Chapter 4

Differential Expression Analysis

4.1 Introduction

Differentially Expressed Genes (DEG) information is a key step in the identification of biomarkers for a critical disease of interest. This is accomplished by Differential Expression Analysis (DEA), which accelerates the search for biomarkers by providing a candidate list of these discriminative candidate genes and tracks the behavior of each gene separately under normal and pathological settings. RNA sequencing (RNA-Seq) and DNA microarrays are essential data sources of DEA techniques. Although, microarray technology was one of the commonly used strategy, there are, however, certain built-in restrictions, such as the requirement for prior sequence knowledge for array design or the fact that cross-hybridization makes it challenging to analyze strongly correlated sequences. Major obstacles include the lack of reproducibility across platforms and laboratories as well as the lack of sensitivity to highly or lowly expressed genes. RNA sequencing technology overcomes these restrictions. In order to support both RNA-Seq and microarray gene expression data, many DEA methods have been introduced. Additionally, there are numerous datasets on crucial diseases that are accessible and interoperable with both technologies. We suggest a consensus-based integrative strategy that ensembles a few of these methods with the aim of achieving enhanced performance, while keeping in mind that the majority of methods created for these technologies are not beneficial for all circumstances.

4.1.1 Differential Expression Analysis (DEA)

In genomics and transcriptomics research, DEA approach has already been established as useful to find genes that are expressed differently in two or more situations or conditions. By comparing the levels of gene expression in several samples or groups of samples, DEA aims to find genes that are linked to a specific biological process or disease. The following steps are commonly used in the DEA process.

1. Pre-processing is the first step towards DEA and is specific to the input gene expression data. This step generally involves, removal of noise, normalization, and missing value estimation.
2. After listing of the conditions under which the experiment is carried out so as to observe the differential behavior exhibited each gene under these conditions the statistical test are performed. By employing various statistical tests such as t-tests [761, 615], ANOVA [168, 167], gene expression levels of the genes between conditions are analyzed.
3. Testing of numerous genes at once has certain drawbacks that can be overcome using multiple testing correction. Widely used methods for multiple testing correction such as Benjamini-Hochberg technique [43, 764], the Bonferroni adjustment [49], held in the reduction of the quantity of false positives.
4. As an indication of the degree of gene expression variation between conditions, the fold change for each gene is computed using $log2$ scale as the ratio of expression values across conditions.
5. To identify genes that are differentially expressed, ie, DEGs, a significance threshold is chosen. The most widely used significance value is 5% (i.e., *p-value*=0.05) or 1% (i.e., *p-value*=0.01).

Due to chance, technological turbulence, or confounding elements like batch effects or sample heterogeneity, DEA might result in false positives. The probability of false positives should be reduced by using appropriate statistical approaches and repeated testing correction procedures. In order to attain statistical power and accuracy, DEA needs a large enough sample size. In some circumstances, having fewer samples available may result in diminished power and higher false discovery rates. The assumption behind DEA is that gene expression levels are regularly distributed, which may not always be the case. Some datasets may not benefit from the normalization techniques used to take into account technological variation. Variability can be introduced by differences in sample preparation, sequencing equipment, or other technical issues, which could muddle the analysis. Technical variability can be reduced with the use of proper normalization techniques and careful experimental design. Even though DEA might pinpoint genes that are linked to particular biological pathways or activities, it could not fully explain the underlying mechanisms.

4.2 Related Works

Finding genes or other traits that demonstrate noticeable changes in expression levels across multiple contexts or groups is the aim of DEA. The identification of genes or traits that show notable differences in expression levels between various conditions or groups depends critically on statistical approaches, which offer a rigorous framework for doing so. The following are a few major implications of statistical techniques in DEA.

- Statistical approaches helps in distinguishing between significant changes in gene expression and random fluctuations as well as helps in ascertaining the statistical significance of these variations.
- Measures of uncertainty provided by statistical approaches, such as *p-values*, helps in the evaluation of the validity of the results.
- Statistical approaches combine multiple testing correction processes with the aim to reduce the probability of incorrectly classifying genes as DEGs.
- Statistical methods have taken into consideration distinct characteristic and assumptions corresponding to various gene expression data, such as microarray, bulk RNA-Seq, or scRNA-Seq to ensure reliable analysis and interpretation.
- Integration of statistical methods related to DEA into bioinformatics pipelines and software ensures simple and effective analysis across various studies.

4.2.1 Statistical Tests

Analysis of variance (ANOVA) [168, 167] is a parametric statistical technique for comparing the means of three or more groups or situations. Based on the variation within and across groups, it determines whether there are any significant variations in the group means. In ANOVA, with the presumptions of the normality of the data and the homogeneity of variances , the null hypothesis, H_0, that there are no variations in population means between groups or conditions is tested. The alternative hypothesis, H_1 asserts that at least one group mean differs from the rest. The variability between groups and the variability within groups are the two parts of the overall variability that was seen in the data after the ANOVA. While the within-group variability reflects variance within each group, the between-group variability represents variations in the group means. Two measures, a) within-group mean square (MSW) that reflects variation within groups and b) between-group mean square (MSB) that reflects the variation between groups is used to compute F-statistics. F-Statistics is the ratio of MSB to MSW. Critical value of F-distribution at a selected level of significance (e.g, $alpha = 0.05$) is compared to the F-statistics to determine its statistical significance. If the F-value > critical value, the null hypotheses, H_0 is rejected and it can be concluded that there is a significant difference between atleast one pair of group means. ANOVA is found effective in analyzing variations in the data under more than two groups.

The t-test [761, 615] is a statistical method that has the ability to determine significant differences between two groups through comparison of the means. With the assumption that data is normalized and the variances are homogeneous, t-test is a parametric test that assesses whether the differences observed in means is due to chance variability or are statistically significant. Assuming, the null hypothesis, H_0: There is no difference in the population means of the two groups and alternative hypothesis, H_1: There is a significant difference between the means of the two groups. The $t - value$, which is a test statistic based on the differences in means and variability within the groups, is compared to the critical values from the t-distribution for a pre-determined significance level (α). If the $t - value >$ critical value or $p - value < \alpha$, H_0 is rejected, implying that there isa significant difference between the means of the two groups.

4.2.2 Microarray Methods

Linear Models for Microarray Analysis (Limma)[637, 638] is a widely used R/ Bioconductor package for analyzing microarray gene expression data. Limma is a sophisticated statistical framework based on linear models that identify differentially expressed genes across experimental groups. It's ideal for analyzing studies with a small number of replicates, which is a common circumstance in genomics research. Limma includes functions to accomplish pre-processing and normalization of raw microarray data. Limma's linear model requires the design matrix as an input. The design matrix incorporates information regarding treatment conditions, time points, and other important elements to define the experimental design and sample groups of interest. Limma uses the design matrix to fit a linear model to the expression data of each gene. This model takes into account both within-group and between-group variation. Limma's moderated t-statistic approach reduces gene-wise variance estimations to a common value to improve stability and robustness. Because gene expression study often entails assessing thousands of genes at the same time, it is critical to account for the issue of multiple comparisons. Limma includes numerous ways for changing p-values, including the commonly used Benjamini-Hochberg [43, 764] method, to control the false discovery rate (FDR).

Tusher et. al [697] developed Significance Analysis of Microarrays (SAM) as an alternative to widely used t-test [761, 615] and fold-change approaches. It is necessary that the input data to SAM is divided into two or more conditions or groups. To measure the differences in gene expression under varying groups modified t-statistics incorporates both variability between and within groups and the mean difference. From randomized permutations of the data, SAM compares the observed test statistics with the distribution of test statistics to compute the significance of each gene. SAM allows estimation of FDR and ranks the genes based on the strength of evidence for differential expression. Significant genes or DEGs are determined through implementation of a threshold that controls the FDR to a desired level. SAM provides a list of DEGs and their statistical significance and fold-change values.

The simultaneous measurement of the expression levels of thousands of genes is made possible by microarray technology. However, because to noise, variability, and repeated testing concerns, analyzing such high-dimensional data is extremely difficult. By combining Bayesian statistics and empirical Bayes methodologies, empirical Bayes analysis of microarrays (EBAM) [148] tackles these difficulties. The main principle of EBAM is to use information from diverse genes to improve the identification of genes with differential expression. It accomplishes this by modeling the distribution of gene-specific parameters using hierarchical models, such as mean expression levels and variances. The gene-specific metrics, such as mean expression levels and variances, are presumptively distributed according to EBAM. An empirical Bayes method is used to estimate these distributions from the available data. Usually, these characteristics are believed to have a normal or gamma distribution. Following the estimation of the parameters, EBAM determines a score or statistic for each gene that assesses the presence of differential expression. This rating is frequently based on a moderated t-statistic, which takes into consideration both the variation within and between genes. Multiple testing correction is essential to manage the FDR, as microarray data analysis entails simultaneously evaluating a large number of hypotheses (one for each gene). The gene-specific p-values or scores for multiple testing can be modified using a variety of techniques, including the Benjamini-Hochberg [43, 764] method. Finally, using the modified p-values or scores, genes with strong evidence of differential expression are found. A list of genes that are differentially expressed can be created by setting a threshold on the adjusted p-values or managing the FDR. When working with small sample sizes or highly variable data, EBAM's capacity to borrow information across genes makes it particularly useful.

Rank Product (RankProd) [239] is a non-parametric method based on the idea that genes that consistently ranks high in one group but consistently ranks low in the other has high probability of being differentially expressed. For each gene a ranking score is calculated based on its expression level in each sample. Rank product statistics measure the consistency of ranking across samples and groups. FDR is estimated using permutation based methods and the FDR cut-off determines the genes whose expression differs significantly between the two groups. RankProd is robust to outliers and can handle data that doesn't fall under normal distribution but is sensitive to parameter selection.

Bayesian hierarchical modelling [57] is a statistical method to analyze microarray data by combining data from several levels of data heirarchy, such as genes, samples and experimental conditions. It employs the Bayesian framework that uses previous knowledge and likelihoods to draw inferences about the data. For each level of hierarchy, a prior distribution that reflects the data distribution based on previous assumptions and information is provided. Based on the observed data, the likelihood function is established. The objective is to estimate the posterior distribution of the model parameters based on prior knowledge and the observed data. Prior knowledge and uncertainty is incorporated into the model in Bayesian hierarchical modelling thus increasing the accuracy and precision of the estimation. Bayesian hierarchical modelling facilitates estimation of the variance and correlation structures over the layers of data hierarchy so as to account heterogeneity and dependence in the data.

4.2.3 Bulk RNA-Seq Methods

In order to overcome the difficulties presented by count-based sequencing data, where the distribution of expression levels is frequently skewed and the variability is dependent on the mean expression, Voom (Variance modelling at the observational level) [332] was specifically designed. The count data is first converted by Voom into approximative log-counts per million (log-CPM) values. This adjustment makes the variance more accessible to common statistical modelling approaches by stabilizing it and reducing its dependence on the mean expression. Inferring precision weights from the mean-variance relationship in the data, Voom calculates precision weights for each observation (gene). Low variability genes have higher weights compared to high variability genes, and vice versa. By lowering their weight, genes with high variability have less of an impact on the statistical analysis that comes after. The correlation between the precision weights corresponding to the log-CPM values is modeled by Voom. It conforms to a linear model that takes into account heteroscedasticity and mean-dependent variance. To effectively estimate the mean-variance relationship, this modeling phase is essential. Based on the fitted mean-variance relationship, Voom calculates the precision (inverse of the variance) at the observation (gene) level. In order to produce more trustworthy differential expression statistics, these precision estimates are applied. To evaluate differential expression between various experimental conditions or groups, a moderated t-test or linear model is run using the estimated observation-level precisions. The results of the moderated t-test are more solid and trustworthy since it takes into consideration both the variability within and between genes. Due to its versatility in handling count-based data and tolerance against variability-dependent mean expression, Voom has grown in popularity in the analysis of RNA-seq data.

edgeR [584] is a statistical software program and approach for analyzing differential gene expression in high-throughput sequencing experiments, notably RNA-seq data. It was created especially for count-based data, where the amount of reads that correspond to each gene is utilized to calculate the level of gene expression. EdgeR's first step is to normalize the raw count data in order to eliminate systematic biases. The normalization process makes allowances for variations in library size and sample makeup. The most popular normalization technique in edgeR is the trimmed mean of M-values (TMM) [584], which takes into consideration both the differences between genes and the total number of reads. Using a negative binomial distribution, edgeR models the data's biological and technical variability. EdgeR uses an empirical Bayes approach [583] to estimate the dispersion parameter, which represents the level of variability for each gene. Utilizing the knowledge shared across genes, this phase enables the estimation of the dispersion for a given gene. EdgeR models the association between experimental variables or relevant conditions of interest and gene expression using a generalized linear model (GLM) [518] framework. The GLM can include a variety of design elements, including time points, treatment conditions, and batch effects. To evaluate the significance of differential expression, a likelihood ratio test or a quasi-likelihood F-test is used. EdgeR tests many genes at once, thus it's crucial to account for multiple testing to keep the FDR under control. EdgeR frequently use the Benjamini-Hochberg [43, 764] approach to correct the p-values or test statistics for multiple testing. Due to its strong statistical technique, capacity to handle count-based data, and ability to detect differential expression with very small sample sizes, edgeR has gained widespread application in the field of genomics.

The popular packages DESeq [23] and DESeq2 [449, 450] are used for differential gene expression analysis in RNA-seq research. To compare the levels of gene expression across various situations or populations, they offer statistical techniques and tools. Data normalization is done by DESeq to adjust for variations in library size and make-up between samples. To account for variations in library sizes, it employs the median of ratio normalization (MRN) [494] technique. To make count data more suitable for later statistical analysis, DESeq also performs a variance stabilizing transformation (VST) [28]. A negative binomial distribution is used by DESeq to estimate the dispersion parameter, which symbolizes the variability of gene expression. To precisely estimate the dispersion for each gene, a model describing the connection between the mean and variance of the count data is used. To get more accurate estimates of dispersion, DESeq employs the method of maximum likelihood estimation (MLE) [13], also known as shrinkage estimation. To model the count data with a negative binomial distribution and calculate the fold changes in gene expression between various conditions, DESeq uses a generalized linear model (GLM) [518]. The statistical significance of differential expression is evaluated using hypothesis testing, which is frequently based on Wald tests [716]. In order to stabilize the estimation and draw information from several genes, DESeq additionally uses empirical Bayes methods [583]. In order to control the FDR, DESeq modifies p-values or test statistics using the Benjamini-Hochberg approach[43, 764] or other suitable techniques. This adjustment takes into consideration multiple hypothesis testing and aids in the discovery of genes with notable expression differences. A version of DESeq with more features and improvements is called DESeq2. The size factors normalization technique [23], used by DESeq2, is more reliable. The accuracy of DEA is increased by using this method, which takes into consideration the compositional character of RNA-seq data. To enhance the estimation of dispersion and fold change shrinkage, DESeq2 uses a brand-new approach dubbed estimating size factors for generalized linear models (EGM) [23]. This improves the identification of genes that are differently expressed, especially those with low expression levels or low numbers. To calculate the fold changes, DESeq2 uses a Bayesian method referred to as adaptive shrinkage [644]. With the help of this technique, the estimates are shrunk towards a common distribution, increasing their accuracy and lowering their variability. Additional features offered by DESeq2 include gene-level and pathway-level analysis, visualization tools, and compatibility with more intricate experimental designs, such as time-series analysis and multi-factor studies. Due to their reliable statistical approaches, capacity to manage count-based data, and extensive functionality, DESeq and DESeq2 have grown to be widely utilized in the analysis of RNA-seq data.

NOISeq [676] is a non-parametric method for analyzing bulk RNA-Seq data. With the aim to circumvent the dependence of parametric methods such as DESeq2 [449, 450] and edgeR [584] on data distribution and their susceptibility to outliers as well lowly expressed genes, NOISeq was developed. NOISeq compares the empirical read count distribution between two conditions and employs non-parametric techniques for estimation of fold changes and FDR. By using a permutation based approach NOISeq determines a noise threshold and estimates FDR with the aim to distinguish noise from differential expression. It is effective in bulk RNA-Seq data with low read depth or substantial technical variance and does not require normalization or batch effect removal as it does not presume normal distribution.

Sleuth [55] employs maximum likelihood estimation [13] to compute differential expression and calculates gene level abundance estimates. Sleuth fits a linear model to the read counts with the aim to estimate mean and variation level for each gene and sample. Like in Limma, Voom [332] is employed by Sleuth to transform the bulk RNA-Seq data so as to facilitate linear model analysis. Sleuth makes exploration of data and discernment of unusual patterns easier by facilitating interactive real time visualization of the results. Sleuth offers various built in statistical tests, such as Wald tests [716], that can be implemented for comparison of expression levels under two or more conditions. Sleuth further offers various quality control and normalization tools such as, the capability to remove batch effects, apply spike in controls for normalization.

Tab. 4.1: DE methods for Microarray and bulk RNA-Seq data.

	Method	Pros	Cons
Statistical Test	t-test [761, 615]	• Simplicity and accessibility. • Performs well even with comparatively small sample sizes. • Presumes that the data have a normal distribution. Thus resilient to outliers, especially when the sample sizes are sufficiently big. • Yields results that are simple to understand in the form of $p-value$ that shows the statistical significance of the difference that was noticed.	• Presumes that the data have a normal distribution. Thus, deviations from normality, can lead to erroneous or misleading results. • Sensitive to outliers particularly with small sample sizes. • Built specifically for comparing the means of two groups. • Focuses on the disparities between two groups according to a single significant variable. • Less accurate and more likely to produce false-positive or false-negative results when the sample sizes are very small.
	ANOVA [168, 167]	• The simultaneous comparison of means across three or more groups is possible. • One-way, two-way, and factorial study designs can all be accommodated • The pooled variability across groups can be utilized thus improving the capacity to identify real differences between conditions. • Covariates or confounding variables can be incorporated. • Provides a framework for post-hoc tests to identify which particular group differences are statistically significant.	• Assumptions of normality and homogeneity of variances may result in outcomes that are incorrect or biased. • Sensitive to outliers, particularly in small sample sets. • Assumes that all groups being compared have similar sample sizes. • Does not specify which particular categories are most distinct from one another. • Not suited for nested experimental designs or random effects, where the grouping structure is more intricate.
Microarray	Linear Models for Microarray Analysis (Limma)[637, 638]	• Provides a strong framework that incorporates statistical modelling techniques to create results that accounts for causes of volatility. • Employed shrinking process used produces estimates of differential expression that are more stable and exact. • Capable of managing intricate experimental designs with several factors, confounders, and interaction terms. • Provides empirical Bayes statistics and moderated t-statistics, which have better precision and robustness.	• Requires a solid grasp of statistical and linear modelling principles in order to be used correctly. • Assumes that the data are modelled linearly, with the covariates or factors being combined linearly to get the gene expression levels. • Computationally intensive. • Can cause biases and effect the results if batch effects are not effectively addressed. • Does not automatically do multiple testing correction, thus users must make the necessary modifications.
	Significance Analysis of Microarrays (SAM)[697]	• Performs effectively even when the sample size is small. • Built to handle data that might not always follow a normal distribution. uses non-parametric techniques to evaluate differential expression, making it appropriate for gene expression data with outliers or deviations from the norm. • Uses the FDR concept to allow for multiple testing. • Enables researchers to define parameters based on the requirements and specifics of their own study design.	• It is computationally demanding because it uses permutation-based resampling to estimate the null distribution and determine the statistical significance of differential expression. • Assumes that genes are independent of one another, i.e., that there is no correlation between the expression levels of various genes. • Assumes that the gene expression data have a normal distribution or that a normal distribution can be appropriately used to approximate the data. • The amount of discovered DEGs may vary depending on how the δ threshold is chosen. • Primarily intended for paired-sample or two-class comparisons.

	Method	Pros	Cons
Microarray	Empirical Bayes Analysis of Microarrays (EBAM)[148]	• The employed shrinkage method incorporates data from various genes, making estimates of differential expression more dependable and consistent. • Compared to other approaches, is more likely to correctly identify DEGs. • It can be adapted to diverse experimental designs because it allows for the insertion of different design matrices and variables.	• Assumes that gene expression measurement variances are comparable across genes or that they can be accurately approximated by using data from other genes. • As it uses prior distribution hypotheses to calculate the variances particular to each gene. the outcomes may be effected resulting in skewed results. • Sensitive to the choice of previous parameters. • Includes computationally expensive iterative calculations and estimate techniques,
	Rank Product (RankProd)[239]	• a non-parametric technique that doesn't rely on presumptions about how the data are actually distributed. • Capable of handling a variety of experimental designs, including ones with numerous factors, confounders, or varied treatment conditions. • Focuses on finding genes that persistently exhibit differential expression after going through several iterations of resampling or permutations. • Recognised for it's capacity to deal with bias-introducing batch effects and other errors. • Identifies genes that are expressed differently in either direction by taking into account both upregulation and downregulation of genes.	• Can be sensitive to ties in the ranked gene expression values. • Its capacity to measure the amplitude of gene expression changes and assess the biological significance of DEGs is constrained by the fact that it primarily focuses on finding consistently rated genes across permutations or resampling iterations. • Generating the null distribution needs choosing the amount of permutations or resampling iterations. Choosing the right amount of permutations might be difficult. • Ideal for two-class comparisons, but more difficult to interpret in comparisons of multiple classes. • Involves computationally expensive resampling or permutation methods,
	Bayesian hierarchical modeling[57]	• Provides modeling complex dependencies and linkages in the data some flexibility. • Enables the inclusion of past understanding or assumptions regarding the data. • Provides estimates of posterior distributions for model parameters will allow you to quantify uncertainty and get more detailed, illuminating findings. • Assists the fusion of data from several sources, such as omics data from various studies or data from related ones. • Enables posterior predictive checks, where scientists can create brand-new data based on the estimated model parameters and contrast it with the observed data.	• Includes the estimation of a large number of parameters, particularly when working with complex experimental designs or large-scale datasets. • Requires that previous distributions for the model parameters be specified. • Can get too complicated, particularly when attempting to take into account different sources of variation in gene expression data. • Rely on presumptions about the data, such as the presumption of the normality of the distribution of gene expression values or the presumption of the conditional independence between the variables.
Bulk RNA-Seq	Limma+Voom[332]	• A flexible framework for modeling the mean-variance relationship in RNA-Seq data. • Allows the use of conventional linear modeling techniques on RNA-Seq data by converting the raw count data into a logCPM format. • Loosens up the strict RNA-Seq count distribution assumptions. • Even with fewer repeats, the precision weight model's estimations of differential expression are more accurate because it takes into account sample variability both within and between samples.	• The results might not be accurate if the data are severely skewed or contain extreme outliers. • Depends on precise gene abundance prediction from RNA-Seq data. • Assumes that a reference design is available, in which samples are compared to a standard reference or control group. • Involves additional computational steps compared to some other DEA methods

	Method	Pros	Cons
Bulk RNA-Seq	edgeR[584]	• Utilizes a negative binomial model that takes into account the intrinsic over-dispersion seen in count data, which is frequently not sufficiently reflected by conventional approaches that use the normal distribution as their basis. • Incorporates the empirical Bayes method to increase the precision and power of the differential expression detection by adaptively shrinking the gene-wise dispersion estimates. • Uses the TMM (Trimmed Mean of M-values) approach to normalise the count data in order to take into consideration sample and library size variations.	• Sensitive to low counts because it presumes a negative binomial distribution that might not hold for genes with extremely low counts. • Performs most effectively only when there are enough replicates in each group. • Assumes that the genes are interchangeable, which enables the pooling of dispersion estimates across genes. • Might not be directly applicable to other kinds of high-throughput data because it was created specifically for count-based RNA-Seq data.
	DESeq2[450]	• Enables more precise modeling of the variation in gene expression across samples by estimating gene-specific dispersion. • Consists of a regularized log transformation (rlog) that reduces the variance at different mean expression levels. Low-count genes, which frequently exhibit higher levels of noise and unpredictability, benefit especially from this change. • Supports a wide range of experimental designs, including as time series studies, multifactorial designs, and designs using variables.	• Makes the assumption that the count data have a negative binomial distribution, which might not be true if the data have a lot of extreme outliers or are heavily skewed. • Depends on the precise measurement of the size factors, which are used for normalization to take into account variations in the size of the library. • Performs best when each group has an appropriate number of duplicates. • Increases the computational needs by utilizing intricate statistical modeling and estimation techniques.
	NOISeq[676]	• Employs a nonparametric approach, which implies it doesn't rely on certain distributional presumptions about the data. • Because of its nonparametric approach, it is resistant to outliers and data skewness. • Suitable for experiments with few duplicate samples since it can handle datasets with tiny sample sizes.	• Employs resampling methods, which can be computationally demanding, especially when analyzing a large number of genes or large-scale datasets. • The sample size might have an impact on performance. • Does not specifically model the variation in gene expression, which could affect the DEA's accuracy. • When working with very big datasets, its resampling feature may hinder its ability to scale.
	Sleuth[55]	• Estimate the abundance of transcripts, allowing for accurate assessment of gene expression levels even in the presence of complicated transcript architectures and alternative splicing events. • mploys a comprehensive statistical model that takes into account a variety of sources of technical variability, including batch effects, gene length bias, and sequencing depth • Offers adaptable modeling options to deal with different experimental layouts and data properties.	• Sleuth's reliance on its own method for estimating transcript abundance restricts the flexibility for users whose data have previously been aligned using different techniques. • instead of gene-level analysis, principally focuses on transcript-level analysis. • Computationally taxing, especially for big datasets with many of samples or complicated transcript architectures. • ses bootstrapping to estimate posterior distributions and confidence intervals for model parameters that may not fully capture the uncertainty or variability present in the data.

4.3 Identification Of Potential Biomarkers Using Integrative Approach: Application To ESCC

The proposed framework detailed in Fig 4.1 is a consensus-based approach that incorporates six DEA methods for unbiased and integrative identification of DEGs as potential biomarkers for critical diseases. Based on the input dataset, our framework employs specific DEA methods to detect DEGs independently. We have chosen three methods that work on micro-array (Limma [637, 638], SAM [697] and EBAM [148]) and three on bulk RNA-Seq (Limma+voom [332], DESeq2 [449, 450] and EdgeR [584]).

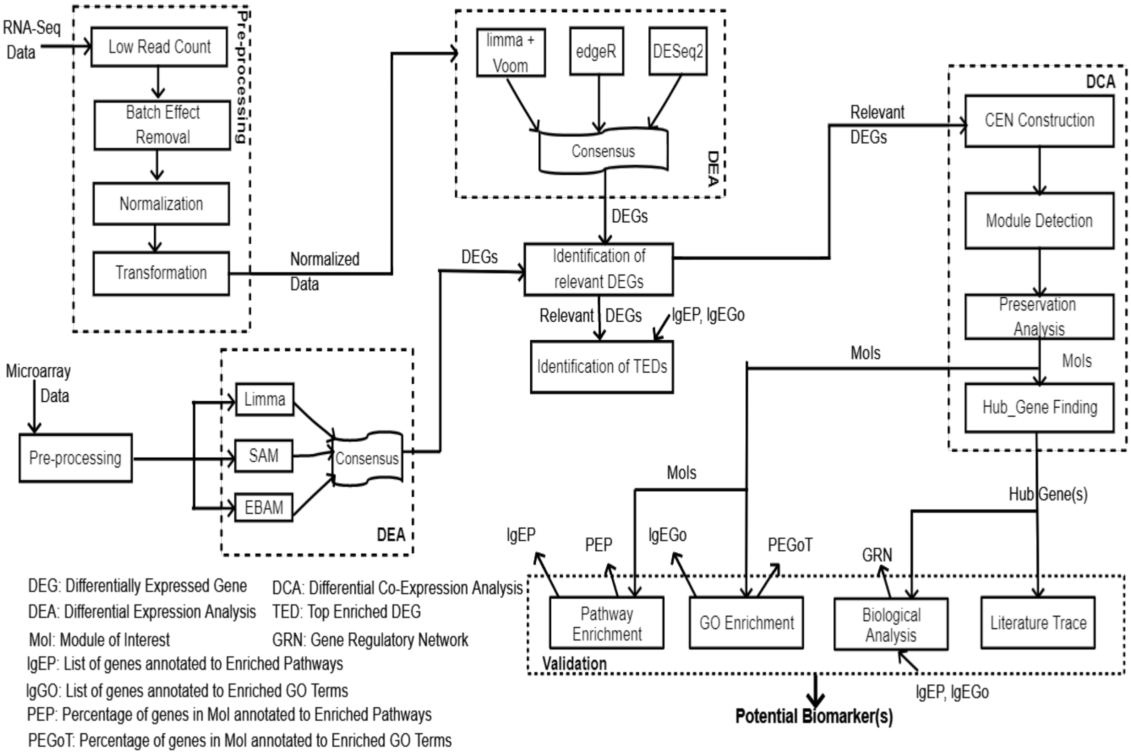


Fig. 4.1: Proposed Integrative Differential Expression Analysis Framework

4.3.1 Pre-processing

The microarray or RNA-Seq data are input to the proposed framework and based on this input data type the pre-processing method is chosen. For microarray data, pre-processing consists of the removal of unwanted and redundant information, normalization of the dataset, missing value estimation while for RNA-Seq data we perform removal of low read counts, normalization, and transformation. the general pipeline we employ for pre-processing of the microarray and bulk RNA-Seq data are discussed in detail in Section 2.7.1 and Section 2.7.2, respectively.

4.3.2 DEA

The pre-processed data is input to the DEA units. Based on the pre-processesd data, the framework employs the respective DE methods that results in identification Differentially Expressed Genes (DEGs). For each data type, we employ a consensus function that filters out all common DEGs for each dataset. In other words, depending on the type of the input dataset, the DEA unit detects DEGs using three corresponding methods, followed by a consensus function that filters the DEGs common to all three methods as well as identify other relevant DEGs. Our consensus function is given by equation 4.1.

$$DEGs_{relevant} = DEGs_{common} \cup DEGs_{others} \tag{4.1}$$

where

$$DEGs_{common} = \begin{cases} DEGs_{limma} \cap DEGs_{SAM} \cap DEGs_{EBAM}, & \text{for } Microarray, \\ DEGs_{limma-voom} \cap DEGs_{edgeR} \cap DEGs_{DESeq2}, & \text{for } RNA\text{-}Seq \end{cases}$$

and

$$DEGs_{others} = DEGs \text{ such that} \begin{cases} DEGs \notin DEGs_{common} \text{ and } q\text{-}value \leq \alpha, & \text{for } RNA\text{-}Seq, \\ DEGs \notin DEGs_{common} \text{ and } lFDR \leq \beta, & \text{for } Microarray \end{cases}$$

Here, α and β are *q-value* (Section 2.1.3) and *lFDR* (Section 2.1.4) significance values that are chosen according to their relevance to the experiment. Through multiple iterations of implementation, we observed that consideration of only genes common to all three methods leads to information loss. Thus, to overcome this we introduced *q-value* into the consensus function. The main idea behind this is that while a *p-value* (Section 2.1.3) of 0.05 gives the implication that 5% of the tests will be false positive (FP), *q-value*, which is an FDR adjusted p-value, implies that 5% of the test found to be significant will be FP. *q-value* requires a very important adjustment for multiple tests on the same data sample. Our consensus function considers all genes common to all three methods with $p = 0.05$ as detected DEGs. Furthermore, all genes that are not among the common genes but have a $q = 0.05$, i.e. α (Equation 4.1), are added to the list of DEGs. However, in the microarray datasets, we implement the proposed consensus function given by Equation 4.1 to start off by taking the DEGs common to all three methods.

Unlike RNA-Seq, instead of *q-value*, we incorporate its useful counterpart *local False Discovery Rate* (β in Equation 4.1). *lFDR* (Section 2.1.4) is a measure of the posterior probability that the null hypothesis is true. We use *lFDR* since Limma and SAM calculate *p-value*. EBAM, on the other hand, estimates the posterior probability and *lFDR*. It is worth mentioning that posterior probability and *p-value* are not interchangeable. However, *lFDR* can be estimated from *p-values*. Our consensus function (Equation 4.1) considers all genes common to all three methods with $p = 0.05$ as detected DEGs. Furthermore, all genes that are not among the common genes but have a $lFDR = 0.05$ (β) are added to the list of DEGs. All DEGs that are determined as relevant DEGs by equation 4.1 are considered for downstream analysis.

4.3.3 DCA

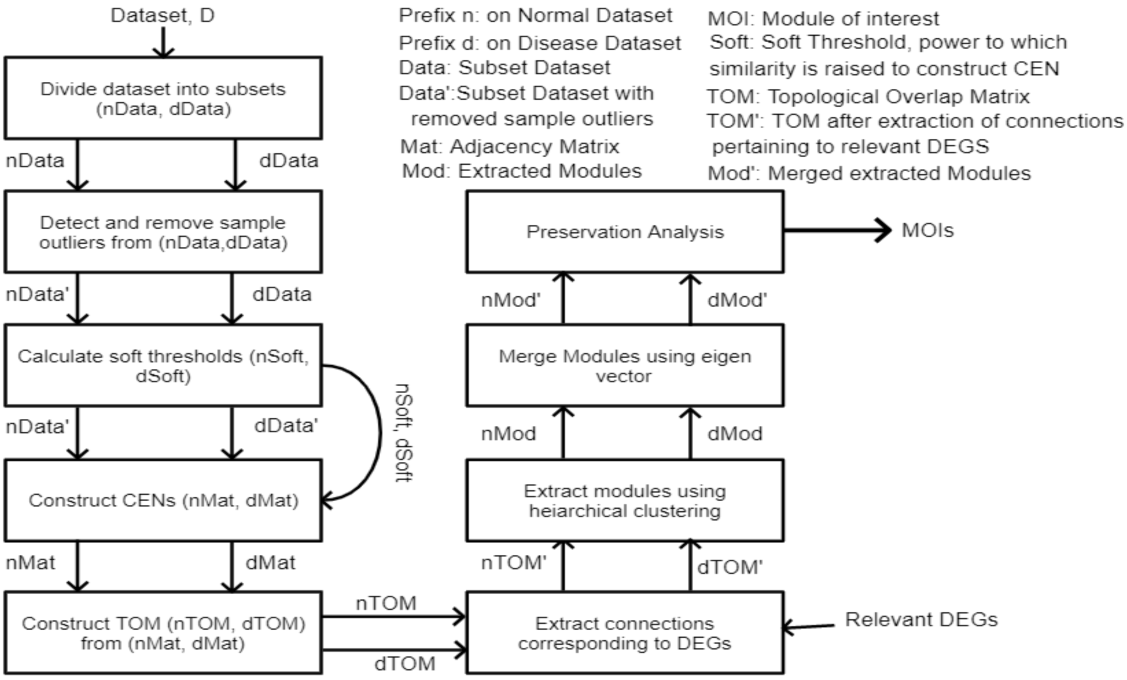

Dataset, D
Divide dataset into subsets (nData, dData)
nData
dData
Detect and remove sample outliers from (nData,dData)
nData'
dData
Calculate soft thresholds (nSoft, dSoft)
nData'
dData'
nSoft, dSoft
Construct CENs (nMat, dMat)
nMat
dMat
Construct TOM (nTOM, dTOM) from (nMat, dMat)
nTOM
dTOM
Extract connections corresponding to DEGs
Relevant DEGs
nTOM'
dTOM'
Extract modules using heiarchical clustering
nMod
dMod
Merge Modules using eigen vector
nMod'
dMod'
Preservation Analysis
MOIs
Prefix n: on Normal Dataset
Prefix d: on Disease Dataset
Data: Subset Dataset
Data':Subset Dataset with removed sample outliers
Mat: Adjacency Matrix
Mod: Extracted Modules
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
Mod': Merged extracted Modules

Fig. 4.2: Pipeline for DCA

All relevant DEGs identified by the DEA unit(s) are taken as input to the DCA unit. The idea behind performing DCA is that it leads to the creation of biologically relevant modules which are easier for further analysis and validation. The DCE unit firstly constructs two CENs corresponding to normal and disease subsets of the input dataset. This is then followed by the extraction of all connections corresponding to the relevant DEGs. The pipeline for DCA in our framework is given by Fig. 4.2.The DCA unit identifies differentially co-expressed modules and performs preservation analysis (Section 2.1.9) on these modules to identify biologically relevant modules. These modules are termed as "Modules of Interest" (MoI) (Definition 4.3.1). This is followed by the identification of hub-genes (Definition 3.4.3 in Chapter 3) in these modules using WGCNA [327] intramodular connectivity.

Definition 4.3.1 (Module of Interest (MoI))**.** A module is defined as 'module of interest' if (i) its $size \geq 100$, and (ii) it is not highly preserved , i.e., it is either non-preserved ($Z_{summary} < 2$) or moderately preserved ($2 \leq Z_{summary} \leq 10$).

4.3.4 Identification of TEDs

We consider all the hub-genes in the biologically relevant modules identified by the DCE unit as biomarker candidates and term them as biomarker candidate genes (BCGs) (Definition 4.3.3). Furthermore, all DEGs that are annotated with the most enriched GO term in all three databases as well as the most enriched KEGG pathway are termed as Top enriched DEGs (TEDs) (Definition 4.3.2). TEDs are also added to the list of BCGs. The validation unit of the framework validates modules in general and BCGs in specific.

Definition 4.3.2 (Top Enriched DEG (TED))**.** A DEG deg_i is defined as a Top Enriched DEG (TED), if deg_i is annotated to the most enriched GO term in all three GO databases (BP, CC, and MF) as well as annotated to the most enriched KEGG pathway.

Definition 4.3.3 (BCG)**.** A gene g_i is defined as a Biomarker Candidate Gene (BCG) if it is identified as a hub-gene or a TED or both in a given MoI extracted by Integrative DEA.

4.3.5 Validation

We employ two approaches for validation. We initially assess the quality of the module(s) retrieved by the DCA unit as MoI (Definition 4.3.1) in order to identify the BCGs identified by our framework as a potential biomarker(s). The procedures for module validation are as follows:

(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.

A module is pathway enriched and GO enriched if at least one enriched pathway and one enriched GO term are present in the module. MoIs identified by the preservation analysis unit are validated by performing Gene Ontology (GO) enrichment (Section 2.4.1.1) and pathway enrichment analysis (Section 2.4.1.1). In the framework, all identified MoIs are taken as input into the pathway enrichment analysis and GO enrichment sub-unit of the validation unit. For each MoI, these subunits compute the percentage of enriched GO terms (PEGoT) for all three GO databases. These three databases are Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) and the percentage of enriched pathways (PEP) in KEGG with $p-value = 0.05$.

For validation of each BCG identified by the framework, we first find lgEGo and lgEP with $p-value = 0.05$. The DEGs identified by the identification of DEGs unit in the framework are input to the GO enrichment and pathway enrichment sub-units. The output is two lists, lgEGo and lgEP. To validate the BCGs identified by the hub-gene finding unit of the framework lgEGo and lgEP as well as the list of BCGs are input to the biological analysis unit. The biological analysis unit identifies BCGs that are annotated to enriched GO terms and enriched pathways. In other words, the biological analysis unit identifies the BCGs that are present in lgEGo and lgEP. This unit further identifies BCGs that are TFs and constructs GRN to establish the regulatory behavior of these BCGs in the network. The literature trace sub-unit of the validation unit in the framework identifies BCGs that have traces of published literature that establish them as biomarkers for ESCC or other SCCs closely related to ESCC. Based on our biomarker criteria (Section 2.5) we identify the BCGs that fall under Cases 1 and 2 and identify them as potential biomarkers.

4.4 Experimental Results

As our primary focus is on ESCC and two microarray ESCC datasets, GSE20347 and GSE23400, and one bulk RNA-Sequencing ESCC dataset, GSE130078 were chosen to validate our proposed framework. Details of each dataset (Table 2.1) is described in details in Section 2.6.1 and Section 2.6.2. The test platform is a DELL workstation running Windows 10 Pro for workstations with an Intel(R) Xeon(R) W-2145 CPU with 3.70GHz processor and 64 GB of RAM. We conduct the results in R programming environment (Section 2.2.1).

4.4.1 Pre-processing

RNA-Seq dataset GSE130078 has 57,783 genes and 46 samples. Large datasets tend to add complications to the analysis and as such, we filter out genes with low read counts. We achieve this by calculating the counts per million (CPM) for each sample for each gene and keep only those genes that have $CPM > 1$ for at least two samples. This reduces the dataset size from 57,783 to 22,270. We then follow up by normalization of the dataset. We also consider two microarray datasets GSE20347 and GSE23400 for analysis. The inputs to these datasets are expression values of genes across samples. First, we pre-process the data through the removal of unwanted and redundant genes, missing value estimation, and normalization. However, for both GSE20347 and GSE23400, there are no missing values and as such we proceed further down the pipeline.

4.4.2 Identification of DEGs

For the microarray datasets, Limma takes the pre-processed dataset as input and outputs the equivalent DEGs with a significance of 5%(p-value ≤ 0.05) and FDR of 0.05. On the other hand, for the other two methods SAM and EBAM, we employ $findDelta$ with $FDR = 0.05$ giving us an estimate of the delta values at which FDR is closest to 0.05 and chose accordingly. In SAM, delta is the distance between the observed and the expected test scores, whereas in EBAM, delta is the probability that a gene with a specific test score is differentially expressed. Table 4.2 summarizes the DEGs detected by all three methods on all three datasets.

Tab. 4.2: Summary of detected DEGs by the three RNA-Seq methods and the three microarray methods for three datasets

Dataset	Method	No. of DEGs with $p \leq 0.05$	Common DEGs
GSE20347	Limma	8,689	7,706
	SAM	10,642	
	EBAM	9,565	
GSE23400	Limma	13,558	3,431
	SAM	14,301	
	EBAM	3,431	
GSE130078	Limma +Voom	6,858	2,765
	edgeR	12,623	
	DESeq2	12,766	

In the case of the bulk RNA-Seq dataset, the pre-processed data are the input to all three methods, i.e., Limma+Voom, edgeR and DESeq2. However, it is to be noted that while DESeq2 directly takes the count data as input, the other two methods require the count data to be transformed into a DGEList (Digital Gene Expression Data) object. All the methods perform multiple tests on all the 22,270 genes in the dataset across 46 samples.

We consider a significance of 5%, i.e., p-value ≤ 0.05 and the corresponding DEGs detected by the three methods are summarized in Table 4.2.

4.4.2.1 Consensus Function

We implement the proposed consensus Equation 4.1 to identify the common genes detected by these three methods. First, we consider the DEGs detected by all three methods, i.e. common genes. In GS20347, there are such 7,706 DEGs. So as not to bypass crucial information, we use β in Equation 4.1, i.e., the consensus function. With $lFDR = 0.05$ (β) another 662 genes are considered DEGs resulting in a list of 8,368 DEGs. Similarly, in GSE23400, Limma, SAM, and EBAM find 3,431 common DEGs. With $lFDR = 0.05$ (β), another 4,066 genes are considered as DEGs, resulting in a list of 7,497 DEGs. In the case of GSE130078, the three methods Limma+Voom, edgeR, and DESeq2 discover 2,765 common DEGs and a *q-value* (α) adds another 9,945 genes resulting in a list of 12,710 DEGs.

In GSE130078, we find 2,765 common DEGs. However, it is to be noted that filtering genes based on this criterion alone might result in the loss of relevant information. For example in the case of GSE130078, TUSC2 and HOTAIR are the only two known ESCC causal genes detected among the 5,337 common genes, out of 10 causal genes present in the entire dataset. However, in addition to these two genes, Limma+Voom was able to detect another one of the 10 causal genes named TINCR among the 6,858 DEGs. Similarly, edgeR detected CDK14 and MEG3 and DESeq2 detected TUG1, MEG3, and CDK14 among the 12,623 and 12,766 DEGs respectively.

4.4.3 DCA

To analyze the interactions among the DEGs as well as the variations in behavior under normal and disease circumstances, we construct co-expression networks (CEN) using WGCNA [327]. The pipeline for DCA done by the framework is described in detail in Fig. 4.2 and Section 4.3.3.

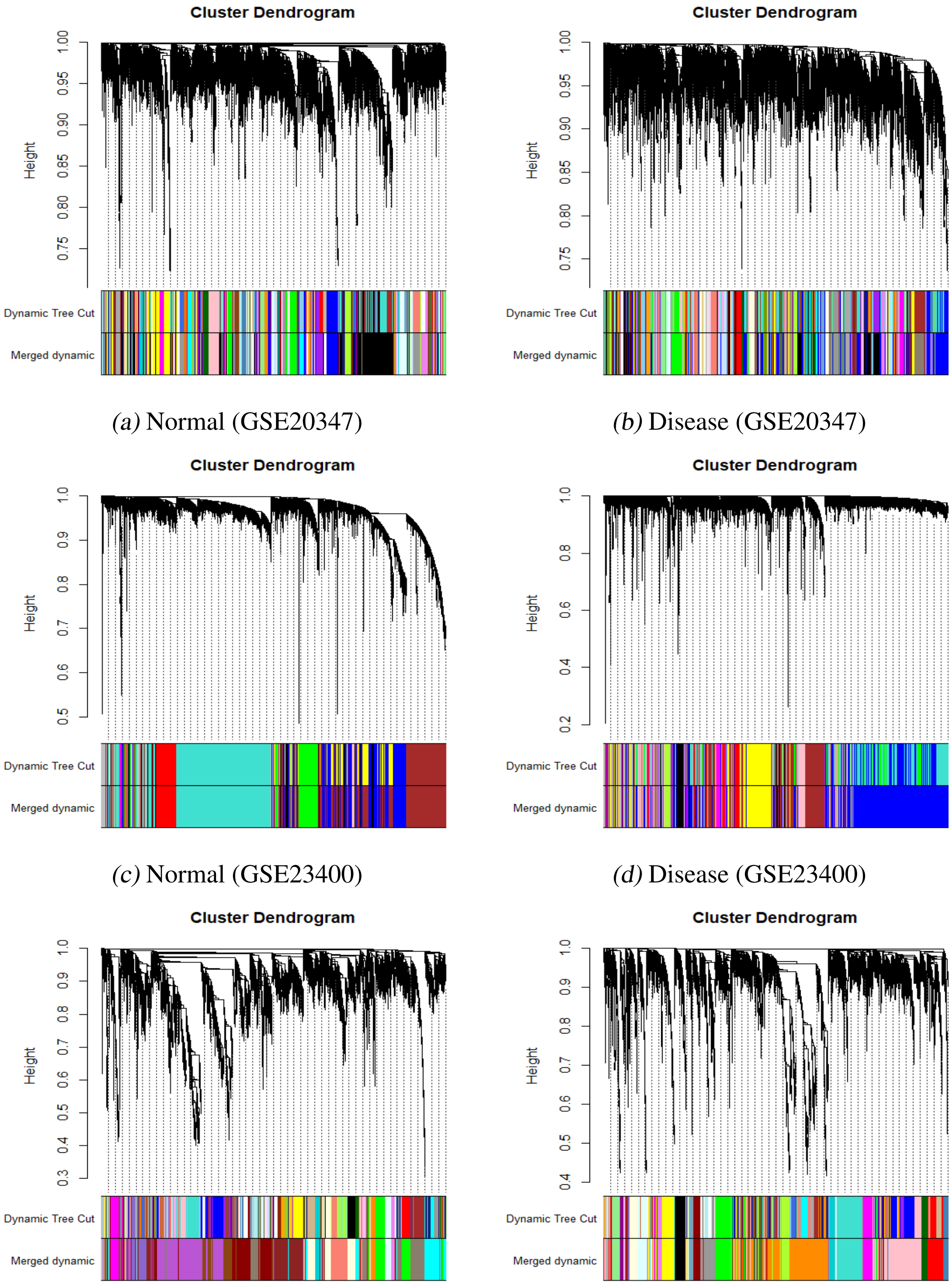
We start DCA by clustering the samples using the hierarchical approach to detect outlier samples. We remove the outlier samples with the aim of creating a more robust CEN. For GSE23047, as seen in Fig 3.4a and Fig 3.4b in Section 3.5.3, we find a single outlier in the case of normal samples with a tree cut at height $h = 70$ (Blue). However, in disease samples, there are two outliers with a cut at $h = 130$ (Red). Similarly, in the case of GSE23400, as seen in 3.4c and Fig 3.4d in Section 3.5.3, a tree cut at height $h = 105$ (Blue) and at $h = 95$ (Red) removes one and two outliers from normal and disease samples, respectively. In the case of GSE130078, a cut at $h = 1500000$ (Blue) and $h = 2000000$ (Red) removes one normal (Fig 3.4e in Section 3.5.3) and one disease Fig (3.4f in Section 3.5.3) sample.

4.4.3.1 Soft Threshold

We apply soft-threshold to the normal (Blue) and disease (Red) samples of dataset GSE20347. Nine is the lowest power for which the network maintains scale-free topology, as can be shown in Fig. 3.5a and Fig. 3.5b in Section 3.5.3.1. As shown in Fig. 3.5c and Fig. 3.5d in 3.5.3.1, the soft thresholding for normal (Blue) and disease (Red) samples in GSE23400 is set at nine. In contrast, for GSE130078, normal (Blue) and disease (Red) samples are selected with soft thresholds of twelve (Fig. 3.5e in Section 3.5.3.1) and nine (Fig. 3.5f in Section 3.5.3.1), respectively.

4.4.3.2 CEN Construction

Using the soft threshold exponent nine, 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 nine. The number of genes in GSE130078 is decreased to 22,270 after CPM filtering, resulting in two adjacency matrices with soft thresholds of twelve (normal) and nine (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.



(a) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

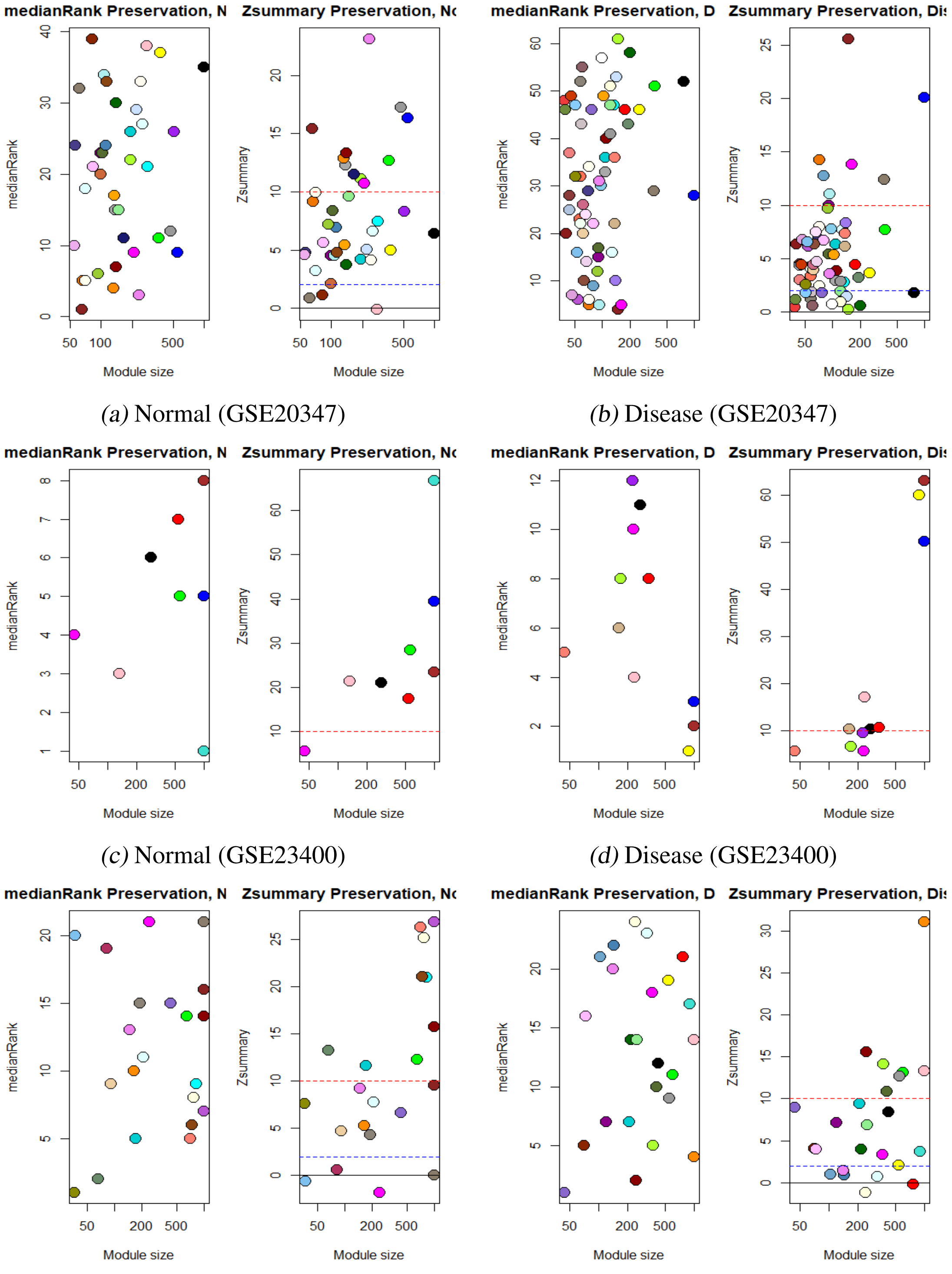
Fig. 4.3: Dendrograms for a) normal and b) disease in GSE20347, c) normal and d) disease in GS23400, and e) normal and e) disease in GSE130078. The first strip of colors represents the corresponding module colors assigned after hierarchical clustering while the second color strip of colors represents the corresponding module colors after merging.

In GSE20347, hierarchical Clustering (Section 2.1.5) and tree cut results in 50 and 75 normal and disease modules, respectively. Fig 4.3a shows the dendrogram while the first strip of colors below represents the corresponding module colors for the normal dataset.

Similarly, Fig 4.3b shows the dendrogram for the disease dataset. To merge modules, we choose a height cut of 0.25, corresponding to a correlation of 0.75. Merging of the modules with tree cut at h=0.25 further reduces the number of modules to 38 and 61 for normal and disease datasets, respectively. The second color strip in Fig 4.3a and Fig 4.3b shows the colors for the merged normal and disease modules respectively. In GSE23400, hierarchical clustering results in 9 normal (the first color strip in Fig4.3c) and 13 (the first color strip in Fig4.3d) disease modules, which are then reduced to 8 normal (the second color strip in fig 4.3c) and 11 disease (second color strip in fig 4.3d) modules after merging. Finally in GSE130078, hierarchical clustering results in 65 normal (the first color strip in Fig 4.3e) and 40 disease (the first color strip in Fig 4.3f) modules, which are then reduced to 21 normal (the second color strip in fig 4.3e) and 24 disease (the second color strip in fig 4.3f) modules after merging.

4.4.3.3 Preservation Analysis

We follow module extraction by module preservation analysis 2.1.9 with the aim of analyzing the distinction between preserved and non-preserved modules. According to Langfelder et al. [329], while the preserved modules retain a majority of their co-expressed connections (or edges between two genes), the same cannot be perceived from non-preserved modules. According to Langfelder et al. [329], a module with $Z_{\text{summary}} < 2$ is considered non-preserved [329]. It is noteworthy that, GSE23400 due to its inherent nature, extracts a smaller number of modules with significantly larger sizes and higher densities. There are no non-preserved modules with $Z_{\text{summary}} < 2$ and most modules are either moderately preserved or highly preserved. We take into consideration moderately preserved modules with $Z_{\text{summary}} < 10$ [329]. In all Z_{summary} plots below, 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) Normal (GSE20347)

(b) Disease (GSE20347)

(c) Normal (GSE23400)

(d) Disease (GSE23400)

(e) Normal (GSE130078)

(f) Disease (GSE130078)

Fig. 4.4: 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.

Table 4.3 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 *pink* in Normal/Disease subset of dataset GSE20347 can be interpreted module *pink* of size 276 detected in the normal network that is non-preserved in disease network with a Z_{summary} value of -0.118842161. We only consider non-preserved modules of substantial size (*size* $\geq$ 100) as MoIs for further downstream analysis and validation.

Tab. 4.3: Preservation analysis of modules detected by our Integrative DEA method in the two microarray and one RNA-Seq datasets

	Ref/Test	Module	Size	Z_{summary}		Ref/Test	Module	Size	Z_{summary}
GSE20347	Normal/ Disease	***pink***	276	-0.11884	GSE23400	Normal/ Disease	*magenta*	45	5.63610
		bisque4	62	0.84896		Disease/ Normal	***magenta***	231	5.59355
		orangered4	82	1.10844			*salmon*	44	5.64756
		grey	17	1.38810			***greenyellow***	172	6.47181
	Disease/ Normal	*grey*	3	-0.11692			*grey*	891	9.22843
		greenyellow	149	0.21296			***purple***	225	9.42312
		brown2	39	0.40356	GSE130078	Normal/ Disease	*magenta*	248	-1.80628
		darkgreen	201	0.57638			*skyblue2*	37	-0.62266
		lightpink4	61	0.58348			***bisque4***	1000	0.01986
		white	99	0.72904			*maroon*	82	0.58940
		lightyellow	122	0.88046			*grey*	70	1.90097
		darkolivegreen4	40	1.13783		Disease/ Normal	***lightyellow***	240	-1.18379
		antiquewhite4	58	1.19766			***red***	759	-0.18883
		lightsteelblue1	143	1.42401			***lightcyan***	321	0.68868
		mediumpurple3	77	1.74769			***steelblue***	145	0.92649
		black	775	1.79867			***skyblue3***	104	1.01960
		skyblue1	51	1.79952			***violet***	142	1.38497
		lavenderblush3	60	1.83005					
		lightgreen	123	1.88961					

4.4.3.4 Hub-genes

To find the hub-genes for each MoI extracted previously we employ WGCNA intramodular connectivity proposed by Langfelder et al. [327]. Intra-modular connectivity calculates the connectivity of a node to other nodes in the same module.

Tab. 4.4: Top 20 hub-genes for each extracted MoI in the two microarray and one RNA-Seq datasets using WGCNA [327] intramodular 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, LaSCC, LSCC, HNSCC, OSCC, and TSCC, are marked in Blue

	Module	hub-genes
GSE20347	*pink*	PTMA, MED1, TRIO, TERF1,BRD2, PWP1, HSD17B10, PPFIA1, EEF1B2, ZNF148, TCOF1 NSD2, SLC25A36, RUFY3, PIK3CB, VGLL4, LYN, DDX24, EPB41L1.
	greenyellow	HOMER3, SHC1, EXT2, PSMD4, CLIC4, MAP3K20, DNMT3B, TGFB2, SELENOP, PSMD11, EXOSC4, SARS1, NABP1, ENTPD7, MYO1B, RAB8B, PSAT1.
	darkgreen	SLC3A2, IMP4, MAPRE1, RALY, PSMB5, UQCRC2, NONO, GNB5, TFRC, GNAPDA1, ODF2, NMD3, RPL22 NEU1, SENP5, NID1, ITSN2, ABI2.
	lightyellow	ANP32E, NEB, AHDC1, RPRM, HOXC11, ENOX2, TNS1, MAN1C1, RCN1, CNPY2, APOOL, HAUSS, SBF1, ESF1, GNAQ, LSS, MCL1.
	lightsteelblue1	DBF4, POP7, MCM7, RFC2, DUS4L, POM121, ZKSCAN5, ORC3, PUS7, GMCL2, PSMC2, ITPKC, TRRAP, TIMELESS, EPHA2, CRYBG2, POM121C, CEP290.
	black	KPNA2, RRP7A, EBNA1BP2, KIF4A, TMEM97, CYP3A5, CCT4, CKS2, HAUS7, CIAPIN1, RANBP1, PITX1, PRMT1, PNO1, MAGOHB, JPT2, SPAG5, VPS13D.
	lightgreen	ITGB7, CXCR3, HPRT1, TARP, NPIPB3, CD48, NEDD4L, CASP10, TP63, UBA7, ITM2A, CD3D, MSRA, ECHDC2, LST1, CD2, UBASH3A, CD52.
GSE23400	*greenyellow*	TAP1, PSMB9, IFIH1, HLA-F, IFIT3, HLA-G, HLA-J, IFI44L, UBE2L6, HLA-C, IFI35, CXCL10, OAS3, IFIT1, PSMB8, ISG15, GZMB, SCO2, CXCL11.
	magenta	CDKN3, PHB, MTHFD1, DLGAP5, EIF2S1, ZNRD2, MNAT1, TIMM9, VRK1, YIF1A, PSMA3, NASP, SRM, PSMC1, EBNA1BP2, C12orf29, GLRX5, PLEK2, TUBG1, TIMM10.
	purple	FCER1G, HNMT, CD14, CD163, TYROBP, LAPTM5, C1QB, MS4A4A, PLXNC1, C1QA, ENTPD1, SRGN, CD53, TFEC, ITGB2, CD86, MS4A6A, FCGR2A, C3AR1, MNDA.
GSE130078	*lightyellow*	PCNX1, CAV1, RRAS2, IGF2BP2, CAVIN1, PI4K2A, PPP4R4, HRH1, SAMD4A, VEGFC, FJX1, SGPP1, LINC01998, PGF, LINC02454, HIF1A, ANO4, FOLR3, FEZ1, CSF2.
	red	COA6, GNA13, LIN52, POLR2D, APPBP2, PPP2R5A, PPP6C, RIT1, RBBP5, MEGF9, RALB, MEF2A, ERCC3, CDC42SE1, SDE2, STARD7, CTDSPL2, BLOC1S2, DDX59, COQ10B.

	Module	hub-genes
GSE130078	*lightcyan*	ANKRD20A8P, LINC01287, SOHLH1, CDH22, DHRS2, CRLF1, TENM1, EMILIN3, ADGRL3, AGGF1P8, CCDC144NL, RHBDL1, HCG23, LOC105370792, ADAMTS20, RPL31P25, RBMS3, TESMIN, OR11J2P, NFIB.
	steelblue	GMNN, RFC5, TMTC1, UBE2T, LIMK1, OSR2, CLUAP1, HMGB3, DTL, DNA2, LMO4, SENP1, ZNF367, CDK4, EXO1, MSH2, SUMO3, ARL4A, H1-2, TMEM270.
	skyblue3	MANEAL, CCT2, PCSK1, GNS, ZNF737, ZNF85, PANK2, NAT8B, TBK1, TBC1D15, SYT15, MON2, CXCL13, ZNF91, TDRD1, NEXMIF, TMBIM4, DLGAP1-AS5, RHOXF1-AS1, MUCL1.
	violet	GRID2IP, ZNF568, ZNF239, PIWIL1, HPDL, ZNF233, ELP6, ZNF470, MST1L, ZNF232, ZNF790-AS1, LRP6, ESRRG, CFAP91, ZNF829, THUMPD3, GSE1, LINC01205, ZNF667, KIF15.
	magenta	FAM155A, SORCS1, IGFN1, ZAN, ACAN, XIRP1, CACNA1B, DNAH10, EPHA3, CDH4, PCDH10, CACNA1E, RNF112, ST6GALNAC5, TGM4, DSCAM, CFAP61, CDH23, FNDC1, KCNH3.
	bisque4	CHPF, TRAM2, IGFBP3, CXCL16, PIGT, CRELD2, SEPTIN9, MFHAS1, TOR3A, PDIA4, CARMIL1, MOGS, ORAI2, CLPTM1L, ARSB, CHST15, ARFGAP1, ST6GAL1, CDK18, CSF2RB.

4.5 Validation

We achieve validation through various approaches. Foremost we validate whether the DEGs and MoIs identified by our framework are biologically relevant are highly enriched. 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. Furthermore, we identify TEDs through enrichment analysis (Section4.3.4). All hub-genes of the biologically relevant MoIs and the TEDs 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 existing literature that establish 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.

4.5.1 Enrichment Analysis of DEGs and Modules

For a MoI to be regarded as Gene Ontology (GO) or pathway enriched, at least one gene in the module must be assigned to an enriched GO term or pathway, respectively with a significance of 5% (i.e., $p \leq 0.05$). To perform functional enrichment analysis, we use the online tool DAVID [628, 253] (Section 2.2.3).

Tab. 4.5: Percentages of genes in each MoI 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
			(%)	(%)	(%)	(%)				(%)	(%)	(%)	(%)
GSE130078	*lightyellow*	240	78.9	80.5	78.9	38.3	GSE20347	*pink*	276	90.1	96.0	95.2	47.2
	red	759	83.3	87.6	84.8	38.9		*greenyellow*	149	98.4	98.4	97.7	49.2
	skyblue3	104	82.6	87.2	85.3	46.8		*darkgreen*	201	95.6	96.7	95.6	55.0
	steelblue	145	73.1	76.6	71.9	26.9		*lightyellow*	122	95.0	96.0	97.0	53.5
	violet	142	82.0	84.7	80.0	37.3		*lightsteelblue1*	143	95.2	94.4	94.4	49.2
	lightcyan	321	69.1	70.7	69.4	27.7		*black*	775	94.9	97.0	96.1	52.9
	bisque4	1000	87.5	91.4	89.1	40.5		*lightgreen*	123	90.7	98.1	92.6	56.5
	magenta	249	84.4	89.9	84.9	31.0							
GSE23400	*greenyellow*	172	97.7	98.3	96.5	63.4							
	magenta	231	94.8	96.5	93.4	44.5							
	purple	225	96.6	98.1	96.3	58.2							

Table 4.5 summarizes the percentages of genes in the MoI annotated to enriched GO terms as well as enriched KEGG pathways. We observe that all MoIs identified by our framework are GO and pathway enriched.

4.5.1.1 Biomarker Candidate Genes (BCG)

As mentioned earlier, we select DEGs as candidates for potential biomarkers based on the following two criteria:

1. All hub-genes detected by the DCA unit of our framework in all MoIs are BCGs.
2. DEGs that have been annotated to the most enriched GO terms in all three GO databases (BP, CC and MF) and are also annotated to the most enriched pathway after GO and Pathway enrichment analysis on the entire dataset are also considered as potential biomarkers. We rename these DEGs as TEDs (Top Enriched DEGs)

Thus, alongside all DEGs that are among Top 20 hub-genes in MoIs (as summarized in Table 4.4), our second criterion adds 22, 18 and 11 TEDs to the list of candidate genes in GSE20347, GSE23400 and GSE130078, respectively. We summarize these DEGs (TEDs) in Table 4.6. The numbers of BCGs for GSE20347,GSE23400 and GSE130078 increase from 140, 60 and 160 to 162, 78 and 171, respectively.

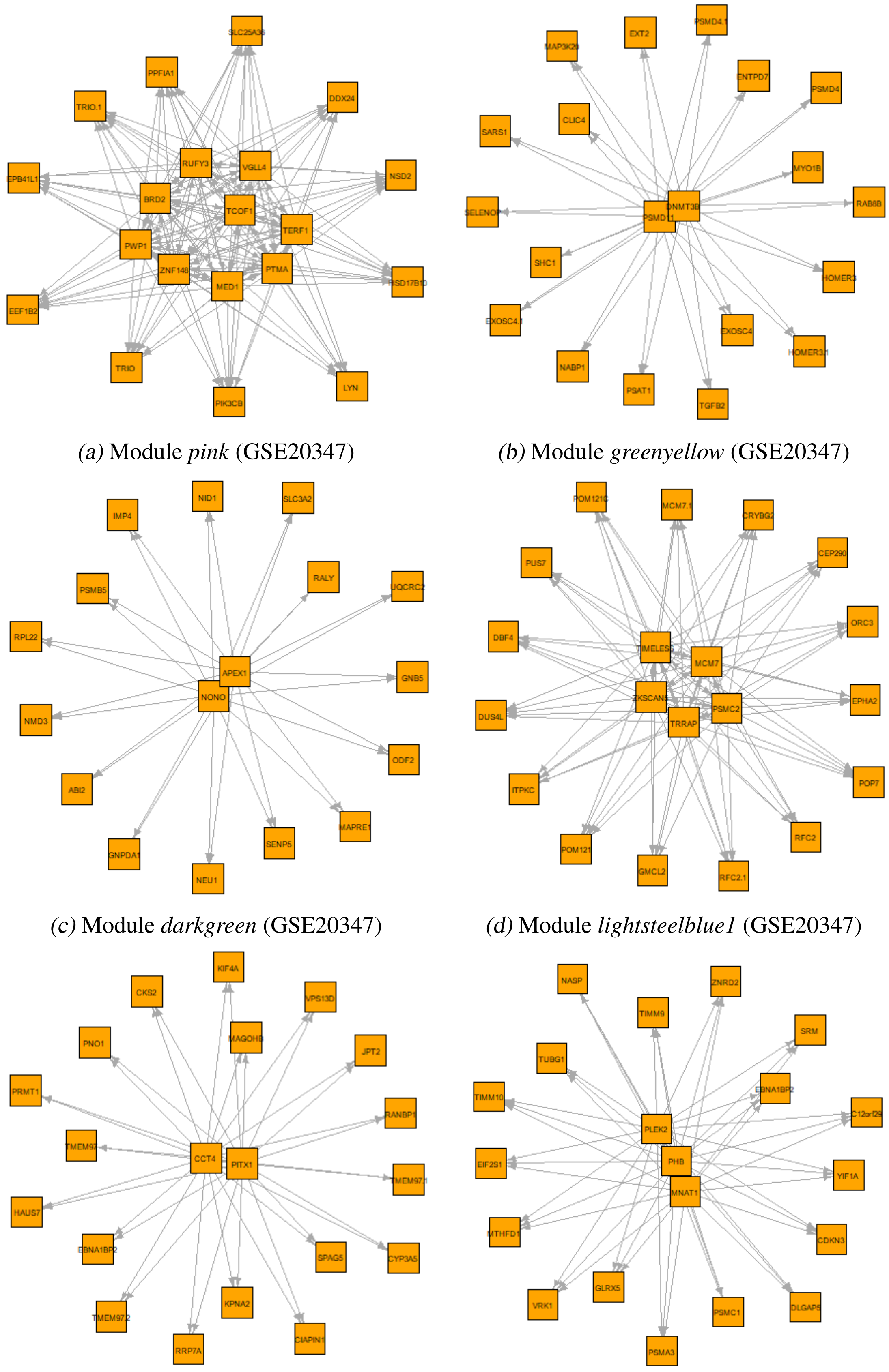
Tab. 4.6: DEGs that are annotated to most enriched GO term in all three GO databases (BP, CC and MF) as well as the most enriched pathway. DEGs with strong literature evidence of association to ESCC are marked in Red while hub-genes with evidence of association to five other SCCs, LaSCC, LSCC, HNSCC, OSCC, and TSCC, are marked in Blue

Dataset	Top Enriched DEGs
GSE20347	*TXNRD1, APPL1, FADD, FAS, MAPK1, PIK3R1, STAT1, RAF1, RARA, MAP2K1, PIK3CD, RAC2, MAPK10, PRKACB, AR, PIK3CB, BCR, KRAS, GSK3B, NFKB2, PIK3R2, FLT3LG*
GSE23400	*RAF1, PIK3R1, APPL1, MAP2K1, AR, PRKCB, PRKACB, STAT1, HIF1A, TXNRD1, FADD, RARA, PIK3CD, IL15, RAC2, GSK3B, STAT2, BCR*
GSE130078	*TYMP, PDE4A, PIK3CD, PIP5K1A, GPI, PDE1B, PDE3B, PDE9A, HPGDS, PDE3A, PI4KA*

We perform the enrichment analysis on the entire dataset or in more specific terms the list of all genes in the dataset. This leads to the observation that as the lists of genes in GSE20347 (22,278 genes) and GSE23400 (22,283 genes) are almost the same, the list of top enriched genes (35 genes) extracted are the same. However, the differences in TEDs are seen (Table 4.6) due to the fact that there are DEGs identified in one dataset that might not be detected in the other.

4.5.2 Biological Analysis

To establish the biological relevance of the BCGs detected by our method, we use functional enrichment analysis and the construction of a gene regulatory network (GRN). Transcription Factors (TF) have remarkable diversity as well potency as drivers of cell transformation. Bhagwat et al. [45] justify the continued pursuit of TFs as potential biomarkers across many forms cancer by the prevalent deregulation of the same. We observe that 26 (hub-genes:21, TEDs:5), 11 (hub-genes:6, TEDs:5) and 23 (hub-genes:23, TEDs:0) BCGs detected by our method in GSE20347, GSE23400 and GSE130078, respectively are TFs. These TFs exhibit regulatory behavior in their respective modules, establishing their biological relevance. For easy visualization, we extract a manageable subset of hub-genes from the non-preserved modules detected by our method (Fig 4.5a-4.6f). We construct a Gene Regulatory Network (GRN) with these hub-genes and associated Transcription Factors (TFs) so as to observe the regulatory behavior of the corresponding genes. The resulting GRN is in the form of an adjacency list with weighted directed edges from TFs to other target genes (TGs).



(a) Module *pink* (GSE20347)

(b) Module *greenyellow* (GSE20347)

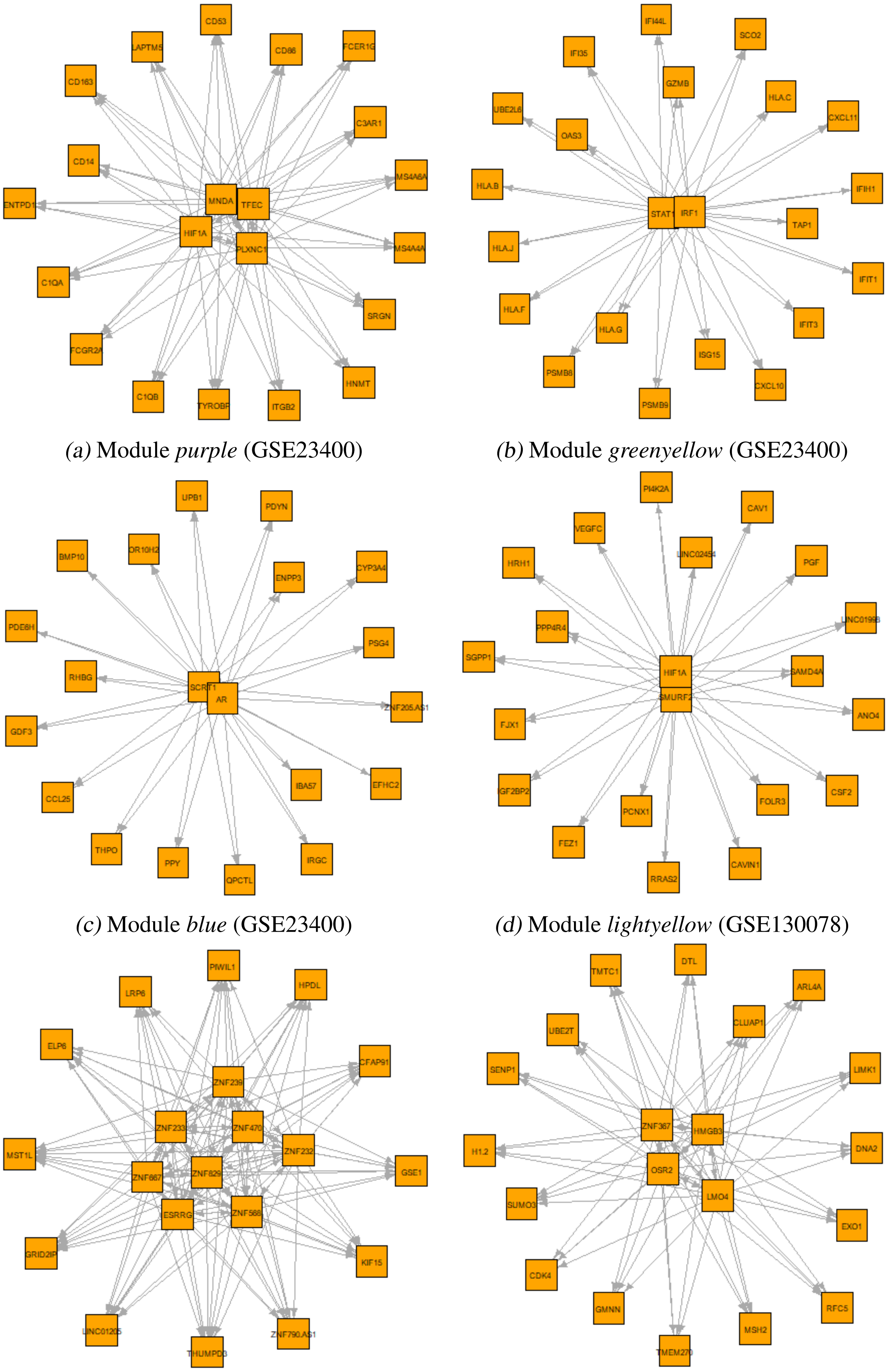
(c) Module *darkgreen* (GSE20347)

(d) Module *lightsteelblue1* (GSE20347)

(e) Module *black* (GSE20347)

(f) Module *magenta* (GSE130078)

Fig. 4.5: GRN for normal module a) *pink* and disease modules b) *greenyellow* in GSE20347, disease modules c) *darkgreen*, d) *lightsteelblue1* e) *black* in GSE20347. GRN for disease module f) *magenta* in GSE23400.



(a) Module *purple* (GSE23400)

(b) Module *greenyellow* (GSE23400)

(c) Module *blue* (GSE23400)

(d) Module *lightyellow* (GSE130078)

(e) Module *violet* (GSE130078)

(f) Module *steelblue* (GSE130078)

Fig. 4.6: GRN for normal modules a) *purple* and b) *greenyellow* in GSE20347, and disease modules c) *blue* in GSE23400. GRN for disease modules d) *lightyellow* e) *violet*, and f) *steelblue* in GSE130078.

Tab. 4.7: Summary of BCGs detected by Integrative DEA 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	*PRKACB, BCR, AR, LYN, APPL1, STAT1, SHC1, NFKB2, PIK3CD, PIK3CB, PIK3R2, PIK3R1, EXT2, RAC2, PPFIA1, RAF1, GNB5, GSK3B, MAPK10, KRAS, RARA, MAP2K1, TXNRD1, FLT3LG, FADD, RANBP1, FAS, MAPK1*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*AR, HOXC11, STAT1, PITX1, NFKB2, TP63, PIK3R2, PIK3R1, PTMA, RAF1, MED1, ZNF148, RARA, FADD, CXCR3*
	GO:0016032 viral process	*ITSN2, POM121, POM121C, LYN, BRD2, PSMB5, STAT1, ABI2, SHC1, PIK3R1, FADD, RANBP1, MAPK1*
GO_CC	GO:0005829 cytosol	*ITSN2, PRKACB, BCR, CASP10, MSRA, RAC2, PTMA, RAF1, UBA7, HOMER3, RARA, MAP2K1, MAP3K20, MCL1, MCM7, UBASH3A, AR, LYN, APPL1, STAT1, KPNA2, PIK3CD, PIK3CB, PIK3R2, PIK3R1, TRIO, RPL22, NEDD4L, PNO1, ABI2, NEB, SARS1, TXNRD1, FADD, CLIC4, PRMT1, FAS, SENP5, EPB41L1, HOXC11, SHC1, ODF2, CEP290, EEF1B2, NABP1, GNB5, KRAS, HAUS7, PSAT1, FLT3LG, MAPK1, PSMD4, RUFY3, PSMB5, TCOF1, PSMC2, NFKB2, JPT2, EXOSC4, CCT4, PPFIA1, SBF1, KIF4A, GSK3B, MAPK10, HPRT1, ITPKC, ENOX2, SPAG5, MAPRE1, RANBP1*
	GO:0005654 nucleoplasm	*PRKACB, CIAPIN1, TP63, MSRA, PTMA, UBA7, SLC3A2, RRP7A, RARA, ORC3, MCL1, MCM7, TIMELESS, NONO, UBASH3A, AR, STAT1, KPNA2, PIK3CB, DBF4, MED1, NEDD4L, PNO1, ABI2, DNMT3B, TXNRD1, PRMT1, SENP5, POP7, MAGOHB, HOXC11, NABP1, TERF1, NSD2, RFC2, NPIPB3, RAB8B, MAPK1, PSMD4, UQCRC2, PSMB5, TCOF1, PSMC2, NFKB2, IMP4, EXOSC4, CCT4, KIF4A, GSK3B, MAPK10, TRRAP, POM121, ZNF148, BRD2, ESF1, NMD3*
	GO:0016020 membrane	*TFRC, BCR, ITGB7, CEP290, NEU1, EXT2, RAC2, GNAQ, SLC3A2, MAN1C1, KRAS, MCL1, MCM7, FLT3LG, NONO, LSS, RUFY3, APPL1, PSMC2, KPNA2, PIK3CD, PIK3CB, ENTPD7, PIK3R1, CD52, CD48, SBF1, KIF4A, DDX24, MED1, LST1, NMD3, CLIC4, FAS*
GO_MF	GO:0005515 protein binding	*TFRC, PRKACB, BCR, ZKSCAN5, NEU1, MSRA, RAC2, RAF1, RALY, CD2, HOMER3, RRP7A, APOOL, RARA, MAP2K1, MCL1, MCM7, TIMELESS, UBASH3A, AR, APPL1, STAT1, PITX1, PIK3CD, PIK3CB, HSD17B10, EPHA2, PIK3R2, PIK3R1, RCN1, MED1, RPL22, NEDD4L, PNO1, ABI2, DNMT3B, TXNRD1, GMCL2, FADD, ITM2A, CLIC4, PRMT1, FAS, TGFB2, POP7, EPB41L1, MAGOHB, HOXC11, CNPY2, SHC1, CEP290, MAN1C1, KRAS, TERF1, RFC2, PSAT1, FLT3LG, RAB8B, VGLL4, MAPK1, UQCRC2, CKS2, TCOF1, NFKB2, IMP4, EXOSC4, KIF4A, GSK3B, MAPK10, TNS1, RANBP1*
	GO:0042802 identical protein binding	*TFRC, PSMD4, APPL1, STAT1, CEP290, TP63, CD3D, RAF1, RALY, CD2, KRAS, HOMER3, RPL22, HPRT1, TERF1, ABI2, PSAT1, FADD, PRMT1, MAPRE1, FAS, TIMELESS, NONO, MAPK1*
	GO:0003723 RNA binding	*POP7, TFRC, PSMD4, MAGOHB, TCOF1, KPNA2, HSD17B10, CCT4, NABP1, DDX24, SLC3A2, RALY, RPL22, RRP7A, PNO1, EBNA1BP2, MAP3K20, SARS1, ENOX2, ESF1, TNS1, NMD3, PRMT1, MAPRE1, PUS7, NONO*

Tab. 4.8: Summary of BCGs detected by Integrative DEA 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	*BCR, AR, APPL1, STAT1, STAT2, CXCL10, CXCL11, PIK3CD, PIK3R1, RAC2, CD53, IL15, TYROBP, RAF1, GSK3B, PRKCB, HIF1A, RARA, MAP2K1, VRK1, TXNRD1, FADD*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*AR, STAT1, STAT2, CXCL10, PIK3R1, RAF1, HIF1A, RARA, FADD, TFEC*
	GO:0045893 positive regulation of transcription	*PRKCB, AR, HIF1A, RARA, STAT1, MAP2K1, CD86*
GO_CC	GO:0005829 cytosol	*BCR, MTHFD1, SCO2. SRM, GZMB, RAC2, RAF1, PRKCB, IFIH1, RARA, MAP2K1, IFIT1, IFIT3, OAS3, AR, APPL1, STAT1, STAT2, HNMT, PIK3CD, UBE2L6, PIK3R1, LAPTM5, CD163, TXNRD1, MNDA, FADD, DLGAP5, CDKN3, IL15, VRK1, EIF2S1, PSMA3, PSMB8, PSMB9, PSMC1, IFI35, ISG15, GSK3B, HIF1A, TUBG1*
	GO:0005654 nucleoplasm	*MNAT1, PRKCB, RARA, OAS3, AR, STAT1, STAT2, HNMT, UBE2L6, TXNRD1, MNDA, IL15, NASP, VRK1, PSMA3, PSMB8, PSMB9, PSMC1, ISG15, GSK3B, HIF1A, TFEC*
	GO:0016020 membrane	*BCR, MTHFD1, ITGB2, GZMB, RAC2, PLXNC1, HLA-C, HLA-F, HLA-G, EIF2S1, OAS3, APPL1, PSMC1, PIK3CD, IFI35, ENTPD1, PIK3R1, TAP1, CD163*
GO_MF	GO:0005515 protein binding	*BCR, SCO2, MNAT1, SRM, RAC2, FCER1G, RAF1, PRKCB, RARA, MAP2K1, AR, APPL1, STAT1, STAT2, PIK3CD, UBE2L6, ENTPD1, PIK3R1, GLRX5, CD163, TXNRD1, FADD, IL15, TIMM9, PLXNC1, VRK1, EIF2S1, FCGR2A, IFI35, GSK3B, HIF1A, TUBG1*
	GO:0042802 identical protein binding	*APPL1, STAT1, STAT2, SRM, IFI35, CD53, FCER1G, TYROBP, RAF1, IFIH1, IFIT3, HLA-G, FADD, TUBG1, C1QB*
	GO:0003723 RNA binding	*PSMC1, IFIH1, EBNA1BP2, IFIT1, IFIT3, EIF2S1*

Tab. 4.9: Summary of BCGs detected by Integrative DEA 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	*PIP5K1A, RHBDL1, RRAS2, CXCL16, HPGDS, PIK3CD, PIK3CB, VEGFC, RIT1, RALB, CDC42SE1, HIF1A, GPI, PGF, CDK4, PDE1B, PDE2A, PDE3A, TENM1, TYMP, PDE9A, CSF2RB, PI4KA, PI4KB, PDE3B, PDE4A, PDE4D, PPP2R5A, PDE5A, GNA13, LIMK1*
	GO:0000122 negative regulation of transcription from RNA polymerase II promoter	*MEF2A, H1-2, ZNF239, ZNF85, OSR2, CAV1, NFIB, ZNF568, PDE2A*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*TBK1, MEF2A, LRP6, OSR2, ZNF91, HIF1A, NFIB, ESRRG, LMO4, SENP1*
GO_CC	GO:0005886 plasma membrane	*FEZ1, KCNH3, ARL4A, SDE2, RIT1, RALB, PI4K2A, COA6, TENM1, CDH22, CDH23, RRAS2, CXCL16, PIK3CD, PIK3CB, LRP6, EPHA3, CAVIN1, CARMIL1, PDE2A, PDE9A, PDE4A, PDE4D, PIP5K1A, ADGRL3, EXO1, SYT15, CACNA1B, CACNA1E, CAV1, GPI, CSF2, HRH1, PI4KA, GNA13, ZAN, PCDH10, ANO4, CDC42SE1, CDH4, MUCL1*
	GO:0005829 cytosol	*ARFGAP1, ARL4A, SDE2, PI4K2A, ELP6, TYMP, NEXMIF, SAMD4A, HPGDS, PIK3CD, PIK3CB, DTL, EPHA3, CAVIN1, CHPF, PPP6C, PDE1B,CARMIL1, PDE2A, PDE3A, PDE9A, PDE3B, PDE4A, PDE4D, PDE5A, TBK1, PIP5K1A, CRLF1, GMNN, IGF2BP2, MON2, GPI, POLR2D, PANK2, HRH1, PI4KA, PI4KB, PPP2R5A, RBMS3, GNA13, LIMK1, PPP4R4, MEF2A, THUMPD3, KIF15, CCT2*
	GO:0005654 nucleoplasm	*MSH2, ARL4A, SDE2, ZNF85, ERCC3, NFIB, COA6, NEXMIF, HPGDS, PIK3CB, SUMO3, DTL, EPHA3, CAVIN1, APPBP2, RNF112, ESRRG, PPP6C, RBBP5, CARMIL1, PDE9A, CTDSPL2, PDE4A, SENP1, TBK1, PIP5K1A, LIN52, ZNF470, GMNN, EXO1, UBE2T, GPI, ZNF367, DHRS2, RFC5, POLR2D, DNA2, MEF2A, ZNF232, CLUAP1, HIF1A, CDK4*
GO_MF	GO:0005515 protein binding	*ORAI2, KCNH3, MSH2, PIWIL1, SDE2, RALB, XIRP1,GSE1, TYMP, RRAS2, HPGDS, MFHAS1, PIK3CD, PIK3CB, DTL, LRP6, EPHA3, CRELD2, APPBP2, RNF112, ESRRG, RBBP5, PDE1B, PDE2A, PDE3A, IGFN1, PDE9A, SOHLH1, PDE3B, PDE4A, PDE4D, PDE5A, ACAN, PIP5K1A, LIN52, CRLF1, TMBIM4, VEGFC, CACNA1B, UBE2T, GNS, GPI, RFC5, DNA2, PI4KA, PI4KB, PPP2R5A*
	GO:0042802 identical protein binding	*TBK1, CAV1, PGF, EMILIN3, LRP6, PCSK1, CAVIN1, ESRRG, PDE2A, PDE9A*
	GO:0004712 protein serine/threonine/tyrosine kinase activity	*TBK1, CDK18, CDK4, EPHA3, LIMK1*

Tab. 4.10: Summary of BCGs detected by our method, Integrative DEA, in the two microarray and one RNA-Seq datasets that have been annotated to the top 5 KEGG enriched pathways

	KEGG Pathways	Annotated BCGs
GSE20347	hsa05200:Pathways in cancer	*BCR, AR, APPL1, CKS2, RARA, STAT1, MAP2K1, NFKB2, PIK3CD, PIK3CB, TXNRD1, FLT3LG, PIK3R2, PIK3R1, FADD, RAC2, FAS, RAF1, TGFB2, GNAQ, GNB5, GSK3B, MAPK1, MAPK10, KRAS*
	hsa04010:MAPK signaling pathway	*PRKACB, FLT3LG, EPHA2, RAC2, FAS, MAP2K1, RAF1, MAP3K20, TGFB2, NFKB2, MAPK1, MAPK10, KRAS*
	hsa05169:Epstein-Barr virus infection	*PSMD4, PIK3R2, PIK3R1, FADD, LYN, CD3D, STAT1, PSMC2, FAS, NFKB2, PIK3CD, PIK3CB, MAPK10*
	hsa04151:PI3K-Akt signaling pathway	*FLT3LG, EPHA2, PIK3R2, PIK3R1, MAP2K1, RAF1, MCL1, GNB5, GSK3B, PIK3CD, PIK3CB, ITGB7, MAPK1, KRAS*
	hsa05171:Coronavirus disease - COVID-19	*RPL22, PIK3R2, PIK3R1, STAT1, PIK3CD, PIK3CB, MAPK1, MAPK10*
GSE23400	hsa05200:Pathways in cancer	*PRKCB, BCR, AR, HIF1A, APPL1, RARA, STAT1, STAT2, MAP2K1, PIK3CD, TXNRD1, PIK3R1, FADD, RAC2, IL15, RAF1, GSK3B*
	hsa04010:MAPK signaling pathway	*PRKCB, CD14, RAC2, MAP2K1, RAF1*
	hsa04151:PI3K-Akt signaling pathway	*PIK3R1, MAP2K1, RAF1, GSK3B, PIK3CD*
	hsa05169:Epstein-Barr virus infection	*ENTPD1, HLA-C, HLA-F, HLA-G, PIK3R1, FADD, STAT1, STAT2, PSMC1, TAP1, ISG15, CXCL10, PIK3CD, OAS3*
	hsa05171:Coronavirus disease - COVID-19	*C3AR1, PRKCB, PIK3R1, IFIH1, C1QB, C1QA, STAT1, STAT2, ISG15, CXCL10, PIK3CD, OAS3, FCGR2A*
GSE130078	hsa01100:Metabolic pathways	*PIP5K1A, PIGT, HPGDS, PIK3CD, PIK3CB, ST6GALNAC5, ST6GAL1, MOGS, ARSB, NAT8B, PI4K2A, GNS, GPI, CHPF, PANK2, PDE1B, PDE2A, PDE3A, TYMP, PDE9A, PI4KA, PI4KB, PDE3B, PDE4A, PDE4D, PDE5A, SGPP1*
	hsa05200:Pathways in cancer	*HIF1A, PGF, MSH2, CDK4, PIK3CD, PIK3CB, LRP6, VEGFC, CSF2RB, RALB, GNA13*
	hsa04144:Endocytosis	*PIP5K1A, CAV1, ARFGAP1, FOLR3*
	hsa04010:MAPK signaling pathway	*RRAS2, VEGFC, PGF, CACNA1B, CACNA1E*
	hsa05165:Human papillomavirus infection	*TBK1, CDK4, PIK3CD, PIK3CB, PPP2R5A*

As in the the case of validation of modules, we employ DAVID [628, 253] to perform functional enrichment analysis of all BCGs detected by our method. A BCG can be regarded as GO enriched considering a GO database (GO_BP, GO_CC, GO_MF) if it is annotated to at least one GO term in that database with significance of 5% ($p \leq 0.05$). Table 4.7, Table 4.8 and Table 4.9 summarize the BCGs annotated to the top 3 GO terms in each GO database in GSE20347, GSE23400 and GSE130078, respectively. Similarly, a BCG is KEGG pathway enriched if it is annotated to at least one KEGG pathway term with significance of 5%. Table 4.10 summarizes the BCGs annotated to top 3 enriched KEGG pathways in GSE20347, GSE23400 and GSE130078.

4.5.3 Literature Trace

Following are the literatures where BCGs identified by our framework in all three datasets have evidence of association to ESCC.

- Caveolin-1 (CAV1) is a biomarker for ESCC, according to Kato et al. [301], Ando et al.[26], and Jia et al.[283].
- Yu et al. [835] and Wang et al. [738] identified that cyclin-dependent kinase inhibitor 3 (CDKN3) controls tumour growth in ESCC via activating the AKT signalling pathway. According to Liu et al., [424] CDKN3 behaved as an oncogene in human ESCC.
- Cyclin-dependent kinase 4 (CDK4) amplification was discovered to be a distinct prognostic factor for survival, which could be incorporated into the tumor-node-metastasis staging system to improve risk stratification of patients with ESCC, according to Huang et al. [255].
- Carbohydrate sulfotransferase 15 (CHST15), according to Wang et al. [744], stimulates the growth of TE-1 cells in ESCC through a variety of mechanisms.
- According to Zheng et al. [904], cytokine induced apoptosis inhibitor 1 (CIAPIN1) expression was statistically correlated with the degree of differentiation, depth of invasion, and lymph node metastasis of ESCC and has since been regarded as an important prognostic indicator in ESCC.
- According to Kita et al. [310], cyclin-dependent kinase subunit 2 (CKS2) expression in ESCC was higher than it was in normal tissue, and CKS2 overexpression is linked to the depth of the tumor's lymphatic invasion, clinical stage, distant metastasis, and a poor prognosis.
- Making use of Cox regression Canopy FGF Signalling Regulator 2 (CNPY2) was

shown to be useful in predicting ESCC outcomes by He et al [228] in their 2017 decision.

- High CXCL10 expression has the potential to be a clinically useful marker of the need for adjuvant chemotherapy after surgery in patients with advanced thoracic ESCC, according to Sato et al. [609]. This is because high CXCL10 expression is an independent prognostic factor.
- Component 3a Receptor 1 (C3AR1) may contribute to the development of an immunosuppressive microenvironment by influencing the polarization of macrophages to M2 phenotype thus leading to the progression of ESCC, according to Qu et al [564] .
- Dehydrogenase/reductase member 2 (DHRS2) was demonstrated by Zhou et al. [915] to play a significant role in the initiation and progression of ESCC.
- Disks large-associated protein 5 (DLGAP5) may promote cell poliferation in ESCC, according to preliminary research by Hu et al [245].
- In ESCC, Chen et al. [83] discovered that overexpression of DNA methyltransferase 3b (DNMT3b) is connected to increased STAT3 signalling and is the cause of more aggressive tumour growth and treatment resistance.
- EPH receptor A2 (EphA2) overexpression appears to be associated with a low degree of tumour differentiation and lymph node metastasis in ESCC, according to Miyazaki et al. [504].
- According to Chen et al. [84], silencing EPH receptor A3 (EphA3) in *KYSE410* cells causes the epithelial-mesenchymal transition and promotes cell migration and invasion in ESCC.
- Estrogen-related receptor gamma (ESRRG) is one of four molecular markers that may be useful in the diagnosis and therapy of ESCC, according to Xu et al. [795].
- Bolidong et al. [52] proposed that glycogen synthase kinase 3 beta (GSK3B) has a tumor-promoting effect in ESCC via *cyclin D1/CDK4-mediated cell cycle* progression. Gao et al. [182] established that GSK3$beta$-STAT3 signalling could be a viable therapeutic target for ESCC treatment since GSK3B expression enhances ESCC progression through STAT3 in vitro and in vivo.
- Hypoxia-inducible factor 1 alpha (HIF1A), p53, and vascular endothelial growth factor (VEGF) are significant variables that promote tumour progression, according Shao et al. [619]. Study done by Hu et al. [251] revealed that HIF1A enhances ESCC

metastasis by targeting SP1 in a hypoxic microenvironment, according to the findings.

- Human leukocyte antigen-F (HLA-F) antigen expression was shown to be related to survival in patients with ESCC by Zhang et al.[876].
- Human leukocyte antigen-G (HLA-G) expression in human ESCC has been proven by Yie et al. [827] to have a strong and independent prognostic value.
- Homer scaffolding protein 3 (HOMER3) is one of the three genes put forth as potential cancer-associated genes by Shen et al. [623] and may contribute to tumorigenesis in ESCC.
- High mobility group box 3 (HMBG3) has been shown by Gao et al. [180] to have potential as a molecular marker for ESCC patient prognosis prediction.
- Insulin-like growth factor 2 mRNA-binding protein 2 (IGF2BP2) plays a significant carcinogenic effect in ESCC, according to studies by Lu et al. [452] and Shu et al. [635].
- According to Luo et al.[458]'s research, insulin-like growth factor binding protein-3 (IGFBP3) knockdown gives resistance to the cell-killing effects of IR on ESCC both in vitro and in vivo. According to Zhao et al. [892], the elevated ESCC chemosensitivity may be dependent on IGFBP-3 upregulation via *EGFR-dependent* pathway. Additionally, Luo et al. [459] state that high levels of IGFBP3 expression in ESCC are associated with early clinical stages and are indicative of favourable patient outcomes after radiation.
- Interferon-stimulated gene 15 (ISG15) has been linked to the promotion of tumours in ESCC via *c-MET/Fyn/alpha-catenin pathway*, according to Yuan et al. [840].
- Kinesin family member 4A (KIF4A) was discovered by Wang et al. [730] as a facilitator of ESCC proliferation, cell cycle, migration, and invasion both in vivo and in vitro. Similar to this, Sun et al.[657] claimed that KIF4A regulates the biological function of ESCC cells through the Hippo signalling pathway, boosting ESCC cell proliferation and migration.
- 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.

- Myeloid cell leukemia 1 (MCL-1) has been shown to contribute to the development of ESCC by Yu et al. [838].
- According to research done by Qiu et al. [563], MCM7 (maintenance complex component 7) aids in the proliferation of tumour cells, colony formation, and ESCC cell migration via activating the AKT1/mTOR signalling pathway. Further recommendations from Choy et al. [102] and Zhong et al. [906] pointed to MCM7 as a more accurate proliferation marker for assessing and forecasting various clinical outcomes of ESCC, respectively.
- MutS homolog 2 (MSH2) methylation in the plasma was suggested by Ling et al. [408] to be a reliable indicator of DFS for these ESCC patients prior to oesophagectomy.
- Findings by Cheng et al. [94] revealed that non-POU domain containing octamer binding (NONO) plays a significant role in numerous biological features of ESCC through activation of the Akt and Erk1/2 signalling pathways.
- By regulating lipocalin 2 (LCN2), Wang et al. [719] identified pleckstrin-2 (PLEK2) as the primary factor causing metastasis and chemoresistance in ESCC.
- Lower processing of precursor 7 (POP7) expression is associated with a worse prognosis in esophageal cancer, according to Yang et al. [819].
- The expression of PTPRF interacting protein alpha 1 (PPFIA1) is significantly increased and is associated with some malignant clinical features and poor outcomes in ESCC patients, according to Tang et al. [674], establishing it as a valuable biomarker for early detection, treatment planning, and prognosis evaluation for ESCC.
- According to Zhao et al. [900], protein arginine methyltransferase 1 (PRMT1) mediates transcriptional modification through histone H4 arginine methylation, which activates and maintains esophageal TICs. According to Zhou et al. [913], PRMT1 has an oncogenic function in the development of ESCC by activating Hedgehog signalling and up-regulating the expression of target genes that are downstream of Hedgehog signalling.
- Phosphoserine aminotransferase 1 (PSAT1) expression was found to be higher in ESCC tissues compared to normal esophageal tissues, and Liu et al. [411] found that this increase is significantly correlated with disease stage, lymph node metastasis, distant metastasis, and poor prognosis.
- A potential target for the immuno-oncology action of proteasome 20S subunit alpha

3 (PSMA3) in ESCC therapy was offered by Liu et al. in their publication [422].

- According to Ma et al. [464], overexpression of proteasome 26S subunit non-ATPase 4 (PSMD4) accelerates the development of ESCC.
- Prothymosin alpha (PTMA) is presented as a viable candidate for ESCC because Zhu et al. [920] emphasize that PTMA expression was up-regulated in ESCC tissues.
- A tumor-suppressive role for the RNA binding motif single stranded interacting protein 3 (RBMS3) gene in ESCC was proposed by Li et al. [375].
- According to Feng et al. [162], Ras-like without CAAX1 (RIT1) exhibits tumor-suppressing functions in ESCC. These functions were carried out by inhibiting the *MAPK* and *PI3K/AKT signalling* pathway, inhibiting EMT, and down-regulating the cancer stemness of ESCC cells.
- Signal transducer and activator of transcription-1 (STAT1) may act as a tumor suppressor in ESCC, according to Zhang et al [884].
- The clinically significant implications of the Transferrin Receptor (TFRC) were highlighted by Wada et al. [715], who came to the conclusion that it provides an independent prognostic factor.
- Ubiquitin conjugating enzyme E2 T (UBE2T) plays a role in the emergence of ESCC, and gene signatures formed from UBE2T-associated genes are prognostic in ESCC, as suggested by Wang et al. [743].
- Vascular endothelial growth factor C (VEGF-C) expression is correlated with lymph node metastases and a poor prognosis, according to Tanaka [670]. Similarly, as suggested by Kimura et al.[305], Vascular implies that VEGF-C expression in ESCC may play a significant role in lymphatic propagation.
- According to Jiang et al. [284], vestigial like family member 4 (VGLL4)'s down-regulation was crucial to the development of ESCC, and regaining its functionality could be a viable treatment for the disease.
- Vaccinia-related kinase 1 (VRK1), according to Liu et al. [439], increases CDDP resistance through *c-MYC* by activating *c-Jun* and amplifying a malignant phenotype in 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.

Tab. 4.11: Summary of potential biomarkers identified by our proposed framework, Integrative DEA. Here, All 3 under GO databases imply all three databases, BP, CC, and MF. HG: hub-gene. TED: Top Enriched DEG. +: Upregulated DEG and -: Downregulated DEG

	BCG	GO Database	Enriched Cancer Pathway(s)	HG/ TED	TF?	+/-	Literature Evidence
GSE20347	*CCT4*	All 3	Nil	HG	Yes (Fig 4.5e)	+	ESCC[152], HNSCC[141]
	CIAPIN1	BP, CC, MF	Nil	HG	No	+	ESCC[904]
	CKS2	All 3	hsa05200, hsa05222	HG	No	+	ESCC[310], TSCC[178]
	CLIC4	All 3	Nil	HG	No	+	HNSCC [797]
	CNPY2	All 3	Nil	HG	No	+	ESCC[228]
	DNMT3B	All 3	hsa01100	HG	Yes (Fig 4.5b)	+	ESCC[79], HNSCC[441], OSCC[83]
	EPHA2	All 3	hsa04010, hsa04151, hsa04015, hsa04014, hsa04360	HG	No	+	ESCC[504, 660], LSCC[668, 155], HNSCC[437, 582]
	HOMER3	All 3	hsa04068, hsa04724	HG	No	+	ESCC [623]
	HPRT1	BP, CC, MF	hsa01100	HG	No	+	OSCC[775], HNSCC[8]
	KIF4A	All 3	Nil	HG	No	+	ESCC[730, 657], OSCC[497]
	KPNA2	All 3	hsa05164, hsa03013, and hsa05207	HG	No	+	ESCC[475, 596], OSCC[403]
	KRAS	All 3	hsa05200, hsa05205, hsa05215, hsa05210, hsa05212, hsa05203, and 74 others	TED	No	-	LSCC[6]
	MCL1	All 3	hsa04151, hsa04210, and hsa04630	HG	No	+	ESCC[838], OSCC[483]
	MCM7	All 3	hsa04110, and hsa03030	HG	Yes (Fig 4.5d)	+	ESCC[563, 102, 906], OSCC[157]
	MYO1B	All 3	hsa05130	HG	No	+	HNSCC [530]
	NONO	All 3	Nil	HG	Yes (Fig 4.5c)	+	ESCC[94]
	PIK3CB	All 3	hsa05200, hsa05205, hsa05222, hsa05215, hsa05210, hsa05221, and 35 others	HG	No	+	OSCC [14]
	PIK3R2	All 3	Nil	TED	No	-	LSCC[706]
	POM121	All 3	hsa03013and hsa05014	HG	No	+	OSCC[468]
	POP7	All 3	Nil	HG	No	+	ESCC[819]
	PPFIA1	BP,CC	Nil	HG	No	+	ESCC [674], HNSCC [667]
	PRMT1	All 3	hsa04068 and hsa04922	HG	No	+	ESCC [900, 913], HNSCC[103]
	PSAT1	CC,MF	hsa01100and hsa01200	HG	No	+	ESCC[411]
	PSMC2	All 3	hsa05169, hsa05017, hsa05022, hsa05010, hsa05020, hsa05016, and 3 others	HG	Yes (Fig 4.5d)	+	OSCC[755]
	PSMD4	All 3	hsa05169, hsa05017, hsa05022, hsa05010, hsa05020, hsa05016, and 3 others	HG	No	+	ESCC [464]
	PTMA	All 3	Nil	HG	Yes (Fig 4.5a)	+	ESCC [920], HNSCC [688]

	BCG	GO Database	Enriched Cancer Pathway(s)	HG/ TED	TF?	+/-	Literature Evidence
GSE20347	*RCN1*	All 3	Nil	HG	No	+	OSCC[699]
	SBF1	All 3	Nil	HG	No	-	HNSCC[918]
	SENP5	BP,CC	Nil	HG	No	+	OSCC[139]
	SHC1	All 3	hsa05220, hsa05225, hsa05226, hsa05224, ,hsa05214, hsa04510, and 14 others	HG	No	+	LSCC [390]
	SLC3A2	All 3	hsa04150 and hsa04216	HG	NO	+	OSCC[387]
	TFRC	All 3	hsa04066, hsa04145, hsa04640, hsa04144, and hsa04216	HG	No	+	ESCC[715]
	VGLL4	All 3	Nil	HG	Yes (Fig 4.5a)	+	ESCC [284]
GSE23400	*CDKN3*	All 3	Nil	HG	No	+	ESCC[835, 424, 738]
	CXCL10	All 3	hsa05169, hsa05171, hsa05164, hsa05160, hsa04060, hsa04062, and 6 others	HG	No	+	ESCC[609], TSCC[578], HNSCC[380]
	DLGAP5	All 3	Nil	HG	No	+	ESCC[245]
	HLA-F	All 3	hsa05169, hsa05165, hsa05166, hsa05163, hsa05167, hsa05203, and 11 others	HG	No	+	ESCC[876]
	HLA-G	All 3	hsa05163, hsa05167, hsa05170, hsa05203, hsa04218, hsa04145 and 8 others	HG	No	+	ESCC[827] , HNSCC[605, 54], OSCC[267]
	HIF1A	All 3	hsa05200, and hsa05205	TED	Yes (Fig 4.6a)	+	ESCC[619, 251], OSCC[603, 164], TSCC[388]
	IFIT1	All 3	hsa05160	HG	No	+	OSCC[552], HNSCC[353]
	IFIT3	All 3	Nil	HG	No	+	OSCC[552, 341]
	IFI44L	All 3	Nil	HG	No	+	OSCC[535]
	ISG15	All 3	hsa05169, hsa05171, hsa05165, and hsa04622	HG	No	+	ESCC[840] , OSCC[326, 872]
	PLEK2	BP,CC	Nil	HG	Yes (Fig 4.5f)	+	ESCC[719], HNSCC[726]
	PSMA3	All 3	hsa05017, hsa05022, hsa05010, hsa05020, hsa05016, hsa05012 and 2 others	HG	No	+	ESCC[422], OSCC[65]
	TAP1	BP, CC, MF	hsa05169,hsa05163, hsa05170, hsa04145, and hsa04612	HG	No	+	TSCC[32]
	VRK1	All 3	Nil	HG	No	+	ESCC[439], HNSCC[602]
GSE130078	*CAV1*	All 3	hsa04144, hsa04510, and hsa05205	HG	No	+	ESCC[301, 26], OSCC[590], HNSCC[294], TSCC[798]
	CDK4	All 3	hsa05200, hsa05165, hsa04151, hsa04530, hsa05224, and hsa04110	HG	No	+	ESCC [255], HNSCC[708, 320]
	CHST15	CC	Nil	HG	No	+	ESCC[744]
	CD163	All 3	Nil	HG	No	+	ESCC[247], HNSCC[689], OSCC[229, 735]
	CLPTM1L	BP,CC	Nil	HG	No	+	OSCC[215, 241]
	C3AR1	All 3	hsa05171, hsa04936, hsa04080, and hsa04610	HG	No	+	ESCC[564]
	DHRS2	BP,CC	Nil	HG	No	+	ESCC[915]
	ESRRG	All 3	Nil	HG	Yes (Fig 4.6e)	-	ESCC[795], LaSCC [627]

	BCG	GO Database	Enriched Cancer Pathway(s)	HG/ TED	TF?	+/-	Literature Evidence
GSE130078	*EPHA3*	All 3	Nil	HG	Yes	+	ESCC[84]
	EXO1	All 3	Nil	HG	No	+	HNSCC[448]
	FCGR2A	All 3	hsa05171, hsa05152, hsa04380, hsa05135, hsa04145, hsa05130 and 4 others	HG	No	+	ESCC[453], HNSCC[118, 482]
	GNA13	All 3	hsa05200 and hsa04270	HG	No	+	LSCC[515]
	HIF1A	All 3	hsa05200 and hsa05205	HG	Yes (Fig 4.6d)	+	ESCC[619, 251], OSCC[603, 164], TSCC [388]
	HMGB3	All 3	Nil	HG	Yes (Fig. 4.6f)	+	ESCC[180]
	HPDL	CC,MF	Nil	HG	No	+	TSCC[864]
	IGFBP3	All 3	Nil	HG	Yes	+	ESCC[458, 892, 459], OSCC[737, 599]
	IGF2BP2	All 3	Nil	HG	No	+	ESCC[452, 635], OSCC[911, 740]
	LMO4	All 3	Nil	HG	Yes (Fig 4.6f)	+	HNSCC[636], TSCC[325]
	LRP6	All 3	hsa05200, hsa05224	HG	No	+	OSCC[845]
	MSH2	All 3	hsa05200	HG	No	+	ESCC[408], HNSCC[550]
	ITGB2	All 3	hsa04015, hsa05166, hsa05152, hsa05146, hsa04810, hsa04145, and 12 others	HG	No	+	OSCC[879]
	RBMS3	All 3	Nil	HG	No	+	ESCC[375], LSCC[389]
	RFC5	All 3	Nil	HG	No	+	LSCC[731]
	RIT1	BP,CC	Nil	HG	No	-	ESCC[162]
	TBK1	All 3	hsa05165	HG	No	+	HNSCC[749]
	TRAM2	CC	Nil	HG	Yes	+	OSCC[175]
	UBE2T	All 3	Nil	HG	No	+	ESCC[743]
	VEGFC	All 3	hsa05200, hsa04010, hsa04151, hsa04015, hsa04510, and 2 others	HG	No	+	ESCC[670, 305], OSCC[571, 20]
GSE20347 & GSE23400	*AR*	All 3	hsa05200, hsa05215, and hsa05207	TED	Yes (Fig 4.6c)	-	OSCC[685, 432]
	FADD	All 3	hsa05200, hsa05169, hsa05165, hsa05161, hsa05162, hsa05164 and 19 others	TED	No	+	OSCC[99], HNSCC[195, 572]
	GSK3B	All 3	hsa05200, hsa05215, hsa05225, hsa05210, hsa05226, hsa05224, and 37 others	TED	No	+	ESCC[52, 182], OSCC[501, 492]
	MAP2K1	All 3	hsa05200, hsa05205, hsa05215, hsa05210, hsa05221, hsa05212, and 80 others	TED	No	-	HNSCC [279]
	RAF1	All 3	hsa05200, hsa05205, hsa05215, hsa05210,hsa05221, hsa05220, and 74 others	TED	No	-	OSCC[319]
	STAT1	All 3	hsa05200, hsa05212, hsa05169, hsa05171, hsa05165, hsa05161, and 22 others	TED	Yes (Fig 4.6b)	+	ESCC[884], HNSCC[312, 779]
	TXNRD1	All 3	hsa05200 and hsa05225	TED	No	+	OSCC[275], HNSCC[159]

In Table 4.11, we give a detailed summary of all DEGs that have been identified by our method as candidates for potential biomarkers for ESCC. In our method, 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 4.11. In the table, we also highlight the enriched GO terms and pathways to which the BCGs has been annotated. Furthermore, it also details whether the same is a hub-gene, a transcription factor (TF) or whether it is upregulated or down-regualted. A DEG is upregulated if $logFC > 0$ and downregulated when $logFC < 0$. We take into consideration $logFC$ values calculated by limma for the microarray datasets, and edgeR in the bulk RNA-Seq dataset.

4.6 Discussion

We employ our biomarker criteria (Section 2.5) to determine the potential biomarkers. Table 4.12 gives a summary of all the cases the BCGs are annotated to.

Tab. 4.12: Summary of potential ESCC biomarkers identified by Integrative DEA using the biomarker criteria (Section 2.5), Here, HG: Hub-gene, and TED: Top Enriched DEG.

	GSE20347		GSE23400		GSE130078	
	HG	TED	HG	TED	HG	TED
Case 1	DNMT3B, MCM7	STAT1		STAT1		HIF1A
Case 2	HOMER3, PSMD4, PSAT1, TFRC, MCL1, EPHA2, KPNA2, CKS2, PRMT1	GSK3B	HLA-F, HLA-G, CXCL10, ISG15, PSMA3, FCGR2A, C3AR1	GSK3B	CAV1, VEGFC, CDK4, MSH2	
Case 3	PTMA, VGLL4, NONO		PLEK2		HMGB3, ESRRG	
Case 4	PSMC2	AR		AR		

All BCGs that fall under Case 1 and Case 2 are considered potential biomarkers for ESCC because of existing evidence of association with ESCC in the form of other works while our biological validation of these genes establishes their relevance to their respective datasets. For BCGs that fall under Case 3, although there is strong literature evidence of association with ESCC, we have weak evidence of their biological relevance to their datasets. On the other hand, for BCGs that fall under Case 4, although we strongly validate their biological relevance to their datasets, there is only literature evidence of association with other SCCs related to ESCC. For both these cases, the candidates can be considered probable potential biomarkers, but need further in-depth analysis.

Top Enriched DEGs (TEDs), STAT1 and HIF1A detected in both microarray datasets (GSE20347 and GSE23400) and GSE130078, respectively, belong to Case 1. In GSE20347, two candidates DNMT3B and MCM7 also belong to Case 1. Thus, STAT1, HIF1A, DNMT3B and MCM7 are potential biomarkers for ESCC. GSK3B is a TED detected in both microarrays, and belongs to Case 2. In dataset GSE20347, 9 BCGs HOMER3, PSMD4, PSAT1, TFRC, MCL1, EPHA2, KPNA2, CKS2 and PRMT1 belong to Case 2, and thus are potential biomarkers for ESCC. Similarly, 7 BCGs HLA-F, HLA-G, CXCL10, ISG15, PSMA3, FCGR2A and C3AR1 are potential biomarkers for ESCC as they fall under Case 2. Four BCGs in the RNASeq dataset GSE130078, CAV1, VEGFC, CDK4 and MSH2 fall under Case 2 and are potential biomarkers for ESCC.

Three candidates genes in GSE20347, PTMA, VGLL4 and NONO 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 can still be regarded as probable potential biomarkers for ESCC, but need further in-depth validation. Similarly in GSEE23400 and GSE130078, one (PLEK2) and 2 (HMGB3 and ESRRG) genes fall under Case 3. PSMC2 detected in GSE20347, on the other hand falls under Case 4. We validate its strong association with the dataset as this BCG has been annotated to GO terms in all three GO databases as well as several enriched pathways. They further exhibit regulatory behavior in a GRN, but there are no previous works that relate the same to ESCC. However, its worth mentioning that there is literature evidence that identify PSMC2 as potential biomarker for OSCC. Similarly, the TED identified in the two microarray datasets, AR, also falls under Case 4. Both, PSMC2 and AR are probable potential biomarkers for ESCC, but need further in-depth analysis.

In Table 4.13, we put forward a comparison between our work and two recent works presented by Patowary et al. [541] and Hu. et al.[245] that perform DEA by employing approaches and methods similar to our work.

Tab. 4.13: Comparison of our method, Integrative DEA with two recent works that employ DEA on ESCC datasets

Parameter	Our Method	Patowary et al. [541], 2020	Hu. et al.[245], 2020
Datasets	Microarray: GSE20347 and GSE23400, RNA-Seq: GSE130078	Microarray: GSE20347 and GSE23400, RNA-Seq: SRP064894	Microarray: GSE20347 and GSE26886
Identification of DEGs	DEA methods: Limma, SAM, EBAM, limma-voomm, edgeR and DESeq2 with *p-value* ≤ 0.05	DEA methods: Limma, SAM, EBAM, limma-voomm, edgeR and DESeq2 with *p-value* ≤ 0.01	Limma; $\lvert log2$fold change (FC)$\rvert) > 2$ and *adjusted p-value* < 0.05
Consensus function	(a) Common genes given by the methods with *p-value* ≤ 0.05 (b) For each method, DEGs not in common genes with *q-value* ≤ 0.05(RNA-Seq) or $lFDR \leq 0.05$ (Microarray)	(a) Common genes given by the methods with *p-value* $\leq$ 0.01 and (b) top ranked DEGs (other than common) given by each method with *p-value* ≤ 0.001	Up- and downregulated DEGs common to both microarray datasets
Module Extraction	Heiarchical Tree Classification; Eigenmodule selection and MEDissThres threshold merging	Heiarchical Tree Classification; Eigenmodule selection and MEDissThres threshold merging	(a) PPI network; Nodes with edge of > 20 as hub-genes; (b) WGCNA and Heiarchical Tree Classification
Preservation Analysis	Yes	Yes	No
Modules of Interest	All non-preserved modules of size larger that 100 nodes	Least Preserved module	Relevance between each module and hub-genes idenfied in PPI network < 0.5
hub-gene	Intramodular Connectivity	Intramodular Connectivity; Degree and confidence in PPI Network	Nodes with edge of > 20 in PPI network
Candidate Identification	Top 20 hub-genes in all modules of interest; DEGs that are annotated to most enriched GO term in all three GO databases as well as the most enriched pathway	DEG with highest intramodular connectivity in each MoI	Top hub-gene in MoIs
Enrichment Analyses	DAVID; GO and KEGG pathway	DAVID; GO and pathway	DAVID; GO and pathway and GSEA
Diseases considered for literature trace	ESCC and five other SCCs, namely oral SCC, Lung SCC, Tongue SCC, Head and Neck SCC, Laryngeal SCC	ESCC and all other cancers	ESCC and all other cancers
Identified potential biomarkers	*STAT1, HIF1A, DNMT3B, MCM7, GSK3B, HOMER3, PSMD4, PSAT1, TFRC, MCL1, EPHA2, KPNA2, CKS2, PRMT1, HLA-F, HLA-G, CXCL10, ISG15, PSMA3, FCGR2A, C3AR1, CAV1, VEGFC, CDK4 and MSH2*	*COL27A1, SOX11, BAG6, TOP3, CDC6, EZH2, COL7A1, G6PD and AKR1C2*	*DLGAP5*

4.7 Chapter Summary

The proposed framework has been found successful in identifying several interesting differentially expressed genes (DEGs) with a *p-value* of 0.05. Our consensus function that uses *lFDR* (for microarray) and *q-value* (for RNA-Seq) also takes into account the information loss caused by the DEGs that are common to all three methodologies. We investigated the behavioral alterations among the DEGs in both normal and disease conditions using Differential Co-expression (DCE) analysis and preservation analysis. All reasonably sized non-preserved modules are considered as modules of interest and are analyzed later on in the pipeline. When DEGs are either (a) hub-genes in the modules of interest, or (b) Top Enriched DEGs (TED), which are DEGs annotated to the most enriched GO term in each of the three GO databases as well as the most enriched KEGG pathway in their respective datasets, they are considered candidates for potential biomarkers for ESCC. Two microarray datasets (GSE20347 and GSE23400) and one bulk RNA-Seq dataset (GSE130078) were used to validate our Integrative DEA framework. With a *p-value* of 0.05, Limma+Voom, edgeR, and DESeq2 were each able to extract 6,858, 12,623, and 12,766 DEGs for GSE130078. 9,225 DEGs were found by our consensus function with the additional parameter (q-value). SAM, EBAM, and Limma were successful in extracting 8,689, 10,642, and 9,565 DEGs for GSE20347 at $p = 0.05$. 8,318 were found by the consensus function with the *lFDR* option added. Similar DEGs were found in GSE23400, including 13,558 (Limma), 14,301 (SAM), and 15,748 (EBAM). These DEGs have been found to be highly GO enriched, including 2,418 (GSE130078), 5,860 (GSE20347), and 7,882 (GSE23400). One hundred and twenty four, 59, and 160 hub-genes were discovered from 7, 3, and 8 modules of interest in GSE20347, GSE23400, and GSE130078, respectively. 146, 77, and 176 are candidates for putative ESCC biomarkers when the 22, 18 and 16 TEDs discovered by GSE20347, GSE23400, and GSE130078, respectively, are taken into account. The biological relevance of each candidate to 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, was a very important factor we took into consideration when deciding whether a BCG should be a potential biomarker.

Our method identified four BCGs, including STAT1, HIF1A, DNMT3B, and MCM7, which are Transcription Factors (TFs), have significant biological significance to their respective datasets, and may serve as ESCC biomarkers. These BCGs were found using previous research works. Our method identified GSK3B, reported as a DEG by both microarray datasets (GSE20347 and GSE23400), as a TED because it has substantial biological relevance to both microarray datasets and significant literature support as a possible biomarker of ESCC. Similar to this, nine BCGs, including HOMER3, PSMD4, PSAT1, TFRC, MCL1, EPHA2, KPNA2, CKS2, and PRMT1, seven BCGs, including HLA-F, HLA-G, CXCL10, ISG15, PSMA3, FCGR2A, and C3AR1, and four BCGs, including CAV1, VEGFC, CDK4, and MSH2, have been identified as potential biomarkers for ESCC in the datasets GSE20347, GSE23400, and GSE130078, respectively. Additionally, we discovered that 3 TFs, PTMA, VGLL4, and NONO, 1 TF, PLEK2, and 2 TFs, HMGB3 and ESRRG, in the datasets GSE20347, GSE23400, and GSE130078, respectively, had moderate biological relevance but substantial literature support as possible ESCC biomarkers. Therefore, these TFs can be thought of as probable ESCC biomarkers but requires further in-depth analysis to further establish their relevance to ESCC. On the opposite end of the spectrum, despite their great biological significance to their separate datasets, the transcription factor AR, a TED that is recognized as a DEG in both microarray datasets, and PSMC2 have been identified as possible biomarkers for further SCC related to ESCC.

Next Chapter presents a centrality-based hub-gene centric method called Centrality Based Differential Co-Expression Method (CBDCEM), for crucial gene finding for critical diseases. The identification of hub-genes for each differentially co-expressed module is a key task of differential co-expression (DCE) analysis. We develop a consensus-based approach that identify hub-genes using seven centrality measures.