Chapter 5

Differential Co-expression Analysis

5.1 Introduction

Within biological systems, an approach to unravel the relationships among genes is network analysis. Construction and analysis of networks that represent associations or interactions between genes is an integral part of network analysis. By identifying key genes in a network that play critical roles in biological process, disease or condition, network analysis can facilitate biomarker identification. Co-expression network (CEN) is an approach to network analysis that investigates trends and connections in gene expression across conditions or samples. Genes associated with a CEN exhibit similar patterns and thus imply shared regulatory mechanisms as well as functional connection. Weighted Gene Co-expression Network Analysis (WGCNA) [327] is a widely used method that constructs a weighted CEN through estimation of similarity score based on the expression profiles of the genes. The primary goal of differential co-expression analysis (DCA) is to uncover the condition specific changes in co-expression patterns. In other works, through DCA we compare the changes in relationships and interactions among genes under varying conditions. While DEA identifies individual genes that are differentially expressed between conditions, DCA takes into consideration the interactions among genes under varying conditions.

5.1.1 Differential Co-expression Analysis (DCA)

By detecting variations in co-expression patterns among genes in different populations or conditions, DCA aids the discovery of crucial modules or genes associated with specific biological functions or conditions. Following are the key steps for DCA.

1. Pre-processing is the first step towards DCA and is specific to the input gene expression data. This step generally involves, removal of noise, normalization, and missing value estimation.

2. Various similarity approaches such as Pearson correlation[545], Spearman correlation [643] or mutual information [107] are employed to find pair-wise gene-gene similarity. Based on these pair-wise gene-gene similarity, CENs are constructed.
3. DCA is conducted to find differences in gene co-expression patterns between CENs separated based on samples (e.g., control vs. disease).
4. Statistical tests such as t-tests [761, 615], ANOVA [168, 167] are employed for comparison of the co-expression correlations and evaluation of the significance of variations between groups.
5. Modules or gene-pairs that exhibit significantly different co-expression patterns under varying conditions are chosen for further biological validation.

As genes tend to interact in intricate networks as opposed to functioning in isolation, DCA helps uncover changes in these networks. This helps unravel more nuanced underlying mechanisms and interactions that might be overshadowed if focus is solely on changes in expression of genes individually. Furthermore, DCA helps highlight genes that act as hubs or key regulators in the network and control interaction among genes. DCA, however, entails sufficient sample sizes for accuracy as smaller sample sizes may lead higher FDR. In large datasets, construction of CEN for DCA can be computationally intensive. The accuracy of correlation calculation in DCA can be impacted by missing gene expression values. Interpretation of biological significance of large modules with complex interactions in the CENs can be challenging.

Our work on DCA examines pair-wise gene expression changes in disease tissue vs normal tissue with the goal of identifying important genes for serious diseases like Esophageal Squamous Cell Carcinoma (ESCC). Due to its capacity to detect changes in regulatory connections between genes that would not be picked up by conventional CEN or Differential Expression (DE) research, DCA can aid in the identification of biologically significant DCA gene modules. Various steps are included in a DCA. On the basis of pair-wise gene-gene similarity, a CEN is initially built. The most often utilized tool for CEN is the Weighted Gene Co-expression Network Analysis (WGCNA) [327]. Several techniques have been introduced to extract relevant modules, combining a) Clustering techniques such as those proposed by Kisilevi et al.[309], Chen et al. [89], Langfelder et al. [330] and Fukushima et. al[174], b) Guilt by association (GBA) approaches such as those proposed by Oliver et al. [533], Gillis et al. [189] and Wolfe [765], c) Hub-gene finding approaches such as those presented in Albert et al. [15], Oh et al. [529], Keller et al. [304], Voineagu et al. [713], Das et al.[120] and Azuaje et al. [35], d) Enrichment analyses such as those proposed by Huang et al. [253], Glaab et al. [190], DAVID ([628], [253]) and Creixell [108], and e) Regulatory network identification presented in Linde et al. [407], Margolin et al. [487], Wang et al.[752] and Irrthum et al. [268]. These modules are then subjected to downstream analysis, which identifies potential biomarker(s). One of the most important processes in identifying a biomarker is hub-gene finding. Existing techniques for locating biomarkers include using p-value cut-off [120], weighted gene score [120], and intra-modular connectivity [327]. As far as we are aware, no study has employed centrality metrics to pinpoint hub-genes. Our research primarily focuses on the discovery of hub genes, and we suggest the Centrality Based Differential Co-Expression Method (CBDCEM), which is a method based on the centrality metric.

5.2 Related Works

Many methods and tools have been created to analyze gene expression data and identify differentially co-expressed genes (DCGs). These methods can be divided into two groups: a) supervised and b) unsupervised. When DCA methods are influenced by the prior knowledge or details about the conditions under which the comparisons are carried out, those methods are supervised. Supervised DCA methods detect different co-expression patterns based on pre-specified list of categories. Unsupervised DCA methods, on the other hand, seek to identify natural or intrinsic co-expression patterns in the data and do not depend on specified conditions or groups. Unsupervised methods tend to explore co-expresssion patterns that are otherwise overlooked by methods that rely on prior knowledge. Unsupervised approaches includes Weighted Gene Co-expression Network Analysis (WGCNA) [327], Differential Co-expression Graph Learning (DCGL)[410, 807], Co-expression Explorer (co-Xpress)[759], and Differential Co-Expression Analysis (DiffCoEx) [679], whereas Cognition and Genetics of Aging (CoGA)[604], Gene Set Co-expression Analysis (GSCA)[101], Gene Sets Net Correlations Analysis (GSNCA)[567], and Differential Co-expression Analysis for REmodeling (DICER)[22] falls under supervised approaches.

5.2.1 Unupervised DCA approaches

A well-known and often-used method for finding differentially co-expressed modules is Weighted Gene Co-expression Network Analysis (WGCNA) [327]. WGCNA creates CENs utilizing a gentle thresholding method and a well-defined dissimilarity measure. In order to ensure that gene expression levels are comparable across samples, WGCNA starts by normalizing gene expression data. Typically, this is done using techniques like log transformation or quantile normalization. By computing pairwise correlations between genes and translating those correlations into a weighted adjacency matrix using a power function to emphasize strong correlations and down-weight weak ones, WGCNA creates a weighted network. This is done to create a scale-free network with a power-law distribution of node connectivity. Genes with comparable expression patterns are organized into modules by WGCNA using hierarchical clustering. Based on the topological overlap measure, which shows the shared connection of genes in the network, clustering is done. Using metrics like module eigengene-based connectivity [327] and $Z_{summary}$ scores (Section 2.1.10), WGCNA assesses the stability and preservation of modules across various datasets or conditions. Using module eigengenes (MEs), the initial principal elements of gene expression inside a module, WGCNA assesses the association between modules and clinical features, environmental factors, or experimental settings. WGCNA determines hub genes, which are highly connected genes that are biologically related to the trait of interest, by calculating the correlation between gene expression levels and clinical traits or other relevant parameters.

Differential co-expression networks between various experimental circumstances can be found using the Differential Co-expression Graph Learning (DCGL)[410, 807] approach. To ensure that gene expression levels are consistent across samples, DCGL first normalizes gene expression data, generally using quantile normalization or log-transformation techniques. DGCL identifies distinct co-expression patterns between two groups of samples by calculating the differential co-expression score (DCES) for each gene pair. DCGL constructs a graph that represents the differential co-expression network by creating edges between genes with substantial DCES values. The co-expression connection along the edges represent the genes that differ significantly between two groups. A set of genes that are most useful in differentiation between the two groups are chosen by ranking the genes based on DCES values. DGCL employs graph embedding to create low dimensional representations of the differential CEN. Underlying structure of the relationships between genes with differential co-expression are captured by a vectors that represent the embedding.

DiffCoEx [679] approaches DCG identification by providing two types of DCA, namely, a) intra-module DCE and b) inter-module DCE. To ensure that gene expression levels are consistent across samples, DiffCoEx starts by normalizing gene expression data, generally using techniques like log-transformation or quantile normalization. In order to create a co-expression network, DiffCoEx computes the pairwise correlation coefficients between all the genes in all samples. By comparing the correlation coefficients of each gene pair between the two groups, DiffCoEx finds gene pairs that are differentially co-expressed in two groups of samples. Utilizing statistics like the t-statistic or fold change, it is possible to quantify the variance in correlation coefficients. By regulating the false discovery rate (FDR) (Section 2.1.2) using strategies like the Benjamini-Hochberg [43, 764] method, DiffCoEx accounts for multiple hypothesis testing. DiffCoEx analyses the differentially co-expressed gene pairs and uses clustering algorithms like hierarchical clustering [756, 291] or k-means clustering [444] to identify co-expressed gene modules. Using correlation or regression analysis, DiffCoEx measures the correlation between co-expression modules and outside factors like clinical characteristics or experimental circumstances.

Co-expression Explorer (co-Xpress)[759] employs clustering methods such as hierarchical clustering [756, 291] or k-means clustering [444] to identify co-expressed gene modules. Gene pair-wise correlations are the basis of highly correlated modules. To assess connections between co-expression modules as well as external variables such as experimental circumstances or clinical characteristics, Co-Xpress uses regression analysis. Instead of module eigengene, Co-Xpress utilizes module expression profiles. Functional enrichment analysis (Section 2.4.1) and gene set enrichment analysis [650] are employed to analyze biological relevance of the co-expression modules. It is also essential to choose a set of genes that are most useful in predicting external variables and as such Co-Xpress chooses the genes that exhibit strongest correlations with the outside variables. Co-Xpress divides the data into training and testing sets to assess the predictive performance of the chosen genes and evaluate the prediction precision.

5.2.2 Supervised DCA Approaches

A probability score is used by Differential Co-expression Analysis for Remodelling (DICER) [22] to identify DCE gene sets, and a probability-based framework is also used for significance assessment. The goal of DICER is to identify changes in the co-expression connections between genes, which can shed light on functional alterations in biological systems. DICER identifies diverse co-expression patterns between various conditions or groups by comparing the CENs and observing changes in the strength of gene-gene connections. Differential network analysis and differential module analysis are used to find significant changes in co-expression patterns.

A computational technique called Gene Set Co-expression Analysis (GSCA) [101] is used to find coordinated patterns of gene expression within predetermined gene sets or gene modules. A gene set in GSCA is a predetermined collection of genes that have been assigned to the same biological region, function, or regulatory mechanism. Pathways, gene ontologies, or gene modules discovered using different clustering or co-expression research techniques are a few examples of gene sets. From the samples of interest, gene expression data, such as microarray or RNA sequencing data, is gathered. The expression levels of the genes across the samples are quantified by this data. Based on prior information or annotations, predefined gene sets or modules are chosen or created that are associated to particular biological functions, pathways, or gene ontologies. These gene sets can be found in databases like Gene Ontology or KEGG, the Kyoto Encyclopedia of Genes and Genomes.Pairwise correlations or other measures of relationship between the gene expression profiles across the samples are computed to form a co-expression network. The co-expression interactions between genes are captured by this network in terms of their intensity and direction. The goal of GSCA is to locate gene sets inside the co-expression network that have notable co-expression patterns. If the expression patterns within a gene set are more correlated than would be predicted by chance, it can be determined using statistical techniques like enrichment analysis or permutation tests. The outcomes of GSCA are frequently represented as networks or heatmaps, where gene sets with noteworthy co-expression patterns are emphasized. Finding functional modules or pathways that exhibit coordinated expression changes in this way enables researchers to gain understanding of the biological mechanisms or regulatory processes behind the phenotype or condition being researched.

In Gene Sets Net Correlations Analysis (GSNCA) [567] relationships and co-expression patterns between gene sets or pathways are assessed and their strength are determined. Firstly, GSNCA identifies relevant pathways and gene sets. Either prior biological knowledge or GSEA [650] is employed to generate gene sets. After the specification of the gene sets, within each gene set, pairwise correlations between the genes are determined. This is implemented to evaluate the directionality and the intensity of the links as well as to measure the comparability of the expression profiles across samples. Statistical testings are used to evaluate the importance of the observed connections within a gene set and the strength of these connections over randomly predicted connections.

5.3 Basics Of Centrality Measures

In this section, we discuss the seven measures used by our proposed hub-gene finding algorithm employed in our DCA framework, CBDCEM: An effective Centrality Based Differential Co-Expression Method for crucial Gene Finding. According to network theory, a node's prominence or importance inside a network is referred to as its centrality. Based on a node's structural location in the network and its connections to other nodes, it measures the relative influence or relevance of that node. Node centrality measurements enable the detection of functional roles of genes in biological networks. The network's unique properties and the current research question influence the choice of centrality measure. Different centrality measures can provide light on a variety of network phenomena, including information flow, impact, and control within a network, by capturing different facets of node importance. According to Azuaje et al., highly linked genes (also known as hub-genes) in gene CENs frequently associate with important disease pathways. We use WGCNA [327] to create a CEN and extract important modules for a specific dataset. We try to find significant nodes (or genes) in the collected modules that might serve as biomarkers.

A graph's centrality can be a reliable indicator of key nodes. To represent centrality indices, real valued functions on vertices are utilized, and as a result, these values can provide a ranking that makes it easier to identify the most significant nodes on the graph. Classification is possible depending on how specific centralities assess cohesion. The walk structure is emphasized by categorizing centralities to emphasize cohesiveness. These centralities fall into two subcategories: radial and medial, which can be deduced from the way they are built. Radial centrality is a type of centrality that considers treks that originate at or conclude at a certain vertex. Radial centrality is exemplified by *eigenvector centrality* [519] and *degree centrality* [171]. In contrast, a vertex's median centrality is calculated by counting the walks that go through it. One example of this subcategory is *betweeness centrality* [170]. It is possible to group centralities that measure the quantity or duration of walks. *Closeness centrality* [39] serves as the best illustration of this group.

Tab. 5.1: Centrality Measures[621] for hub-gene findin employed in CBDCEM

Centrality Measure	Function	Formula
Betweeness Centrality [170]	The ability of a given node to track information flow between other vertices is measured by its betweeness centrality..	When the number of shortest pathways from node v_a to node v_b is $\alpha_{v_a v_b}$ and the number of those shortest paths that pass via node v_i is $alpha_{v_a v_b}(v_i)$, the betweeness centrality is given by: $C_{betC_{v_i}} = \sum_{v_a \neq v_b \neq v_i} \frac{\alpha_{v_a v_b(v_i)}}{\alpha_{v_a v_b}}$
Closeness Centrality [39]	The degree of closeness between nodes determines a node's significance in the network. Based on its ability for rapid communication with other nodes, a node is given a higher value.	If the shortest distance between nodes v_i, and v_j is $d(v_i, v_j)$, then closeness centrality is defined as follows. $C_{cC_{v_i}} = \frac{1}{\sum_{v_j} d(v_i, v_j)}$
Degree Centrality [171]	The amount of other nodes that a given node is connected to serves as a measure of its degree centrality.	Degree centrality is determined by the formula: $C_{degC}(v_i) = deg(v_i)$, where $deg(V_i)$ is the degree of node i.
Eigenvector Centrality [519]	A node is given a greater value if its connections to nearby neighbors are thought to be significant. This is done to make sure that every node's neighbors experience the same effects.	If λ is a constant such that $\lambda \neq 0$ and the entry in the u^{th} row and i^{th} column of the adjacency matrix of the network are represented as $d(v_u, v_i)$, then the eigen vector of a node $v_i \in V$ is given by. $C_{eigenC_{v_i}} = \frac{1}{\lambda} \sum_{v_u} d(v_u, v_i) E_u$.
Katz Centrality [302]	The total number of walks between any two nodes is taken into consideration when calculating a node's influence. By adding a penalizing attenuation factor, α that distinguishes between direct and indirect connections, the measure makes a distinction between the two.	If the total number of k degree connections between node i and node j is reflected by the element at location (i, j) of the adjacency matrix A raised to the power of k degrees than the Katz centrality of node i is given by: $C_{katzC_{v_i}} = \sum_{i=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$. The magnitude of the attenuation factor, α, is selected so that it is less than the reciprocal of the absolute value of the biggest eigenvalue of the A matrix.
Page Rank [652]	An adaptation of the eigenvector radiality metric that assigns a node's score based on both the node's quality and the number of linkages.	If the set of all nodes linking to node v_i is B_u and $L(v_u)$ is the number of links from node v_u, then Page Rank is determined by. $C_{pageR_{v_i}} = \sum_{v_u \in B_u} \frac{pageR(v_u)}{L(v_u)}$

Continued on next page

Centrality Measure	Function	Formula
Radiality [705, 766]	Based on a node's reachability to every other node in the network, a value is assigned to it.	Radiality is determined using the formula. $C_{radC_{v_i}} = \frac{\sum_{v_u \neq v_i} R_v D_{v_u v_i}}{n-1}$, where $R_v D_{v_u v_i}$ is the reverse distance between nodes v_u and v_i, and n is the total number of nodes.

Degree centrality [171], which is determined by the number of linkages occurring on a network node, is the most basic and perhaps oldest type of centrality measure. However, this measure focuses on each node separately rather than taking into account the network's overall structure. The average length of the shortest paths connecting a node to every other node in the graph is known as a node's normalized *closeness centrality* [39]. The node is closer to all other nodes when the average journey length is shorter. The interpretation states that a node is more central the closer it is to all other nodes. However, graphs with disconnected components cannot use this metric. A vertex's *betweeness centrality* [170] , which measures how frequently it acts as a bridge along the shortest path connecting any two other nodes, is a metric of in a graph. In other words, the vertices with higher *betweeness centrality* have higher probability of occurring on the shortest path between any two randomly chosen vertices in a graph. *Eigenvector centrality* [519] in a graph assigns relative scores to each node in order to measure each node's influence. This rating gives more weight to the idea that a node's connections to other high-scoring nodes contribute more significantly than equivalent connections to nodes with lower ratings. A variation of *eigenvector centrality* called *page rank centrality* [652] counts the quantity and quality of connections to a node in order to roughly gauge the significance of that node. While *degree centrality* [171] and *eigenvector centrality* [519] can be used to evaluate the local and global importance of a node inside the network, respectively, *katz Centrality* [302] accounts for both of these influences. *Katz Centrality* [302] counts the number of a node's immediate neighbors and the connections that node has made to other nodes through those neighbors to determine a node's influence within a network. *Katz Centrality* [302] includes an attenuation factor to penalize links to far-off neighbors (i.e., indirect connections made through close neighbors). In Table 5.1, we summarize these measures.

Tab. 5.2: Comparison of the seven centrality measures employed by CBDCEM

Centrality Measure	Pros	Cons
Betweeness Centrality [170]	• Nodes with a high betweenness centrality serve as crucial bridges or mediators between various network nodes, providing insight into the network's potential weak points structurally and as points of control or disruption. • Nodes with a high betweenness centrality have links to many parts of the network, which makes it easier to understand how cohesive and integrated the network is overall. • The efficient diffusion or flow of information is likely to occur at nodes with high betweenness centrality, making them ideal targets for interventions or resource allocation.	• A node's value, influence, or relevance based on its content, quality, or expertise are not taken into account by betweenness centrality, which only considers the network structure. • Due to the fact that betweenness centrality is predicated on the idea of shortest paths between nodes, it may ignore alternate paths resulting in inaccurate assessments. • When there are numerous connections with equal lengths or alternate paths for information flow, the calculation of betweenness centrality is sensitive to the structure of the network. • High degrees of connectedness are typically given a higher centrality rating by it.
Closeness Centrality [39]	• Because they are easily accessed by other nodes in the network, nodes with high closeness centrality can be targeted for effective resource allocation, communication, or intervention. • For the network to remain connected and reliable, nodes with high closeness centrality are essential. • High closeness centrality nodes serve as the focal points for information gathering, coordination, or decision-making in decentralized networks. • High closeness centrality nodes frequently have a big influence in the network.	• Closeness centrality requires that all nodes in the network can be reached from one other, which can lead to distorted centrality results, particularly in networks containing isolated or unreachable nodes. • It is skewed towards smaller networks since the average distance between nodes in bigger networks is greater, resulting in lower closeness centrality values overall. • Because of the small distances to other nodes, isolated nodes with no or limited connections to the rest of the network may have unnaturally high proximity centrality values.
Degree Centrality [171]	• Degree centrality is simple to perceive and comprehend because it represents the number of direct network connections a node has, indicating its immediate reach and potential influence. • In a network, nodes with a high degree of centrality are frequently crucial hubs or connectors. • It can help uncover communities and identify heavily connected locations within a network.	• Degree centrality measures only the amount of connections a node has in the network and ignores any node traits or characteristics. • It considers all connections to be equal, which may oversimplify the network structure and overlook key variations in the effect or flow of information. • It does not take into consideration indirect connections or the influence of nodes other than their immediate neighbors. • The presence of communities, cliques, or other complicated patterns is not taken into account when calculating degree centrality.

Centrality Measure	Pros	Cons
Eigenvector Centrality [519]	• Eigenvector centrality considers both the amount of connections a node has and the importance of those connections. • It aids in the identification of nodes having a high impact or influence inside the network since nodes with a high eigenvector centrality are not only well-connected but also related to other influential nodes. • It is a recursive method that allocates importance to nodes based on their neighbors' importance.	• Even if nodes have major roles or impact in the network, nodes that are not well connected to prominent nodes may have low eigenvector centrality scores, and this sensitivity to network topology can lead to incorrect judgements of node importance in some cases. • Computing eigenvector centrality for large networks can be computationally and computationally expensive. • If the network contains unconnected components, the eigenvector centrality ratings may not effectively reflect the importance or influence of nodes in those components. • It gives equal weight to all sorts of impact or importance.
Katz Centrality [302]	• Katz centrality considers not only a node's near neighbours, but also the influence of all nodes in the network, even those located further away. • When compared to other centrality measures that rely primarily on direct connections, Katz centrality provides a more comprehensive perspective of node importance and influence by including many pathways. • Users can fine-tune the importance of indirect connections by modifying the value of α, and balancing the influence of near neighbors and the overall structure of the network. • It handles isolated nodes better than other centrality methods.	• Calculating Katz centrality in large networks can be computationally and time-consuming. • Because different values of may produce different findings, it is critical to carefully choose an acceptable value for the given network and study topic. • It is calculated by scaling the adjacency matrix, which can cause numerical instability or amplification of small values. • It considers all sorts of influence or importance to be equal.
Page Rank [652]	• PageRank centrality considers the relevance of nodes that link to a certain node in order to account for the global structure of the network. • It performs better in networks with isolated nodes or unconnected components than other centrality methods. • It is intended to be resistant to attempts at manipulation, such as artificially adding or removing links to alter node ranks.	• Calculations of PageRank centrality might be affected by the initial conditions or starting values. • For very large networks, PageRank centrality computations can be computationally expensive, especially if the network structure is highly connected or dense. • It prioritizes nodes with a greater number of incoming connections and may ignore nodes with strong influence or relevance but fewer inbound links.
Radiality [766]	• Radiality centrality quantifies the relevance of nodes in the network based on their proximity to other nodes. • It is very good at capturing the influence and relevance of nodes in localized clusters or communities, providing insights into the network's localized impact. • When compared to other centrality measures, such as eigenvector centrality or betweenness centrality, calculating radiality centrality is more computationally efficient.	• Radiality does not take into account the network's general global structure or long-range connections, and it may ignore nodes that are not central within their neighborhoods but serve critical roles in connecting different areas of the network. • Because of their near closeness to many other nodes, nodes in dense networks are likely to have similar radiality scores. • Even though a node's overall influence or relevance in the network is limited, nodes that are well-positioned within their immediate neighborhoods may earn high radiality ratings.

5.4 CBDCEM: An Effective Centrality Based Differential Co-Expression Method For Crucial Gene Finding

We analyze the transcriptional changes in gene connections rather than individual genes in the proposed framework shown in Fig 5.1. In order to identify the disease-induced topological and functional alterations in the networks, we first generate gene CENs. Two separate networks that correspond to healthy and diseased states are built by the identification of co-expressed pairings across the circumstances. Identification of the biological alterations is aided by a comparison of the normal and disease CENs, or from the normal condition to the disease condition and vice versa. Our approach accepts microarray or bulk RNA-Seq data as input datasets. The gene expression dataset(s) are initially split into two subsets according to the kind of tissue: normal adjacent tissue and disease tissue. These subsets are used as input to the framework and can either be microarray or bulk RNA-Seq data.

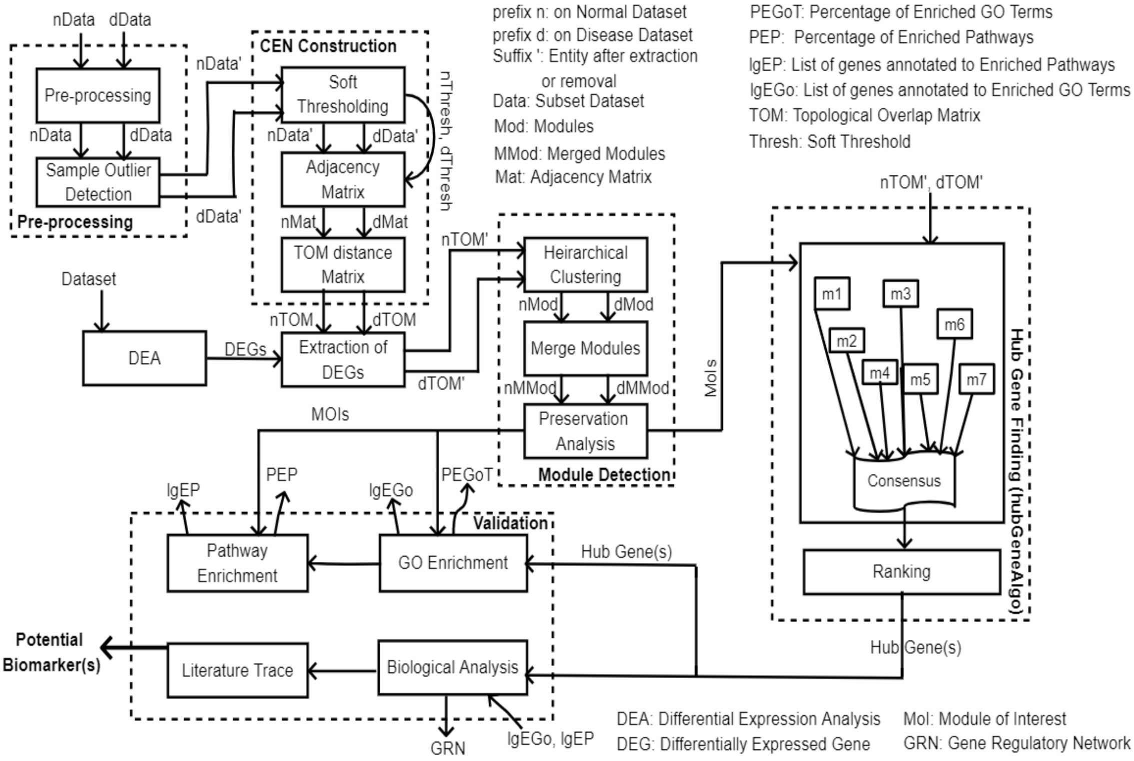


Fig. 5.1: Proposed Centrality Based DCA Framework, CBDCEM

5.4.1 Pre-processing

The pre-processing part of CBDCEM carries out all fundamental tasks like removing unnecessary and redundant data, normalizing the dataset, and estimating missing values while guaranteeing that the data meet the requirements for subsequent analysis. However, this unit also handles low read count and batch effect removal for bulk RNA-Seq data. The pre-processing pipeline for microarray and bulk RNA-Seq data employed by CBDCEM are described in detail in Section 2.7.1 and Section 2.7.2, respectively. We eliminate the outlying conditions (samples) after we are certain the facts meet all the prerequisites. We achieve this by clustering the dataset conditions (samples) and removing the outliers.

5.4.2 CEN Construction

We use weighted gene network analysis (WGCNA) [327] to continue with the construction of the CENs. In order to calculate an adjacency matrix while employing WGCNA, the soft threshold power to which co-expression similarity is raised must be chosen. On the basis of the approximate scale-free topology [38] criteria , we select the soft threshold power. We convert the adjacency matrix into a topological overlap matrix (TOM [574]), which yields a comparable dissimilarity matrix of the same size, in order to reduce the impact of noise and erroneous associations.

5.4.3 Module Extraction

We take into account the cardinality of the genes in a module (Definition 3.4.1 in Chapter 3) when separating the normal from the disease modules. Unwanted challenges in the downstream analysis may result from a significant imbalance in the module sizes. In order to produce modules with manageable sizes, we balance the amount of instances using differential expression analysis. We next extract all the connections related to the differentially expressed genes (DEGs) from the CEN that has been created. After that, we use hierarchical clustering (Section 2.1.5) to roughly extract the modules for normal and disease states. On the normal and disease datasets, the dynamic tree cut technique can be used to further extract comparable modules with similar expression profiles. Additionally, it would be a good idea to combine some modules because the genes in those modules are significantly co-expressed. To measure the similarity of co-expression across entire modules, eigengenes are computed and clustered based on their correlation. The DCA unit identifies differentially co-expressed modules and performs preservation analysis (Section 2.1.9) on these modules to identify biologically relevant modules. These modules are termed as "Modules of Interest" (MoI) (Definition 4.3.1).

This is followed by the identification of hub-genes (Definition 3.4.3 in Chapter 3) in these modules using our proposed centrality based hub-gene finding algorithm described in the next subsection.

5.4.4 Hub-gene Finding

In order to identify potential biomarkers, the significant modules recovered from both conditions are further examined using centrality measures . To find crucial genes that can be regarded as hub-genes (Definition 3.4.3 in Chapter 3), we use seven centrality measures (given in Table 5.1). Our strategy for locating hub-genes using these centrality indicators is presented in the algorithm 1. Symbols used in Algorithm 1 are given in Table 5.3

Tab. 5.3: Symbols used in proposed Hub-gene finding algorithm

Symbol used	Meaning
$modGenes$	Network connections of a module
M	Set of seven centrality measures
$S[m,g]$	Centrality score for centrality measure, m for gene, g
$L[[m]]$	List of corresponding centrality scores of genes for measure, m.
$Topk[m,g]$	Score of 1 or 0 assigned to gene, g based on its presence or absence among top k genes for measure, m.
$Score[g]$	Consensus score for gene, g
$lHub$	List of hub-genes

According to our proposed algorithm's description:

1. Accept as input the important module's matching gene co-expression network, $modGenes$.
2. Seven score lists, $l_1, l_2, l_3, l_4, l_5, l_6$, and l_7, with associated centrality measures $m_1, m_2, m_3, m_4, m_5, m_6$, and m_7 for each of the genes, are produced by estimating each centrality measure separately on each gene in $modGenes$.
3. Sort the seven gene lists $L_1, L_2, L_3, L_4, L_5, L_6$, and L_7 in decreasing order of the score values of the individual genes in the lists $l_1, l_2, l_3, l_4, l_5, l_6$, and l_7. .
4. The top k genes are given the flag value 1, designating their inclusion in the $Topk$ list for the relevant centrality measure (m) score. This creates seven $Topk$ (for i= 1 to 7) lists for the genes.
5. Calculate the consensus score for each gene in $modGenes$ by adding up the number

Algorithm 1: Proposed Hub-gene Finding employed in CBDCEM

```
Input  : modGenes: the network connections of a module, $M$: the set of seven centrality
         measures
Output: lHub: List of detected hub genes in the given network module.
// Compute the centrality measure scores of each of the genes in the
   given gene module connectivity graph for each of the seven centrality
   measures in $M$.
foreach $m \in M$ do
    foreach $g \in modGenes$ do
        $S[m,g]$ = centrality score for measure $m$ of $g$ in $modGenes$
    end foreach
end foreach
// // Create separate sorted lists of the genes in $modGenes$, one for each
   $m$ in $M$, based on the corresponding centrality measure (m) scores.
foreach $m \in M$ do
    foreach $g \in modGenes$ do
        $L[m,g].score = S[m,g]$
        $L[m,g].gene = g$
    end foreach
end foreach
foreach $m \in M$ do
    Sort $[L[m]]$ in descending order based on the score field
end foreach
// For the topk genes in in sorted $[L[m]]$ assign flag value 1, marking its
   inclusion in the Top k list for the corresponding centrality measure
   ($m_i$) score
foreach $m \in M$ do
    foreach $i = 1$ to $k$ do
        $Topk[m, L[m,i].gene] = 1$
    end foreach
end foreach
// Compute consensus rank score for each gene in $modGenes$ by counting the
   centrality measures in which it ranks in top $k$ genes
foreach $g \in modGenes$ do
    $Score[g] = 0$
    foreach $i = m$ do
        $Score[g] = Score[g] + Topk[m,g]$
    end foreach
end foreach
// Determine the hub-genes as those that have consensus score $\geq 4$
foreach $g \in modGenes$ do
    if $Score[g] \geq 4$ then
        $lHub = lHub \cup g$
    end if
end foreach
```

of times the gene has appeared in the $Topk$ lists using the formula

$$Score_g = \sum_{m=1}^{7} Topk[m,g]. \tag{5.1}$$

6. Include the hub-genes ($lHub$) that have a consensus score of ≥ 4.

In other words, all genes are regarded as hub-genes and investigated as possible biomarkers if they rank in $Topk$ for at least 4 centrality measurements. The set of potential biomarkers for further downstream investigation is taken from this list, or $lHub$.

5.4.4.1 Choosing the value of k

For the algorithm to work, the value of k must be selected appropriately. It is clear that the genes listed in the individual centrality lists as being within $Topk$ may not be present in $lHub$. In many cases, especially in denser modules, experimental research reveals that the list of $Topk$ genes in each centrality list varies dramatically. As a result, only a small number of genes have a consensus score of ≥ 4, while a greater number have a consensus score of 1 or 2. When looking for K hub-genes in $lHub$, the experimental investigation leads us to the following value for k:

$$k = \begin{cases} K, & \text{if } 10\% of MS \leq K \\ 10\% of MS, & \text{otherwise} \end{cases} \tag{5.2}$$

where, MS is the module size in terms of no. of genes belonging to the module.

5.4.4.2 Correctness, completeness and complexity of the Algorithm

Correctness: Due to the unavailability of adequate ground truths and the algorithm being an unsupervised problem, it is not possible to evaluate the correctness of the same. We aim to identify novel biomarkers and the biomarkers have been validated based on literature trace, GO enrichment, and pathway enrichment using the proposed biomarker criteria as discussed in detail in Section 2.5.

Completeness: We attempt to extract hub-gene by considering seven centrality measures in an unbiased manner. The consensus measure of ≥ 4 ensures that the genes deemed as hub-genes by our algorithm rank among the top k genes for atleast four out of seven centrality measures. This ensures that a gene that has a high centrality score for only one centrality measure does not show up among the final identified hub-genes.

None of the hub-genes are left out that fulfill the pre-specified criteria such as: 1) it is a hub-gene, it is among the top k genes in atleast 4 centrality measures, and 3) its consensus score ≥ 4.

Complexity: To evaluate the running time of the algorithm, we initially analyze the algorithm in parts. For each biologically relevant module, *modGenes* is the network connections of a module with n genes, and e edges. The running time for each component of the algorithm is analyzed below.

1. The first component computes the centrality measure score of each gene in a given module. For each centrality measure , $m \in M$, the complexity can be computed as O(complexity of centrality measure, m). The centrality score is computed for all genes in the connectivity graph (i.e, module). As such the final complexity of this module can be analyzed as $O(n \times (O(m_1) + O(m_2) + \ldots O(m_7)))$. The time complexities of betweeness centrality, closeness centrality , degree centrality, eigenvector centrality, katz centrality, page rank, and radiality are $O(ne+n^2)$, $O(ne+n^2)$, $O(n^2)$, $O(n^3)$, $O(n^3)$, $O(e)$, and $O(ne+n^2)$, respectively. It is noteworthy that for very dense networks e is equivalent to n^2. As such we can conclude that for most centrality measures $T(n) = O(n^3)$. Taking the worst time complexity of $O(n^3)$, we can conclude that the time complexity of this component is $T(n) = O(n \times n^3) \sim O(n^4)$.
2. The second component creates separate lists of genes (n) in *modGenes* for each $m \in M$ based on the corresponding centrality measure scores. Thus, $T(n) = O(n)$. These m lists are then sorted in descending order based on centrality scores. By considering efficient sorting algorithms such as merge sort, these seven lists can be sorted in $O(nlog(n))$. Thus, for this component $T(n) = O(7 \times nlog(n)) \sim O(nlog(n))$.
3. The third component assigns flag value 1 to the the top k genes in the seven sorted lists. This can be achieved in $O(7 \times k) \sim O(k)$ time. However, as discussed in previous subsection, we choose k using equation 5.2. As the choice of k is not constant and the number of genes (n) can determine the value of k, $T(n) = O(n)$.
4. The fourth component computes the consensus rank score for each gene in *modGenes*. This is done for all seven centrality measures and as such $T(n) = O(7 \times n) \sim O(n)$.
5. The fifth and final component determines the hub-genes with consensus score $geq4$ by analyzing the consensus scores of all genes in *modGenes*. As such, $T(n) = O(n)$.

Finally, we can conclude that the time complexity of the algorithm is determined as follows.

$$T(n) = O(n^4) + O(nlog(n)) + O(n) + O(n) + O(n) \sim O(n \times n^3) \sim O(n^4) \quad (5.3)$$

In other words, for each biologically relevant module *modGenes* with n genes, the time complexity is $O(n^4)$. Here, n is of moderate size, since it represents the cardinality of the set of genes included in an MoI. GPU based implementation of the algorithm can further reduce the computation cost.

5.4.5 Validation

We take two approaches to validation. In order to determine the hub-genes ($lHub$) indicated by our proposed framework as potential biomarker(s), we first evaluate the quality of the module(s) retrieved by the module identification unit of the framework as ‘Module of Interest’ (MoI) (Definition 4.3.1 in Chapter 4). The following steps are taken to validate modules:

(a) GO enrichment analysis is used to evaluate the quality of an extracted module, and

(b) Enhanced pathway presence is used to further evaluate the quality of modules.

All hub genes found in biologically significant modules found by the Hub-gene discovery unit are regarded as potential biomarker candidates and are referred to as Biomarker Candidate Genes (BCG) (Definition 5.4.1). A module is pathway and GO enriched if it contains at least one enriched pathway and one enriched GO word. Gene Ontology (GO) enrichment analysis and pathway enrichment analysis are used to validate MoIs found by the preservation analysis (Section 2.1.9) unit. All detected MoIs are used as input in the validation unit’s pathway enrichment analysis and GO enrichment sub-unit in the framework. These subunits calculate the percentage of enriched GO words (PEGoT) across the three GO databases for each MoI. These three databases include the percentage of enriched pathways (PEP) in KEGG with a $p-value = 0.05$ and the biological process (BP), cellular component (CC), and molecular function (MF) databases.

Definition 5.4.1 (BCG)**.** A gene g_i is defined as a Biomarker Candidate Gene (BCG) if it is identified as a hub-gene in a given MoI extracted by CBDCEM.

First, we find lgEGo and lgEP with $p-value = 0.05$ for each BCG identified by the framework that needs to be validated. The GO enrichment and pathway enrichment sub-units in the framework receive input from the DEGs discovered by the identification of DEGs unit. Two lists, lgEGo and lgEP, are the output. The list of BCGs, along with lgEGo and lgEP, are input to the biological analysis unit in order to validate the BCGs found by the hub-gene discovery unit of the framework. The biological analysis unit locates BCGs that have enriched GO keywords and enriched pathways associated to them. In other words, the BAU recognizes the BCGs that are present in lgEGo and lgEP. For the purpose of establishing the regulatory behaviour of these BCGs in the network, this unit further detects BCGs that are TFs and builds GRN. The validation unit of the framework's literature trace sub-unit finds BCGs that have published literature traces that support their status as biomarkers for ESCC or other SCCs that are closely related to ESCC. We select the BCGs that come under Cases 1 and 2 and classify them as potential biomarkers based on our biomarker criteria (Section 2.5).

5.5 Experimental Results

In order to assess the effectiveness of our method, CBDCEM, we examine the critical disease, ESCC. To assess the efficacy of our technique, three ESCC datasets were chosen, including GSE130078 for bulk RNA-Seq data, GSE20347, and GSE23400 from microarray data. The details of each dataset are described in Sections 2.6.1 and 2.6.2 (Table 2.1). The experimental evaluation is conducted on a DELL workstation running Windows 10 Pro and equipped with a 3.70GHz Intel(R) Xeon(R) W-2145 CPU and 64 GB of RAM. We run the experiments in the R programming environment (Section 2.2.1). The gene expression of cancers has been examined in all three datasets and contrasted with that of surrounding contrast tissue.

5.5.1 Pre-processing

Pre-processing for the two microarray datasets, GSE20347 and GSE23400, begins with the elimination of unnecessary redundant data. However, there are no missing values for either GSE20347 or GSE23400, so we continue down the pipeline. We begin by removing low read counts from the bulk RNA-Seq dataset using counts per million (CPM). Genes with $CPM > 1$ in at least two samples are filtered. By doing so, the number of genes is decreased from 57,783 to 22,183. In Section 2.7.1 and Section 2.7.2, the overall workflow we use for pre-processing the microarray and bulk RNA-Seq data is covered in depth, respectively. We implement these pipelines to prepare the data for subsequent downstream analysis.

5.5.1.1 Outlier Gene Detection

To find outliers, we begin by hierarchically clustering the samples. In the case of normal samples with a tree cut at height h=70 (Blue), we discovered a single outlier for GSE23047 as shown in Fig. 3.4a and Fig. 3.4b in Section 3.5.3. However, there are 2 outliers with a cut at h=130 (Red) in disease samples. Similarly, in GSE23400, tree cuts at heights of h=105 (blue) and h=95 (red) eliminate one and two outliers from the normal (Fig. 3.4c in Section 3.5.3) and disease (3.4d in Section 3.5.3) samples , respectively. Cuts at h=1500000 (Blue) and h=2000000 (Red) in the case of GSE130078 remove one sample of normal 3.4e in Section 3.5.3) and one sample of disease (Fig. 3.4f in Section 3.5.3).

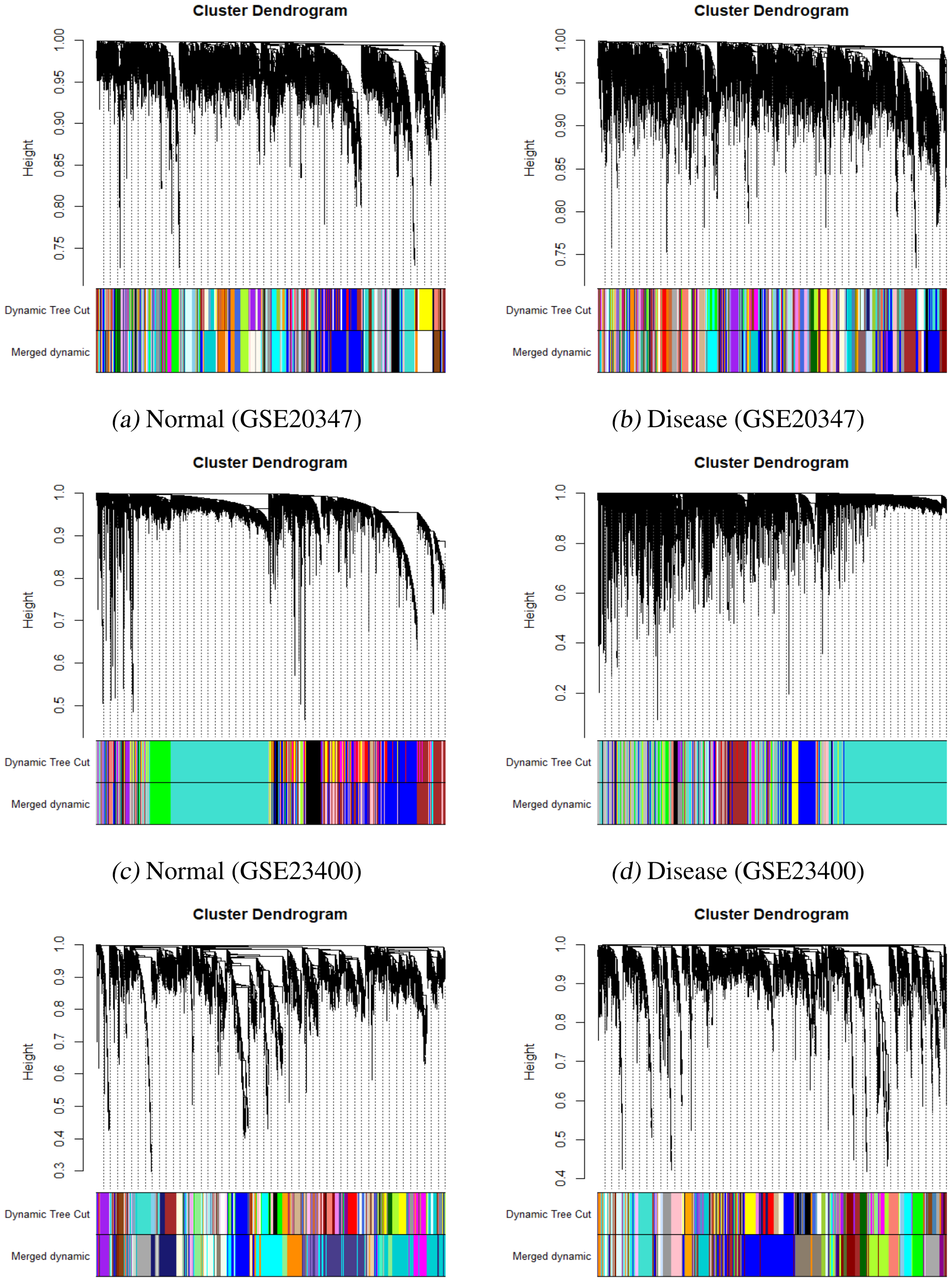
5.5.2 CEN Construction

We apply soft-thresholding to the normal (Blue) and disease (Red) samples of dataset GSE20347. Nine is the lowest power for which the network maintains scale-free topology, as can be shown in Fig. 3.5a and Fig. 3.5b in Section 3.5.3.1. As shown in Fig. 3.5c and Fig. 3.5d in 3.5.3.1, the soft thresholding for normal (Blue) and disease (Red) samples in GSE23400 is set at nine. In contrast, for GSE130078, normal (Blue) and disease (Red) samples are selected with soft thresholds of twelve (Fig. 3.5e in Section 3.5.3.1) and nine (Fig. 3.5f in Section 3.5.3.1), respectively. Using the soft thresholding exponent nine, we compute the adjacency matrices for the normal and disease samples of the GSE20347 dataset, yielding two corresponding matrices with a size of $22,277 \times 22,277$. Similar to this, GSE23400 produces two corresponding matrices of size $22,283 \times 22,283$ each with a soft thresholding power of nine. The number of genes in GSE130078 is decreased to 22,183 after CPM filtering, resulting in two adjacency matrices with soft thresholds of twelve (normal) and nine (disease). The adjacency matrices used to create the associated TOMs have the same size as the relevant adjacency matrix.

5.5.3 Module Extraction

As previously stated, we use DE analysis to find modules with sizeable dimensions. We use DESeq2 [450] for the bulk RNA-Seq dataset and Limma [638] for the microarray datasets (GSE20347 and GSE23400) for DE analysis. When DE analysis is applied to the whole dataset for GSE20357, the number of normal instances decreases from $22,277\times16$ to $8,474\times16$ and the number of disease instances decreases from $22,277\times15$ to $8,474\times15$. Similar reductions in normal and disease instances are made for GSE23400, with normal instances dropping from $22,283\times52$ to $13,338\times52$ and disease instances falling from $22,283\times51$ to $13,338\times51$ respectively. However, in the example of GSE130078, DE on the complete dataset yields $11,537$ DEGs, but the cases of normal behavior are decreased from $22,183\times22$ to $10,436\times22$ and the occurrences of disease are reduced from $22,183\times22$ to $11,316\times22$. The extraction of TOM values corresponding to these DEGs is then performed, resulting in smaller TOMs of sizes $8,474\times84,74$ (GSE20347) , $13,338\times13,338$ (GSE23400) , $10,436\times10,436$ (normal GSE130078) and $11,316\times11,316$ (disease GSE130078).

We employ hierarchical clustering to create a dendrogram of genes, resulting in 55 normal modules and 74 disease modules, in order to extract relevant modules. The first strip of colours below the dendrogram in Fig. 5.2a depicts the matching module colours from the normal dataset. Similarly, the dendrogram for the disease dataset is shown in Fig. 5.2b. Heirarchical clustering produces 17 normal and 18 disease modules for GSE23400. The dendrograms for the normal and disease datasets are shown in Figs. 5.2c and Fig. 5.2d, respectively, whereas the first strip of colours represents the colours allocated to these modules. For GSE130078, we obtain 48 disease (Fig. 5.3a) and 62 normal modules (Fig. 5.3b).



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

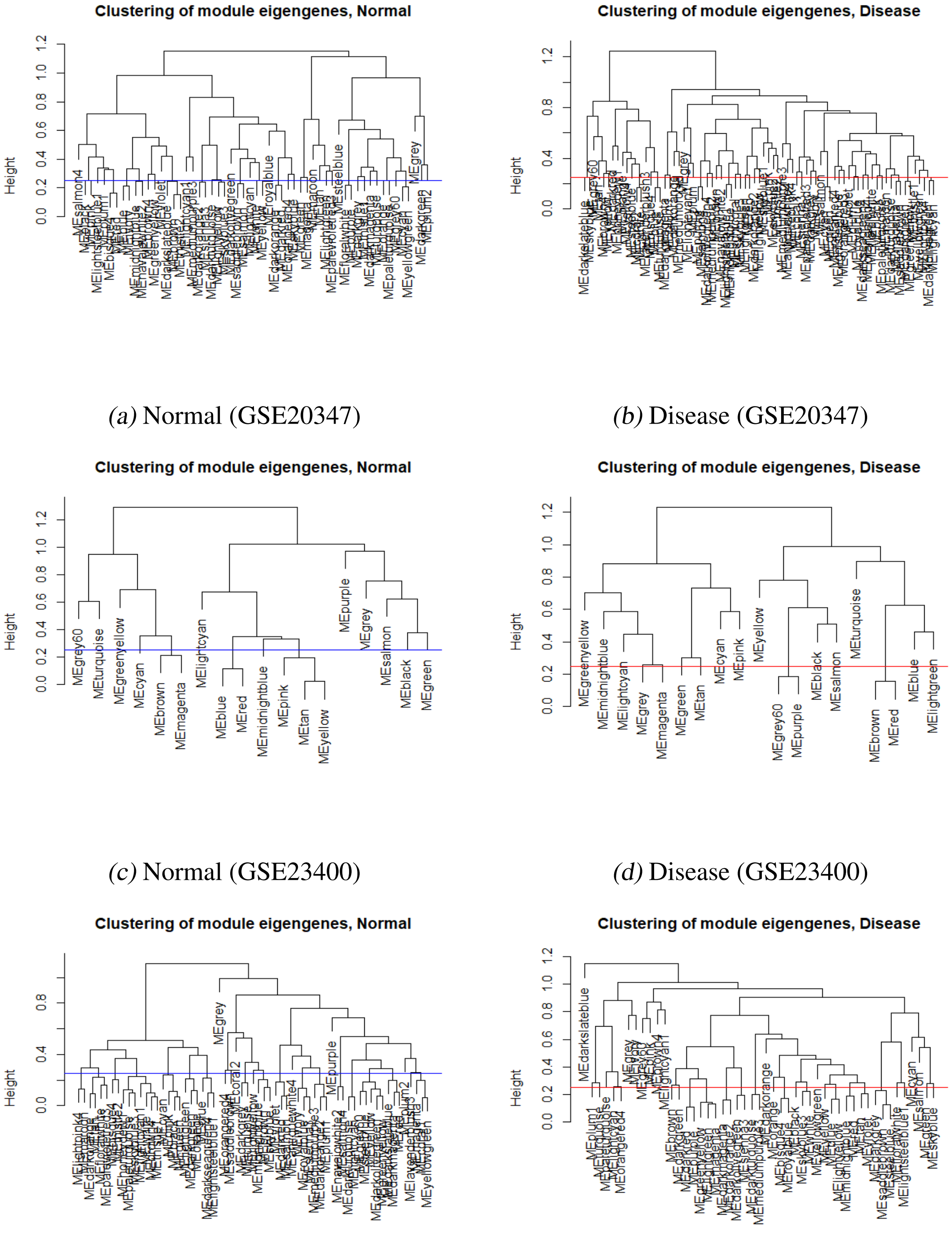
(d) Disease (GSE23400)

(a) Normal (GSE130078)

(b) Disease (GSE130078)

Fig. 5.3: Dendrograms for a) normal and b) disease in GSE20347, c) normal and d) disease in GS23400, and e) normal and e) disease in GSE130078. The first strip of colors represents the corresponding module colors assigned after hierarchical clustering while the second color strip of colors represents the corresponding module colors after merging.

We select a height cut of 0.25, which corresponds to a correlation of 0.75, to merge modules. For the normal and disease datasets in GSE20347, merging the modules with a tree cut at h=0.25 further reduces the number of modules to 40 and 63, respectively, as shown in Fig. 5.4a and Fig 5.4b. The integrated normal and disease modules are represented by the colors in the second color strip in Fig5.2a and Fig. 5.2b.



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

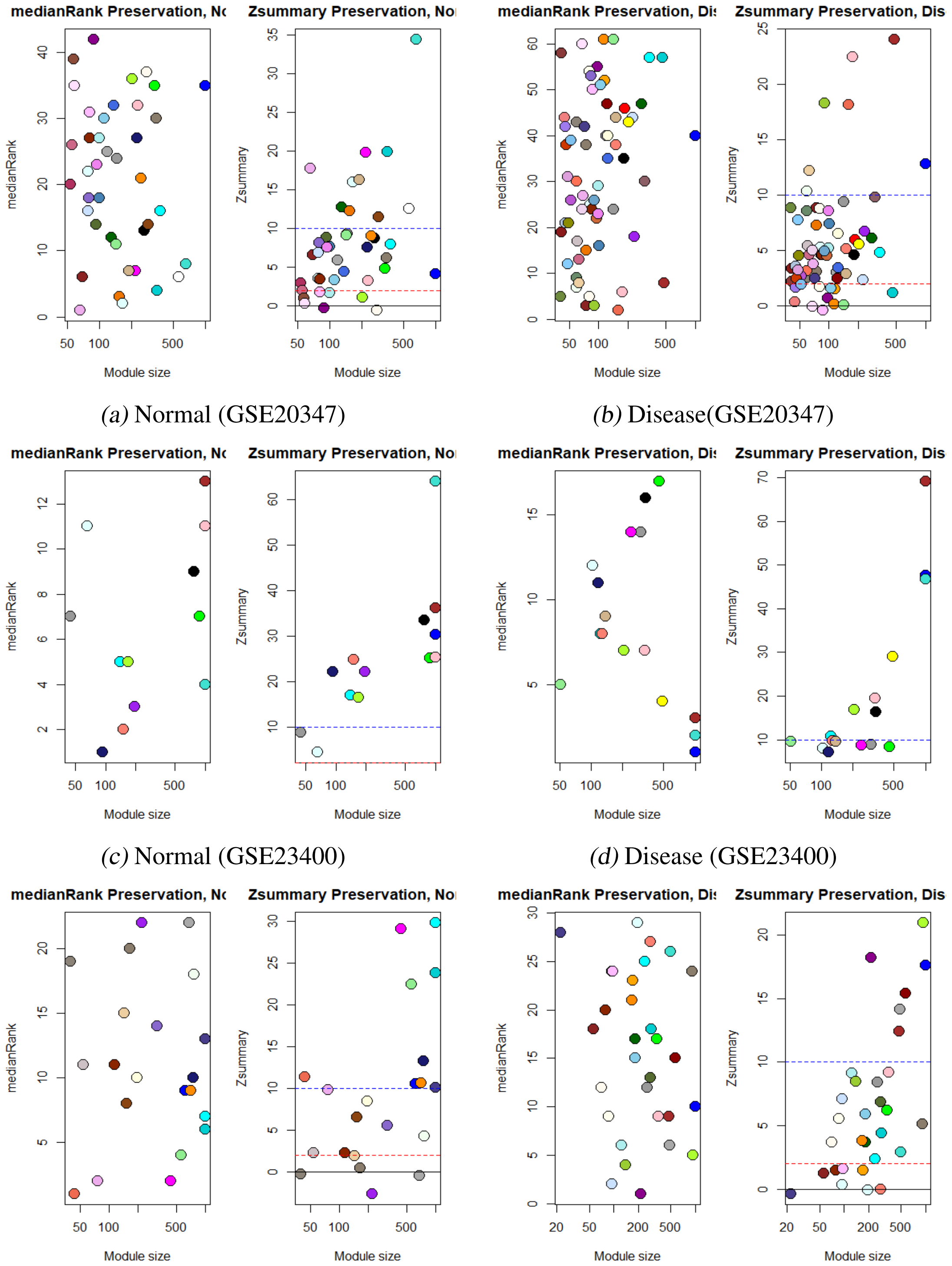
(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 5.4: Heiarchical Trees for module detection for a) normal and b) disease in GSE20347, c) normal and d) disease in GS23400, and e) normal and e) disease in GSE130078. The chosen tree cut is at height, h=0.25

A comparable tree cut at h=0.25 reduces the number of healthy modules for GSE23400 to 13 (Fig. 5.4c) and the number of disease modules to 16 (Fig. 5.4d). The merged module dendrograms in the normal and disease dataset for GSE23400 are shown in Fig. 5.2c and Fig. For GSE130078, h=0.25 yields 21 normal (Fig.5.4e) and 30 disease modules (Fig. 5.4f).



(a) Normal (GSE20347)

(b) Disease(GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 5.5: Zsummary plots for a) normal and b) disease in GSE20347, c) normal and d) disease in GS23400, and e) normal and e) disease in GSE130078. All modules below the red line are non-preserved, all modules between the red and blue lines are weak to moderately preserved and all modules above the blue line have strong evidence of being preserved.

In the Z_{summary} plots above all modules below the red line are non-preserved, all modules between the red and blue lines are weak to moderately preserved, and all modules above the blue line have strong evidence of being preserved.

In order to distinguish between preserved and non-preserved modules, preservation analysis is performed after module extraction. Contrary to preserved modules, the bulk of co-expression linkages in non-preserved modules vary (i.e., they are not preserved), and as a result, they might offer useful information for identifying causative genes. A module with $Z_{\text{summary}} < 2$ is deemed non-preserved according to the discussion in subsection 2.2.5 [329]. For GSE20347, GSE23400, and GSE130078, normal Z_{summary} plots are shown in Fig. 5.5a, Fig. 5.5c, and Fig. 5.5e, respectively. Similarly, Fig. 5.5b, Fig. 5.5d, and Fig. 5.5f are the disease Z_{summary} plots for GSE20347, GSE23400, and GSE130078, respectively. Table 5.4 compiles the preservation analysis for non-preserved modules across all three datasets. The module preservation reference and test networks are highlighted in the second column of both tables. In the Normal/Disease subset of dataset GSE130078, for instance, the table reading for module *purple* is of size 221, recognized in the normal network, but is non-preserved in the disease network with a Z_{summary} value of -2.61673. Dataset GSE23400 naturally extracts a number of modules with noticeably greater sizes and higher densities. As a result, no modules with $Z_{\text{summary}} < 2$ are non-preserved, and the majority of modules are either highly preserved ($Ztextsubscriptsummary > 10$) or moderately preserved ($2 \leq Z_{\text{summary}} \leq 10$). We have therefore considered moderately preserved modules with $Z_{\text{summary}} < 10$ [329, 569] in the absence of non-preserved modules in dataset GSE23400.

There are 14 and 25 non-preserved modules in datasets GSE130078 and GSE23400, respectively, as shown in Table 5.4. Similarly, GSE20347 has 22 non-preserved modules in Table 5.4. In Table 5.4, we consider modules of size $geq100$ to be modules of interest and streamline them for further analysis. These modules are marked by blue and **bold** in Table 5.4. Thus, the relevant modules for GSE130078, GSE23400, and GSE20347 are 7, 7, and 8 correspondingly. Modules with the color *grey* are background genes and are not taken into account. An example of a module with this color is module grey in the Disease/Normal section of Table 5.4.

Tab. 5.4: Preservation Analysis (Z_{summary}) of normal modules in disease dataset and vice versa in the two microarray ESCC datasets, GSE20347 and GSE23400, and the bulk RNA-Seq dataset, GSE130078. Modules with $\mathit{Size} \geq 100$ and at least moderately preserved (i.e, $Z_{\text{summary}} \leq 10$) are considered for further downstream analysis and highlighted in blue and **bolded**.

	Ref/Test	Module	Size	Z_{summary}
GSE130078	Normal/ Disease	***purple***	263	-2.61673
		darkgrey	691	-0.49294
		antiquewhite4	39	-0.23228
		bisque4	173	0.46919
		navajowhite2	146	1.95233
	Disease/ Normal	*grey*	34	-0.53381
		darkslateblue	22	-0.36112
		lightcyan	290	-0.09802
		salmon	320	-0.00052
		lightcyan1	96	0.324719
		brown4	56	1.22299
		orangered4	80	1.45160
		orange	212	1.45246
		plum1	97	1.59667
GSE23400	Normal/ Disease	*grey60*	44	8.70877
		lightcyan	66	4.43938
	Disease/ Normal	*grey*	1000	1.37004
		midnightblue	118	7.19999
		lightcyan	105	7.98239
		green	450	8.39948
		magenta	243	8.78524
		grey60	300	8.84688
		tan	139	9.52371
		lightgreen	51	9.58615
		salmon	128	9.83334
GSE20347	Normal/ Disease	***floralwhite***	279	-0.57028
		grey	24	-0.30498
		darkmagenta	88	-0.24808
		thistle1	58	0.27944
		salmon4	57	0.95925
		greenyellow	204	1.11967
		paleturquoise	100	1.68276
		plum1	81	1.79234
		palevioletred3	55	1.92806
	Disease/ Normal	*grey*	2	-0.58560
		plum1	86	-0.38140
		thistle1	68	-0.02133
		lightgreen	142	0.06532
		darkorange	113	0.12787
		lightcoral	44	0.35072
		darkmagenta	98	0.67363
		darkturquoise	456	1.16855
		orange	116	1.48801
		skyblue	105	1.61022
		mediumpurple2	45	1.67677
		ivory	81	1.70243
		skyblue2	52	1.94493

5.5.4 Hub-gene Finding

We apply our suggested hub-gene discovery approach, which uses seven centrality measures on all 22 modules of interest found in the three datasets, as explained in subsection 5.4.4. From the CEN created on the full dataset, we extract the network corresponding to a module of interest, which is subsequently used as input by our hub-gene discovery approach. A list of the hub-genes found in the module is the algorithm's output. We test our findings using the three datasets and a $K = 20$ threshold. We use our hub-gene identification algorithm to select the top 20 ($K = 20$) hub-genes for each of the 22 modules of interest in order to locate possible biomarkers for ESCC. Further testing is done on these hub-genes to find potential ESCC biomarkers.

Tab. 5.5: Top 20 hub genes for each extracted module of interest in all three datasets using our hub-gene finding algorithm. Hub genes with strong literature evidence of association to Esophageal Squamous Cell Carcinoma (SCC) are marked in Red while hub genes with evidence of association with five other SCCs namely, HNSCC, LaSCC, LSCC, OSCC, and TSCC are marked in Blue

	Module	hub genes
GSE20347	*darkturquoise*	*USP7, AAMP, GALNT1, ZNF107, KRT6A, LEPR, GATD3A, HOXA10, CAMKK2, ZNF273, PHLDA1, DHX32, THEM6, SRR, CHODL, MARK1, RAB35, TRIB2, SPRY2*
	lightgreen	*SDHC. PNRC2, H2AJ, SEC24C, BASP1, ZBTB1, SNURF, GRB10, WNK1, DST, DOP1A, PPP6R3, RWDD2B, GMEB1, GPSM2, PEG3, CEP152, VPS13A*
	darkorange	*KMT2A, NFYA, MLX, RABGEF1, THNSL1, PDPK1, TSG101, HOXD4, CALB1, PNMA2, SUZ12P1, ANKRD36, SUGP1, ACSF2, GALNT12, PEX26, TMEM80, PRDM2*
	orange	*TCOF1, ACO1, FXR1, DHRS12, SPTBN2, SLC18A2, SLC16A6, PWP2, DGKA, AHNAK2, BCL6, PIAS1, TTC31, SLC24A3, AHNAK, ABHD17B, AUP1, HSF1, CCNI*
	skyblue	*DPYSL3, WSB2, ARPC3, POFUT2, RFC5, PRDM4, DDX54, TNS1, JMJD7, MAP4, PLD3, TDG, PFN2, HSPA4, PRKAA1, DGCR11, PSENEN, RPL22, CACNB3*
	floralwhite	*UBE2L3, TMX4, CD163, ZKSCAN5, AURKA, EIF2AK2, RGS14, PTGS1, VNN2, GINS1, PLXNC1, DUSP6, RAPSN, BBC3, SCD, MK167, HOXA10, SLC49A3, DGCR11, NR1H2*
	greenyellow	*RALY, SLC22A4, HMMR, FST, TNFRSF9, NR2C1, MARK2, FMO2, SYDE1, OSGIN2, RLN2, IGKC, ITGA2B, RANGRF, TSPAN15, ARMH3, DNAJC12, PIMREG, EPOR*
	paleturquoise	*AP1B1, ABR, ZNF556, SDC1, ANCBP2, UQCR2, SRSF5, NRG2, ACVRIB, SPP1, LST1, UPF1, UBE3B, IP6K1, CEP170, CYP3A5, GOLT1B, MCHR1, DHRS7B, ARMH3*
GSE23400	*grey60*	*APRC1B, DDOST, VGLL4, CAMSAP1, TNFSF10, SFTB, CD38, NCSTN, SMARCC1, PSMA5, CTSC, SND1, DBN1, MAPK9, IGFBP5, PIK3CD, MMP1, COCH, TTLL1, FOXA2*

Continued on next page

	Module	hub genes
GSE23400	*lightcyan*	*VARS1, NDUFB7, MAD2L1BP, TCF3, KIFC1, USP39, MEA1, DNAAF2, ATP13A3, SINHCAF, SPDL1, CDC6, YIPF2, PAGR1, STMN1, TMPO1, MCM7, TPT1, DNAJB12, ARHGEF3*
	green	*CALM3, CTSA, PRDX6, SREBF2, TMEM109, NIPSSNAP1, DCN, PRDX4, UBE2B, APPBP2, DNTTIP2, FECH, MMP15, PSMG1, STK19, PSEN2, RGS10 , SOX12, GLI3, FGF8*
	magenta	*DGCR2, LPIN2, UBTF, ZNF74, CYP2R1, NOLC1, SPIDR, NUS1P3, ACOT7, UBE2J1, DYNC1LI1, PDP1, PLXNA1, ISYNA1, IDS, GEMIN2, SNRNP40, DSCC1, GAK, TOMM70*
	tan	*FTSJ1, NUP62, SUMO4, SAMD4B, APP, PRKAR1A, MYL12A, UBE2D2,TMED5, CD2BP2, RNGTT, LIMK1, ANXA3, COA1, DBP, MSC, CMC4, KHK, RAB7A, MZF1*
	salmon	*ARHGEF1, HUS1, NCF1, GNG7, TRAF3IP1, GK, HACD3, YTHDF2, CDC73, C1orf109, LRRC2, SMG5, TAF6L, IPCEF1, RNF121, AK1, MTRR, PARM1, POLA1, HIC1*
	midnighblue	*YBX3, ZNF200, HAUS3, GOLGA8A, RBM25, MCFD2, PUS7, ETNK1, SUPT7L, SEL1L, MYC, FNTB, EHD2, RAD52, NRG1, HINFP, TNFSF4, ATXN10, LGALS8, ITGB4*
GSE130078	*lightcyan*	*DDHD1, CSRNP2, ST8SIA1, PRELID3A, DIP2C, SPATA33, CIQTNF1, RPS10P7, MTND5P26, CD28P2-DT, ATP2C2, FAM128A, PRSS53, ADAMTS9, DNAH17, KRTT8P42, GRIN3B, C2CD4C*
	salmon	*OTUB2, PRNP, PARP12, JMJD6, TXLNG, BIRC5, MAP3K1, OAS2, KYNU, CAAP1, PRRG1, RSAD2, CMPK2, SQOR, PML, IL18, MIER3, RRP1B, SGO2, TGM4*
	orange	*ADIPOR2, AGPAT4, CHI3L2, NOP2, DNPH1, CHTF18, EBP, SH2B2, TBC1D24, ZNF75A, BANP, INPP5F, MYL5, AP4M1, NRSN2-AS1, MINCR, EEF1A1P38, CBX3P2, C17orf67*
	purple	*ARHGAP33, CCL26, CATSPERG, SAMD15, PTCH2, TESMIN, HLX, REN, ZNF474, KRT75, EIF3J-DT, DI01, PLA2G12AP1, CAMK2A, AVPI1, KLHL31, FBXO43, HYDIN, KLF11, NOX5*
	darkgrey	*TENM1, PCDHB4, ANGPT2, ATPSF10, TXN2, PIEZO1, UBE25, SRM, IRF2BPL, SORBS3, SDHC, ARF1, PPP4C, GXYLT1, JOSD2, STX5, NFIL3, CD3002, FAM241B, WIPF2*
	navajowhite2	*NEK11, ALDH8A1, STXBP1, SLC2A4, MAD2L1P1, CSPG4P11, LOC100287042, PI15, PDLIM3, TUBA1A, UCP3, NANOS1, MBNL1-AS1, LOC392266, DNAJB5*
	bisque4	*PLEKHG6, CXCL2, P3H2, CD4OLG, EDF1, TNFSF11, LINC00243, H2BP1, MYLK-AS1, PRKXP1, AK4P1, MAGI1-I71, LEKR1, SNORA80B, RPS9P2, CPEB2-DT, GOLGA8B*

5.6 Validation

Several approaches have been used to validate the biological significance of modules found by CBDCEM's module detection unit as well as the validation of hub-genes found by the hub-gene discovery unit to establish them as possible biomarkers.

5.6.1 Enrichment Analysis of Modules

GO and pathway enriched genes are evidence of the biological relevance of each module of interest. We employ the simple web programme DAVID [628, 253] to carry out functional enrichment analysis. The percentage of genes in the relevant module annotated to enriched GO keywords and enriched KEGG pathways is summarized in Table 5.6.

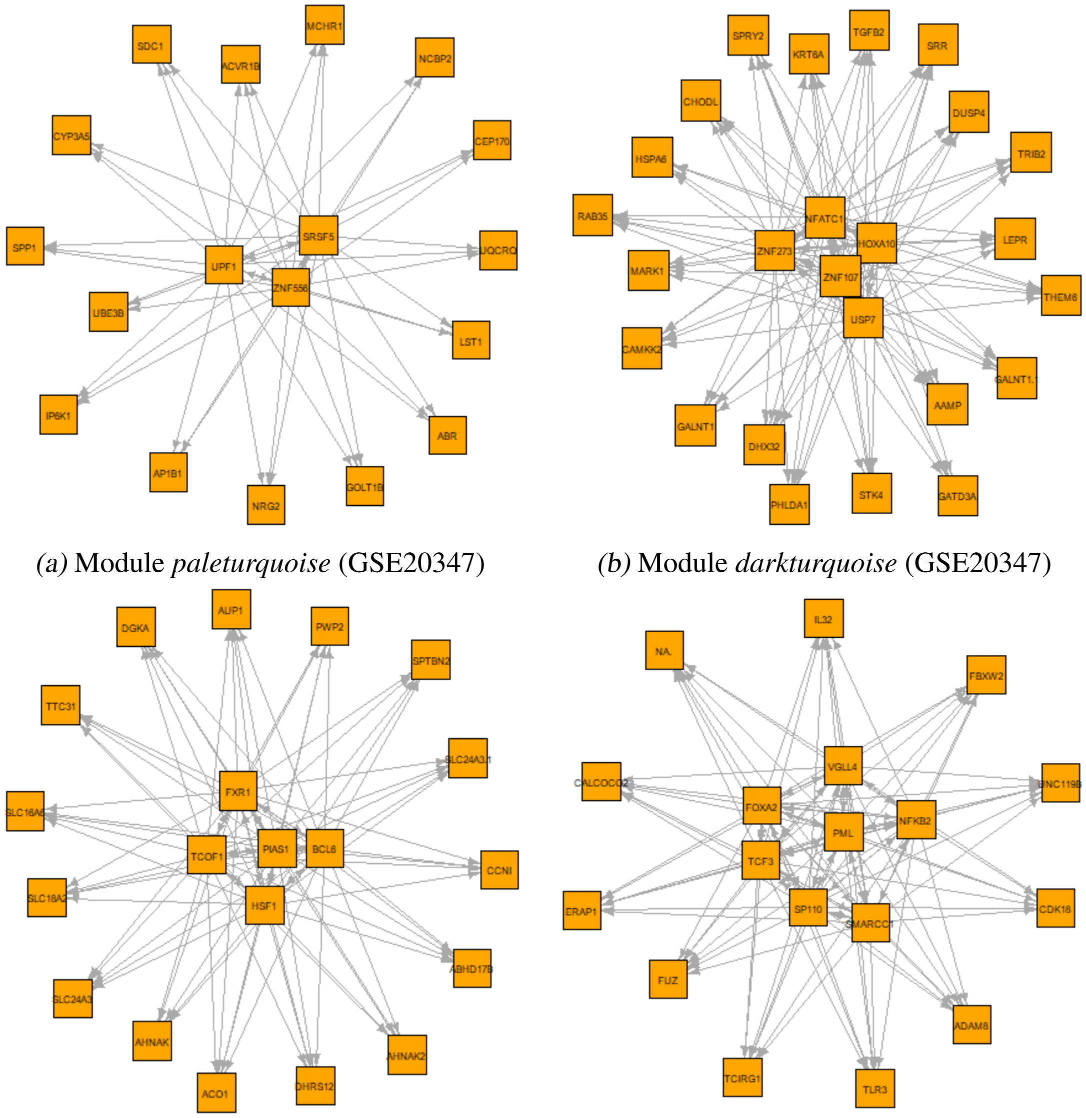
Tab. 5.6: Percentage of genes in each module that are annotated in the Gene Ontology (GO) databases (BP: Biological Processes, CC: Cellular components or MF: Molecular function) and KEGG pathways.

	Module	Size	BP	CC	MF	KEGG		Module	Size	BP	CC	MF	KEGG
			(%)	(%)	(%)	(%)				(%)	(%)	(%)	(%)
GSE20347	*darkturquoise*	456	93.5	94.0	96.4	48.2	GSE23400	*grey60*	300	95.2	94.3	95.7	53.5
	lightgreen	142	89.3	90.1	93.4	44.6		*green*	450	92.6	94.2	96.4	47.0
	darkorange	113	91.8	92.8	96.9	40.2		*midnightblue*	118	96.8	94.6	98.9	44.1
	orange	116	95.0	93.0	97.0	43.0		*lightcyan*	105	93.9	92.7	95.1	48.8
	skyblue	105	92.0	92.0	93.1	44.8		*magenta*	243	93.7	92.1	95.3	44.7
	floralwhite	279	90.5	94.2	96.3	43.2		*tan*	139	87.9	86.9	93.5	37.4
	greenyellow	204	92.0	94.8	94.3	42.5		*salmon*	128	88.3	91.5	84.0	44.7
	paleturquoise	100	92.3	91.2	95.6	35.2							
GSE130078	*lightcyan*	290	70.4	68.7	74.3	21.2	GSE130078	*darkgrey*	691	83.5	83.1	88.9	32.2
	salmon	320	81.3	81.0	86.1	37.0		*navajowhite2*	146	57.8	63.7	62.7	25.5
	orange	212	71.1	71.7	75.9	28.9		*bisque4*	173	76.9	76.2	80.0	33.1
	purple	263	79.7	78.1	90.3	27.4							

5.6.2 Biological Analysis

As mentioned in subsection 5.4.5, we employ functional enrichment analysis and the creation of gene regulatory networks to determine the biological relevance of the hub-genes discovered by CBDCEM.

The diversity and power of transcription factors (TF) as agents of cell change is astounding. The continued search for TFs as possible biomarkers [45] is justified by the fact that the deregulation of TFs is a common trend across many types of human cancer. We have found that the hub-genes identified by CBDCEM in GSE20347, GSE23400, and GSE130078, respectively, are TFs in 41, 45, and 23 cases. The biological importance of these TFs is demonstrated by their regulatory behavior in their respective modules. We have taken a reasonable selection of hub-genes from the non-preserved modules discovered by CBDCEM for simple visualization. In order to track the regulatory behavior of the corresponding genes, we build a Gene Regulatory Network (RN) using these hub-genes and related TFs. An adjacency list with weighted directed edges from TFs to other target genes (TGs) makes up the RN that results from this process.



(a) Module *paleturquoise* (GSE20347)

(b) Module *darkturquoise* (GSE20347)

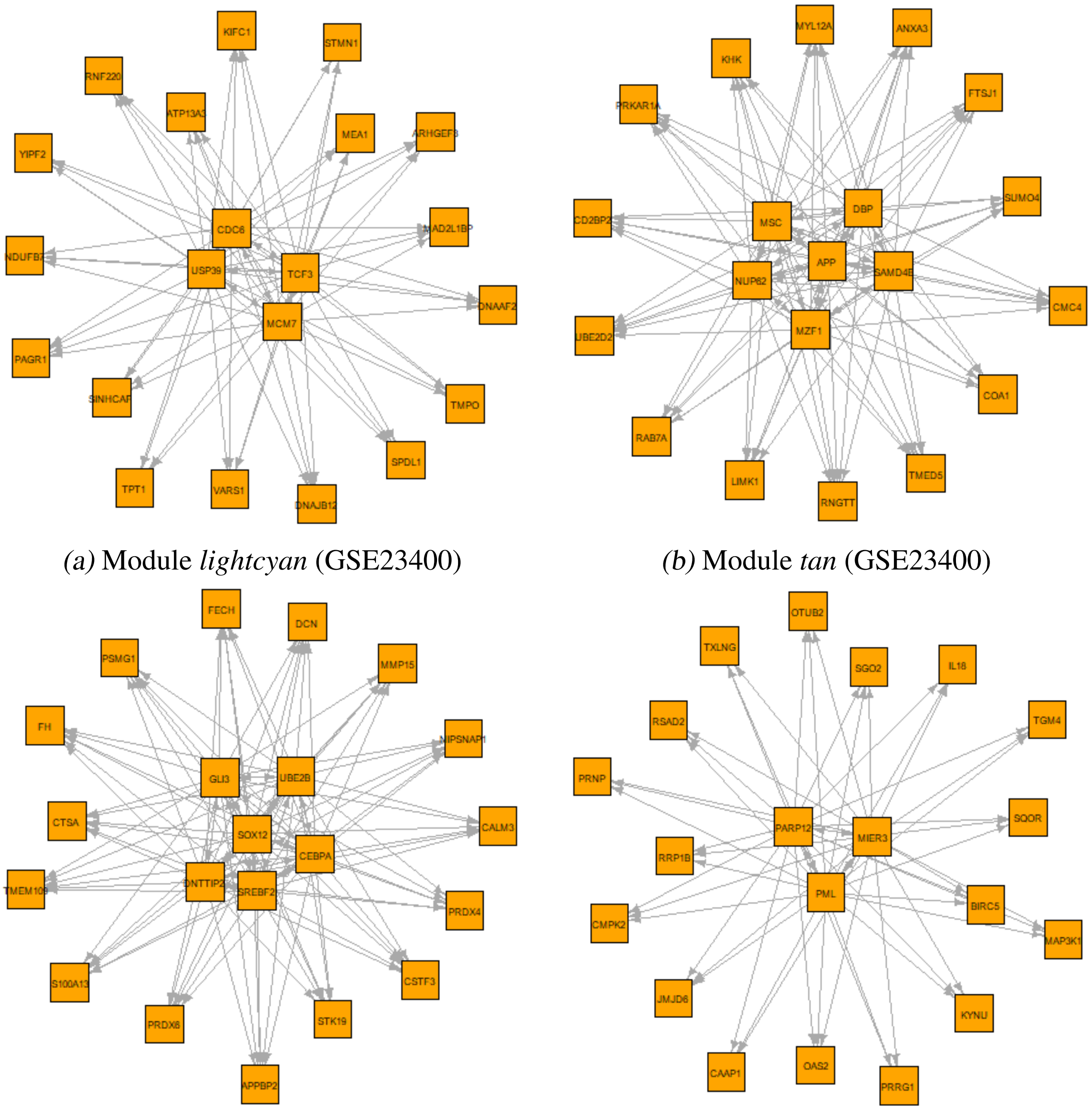
(c) Module *orange* (GSE23400)

(d) Module *grey60* (GSE23400)

Fig. 5.6: GRN for normal module a) *paleturquoise*, disease modules b) *darkturquoise* and c) *orange* in GSE20347. GRN for disease module d) *grey60* in GSE23400

As shown in Fig 5.6d, seven hub-genes in the module *grey60* (GSE23400) are TFs:

VGLL4, FOXA2, PML, NFKB2, SMARCC1, SP110, TCF3, and NFKB2. All hub-genes that are TFs control both other genes in the module and hub-genes that are not TFs. The hub-genes that are TFs also control one another, in addition. Similar to this, 5 hub-genes identified by CBDCEM in module *orange* (GSE20347), namely FXR1, PIAS1, BCL6, TCOF1, and HSF1, are TFs (Fig. 5.6c.



(a) Module *lightcyan* (GSE23400)

(b) Module *tan* (GSE23400)

(c) Module *green* (GSE23400)

(d) Module *salmon* (GSE130078)

Fig. 5.7: GRN for disease module a) *lightcyan*, d) *tan* e) *green* in GSE23400, and disease module f) *salmon* in GSE130078.

We monitor the regulatory behavior exhibited by the BCGs detected by CBDCEM. We further perform GO enrichment (Section 2.4.1.1) and the pathway enrichment analysis (Section 2.4.1.2) on these BCGs. We employ the web tool DAVID [628, 253] (Section 2.2.3), just as when analyzing the enrichment of the modules.

Tab. 5.7: Summary of hub-genes detected by CBDCEM in GSE20347 , and GSE23400 annotated to the top 20 KEGG enriched pathways

KEGG Pathways	GSE20347	GSE23400
hsa05200:Pathways in cancer	*ITGA2B, STK4*	*MYC, GLI3, ARHGEF1, FH, FGF8, PIK3CD, MMP1, CEBPA, MAPK9, GNG7*
hsa04151:PI3K-Akt signaling pathway	*SPP1, ITGA2B, PDPK1, PRKAA1, EPOR*	*MYC, FGF8, ITGB4, PIK3CD, GNG7, EIF4B*
hsa04510:Focal adhesion	*SPP1, ITGA2B, PDPK1*	*MYL12A, ITGB4, PIK3CD, MAPK9*
hsa04010:MAPK signaling pathway	*CACNB3, STK4, DUSP6*	*MYC, STMN1, FGF8, MAP2K6, HSPA8, MAPK9*
hsa05205:Proteoglycans in cancer	*SDC1, PDPK1*	*MYC, EZR, PIK3CD, DCN, ARHGEF1, EIF4B*
hsa05161:Hepatitis B	None	*MYC, PIK3CD, MAPK9*
hsa05166:HTLV-I infection	None	MYC, TCF3, PIK3CD
hsa05164:Influenza A	*EIF2AK2*	*TNFSF10, MAP2K6, HSPA8, PIK3CD, MAPK9*
hsa05162:Measles	*EIF2AK2, BBC3*	*TNFSF10, HSPA8, PIK3CD*
hsa04380:Osteoclast differentiation	None	*NCF1, MAP2K6, PIK3CD, MAPK9*
hsa05152:Tuberculosis	*NFYA*	*CALM3, RAB7A, MAPK9*
hsa04080:Neuroactive ligand-receptor interaction	*LEPR, ADRB2, MCHR1*	*GNRHR*
hsa04015:Rap1 signaling pathway	*PFN2, ITGA2B, RGS14*	*CALM3, FGF8, MAP2K6, PIK3CD*
hsa05168:Herpes simplex infection	*USP7, SRSF5, EIF2AK2*	*SRSF6. TAF6L, MAPK9*
hsa05169:Epstein-Barr virus infection	*USP7, EIF2AK2*	*MYC, MAP2K6, PIK3CD, MAPK9*
hsa04810:Regulation of actin cytoskeleton	*PFN2, ITGA2B, ARPC3*	*EZR, MYL12A, ARPC1B, ARHGEF1, FGF8, ITGB4, PIK3CD, ITGAE, LIMK1*
hsa05146:Amoebiasis	None	*PIK3CD, RAB7A*
hsa04640:Hematopoietic cell lineage	*ITGA2B, EPOR*	*CD38*
hsa04020:Calcium signaling pathway	*ADRB2*	*CALM3, CD38*
hsa04060:Cytokine-cytokine receptor interaction	*LEPR, TNFRSF9, ACVR1B, EPOR*	*TNFSF10, TNFSF4*

Tab. 5.8: Summary of hub-genes detected by CBDCEM in GS130078 annotated to the top 20 KEGG enriched pathways

KEGG Pathways	Hub Genes
hsa04151:PI3K-Akt signaling pathway	*ANGPT2*
hsa05200:Pathways in cancer	*BIRC5, PML, PTCH2*
hsa04510:Focal adhesion	*ACTN4, MYL5*
hsa04010:MAPK signaling pathway	*MAP3K1*
hsa04024:cAMP signaling pathway	*GRIN3B, CAMK2A*
hsa04810:Regulation of actin cytoskeleton	*ACTN4, MYL5*
hsa04015:Rap1 signaling pathway	*ANGPT2*
hsa04020:Calcium signaling pathway	*CAMK2A*
hsa04261:Adrenergic signaling in cardiomyocytes	*CAMK2A*
hsa05205:Proteoglycans in cancer	*CAMK2A*
hsa05166:HTLV-I infection	*MAP3K1*
hsa04014:Ras signaling pathway	*ANGPT2*
hsa04728:Dopaminergic synapse	*CAMK2A*
hsa01130:Biosynthesis of antibiotics	*SDHC*
hsa04142:Lysosome	*AP4M1*
hsa04722:Neurotrophin signaling pathway	*IRAK1, CAMK2A, SH2B2, MAP3K1*
hsa04919:Thyroid hormone signaling pathway	*DIO1*
hsa04724:Glutamatergic synapse	*GRIN3B*
hsa04725:Cholinergic synapse	*CAMK2A*
hsa04114:Oocyte meiosis	*FBXO43, CAMK2A*

Contrary to modules, where enrichment analysis was carried out on the list of genes for each module separately, we do the analysis on whole datasets. The significance level that we used was 0.05. Alternatively stated, a GO term or pathway is deemed highly enriched if its p-value is ≤ 0.05.

The top 20 enriched KEGG pathways and the associated hub-genes discovered by CBDCEM and annotated to these pathways in the two microarray datasets (GSE20347, GSE23400) and the bulk RNA-Seq dataset (GSE130078) are summarized in the tables 5.7 and 5.8, respectively. Due to the similar gene sets in both microarray datasets, enriched KEGG pathways are present in both. Similarly, Tables 5.9 and 5.10 give a summary of the top 10, 3 and 3 enriched GO Terms in GO_BP, GO_CC and GO_MF databases and the corresponding set of hub-genes detected by CBDCEM that are annotated to these GO terms in the two microarray (GSE20347, GSE23400) and one bulk RNA-Seq dataset (GSE130078), respectively.

Tab. 5.9: Summary of hub-genes detected by CBDCEM in GS20347, and GSE23400 annotated to top GO terms in the three GO databases

	GO Term	GSE20347	GSE23400
GO_BP	GO:0007165 signal transduction	*PPP4R1, FST, NRG2, PRKAA1, PRDM4, NCKIPSD, ACVR1B, EPOR, ABR, GRB10, GOLT1B, SYDE1, STK4*	*PIK3CD, SH3BP2, TNFSF10, CD38, STMN1, TNFSF4, TLE3, IGFBP5, MAP2K6, LIMK1*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*KMT2A, PEG3, HSF1, ACVR1B, GMEB1, NR1H2, HOXA10,ADRB2*	*SS18, CDC73, NRG1, SMARCC1, SUB1, EAF2, SREBF2, FOXA2, MZF1, TCF3, MYC, SOX12, GLI3, ZEB1, KAT6B, PAGR1, CEBPA, APP, DBP, DCN, LPIN2*
	GO:0042493 response to drug	*SRR, DUSP6*	*MYC, CD38, SS18, TERF1, FGF8, FECH, MAP2K6, UBE2B, MCM7*
	GO:0045893 positive regulation of transcription	*NR1H2, PIAS1, KMT2A, CAMKK2, NFYA*	*MYC, ZXDC, GLI3, HINFP, KAT6B, SMARCC1, CD38, FOXA2, TCF3, NUP62*
	GO:0006366 transcription from RNA polymerase II promoter	*GMEB1, KMT2A, NFYA, PRDM4, HOXA10, NCBP2*	*MYC, SOX12, GLI3, CEBPA, RNGTT, SUB1, EAF2, MZF1, MSC, DBP, NCBP1*
	GO:0016032 viral process	*USP7, PDZD8*	*SND1, DYNC1LI1, POLA1, MICB, HSPA8, NUP62, MMP1, CEBPA, RNGTT*
	GO:0010628 positive regulation of gene expression	*RLN2, PRKAA1, SPRY2*	*MYC, EZR, HINFP, FGF8, PIK3CD, MAPK9*
	GO:0043066 negative regulation of apoptotic process	*EIF2AK2, PIAS1, PRKAA1, AURKA, SPRY2*	*MYC, GLI3, BLK, TPT1, CD38, PSEN2, UBE2B, YBX3, NUP62*
	GO:0008283 cell proliferation	*MKI67, UBE2L3, PRDM4*	*MYC, CMC4, POLA1, ZEB1, ARHGEF1, MCM7, NRG1*
	GO:0043065 positive regulation of apoptotic process	*ABR, PSENEN, PNMA2, STK4, BCL6, DUSP6*	*TNFSF10, NCSTN, TERF1, ARHGEF3, PSEN2, MAP2K6*
GO_CC	GO:0005737 cytoplasm	*DPYSL3, EIF2AK2, SRSF5, KMT2A, IP6K1, PRKAA1, SRR, SDC1, PRDM4, TSG101, HOMER2, SMG5, HECTD3, MKI67, FXR1, GINS1, GPSM2, DST, AAMP, UBE2L3, DHX32, MLX, RANGRF, RPL22, GRB10, LST1, PFN2, RGS14, CHODL, ARPC3, ACO1, PPP6R3, TRIB2, PNPLA3, UBE3B, PDPK1, GMEB1, NR1H2, THNSL1, AP1G1, CEP170, SPTBN2, HOXA10, UPF1, PTGS1, TCOF1, SNURF, WNK1, CAMKK2, HSF1, QDPR, NCBP2, AHNAK, AHNAK2, TNS1, MAP4, SYDE1, BASP1, STK4, MARK1, RMND5A*	*URI1, POLA1, KHK, ARHGEF1, DNAAF2, SMG5, FGD6, PARM1, MAP2K6, COA1, MCM7, TPT1, MTRR, ISYNA1, CAMSAP1, DCN, SND1, PRKAR1A, HACD3, FH, SAMD4B, TTLL1, UBE2J1, DBN1, FOXA2, APPBP2, RBM25, MEA1, PRDX6, GLI3, EZR, RNF220, ETNK1, CDC73, NRG1, ATXN10, SREBF2, MORC2, UBE2B, TCF3, GNE, TERF1, ZEB1, YTHDF2, HIC1, GOLGA8A, PLXNA1, SNRNP40, TRAF3IP1, TUBB, YBX3, LIMK1, PSMG1, PSMA5, NCF1, FDYNC1LI1, TSJ1, TMPO, CD2BP2, CALM3, GEMIN2, PPP1R9A, NUP62, NCBP1, STMN1, ACOT7, NOLC1, S100A13, MAD2L1BP, CDC6, PRSS21, ANXA3, AK1, CROCC, APP, TRAK1, IPCEF1*

Continued on next page

	GO Term	GSE20347	GSE23400
	GO:0005829 cytosol	*DPYSL3, SAR1B, EIF2AK2, BBC3, IP6K1, ARPC3, ACO1, PPP6R3, VPS13A, PCCA, PRKAA1, NCKIPSD, PDPK1, SMG5, USP7, DUSP6, HMMR, AP1G1, CEP152, AP1B1, SPTBN2, UPF1, DMD, DST, SEC24A, SEC24C, WNK1, HSF1, QDPR, CALB1, NCBP2, AURKA, RPL22, ABR, AHNAK, DGKA, GRB10, PHLDA1, CACNB3, CUL3, SYDE1, STK4, HSPA4, SPRY2*	*SULT1C4, ETNK1, CENPM, FNTB, SPDL1, EIF4B, ATXN10, LGALS8, KHK, SREBF2, GAK, ARHGEF3, ARHGEF1, MORC2, HUS1, GNE, EHD2, ARPC1B, SMG5, YTHDF2, MAP2K6, MCM7, HAUS3, MTRR, MYL12A, RAB7A, ISYNA1, MAPK9, LIMK1, LPIN2, PRKAR1A, PSMA5, FH, GK, NCF1, UBE2D2, PIK3CD, CALM3, GEMIN2, PPP1R9A, NCBP1, MYC, STMN1, ACOT7, PRDX4, S100A13, CDC6, PRDX6, GLI3, AK1, LSM7, APP, EZR, IPCEF1, HSPA8, RGS10*
	GO:0016020 membrane	*EIF2AK2, SLC18A2, ARPC3, SDC1, PNPLA3, SDHC, DHRS7B, PDPK1, AUP1, PLXNC1, HMMR, AP1G1, MKI67, FXR1, ELOVL1, SLC16A6, LEPR, WNK1, DDX54, ACVR1B, GALNT1, ABHD17B, ABR, AHNAK, FMO2, DGKA, CACNB3, CUL3, CD163, LST1, GOLT1B, KRT6A, PDZD8, MARK2, SCD, SPRY2*	*ETNK1, NRG1, ATXN10, LGALS8, GNRHR, SREBF2, GAK, DNAJB12, NCSTN, EHD1, KIFC1, MCM7, NOL9, LIMK1, ATP13A3, SND1, PRKAR1A, DDOST, TMPO, SLC12A9, CTSA, CD38, CTSC, PSEN2, STMN1, DYNC1LI1, ENDOD1, PRDX6, PRSS21, ANXA3, GOLGA5, EZR, TOMM70, HSPA8*
GO_MF	GO:0005515 protein binding	*BBC3, PIAS1, ZKSCAN5, VPS13A, PRKAA1, PEX26, ZBTB1, SDC1, PLD3, TDG, RALY, HOMER2, BCL6, HECTD3, MKI67, SLC22A4, NFYA, DMD, DST, UBE2L3, DHX32, ACVR1B, RPL22, GRB10, CD163, PNRC2, GOLGA2, HSPA4, RGS14, LCAT, FST, ACO1, TSPAN15, SUGP1, NCKIPSD, EPOR, AUP1, PLXNC1, RAB35, AP1G1, CEP170, RFC5, CEP152, AP1B1, PSENEN, CDCP1, HOXA10, UPF1, TCOF1, LEPR, SEC24A, SEC24C, CALB1, NCBP2, AURKA, AHNAK, RWDD2B, BASP1, ADRB2, MARK2, DPYSL3, EIF2AK2, SRSF5, KMT2A, IP6K1, PRDM4, TSG101, SMG5, USP7, HMMR, FXR1, GPSM2, ELOVL1, MLX*	*MMP15, FNTB, BLK, SPDL1, TNFSF10, LGALS8, SUB1, DNAAF2, HUS1, MAP2K6, MCM7, CEBPA, MTRR, TMED5, ISYNA1, DCN, PRKAR1A, FH, GK, UBE2D2, PIK3CD, YIPF2, APPBP2, MEA1, LSM7, RNF24, EZR, HSPA8, CDC73, EIF4B, ATXN10, GAK, UBE2B, SPIDR, GTPBP3, GNE, TERF1, TAF6L, YTHDF2, TLE3, HIC1, SNRNP40, TRAF3IP1, MYL12A, VGLL4, RAB7A, MAPK9, LIMK1, DDOST, SH3BP2, CD2BP2, EAF2, CALM3, MZF1, SEL1L, PSEN2, PPP1R9A, NCBP1, ACOT7, DYNC1LI1, S100A13, MAD2L1BP, IGFBP5, FECH, NIPSNAP1, RAD52, CROCC, URI1, SRSF6, POLA1, KHK, ARHGEF3, ARHGEF1, NCSTN, SMG5, UBTF, COCH, TPT1, CSTF3, NOL9, PDP1, SND1, HACD3, HACD2, DBN1, RBM25*
	GO:0044822 poly(A) RNA binding	*EIF2AK2, UPF1, TCOF1, SRSF5, UBE2L3, DHX32, SUGP1, NCBP2, DDX54, RALY, RPL22, AHNAK, MKI67, PWP2, TNS1, FXR1, MAP4, MARK2*	*DNTTIP2, SND1, SAMD4B, SRSF6, MAK16, EIF4B, SUB1, ARHGEF1, RBM25, NCBP1, DDX10, DYNC1LI1, NOLC1, YTHDF2,UBTF, SNRNP40, TPT1, EZR, CSTF3, PUS7, HSPA8, YBX3, DCN*
	GO:0005524 ATP binding	*EIF2AK2, UPF1, IP6K1, PCCA, PRKAA1, SRR, UBE2L3, TRIB2, DHX32, WNK1, CAMKK2, PDPK1, DDX54, ACVR1B, AURKA, DGKA, RFC5, MKI67, SLC22A4, STK4, MARK2, MARK1, ACSF2, HSPA4*	*GK, ETNK1, TTLL1, UBE2D2, STK19, BLK, PIK3CD, UBE2J1, KHK, GAK, UBE2B, DDX10, DYNC1LI1, GNE, NOLC1, CDK18, CDC6, EHD1, EHD2, KIFC1, AK1, MAP2K6, MCM7, NOL9, HSPA8, MAPK9, LIMK1, ATP13A3*

Tab. 5.10: Summary of hub-genes detected by CBDCEM in bulk RNA-Seq dataset, GS130078 annotated to top GO terms in the three GO databases

	GO Term	Hub Genes
GO_BP	GO:0045893 positive regulation of transcription DNA-templated	*EDF1, TBX1, BANP, TFE3, IRAK1*
	GO:0006357 regulation of transcription from RNA polymerase II promoter	*TBX1, CUX1*
	GO:0007507 heart development	*JMJD6, ADIPOR2, PDLIM3, TBX1*
	GO:0006366 transcription from RNA polymerase II promoter	*SALL4, CEBPD, CSRNP2, NFIL3, KLF11*
	GO:0032496 response to lipopolysaccharide	*IRAK1, REN, CXCL2*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*TBX1, TNFSF11, CTR9, CSRNP2, IL18, SALL4, CEBPD, PML, TFE3, IRF2BPL*
	GO:0055085 transmembrane transport	*SLC12A8*
	GO:0043065 positive regulation of apoptotic process	*KLF11*
	GO:0006810 transport	*SLC27A4, AP4M1*
	GO:0007165 signal transduction	*SH2B2, EBP, IRAK1, PDAP1, CCL26, SALL4, ANGPT2, ANTXR1, TENM1, GDI1, HGS, CD300C, PTCH2, ARHGAP33*
GO_CC	GO:0005737 cytoplasm	*TNFSF11, MAP3K1, EDF1, STXBP1, TENM1, UBXN1, OAS2, DNPH1, PPP4C, DNAH17, GDI1, KYNU, ACTN4, TBC1D24, IRAK1, UBE2S, PCDHB4, SALL4, GOLGA8B, PLEKHG6, PML, NANOS1, HGS, PSMD3, BIRC5, SH2B2, PRNP, PDLIM3*
	GO:0005829 cytosol	*STX5, AP4M1, SRM, MAP3K1, WIPF2, RRP1B, DNAJB5, STXBP1, UBXN1, OAS2, ARF1, TUBA1A, ARHGEF28, LAMTOR5, JMJD6, GDI1, SGO2, KYNU, JOSD2, IL18, IRAK1, SYT7, EHD2, TCTN1, PML, SORBS3, HGS, PSMD3, MYL5, BIRC5, SH2B2, ALDH8A1, TXLNG, SPATA33, DDHD1, IMPDH2, CUX1, CMPK2, ARHGAP33, SLC2A4, CAMK2A, USP18*
	GO:0016020 membrane	*STX5, SLC27A4, TNFSF11, SDHC, SYT7, TCTN1, CSGALNACT2, KRT1, CD40LG, OAS2, PSMD3, IMPDH2, SIRPA, SLC2A4, CHTF18*
GO_MF	GO:0005515 protein binding	*EXOC3-AS1, STX5, SRM, MAP3K1, TXN2, CEBPD, AVPI1, SCAMP2, ARF1, TUBA1A, GDI1, SGO2, CIZ1, IL18, NFIL3, SYT7, AGPAT4, SALL4, PLEKHG6, KRT1, SORBS3, NANOS1, HGS, HLX, SH2B2, PRNP, PDLIM3, ANTXR1, ARHGAP33, FAM118A, AP4M1*
	GO:0046872 metal ion binding	*ZNF474, FBXO43, NEK11, SDHC, SALL4, CSGALNACT2, ZNF316, ZNF75A, ANGPT2, PDP1, HGS, IRF2BPL, OAS2, BIRC5, RSAD2, ARHGEF28, PRNP, ATP2C2, PPP4C, TESMIN, DDHD1, IMPDH2, ANTXR1, PARP12, ADIPOR2, KLF11, CAMK2A*
	GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding	*TFE3, CEBPD*

5.6.3 Literature Trace

As a final step, we validated the detected hub-genes based on existing literature, establishing that these hub-genes may serve as potential biomarkers for six types of SCCs which are ESCC, HNSCC, LaSCC, LSCC, OSCC and TSCC. Based on CBDCEM analysis and existing literature that correlated these hub-genes with the six previously mentioned SCCs, Table 5.11 summarizes the hub-genes detected by CBDCEM and existing literature.

- According to Loomans et al. [447], Activin A's suppression of ESCC development depends on ACVR1B. Loomans et al. [446] further emphasize how the absence of ARCVIB can cause Squamous Cell Carcinoma in general to become more aggressive.
- Gao et al. [183] note that annexin A3 (ANXA3) reduction greatly reduces ESCC cell proliferation and propose it as a possible biomarker.
- According to Shang et al. [618], down-regulation of Baculoviral IAP Repeat Containing 5 (BIRC5) is observed to prevent both migration and invasion in ESCC.
- Hu et al. [247] provide CD163 as a marker of M2 macrophage, helping to predict the aggressiveness and prognosis of Kazakh esophageal squamous cell carcinoma.
- Diacylglycerol kinase α (DGKA) is crucially involved in the progression of ESCC, according to Chen et al. [76], who suggest DGKA as a viable target for ESCC treatment.
- The work of Wong et al. [767] and Ma et al. [469] suggest that Dual-specificity phosphatase 6 (DUSP6) plays a role in the metastasis, carcinogenesis of ESCC.
- Li et al. [358] suggest EH domain-containing protein 2 (EHD2) as a promising independent prognostic biomarker for ESCC.
- According to Gao et al. [179], Forkhead Box A2 (FOXA2) is crucial to the development of ESCC.
- Overexpression of heat-shock factor 1 (HSF1) is a biomarker for ESCC as suggested by Tsukao et. al [694].
- Imai et al. [266] emphasize the critical function of Kinesin Family Member C1 (KIFC1) in the carcinogenesis of ESCC.
- Microtubule-associated protein 4 (MAP4) has been identified by Jiang et. al [285] as a key regulator of cell invasion and migration in ESCC, making it a potential prognostic biomarker.

- Mitogen-activated protein kinase 9 (MAPK9) is down-regulated in ESCC, which Song et al. [641] theorizes promotes carcinogenesis.
- The results presented by Qiu et. al [563] show that mini-chromosome maintenance complex component 7 (MCM7) activates the AKT1/mTOR signalling pathway, promoting colony formation, migration, and tumour cell proliferation in ESCC cells. Recommendations from Choy et. al [102] and Zhong et. al [906] pointed to MCM7 as a biomarker for ESCC.
- The biological significance of NADPH oxidase 5 (NOX5) in the emergence of ESCC is discussed by Chen et. al [75].
- Profilin-2 (PFN2), which Cui et. al [114] show has a novel role in increasing ESCC progression and also present as a biomarker of high-risk population.
- Since its down-regulation inhibits the growth of ESCC, Gao et al. [181] suggests that Piezo Type Mechanosensitive Ion Channel Component 1 (PIEZO1) presents as a new therapeutic target for ESCC.
- In patients with ESCC, Yen et al. [825] demonstrate that the promyelocytic leukaemia gene (PML) protein serves as an independent prognostic predictor.
- Peroxiredoxin 6 (PRDX6) overexpression is shown by He et al. [231] to contribute to the development of ESCC via Erk1/2.
- According to the study presented by Granelli et al. [197], SEL1L aids in identifying patients who are at a high risk of developing ESCC.
- According to Li et al. [346], sex-determining region Y box 12 (SOX12) promotes the JAK2/STAT3 signalling pathway, which improves the motility of ESCC cells.
- Cell proliferation, migration, and invasion are all significantly reduced when Stathmin 1 (STMN1) is down-regulated, according to Ma et al. [467], but these processes are increased when STMN1 is up-regulated.
- According to Sheyhidin et al. [629], Toll-like receptor (TLR) 3 is a potential target for the treatment of ESCC..
- Zhang et al. [852] introduce Tetraspanin 15 (TSPAN15) as a new therapeutic biomarker.
- Ubiquitin-specific protease 39 (USP39), which Zhao et al. [899] identifies as an oncogenic factor in ESCC.
- According to Jiang et al. [284], restoring the function of VGLL4 may be a promising therapeutic approach for treating ESCC. VGLL4 down-regulation is thought to be crucial in the development of ESCC.

Tab. 5.11: Summary of potential biomarkers identified by CBDCEM. Here, All 3 under GO databases imply all three databases, BP, CC, and MF.

	Hub-Gene	GO Database	Enriched Pathway(s)	TF ?	ESCC Literature Evidence
GSE130078	ANGPT2	BP,MF	hsa04151,hsa04015, and hsa04014	No	OSCC [295], LSCC [561]
	BIRC5	All 3	hsa05200	No	ESCC [618], OSCC [690]
	IL-18	BP,CC	None	No	HNSCC [581], OSCC [382]
	JMJD6	BP,MF	None	No	HNSCC [204] , OSCC[334]
	NOX5	All 3	None	No	ESCC [75], OSCC[261], HNSCC [80]
	PIEZO1	CC	None	No	ESCC [181], OSCC [223]
	PML	All 3	hsa05200	Yes (Fig 5.7d)	ESCC [825]
GSE20347	ACVR1B	All 3	hsa04060, hsa04550, and hsa04350	No	ESCC [447, 446]
	AURKA	All 3	hsa04512	No	OSCC [123], HNSCC [770]
	BASP1	All 3	None	No	HNSCC [568], TSCC [379]
	CD163	BP,CC	None	No	ESCC [247] , HNSCC[689], OSCC [229, 735]
	DGKA	All 3	hsa05231, hsa04070, and hsa01100	No	ESCC [76]
	DUSP6	All 3	hsa04010	No	ESCC [469, 767]
	EPOR	All 3	hsa04151, hsa04640, and hsa04060	No	OSCC [406]
	FMO2	All 3	None	No	OSCC [163]
	HMMR	BP,CC	hsa04512	No	HNSCC [455]
	HOXA10	All 3	hsa05202	Yes (Fig 5.6b)	OSCC [67, 800], LaSCC [353]
	HSF1	All 3	hsa05231, hsa04070, and hsa01100	Yes (Fig 5.6c)	ESCC [694, 396]
	MAP4	All 3	hsa04010	No	ESCC [285]
	PFN2	All 3	hsa04010, hsa04810, and hsa05131	No	ESCC [114], OSCC [465]
	PHLDA1	BP,CC	-	No	OSCC [106]
	SDC1	All 3	hsa05205,hsa04512,hsa04514, and hsa05144	No	OSCC [768, 742]
	SPP1	All 3	hsa04151, hsa04510, hsa04512, and hsa04620	No	TSCC [794], LSCC [760]
	SRSF5	All 3	hsa05168, and hsa03040	Yes (Fig 5.6a)	OSCC [816]
	TSPAN15	All 3	None	No	ESCC [852]
	USP7	All 3	hsa04010, hsa05168, hsa05169, and hsa04068	Yes (Fig 5.6b)	LSCC [891], LaSCC [870]
	ACOT7	CC,MF	None	No	HNSCC [289]
	ANXA3	All 3	-	No	ESCC [183]

Hub-Gene	GO Database	Enriched Pathway(s)	TF ?	ESCC Literature Evidence
ATP13A3	All 3	None	No	HNSCC [343]
CALM3	All 3	hsa05152, hsa04015, hsa04020, hsa04722, and hsa04910	No	OSCC [322]
CD38	All 3	None	No	OSCC [140], HNSCC [869]
CDC6	All 3	hsa05200, hsa04110, and hsa05216	Yes (Fig 5.7a)	OSCC [157], TSCC [158]
CDC73	All 3	None	No	OSCC [573]
EHD2	All 3	hsa04144	No	ESCC [358]
FGF8	All 3	hsa05200, hsa04151, hsa04010, hsa04015, hsa04810, hsa004014, and hsa05218	No	OSCC [218]
FOXA2	All 3	None	Yes (Fig 5.6d)	ESCC [179]
GLI3	All 3	hsa05200, and hsa04024	Yes (Fig 5.7c)	OSCC [586]
GNG7	All 3	hsa05200, hsa04151, hsa04014, hsa04062, hsa04728, hsa04725,hsa05032 and 5 others	No	HNSCC [221, 130]
HIC1	All 3	None	No	ESCC [360], HNSCC [56]
IGFBP5	All 3	None	No	HNSCC [263]
ITGB4	All 3	hsa04151, hsa04510, hsa04810, hsa04512, hsa05410, hsa05414, and hsa05412	No	OSCC [517], HNSCC [347]
KIFC1	All 3	hsa03040	No	ESCC [266]
MCFD2	All 3	None	No	OSCC [173]
MCM7	All 3	hsa04520, and hsa03030	Yes (Fig 5.7a)	ESCC [563, 102, 10, 906], OSCC [157]
MAPK9	All 3	hsa05200, hsa04510, hsa04010, hsa05161, hsa05164, hsa04380,hsa05152, and 2 others	No	ESCC [641]
MMP15	All 3	None	No	LaSCC [51]
MZF1	BP,MF	None	Yes (Fig 5.7b)	OSCC [313]
PRDX4	All 3	None	No	OSCC [71], LSCC [265]
PRDX6	All 3	hsa01100	No	ESCC [231]
RAD52	All 3	None	No	LSCC [397]
SEL1L	BP,CC	hsa04141	No	ESCC [197]
SOX12	All 3	None	Yes (Fig 5.7c)	ESCC [346]
STMN1	All 3	hsa04010	No	ESCC [521], OSCC [467]
USP39	All 3	None	Yes (Fig 5.7a)	ESCC [899], OSCC [355]
VGLL4	None	None	Yes (Fig 5.6d)	ESCC [284]
YTHDF2	All 3	None	No	LSCC [790]

Table 5.11 gives a summary for all hub-genes detected by CBDCEM that have literature evidence as potential biomarkers for six SCCs as mentioned earlier.

Tab. 5.12: Summary of potential ESCC biomarkers identified by CBDCEM using the biomarker criteria (Section 2.5).

	GSE20347	GSE23400	GSE130078
Case 1	HSF1	MCM7	PML
Case 2	DGKA, MAP4, PFN2, DUSP6, ACVR1B	PRDX6, MAPK9, SEL1L, EHD2, KIFC1, STMN1	BIRC5
Case 3	TSPAN15	ANXA3, HIC1, SOX12, FOXA2, USP39	NOX5
Case 4	USP7, HOXA10, SRSF5	GLI3, CDC6, GNG7	

There exists strong literature evidence that associate all BCGs that falls under Cases 1 and 2 to ESCC. In all three datasets, they have also been annotated to highly enriched GO terms and enriched pathways. All hub-genes that fall under Cases 1 and 2 are potential biomarkers as discussed in the biomarker criteria (Section 2.5). As shown in Table 5.12, twelve BCGs including DGKA, MAP4, PFN2, DUSP6, ACVR1B, PRDX6, MAPK9, SEL1L, EHD2, KIFC1, STMN1, and BIRC5 are potential biomarkers for ESCC as they fall under Case 2. Case 1 includes three BCGs, HSF1, MCM7, and PML, which are also TFs (thus they fall under Case 1).

TSPAN15 in GSE20347, NOX5 in GSE130078, and 5 hub genes namely, ANXA3, HIC1, SOX12, FOXA2, and USP39 in GSE23400 fall under Case 3. Thus , while there exists literature evidence that tie these seven BCGs to ESCC, further in-depth analysis is necessary to establish them as potential biomarkers. Similar to Case 3, Case 4 pertains to three BCGs, USP7, HOXA10, and SRSF5)in GSE20347 and three BCGs GLI3, CDC6, and GNG7 in GSE23400. USP7, HOXA10, SRSF5, GLI3, GNG7 and CDC6 are five BCGs that have been identified as possible biomarkers for the other five SCC (but not ESCC). All of these hub-genes are TFs, indicating their regulatory function in the network. Additionally, they are linked to highly enriched pathways and GO terms in at least two out of three GO databases, demonstrating their biological relevance.

Although 15 BCGs detected by CBDCEM, namely JMJD6, IL18, PHLDA1, BASP1, FMO2, PRDX4, MMP15, ACOT7, CD38, IGTFBP5, MCFD2, RAD52, ATP13A3, YTHDF2 and CDC73 are potential biomarkers for five other previously mentioned SCC, they have only been annotated to enriched GO terms in one or more GO databases and are neither TFs nor annotated to any enriched pathway. As a result, they lack sufficient biological or literary support to be taken into consideration as potential biomarkers for ESCC. The two BCGs KRT75 and MINCR, which are biomarkers for other SCC but have not been linked to any GO terms or pathways, and cannot be counted as a potential ESCC biomarker.

Finally, it can be said that validation has confirmed the potential biomarker status of 15 hub-genes identified by CBDCEM, including HSF1, MCM7, PML, DGKA, MAP4, PFN2, DUSP6, ACVR1B, PRDX6, MAPK9, SEL1L, EHD2, KIFC1, STMN1, and BIRC5. Additionally, 18 additional hub-genes identified by CBDCEM, including TSPAN15, ANXA3, HIC1, SOX12, FOXA2, USP39, NOX5, USP7, HOXA10, SRSF5, HMMR, SDC1, SPP1, GLI3, CDC6, GNG7, CALM3, and ITGB4, have weak support for their potential as ESCC biomarkers and need further biological investigation.

5.7 Discussion

Our research is compared to four other hub-gene discovery techniques. Hub-genes can be located using two of the simplest and most popular techniques: Genes with the highest degrees within the module and those with the highest intra-modular connectivity [327]. Additionally, Das et al. in their publication DHGA [120] offered two approaches for locating hub genes, a) Weighted Gene Score and b) p-value Cut Off, both of which are extensively applied. For the purpose of hub-gene discovery, we give a straightforward comparison between CBDCEM and these four approaches. Incorporating CBDCEM into a comparison with these four hub-gene discovery techniques is inappropriate. Therefore, we compare these approaches to our hub-gene discovery algorithm utilizing the procedures listed below:

1. Pre-processing to module discovery still follows the same pipeline.
2. We use the four techniques simultaneously for each module taken into account by CBDCEM, and we compile a list of the 20 hub-genes found by each module.
3. We discover the hub-genes from these hub-genes that have literary evidence of relationship with the following types of squamous cell carcinoma (SCC): Esophageal, oral, laryngeal, lung, tongue, and head and neck SCCs.

Table 5.13 summarizes in detail, the comparison between CBDCEM and four other previously mentioned methods.

Tab. 5.13: Summary of potential biomarkers detected by CBDCEM, WGS: DHGA [120] Weighted Gene Score, PCO:DHGA [120] p-value Cut Off, IMC: WGCNA [327] Intra-modular Connectivity and Degree with strong literature evidence of relation to ESCC(marked in Red), HNSCC, LaSCC, LSCC, OSCC, TSCC

	Module	CBDCEM	WGS	PCO	IMC	Degree
GSE10078	*lightcyan*	None	None	None	None	None
	salmon	JMJD6[204, 334], BIRC5[618, 690], PML [825], IL18[581, 382]	CAV1[26]	CAV1[26, 301], CAV2[26]	CAV1[26, 301], CAV2[26], KPNA2[475, 403]	JMJD6[204, 334], BIRC5[618, 690], IL18[581, 382]
	orange	MINCR[463]	CDK4[255], GIHCG[473], DUSP2[522], MAGEB2[543]	MCM7[563, 906, 10, 157], ORC5 [649], PSMC2 [649, 754], POP7 [819]	HMGB3 [180]	MINCR[463]
	purple	KRT75[526], NOX5[75, 261, 80]	CDH23[601]	CDH23[601]	CDH23[601], ADAMTS16[597]	KRT75[526], NOX5[75, 261, 80]
	darkgrey	ANGPT2[295, 561], PIEZO1[181, 223]	-	DVL3[87, 288], PPFIA1[861, 50]	MDC1[121, 880], PPFIA1[861, 50]	SYT7[172], LAMTOR5[814]
	bisque4	-	PAX5[323], BTLA[64]	PAX5[323], BTLA[64]	IGFBPL1[436], IGFBP5[263]	-
	navajowhite2	-	MIR145[718, 620], SCN7A[842], TWIST1[340], CCN5[656]	SCN7A[842], BCHE[280]	SCN7A[842], MIR145[718, 620], TGFBR3[882],	-
GSE20347	*darkturquoise*	USP7[891, 870], HOXA10[67, 800, 353], PHLDA1[106]	-	TEAD4[875], PXN[356], RGS5[4]	PRMT1[913, 900], RUVBL1[402], RGS5[4]	PHLDA1[106], USP7[891, 870], HOXA10[67, 800, 353]
	lightgreen	BASP1[568, 379]	UAP1L1[782]	METTL3[85, 12]	UAP1L1[782]	BASP1[568, 379]
	darkorange	-	-	-	WRAP53[570]	-
	orange	DGKA[76], HSF1[694, 396]	HES1[663], TRIM29[793]	PRDX3[863]	PRDX3[863], HES1[663]	DGKA[76], HSF1[694, 396]
	skyblue	MAP4[285], PFN2[114, 465]			SHMT2[394, 393]	PFN2[114, 465], MAP4[285]
	floralwhite	CD163[247, 689, 229, 735], AURKA[123, 770], DUSP6[469, 767], HOXA10[67, 800, 353]	NFAT5[831]	-	PTMA[920]	CD163[247, 689, 229, 735], HOXA10[67, 800, 353], AURKA[123, 770]
	greenyellow	HMMR[455], FMO2[163], TSPAN15[852], EPOR[406, 349]	RAB2A[134], MYO6[881], CEACAM6[40, 98]	RAB2A[134], MYO6[881]	GSTO1[377], CEACAM6[40, 98], RAB2A[134], MYO6[881]	TSPAN15, HMMR[455], EPOR[406, 349], FMO2[163]

Continued on next page

	Module	CBDCEM	WGS	PCO	IMC	Degree
	paleturquoise	SDC1[768, 742], SRSF5[816], ACVR1B[447, 446], SPP1[794, 760]				ACVR1B[447, 446], SRSF5[816], SDC1[768, 742], SPP1[794, 760]
GSE23400	*green*	CALM3[322], PRDX6[231], PRDX4[71, 265], MMP15[51], SOX12[346], GLI3[586], FGF8[218]	AKR1C2[888, 366], GPX2[342], G6PD[745]	GPX2[342], ITGA3[156, 517]	AKR1C2[888, 366], GPX2[342], ITGA3[156, 517]	CALM3[322], PRDX6[231], PRDX4[71, 265]
	magenta	ACOT7[289]	MCM4[102], ATAD2[63, 741], SLC39A4[780, 287], EIF3E[788]	EIF3E[788], MCM4[102], ATAD2[63, 741], HSF1[694, 396]	MCM4[102], ATAD2[63, 741], STX12[213]	ACOT7[289]
	grey60	VGLL4[284], MAPK9[641], FOXA2[179], CD38[140, 869], IGFBP5[263]	HLA-G[267, 54, 605], IFIT3[552]	IFI144L[535], IFIT3[552], HLA-G[267, 54, 605]	HLA-G[267, 54, 605] , IFIT3[552]	MAPK9[641], FOXA2[179], CD38[140, 869], IGFBP5[263]
	midnightblue	MCFD2[173], SEL1L[197], EHD2[358], RAD52[397], ITGB4[517, 347]	CD44[83, 149]	CD44[83, 149]	CD44[83, 149], MYO1B[530], ITGB4[517, 347]	SEL1L[197], EHD2[358], RAD52[397]
	lightcyan	KIFC1[266], USP39[899, 355], ATP13A3[343], CDC6[157, 158],STMN1[521, 467], MCM7[563, 906, 10, 157]	-	KEAP1[133]	CDC6[157, 158], DNMT1[896, 112]	KIFC1[266], USP39[899, 355], ATP13A3[343], STMN1[521, 467], CDC6[157, 158], MCM7[563, 906, 10, 157]
	tan	ANXA3[183], MZF1[313]	SULTIA1[617, 119], SULT1A2[617]	SULTIA1[617, 119]	SULTIA1[617, 119], SULT1A2[617]	ANXA3[183], MZF1[313]
	salmon	GNG7[221, 130], YTHDF2[790], CDC73[573], HIC1[360, 56]	-	-	-	GNG7[221, 130], YTHDF2[790], CDC73[573], HIC1[360, 56]

We compare CBDCEM's performance to each of the four methods using two parameters: a) quantity (which measures how many potential biomarkers have been identified by a method for the six categories of SCC already mentioned) and b) quality (which measures how many potential biomarkers have been identified by a method for ESCC specifically). It is advantageous if a method performs well overall across the board for both parameters.

5.7.1 Comparison of with WGS, PCO, IMC and Degree

5.7.2 CBDCEM vs. WGS

The DHGA Weigted Gene Score (WGS) [120] identifies a list of hub genes in a co-expressed gene network based on Weighted Gene Score and does not use statistical significant values.

As can be observed in Table 5.13, CBDCEM outperforms WGS for the majority of modules. In ten modules, *salmon* (GSE130078), *darkgrey* (GSE130078), *skyblue* (GSE20347), *floralwhite* (GSE20347), *greenyellow* (GSE20347), *paleturquoise* (GSE20347), *grey60* (GSE23400), *midnightblue* (GSE23400), *lightcyan* (GSE23400), and *salmon* (GSE23400), CBDCEM performs better than WGS in terms of both quality and quantity. Furthermore, both CBDCEM and WGS find two possible biomarkers in the *orange* module (GSE20347). CBDCEM has discovered two putative ESCC biomarkers, DGKA and HSF1, although WGS can only identify HES1. On the other hand, both WGS and CBDCEM identify two hub-genes that could serve as biomarkers in the *tan* module (GSE23400). However, only one of the two possible biomarkers identified by CBDCEM (ANXA3) and one of the two identified by WGS are connected to ESCC. In the *purple* module (GSE130078), CBDCEM and WGS find two and one possible biomarkers, respectively; nevertheless, both detect one biomarker associated with ESCC. WGS has outperformed CBDCEM in terms of both quality and quantity in four modules: *orange* (GSE130078), *bisque4* (GSE130078), *navajowhite2* (GSE130078), and *magenta* (GSE23400). In the *green* module (GSE23400), CBDCEM outperforms WGS in terms of quantity, detecting seven possible biomarkers as opposed to three. However, only two of the seven hub-genes found by CBDCEM, PRDX6, and SOX12 are possible ESCC biomarkers, whereas all three potential biomarkers detected by WGS (AKR1C2, GPX2, and G6PD) exhibit evidence of correlation with ESCC. While WGS is unable to identify any prospective biomarker in the *darkturquoise* module

(GSE20347), CBDCEM finds three potential biomarkers, none of which are associated with ESCC.

5.7.3 CBDCEM vs. PCO

p-value for DHGA Based on gene connection significance values, Cut-off (PCO) [120] finds a list of hub genes in a co-expressed gene network.

As can be observed in Table 5.13, CBDCEM outperforms PCO for the majority of modules. CBDCEM performs better than PCO in terms of both quality and quantity in ten modules, *orange* (GSE20347), *skyblue* (GSE20347), *floralwhite* (GSE20347), *greenyellow* (GSE20347), *paleturquoise* (GSE20347), *green* (GSE23400), *grey60* (GSE23400), *midnightblue* (GSE23400), *lightcyan* (GSE23400) and *salmon* (GSE23400). In the module *salmon* (GSE130078), CBDCEM outperforms PCO in terms of quantity since it identifies one more potential biomarker; however, in terms of quality, CBDCEM and PCO are equal because each discovers two potential biomarkers for ESCC. In the *tan* module (GSE23400), CBDCEM outperforms PCO in terms of quantity, detecting just two possible biomarkers as opposed to three for PCO. However, in terms of quality, CBDCEM and PCO are comparable because both detect a putative ESCC biomarker. In module *purple* (GSE130078), as seen in the case of WGS, CBDCEM and PCO find two and one possible biomarkers, respectively, but only one of them is associated with ESCC. Similar to WGS, PCO has outperformed CBDCEM in terms of both quality and quantity in four modules: *orange* (GSE130078), *bisque4* (GSE130078), *navajowhite2* (GSE130078), and *magenta* (GSE23400). CBDCEM and PCO each identify four and three possible biomarkers in the module *darkturquoise* (GSE20347), but none of them are connected to ESCC. There is no evidence of a connection between any of the only putative biomarkers identified by CBDCEM and PCO in module *lightgreen* (GSE20347) and ESCC.

5.7.4 CBDCEM Vs IMC

With the assumption that a highly connected node has larger flow of relevant information through it, genes with high intra-modular connectivity can be considered hub-genes. WGNA [327] intra-modular connectivity (IMC) calculates connectivity of a node to other nodes in the same module.

As seen in Table 5.13, there are different scenarios when CBDCEM and IMC are compared, unlike in previous cases where there are clear indications that CBDCEM performing better than WGS and PCO. CBDCEM outperforms IMC in terms of both quality and quantity in seven modules: *skyblue* (GSE20347), *floralwhite* (GSE20347), *paleturquoise* (GSE20347), *grey60* (GSE23400), *midnightblue* (GSE23400), *lightcyan* (GSE23400), and *salmon* (GSE23400). It can be said that the performance of CBDCEM and IMC in three modules, *salmon* (GSE130078), *dark grey* (GSE130078), and *greenyellow* (GSE20347), is comparable in terms of both quality and quantity. Four possible biomarkers are found by CBDCEM and IMC in the module *salmon* (GSE130078), with PML and BIRC5 being detected by the former while CAV1 and CAV2 being detected by the latter. Out of the two potential biomarkers found in *darkgrey* (GSE130078), PIEZO1 and PPFIA1 were found by CBDCEM and IMC, respectively, and have indications of correlation with ESCC. In the module *green* (GSE23400), CBDCEM outperforms IMC in terms of quantity, detecting seven and two possible biomarkers, respectively. However, both are equivalent in terms of quality as both CBDCEM and IMC found two potential ESCC biomarkers. One putative biomarker in *orange* (GSE130078), is detectable by both CBDCEM and IMC, however HMGB3, found by the latter, may be a potential ESCC biomarker. Both CBDCEM and IMC identify two potential biomarkers in the *purple* module (GSE130078), but only one of the two is a potential ESCC biomarker according to CBDCEM, whilst both are in IMC. One of the four potential biomarkers identified by IMC has an association with ESCC, however none of the four potential biomarkers identified by CBDCEM in the module *darkturquoise* (GSE130078) do. Similar to this, in *lightgreen* (GSE20347), CBDCEM and IMC found possible biomarkers, BASP1 and UAP1L1 respectively, of which only the latter has evidence of relation to ESCC. Similar to WGS, CBDCEM also identifies two hub-genes in the *tan* module (GSE23400), which have the potential to serve as biomarkers. However, only one of the two potential biomarkers identified by CBDCEM (ANXA3) and both identified by IMC (SULTIA1 and SULT1A2) are connected to ESCC. On the other hand, both CBDCEM and IMC identify two possible biomarkers in the *orange* module (GSE20347). CBDCEM has discovered two putative ESCC biomarkers, DGKA and HSF1, although WGS can only identify HES1.

5.7.5 CBDCEM Vs Degree

Genes with the most degrees can be regarded as hub-genes on the premise that nodes in a network with high degrees, which have a high volume of incoming and outgoing degrees, hold the most information. This is the simplest and most basic technique for finding hub-genes.

The list of identified hub-genes in the majority of modules is pretty comparable between CBDCEM and Degree. Particularly in smaller and sparser modules discovered by GSE20347, this is the case. As a result, the bulk of hub-genes that have the potential to serve as biomarkers are shared by the two approaches. However, there are circumstances in which CBDCEM outperforms Degree and is able to identify distinct hub-genes that are missed by Degree. Although there are no shared hub-genes in the module *darkgrey* (GSE130078), both CBDCEM and Degree identify two potential biomarkers, with CBDCEM providing higher-quality results because it identifies PIEZO1, a potential ESCC biomarker. A further possible biomarker for ESCC, PML, is discovered by CBDCEM in the *salmon* module (GSE130078). In the *floralwhite* module (GSE20347), CBDCEM detects the possible biomarker DUSP6 while Degree misses it. We have seen that in dataset GSE23400, CBDCEM generally outperforms Degree. A biomarker for ESCC, SOX12, and two for additional SCCs, MMP15, GLI3, and FGF8, which are not detected by Degree, are found in the module *green* (GSE23400). Similar to this, the possible ESCC biomarker VGLL4 is discovered by CBDCEM in the module *grey60* (GSE23400). The module *midnightblue* (GSE23400) contains two more biomarkers for different SCCs, MCDF2 and ITGB4.

We conclude from the experimental findings that CBDCEM works satisfactorily across the board for all datasets. The percentage of modules in which CBDCEM performs better, similar, or worse than WGS, PCO, IMC, and Degree when considering both parameters—Quantity and Quality—is summarized in Table 5.14. All modules of interest can be extracted by CBDCEM with significant levels of GO and pathway enrichment, proving their biological importance. In the majority of modules, CBDCEM is able to pinpoint at least one hub gene as a potential biomarker. Only the modules *lightcyan*, *bisque4*, *navajowhite2*, and *darkOrange* are an exception to this rule. CBDCEM's performance is satisfactory because it can identify a sizable number of hub-genes that could serve as potential biomakers for Squamous Cell Carcinoma in general. Many of these hub-genes also show evidence of a specific relationship with ESCC.

Tab. 5.14: Summary of performance of CBDCEM vs. other methods in terms of proportion of modules. We compare these methods on 8, 7 and 7 MoIs in GSE20347, GSE23400 and GSE130078. Quantity measures the number of potential biomarkers identified by a method for the six previously mentioned categories of SCC in general and Quality measures the number of potential biomarkers identified by a method for ESCC in particular

		CBDCEM>Method		CBDCEM≡Method		CBDCEM<Method	
	Dataset	Quantity (%)	Quality (%)	Quantity (%)	Quality (%)	Quantity (%)	Quality (%)
WGS	GSE20347	62.5	62.5	37.5	25	0	12.5
	GSE23400	57.1	42.9	28.6	0	14.3	57.1
	GSE130078	42.9	28.6	0	14.3	42.9	42.9
PCO	GSE20347	62.5	62.5	37.5	37.5	0	0
	GSE23400	71.4	57.1	14.3	28.6	14.3	14.3
	GSE130078	28.6	0	14.3	28.6	42.9	57.1
IMC	GSE20347	37.5	50	50	12.5	12.5	37.5
	GSE23400	57.1	42.9	28.6	14.3	14.3	42.9
	GSE130078	14.3	0	14.3	42.9	57.1	42.9
Degree	GSE20347	12.5	12.5	87.5	87.5	0	0
	GSE23400	28.6	14.3	71.4	85.7	0	0
	GSE130078	14.3	28.6	71.4	57.1	0	0

It is interesting, however, that CBDCEM has the capacity to identify possible biomarkers that are missed by other methods in several modules (Table. 5.13). With strong evidence of association with ESCC, CBDCEM detects PIEZO1, PML, DUSP6, VGLL4, and SOX12 in the modules *dark grey* (GSE130078), *salmon* (GSE20347), *floralwhite* (GSE20347), and *green* (GSE23400), but not by the other four methods. Additionally, the hub-genes ANGPT2, MMP15, GLI3, FGF8, and MCFD2, respectively, with strong evidence of association with five other SCCs (excluding ESCC) are detected by CBDCEM in the modules *darkgrey* (GSE130078), *green* (GSE23400), and *midnightblue* (GSE23400), but are not detected by the other four methods.

5.8 Chapter Summary

We demonstrate that our proposed differential expression analysis method, CBDCEM, performs effectively in terms of extracting differentially co-expressed modules and identifying hub genes. We have demonstrated that CBDCEM is capable of extracting relevant modules that GO enriched and pathway enriched from two microarray datasets , GSE20347 and GSE23400 and one bulk RNA-Seq dataset, GSE130078. We investigated the behavioral alterations among the DEGs in both normal and disease conditions using Differential Co-expression (DCE) analysis and preservation analysis. All reasonably sized non-preserved modules are considered as ‘Modules of Interest’ (MoI) and are further analyzed. CBDCEM identifies 22 MoIs across all three datasets. Using the proposed hub-gene finding method, CBDCEM identifies 20 hub-genes from each of the 20 MoIs. CBDCEM considers all these hub-genes as biomarker candidate genes (BCGs). CBDCEM performs biological analysis on each BCG and finds literature evidence that associates that BCG with either ESCC or five other SCCs that are associated with ESCC, namely head and neck SCC, larygeal SCC, lung SCC, oral SCC and tongue SCC. Three transcription factors (TFs) HSF1, MCM7 and PML fall under Case 1 of the biomarker criteria (Section 2.5) and are potential biomarkers for ESCC. Twelve BCGs DGKA, MAP4, PFN2, DUSP6, ACVR1B, PRDX6, MAPK9, SEL1L, EHD2, KIFC1, STMN1, BIRC5 fall under Case 2 of the biomarker criteria and thus are potential biomarkers for ESCC. Seven BCGs TSPAN15, ANXA3, HIC1, SOX12, FOXA2, USP39 and NOX5 fall under Case 3 and thus require further in depth analysis to establish them as potential biomarkers for ESCC. Similarly, six BCGs USP7 , HOXA10, SRSF5, GLI3, CDC6, and GNG7 fall under case 4 and thus are biologically relevant and have literature evidence that associate them to five SCCs related to ESCC. Thus, they are probable biomarkers for ESCC.

In most scenarios, CBDCEM performs satisfactorily, according to a comparison of CBDCEM with four other hub-gene methods. In addition, CBDCEM can identify ten unique potential biomarkers that the other four methods are unable to detect, five of which, namely PIEZO1, PML, DUSP6, VGLL4, and SOX12 have strong evidence of association with ESCC. It is observed that in the majority of MoIs detected in the two microarray datasets, GSE20347 and GSE23400, CBDCEM outperforms Weighted Gene Score (WGS) [120] and p-value cut-off (PCO) [120]. However, they perform at par or better than CBDCEM in the bulk RNA-Seq dataset, GSE130078. In seven modules, CBDCEM performs better than Intra-modular Connectivity (IMC) [327], while the latter performs better in three modules . Additionally, their performance is at par across three modules. CBDCEM performs at par with or better than IMC in terms of the quantity of potential biomarkers that are discovered. In eight modules, CBDCEM outperforms IMC, however, IMC outperforms CBDCEM when the number of potential ESCC biomarkers found is taken into account. It is also noteworthy that, in the majority of cases, there are no hub-genes shared by CBDCEM and Intra-modular Connectivity.

In the following chapter, a framework for DCA of single cell RNA-Seq (scRNA-Seq) data is presented. Through this, we aim to gain an insight into how intrinsic biological processes interact under various conditions (or states) provided by scRNA-Seq data. In handling scRNA-Seq data, we examine and address the issues and challenges that may arise. We employ a variation of our proposed hub-gene finding algorithm presented in this chapter and compare the same against four hub-gene finding methods.