whether modules found from one stage was preserved in other stages. NetRep [46], an R package, was used to find the preservation property in different networks. It uses a permutation approach, takes gene expression matrix, correlation matrix, and adjacency matrix of first (discovery) and second(test) network and the module genes as input. It calculates seven preservation statistics to see whether modules found from the first network (discovery network) is preserved in the second network (test network). It generates empirical null distributions for each of these test statistics and calculate their corresponding P-values. For the experiment, the number of permutations was set to 1000. The P-value of the seven statistics less than 0.001 was used as the criteria to be preserved in the test network. The definitions of these seven statistics, as explained in [46], are given in Table II.

TABLE II. MODULE PRESERVATION STATISTICS USED IN THIS ANALYSIS.

| **MPA Statistics** | **Equation** |
|---|---|
| Module Coherence | $mean((cor(g_i^{[t](w)}, Eig_1^{[t](w)}))^2)$ |
| Average Node Contribution | $mean(sign\left(cor\left(g_i^{d(w)}, Eig_1^{[d](w)}\right)\right) \cdot cor\ (g_i^{[t](w)}, Eig_1^{[t](w)}))$ |
| Concordance of node contributions | $cor\ (cor\left(g_i^{d(w)}, Eig_1^{[d](w)}\right), cor(g_i^{[t](w)}, Eig_1^{[t](w)}))$ |
| Density of correlation structure | $mean(sign\left(C^{[d](w)}\right) \cdot C^{t(w)}))$ |
| Concordance of correlation structure | $cor_{i \neq j}(C^{[d](w)}, C^{[t](w)})$ |
| Average edge weight | $mean_{i \neq j}(g_{ij}^{[t](w)})$ |
| Concordance of weighted degree | $cor((\sum_{i \neq j}^{j} a_i)^{[d](w)}, (\sum_{i \neq j}^{j} a_i)^{[t](w)})$ |

The mathematical symbols are as follows: $G$ is $m \times n$ gene expression profiles over $n$ genes and $m$ samples, $C$ is $n \times n$ pairwise correlation matrix, and $A$ is $n \times n$ adjacency matrix. Lowercase $g, c, a$ refer to individual elements of matrices denoted by their corresponding uppercase letters G, C and A (gene expression matrix, correlation matrix and adjacency mtarix). Superscripts $[d]$ and $[t]$ refer to discovery and test networks. The subscripts $i, j$ denote individual variables/nodes in module $w$. $Eig_1^{(w)}$ refers to the 1st principal component of module $w$, otherwise known as the module eigengene or summary profile. The $sign$ evaluates to 1 if its argument is positive and -1 if its negative value.

### E. GO and Pathway Enrichment Analysis

GSEApy (Gene Set Enrichment Analysis in Python) was used for pathway and GO enrichment analysis of biomarker genes. GSEApy is a python wrapper for Enrichr [47] and GSEA (Gene Set Enrichment Analysis) [48]. The analysis was implemented in python 3.7.3 using gseapy package 0.9.5. The adjusted p-value < 0.05 and Benjamini & Hochberg correction for multiple testing was used as a statistical measure. The pathway analysis was performed using Enrichr 'KEGG_2019_Human' library. The enrichr libraries used for GO enrichment are –

- 'GO_Biological_Process_2018',
- 'GO_Cellular_Component_2018' and,
- 'GO_Molecular_Function_2018'.

## III. Results and Discussion

### A. Results of Network Construction and Module Detection

For network construction, different values of $\beta$ were chosen for samples of each of the four stages and normal samples to meet the scale-free criteria, which are given in Table III.

TABLE III. Soft threshold exponent $\beta$ for each group of patients, from stage I to Stage IV and normal samples.

| Cancer Stage/ Normal | Soft Thresholding Exponent $\beta$ |
|---|---|
| Stage I | 7 |
| Stage II | 9 |
| Stage III | 13 |
| Stage IV | 12 |
| Normal | 15 |

For each group of samples, the value of beta was varied between 1 to 30 and their corresponding R^2 value was calculated. The first beta to cross the threshold, which is R^2 >= 0.90 is considered for construction of co-expression network. This process for stage-I is shown in Figure 3.

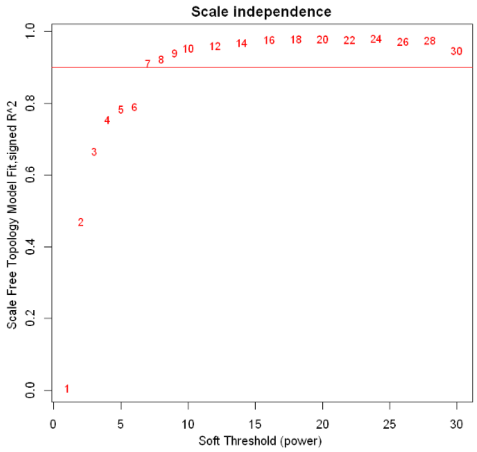

**Fig 3.** Soft Thresholding Exponent β versus Model Fitting Index $R^2$ for Stage-I.

Figure 3 shows how the value of beta was chosen to construct the co-expression network of stage-I samples. Although the $\beta$ value of 24 would give the highest value of $R^2$, meaning closer to be a scale-free network, it would result in a network with very low edge weights. So, 7 is chosen here as it is the first one to cross the 0.90 threshold.

Adjacency matrix of co-expression network were then converted to TOM-based dissimilarity matrix and hierarchical clustering was used to find the modules. The summary of modules, in terms of number of modules, the size of the smallest and largest module for each stage is given in Table IV.

TABLE IV. Summary of the Modules in four stages.

| Stages | Total # of modules | Size of the largest modules | size of the smallest modules |
|---|---|---|---|
| I | 293 | 1325 | 4 |
| II | 267 | 1229 | |
| III | 150 | 940 | |
| IV | 733 | 1305 | |

From Table IV, we can see that the lowest number of modules was found from the co-expression network of stage-III, and the highest number of modules was found from the network of stage-IV patients. It might correlate with the number of patients and the value of β that was chosen. Stage-IV has the lowest number of patients which may be the reason of highest number of modules. The smallest module size is set as 4 in the parameters of module detection.

### B. Results of Survival Analysis

We used all the modules coming from each stage to perform survival analysis. We used K-means clustering to divide all the patients using the module genes as covariates for all modules. The Log-rank test and Kaplan Meier Test were performed to check the significant difference between the two groups in terms of overall survival. Those having a Log-rank test P-value below 0.05 and the Hazard ratio above or below 1.00 were the criteria used to measure the significance. Table IV contains the number of all the modules and the number of modules which passed the survival analysis test, thereby, considered as prognostically significant modules.

TABLE V. Summary of Survival Analysis of Modules with Significant prognostic capability (P-value <= 0.05 and HR > 1.0 or HR < 1.0) for each stage.

| Stages | # modules | # prognostically significant modules |
|---|---|---|
| I | 293 | 26 |
| II | 267 | 28 |
| III | 150 | 20 |
| IV | 733 | 97 |

Stage-IV has the highest number of prognostically significant modules. As there are more modules coming from stage-IV, the number of prognostically significant modules are also much higher than the other stages. Modules are indexed from #1 to #293 for Stage-I modules, and the same is done for other stage-specific modules. For example, module #103 from Stage-I could differentiate high-risk patients and low-risk patients quite well with P-value 0.00356 and hazard ratio 1.44, as shown in Figure 4. From the Kaplan Meier curve, it can be seen that as time progresses, the proportion of alive people in each group decreases at different rates and group B has higher rate of death at a particular time point. So, group B is the high-risk group and group A is the low-risk group.

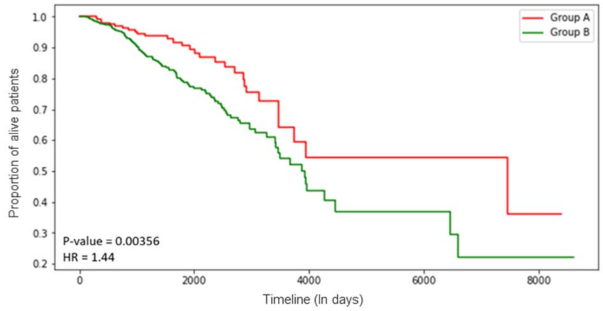

**Fig 4.** Kaplan Meier Curve of Module #103 from Stage-I. The X-axis represents the timeline in days, and the Y-axis represents the proportion of alive samples.

## C. *Results of Module Preservation Analysis*

We performed module preservation analysis (MPA), using NetRep [46], on all the prognostically significant modules on all the other networks to find which of them are preserved in co-expression networks of other stages. Table VI contains the summary of the MPA analysis for each stage. The table lists number of modules based on four conditions – (i) The modules which are preserved in other cancer stages but not in normal samples, (ii) the modules which are not preserved in any other stages, (iii) the modules which are preserved in other cancer stages but not in normal samples and prognostically significant (potential trajectory-specific biomarkers) and (iv) the modules which are not preserved in any other stages and prognostically significant (stage-specific biomarkers).

Of 293 stage-I modules, 74 happened to be preserved in co-expression network of stage-II, III, and IV, but not in normal stage. Of 74 modules, 7 are prognostically significant. Additionally, of 293 modules, 30 are not preserved in any other stages and normal samples and from them only 1 is prognostically significant.

TABLE VI. Summary of Module Preservation Analysis (MPA) of prognostically significant Modules from each stage. first row contains the number of modules that are prserved in other stages but not in normal stage. second row contains the number of modules that are not preserved in any other stages. 3rd and 4th row contains trajcetory and stage specific biomarkers.

| Conditions | Stage I | Stage II | Stage III | Stage IV |
|---|---|---|---|---|
| (i) # of modules preserved in other stages (excluding normal) | 74 | 60 | 25 | 67 |
| (ii) # of modules not preserved in any stages | 30 | 27 | 32 | 560 |
| (iii) # of potential trajectory-specific biomarkers | 7 | 6 | 3 | 9 |
| (iv) # of potential stage-specific biomarkers | 1 | 3 | 2 | 77 |

So, out of 293 stage-I modules, seven modules are preserved in four cancer stages but not in normal samples and can be considered as trajectory-specific biomarkers and one module is not preserved in any other stages, which can be considered as stage-specific biomarkers. Similarly, from stage II, III, and IV we found 6, 3, and 9 trajectory-specific biomarkers and 3, 2, and 77 stage-specific biomarkers, respectively. From stage IV, we found 77 stage-specific biomarkers because the number of modules found from stage IV co-expression network was much higher than the other stages.

## D. *GO Term and Pathway Enrichment Analysis*

Gene Ontology Term and Pathway Enrichment Analyses were conducted for 25 trajectory-specific and 83 stage-specific biomarkers found in this study. Table VII shows the enriched GO terms and KEGG pathways for different modules. It is noticeable that only 14 modules out of 108 (25 + 83) are enriched with some GO terms and KEGG pathways as shown in table VII. The module index, associated cancer stage, type of biomarker (trajectory- or stage-specific) and the enriched GO terms and KEGG pathways are enlisted in Table VII.

# IV. Conclusion

Cancer genes exhibit changes over time. This work presents a computational pipeline that works as a way of finding biomarkers that persist through these changes by doing a stage-specific co-expression network analysis of breast cancer, which are labeled as trajectory specific biomarkers. Also, network biomarkers that are only associated with one stage and not with any other stages are considered as stage-specific biomarkers, which are also found using this pipeline. A total of 25 trajectory-specific and 83 stage-specific biomarkers were found using this pipeline.

TABLE VII. SUMMARY OF GO TERM AN KEGG PATHWAY ENRICHMENT ANALYSIS OF BIOMARKER MODULES.

| Stage | Module # | Biomarker Types | GO Term & KEGG Pathway |
|---|---|---|---|
| I | 13 | Trajectory | extracellular matrix organization (GO:BP:0030198), endoplasmic reticulum lumen (GO:CC:0005788), metalloendopeptidase activity (GO:MF:0004222), Protein digestion and absorption (kegg:04974 ) |
| I | 151 | Trajectory | DNA Binding (GO:MF:0003677) |
| I | 261 | Trajectory | ribonucleoside monophosphate biosynthetic process (GO:BP:0009156) |
| II | 2 | Trajectory | cytokine-mediated signaling pathway (GO:BP:0019221), cytokine receptor activity (GO:0004896),Hematopoietic cell lineage (kegg:04640) |
| II | 32 | Trajectory | mitochondrial inner membrane (GO:CC:0005743) |
| II | 103 | Trajectory | regulation of cellular macromolecule biosynthetic process (GO:BP:2000112) |
| II | 164 | Trajectory | regulation of type I interferon production (GO:BP:0032479), RIG-I-like receptor signaling pathway(kegg:04622) |
| II | 208 | Trajectory | large ribosomal subunit (GO:CC:0015934) |
| III | 3 | Trajectory | cytokine-mediated signaling pathway (GO:BP:0019221). Cytokine-cytokine receptor interaction (kegg:04060) |
| III | 117 | Stage | Ribosome (GO:CC:0005840) |
| III | 134 | Stage | antigen processing and presentation of peptide antigen via MHC class II (GO:BP0002495), MHC class II protein complex (GO:CC:0042613), MHC class II receptor activity (GO:MF:0032395) |
| IV | 3 | Trajectory | T cell activation (GO:BP:0042110), T cell receptor complex (GO:CC:0042101) |
| IV | 7 | Trajectory | ameboidal-type cell migration (GO:BP:0001667) |
| IV | 69 | Trajectory | Hematopoietic cell lineage (kegg:04640) |

The immediate extension of this work would be the verification of trajectory-specific modules using pseudotime-based approach. In this work, only the gene co-expression network was used. Future work could use other types of biological networks such as protein-protein networks, gene regulatory networks, etc. An aggregation of these different networks would yield a heterogeneous network, and dense modules found from such networks might have biological significance. Another possible future work could be graph-based machine learning techniques to find biomarker modules. This work used gene expression data only. An integration of multi-omics data to construct the network might lead to more insightful findings.

## ACKNOWLEDGMENT

This research was partially funded by NSF CAREER award #1651917 (transferred to #1901628) and NSF RAPID award #2037374.

## REFERENCES

[1] A. M. Mondal, C. A. Schultz, M. Sheppard, J. Carson, R. B. Tanvir, and T. Aqila, "Graph Theoretic Concepts as the Building Blocks for Disease Initiation and Progression at Protein Network Level: Identification and Challenges," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2713–2719.

[2] R. B. Tanvir, T. Aqila, M. Maharjan, A. Al Mamun, and A. M. Mondal, "Graph Theoretic and Pearson Correlation-Based Discovery of Network Biomarkers for Cancer," *Data*, vol. 4, no. 2, p. 81, Jun. 2019.

[3] R. B. Tanvir, M. Maharjan, and A. M. Mondal, "Community Based Cancer Biomarker Identification from Gene Co-expression Network," in *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB'19)*, 2019, pp. 545–545.

[4] A. Palaniappan, K. Ramar, and S. Ramalingam, "Computational Identification of Novel Stage-Specific Biomarkers in Colorectal Cancer Progression," *PLoS One*, vol. 11, no. 5, pp. 1–21, 2016.

[5] L.-H. Yu, Q.-W. Huang, and X.-H. Zhou, "Identification of Cancer Hallmarks Based on the Gene Co-expression Networks of Seven Cancers," *Front. Genet.*, vol. 10, p. 99, 2019.

[6] J. Tang *et al.*, "Overexpression of ASPM, CDC20, and TTK Confer a Poorer Prognosis in Breast Cancer Identified by Gene Co-expression Network Analysis," in *Front. Oncol.*, 2019.

[7] A. Sarathi and A. Palaniappan, "Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma," *BMC Cancer*, vol. 19, no. 1, p. 663, 2019.

[8] M. Maharjan, R. B. Tanvir, K. Chowdhury, and A. M. Mondal, "Determination of Biomarkers for Diagnosis of Lung Cancer Using Cytoscape-based GO and Pathway Analysis," *20th Int. Conf. Bioinforma. Comput. Biol.*, Jul. 2019.

[9] R. B. Tanvir and A. M. Mondal, "Cancer Biomarker Discovery from Gene Co-expression Networks Using Community Detection Methods," in *2019 IEEE International Conference on Bioinformatics & Biomedicine (IEEE BIBM )*, 2019, pp. 2097–2104.

[10] T. Aqila, A. Al Mamun, and A. M. Mondal, "Pseudotime Based Discovery of Breast Cancer Heterogeneity," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2019)*, 2019.

[11] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, Sep. 2004.

[12] A. Al Mamun and A. M. Mondal, "Long Non-coding RNA Based Cancer Classification using Deep Neural Networks," in *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB'19)*, 2019, pp. 541–541.

[13] A. Al Mamun and A. M. Mondal, "Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)*, 2019, pp. 2825–2831.

[14] X.-J. Ma *et al.*, "Gene expression profiles of human breast cancer progression," *Proc. Natl. Acad. Sci.*, vol. 100, no. 10, pp. 5974–5979, 2003.

[15] H. S. Lee, H. K. Lee, H. S. Kim, H.-K. Yang, and W. H. Kim, "Tumour suppressor gene expression correlates with gastric cancer prognosis," *J. Pathol.*, vol. 200, no. 1, pp. 39–46, 2003.

[16] P. H. Johnson *et al.*, "Multiplex Gene Expression Analysis for High-Throughput Drug Discovery: Screening and Analysis of Compounds Affecting Genes Overexpressed in Cancer Cells 1 Supplementary material for this article is available at Molecular Cancer Therapeutics Online (http:{," *Mol. Cancer Ther.*, vol. 1, no. 14, pp. 1293–1304, 2002.

[17] A. M. Mondal and J. Hu, "NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks," in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 142–148.

[18] A. Mondal and J. Hu, “Network based prediction of protein localisation using diffusion kernel.,” *Int. J. Data Min. Bioinform.*, vol. 9, no. 4, pp. 386–400, 2014.

[19] A. M. Mondal and J. Hu, “Protein Localization by Integrating Multiple Protein Correlation Networks,” in *The 2012 International Conference on Bioinformatics & Computational Biology*, 2012, p. 7.

[20] B. Zhang and S. Horvath, “A general framework for weighted gene co-expression network analysis,” *Stat. Appl. Genet. Mol. Biol.*, 2005.

[21] C. J. Wolfe, I. S. Kohane, and A. J. Butte, “Systematic survey reveals general applicability of ‘guilt-by-association’ within gene coexpression networks,” *BMC Bioinformatics*, vol. 6, no. 1, p. 227, 2005.

[22] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, “A gene-coexpression network for global discovery of conserved genetic modules,” *Science (80-. ).*, 2003.

[23] X. Zhou, M. C. J. Kao, and W. H. Wong, “Transitive functional annotation by shortest-path analysis of gene expression data,” *Proc. Natl. Acad. Sci. U. S. A.*, 2002.

[24] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 25, pp. 14863–14868, 1998.

[25] P. Langfelder and S. Horvath, “WGCNA: An R package for weighted correlation network analysis,” *BMC Bioinformatics*, 2008.

[26] M. Kotlyar, S. Fuhrman, A. Ableson, and R. Somogyi, “Spearman Correlation Identifies Statistically Significant Gene Expression Clusters in Spinal Cord Development and Injury,” *Neurochem. Res.*, vol. 27, no. 10, pp. 1133–1140, 2002.

[27] P. O. Van Trappen *et al.*, “A model for co-expression pattern analysis of genes implicated in angiogenesis and tumour cell invasion in cervical cancer,” *Br. J. Cancer*, vol. 87, no. 5, pp. 537–544, 2002.

[28] C.-H. Zheng, L. Yuan, W. Sha, and Z.-L. Sun, “Gene differential coexpression analysis based on biweight correlation and maximum clique.,” *BMC Bioinformatics*, vol. 15 Suppl 1, no. Suppl 15, p. S3, 2014.

[29] R. Xulvi-Brunet and H. Li, “Co-expression networks: graph properties and topological comparisons.,” *Bioinformatics*, vol. 26, no. 2, pp. 205–214, Jan. 2010.

[30] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, “From molecular to modular cell biology,” *Nature*, vol. 402, no. 6761, pp. C47–C52, 1999.

[31] A. Maind and S. Raut, “Identifying condition specific key genes from basal-like breast cancer gene expression data.,” *Comput. Biol. Chem.*, vol. 78, pp. 367–374, Feb. 2019.

[32] H. Shi, L. Zhang, Y. Qu, L. Hou, L. Wang, and M. Zheng, “Prognostic genes of breast cancer revealed by gene co-expression network analysis,” in *Oncology letters*, 2017.

[33] J. Zhang, Q. Lan, and J. Lin, “Identification of key gene modules for human osteosarcoma by co-expression analysis.,” *World J. Surg. Oncol.*, vol. 16, no. 1, p. 89, May 2018.

[34] S. Deng, L. Zhu, and D. Huang, “Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, pp. 27–35, 2016.

[35] 丛 CongZhang张 and 茜 QianSun孙, “Weighted gene co-expression network analysis of gene modules for the prognosis of esophageal cancer,” *J. Huazhong Univ. Sci. Technol. [Medical Sci.*, vol. 37, pp. 319–325, 2017.

[36] Q. Li, D. Ghersi, I. Thapa, L. Zhang, H. Ali, and K. Cooper, “Identifying Structural Changes in Correlation Networks Models of Cancer Gene Expression by Stage,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 2075–2082.

[37] T. Huo, R. Canepa, A. Sura, F. Modave, and Y. Gong, “Colorectal cancer stages transcriptome analysis,” *PLoS One*, vol. 12, no. 11, pp. 1–11, 2017.

[38] N. Misra, E. Szczurek, and M. Vingron, “Inferring the paths of somatic evolution in cancer,” *Bioinformatics*, 2014.

[39] S. Pang *et al.*, “Reconstruction of kidney renal clear cell carcinoma evolution across pathological stages,” *Sci. Rep.*, 2018.

[40] A. Palaniappan, K. Ramar, and S. Ramalingam, “Computational identification of novel stage-specific biomarkers in colorectal cancer progression,” *PLoS One*, 2016.

[41] M. P. Pradhan, A. Desai, and M. J. Palakal, “Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma,” *BMC Syst. Biol.*, 2013.

[42] H. Lalremmawia and B. K. Tiwary, “Identification of Molecular Biomarkers for Ovarian Cancer using Computational Approaches.,” *Carcinogenesis*, 2019.

[43] M. Goldman *et al.*, “The UCSC Xena platform for public and private cancer genomics data visualization and interpretation,” *bioRxiv*, 2018.

[44] P. L. and S. Horvath, “WGCNA package FAQ.” [Online]. Available: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html.

[45] P. Langfelder, B. Zhang, and S. Horvath, “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R,” *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2007.

[46] S. C. Ritchie, S. Watts, L. G. Fearnley, K. E. Holt, G. Abraham, and M. Inouye, “A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets.,” *Cell Syst.*, vol. 3, no. 1, pp. 71–82, Jul. 2016.

[47] M. V Kuleshov *et al.*, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W90–W97, Jul. 2016.

[48] A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.