Chapter 3

Biclustering Approach

3.1 Introduction

Gene expression data analysis techniques such as clustering and biclustering are used with the aim to find patterns or groupings. Clustering of genes identifies groups of co-expressed genes based on the expression levels they exhibit under all experimental conditions. In other words, two different genes that have similar experimental tendencies across samples (or conditions) tend to exhibit common patterns of regulation and thus reflects relations or interaction between their functions. A subset of co-expressed genes under certain experimental conditions may behave almost independently over another set of conditions. Thus, a two mode clustering approach known as biclustering [96] that clusters genes over a subset of conditions and operates in both dimensions simultaneously was introduced. While biclustering groups genes and features simultaneously thus identifying subsets that exhibit coherent patterns within particular context or setting, clustering primarily focuses on grouping similar genes together based on their feature similarity. Thus, biclustering is very useful for identifying context specific patterns as well as subsets of characteristics that behave consistently in different subsets of samples. While biclustering seeks to concurrently identify subgroups of samples and characteristics that exhibit comparable patterns, clustering does not take into account the context and seeks to identify homogeneous groups typically based on similarity of their features. Clustering does not always consider whether the similarity within the cluster is due to the same group of attributes.

Major limitations with clustering are mentioned below.

- Clustering results in creation of groups with no contextual or biological significance as clustering methods treat all features without consideration of the context or feature subsets under which they exhibit distinct patterns.
- When working with high dimensional data, noise or irrelevant features may impact

clustering outcomes as all features are treated similarly.

- Clustering is limited in it's capacity to recognize more intricate patterns that require connections between different feature subsets or characteristics as it operates on single dimension and primarily focuses on patterns exhibited by particular feature or variable.
- Specification of initialization parameters are often essential for clustering algorithms. However, different initialization parameter selection may lead to varying clustering results.

By explicitly taking context specific patterns into account, biclustering identifies subsets of samples and features that behave coherently. It makes feature selection easier and reduces dimensions by discovering feature subsets specifically associated with particular samples. Furthermore, biclustering has the ability to find subsets of attributes and samples that exhibit similar patterns making management of overlapping clusters easier. Like clustering, biclustering methods often require parameter setting but these parameters are often context specific and may not be sensitive to changes. Due to their participation in more than one function, genes tend to exhibit varying regulation patterns under varying conditions. In cases such as cellular processes that are active only under specific conditions or genes that participate in differentially regulated multiple pathways, traditional clustering often falls short. Co-regulated genes involved in specific cellular pathways or biological processes can be identified through biclustering analysis.

3.1.1 Biclustering Analysis

With the aim to identify subsets of genes that are co-expressed across a subset of samples or genes, the computational technique for gene expression data analysis, biclustering analysis is very useful. While traditional clustering methods identify groups of genes based on their expression profiles across all samples or conditions, biclustering simultaneously considers both genes and conditions to identify subsets of genes such that they exhibit similar expression patterns across a subset of samples or conditions. Some of the advantages of biclustering are noted below.

(a) By identifying sets of genes that exhibit similar expression patterns across a subset of samples and are co-regulated, biclustering can provide crucial insight into the underlying biological processes.

(b) By identifying subsets of samples that exhibit similar expression patterns, biclustering can provide crucial insight into the different disease states or phenotypes corresponding to the sample subtypes. This can further facilitate treatment planning and personalized medicine.

(c) By identifying subsets of genes and samples most relevant to the biological question of interest, biclustering can reduce the dimensionality of high-dimensional datasets thus making subsequent downstream analysis and interpretation efficient and effective.

(d) In genomic and transcriptomic data where noise and missing data is quite common, biclustering can help reduce false positives and improve result accuracy as they are often designed to be robust to them.

Throughout the years various biclustering methods have been developed that are applicable on different types of gene expression data, such as microarray and bulk RNA-Seq data. Several approaches to biclustering analysis, such as model based methods, matrix factorization methods, heuristic methods, employ different methods to identify biclusters. The choice of the biclustering method often hinges on the research question, availability of computational resources as well as the size and complexity of the data.

3.2 Related Works

In the literature, biclusters are represented in different ways depending on the genes being placed in the rows or the columns in the data matrix. Similarly, the same expression sub-matrix is also given different names.

Let, γ represents a bicluster that consists of a set of K of $||K||$ genes and a set of L of $||L||$ conditions, such that, the expression levels of gene k is represented by b_{kl} under sample l. Thus γ can be represented as follows[553]:

$$\gamma = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1|L|} \\ b_{21} & b_{22} & \dots & b_{2|L|} \\ \vdots & \vdots & \ddots & \vdots \\ b_{|K|1} & b_{|K|2} & \dots & b_{|K||L|} \end{pmatrix} \tag{3.1}$$

where, the gene g_{k} is the k^{th} row, i.e., $g_{\mathrm{k}} = b_{k1}; b_{k2}, \dots, b_{k|L|}$, and the condition c_{l} is the l^{th} column, i.e., $c_{\mathrm{l}} = b_{1l}; b_{2l}, \dots, b_{|K|l}$. To define evaluation measures, means of genes and samples in biclusters are used frequently.These values are represented as $b_{k|L|}$ and

$b_{|K|l}$ referring to the k^{th} row or gene and l^{th} column or sample means, respectively. Furthermore, the mean of all the expression values in γ is referred to as $b_{|K||L|}$.

According to Pontes et al. [553], biclustering algorithms can be grouped as: (i) algorithms based on evaluation measures that form biclusters according to the type of meta-heuristics used and (ii) non-metric biclustering that forms biclusters according to most distinctive properties. We present a taxonomy of various biclustering approaches in Fig.3.1, followed by a discussion on each of these approaches.

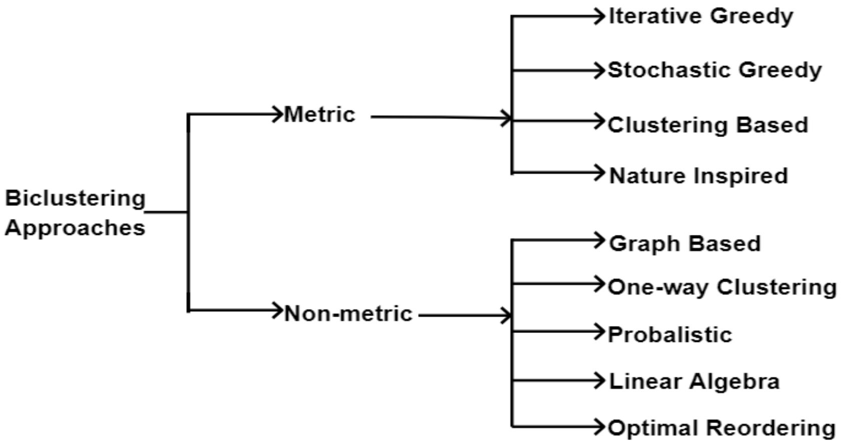


Fig. 3.1: Biclustering Approaches: A Taxonomy

3.2.1 Metric-based Biclustering Approach

Several biclustering strategies, including iterative greedy, stochastic greedy, nature inspired, and clustering based, have evolved in response to a metric-based approach.

3.2.1.1 Iterative Greedy Search

By building a group of objects from their simplest components, either iteratively or recursively, and making tactical local optimal decisions at each step, this method comes close to the ideal global answer.

Direct Clustering (DC) on data matrix was one of the earliest works on biclustering to be published. It was proposed by Hartigan et al.[220]. However, it was never utilized with genetic information. This technique uses a divide and conquer approach to produce k matrices by iteratively dividing the input matrix into a subset of sub-matrices, taking into consideration the desired number of biclusters as an input parameter k. During the partitioning process, variance is used as an evaluation metric. The variance of a bicluster

β is calculated as using Equation 3.2.

$$VAR(\beta) = \sum_{i-1}^{I} \sum_{j-1}^{J} (b_{\mathrm{ij}} - b_{\mathrm{IJ}}) \tag{3.2}$$

where b_{ij} and b_{IJ} represent the element in the i^{th} row and j^{th} column, and the mean of β, respectively. The algorithm choses the rows and columns that improve the overall partition variance at each iteration thus leading the algorithm towards constant biclusters. The characteristics of the search allows overlapping among biclusters.

The most important piece of research in this category was Mean Squared Residue (MSR) proposed by Cheng et al. [96]. Cheng and Church [96] were the first to apply biclustering to gene expression data. Cheng and Church (CC) [96] is a deterministic greedy algorithm that finds biclusters with minimal variance. This work attempts to find n biclusters in an expression data matrix using sequential cover. The most significant contribution of this work is the formulation of Mean Squared Residue (MSR), a measure to assess the quality of a bicluster of expression data. As the name suggests, MSR uses the means of gene and condition expression values to evaluate the coherence of the genes and conditions in a bicluster α and is given by Equation 3.3

$$MSR(\alpha) = \frac{1}{|I|.|J|} \sum_{i=1}^{i-|I|} \sum_{j=1}^{j-|J|} (a_{\mathrm{ij}} - a_{\mathrm{iJ}} - a_{\mathrm{Ij}} + a_{\mathrm{IJ}})^2 \tag{3.3}$$

where a_{ij}, a_{iJ}, a_{Ij} and a_{IJ} represent the element in the i^{th} row (condition) and j^{th} column (gene), the row and column means, and the mean of α, respectively. MSR is widely considered to be the earliest quality metric to be defined for biclusters of expression data [553]. However, because MSR is unable to capture shifting tendencies many approaches have incorporated modified versions of MSR. Despite the fact that MSR has been shown to be ineffective because it can only capture shifting patterns, many authors continue to utilize it. A search heuristic called SMSR-based biclustering (SMSR-CC) by Mukhopadhyay et al. [513] incorporates their evaluation measure SMSR (Scaling MSR), a variant of MSR. This is how SMSR is described in Equation 3.2.

$$SMSR(\beta) = \frac{1}{|L| \times |M|} \sum_{i=1}^{|L|} \sum_{j=1}^{|M|} \frac{(b_{\mathrm{lJ}} X b_{\mathrm{Lm}} - b_{\mathrm{lm}} \times b_{\mathrm{LM}})^2}{{b_{\mathrm{lM}}}^2 \times {b_{\mathrm{Lm}}}^2} \tag{3.4}$$

Unlike MSR, SMSR can identify scaling patterns but is unable to detect shifting correlations. The procedure is modified to apply CC [96] twice: once using MSR for evaluation and once using SMSR to find shifting and scaling patterns (not concurrently).

In Intensive Correlation Search (ICS), Ahmed et al. [9] accepts that MSR cannot identify simultaneous shifting and scaling patterns, and as a result, introduces a new algorithm that includes a new metric called the SSSim score. The SSSim score is defined as follows for two gene expression patterns, let's say $g_1 = a_2 1, a_2, \ldots, a_m$ and $g_2 = b_1, b_2, \ldots, b_n$, as given by Equation 3.5.

$$SSSim(g_1,g_2) = 1 - \frac{\sum_{i=2}^{n-1} \frac{|\frac{a_{i+1}-a_i}{a_2-a_1} - \frac{b_{i+1}-b_i}{b_2-b_1}|}{2xmax(|lmean_i - \frac{a_{i+1}-a_i}{a_2-a_1}|.|lmean_i - \frac{b_{i+1}-b_i}{b_2-b_1}|)}}{n-2} \tag{3.5}$$

where,

$$lmean_i = \begin{cases} mean(\frac{a_{i+1}-a_i}{a_2-a_1}, \frac{b_{i+1}-b_i}{b_2-b_1}, \frac{a_{i+2}-a_{i+1}}{a_2-a_1}, \frac{b_{i+2}-b_{i+1}}{b_2-b_1}), \\ \text{if i=2} \\ mean(\frac{a_i-a_{i-1}}{a_2-a_1}, \frac{b_i-b_{i-1}}{b_2-b_1}, \frac{a_{i+1}-a_i}{a_2-a_1}, \frac{b_{i+1}-b_i}{b_2-b_1}), \\ \text{if i=n-1} \\ mean(\frac{a_i-a_{i-1}}{a_2-a_1}, \frac{b_i-b_{i-1}}{b_2-b_1}, \frac{a_{i+1}-a_i}{a_2-a_1}, \frac{b_{i+1}-b_i}{b_2-b_1}, \\ \frac{a_{i+2}-a_{i+1}}{a_2-a_1}, \frac{b_{i+2}-b_{i+1}}{b_2-b_1}), \text{otherwise.} \end{cases} \tag{3.6}$$

SSSim score for a pair of gene expressions spans from 0 to 1, with a value of 1 denoting the ideal demonstration of shift-and-scaling correlation.

Developed by Yip et al.[828], HARP (Hierarchical approach with Automatic Relevant dimension selection for Projected clustering) offers an efficient statistic (Relevance Index,RI). By adding the column-wise relevance indices, this metric assesses the quality of the bicluster. For a column $j \in J$ RI is defined as[828] in Equation 3.7.

$$R_{\mathrm{Ij}} = 1 - \frac{{\sigma_{\mathrm{Ij}}}^2}{{\sigma_{\mathrm{j}}}^2} \tag{3.7}$$

where ${\sigma_{\mathrm{Ij}}}^2$, is the variance of the values in column j for the bicluster and ${\sigma_{\mathrm{j}}}^2$ represents the variance of the whole data set. According to the experimental conditions, the bottom-up merging approach HARP [828] iteratively merges biclusters that meet a RI threshold. Constant biclusters are a situation where HARP maximizes quality and avoids producing overlapped solutions.

The Maximum Similarity Biclustering (MSB) algorithm and a similarity score for biclusters were proposed by Liu et al. [431]. It is a top-down, iterative procedure that starts with the assumption that the entire matrix is the bicluster. It eliminates the row or column with the lowest similarity score at each stage until only one element of the bicluster is left. The technique computes $n+m-1$ sub-matrices [431] for a data matrix with n rows and m columns. The sub-matrix with the highest similarity score is the only bicluster that MSB returns. RMSBE makes an effort to fix a problem MSB had with the unique case of roughly squared biclusters. Additionally, by applying random selection for the reference gene and averaging the similarity scores between the gene pairs, the process is sped up [553].

Weighted Fuzzy-based Maximum Similarity Biclustering (WF-MSB), a fuzzy-based extension of MSB, was proposed by Chen et al. [77] . The method seeks to extract biclusters based on user-defined reference genes. Due to the occurrence of large differences from the baseline of all expression values, it has been noted that biclusters produced using this method are quite significant. Additionally, WF-MSB frequently identifies both the most similar and the most dissimilar bicluster in relation to the reference gene.

In Biclustering by Iteratively Sorting Weighted Co-efficients (BISWC), Teng et al. [678] use a Pearson Correlation Coefficient (PCC) [545] variant to define a metric (the weighted correlations index, or WCI). Using WCI, sorting and transposing are applied alternately to the gene expression data in order to compare the genes and conditions. The main objective was to emphasize attributes with substantially greater influence.

BIclustering by Correlated and Large number of Individual Clustered seeds (BICLIC) [848] likewise makes use of PCC [48] to assess the biclusters. Three stages make up the search procedure. The first stage of seed discovery uses individual dimension-based clustering to produce an unknown number of bicluster seeds by labeling and combining the genes according to their expression levels under various situations. Iteratively attempting to combine each gene or condition until the PCC [48] is exceeded by a predetermined threshold, the second phase seeks to expand these seeds, either gene-wise or condition-wise. Although not every gene in the candidate bicluster matrix may exhibit correlated patterns under every condition, the seed expansion process ensures that the candidate biclusters are correlated with the seed. By removing the less linked groups of genes and circumstances, the third phase filters the data to produce correlation-based biclusters. However, it's possible that various seed biclusters will produce products based on correlation that overlap. Therefore, the duplicate biclusters are eliminated in the fourth and final phase by sorting and comparing them. A candidate bicluster is discarded if it turns out that every gene and condition is present in another prospective bicluster.

Ahmed et al. [9] propose a biclustering technique Intensive Correlation Search (ICS) and a similarity measured called SSSim that can analyze biclusters. The core concept of ICS is the iterative extraction of correlated subspaces for various genes. The largest sample set for which gene expressions are correlated is regarded as a subspace for a user-defined threshold, τ. This method uses greedy search with a stochastic technique to add a random element, rendering the algorithm non-deterministic.

Spectral Biclustering (SB) [311] was designed specifically for analyzing microarray cancer datasets. The basic assumption is that with blocks of high expression as well low expression levels, the expression matrix conforms to a checkerboard-like structure. Hence, this approach primarily focuses on finding these distinct patterns using Singular Value Decomposition (SVD) [192] and eigenvectors. The search further incorporates normalization of genes. SB prevents overlapping among biclusters and also makes attempts to ensure that every gene and condition is included in at least one bicluster.

3.2.1.2 Stochastic Iterative Greedy Search

This approach uses greedy search with a stochastic technique to add a random element, rendering the algorithm non-deterministic.

In order to speed up the biclustering process, Yang et al. [809] present FLexible Overlapped biClustering (FLOC), which uses a heuristic to combine with the random masking of the values in the data matrix. The algorithm starts with the creation of k initial biclusters, then adds rows and columns according to a certain probability, before iteratively eliminating one row or column at a time. The goal is to select an action that will increase the average MSR values throughout the k biclusters, and the process will only come to an end if no action can be discovered to raise the overall standard. The algorithm favours larger biclusters, while constant biclusters are rejected using the variance.

Random Walk Biclustering (RWB) by Angiulli et al. [27] employ user-provided weights to adjust for a composite game function that combines three distinct goals—MSR, gene variance, and bicluster size. The Greedy Randomized Adaptive Search Procedure (GRASP) [135] algorithm, which takes into account combinatorial problems, employs an iterative multi-start meta-heuristic that is implemented in two phases: first, the construction of a greedy randomized solution, and then a search of the solution's immediate vicinity for a local minimum. The best of the solutions is the result [553]. The self-tuning threshold value used by the GRASP variation known as RGRASP-B [136] is connected to the candidate solutions. Based on the calibre of the most recent computed solutions, the threshold value is recalculated on a regular basis.

The Pattern Driven Neighborhood Search (PDNS) method put forth by Ayadi et al. [34] replaces a lower quality neighbor with a higher quality neighbor in order to incrementally improve a first candidate answer. The input data matrix is pre-processed by this technique into a behavior matrix M'', where each gene's trajectory patterns across all conditions make up the rows and all of the genes' patterns across specific samples make up the columns. To obtain the initial bicluster, the algorithm uses two greedy algorithms, CC [96] and OPSM [41], and then encodes it in its behavior matrix M''. The algorithm alternates between two basic components.

3.2.1.3 Nature Inspired Meta-heuristics

This approach involves algorithms that are based on effective natural phenomena, such as ant colony optimization, artificial immune system, swarm optimization, and evolutionary computation to solve complex optimization issues.

Simulated Annealing (SA) [308], a well-known stochastic technique that simulates the natural crystallization process, is included in Simulated Annealing Biclustering (SA-B) [58]. The Particle Swarm Optimization approach, which simulates the behaviour of a flock of birds or a school of fish, is the foundation of CMOPSOB [425]. Multi-Objective Multipopulation Articial Immune Network (MOM-aiNet) [104] develops a number of sub-populations that explore various parts of the search space and models the collection of solutions as a multi-population space. The algorithm begins by selecting people made up of one row and one column at random, and then iterates sub-populations by cloning and changing the individuals.

3.2.2 Non-metric Biclustering Approach

In non-metric biclustering algorithms the search is not guided by any evaluation measure. These algorithms are divided into groups based on what are thought to be their main distinguishing traits, while some of them may be included in more than one group [553]. Graph theoretic approach is used in graph-based techniques, which either (a) use nodes to represent individual bicluster components, such as genes or samples, or both, or (b) use nodes to represent entire biclusters. In addition to statistical modeling of the expression data presented as a bipartite graph, Statistical Algorithmic Method for Bicluster Analysis (SAMBA) [671] employs a graph theoretic approach. The bipartite graph is made up of genes and circumstances, with edges denoting substantial changes. Vertex pair weights are assigned using a probabilistic model, and biclusters with higher likelihood are given heavier weights. In addition to finding maximal biclusters that frequently match certain identical requirements and may have overlapping sections, the MicroCluster [893] approach may also find biclusters that display changing patterns by utilizing the appropriate exponential transformations. Utilizing probability theory to statistically analyze the dataset, probabilistic biclustering algorithms build probabilistic models. For exploratory analysis on multivariate data, Plaid Model [333] was proposed. Here, the gene-condition matrix is made up of the superposed layers that represent the biclusters. When there were multiple biclusters, there was a significant amount of overlap between the biclusters in the original Plaid Model [333]. This is in contrast to a single bicluster, when the Plaid Model's model can be applied. To remedy the problem, the Bayesian biclustering (BBC) [199] model allows overlap in only one of the dimensions of gene and condition. Through linear mapping between vector spaces, some biclustering algorithms that use a linear algebraic approach find the correlated sets of submatrices in the input data set. Optimal permutations of the rows and columns are carried out to improve the organization of the components in the data matrix. Biclusters are defined as Order-Preserving Sub-Matrices (OPSMs) [41] when all of the genes' expression levels result in the same linear ordering for one of the experimental orderings. A sub-matrix is referred to as an OPSM [41], in other words, if there is a permutation of its columns such that the values in every row are strictly ordered in ascending or descending order. By checking for a large linear order tendency in the rows of the expression data matrix, one may be able to loosen this condition somewhat.

3.3 Some Selected Biclustering Algorithms

From the four distinct approaches stated in the preceding part, we chose a few widely used biclustering techniques for our biclustering-based analysis framework. We include eight biclustering algorithms in this section based on their behavior, popularity and effectiveness in handling gene expression data. Table 3.1 lists the pros and cons of these algorithms and the datasets they were validated on Following is brief description of all eight algorithms.

3.3.1 Bimax

Using a divide-and-conquer method, Bimax lists every conceivable bicluster in the input data matrix. The input data must be binarized because the algorithm prefers to look for rectangles of 1s in a binary matrix. The method starts by selecting any row that contains a mix of 0s and 1s. Either the incoming data matrix satisfies the requirements, or no such row is present. This is true if either all of the entries in the matrix are 1 (in which case the entire matrix is a single bicluster) or all of the elements in the matrix are 0 (in which case there is no bicluster). The first row, $r*$, of the input $m \times n$ matrix, M, is randomly selected as the algorithm's starting point. $r*$ is then used to partition M into two submatrices, each of which is processed independently. By splitting the columns C=$1,\ldots,n$ into two sets, C_{U}=$C : M[r*,c] = 1$ and C_{V}=$C - C_{\mathrm{U}}$, submatrices are discovered. The algorithm then divides m rows into three sets as follows: 1) R_{U}: Rows with 1s only in C_{U} , 2) R_{W}: Rows with 1s in both C_{U} and C_{V} and 3) R_{v}: Rows with 1s only in C_{V}. Following are some observations made when the rows and columns of M are adjusted to make each set contiguous: Because both C_{U} and C_{V} are empty, the submatrix by (R_{U},C_{V}) cannot form any bicluster. Alternatively, the submatrix by (R_{U},C_{V}) contains all possible biclusters in M. As long as there are any rows with mixed 0s and 1s, this process will recursively process U, followed by V.

In particular, the Bimax method concentrates on datasets with entries that are binary values, denoting the presence or absence of features or events. It works by repeatedly looking for biclusters with high co-occurrences of 1s (presence) in particular subsets of rows and columns. It begins with a small starting bicluster and gradually grows it by adding rows and columns that make it more likely that 1s will occur together. The algorithm analyzes the importance and quality of the biclusters using statistical tools like *p-values* or significance scores. Bimax is a greedy method that gives preference to smaller, more frequent biclusters over bigger ones. It looks for biclusters inside subsets of rows and columns that are both dense (have a high density of 1s) and homogeneous (have comparable patterns of 1s). Until a stopping requirement is satisfied, such as a predetermined number of iterations or reaching a maximum permissible size for biclusters, the algorithm keeps improving and growing the biclusters. The output of the Bimax algorithm is displayed as biclusters, which are collections of rows and columns. A subset of rows and columns with consistent patterns of 1s are represented by each bicluster. These biclusters can indicate functional correlations or co-occurring occurrences in binary data, as well as insights into the interactions between features and samples.

3.3.2 XMOTIFs (conserved gene expression Motifs)

A subset of genes is an xMOTIF in the context of xMOTIFs if it meets the criteria listed below: a) a subset is simultaneously conserved across a subset of samples, and b) the gene expression level in the subset is conserved across a set of samples if it is in the same state in every sample [553]. A list of the intervals that correspond to the states in which each gene is expressed in the samples is a fundamental prerequisite for each gene [553, 554]. However, various restrictions are made to the conservation, maximality, and size in order to prevent the discovery of extremely small or extremely huge xMOTIFs. The objective is to determine the largest xMOTIF using a probabilistic technique that takes advantage of the xMOTIF's mathematical structure [553]. To locate several xMOTIFs in the data, samples that match each xMOTIF's expectation are successively eliminated, followed by a search for the next greatest xMOTIF, until all samples are connected to at least one xMOTIF.

The goal of xMOTIF is to identify recurring patterns or motifs in various biological situations. The xMOTIFs method works by breaking down a gene expression matrix into biclusters, or groups of genes and conditions with comparable expression patterns. In contrast to conventional biclustering methods, which take into account only one matrix, xMOTIFs makes use of numerous matrices that represent gene expression data under various experimental settings. Using currently available biclustering techniques, the approach starts by producing a starting set of seed biclusters. Then, by taking into account both intra-matrix and inter-matrix interactions, these seed biclusters are enlarged and refined. Inter-matrix interactions capture shared patterns across several matrices, whereas intra-matrix relationships focus on patterns seen inside specific gene expression matrices. For the purpose of determining the relevance of the discovered biclusters, xMOTIFs uses statistical methods and optimization approaches. It seeks to identify conserved motifs with strong co-expression and enrichment for biological pathways or functional gene sets. The outcomes of xMOTIFs are frequently represented as collections of genes and circumstances that collectively constitute conserved motifs, signifying co-expression patterns that are remarkably persistent across several experimental situations. These patterns can shed light on the molecular or regulatory mechanisms that underlie gene expression.

3.3.3 Plaid models(PM)

The gene-condition matrix is represented using plaid models [333] as a superposition of layers that correspond to biclusters [553, 554]. In this study, Lazzeroni and Owen provided numerous iterations of the model and enabled a gene to reside in more than one bicluster. According to Pontes et al.[553, 554] and Lazerroni [333], the most generic model is as given Equation 3.8.

$$X_{ij} = \sum_{k=0}^{K} (\theta_{ijk} - \sigma_{ijk} - \kappa ijk) \tag{3.8}$$

where, X_{ij} = the expression level of the i^{th} gene in the j^{th}sample, κ = the number of biclusters, θ_{ij0}= the background layer and θ_{ijk}= the four types of models, corresponding to the type of biclusters (overlapped, exclusive, ...). Each $\sigma_{ik} \in 0,1$= { 1, if i^{th} gene is in the k^{th} bicluster; 0 otherwise} [553]. Each $\kappa jk \in 0,1$= { 1, if j^{th} sample is in the k^{th} bicluster; 0 otherwise}. By looking for a plaid model that minimizes the sum of squared errors while approximating the data matrix to the model, this greedy approach finds k biclusters by adding one layer at a time.

PM employs a set of criteria with the aim to find subgroups of data where patterns exhibited may vary between different subsets. Rectangular subsets of rows and columns with similar behavior are assigned as biclusters in the data matrix and are used as the basis for the operation of the algorithm. Identification of subsets of rows and columns that are similar to one another in particular areas of the data matrix is the primary goal of PM and it is accomplished by focusing on identifying local patterns in the data. PM is capable of detecting complicated patterns and capturing overlapping biclusters with flexibility. PM decreases bicluster complexity through an optimization framework and increases coherency in the bicluster.

3.3.4 Iterative Signature Algorithm (ISA)

In Iterative Signature Algorithm [44], biclusters are primarily defined as transcription modules retrievable from the expression data [553, 554]. A set of co-regulated genes alongwith a set of conditions where co-regulation is the most stringent [553] constitutes a transcription module(TM). TMs in the data are found by applying a generalized Singular Value Decomposition (SVD). In a module, the gene and condition similarity is determined by two thresholds which in turn regulate its size. The algorithm starts with a random selection of a set of genes or conditions which are then iteratively refined until they match the definition of a TM producing one bicluster for every iteration. As initial selection of seeds is random and lack overlap restrictions, the resulting biclusters might have overlapped genes and/or conditions.

When genes within a module are co-regulated by similar regulatory mechanisms, the ISA method is particularly helpful for identifying transcriptional regulatory modules. It works iteratively, beginning with a single seed module and progressively growing it into a bicluster. It incorporates aspects of non-negative matrix factorization and clustering algorithms. The basis matrix and the activation matrix are updated by ISA after each iteration. The basis matrix shows the expression patterns of genes within the bicluster, whereas the activation matrix represents the activity levels of genes in a bicluster across samples. The activation matrix is encouraged by the sparsity constraint used by ISA to contain a small number of highly active genes in a subset of samples. This restriction encourages the discovery of distinct and coherent gene groups. The algorithm keeps going through iterations where it improves the bicluster, updates the activation and basis matrices, and thinks about the value and importance of the bicluster. When a stopping requirement, such as a set number of iterations or when further expansion does not significantly enhance the bicluster quality, is satisfied, the process comes to an end. Genes can take part in several modules since ISA permits overlapping biclusters. The ability to capture intricate interactions and shared regulatory processes between genes is a benefit of this trait. The ISA results are shown as collections of genes and samples that collectively make up biclusters. Each bicluster is made up of a set of co-expressed genes with comparable patterns in a subset of data. These biclusters can shed light on the regulating mechanisms, functional relationships, or biological processes that underlie them.

3.3.5 Factor Analysis for BIcluster Acquisition (FABIA)

For the analysis of gene expression data in Factor Analysis for BIcluster Acquisition (FABIA) presented by Hochreiter et al. [237], Multiplicative Model (MM) is used since it enables modeling of heavy-tailed data. Furthermore, by standardizing expression values, data pre-processing produces fictitious multiplicative effects. A bicluster is described in this work as a pair of row sets or column sets with similar rows on similar columns, and vice versa. If one of two vectors in an MM is a multiple of the other or if their correlation as realized random variables is negative, then the vectors are comparable. The outer product $\lambda : ZT$ of two vectors, where $\lambda =$ prototype column vector and $Z =$ vector of factors with which prototype column vectors is scaled, can be used to show linear dependence on subset of rows and columns. For genes and samples not included in the bicluster, $lambda$ and Z are 0, respectively. With additive noise, p biclusters can be described, as given in Equation 3.9.

$$X = \sum_{i=1}^{p} \lambda_i {Z_i}^T + \gamma = \Lambda Z + \gamma \tag{3.9}$$

where, $\gamma in \mathbb{R}^{nxl}$ = additive noise, $\lambda_i \in \mathbb{R}^n$ =sparse prototype vector and $Z_i \in \mathbb{R}^l$ = sparse vector of factors of the i^{th} bicluster. FABIA model utilizes the concepts of Laplacian Distribution, Gaussian and Bessel functions so as to exactly get the desired model characteristics of heavy tails. Generative MM allows ranking of biclusters according to their information content.

3.3.6 QUalitative BIClustering algorithm(QUBIC)

QUBIC [348] initially represents the input data in the form of a matrix with integer values either qualitatively or semi-qualitatively. Under a subset of conditions, if two rows of the matrix have identical integer values, then the two corresponding genes can be considered correlated. The qualitative (or semi-qualitative) representation helps the algorithm in detecting different kinds of biclusters such as scaling patterns. Furthermore, it is well suited for detection of positive and negative correlation expression patterns. The algorithm initiates by taking into account the qualitative or semi-qualitative matrix with vertices for genes and constructing a weighted graph. An Edge that connects a pair of genes, however, has a corresponding weight computed based on the similarity between the two corresponding rows. Initially, a bicluster is built using the heaviest unused edge as the seed and the algorithm proceeds to add additional genes to the current solution iteratively. QUBIC employs a method known as 'merge and split' to iteratively improve the biclusters through progressive bicluster refinement of assignments. Through merging of biclusters FABIA seeks to capture substantially coherent patterns while in the split phase it seeks to separate subpatterns within the biclusters. Until the a convergence requirement is met, this refinement of the biclusters continue. QUBIC assesses the coherence of the biclusters by using 'modularity score' that measures how closely the actual pattern resembles the predicted pattern. Biclusters with higher modularity score are consistent and thus deemed as significant.

3.3.7 Iterative Binary Biclustering algorithm with greedy search (iBBiG)

iterative Binary Biclustering algorithm with Greedy search (iBBiG) [208] is an iterative greedy search algorithm that starts with a random initial bicluster and employs greedy search to identify neighboring biclusters that share a large proportion of genes and samples. Until convergence is achieved, the algorithm iteratively combines these neighboring biclusters to find larger biclusters. By gradually deleting genes and samples that do not significantly improve the bicluster's coherence, iBBiG achieves refinement and seeks to minimize the proportion of false positives. iBBiG is effective on binary datasets as well as in finding biologically significant biclusters. Initial biclusters identified by iBBiG consists of a single row and a single column. Quality score measures the resemblance of the pattern exhibited by the identified subset as compared to the predicted pattern for the random assignment. Quality score corresponding to each row and column is used as a rating. After each iteration, the quality score of the resulting biclusters is assessed which takes into consideration all potential row and column additions into the existing biclusters. The row and column addition that optimizes the improvement in quality score are chosen. This process is repeated until further improvement in the quality score is not possible. Bicluster expansion stage is followed by the refining stage where the reliability and significance of the biclusters are assessed using statistical tests. Until a predetermined criteria such as the number of iterations is met or there is a lack of progress, iBBiG iteratively repeats the bicluster expansion and the refinement steps. Post processing in iBBiG further eliminates biclusters that do not meet minimal size or significance threshold, are redundant or have heavy overlappings.

3.3.8 FLexible Overlapped biClustering (FLOC)

FLexible Overlapped biClustering (FLOC) [808, 809] employs a information theoretic measure , mutual information [107], to measure the similarity between genes and conditions. By utilizing mutual information, FLOC algorithm separates the genes and conditions into groups until a halting criterion is encountered. Unlike most biclustering algorithms, FLOC uses a probabilistic approach for the creation of biclusters and accounts for the prior probabilities of the biclusters. FLOC algorithm is capable of finding adaptable as well as overlapping biclusters. It seeks to capture complicated relationships between subsets of rows and columns by allowing overlapping patterns. Randomly chosen subsets of rows and columns constitute the initial biclusters for FLOC. Using an iterative optimization process, the biclusters are improved. FLOC algorithm consists of two stages: a) expansion and b) shrinkage. During the expansion stage, FLOC adds or removes rows and columns to or from the existing biclusters. The quality of the biclusters are assessed based on a scoring scheme that captures the relevance and coherence of the biclusters. The configurations of addition or removal that leads to optimal scoring are chosen. During the shrinkage stage, FLOC gradually reduces the size of the biclusters thus enabling the occurrence of overlappings. By taking into consideration their contribution to the overall quality of the bicluster, FLOC eliminates the rows and columns. Until a halting criterion .,such as the number of iterations, stability of the biclusters or the convergence of the scoring function is encountered, FLOC iteratively repeats expansion and shrinkage stages. Post processing in FLOC eliminates bicluters with poor quality, or are redundant or do not meet certain requirements like minimal size or significance threshold.

Tab. 3.1: Eight chosen biclustering methods: A Comparison

Approach	Algorithm	Pros	Cons	Dataset Used
Iterative Greedy Search	Iterative Binary Biclustering algorithm with greedy search (iBBiG) [208]	• With discrete or categorical values for the variables, iBBiG is specifically made to handle binary or binary-like datasets. • To examine the coherence of the biclusters, iBBiG uses quality scores. These ratings measure how closely the observed patterns inside the biclusters resemble the expected patterns based on random assignment. • To improve the robustness of the biclustering results and aid in the identification of more significant and trustworthy biclusters, iBBiG adds a refinement step that assesses the statistical significance of the detected biclusters. • iBBiG allows for flexibility in in bicluster size and shape.	• It can be difficult and may require manual adjustment or lengthy testing to choose the right parameters for iBBiG. • When used on huge datasets with a lot of rows and columns, iBBiG could have scaling issues. • BBiG is not necessarily appropriate for datasets containing continuous or mixed-type variables because it is created especially for binary or binary-like datasets. • iBBiG concentrates on finding biclusters with comparable binary patterns, but it may not explicitly capture more complex structures, including overlapping or hierarchical biclusters.	Single sample GSA data: GSE20685 [300], breastCancerNKI[710], breastCancerVDX [499], GSEAlm data: 21 normalized breast cancer gene expression datasets annotated with 107 clinical covariates from GeneChip Ontology Database [413]
Divide & Conquer	Bimax [556]	• The dataset is represented using a binary matrix in the method, which simplifies comprehension and application. • Large-scale datasets with thousands of rows and columns may be handled, and it is computationally efficient. • Binary can spot biclusters in binary datasets or continuous datasets with consistent patterns of presence or absence. • Bimax can find overlapping biclusters thus indicating that it can record intricate dependencies and linkages between different sets of rows and columns. • Highly interpretable biclusters are produced by Bimax using binary matrix representation. • Bimax can produce numerous biclusters, each of which can capture a different pattern in the data. • Parameter tweaking is not necessary with Bimax because no parameter values must be specified.	• Working with continuous or multi-valued datasets can provide limitations because Bimax may not be directly applicable or may call for pre-processing in order to convert the data to a binary format. • Although Bimax is more scalable and economical its computational complexity may become a limiting issue as the number of rows and columns rises. • Missing values may interfere with the representation of the binary matrix and compromise the results. • It might have trouble identifying biclusters with more intricate or overlapping structures, which could restrict its capacity to identify specific kinds of patterns in the data. • The output may differ based on the original seed selection, which could result in variability and poor reproducibility.	Yeast Saccharomyces Cerevisiae cell cycle [677]

Approach	Algorithm	Pros	Cons	Dataset Used
Probabilistic Models	Plaid Models (PM) [333]	• By enabling overlapping biclusters, it can handle datasets with heterogeneous patterns and can manage complex datasets with various patterns or when separate clusters have some commonalities. • It primarily focuses on identifying local patterns in the data, i.e., subsets of rows and columns that behave similarly in particular areas of the data matrix. • Plaid Models provides a variety of optimization criteria, giving users choice in selecting the best model. • It makes use of optimization techniques to be scalable to datasets with many rows and columns.	• In order to establish the ideal parameter values, which can be time-consuming and subjective, it is necessary to choose a number of parameters, such as the number of biclusters and the optimization criterion. • Because it lacks integrated automatic model selection methods, users must manually choose the best model and provide its parameters in accordance with their comprehension of the data. • Plaid Models makes the assumption that the input data matrix is full and error-free. • For non-continuous data, such as categorical or binary data, which call for further adaptations or extensions to handle properly, it might not be as suited.	Yeast Saccharomyces Cerevisiae [677]
	Conserved gene expression Motifs (xMOTIFs) [514]	• Smaller and bigger gene sets that display conserved expression patterns can be accommodated by x-Motif, which can find motifs of various sizes. • The x-Motif method uses statistical approaches and metrics to increase the reliability and accuracy of motif discovery. • Multiple conserved motifs can be found using x-Motif within a dataset, allowing for the identification of distinctive expression patterns and regulatory processes in various situations or sample sets. • The motifs found by x-Motif are frequently quite interpretable and have biological significance. • Biological knowledge from the past, such as gene annotations, pathway details, or functional databases, can be combined with x-Motif.	• The computational complexity of rises with the number of samples and genes, which can make it costly and time-consuming to analyze huge datasets. • Due to noise interference, this may result in the detection of false positive motifs or the inability to detect pertinent motifs. • It can be difficult to specify motif lengths in advance for x-Motif and the inclusion of unnecessary motifs or the exclusion of important ones can result from selecting motif lengths that are too short. • x-Motif often identifies non-overlapping motifs, which could be insufficient to fully represent intricate gene regulatory linkages and functional interactions. • It may not be suitable for analyzing dynamic or temporal gene expression patterns because it is primarily made for static gene expression data.	Leukemia [193], Colon cancer [21], B-cell lymphoma [19]

Approach	Algorithm	Pros	Cons	Dataset Used
Graph Based Approaches	QUalitative BIClustering algorithm(QUBIC) [348]	• The datasets with binary or qualitative values, where the variables are discrete rather than continuous, are the ones that QUBIC is specifically built to handle. • The coherence of biclusters is evaluated by QUBIC using a modularity score, which quantifies the similarity between observed patterns inside biclusters and expected patterns based on random assignment. • Iteratively merging and dividing the biclusters is used by QUBIC to improve the biclusters and thus can find both fine-grained subpatterns and large-scale coherent patterns. • Due to its adaptability, QUBIC can detect biclusters with more intricate structures or missing values in the binary matrix.	• The number of desired biclusters or the criteria for merging and splitting must be chosen when using QUBIC. It can be difficult to determine the ideal parameter values; manual tuning or exploratory analysis may be needed. • Since QUBIC is designed especially for binary or qualitative data, datasets with continuous or mixed-type variables might not be a good fit. • Although QUBIC is intended for categorical or binary data, categorical data lacks straightforward numerical meanings, which may make it more challenging to give the patterns that are discovered biological or meaningful interpretations. • It can be computationally demanding, especially for large datasets or datasets with a lot of variables and samples.	E.coli [150], Arabidopsis [451], BRCA tumor [680]
Linear Algebra Based	Iterative Signature Algorithm (ISA) [44]	• Co-expressed gene modules that display comparable expression patterns across a selection of samples are identified using ISA, which aids in identifying groupings of genes that are probably controlled by similar regulatory mechanisms or take part in similar biological processes. • The sparsity constraint in ISA helps identify distinct and important gene subsets within modules, lowering the likelihood of incorporating irrelevant genes by encouraging the activation matrix to include few highly active genes within each bicluster. • Each iteration updates the activation and basis matrices, gradually enhancing the bicluster quality. This iterative refinement aids in capturing subtle expression patterns and strengthens the biclustering results.	• Choosing proper parameter values can be difficult, and different settings may produce different outcomes. To get the best parameter values, it may take some trial and error or domain knowledge, which can take time and be subjective. • ISA can require a lot of processing power, especially for large gene expression datasets or when there are a lot of iterations. • In order to capture non-linear linkages or dependencies, ISA makes the assumption that gene expression patterns and regulatory modules have a linear relationship. • ISA is primarily intended for the analysis of gene expression data and may not be as effective for other omics data types or datasets with unique properties, such as categorical or time-series data.	Yeast Saccharomyces Cerevisiae [677]

Approach	Algorithm	Pros	Cons	Dataset Used
Factor Analysis Model	Factor Analysis for BIcluster Acquisition (FABIA) [237]	• FABIA assists in the identification of coherent groups of genes and samples that display comparable expression patterns by decomposing the data into biclusters ascribed to these latent components. • FABIA uses sparsity constraints to identify compact and informative biclusters while minimizing the inclusion of irrelevant genes or noisy samples. • It provides statistical methods to evaluate the quality and relevance of the discovered biclusters. • FABIA can be modified to analyze various omics data types, such as proteomics or metabolomics data, while being initially developed for gene expression data.	• Numerous parameters, including the quantity of latent components, sparsity restrictions, and convergence criteria, must be chosen for FABIA. • It may be computationally intensive in case of large datasets or datasets with many latent components. • FABIA is vulnerable to noise and problems with the quality of the gene expression data it receives as input. • Due to the fact that FABIA assumes a linear link between gene expression patterns and latent components, it may be ineffective at capturing non-linear relationships or dependencies in the data. • When deciding how many biclusters should be present in the data, FABIA doesn't offer any defined guidelines.	Breast Cancer [680], B-cell lymphoma [19]
	FLexible Overlapped biClustering (FLOC) [808, 809]	• In order to capture more complicated linkages and patterns that may exist across subsets of rows and columns, FLOC is specifically made to find flexible and overlapping biclusters within the data. • FLOC is more suited for datasets with a variety of patterns and structures since it automatically estimates the number of biclusters during the optimization process based on the features of the data and the quality score mechanism. • The identification of overlapping biclusters by FLOC allows the detection of these intricate interactions and offers a more thorough comprehension of the underlying biological mechanisms. • FLOC uses an iterative optimization method that enables the biclusters to be improved over several iterations, improving their quality and relevance and enabling the collection of more precise and significant patterns in the data.	• Due to FLOC's iterative optimization process and the inclusion of overlapping biclusters, the algorithm is more sophisticated and computationally expensive. This can lead to longer execution times and present problems when dealing with resource constraints or time-sensitive studies. • FLOC calls for the specification of a number of parameters, including the scoring function, the termination criterion, and the initialization approach, necessitating substantial experimentation or domain knowledge. Poor parameter selection might result in substandard biclustering outcomes. • If FLOC is used on particularly large or high-dimensional datasets, scaling problems may arise. • When presented with noisy or poor-quality data, or when the underlying patterns are extremely complicated or nuanced, performance of FLOC may suffer.	

3.4 BicGenesis: A Method To Identify ESCC Biomarkers Using Biclustering Approach

The proposed framework, named BicGenesis detailed in Fig 3.2, is a biclustering analysis framework. It employs eight well known biclustering methods across various approaches to generate biclusters. Hub-genes of the relevant biclusters identified by our relevant bicluster selection criteria are candidates for potential biomarkers and are further validated. We choose eight biclustering methods Bimax [556] , x-MOTIF [514] , Plaid Models [333] , ISA [44], FABIA [237], QUBIC [348], iBBiG [208], and FLOC [808, 809] detailed in section 3.3. Table 3.1 gives a detailed comparison of all these eight algorithms.

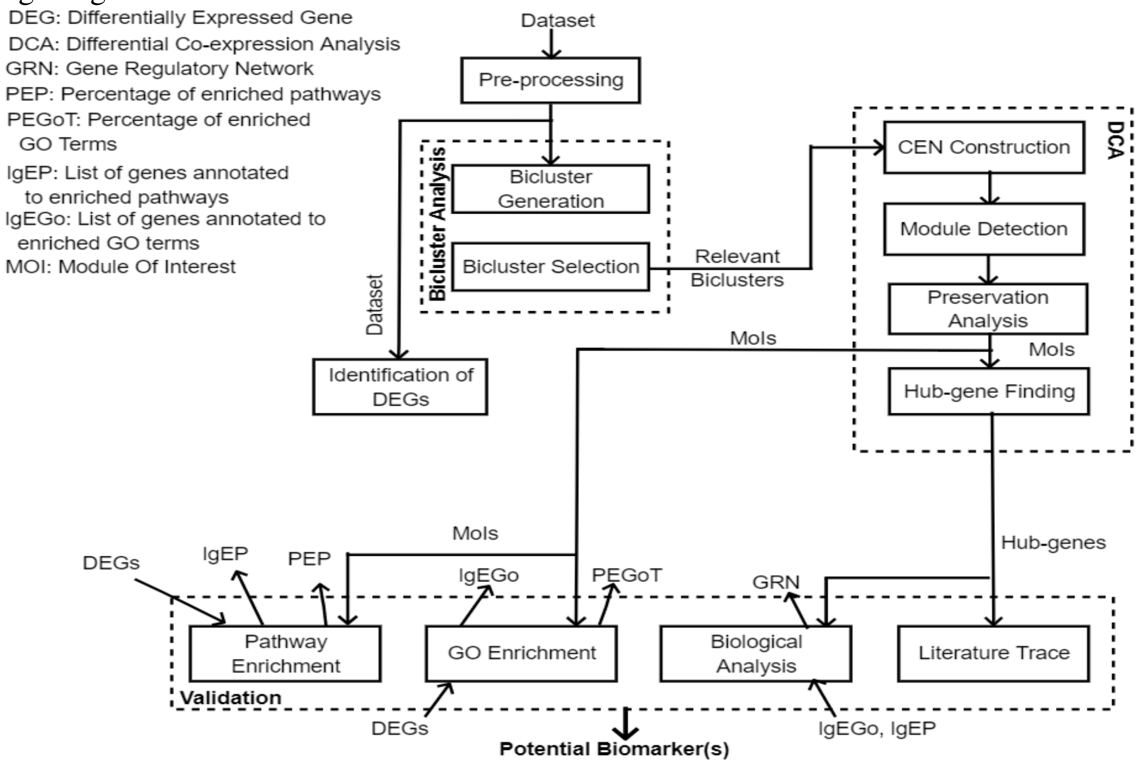


Fig. 3.2: Proposed Biclustering Analysis Framework

The eight biclustering algorithms were chosen based on their popularity, behavior, and efficiency in handling gene expression data. All chosen algorithms namely, Bimax [556] , x-MOTIF [514] , Plaid Models [333] , ISA [44], FABIA [237], QUBIC [348], iBBiG [208], and FLOC [808, 809] are widely used. Furthermore, we also took into consideration the ease of implementation of these algorithms without the incorporation of complications that rise from their application in various platforms. In other words, we wanted that all chosen algorithms can be implemented on a single platform. Few other biclustering algorithms such as Cheng and Church (CC) [96], and spectral biclustering [311] were implemented in the initial stages. However, after multiple iterations with various parameters, we observed that CC generates only one bicluster corresponding to the entire dataset rendering the generated bicluster irrelevant for further analysis. Various iterations of implementation spectral biclustering across all datasets did not produce relevant biclusters and as such the spectral biclustering algorithm was not taken into further consideration.

The BicGenesis framework pipeline is represented in detail in Figure 3.2. BicGenesis framework starts with the dataset as input and based on this input data the corresponding pre-processing technique (Section 2.7.1 and Section 2.7.2) is chosen. This pre-processed data is further taken as input into the Identification of DEGs unit that identifies DEGs from the pre-processed dataset. The identified DEGs are input into the Validation unit to facilitate GO and Pathway enrichment. The pre-processed data is also taken as input into the Biclustering Analysis unit. The bicluster generation sub-unit of this unit employs the eight chosen biclustering algorithms to generate biclusters from the pre-processed dataset. Row and column behavior or patterns within the specified subset of the bicluster as well as comparison of the means of various biclusters are evaluated for the selection of the relevant biclusters by the Bicluster Selection sub-unit . These relevant biclusters are then taken as input into the DCA analysis unit. The DCA unit of BicGenesis divides the pre-processed dataset into normal and disease subsets and constructs co-expression networks corresponding to each subset. This is then followed by extraction of biclusters as modules and preservation analysis (Section 2.1.9) so as to identify relevant modules. All moderately preserved modules (Section 2.1.9) are identified are as relevant and we term them as "Modules of Interest" (MoI). All hub-genes in relevant modules are further considered as "Biomarker Candidate Genes" (BCGs) and thus the DCA employs a hub-gene finding algorithm to identify hub-genes for downstream analysis. As described in the proposed Biomarker Criteria (Section 2.5) we employ GO enrichment, pathway enrichment and gene regulatory behavior analysis to biologically validate BCGs as potential biomarkers. The literature trace sub-unit of the validation unit traces existing literature that establishes the BCGs as potential biomarkers for ESCC and six other SCCs related to ESCC. The details of each unit of the BicGenesis framework are further described in the following five subsections.

3.4.1 Pre-processing

The proposed framework accepts input from either bulk RNA-Seq or microarray data, and the pre-processing technique is selected in accordance with the input data type. Pre-processing for bulk RNA-Seq data entails removing low read counts, normalizing the dataset, and transformation whereas pre-processing for microarray data entails removing redundant and undesirable data, normalizing the dataset, and missing value estimates. 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.

3.4.2 Bicluster Analysis

The pre-processed data is input to the bicluster analysis unit. In the bicluster analysis unit of BicGenesis, we first generate biclusters by employing the chosen eight biclustering methods mentioned in Section 3.3. This is followed by the selection of relevant biclusters for subsequent downstream analysis.

3.4.2.1 Bicluster Generation

The bicluster generation subunit of bicluster analysis unit employs each of the eight bicluster analysis method discussed in Section 3.3 on the pre-processed input data. BicGenesis employs biclustering methods Bimax [556] , x-MOTIF [514] , Plaid Models [333] , ISA [44], FABIA [237], QUBIC [348], iBBiG [208], and FLOC [808, 809] individullay on the pre-processed input data. All biclusters detected by each of the eight methods are considered for subsequent bicluster selection to identify relevant biclusters.

3.4.2.2 Bicluster Selection

All biclusters detected by each of the eight individual bicluster analysis method are input to the bicluster selection subunit. Row and column behavior or patterns within the specified subset of the bicluster are referred to as the row effect and column effect of a bicluster, respectively [481, 96, 556]. It encapsulates the characteristic response or similarity between the rows or columns within the bicluster. A post hoc test called the Tukey test [696], commonly referred to as Tukey's Honestly Significant Difference (HSD) test, is used to compare the means of various groups. To evaluate substantial variations between subgroups or conditions inside a bicluster, it can be utilized. Row effect, column effect, and tukey tests are performed for every bicluster identified using each of the eight biclustering methods. A significant *p-value* supports the idea that there is a real pattern or difference inside the bicluster by indicating that the observed row effect/column impact is unlikely to happen by chance alone. *p-values* for pairwise comparisons between the means of the subgroups will be provided by the tukey test. When taking into account the overall variability within the bicluster, these *p-values* show whether the observed differences between the subgroup means are statistically significant. Thus, the bicluster selection subunit considers all biclusters that have significant *p-values* for row effect, column effect and tukey test as relevant biclusters.

3.4.3 DCA

The DCA unit of BicGenesis receives all relevant biclusters found by the bicluster analysis unit as input. DCA is done with the intention of identifying biologically relevant modules that are more intuitive to analyze and validate. The DCE unit constructs two CENs at the initial stage that correspond to the input dataset's subsets for normal and disease.

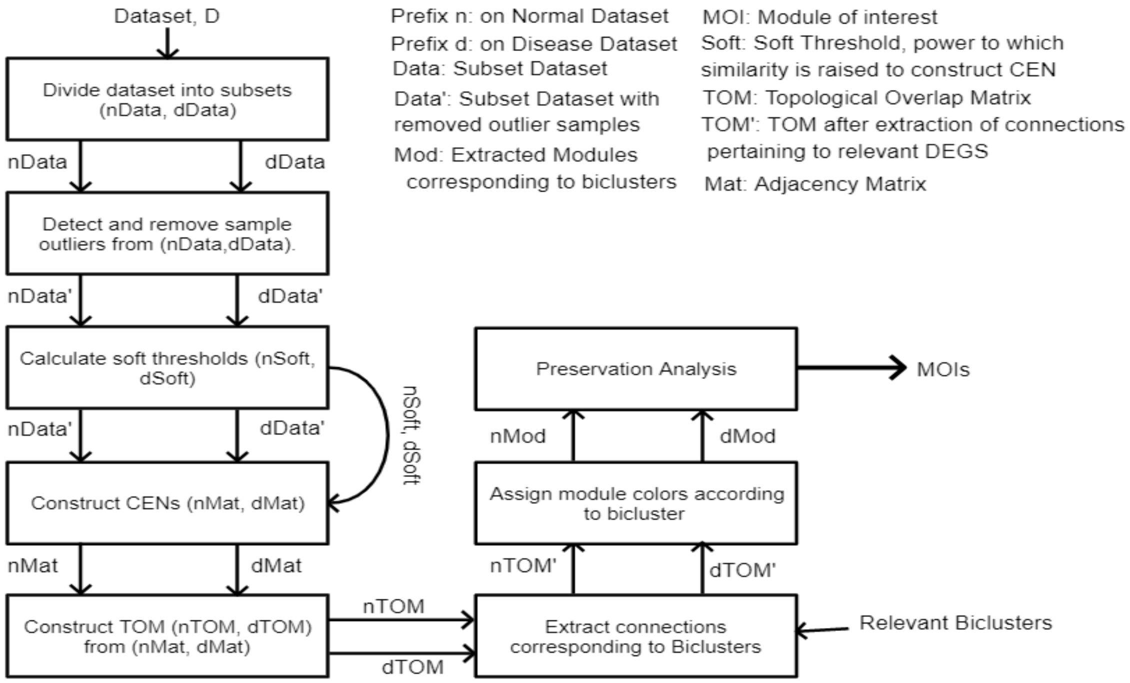

Dataset, D
Divide dataset into subsets (nData, dData)
nData
dData
Detect and remove sample outliers from (nData,dData).
nData'
dData'
Calculate soft thresholds (nSoft, dSoft)
nData'
dData'
nSoft, dSoft
Construct CENs (nMat, dMat)
nMat
dMat
Construct TOM (nTOM, dTOM) from (nMat, dMat)
nTOM
dTOM
Extract connections corresponding to Biclusters
Relevant Biclusters
nTOM'
dTOM'
Assign module colors according to bicluster
nMod
dMod
Preservation Analysis
MOIs
Prefix n: on Normal Dataset
Prefix d: on Disease Dataset
Data: Subset Dataset
Data': Subset Dataset with removed outlier samples
Mod: Extracted Modules corresponding to biclusters
MOI: Module of interest
Soft: Soft Threshold, power to which similarity is raised to construct CEN
TOM: Topological Overlap Matrix
TOM': TOM after extraction of connections pertaining to relevant DEGS
Mat: Adjacency Matrix

Fig. 3.3: Pipeline for DCA

Figure 3.3 provides an insight into the steps involved in the DCA pipeline employed by BicGenesis. The first step of the pipeline divides the dataset, D into its normal (nData) and disease (dData) subsets. This is then followed by the detection and removal of outlier samples from both normal and disease subsets resulting in two subsets nData' and dData'. Construction of co-expression networks (CENs) involves computation and choice of the soft threshold power to which co-expression similarity is raised. Thus, two soft thresholds corresponding to each subset (nSoft, dSoft) are computed and chosen on nData' and dData'. On the basis of the approximate scale-free topology [38] criteria , we select the soft threshold power. Soft thresholding is followed by the construction of CENs in the form of two adjacency matrices corresponding to the normal (nMat) and disease subset (dMat). The fifth step of the pipeline involves conversion of the adjacency matrices (nMat, dMat) into a topological overlap matrix (TOM [574]), which yields a comparable dissimilarity matrix of the same sizes (nTOM, dTOM), in order to reduce the impact of noise and erroneous associations. The two TOMs corresponding to the two subsets are input to the fifth step of the pipeline which involves the extraction of modules. As our primary focus is on bicluster analysis we extract modules that correspond to the relevant biclusters identified by the bicluster analysis subunit of BicGenesis. This is then followed by the assignment of module colors that correspond to the relevant biclusters. The final step of DCA involves preservation analysis described in detail in Section 2.1.9 to identify relevant modules which we term as "Modules of Interest" (MoI).

We consider all relevant biclusters identified by the previous unit as modules of the CENs. However, as all the biclustering methods were performed on the entire dataset, the columns of the biclusters span over both normal and disease subsets. To determine whether a module is a normal or a disease module we consider the majority. In other words, to identify a bicluster as a normal module, the majority of the samples (columns) in the bicluster must be in normal dataset and vice versa.

To extract modules from the CENs, all connections that pertain to each relevant bicluster in the corresponding normal or disease dataset are extracted . These modules are subjected to preservation analysis by the DCE unit in order to find biologically relevant modules. The term "Modules of Interest" (MoI) (Definition 3.4.2) is used to describe these modules. Following this, hub-genes in these modules are found employing intra-modular connectivity in WGCNA [327]. The detailed DCA pipeline is illustrated in Fig. 3.3. All of the hub-genes in the biologically relevant modules found by the DCE unit are taken into consideration as potential biomarker candidates and we term them as biomarker candidate genes (BCG) (Definition 3.4.4).

Definition 3.4.1 (Module)**.** A module is a subset of genes, $M \subset G$, where there exists high coherence or homogeneity among the genes in terms of associations or expression similarities.

Definition 3.4.2 (Module of Interest (MoI))**.** A module is defined as 'module of interest' if (i) its *size* ≥ 50, and (ii) it is not highly preserved , i.e., it is either non-preserved ($Z_{summary} < 2$) or moderately preserved ($2 \leq Z_{summary} \leq 10$).

Definition 3.4.3 (Hub-gene)**.** A gene g_i is defined as a hub-gene in a MoI extracted by our method BicGenesis, if g_i is topologically enriched (i.e., highly connected).

Definition 3.4.4 (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 BicGenesis.

3.4.4 Identification of DEGs

List of genes annotated to enriched GO terms (lgEGo) and list of genes annotated to enriched pathways (lgEP) are essential to establish the biological relevance of the hub-genes identified by the hub-gene finding unit of the framework. To achieve this we employ a DEG finding method on the pre-processed input dataset. The identified list of DEGs are then input to the validation unit of BicGenesis.

3.4.5 Validation

We take two approaches to validation. In order to determine the BCGs indicated by BicGenesis as potential biomarker(s), we first evaluate the quality of the module(s) extracted by the DCA unit as MoI (Definition 3.4.2). The following steps are taken to validate the extracted 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 identified in biologically significant modules by the DCA unit are regarded as potential biomarker candidates and are referred to as Biomarker Candidate Genes (BCG). 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 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 sub-units 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.

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 identifies BCGs that have enriched GO terms and enriched pathways associated to them. In other words, the biological analysis unit recognizes the BCGs that are present in lgEGo and lgEP. For the purpose of establishing the regulatory behavior of these BCGs in the network, this unit further detects BCGs that are TFs and constructs GRN. The validation unit of the framework's literature trace sub-unit finds BCGs that have published literature traces that support being regarded 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).

3.5 Experimental Results

To evaluatethe performance of our method, we consider a critical disease, ESCC. Three ESCC datasets such as GSE130078 for bulk RNA-Sequencing, and GSE20347 and GSE23400 for microarrays have been selected to evaluate the performance of our method, BicGenesis. Each dataset's specifics (Table 2.1) are detailed in Sections 2.6.1 and 2.6.2. A DELL workstation running Windows 10 Pro with a 3.70GHz Intel(R) Xeon(R) W-2145 CPU and 64 GB of RAM is used for experimental evaluation. In the R programming environment (Section 2.2.1), we run the results. Each dataset's characteristics are provided in Table 2.1. In all three datasets, the gene expression of tumors have been analyzed and contrasted with that of surrounding contrast tissue.

3.5.1 Pre-processing

There are 46 samples and 57,783 genes in the bulk RNA-Seq dataset, GSE130078. We eliminate genes with low read counts since large datasets often make analysis more difficult. We do this by counting the number of copies of each gene in a million , i.e., Counts Per Million (CPM) for each sample, and keeping only the genes with $CPM > 1$ for at least two samples. The dataset shrinks from 57,783 to 22,270 as a result. The dataset is then normalized as a next step. For analysis, we also take into account the two microarray datasets GSE20347 and GSE23400 gene expression levels across samples are the datasets' inputs. We first pre-process the data by eliminating redundant and unnecessary genes, estimating missing values, and normalizing the data. However, there are no missing values for either GSE20347 or GSE23400, so we continue down the pipeline. After preproceesing the microarray datasets GSE20347 and GSE23400 are of the dimensions $22,277 \times 34$ and $22,283 \times 106$, respectively.

3.5.2 Bicluster Analysis

For each pre-processed dataset, namely GSE20347, GSE23400 and GSE130078, we detect biclusters by employing each of the eight biclustering methods discussed in section 3.3.

3.5.2.1 Choice of parameters

For all chosen eight biclustering algorithms, we have employed most parameters except the number of biclusters and the number of iterations based on suggestions of the original work. R package Biclust [296] [1] was used to perform three biclustering methods, x-Motif, Bimax, and Plaid Models. For ISA, FABIA, and QUBIC we used the R packages ISA2 [111] [2], FABIA [237] [3], and QUBIC [887] [4], respectively. For iBBiG and FLOC we used the R packages iBBiG [208] [5] and BicARE [187] [6], respectively. These packages implement the parameters suggested by the respective works as default parameters. The number of biclusters and the number of iterations inputs as and when required were determined through exhaustive applications as well as multiple iterations and were chosen based on our perception of the best results.

3.5.2.2 Bicluster Generation

As mentioned earlier BicGenesis employs eight biclustering methods Bimax [556], x-MOTIF [514] , Plaid Models [333] , ISA [44], FABIA [237], QUBIC [348], iB-

[1] https://CRAN.R-project.org/package=biclust
[2] https://CRAN.R-project.org/package=isa2
[3] https://bioconductor.org/packages/fabia/
[4] https://bioconductor.org/packages/QUBIC/
[5] https://bioconductor.org/packages/iBBiG/
[6] https://bioconductor.org/packages/BicARE/

BiG [208], and FLOC [808, 809] to generate biclusters for two microarray datasets, GSE20347 ($22,277 \times 34$), and GSE23400 ($22,283 \times 106$), and one bulk RNA-Seq dataset, GSE130078 ($22,270 \times 46$) . We have observed after pre-processing , all three datasets have approximately around 22,200 genes but varying sizes of samples. With the aim to detect biclusters with sizeable number of genes (rows) and not too many or too few conditions (columns) we perform multiple iterations and exhaustive applications. Most methods across all datasets regardless of the changes in other parameters corresponding to the method can detect 10 biclusters of sizeable number of genes and conditions. As such for all methods across all datasets we aim to determine 10 biclusters. In GSE20347 and GSE23400, while most methods detected 10 biclusters (Table 3.2), ISA detected 7 and 25, respectively. In GSE130078, aside from ISA that detected 5 biclusters we also observed that Plaid also detected only 3 biclusters.

3.5.2.3 Bicluster Selection

As mentioned previously, we take into consideration row effect, column effect and tukey test to determine the significance of each bicluster. Through multiple iterations of application we have observed that a $p-value = 0.05$ for all three parameters determines all biclusters in all three datasets as relevant. With the aim to reduce the number of biclusters as well as to obtain highly significant biclusters we determined set the significance value for all three parameters as $p-value = 0.01$.

Definition 3.5.1 (Relevant Biclusters)**.** A bicluster is defined as 'relevant bicluster' if (i) its $rowEff$ is significant with $p-value \leq 0.01$, (ii) its $colEff$ is significant with $p-value \leq 0.01$, and (iii) its $tukeyTest$ is significant with $p-value \leq 0.01$. Here, $rowEff$, $colEff$ and $tukeyTest$ refers to results of row effect, column effect and Tukey's Honestly Significant Test (HSD) briefly discussed in Subsection 3.4.2.2.

Table. 3.2 summarizes all biclusters detected across all three datasets by the eight biclustering algorithms. Finally, we remove all genes that are not assigned to any bicluster. This step reduces the number of genes in GSE202347 from 22,277 to 20, 941. Similarly, the number of genes in GSE23400 is reduced from 22,283 to 20, 864 while in GSE130078 it reduces from 22,270 to 20,938.

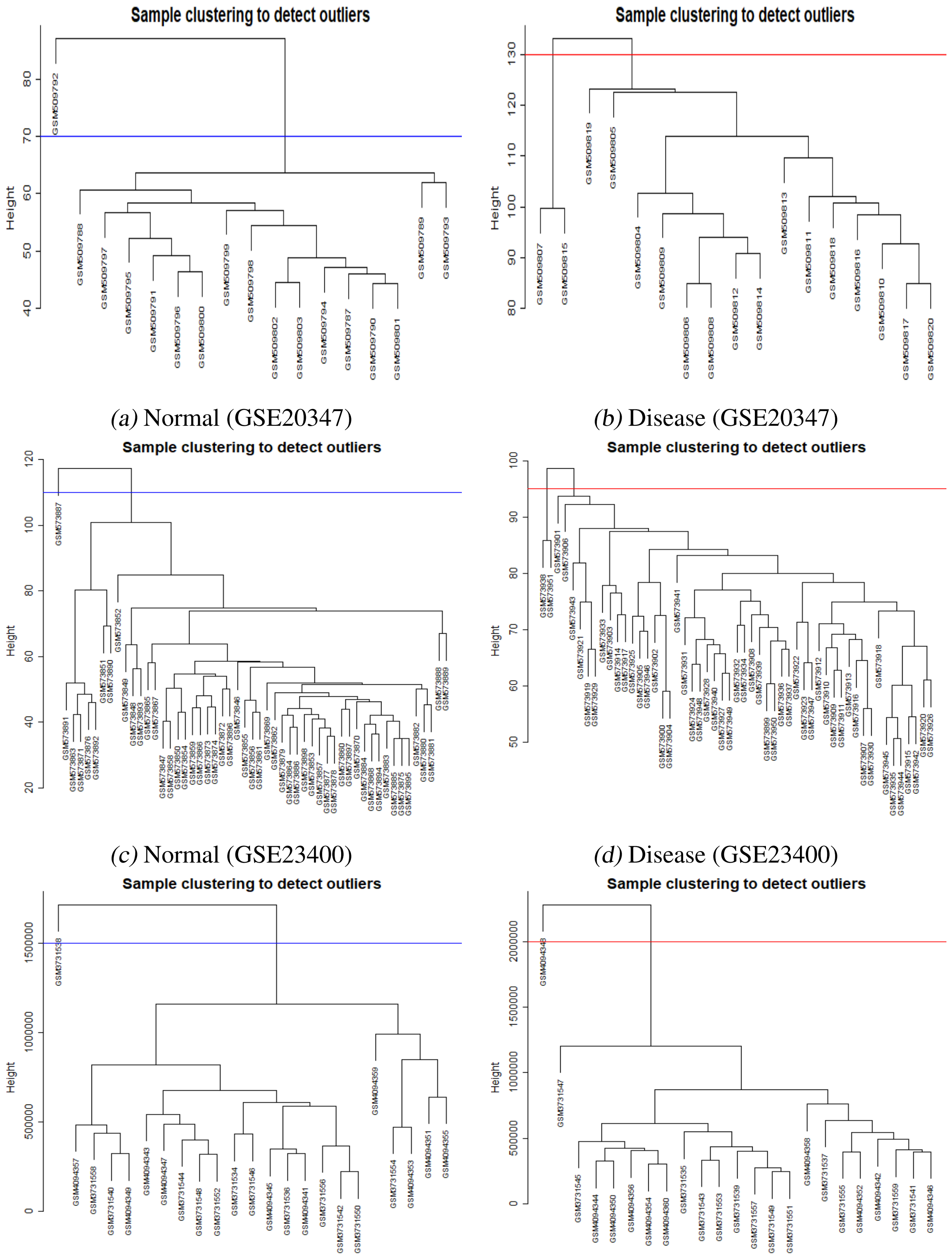
Tab. 3.2: Summary of the biclusters detected by BicGenesis in all three datasets. DB: No. of detected biclusters, RB: No. of relevant biclusters.

	Method	DB/ RB	Relevant Biclusters		Method	DB/ RB	Relevant Biclusters
GSE20347	x-Motif	10/2	1, 2	GSE130078	x-Motif	10/ 8	1, 2, 3, 4, 6, 7, 8, 9
	BiMax	10/5	1, 5, 6, 9, 10		BiMax	10/ 9	1, 2, 3, 4, 6, 7, 8, 9, 10
	Plaid	10/10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10		Plaid	3/ 1	1
	ISA	7/6	1, 2, 3, 4, 6, 7		ISA	5/ 5	1, 2, 3, 4, 5
	FABIA	10/ 6	3, 5, 6, 7, 8, 9		FABIA	10/ 2	2, 3
	QUBIC	10/4	1, 2, 6, 8		QUBIC	10/10	1, 3, 4, 5, 6, 7, 8, 9 , 10
	iBBiG	10/ 9	1, 3, 4, 5, 6, 7, 8, 9 , 10		iBBiG	10/ 7	1, 2, 3, 4, 5, 7, 9
	FLOC	10/ 8	1, 2, 3, 4, 6, 7, 8, 9		FLOC	10/ 4	1, 5, 7, 10
GSE23400	x-Motif	10/ 7	1, 2, 3, 4, 5, 6, 8		FABIA	10/ 9	1, 2, 3, 4, 5, 6, 7, 8, 9
	BiMax	10/ 9	1, 2, 4, 5, 6, 7, 8, 9, 10		QUBIC	10/ 3	6, 7, 10
	Plaid	10/ 9	1, 2, 3, 4, 5, 6, 7, 8, 10		iBBiG	10/ 4	1, 2, 3, 6
	ISA	25/ 19	4, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25		FLOC	10/10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

3.5.3 DCA

We create co-expression networks (CEN) using Weighted Gene Co-expression Network Analysis (WGCNA) [327] to examine the interactions between the genes in a bicluster as well as the variations in behavior under normal and disease conditions. Fig. 3.3 in Section 3.3 provides the detailed pipeline for DCA in our system.

To find outliers, we begin by hierarchically clustering the samples of the three dataset. 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. 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 and disease 3.4d samples, respectively. Cuts at $h = 1500000$ (Blue) and $h = 2000000$ (Red) in the case of GSE130078 remove one sample of normal (Fig. 3.4e) and one sample of disease (Fig. 3.4f).



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

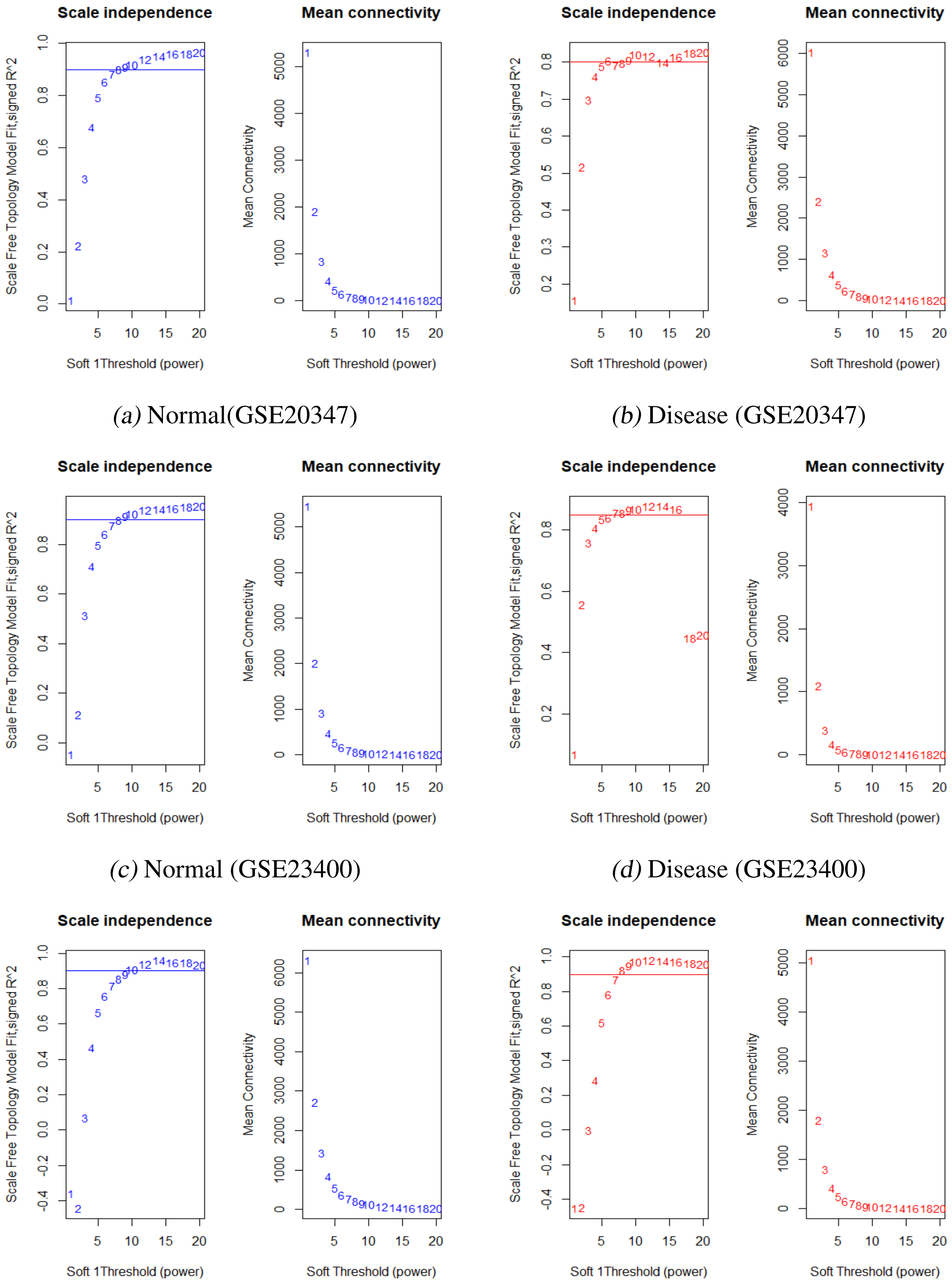
(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 3.4: Hierarchical trees for normal samples in microarray datasets a) GSE20347 and c) GSE23400, e) bulk RNA-Seq dataset, GSE130078. Hierarchical trees for disease samples in microarray datasets b) GSE20347 and d) GSE23400, f) bulk RNA-Seq dataset, GSE130078. The lines represent the height for tree cut in normal (blue) and disease (red) subsets.

3.5.3.1 Soft Threshold



(a) Normal(GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 3.5: Soft thresholds for a) normal and b) disease in GSE20347, c) normal and d) disease in GS23400, and e) normal and e) disease in GSE130078

We apply soft threshold to the normal (Blue) and disease (Red) samples of dataset GSE20347. 9 is the lowest power for which the network maintains scale-free topology, as can be shown in Fig. 3.5a and Fig. 3.5b. As shown in Fig. 3.5c and Fig. 3.5d, the soft threshold for normal (Blue) and disease (Red) samples in GSE23400 is set at 9. In contrast, for GSE130078, normal (Blue) and disease (Red) samples are selected with soft thresholds of 12 (Fig. 3.5e) and 9 (Fig. 3.5f), respectively.

3.5.3.2 CEN Construction

Using the soft threshold exponent 9, we compute the adjacency matrices for the normal and disease samples of the GSE20347 dataset, yielding two corresponding adjacency matrices both with a size of $22,277 \times 22,277$. Similar to this, GSE23400 produces adjacency matrices of size $22,283 \times 22,283$ each with a soft threshold power of 9. The number of genes in GSE130078 is decreased to 22,270 after CPM filtering, resulting in two adjacency matrices with soft thresholds of 12 (normal) and 9 (disease) and sizes $22,270 \times 22,270$ each. The adjacency matrices used to create the associated Topological Overlap Matrix (TOMs) [574] have the same size as the relevant adjacency matrix. Here, it is noteworthy to mention that we construct the CENs from the normal and disease subset of the dataset and then extract the modules corresponding to the biclusters. As, such the number of reduced genes after removal of genes not assigned to any bicluster is not relevant for CEN construction.

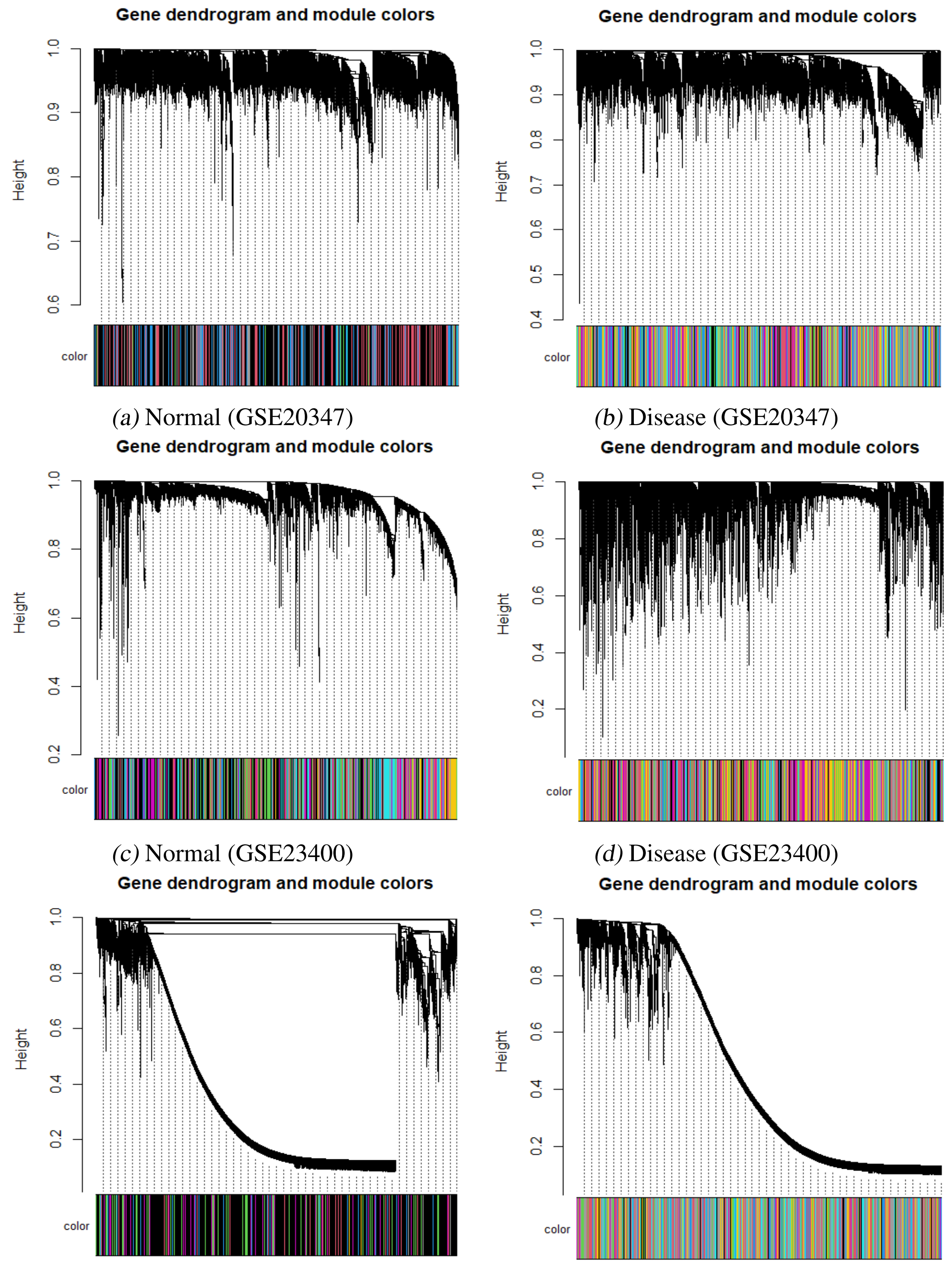
3.5.3.3 Module Extraction

Generally module extraction would entail the employment of methods such as hierarchical clustering on the the normal and disease TOMs to obtain connections in the networks that are highly connected. However, we make the assumption that relevant biclusters detected by the Bicluster analysis unit are highly connected and correspond to biclusters. As such we extract connections corresponding to each relevant bicluster from the normal and disease TOMs and assign them module colors according to bicluster labels. As our aim is primarily to observe the variations in behavior exhibited by biclusters under normal and disease conditions, it is essential that we first first distinguish between normal and disease biclusters. To achieve this we anlyze whether majority of the bicluster conditions (or columns) fall under normal or disease dataset. If most of the columns in a bicluster are part of normal dataset then me consider that as a normal bicluster. Table 3.3 gives a detailed summary of the normal and disease biclusters in each dataset. Furthermore, as we further analyze biclusters of one condition that do not retain majority of its connections under the other condition, i.e, they are not preserved [329], we only consider biclusters that are either normal or disease. As such biclusters that cannot be strictly categorized as normal or disease are not taken into consideration. Bicluster 7 and 10 in FABIA and QUBIC, respectively in GSE23400 are such biclusters.

Tab. 3.3: Subset of Normal and Disease Biclusters. DB: No. of detected biclusters, RB: No. of relevant biclusters

	Method	DB/ RB	Normal Biclusters	Disease Biclusters
GSE20347	x-Motif	10/ 2	1, 2	NULL
	BiMax	10/ 5	NULL	1, 5, 6, 9, 10
	Plaid	10/ 10	7, 8, 9	1, 2, 3, 4, 5, 6, 10
	ISA	7/ 6	NULL	1, 2, 3, 4, 6, 7
	FABIA	10/ 6	NULL	3, 5, 6, 7, 8, 9
	QUBIC	10/ 4	NULL	1, 2, 6, 8
	iBBiG	10/ 9	1, 5, 7, 9	3, 4, 6, 8, 10
	FLOC	10/ 8	2, 7, 8, 9	1, 3, 4, 6
GSE23400	x-Motif	10/ 7	2, 3, 4, 5, 8	1, 6
	BiMax	10/ 9	1, 2, 4, 5, 6, 7, 8, 9, 10	NULL
	Plaid	10/ 9	1, 2, 3, 4, 5, 6, 7, 8	10
	ISA	25/ 19	4, 9, 11, 12, 13, 17, 19, 20, 24	8, 10, 14, 15, 16, 18, 21, 22, 23, 25
	FABIA	10/ 9	1, 2, 6	3, 4, 5, 8, 9
	QUBIC	10/ 3	6	7
	iBBiG	10/ 4	1, 2, 3, 6	NULL
	FLOC	10/ 10	1, 2, 3, 4, 5, 6, 8, 10	7, 9
GSE130078	x-Motif	10/ 8	1, 3, 4, 6, 7, 9	2, 8
	BiMax	10/ 9	NULL	1, 2, 3, 4, 6, 7, 8, 9, 10
	Plaid	2/ 1	1	NULL
	ISA	5/ 5	1, 4	2, 3, 5
	FABIA	10/ 2	2, 3	NULL
	QUBIC	10/ 10	NULL	1, 2 , 3, 4, 5, 6, 7, 8, 9, 10
	iBBiG	10/ 7	3, 5	1, 2, 4, 7, 9
	FLOC	10/ 4	NULL	1, 5, 7, 10

Fig. 3.6a and Fig. 3.6b are the dendrograms for normal and disease subsets in GSE20347 where the strip of color represents the colors assigned to corresponding biclusters. Similarly, Fig. 3.6c and Fig. 3.6d are the dendrograms for GSE23400 while Fig. 3.6e and Fig. 3.6f are the dendrograms for GSE130078.



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

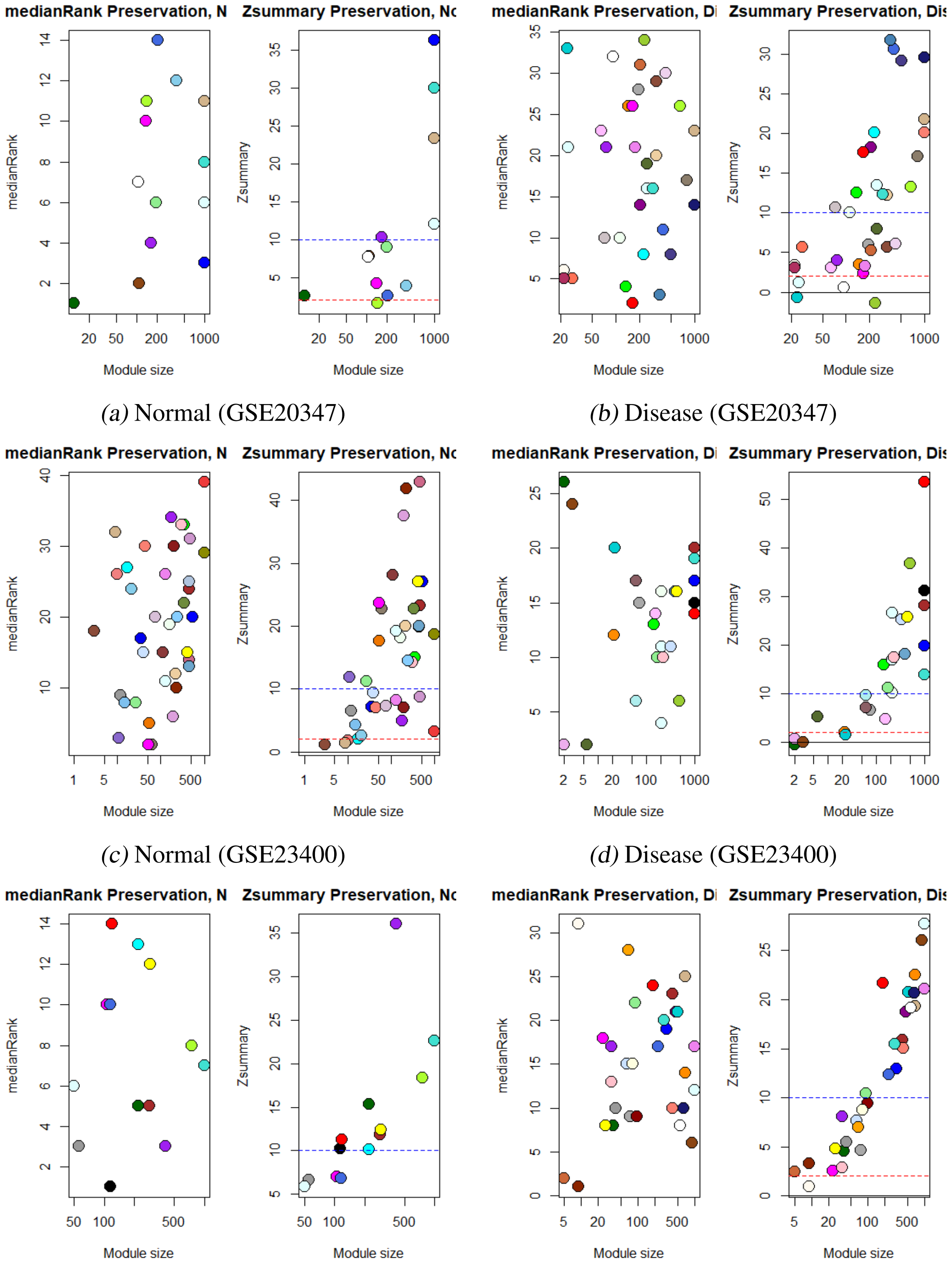
(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 3.6: 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 strip of color represents the colors assigned to corresponding biclusters.

3.5.4 Preservation Analysis

In order to analyze the difference between preserved and non-preserved modules, we follow module extraction by module preservation analysis. While the preserved modules retain the bulk of their co-expressed connections (or edges between two genes), the same cannot be observed from non-preserved modules, according to Langfelder et al. [329].



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 3.7: 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.

A module with $Z_{\text{summary}} < 2$ is regarded as non-preserved, in accordance with Langfelder et al. [329]. We evaluate moderately conserved modules with the $Z_{\text{summary}} < 10$ [329] (Section 2.1.9). Fig. 3.7a, Fig. 3.7c, and Fig. 3.7e are the Z_{summary} plots for datasets, GSE20347, GSE23400, and GSE130078, respectively. In 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.

Tab. 3.4: 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 *Size* $\geq$ 100 and atleast 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}		Ref/Test	Module	Size	Z_{summary}
GSE20347	Normal/ Disease	***greenyellow***	218	1.55407	GSE23400	Normal/ Disease	*salmon4*	3	1.09769
		royalblue	407	2.60446			*tan*	9	1.36990
		darkgreen	12	2.62285			*lightcoral*	10	1.75694
		skyblue	474	3.90280			*cyan*	17	2.01960
		magenta	252	4.16232			*skyblue*	21	2.62263
		white	140	7.69126			***brown2***	1283	3.25331
		saddlebrown	130	7.74030			*skyblue2*	15	4.24438
		lightgreen	498	9.01459			***purple***	226	4.88135
	Disease/ Normal	***yellowgreen***	369	-1.43358			*grey60*	12	6.56521
		darkturquoise	24	-0.67250			***firebrick4***	220	7.07419
		white	127	0.55815			*salmon*	44	7.10097
		lightcyan1	25	1.20614			*blue2*	35	7.15751
		magenta	237	2.33799			***thistle***	93	7.24022
		maroon	22	3.02716			***violet***	187	8.27108
		plum1	97	3.11051			***plum3***	513	8.76015
		violet	299	3.30263			*lightsteelblue1*	39	9.41598
		floralwhite	22	3.42265		Disease/ Normal	*darkgreen*	2	-0.61404
		darkorange	230	3.52126			*saddlebrown*	3	-0.11975
		purple	104	3.95907			*plum2*	2	0.61396
		sienna3	333	5.21882			*darkturquoise*	23	1.41100
		salmon4	545	5.63931			*darkorange*	22	1.97032
		coral1	28	5.71808			***plum1***	244	4.69040
		darkgrey	317	5.96807			*darkolivegreen*	6	5.24692
		thistle2	712	6.07552			***darkgrey***	146	6.54487
		darkolivegreen	418	7.96030			***lightpink4***	111	7.08346

Continued on next page

	Ref/Test	Module	Size	Z_{summary}		Ref/Test	Module	Size	Z_{summary}
							paleturquoise	121	9.53985
GSE130078	Normal/ Disease	***lightcyan***	50	5.83082	GSE130078	Disease/ Normal	darkgreen	37	4.56141
		grey60	56	6.60936			***darkgrey***	74	4.67108
		royalblue	116	6.84477			yellow	27	4.77035
		magenta	106	6.97547			***grey60***	41	5.51684
	Disease/ Normal	floralwhite	9	0.92119			***orange***	68	6.99870
		sienna3	5	2.46668			***lightsteelblue1***	64	7.62333
		magenta	24	2.58236			purple	35	8.11836
		pink	35	2.90630			***lightyellow***	81	8.77480
		orangered4	9	3.31394			***darkred***	98	9.46642

Table 3.4 summarizes the preservation analysis for non-preserved modules in all three datasets. The second column highlights the module preservation reference and test networks. For example, the table reading for module *royalblue* in Normal/Disease subset of dataset, GSE20347, can be interpreted module *royalblue* of size 407 detected in the normal network that is moderately preserved in disease network with a Z_{summary} value of 2.60446. We only consider moderately preserved modules (i.e, $Z_{\text{summary}} < 10$ [329]) of substantial size ($size \geq 50$) as MoIs for subsequent downstream analysis and validation. With the consideration that we identify 20 hub-genes as candidates for potential biomarkers in subsequent downstream analysis we the size of the MoI to be atleast 50. This is to keep the majority of genes as non hub-genes. In GE20347 we have identified 19 MoIs while in GSE23400 9 modules are MoIs.

3.5.5 Hub-gene Finding

To find the hub-genes for each MoI extracted previously we employ WGCNA intra-modular connectivity proposed by Langfelder et al. [327]. Intra-modular connectivity calculates the connectivity of a node to other nodes in the same module. For each MoI (Definition 3.4.2), we compute intra-modular connectivity of each gene and identify genes with high intra-modular connectivity as hub-genes. Table 3.5 gives a detailed summary of top 20 hub-genes in all MoIs identified across each dataset.

Tab. 3.5: Top 20 hub-genes for each MoI in the two microarray and one bulk RNA-Seq datasets using WGCNA [327] intra-modular connectivity. Hub-genes with strong literature evidence of association to ESCC are marked in Red while hub-genes with evidence of association with five other SCCs, HNSCC, LaSCC, LSCC, OSCC, and TSCC are marked in Blue

	Module	hub-genes
GSE20347	*royalblue*	DMXL2, CMKLR2, CNNM2, PCGF1, APBA2, EAPP, APOC1, FEZF2, ZDHHC17, CALD1, DYRK1A, PCLAF, HNRNPA1, NPIPA1, P2RY6, SRSF11, PKD1P1, TPM1, EHBP1, MLF2
	greenyellow	MED21, RPL41, SRSF9, HSDL2, AOPEP, UBE2L3, CTTN, RANBP9, MAST4, IL13RA1, AREG, CNN3, STN1, PDZD2, PDCD4, GMDS, LMOD1, RPL38, TTC9, SLC24A3
	saddlebrown	SET, NDUFS4, SNX3, ADK, DLD, MRPS28, MTFR1, CYB5A, CANX, RAB14, SH2B2, ISCA1, MRPS18A, RNF138, MED13L, ADIPOR1, MSH2, GLUD2, SKAP2, PDLIM5
	skyblue	VAMP3, ARF6, HADHA, DSG2, MAP2K1, CCT4, N4BP2L2, UBE2G1, ECI2, NOTCH2NLA, TSPAN6, BNIP3, PPP1R2, ASCC1, TSPAN6, MACROH2A1, HILPDA, HPGD, ETFDH, BNIP3
	magenta	DIP2A, PSMA2, SPCS1, BLVRA, ATP6V0E1, PSMC1, VPS26A, PPP2CB, SEPTIN10, CHP1, EIF3K, CCNC, SEPTIN10, POLR1D, TMX1, SDCBP, BCAP29, ATP5MC3, RPS27L, REXO2
	white	MCTS1, SDHC, HIKESHI, PCBP1, TRIM33, UBL3, RORA, OCLN, GRB10, AMACR, RCHY1, B3GNT2, TRIP10, POLQ, TPM1, CNPY2, FERMT2, AGFG2, AFF4, CDH3
	lightgreen	DNASE2, SIAH2, UFSP2, ING2, SRP72, TAF1C, ZFAND6, TAB2, CES3, OPN1SW, UBC, SPINT1, CYB5A, MYLK, CLIP3, DAAM2, IL2RA, PICALM, PDE4DIP, BHMT
	yellowgreen	ATP10B, MYO5A, CYP3A5, SASH1, FUT6, MECOM, USF2, OSTM1, CYP3A5, ZC3H15, CRIPT, NFIL3, RUVBL1, DYNC1I2, C2orf49, TMEM38B, ARHGEF10L, GNB5, ELF3, CYP3A5
	whiteD	GFPT1, KIF1A, BPTF, CETN2, UBE3A, ATAT1, COMMD10, GORASP2, ZC3H11A, CYLC2, ZFR, E2F1, ARHGAP4, B2M, SETX, HLA-C, C8orf44, ETV1, DLX5, PHF20
	magentaD	NCR1, NTRK3, GPATCH4, EMILIN2, NODAL, SEMA3G, TNFSF18, TMPRSS2, DMWD, NECTIN1, SPDEF, PART1, ZNF839, DCAF15, MYL1, TRMT44, PTP4A3, MYLPF, RHAG, ABCC5
	plum1	TOMM20, SH3BP4, RPA3, SMARCA5, GCSH, RFC3, PJA1, RNF114, PDGFA, ODF2, HNRNPC, EDEM2, PFKM, CANX, R3HDM1, ABI2, SELENOP, SELENBP1, RPN2, CIAO1

Continued on next page

	Module	hub-genes
GSE20347	*violet*	PRSS3P2, SRI, C6orf120, YTHDC1, TES, UBE2A, AGK, ARPC1A, EXOSC8, CKAP2, NUDT21, GNRH2, IPO5, HTATIP2, SOCS5, RPS6KA3, UCHL5, MTCH2, MTF2, TNFRSF25
	darkorange	SRP14, UBE2M, FSTL1, COPG1, ZNHIT6, SSBP1, TMEM131, RFXANK, PDIA3, SMU1, CALR, PRPF18, IARS2, USP8, WTAP, TNPO1, MRS2, C6orf62, RASA1, GLMN
	purple	SBNO2, PPP6C, EIF3G, HNRNPC, DERA, KSR1, FCN1, CATSPER2, LARP1, TGM4, RPL35, SNTB2, GPR176-DT, PSG3, PSMD4, SHISA6, NFU1, CD4, HMGA2, PDE4A
	sienna3	BUB1, TIPIN, CEACAM1, OIP5, PLAU, TLE2, HDGFL3, CDH12, CALU, TAF1A, HMMR, FEZ2, SLC16A1, MCM10, DBF4, ITGA6, NDUFB7, MTF2, BRD3OS, MSH6
	salmon4	HPX, FGF5, GSN-AS1, TRAPPC4, NNAT, PIN1, PNOC, ACTN3, PDCD10, BUB3, ETNPPL, GLUD2, METAP2, ANP32A, COPS8, ARHGAP11A, RRP12, SNRNP27, KPNA4, NRBF2
	darkgrey	PAK3, POU3F2, ITGA2B, MPZ, CROCCP3, LSAMP, RAB11FIP3, ATRNL1, ELOVL2, POLR2B, CASP2, IQCK, MGC2889, STMN2, NTRK3, CXXC4, ASB1, PADI2, CA6, CD8B
	thistle2	TGFB2, NMBR, SULT1C2, NPAT, CARD14, SUMO2, GCKR, RPL23A, FNDC4, CLCN4, MED18, GBA, ADAM22, BMP3, AGTR2, CARD14, RBMX, SAMD4B, SNTB1, DHODH
	darkolivegreen	ZNF236, COL1A1, RHOQ, FUT2, DACT1, HMGN4, HMGN4, NOX4, CALU, SS18, TRAPPC2L, PLEKHM1, PLAU, RNF114, TOP2A, WASHC3, YKT6, PLA2G12A, TPCN1, DSG2
GSE23400	*brown2*	LGALS1, SDC3, TGFBR2, LOXL1, CAV1, CLCA2, IGFBP7, COL6A2, TGFA, ACTA2, CALD1, RAB11FIP1, FN1, TMEM47, F11R, ABI1, NREP, TAX1BP3, ARAP2
	purple	MYH11, PJA2, PKP3, SAP18, ASL, CYTH1, ABCA7, SEMA3F, SPATA5L1, PLD1, LSM14A, DENND2B, EPS8L1, KRT16, NMT2, TBC1D3E, RIMS2, EHBP1L1, MACO1, ARL6IP5
	firebrick4	SLPI, CALML3, ACOX2, GRN, CADM1, FAM149A, IPO7, CA12, CAMK2G, BRD3, MFF, ELMO3, CLTB, MINK1, BCL2, MAPK3, ACTC1, MYO1D, CA12
	violet	EXTL2, ARHGEF40, ID3, RHOQ, DIPK1A, SOCS5, FTO, ARPP19, PDE4DIP, ROR2, NPTXR, LCAT, SPRY2, SYBU, FOXJ2, PAQR3, RDH14, DEXI, VWA8, DNAJC6
	plum3	KLK11, CAPG, CSK, CAPN1, PTPRM, MAPK13, DGKA, F12, RAB25, DIAPH1, GPR87, PPP1R13L, ACTN1, FUT2, NBEAL2, MFSD5, CYB561, FUT6, GYG1

Continued on next page

	Module	hub-genes
GSE23400	*thistle*	DHCR24, ST14, DNAJB5, PDE4D, TRIM29, RIPK4, CALD1, RHOBTB1, PPARGC1A, SEMA3B, MAP7, EVC, VEGFB, CTSF, CYRIA, PSD4, SRSF9, APRT, SPRR2D, TLE3
	plum1	CD84, VDR, LUC7L3, TEK, BUB1, PPIG, IGKC, USH1C, KCTD12, CETP, IGLC1, TCF4, ATRX, SON, IGLC2, IGLJ3, IGLC1, ESF1, RBM25, PNISR
	darkgrey	EXOSC4, MCM4, PYCR3, TUBB, SLC39A4, FBXL6, HSPD1, UCK2, PRRC2A, PUF60, ACP1, RAD21, G3BP1, NLE1, PSME3, LAPTM4B
	linghtpink4	USH1C, WNT6, CD93, SMG6, EVI2B, IGH, ZC3H13, LCN1, LBP, NEK1, PPWD1, ITK, TARP, GMFG, TRDV2, TOR1AIP1, CAVIN3
	paleturquoise	NME1, WDR74, BOP1, MRGBP, AURKA, NME2, CSE1L, PFDN2, BUB1, GANAB, UBE2C, ATAD2, SLC25A22, ENY2, PAICS, KPNA2, PHB1, TPX2, MYBL2
GSE130078	*lightcyan*	SNX25P1, FOLR3, PLPP4, KRBOX1, AZIN2, PLD6, CCN5, HOXD3, MSL3P1, DTX1, CTXND1, WIPF3, ERVMER34-1, CYP39A1, PRAP1, MMP19, ARNT2, ELN, MAP2K7, TBC1D14
	grey60	IGLVV-58, GLYATL2, RSPO1, SLCO6A1, MYH15, RPS15AP12, STAG3, MCMDC2, HSPA1L, KRT77, SLC13A3, SLC35F3, LINC00319, PNMA2, PRKG2, CMTM3, DNAJC17, LTB, PTGR2, TATDN3
	royalblue	IL6, LINC02904, TCAP, RPS12P7, GSC, TTC34, CLEC18A, LINC02582, STRA8, SNAP25-AS1, EVA1A, AMDHD1, ANKRD19P, CASC8, CBY2, SPANXC, PRAC2, SULT1C2P2, CNIH2, LRRD1
	magenta	VPS37D, ADCY10, LINC00310, SCIRT, OMP, CFHR3, IL1RAPL2, LOC643348, LOC105379109, AMBP, LINC03007, LINC01249, HP, MIR4755, NECAB2, LOC101928682, ARL17A, MRPL53, KLHL14, DOC2A
	darkgrey	KCNMB2, MYH7, STYXL2, CHIA, OMP, MIR4755, ADCY10, SCIRT, NLRP6, CFHR3, IL1RAPL2, LOC643015, RBM34, B3GAT1, CNR2, LOC105379109, BRME1, LNCAROD, SMN1, FBXL8
	orange	ABCC8, HPX, RAPSN, EFCAB12, ZNF648, PGK1P2, TAS2R63P, C4orf48, ZNF415P1, ATOH8, C10orf88B, PRDM6, GDPD1, GFI1, IQCH, TIMM8A, TAS2R4, CACNA1I, ZNF860, RELL2
	lightsteelblue1	SLFNL1, ARMH1, C7orf25, CCN5, MILR1, IGLVV-58, ITPKB-IT1, GPHA2, TIMP4, MTCO3P23, USP3-AS1, DIRAS3, LIME1, HCN4, POTEG, C1orf53, CCR8, FMR1NB, SNAI3-AS1, RHOT1P1
	lightyellow	LOC100422687, MTND5P14, RPL21P53, RNF208, HSPA8P3, LOC100422382, RDM1P5, RBPMS-AS1, RNU6-522P, NLRP6, FCER2, IL1RAPL2, COL23A1, MED15P9, RNF175, SGCA, RTEL1-TNFRSF6B, BBS1, ICAM4, RNF222
	darkred	ADAMTS7P4, MYH7, SULT2A1, BSG-AS1, ALMS1P1, MSLNL, RPSAP12, ABCG4, RRH, RPL18AP6, MIR23B, OXCT2, SNORD20, TNFRSF13C, LINC02404, LOC149935, CLNK, RN7SL587P, IQCD, TLE6

3.6 Validation

We achieve validation through various approaches. Foremost we validate whether the DEGs and MoIs identified by our framework are biologically relevant and highly enriched or not. We achieve this through functional enrichment analysis (Section 2.4.1). Only MoIs that are highly enriched are biologically relevant and considered for further analysis. All hub-genes of the biologically relevant MoIs are considered biomarker candidates genes (BCG). We employ Regulatory Behavior Network analysis (Section 2.4.2) to further validate the biological relevance of these BCGs. Finally, we trace literature for established we-tlab results that help substantiate the BCGs as potential biomarkers for ESCC and five other SCCs associated with ESCC. Through application of our proposed biomarker criteria discussed in Section 2.5 we identify the potential biomarkers for ESCC.

3.6.1 Enrichment Analysis of Biclusters

An MoI must have at least one gene assigned to an enriched Gene Ontology (GO) word or pathway with a significance level of 5% (i.e., $p \leq 0.05$) in order to be considered Gene Ontology (GO) or pathway enriched. We employ DAVID [628, 253] (Section 2.2.3) to carry out functional enrichment analysis. The percentages of genes in the MoI annotated to enriched GO terms and enriched KEGG pathways are shown in Table 3.6. We note that all MoIs identified by BicGenesis are pathway and GO enriched.

Tab. 3.6: Percentages of genes in each MoI extracted from the three datasets that are annotated to the Gene Ontology (GO) databases (BP: Biological Processes, CC: Cellular components or MF: Molecular function) and KEGG pathways.

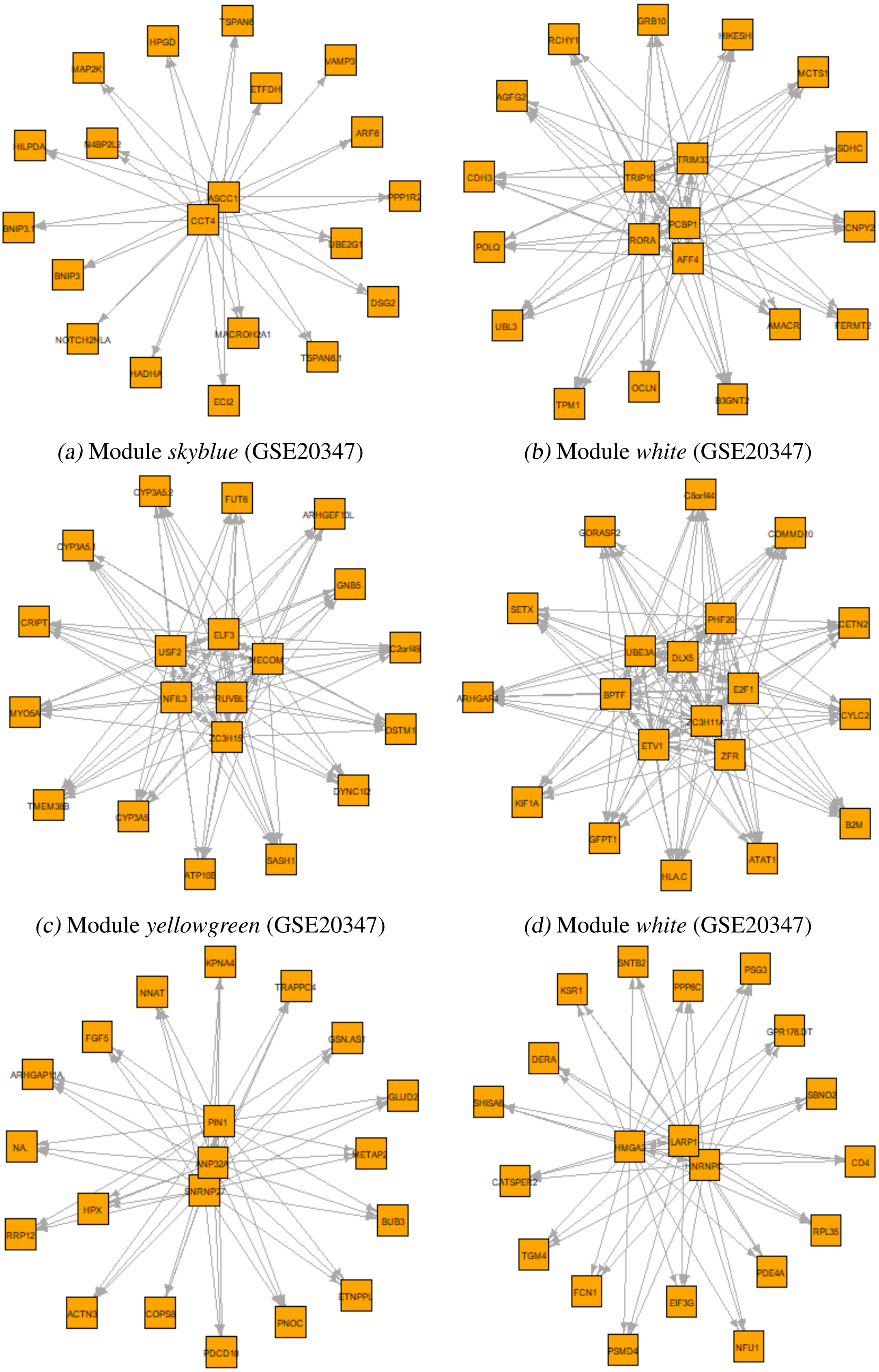
	Module	Size	BP	CC	MF	KEGG		Module	Size	BP	CC	MF	KEGG
			(%)	(%)	(%)	(%)				(%)	(%)	(%)	(%)
GSE20347	*royalblue*	407	93.3	96.0	93.3	48.1	GSE23400	*brown2*	1283	95.8	97.9	96.1	53.7
	greenyellow	218	94.1	94.6	94.1	52.7		*purple*	226	94.4	95.9	96.4	49.7
	saddlebrown	130	98.4	99.2	99.2	63.7		*firebrick4*	220	95.7	97.1	96.2	51.9
	skyblue	474	96.1	97.9	94.7	53.3		*violet*	187	92.1	93.8	94.9	49.7
	magenta	252	94.2	98.8	94.2	56.8		*plum3*	513	94.6	97.6	96.6	54.3
	white	140	96.2	98.5	97.0	51.5		*thistle*	93	95.5	97.8	93.3	53.9
	lightgreen	498	94.5	96.9	95.6	55.2		*plum1*	244	95.5	97.3	96.4	53.6
	yellowgreen	369	93.9	96.6	96.3	41.4		*darkgrey*	146	95.4	97.7	97.7	50.8
	white	127	94.1	95.0	96.6	44.5		*lightpink4*	111	90.2	96.1	92.2	47.1

Continued on next page

	Module	Size	BP	CC	MF	KEGG		Module	Size	BP	CC	MF	KEGG
			(%)	(%)	(%)	(%)				(%)	(%)	(%)	(%)
GSE20347	*magenta*	237	94.0	96.7	95.3	49.8		*paleturquoise*	121	99.1	99.1	100	58.6
	plum1	97	94.6	96.8	97.8	58.1	GSE130078	*lightcyan*	55	90.7	93.0	88.4	51.2
	violet	299	94.0	95.8	94.3	47.2		*grey60*	69	82.5	84.2	86.0	43.9
	darkorange	230	90.8	93.7	91.8	50.2		*royalblue*	541	77.3	82.6	80.5	37.1
	purple	104	94.9	98.0	96.9	53.1		*magenta*	210	81.4	86.2	82.6	40.1
	sienna3	333	96.1	97.7	98.1	51.8		*darkgrey*	102	83.8	87.5	85.0	40.0
	salmon4	545	93.2	97.2	96.6	53.9		*orange*	87	83.6	83.6	82.2	45.2
	darkgrey	317	92.1	93.8	94.5	48.8		*lightsteelblue1*	198	81.5	85.2	85.2	41.4
	thistle2	712	93.3	97.1	95.6	49.7		*lightyellow*	111	77.5	82.0	84.3	40.4
	darkolivegreen	418	94.0	97.1	96.9	50.8		*darkred*	129	80.0	82.7	77.3	36.4

3.6.2 Biological Analysis

We employ gene regulatory network (GRN) construction and functional enrichment analysis to determine the biological relevance of the BCGs found by BicGenesis. The diversity and power of transcription factors (TF) as agents of cell change is astounding. Bhagwat et al. [45] justifies the ongoing search for TFs as possible biomarkers for a variety of cancer types. We note that the BCGs identified by BicGenesis in GSE20347, GSE23400, and GSE130078 are 40, 38 , and 22, respectively, TFs. The biological importance of these TFs is demonstrated by their regulatory behavior in their respective modules. From the non-preserved modules found by our technique, we extract a reasonable subset of hub-genes for straightforward visualization (Fig. 3.8a-3.10). To investigate the regulatory behavior of the corresponding genes, we build a Gene Regulatory Network (GRN) (Section 2.4.2) using these hub-genes and the related Transcription Factors (TFs). An adjacency list with weighted directed edges from TFs to other target genes (TGs) makes up the resulting GRN.



(a) Module *skyblue* (GSE20347)

(b) Module *white* (GSE20347)

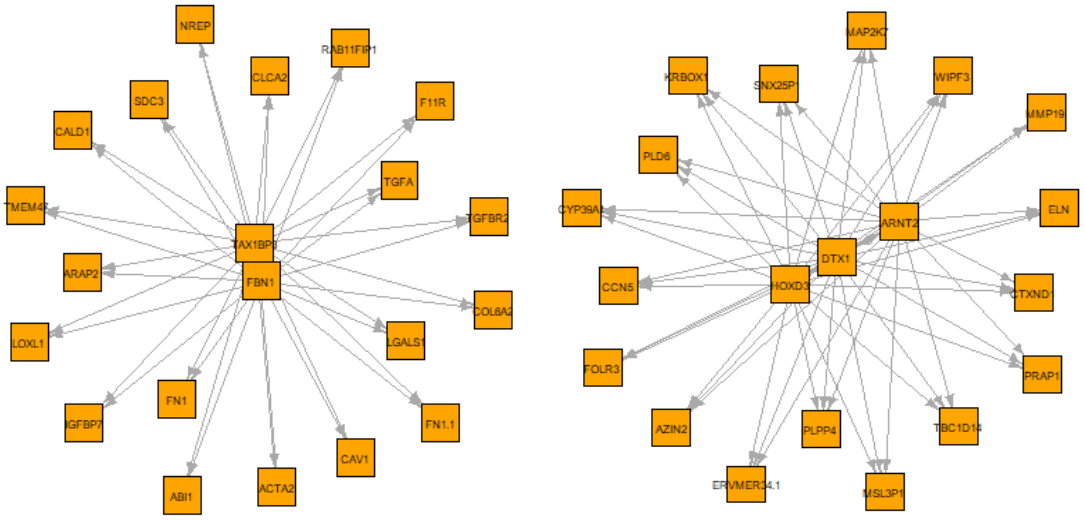
(c) Module *yellowgreen* (GSE20347)

(d) Module *white* (GSE20347)

(e) Module *salmon4* (GSE20347)

(f) Module *purple* (GSE20347)

Fig. 3.8: GRN for normal modules a) *skyblue* and b) *white* in GSE20347, disease modules c) *yellowgreen*, d) *white* e) *salmon4*, and f) *purple* in GSE20347



(a) Module *brown2* (GSE23400)

(b) Module *lightcyan* (GSE130078)

Fig. 3.9: GRN for normal modules a) *brown2* in GSE23400 and b)*lightcyan* in GSE130078

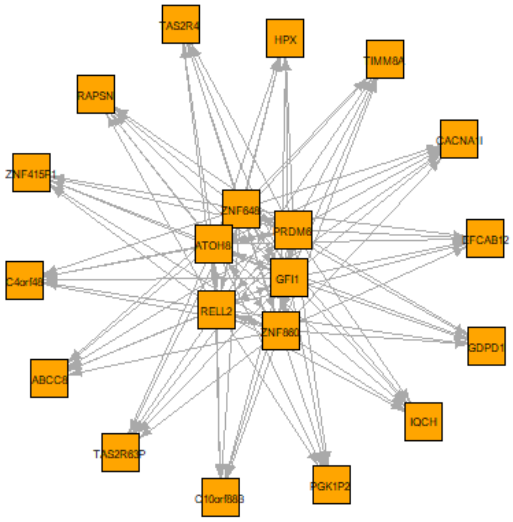


Fig. 3.10: GRN for disease module *orange* in GSE130078

We use DAVID [628, 253](Section 2.2.3) to carry out functional enrichment analysis (Section 2.4.1) of all BCGs found by our method, same as we did for module validation. If a BCG is annotated to at least one GO term in that database with significance of 5% ($p \leq 0.05$), it is considered to be enriched for GO in terms of the GO databases (GO_BP, GO_CC, and GO_MF). The BCGs annotated to the top three GO terms in each GO database, namely BP, CC, and MF, in all three ESCC datasets, GSE20347, GSE23400, and GSE130078, are summarized in Tables 3.7, Table 3.8, and Table 3.9, respectively. Similar to this, a BCG is enriched for KEGG pathways, if it is annotated to at least one term with a significance of 5% ($p \leq 0.05$). The BCGs annotated to the top 5 enriched KEGG pathways in GSE20347, GSE23400, and GSE130078 are included in the Table 3.10.

Tab. 3.7: Summary of BCGs detected by BicGenesis in the microarray dataset, GS20347, that are annotated to top 3 GO terms in the three GO databases.

	GO Term	Annotated BCGs
GO_BP	GO:0007165 signal transduction	*FEZ2, SKAP2, AREG, ARHGAP4, ARHGAP11A, KSR1, TNFRSF25, SRI, SH2B2, TNFSF18, GNRH2, NCR1, PLAU, GNB5, OPN1SW, CD4, PNOC, ING2, GRB10, TLE2, MAP2K1, RPS6KA3, TRIP10, HDGFL3, PDE4A, RASA1*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*SETX, MECOM, SS18, MED21, SPDEF, RORA, NPAT, SBNO2, RFXANK, PIN1, ITGA6, E2F1, CCNC, PCBP1, SMARCA5, HTATIP2, UBE3A, POU3F2, HMGA2, ETV1, ELF3, USF2, MTF2, PFKM, RPS6KA3, NODAL, TOP2A, FEZF2. RHOQ, UBC, BPTF, CKAP2, RBMX*
	GO:0016032 viral process	*SET, HPX, ABI2, IPO5, DYRK1A, MSH6, TNPO1, E2F1, HNRNPA1, HLA-C, ANP32A, SMU1, HTATIP2, UBE3A, BUB1, BNIP3*
GO_CC	GO:0005829 cytosol	*SET, LARP1, MSH6, RFXANK, TNFRSF25, SRI, PDCD4, ECI2, CES3, DYNC1I2, CARD14, CNN3, CA6, ELF3, TES, PHF20, RRP12, OSTM1, USP8, MAP2K1, TPM1, HMMR, COPS8, TNPO1, NFU1, RPS6KA3, PDCD10, PPP2CB, ATAT1, COPG1, UBC, BUB3, BUB1, PADI2, MCTS1, ARF6, SAMD4B, KPNA4, SOCS5, GFPT1, ARHGAP11A, DHODH, UBE2L3, SNX3, PCBP1, RAB11FIP3, MYLPF, CETN2, UBE2G1, GCKR, TAB2, BLVRA, ASB1, RNF114, RPL35, RPL38, VPS26A, RPL41, ABI2, GRB10, CHP1, CLIP3, PPP6C, PFKM, METAP2, SIAH2, TRIP10, SDCBP, ZC3H15, LMOD1, PDE4A, HPGD, GMDS, POLQ, SKAP2, ARHGAP4, SULT1C2, MECOM, HIKESHI, ODF2, RCHY1, EHBP1, EIF3K, EIF3G, ACTN3, SMARCA5, MYO5A, CALR, HTATIP2, ZFAND6, UBE3A, SRP14, UBE2A, GNB5, UBE2M, DACT1, HNRNPC, ARPC1A, TRAPPC4, AOPEP, POLR1D, MTFR1, SELENBP1, RAB14, DERA, LSAMP, TRAPPC2L, AMACR, CASP2, RASA1, MYLK, PSMD4, MYL1, BHMT, FERMT2, PSMA2, VAMP3, YKT6, PSMC1, RUVBL1, PIN1, KSR1, SH2B2, CYB5A, RPL23A, KIF1A, CTTN, EXOSC8, PAK3, CCT4, CALD1, PICALM, ARHGEF10L, PDLIM5, STMN2, ADK, SRP72, AGK, B2M, NMBR, HILPDA, UCHL5, RHOQ, RANBP9, HDGFL3, PDZD2, CKAP2*
	GO:0005654 nucleoplasm	*ZC3H11A, SET, MSH6, MSH2, SRSF9, RFXANK, SRI, E2F1, TRIM33, ELF3, BNIP3, USF2, PHF20, MTF2, PCGF1, OIP5, DYRK1A, COPS8, NFU1, RPS6KA3, UBC, BUB3, BUB1, RBMX, SETX, NOX4, KPNA4, NPAT, MCM10, SUMO2, DBF4, DHODH, UBE2L3, PCBP1, RAB11FIP3, CETN2, GCKR, HMGA2, TAB2, TAF1C, TAF1A, ABI2, MACROH2A1, PPP6C, SIAH2, TRIP10, SDCBP, PDE4A, HPGD, POLQ, WTAP, SKAP2, MECOM, NUDT21, HIKESHI, COMMD10, RCHY1, EHBP1, HNRNPA1, SMARCA5, POU3F2, UBE2A, UBE2M, DACT1, SNTB2, NRBF2, HNRNPC, TLE2, SNRNP27, NPIPA1, RFC3, POLR1D, POLR2B, ASCC1, PCLAF, DERA, TOP2A, BPTF, PSMD4, FERMT2, PSMA2, SRSF11, PSMC1, RORA, RUVBL1, PIN1, CCNC, TMPRSS2, EXOSC8, CCT4, RPA3, YTHDC1, AFF4, ADK, ING2, TIPIN, ANP32A, HILPDA, UCHL5, RANBP9, HDGFL3, STN1*
	GO:0016020 membrane	*SPINT1, PDGFA, LARP1, MTCH2, MSH2, ATP10B, SRI, HNRNPA1, NTRK3, EIF3K, ECI2, MYO5A, CALU, CALR, HTATIP2, CANX, CD4, SNTB1, SNTB2, HNRNPC, IPO5, HMMR, SELENBP1, POLR2B, NECTIN1, HLA-C, MLF2, CASP2, BUB1, RBMX, EDEM2, ARF6, SLC16A1, ATRNL1, PSMC1, RUVBL1, KSR1, CYB5A, BCAP29, PCBP1, PICALM, HSDL2, GORASP2, PDLIM5, CLCN4, STMN2, RPL35, DIP2A, RPN2, AGK, RHAG, ITGA2B, CDH3, PFKM, B2M, AGFG2, CEACAM1, TRIP10, TMEM131, ADIPOR1, SDCBP, LMOD1, PDE4A, ABCC5*

Continued on next page

	GO Term	Annotated BCGs
GO_MF	GO:0005515 protein binding	*AGTR2, SET, LARP1, ADAM22, SPDEF, MTCH2, MSH6, MSH2, RFXANK, SRI, IL2RA, PDCD4, TNFSF18, DYNC1I2, PLAU, SDHC, ELF3, CD4, TES, DAAM2, OSTM1, OIP5, IL13RA1, MAP2K1, COPS8, APOC1, NFU1, RPS6KA3, SPCS1, COPG1, UBC, BUB3, BUB1, RBMX, MCTS1, COL1A1, SETX, AREG, DLD, ARF6, SOCS5, UBE2L3, PCBP1, RAB11FIP3, CETN2, CRIPT, UBE2G1, GCKR, TAF1C, TAF1A, ASB1, GCSH, RNF114, RPL38, VPS26A, RPL41, RNF138, ABI2, GRB10, CHP1, CLIP3, MACROH2A1, SIAH2, B3GNT2, SDCBP, ZC3H15, TGFB2, PDE4A, POLQ, WTAP, PDGFA, CIAO1, HIKESHI, CNPY2, COMMD10, RCHY1, NTRK3, EIF3K, EIF3G, GBA, NFIL3, ZFAND6, UBE3A, SRP14, POU3F2, UBE2A, UBE2M, DACT1, SNTB1, SNTB2, TLE2, SNRNP27, PLEKHM1, RFC3, MTFR1, CATSPER2, SELENBP1, P2RY6, RAB14, ATP6V0E1, TOP2A, APBA2, FCN1, TMEM38B, TRAPPC2L, SHISA6, RASA1, ETFDH, MYLK, HPX, VAMP3, SRSF11, RUVBL1, KSR1, SH3BP4, SH2B2, CYB5A, RPL23A, TSPAN6, KIF1A, ZFR, TMPRSS2, EXOSC8, PAK3, TMX1, YTHDC1, FUT2, CALD1, PICALM, AFF4, EAPP, PDLIM5, DIP2A, ING2, SRP72, RHAG, PTP4A3, SEMA3G, TPCN1, TIPIN, NOTCH2NLA, NODAL, HILPDA, UCHL5, CEACAM1, RHOQ, RANBP9, PDIA3*
	GO:0042802 identical protein binding	*PSMD4, PDGFA, SETX, COL1A1, ARHGAP4, SRSF11, SLC16A1, NUDT21, MCM10, SRI, SH3BP4, HNRNPA1, CCNC, ACTN3, TNFSF18, KIF1A, NFIL3, EXOSC8, POU3F2, BNIP3, CD4, HNRNPC, ABI2, ITGA2B, GRB10, OIP5, TPM1, DYRK1A, TPCN1, PFKM, NECTIN1, B2M, ZNHIT6, ZDHHC17, CEACAM1, TRIP10, APBA2, TMEM38B, PDIA3, ADIPOR1, SDCBP, CASP2, HPGD, GMDS, POLQ, WTAP, SSBP1, RBMX*
	GO:0003723 RNA binding	PSMD4, ZC3H11A, LARP1, SETX, RPS27L, SRSF11, SAMD4B, NUDT21, PSMC1, SRSF9, SUMO2, HNRNPA1, RPL23A, EIF3G, PDCD4, UBE2L3, PCBP1, MYO5A, CALR, MRPS28, ZFR, CANX, CCT4, SRP14, YTHDC1, R3HDM1, SNTB2, TES, GPATCH4, RPL35, RPL38, RPL41, HNRNPC, ADK, RRP12, SRP72, IPO5, TNPO1, POLR2B, ASCC1, METAP2, ANP32A, TOP2A, UCHL5, PDIA3, UBC, ZC3H15, SSBP1, RBMX,

Tab. 3.8: Summary of BCGs detected by BicGenesis in the microarray dataset, GSE23400, that are annotated to top 3 GO terms in the three GO databases.

	GO Term	AnnotatedBCGs
GO_BP	GO:0007165 signal transduction	*GMFG, ROR2, PTPRM, TEK, PHB1, ITK, GRN, DGKA, IPO7, TLE3, PDE4D, ARAP2*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*PPARGC1A, ATAD2, ATRX, FOXJ2, VDR, TCF4, MYBL2, NME2, RHOQ, MAPK3*
	GO:0045893 positive regulation of transcription	*PHB1, PPARGC1A, ATAD2, ROR2, ENY2, FOXJ2, ACTN1, NME2, WNT6, TCF4*
GO_CC	GO:0005829 cytosol	*RDH14, NMT2, NEK1, LGALS1, NBEAL2, FBXL6, APRT, TAX1BP3, ACOX2, CYTH1, SMG6, IPO7, BCL2, DNAJC6, DNAJB5, CSK, SPRR2D, TPX2, DIAPH1, BUB1, UCK2, CSE1L, KPNA2, SOCS5, PPP1R13L, RAB11FIP1, MINK1, CAVIN3, CAPN1, ARHGEF40, VDR, LSM14A, EPS8L1, PRRC2A, ABI1, PNISR, DHCR24, KRT16, PDE4D, HSPD1, FTO, G3BP1, ACP1, MYO1D, ACTN1, PPIG, UBE2C, ACTA2, ACTC1, CLCA2, PYCR3, TUBB, PFDN2, MAPK3, PSME3, EXTL2, USH1C, SAP18, PAICS, EXOSC4, TOR1AIP1, CALD1, AURKA, MAPK13, RAD21, ITK, CLTB, DGKA, GYG1, MAP7, ASL, NME2, NME1, MYH11, TGFBR2, CAMK2G, SPRY2*
	GO:0005654 nucleoplasm	*RDH14, ATAD2, NEK1, PKP3, SRSF9, PPWD1, APRT, MYBL2, WDR74, ARPP19, IPO7, BCL2, MCM4, TPX2, BUB1, PPARGC1A, CSE1L, KPNA2, PPP1R13L, FOXJ2, VDR, PRRC2A, PNISR, ENY2, PUF60, ZC3H13, MRGBP, LUC7L3, FTO, ATRX, CAPG, BOP1, PPIG, UBE2C, PHB1, TLE3, CLCA2, MAPK3, PSME3, GRPEL1, EXTL2, SAP18, ID3, EXOSC4, NLE1, AURKA, RAD21, ESF1, CAMK2G*
	GO:0016020 membrane	*RDH14, ITGA4, SDC3, NBEAL2, VEGFB, EHBP1L1, PLD1, CAV1, ACTC1, PHB1, GRN, IPO7, RIMS2, BCL2, LBP, MCM4, MFSD5, BUB1, PSME3, RIPK4, CSE1L, KPNA2, GANAB, RAB11FIP1, PAICS, CAPN1, TOR1AIP1, CYRIA, RAD21, PRRC2A, DGKA, GYG1, DHCR24, PSD4, ARL6IP5, NME1, PDE4D, HSPD1, CAMK2G, SPRY2*
GO_MF	GO:0005515 protein binding	*NEK1, PKP3, SON, LGALS1, SDC3, TGFA, APRT, TAX1BP3, PLD1, MYBL2, TEK, DENND2B, BCL2, MCM4, CSK, DIAPH1, BUB1, SOCS5, PPP1R13L, VDR, PRRC2A, F12, ABI1, ENY2, PUF60, ELMO3, ZC3H13, PDE4D, LCAT, HSPD1, F11R, FN1, SYBU, LCN1, ACP1, ATRX, KCTD12, VEGFB, PPIG, UBE2C, GRN, TLE3, RAB25, CALML3, PYCR3, MAPK3, EXTL2, USH1C, SAP18, PAQR3, ID3, EXOSC4, FUT2, CALD1, AFF1, AURKA, MAPK13, RAD21, ITK, CLTB, CYB561, GYG1, RHOBTB1, IGFBP7, NME2, NME1, RHOQ*
	GO:0042802 identical protein binding	*PSME3, GRPEL1, FN1, UCK2, PPP1R13L, MFF, FOXJ2, LGALS1, TRIM29, KCTD12, SDC3, PAICS, PTPRM, VEGFB, CD84, CAV1, TCF4, TEK, BCL2, PUF60, CSK, ASL, NME2, NME1, CAMK2G, MAPK3*
	GO:0003723 RNA binding	*PPARGC1A, KPNA2, SRSF9, G3BP1, SON, GANAB, SAP18, LGALS1, KCTD12, BOP1, PPIG, LSM14A, RBM25, PRRC2A, SMG6, GRN, PNISR, PUF60, ESF1, NME1, ZC3H13, DIAPH1, LUC7L3, HSPD1*

Tab. 3.9: Summary of BCGs detected by BicGenesis in the bulk RNA-Seq dataset, GS130078, that are annotated to top 3 GO terms in the three GO databases

	GO Term	AnnotatedBCGs
GO_BP	GO:0007165 signal transduction	*LRRD1, IL1RAPL2, CCN5, OMP, CACNA1I, MAP2K7, PRKG2, LTB, CMTM3*
	GO:0000122 negative regulation of transcription from RNA polymerase II promoter	*PRDM6, GSC, TLE6, GFI1, DNAJC17*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*IL6, HOXD3, STRA8, ARNT2*
GO_CC	GO:0005886 plasma membrane	*ICAM4, CNR2, SLCO6A1, EVA1A, LTB, IL1RAPL2, TNFRSF13C, FCER2, ERVMER34-1, NECAB2, RAPSN, RELL2, COL23A1, AMBP, SLC13A3, NLRP6, KCNMB2, DIRAS3, CCR8, HSPA1L*
	GO:0005829 cytosol	*AMDHD1, BBS1, FBXL8, WIPF3, SMN1, MAP2K7, PRKG2, KLHL14, SULT2A1, RAPSN, KRT77, DTX1, RNF208, CMTM3, OMP, TCAP, TBC1D14, ADCY10*
	GO:0005654 nucleoplasm	*SMN1, DOC2A, RBM34, STAG3, ARNT2, DTX1, RNF208, TBC1D14, ATOH8, HOXD3, HSPA1L*
GO_MF	GO:0005515 protein binding	*BBS1, FBXL8, PLD6, MAP2K7, CLEC18A, HP, VPS37D, CHIA, ELN, STRA8, RNF175, SGCA, NECAB2, RELL2, DNAJC17, RNF208, MCMDC2, CFHR3, CACNA1I, TLE6, SLC13A3, GFI1, LIME1*
	GO:0042802 identical protein binding	*PRDM6, TIMM8A, COL23A1, SMN1, PRKG2, CBY2, PLPP4, NECAB2, ABCG4, HCN4*
	GO:0004712 protein serine/threonine/tyrosine kinase activity	*PRKG2*

Tab. 3.10: Summary of BCGs detected by BicGenesis in the two microarray and one bulk RNA-Seq datasets that have been annotated to the top 5 KEGG enriched pathways

	KEGG Pathways	Annotated BCGs
GSE20347	hsa05200:Pathways in cancer	*PDGFA, MECOM, ITGA2B, FGF5, IL13RA1, MAP2K1, MSH6, MSH2, ITGA6, E2F1, IL2RA, TGFB2, GNB5*
	hsa04010:MAPK signaling pathway	*RPS6KA3, PDGFA, AREG, MECOM, FGF5, MAP2K1, TGFB2, TAB2, RASA1*
	hsa05169:Epstein-Barr virus infection	*PSMD4, B2M, HLA-C, CALR, PSMC1, PDIA3, TAB2, E2F1*
	hsa04151:PI3K-Akt signaling pathway	*PDGFA, COL1A1, AREG, PPP2CB, ITGA2B, FGF5, MAP2K1, GNB5, ITGA6, IL2RA*
	hsa05171:Coronavirus disease - COVID-19	*RPL35, RPL38, RPL41, RPS27L, TMPRSS2, TAB2, RPL23A*
GSE23400	hsa05200:Pathways in cancer	*FN1, BCL2, CALML3, TGFA, VEGFB, PLD1, WNT6, TGFBR2, CAMK2G, MAPK3*
	hsa04010:MAPK signaling pathway	*TGFA, VEGFB, TGFBR2, MAPK3, MAPK13, TEK*
	hsa04151:PI3K-Akt signaling pathway	*TGFA, FN1, VEGFB, BCL2, ITGA4, COL6A2, MAPK3, TEK*
	hsa05169:Epstein-Barr virus infection	*BCL2, MAPK13*
	hsa05171:Coronavirus disease - COVID-19	*MAPK3, MAPK13*
GSE130078	hsa01100:Metabolic pathways	*AMDHD1, PRDM6, OXCT2, CHIA, ADCY10, GDPD1, AZIN2, B3GAT1*
	hsa05200:Pathways in cancer	*ARNT2, IL6*
	hsa04144:Endocytosis	*WIPF3, HSPA1L, VPS37D, FOLR3*
	hsa04010:MAPK signaling pathway	*HSPA1L, CACNA1I, MAP2K7*
	hsa05165:Human papillomavirus infection	

3.6.3 Literature Trace

The literature traces that BicGenesis found to be associated with ESCC in all three datasets are listed below.

- Chen et al.[86] found that overexpression of acylglycerol kinase (AGK) sustained constitutive JAK2/STAT3 activation, consequently promoting the cancer stem cell population and augmenting the tumorigenicity of ESCC cells both in vivo and in vitro and suggets the same as a prognostic biomarker and therapeutic target.
- Cao et al. [63] demonstrate the pro-metastatic function of ATPase family AAA domain containing 2 (ATAD2) and uncovered the new molecular mechanism by regulating C/EBPβ/TGF-β1/Smad3/Snail signaling pathway, thus providing a potential target for the treatment of ESCC metastasis.
- Studies by Ma et al.[478] indicate that BCL2 interacting protein 3 (BNIP3) exerts prodeath effects through the induction of caspase-independent apoptosis under hypoxia in ESCC, though BNIP3-induced autophagy acting as a survival mechanism.
- Cancer susceptibility candidate 8 (CASC8) was found by Wu et al. [773] to have an oncogenic role in the development of ESCC, suggesting that CASC8 may someday serve as a predictive biomarker in ESCC.
- Ando et al.[26] found that the ESCC patients with positive staining for caveolin-1 (CAV1) had significantly shorter survival than those with negative staining and thus CAV1 is a potential prognostic marker of ESCC. According to Kato et al., [301], over-expression of CAV1 is associated with lymph node metastasis and a worse prognosis after surgery in ESCC. Jia et al.[283] found that down-regulation of stromal CAV1 expression in ESCC had high malignant potential and suggests that it could be a powerful prognostic marker for patients with ESCC.
- Results presented by Ochi et al [527] suggest that the expression of carbonic anhydrase XII (CA12) may be a valuable prognostic factor for patients with advanced ESCC.
- Results by Fang et al.[152] indicate that targeting chaperonin containing TCP1 complex 4 (CCT4) may be a therapeutic target in ESCC patients, which provides a theoretical basis to enhance the sensitivity of DDP in ESCC.
- Results of the study done by Qian et al.[558] demonstrate a critical role of CEA cell adhesion molecule 1 (CEACAM1) in angiogenesis of ESCC progression, thus

suggesting that CEACAM1 could be a potential therapeutic target for ESCC.

- Li et al.[376] considers collagen type I alpha 1 chain (COL1A1) as novel potential diagnostic and prognostic biomarkers in patients with ESCC.
- The data presented by Luo et al.[460] suggest that cortactin gene (CTTN, also EMS1) is an oncogene in the 11q13 amplicon and exerts functions on tumor metastasis in ESCC. Hsu et al.[243] observed that CTTN overexpression in early and late stages of human ESCCs and carcinogen-induced murine ESCCs, suggesting a role for cortactin in esophageal carcinogenesis.
- Diacylglycerol kinase alpha (DGKA) is involved in the progression of ESCC, according to Chen et al. [76], who suggest DGKA as a viable target for ESCC treatment.
- Analysis of clinical data by Liu et al.[434] indicate that desmoglein-2 (DSG2) levels were significantly associated with patient age and histological grade in ESCC and may be a diagnostic biomarker for ESCC.
- Ebihara et al.[146] found that over-expression of E2F transcription factor 1 (E2F1) is associated with tumor progression and a worse prognosis after surgery in ESCC. Li et al.[361] also found that E2F1 displays a remarkable potential value for ESCC prognosis, which has improved our understanding of the molecular pathology of E2F1, thus providing a possible therapeutic target for ESCC treatment.
- According to Iwabu et al. [273], ibroblast growth factor 5 (FGF5) methylation is a sensitive marker of ESCC to definitive chemoradiotherapy.
- Ma et al.[472] suggest potential target for the treatment of ESCC as they establish that high expression of FN1 protein in ESCC tumor tissue is an independent poor prognostic factor.
- Results presented by Lau et al.[331] establish the significance of follistatin Like 1 (FSTL1) in driving oncogenesis and metastasis in ESCC by coordinating NFκB and BMP pathway control, with implications for its potential use as a diagnostic or prognostic biomarker and as a candidate therapeutic target in this disease.
- Zhao et al [890] found that knockdown of fat mass and obesity associated (FTO) drastically suppressed the proliferation, migration, and invasion of ESCC cells.
- According to Huang et al. [260], growth factor-independent 1 (GFI1) may be a useful target for ESCC therapy because it indicated how SOCS1 expression was inhibited by GFI1, which allowed ESCC cells to proliferate and migrate more freely.
- Palumbo et al.[538] present high mobility group A 2 (HMGA2) abrogation attenuated

the malignant phenotype of two ESCC cell lines, suggesting that HMGA2 overexpression is involved in ESCC progression.

- Li et al.[373] increased insulin-like growth factor binding protein 7 (IGFBP7) may accelerate ESCC progression by promoting the expression of TGFβ1, α-SMA, and collagen I by activating the TGFβ1/SMAD signaling pathway.
- According to Li et al.[368], importin 5 (IPO5) expression significantly increased in ESCC tissues, which was associated with pathological staging and poor prognosis of ESCC patients and may promote it's malignant progression.
- Kwon et al.[324] present findings that suggest that integrin alpha 6 (ITGA6) plays an important role in tumorigenesis in ESCC and represents a potential therapeutic target in the treatment of ESCC. Ma et al. [466] found that decreased ITGA6 attenuates motility of malignant cells partially through deactivating Akt pathway, which is essential for ESCC cells motility.
- Abbaszadegan et al. [2] found that potassium channel tetramerization domain containing 12 (KCTD12) may exert its inhibitory role in ESCC through the suppression of WNT /NOTCH, stem cell factors, and chromatin remodelers and can be introduced as an efficient therapeutic biomarker.
- He et al [230] found that kallikrein-associated peptidase 11 (KLK11) plays a key role in inhibiting ESCC carcinogenesis and progression and became a potential biomarker for poor prognosis in patients with ESCC.
- Karyopherin alpha 2 (KPNA2) protein levels were shown to be elevated in ESCC tumours, according to Ma et al. [475], and siRNA against KPNA2 was able to limit the proliferation of ESCC cells, suggesting that it may be a novel potent marker and therapeutic target for ESCC. Sakai et al. [596] added that KPNA2 expression is connected to ESCC tumour proliferation, tumour invasiveness, and poor differentiation.
- Tian et al.[683] found that aberrant overexpression of minichromosome maintenance 10 replication initiation factor (MCM10) facilitated the proliferation and metastasis abilities of ESCC cells in vitro and in vivo by inducing DNA over-replication and genomic instability, providing functional evidence to support their population finding that high expression of MCM10 is extensively presented in tumor tissues of ESCC and correlated with inferior survival outcomes.
- Chen et al.[91] suggest that NADPH oxidase 4 (NOX4) overexpression is a poor prognostic factor for patients with ESCC undergoing curative esophagectomy.

- According to Bai et al [37], progestin and adipoQ receptor family member 3 (PAQR3) expression is an independent prognostic indicator for patients with ESCC and may serve an important role in the progress of ESCC and become a potential candidate for ESCC targeted therapy. Bai et al [36] found that PAQR3 is epigenetically silenced in ESCC and restoration of PAQR3 suppresses the aggressive phenotype of ESCC cells and may represent a potential target for the treatment of ESCC.
- Findings by provide a new perspective for understanding the molecular mechanism of esophageal carcinogenesis, and poly(rC) binding protein 1 (PCBP1) is a promising therapeutic target.
- According to Han et al. [212], platelet-derived growth factor A (PDGFA) may serve as an oncogene in ESCC and represent an independent molecular biomarker for prognosis of ESCC patients.
- Meta-analysis done by Guo et al. [207] indicate that high programmed cell death 1 ligand 1 (PD-L1) expression in ESCC is associated with distant metastasis and reduced overall survival.
- Experimental evidence that peptidylprolyl cis/trans isomerase, NIMA-interacting 1 (PIN1) knockdown inhibited proliferation and clonogenicity of ESCC in vitro and tumorigenesis of ESCC in vivo was provided by Lin et al.[404].
- Results presented by Li et al.[352] suggest that high abundance of DNA polymerase theta (POLQ) in ESCC contributes to the malignant phenotype through genome instability and activation of the cGAS pathway.
- A study by Wang et al. [729] indicates a functionally significant regulation mechanism of POTE ankyrin domain family, member G (POTEG) in the aetiology of esophageal cancer, suggesting possible application in the treatment and intervention techniques for ESCC. Findings by Li et al. [374] suggests that POTEG plays a crucial role in the aetiology of ESCC and may serve as a biomarker.
- According to Tong et al. [686], Rab25 may provide a prognostic biomarker for ESCC outcome prediction and a novel therapeutic target in ESCC treatment.
- Tang et al. [673] suggest that semaphorin 3B (SEMA3B) is an important tumor-suppressor gene in the malignant progression of ESCC, as well as a valuable prognostic marker for ESCC patients. Dong et al. [142] suggests SEMA3B as tumor suppressors and may serve as potential targets for antitumor therapy.
- Xie et al. [786] indicate that semaphorin 3F (SEMA3F) serves as a potential prognos-

tic biomarker and tumor suppressor of ESCC and may be involved in the lymph node metastasis development through regulating neuropilin 2.

- Xia et al. [780] shows that solute carrier family 39 member 4 (SLC39A4) knockdown impaired the proliferation and motility capacities of ESCC cells and suggest that it could serve as a novel prognosis biomarker to promote ESCC progression.
- Yang et al. [813] demonstrate the novel function of suppressor of cytokine signaling (SOCS5) in ESCC prognosis and suggests that its's expression could serve as a novel therapeutic biomarker for improving the prognosis of ESCC.
- Li et al. [384] suggest that transforming growth factor alpha (TGFA) as well as three other genes may be associated with angiogenesis, and the progression and metastasis of ESCC.
- According to Ma et al. [476], the high level of methylated CpGs in TGF-beta receptor type II (TGFBR2) in ESCC suggests that DNA methylation in TGFBR2 promoter region would contribute to absent or reduced TGFBR2 mRNA expression, and hence promote ESCC carcinogenesis.
- According to Liu et al. [416], overexpression of TPX2 may be risk factor of lymph node in esophageal carcinoma, and maybe a potential biomarker for early diagnosis and prognosis of ESCC. Hsu et al. [244] demonstrated that TPX2 expression is associated with cell proliferation and poor prognosis among patients with resected ESCC.
- According to Li et al. [364], ubiquitin conjugating enzyme E2 C (UBE2C) mRNA and protein level were highly expressed in ESCC and UBE2C likely plays different roles in different stages of the ESCC. Furthermore, Palumbo et al.[537] reports that UBE2C affects proliferation rates and cell cycle profile of ESCC cell lines, by directly interfering with cyclin B1 protein levels, suggesting its involvement in crucial steps of ESCC carcinogenesis.
- Luo et al.[457] suggest that the WT1 Associated Protein (WTAP), a potential biomarker of ESCC, maybe play an important role in ESCC-genesis through regulating expression of genes related to cell proliferation, migration and apoptosis. Furthermore, Luo et al.[461] identified the significant role of WTAP-catalysed m6AMo in ESCC tumourigenesis, wherein it facilitates ESCC tumour growth and metastasis.
- Zheng et al. [903] demonstrate that cytochrome P450 Family 3 Subfamily A Member 5 (CYP3A5) downregulation, resulting in ZEB2 activation, promoted ESCC invasion and migration.

Tab. 3.11: Summary of potential biomarkers identified by BicGenesis. Here, All 3 under GO databases imply all three databases, BP, CC, and MF.

	BCG	GO Database	Enriched Cancer Pathway(s)	TF?	Literature Evidence
GSE20347	*AMACR*	All 3	hsa01100, and hsa04146	No	OSCC [227]
	AGK	All 3	hsa01100	No	ESCC [86], OSCC [414]
	AFF4	All 3	Nil	No	HNSCC [131]
	ANP32A	All 3	Nil	Yes (Fig. 3.8e)	OSCC [711]
	BNIP3	All 3	hsa05131, hsa04068, hsa05134, hsa04140, and hsa04137	No	ESCC [478]
	COL1A1	All 3	hsa04151, hsa05165, hsa04510, hsa05205, hsa05146, hsa04933, hsa04512, hsa05415, and 2 others	No	ESCC [376]
	CCT4	All 3	Nil	Yes (Fig. 3.8a)	ESCC [152], HNSCC [141]
	CDH3	All 3	hsa04514	No	OSCC [776]
	CEACAM1	All 3	Nil	No	ESCC [558], LaSCC [456], HNSCC [664, 803], OSCC [634]
	CTTN	All 3	hsa05205, hsa05131, hsa05130, hsa05100, and hsa04530	No	ESCC [460, 243], OSCC [438], HNSCC [238]
	CYP3A5	All 3	hsa01100	No	ESCC [903]
	DSG2	All 3	hsa05412	No	ESCC [434]
	DYRK1A	All 3	Nil	No	OSCC [488], HNSCC [566]
	DHODH	All 3	hsa01100	No	ESCC [559]
	DLX5	All 3	hsa04550	Yes (Fig. 3.8d)	OSCC [867]
	E2F1	All 3	hsa05200, hsa05222, hsa05215, hsa05212, hsa05220, hsa05225, hsa05226, hsa05223, and 17 others	Yes (Fig. 3.8d)	ESCC [146, 361]
	FGF5	BP,CC	hsa05200, hsa05224, hsa05218, hsa05226, hsa05207, hsa04151, hsa04810, hsa04020, and 3 others	No	ESCC [273]
	FUT6	BP,CC	hsa01100, and hsa00601	No	HNSCC [732]
	FSTL1	All 3	Nil	No	ESCC [331]
	HMGA2	All 3	hsa05202	Yes (Fig. 3.8f)	ESCC [538], OSCC [598]
	HMMR	BP,CC	hsa04512	No	HNSCC [455]
	IL2Rα	All 3	hsa05200, hsa04151, hsa04015, hsa05166, hsa04630, hsa04060, hsa04640, hsa04659, and 3 others	No	HNSCC[321]
	IPO5	All 3	hsa03013	No	ESCC[368]
	ITGA6	All 3	hsa05200, hsa04151, hsa05165, hsa04510, hsa05222, hsa05145, hsa04810, hsa04640, and 5 others	No	ESCC [324, 466], OSCC [855], HNSCC [156]
	KIF1A	All 3	Nil	No	HNSCC [129]
	MAP2K1	All 3	hsa05200, hsa05205, hsa05210, hsa05221, hsa05220, hsa05212, hsa05211, hsa05225, and 77 others	No	HNSCC[279]

Continued on next page

	BCG	GO Database	Enriched Cancer Pathway(s)	TF?	Literature Evidence
	MCM10	All 3	Nil	No	ESCC[683]
	MYO5A	All 3	hsa05130	No	LaSCC[898]
	MCTS1	All 3	Nil	No	LaSCC[811]
	NOX4	All 3	hsa04933, hsa05022, hsa04936, hsa05010, and hsa05208	No	ESCC[91]
	PCBP1	All 3	hsa04216	Yes (Fig. 3.8b)	ESCC [548]
	PIN1	All 3	hsa04622	Yes (Fig. 3.8e)	ESCC[404], OSCC[503]
	POLQ	All 3	Nil	No	ESCC[352]
	RPN2	BP,CC	hsa01100, and hsa04141	No	LaSCC[910], HNSCC[82]
	RAB14	All 3	hsa04152	No	OSCC[385]
	RUVBL1	All 3	hsa04310	Yes (Fig. 3.8c)	HNSCC [402]
	SDCBP	All 3	Nil	No	HNSCC[500]
	SELENBP1	All 3	hsa01100	No	OSCC[849]
	SPDEF	All 3	Nil	No	HNSCC[748]
	TAB2	All 3	hsa04010, hsa05169, hsa05171, hsa05417, hsa05161, hsa05162, hsa04380, hsa05135, and 12 others	No	HNSCC[418]
	TPM1	All 3	hsa04261, hsa05410, and hsa05414	No	OSCC[539, 662]
	UBE2L3	All 3	hsa05022, hsa05012, and hsa04120	No	OSCC[115]
	WTAP	All 3	Nil	No	ESCC[457, 461]
	YKT6	All 3	hsa04130	No	OSCC[822]
	YTHDC1	All 3	Nil	No	HNSCC[763]
GSE23400	*ACTN1*	All 3	hsa04510, hsa05146, hsa04810, hsa05203, hsa05131, hsa04520, hsa04670, and hsa04530	No	OSCC [784], HNSCC[734]
	ATAD2	All 3	Nil	No	ESCC [63], OSCC [741]
	AURKA	All 3	Nil	No	OSCC [123], HNSCC [770]
	CAV1	All 3	hsa04510, hsa05205, hsa05418, hsa04144, hsa05020, hsa05416, and hsa05100	No	ESCC[26, 301, 283]
	CA12	All 3	hsa01100	No	ESCC [527]
	BUB1	All 3	hsa04110	No	LaSCC [383]
	DGKA	All 3	hsa01100, hsa04072, hsa05231, and hsa04070	No	ESCC [76]
	DIAPH1	All 3	hsa04510, hsa04933, hsa04810, and hsa05131	No	LaSCC [806, 805]

Continued on next page

BCG	GO Database	Enriched Cancer Pathway(s)	TF?	Literature Evidence
FN1	All 3	hsa05200, hsa04151, hsa05165, hsa04510, hsa05205, hsa05146, hsa05222, hsa04933, and 4 others	No	ESCC [472], HNSCC [675, 912], TSCC [794]
FTO	All 3	Nil	No	ESCC [890], OSCC [137], HNSCC [885]
FUT6	BP,CC	hsa01100, and hsa00601	No	HNSCC [732]
GPR87	BP,CC	Nil	No	LSCC [203]
G3BP1	All 3	Nil	No	OSCC [250]
HSPD1	All 3	hsa05417, hsa05152, hsa05134, and hsa04940	No	OSCC [298]
IGFBP7	All 3	Nil	No	ESCC[373, 259]
ITK	All 3	hsa04660, hsa04062, and hsa04670	No	TSCC [534]
KCTD12	All 3	Nil	No	ESCC [2]
KLK11	All 3	Nil	No	ESCC [230], LaSCC [895, 542]
KPNA2	All 3	hsa05164, hsa03013, and hsa05207	No	ESCC[475, 596], OSCC[403]
MINK1	All 3	Nil	No	OSCC [507, 506]
MRGBP	BP,CC	Nil	No	HNSCC [889]
PAQR3	All 3	Nil	No	ESCC [37, 36], LaSCC [771]
PDL1	Nil	Nil	No	ESCC [207]
PKP3	All 3	Nil	No	OSCC[684]
PPARGC1A	All 3	hsa04936, hsa04910, hsa04920, hsa04931, hsa04152, hsa05016, hsa04213, hsa04922, and 2 others	No	ESCC [93]
RAB25	All 3	Nil	No	ESCC [686], HNSCC [616]
SEMA3B	All 3	hsa04360	No	ESCC[673, 142]
SEMA3F	All 3	hsa04360	No	ESCC [786], OSCC [433]
SLC39A4	CC	hsa05010, and hsa05012	No	ESCC [780, 287]
SLPI	All 3	Nil	No	OSCC[762], HNSCC [62]
SOCS5	All 3	hsa04630, and hsa04917	No	ESCC [813]
SPRY2	All 3	Nil	No	HNSCC [401], OSCC[391]
TGF-α	All 3	hsa05200, hsa05215, hsa05210, hsa05212, hsa05225, hsa05223, hsa05214, hsa04010, and 6 others	No	ESCC [384], HNSCC [196]
TPX2	All 3	Nil	No	ESCC [416, 244]
TRIM29	All 3	Nil	No	LSCC [793]

Continued on next page

	BCG	GO Database	Enriched Cancer Pathway(s)	TF?	Literature Evidence
	UBE2C	All 3	hsa04120	No	ESCC [364, 537], HNSCC [290, 820], TSCC [426]
GSE130078	*AMBP*	All 3	Nil	No	OSCC [612]
	ARNT2	All 3	hsa05200	Yes (Fig 3.9b)	OSCC[306]
	CASC8	None	Nil	No	ESCC [773]
	CCN5	All 3	Nil	No	OSCC [656]
	CMTM3	All 3	Nil	No	LaSCC [626], OSCC[860]
	DTX1	All 3	Nil	Yes (Fig 3.9b)	HNSCC [184, 185]
	GFI1	All 3	Nil	Yes (Fig . 3.10)	ESCC [260]
	LINC00319	None	Nil	No	LaSCC [841]
	POTEG	None	Nil	No	ESCC [729, 374]

In Table 3.11, we give a detailed summary of all hub-genes that have been identified by BicGenesis as candidates for potential biomarkers or BCGs for ESCC. In our approach, we consider strong literature evidence for association with ESCC and five other SCCs related to ESCC as the necessary criterion for a BCG to be a potential biomarker, and the findings from literature are summarized in Table 3.11. In Table 3.11, we highlight the enriched GO terms and pathways to which the BCGs have been annotated. Further, it also details whether the BCG is a transcription factor (TF) or not.

3.7 Discussion

To identify potential biomarkers, we use our biomarker selection criteria (Section 2.5). A list of all the cases with BCG annotations is provided in the Table 3.12 . Due to prior evidences of correlation with ESCC in the form of earlier studies and both our statistical and biological validations of these genes, all BCGs that fall under Cases 1 and 2 are considered to be potential biomarkers for ESCC. Although there are considerable literature evidences for a link between Case 3 BCGs and ESCC, there is little support for Case 3 BCGs' biological significance. On the other hand, there is only literature evidence of association with other SCCs related to ESCC for BCGs that fit under Case 4, despite the fact that we substantially support their biological relevance to the respective datasets. The possibilities in both these examples can be thought of as likely prospective potential biomarkers, however more thorough investigation is required.

Tab. 3.12: Summary of potential ESCC biomarkers identified by BicGenesis using the biomarker criteria (Section 2.5)

	GSE20347	GSE23400	GSE130078
Case 1	E2F1, HMGA2, PCBP1, PIN1		
Case 2	AGK, BNIP3, COL1A1, CTTN, FGF5, IPO5, NOX4	CAV1, TGFA, FN1, SEMA3F, CA12, SOCS5, DGKA, PPARGC1A, SEMA3B, UBE2C, KPNA2	
Case 3	FSTL1, ITGA6, CCT4, MCM10, WTAP	IGFBP7, FTO, PAQR3, KLK11, RAB25, KCTD12, SLC39A4, ATAD2, TPX2	GFI1
Case 4	DLX5, RUVBL1		ARNT2

In GSE20347, four BCGs such as E2F1 , HMGA2, PCBP1 , and PIN1 (Table 3.12) belong to Case 1 are recommended as potential biomarkers for ESCC. In dataset GSE20347, seven BCGs uch as AGK, BNIP3, COL1A1, CTTN, FGF5, IPO5, and NOX4 (Table 3.12) belong to Case 2, and thus are recommended potential biomarkers for ESCC. Similarly, eleven BCGs in GSE23400, such as CAV1, TGFA, FN1, SEMA3F, CA12, SOCS5, DGKA, PPARGC1A, SEMA3B, UBE2C and KPNA2 are found and established as potential biomarkers for ESCC as they fall under Case 2. However, none of the BCGs in the RNASeq dataset GSE130078 fall under either Case 1 or Case 2 and thus are not potential biomarkers for ESCC.

Five BCGs in GSE20347 such as FSTL1, ITGA6, CCT4, MCM10, and WTAP fall in Case 3. In other words, although there are other works that establish their role as potential biomarkers for ESCC, the biological relevance to their respective datasets is not that strong. However, they may be further explored for their acceptability as probable biomarkers for ESCC. Similarly in GSE23400 and GSE130078, nine genes such as IGFBP7, FTO, PAQR3, KLK11, RAB25, KCTD12, SLC39A4, ATAD2, and TPX2 and one gene GFI1 (Table 3.12), respectively, fall under Case 3. DLX5 and RUVBL1 identified in GSE20347 and ARNT2 detected in GSE130078 on the other hand falls under Case 4. We establish their biological relevance as these BCGs are annotated to GO terms in all three GO databases as well as several enriched pathways. They further exhibit interesting regulatory behavior, but there are no previous works that relate them to ESCC. However, it is worth mentioning that there is literature evidence that indicate that DLX5 and ARNT2 has strong relevance to OSCC and RUVBL1 has strong relevance to HNSCC (Table 3.11) .

3.8 Chapter Summary

Our biclustering analysis framework, BicGenesis has been found capable of identifying several in crucial genes for ESCC. All eight biclustering methods such as Bimax [556] , x-MOTIF [514] , Plaid Models [333] , ISA [44], FABIA [237], QUBIC [348], iBBiG [208], and FLOC [808, 809] are able to generate relevant biclusters. These biclusters are neither too large nor too small in terms of both rows and colums. Through multiple iterations and exhaustive application we have determined the parametres for extraction of effective biclusters using each method. We perform DCA to observe the changes in behavior exhibited by the relevant biclusters under normal and disease conditions. The modules extracted are equivalent to the relevant biclusters. Using preservation analysis we determine that 19, 9 and 9 modules in the datasets viz GSE20347, GSE23400, and GSE130078, respectively are Modules of Interest (MoI). We identify top 20 hub-genes in each MoI and these hub-genes are considered Biomarker Candidate Genes (BCGs).

The biological relevance of each BCG for each dataset is evaluated based on (a) annotation to enriched GO terms in the GO databases, (b) annotation to enriched KEGG pathways, and (c) whether the BCG is a transcription factor in a gene regulatory network. Previous research that has either (a) established them as potential biomarkers for ESCC itself or (b) established them as potential biomarkers for five other SCCs related to ESCC, namely Oral SCC, Tongue SCC, Lung SCC, Head and Neck SCC, and Laryngeal SCC. With the help of prior literature works, our technique identified four BCGs—E2F1, HMGA2, PCBP1, and PIN1—that are Transcription Factors (TFs), have substantial biological relevance to their respective datasets, and may serve as ESCC biomarkers. Similarly, seven BCGs such as AGK, BNIP3, COL1A1, CTTN, FGF5, IPO5, and NOX4 and seven BCGs such as CAV1, TGFA, FN1, SEMA3F, CA12, SOCS5, DGKA, PPARGC1A, SEMA3B, UBE2C, and KPNA2 in datasets GSE20347, and GSE23400, respectively, are established as potential biomarkers for ESCC. No BCGs in dataset GSE130078 are established as potential biomarkers for ESCC as none fall under Case 1 or Case 2.

Next Chapter presents a consensus-based approach that incorporates six differential expression analysis methods for unbiased and integrative identification of Differentially Expressed Genes (DEGs) as potential biomarkers for critical diseases. To identify DEGs, we employ three microarray and three bulk RNA-Seq differential expression analysis methods. An effective consensus function has also been presented in the next chapter for identification of a set of unbiased yet biologically significant DEGs.