Chapter 6

Differential Co-expression Analysis On Single Cell RNA-Seq Data

6.1 Introduction

Biological systems are often the congregation of meticulously regulated tens of thousands of genes in complex, yet dynamic networks, that change substantially among different tissue types, cell states, or developmental stages. So the first step towards perceiving complex biological processes is by deciphering gene interactions and discovering changes in corresponding networks. The possibility of uncovering the biological and biochemical pathways relevant to disease progression and therapeutic targets can be achieved through the identification of abnormal gene interactions under varying conditions [367]. Network analysis on bulk tissue RNA Sequencing (RNA-Seq) data plays a pivotal role in the identification of genes responsible for similar biological functions, transcriptional regulation targets, and disease-associated pathway regulators. However, with the assumption that cells maintain the same regulatory mechanisms across diverse cell types, tissue level network analysis only explores the gene-gene interactions across multiple samples. Recent years have seen the rapid development of single cell RNA sequencing technology that facilitates construction and subsequent investigation of gene networks across cell types a reality. Network analysis on scRNA-Seq data provides valuable insight into the transcriptional regulation mechanisms underlying various biological processes. As in the case of all technologies, regardless of the fact that exploratory analyses have demonstrated the possibilities of constructing functional gene networks, technical as well as biological complications present significant challenges in scRNA-Seq data. In scRNA-seq, a truly expressed gene may not be detected in some cells due to technical inefficiencies resulting in false zero expressions. Furthermore, zero expression that represents biological variations can be a result of the stochastic gene expression process [367]. Thus, when compared to bulk RNA-Seq data, scRNA-Seq is often much

sparser and thus requires non-conventional computational and statistical tools that are apt in tackling challenges posed by the abundance of zero counts.

6.1.1 Single Cell RNA Sequencing (scRNA-Seq)

Single-cell RNA Sequencing (scRNA-Seq), a revolutionary technique in genomics, enables researchers to examine gene expression at the level of a single cell. scRNA-seq offers a high-resolution view of gene expression within individual cells as opposed to standard bulk RNA sequencing, which examines gene expression across a population of cells. In order to identify and characterize unusual cell types, cell sub-populations, and dynamic changes in gene expression throughout cellular development and response to stimuli, thousands to millions of cells can be analyzed simultaneously using scRNA-seq. Developmental biology, immunology, cancer, neuroscience, and personalized medicine are just a few of the domains where the approach has major implications.

Unlike bulk RNA-Seq, where gene expression is examined in a population of cells, scRNA-Seq analyzes expression profiles of individual cell within a heterogeneous population. Thus, scRNA-Seq analysis provides a nuanced and detailed view of the diversity and function in cells. scRNA-Seq analysis is capable of shedding light on the heterogeneity within cell populations by examining gene expression at single cell level and can lead to detection of sub-populations and uncommon cell types that are not detected by bulk RNA-Seq analysis. scRNA-Seq analysis helps in understanding the complex biological process by investigating co-expression patterns between cells so as to unravel the underlying interactions and associations between cell types within a tissue or organ. scRNA-Seq approaches may introduce technical noise and biases due to sample preparation, sequencing and amplification thus leading to lower accuracy and reproducibility. When compared to bulk RNA-Seq, scRNA-Seq has smaller read depth per cell, thus leading to higher dropout rates and is less sensitive to finding lowly expressed genes. scRNA-Seq can be computationally costly as it extensively investigates thousands of distinct cells. Analysis of scRNA-Seq data entails specialized processing power and knowledge leading to complications in interpretation. Consistent detection of cells with low read counts or cell types that are rare by scRNA-Seq is not possible due to technical limitations.

6.2 Related Works

For co-expression network (CEN) construction and analysis, Salehi et al [600] employ the widely used WGCNA [327]. This work adheres to the basic pipeline of CEN construction with scale-free topology. To quantify the correlation between the expression of each pair of genes and identify only positive correlations, the Pearson correlation coefficient [545] and the signed network options were employed, followed by the production of a topological overlap matrix (TOM) [574].

Li et al [367] proposed scLink, which calculates the correlation between gene pairs followed by the use of a penalized and data-adaptive likelihood method to learn sparse dependencies among genes and construct sparse gene CENs, to improve the construction of gene CENs for single cells. The incapacity of Pearson [545] and Spearman's [643] correlation coefficients to efficiently approach the representation and interpretation of gene-gene relationships in exceedingly sparse scRNA-Seq data was discovered in this study. Using two phases, scLink delivers reliable inference of gene co-expression networks while also capturing functional gene modules. 1) constructing a robust co-expression matrix from gene expression data in order to accurately reflect the co-expression interactions between genes, and 2) identifying a sparse gene network from the co-expression matrix using a penalized and data-adaptive likelihood approach. scLink is intended to detect and predict ligand-receptor interactions among different cell types in a tissue or biological sample.

Algabri et al. [18] proposed Single-cell Gene Expression Network Analysis (scGENA) [1], a systematic pipeline for network analysis of scRNA-Seq data. In scGENA, identification of DEGs through DEA is followed by creation of a CEN from the DEGs, DCA, d=functional enrichment analysis, and finally identification of overlapping genes across samples. scGENA investigates the changes in network topology across cell groups under varying conditions, cell types and stages.

Sekula et. al. [614] offer a hierarchical Bayesian factor model for constructing a gene CEN from scRNA-seq count data. The treatment-dependent parameters in the proposed model determine the activation latent factors in each gene. This permits gene-gene co-expression to be calculated within each treatment group. Although Sekula et al. [614] only consider two group settings labeled as control and treatment for simplicity,

[1] (https://github.com/zpliulab/scGENA)

the model can be extended to additional group scenarios. The proposed count model, which is conditionally Poisson but marginally overdispersed, allows for zero-inflation and high cell-to-cell variability peculiar to scRNA-Seq data.

Chiu et al. [100] developed scdNET [2], where differential gene regulation networks associated with cellular states are analyzed at single cell level. scdNET starts with pre-processing and normalization , as well as deletion of non-informative genes in either state with the aim to reduce inter-cell bias. Fisher transformation is used to reduce sample size related bias while elimination of zeroes is achieved through the computation of gene-gene correlation within each group of cells. Within groups of cells normalized correlation co-efficients are compared so as to assess the changes in the correlation in the Fisher domain. Integration of the significant changes in the gene-gene pairings into the differential network is the final step in scdNet.

While scGENA [18] builds a network from DEGs, Sekula et al. [614] uses a hierarchical bayesian model for network construction, and scdNET [100] follows the pipeline of calculating gene-gene correlation within cell groups and merging gene-gene pairs with significant changes across groups into differential network. There are a few studies that do network analysis [600, 176, 367] as well as DCA [18, 614, 100] on scRNA-Seq data.

To the best of our knowledge, there are no works of DCA on scRNA-Seq data that follow the pipeline of CEN construction, module extraction, identification of biologically relevant modules, detection of hub-genes, and finally identification of biomarkers. With the goal of detecting intrinsic gene-gene interactions at the cellular level, we established a framework for differential co-expression analysis suitable for scRNA-Seq data. We tested the hypothesis on ESCC. In light of the following considerations, the suggested framework for Differential Co-expression Analysis Method on single cell RNA Sequencing data, scDiffCoAM is significant.

- Compared to some of its counterparts, it uses a better hub-gene discovery method.
- It can identify certain crucial ESCC genes that have not been reported by others.
- Due to the evaluation of both statistical and biological factors as well as written evidence, its validation of possible biomarkers is full proof.

[2] https://github.com/ChenLabGCCRI/scdNet

6.3 Background

In this section, we discuss the measures that we use for deciding the significance of genes for hub gene finding and two R packages- Seurat [219] and hdWGCNA [510] that we use in the implementation of scDiffCoAM.

6.3.1 Measures for hub gene finding

Azuaje et al. [35] have observed that there is often an association between key disease pathways and highly connected genes (i.e., hub-genes) in gene CENs. We employ hdWGCNA to construct a CEN for high-dimensional data and to extract significant modules for a given dataset. With the extracted modules, important nodes (or genes) can be identified as potential biomarkers. In Table 6.1, we summarize seven measures that we employ for hub-gene finding.

Tab. 6.1: Centrality Measures for hub-gene finding employed in scDiffCoAM

Measure	Function	Formula
Alpha Centrality [53]	An adaptation of eigenvector centrality with the addition that nodes are imbued with importance from external sources.	Given a graph with adjacency matrix A_i the alpha centrality is defined as follows: $x = (I - \alpha A^T)^{-1} e$ where e_j is the external importance given to node j, and α is a parameter.
Average Distance[128]	Average distance of a node in a strongly connected and loop free graph. It is the inverse of closeness centrality.	Average distance of node u to the rest of nodes in the net defined as: $C_{radC_u} = \frac{\sum_{w \in V} dis(u,w)}{n-1}$
Barycenter Centrality[712]	Barycenter scores are calculated as 1 / (total distance from vertex v to all other vertices) in a strongly connected network. More central nodes in a connected component will have smaller overall shortest paths, and 'peripheral' nodes on the network will have larger overall shortest paths.	If $\sigma(v)$ denotes the sum of the distances from v to all other vertices then Barycenter Centrality for vertex v defined as: $C_{radC_v} = \frac{1}{\sigma(v)}$
Decay Centrality [277]	Decay centrality is a centrality measure based on the proximity between a chosen vertex and every other vertex weighted by the decay.	Decay centrality of a given vertex x of a graph G is defined as: $\sum_{y \in V(G)} \sigma^{d(x,y)}$ where $d(x,y)$ denotes the distance between x and y and $\sigma \in (0,1)$ is a parameter.

A network centrality metric called alpha centrality [53], a subset of eigenvector centrality, is used to evaluate the significance or influence of specific nodes within a network. Alpha centrality measures a node's relative significance in a graph based on its connections to other nodes in the network. Alpha centrality permits the introduction of a parameter (α) to modify the weight of a node's neighbors as opposed to standard eigenvector centrality, which determines centrality based on all connections with equal weight. This parameter offers flexibility in emphasizing or understating particular network edges based on personal preferences or subject-matter expertise. The average number of steps or edges needed to travel from one node to every other node in a network is measured by the average distance [128], which is a network statistic. It measures the typical path length between nodes and offers insights on the network's connectivity or signal transmission efficiency. A network with nodes that are typically closer to one another, enabling quick communication and interaction, has a smaller average distance. On the other side, a network that is more sparse or distant is implied by a longer average distance, which may necessitate more steps for information to spread between nodes. Barycenter centrality [712] determines a node's centrality based on the geometric centers (barycenters) of its nearby nodes. According to this metric, a node's centrality is determined by how close it is to the nodes next to it in terms of mass. The barycenter centrality takes the spatial distribution of nodes within the network into account; nodes with greater centrality scores are those that are closer to the centers of their neighbors. In networks that require the physical position of nodes, such as spatial networks, this centrality measure is very important. Higher centrality scores are given by decay centrality [277] to nodes that are both well-connected and close to other nodes. According to decay centrality, a node's influence on another node declines as their distance increases. Nodes that are closer to one another have a greater influence, whereas nodes further apart have a less impact. This measurement illustrates the notion that a node's influence on its neighbors decreases with increasing distance. Decay centrality is especially helpful in situations where determining the importance of nodes within the network depends on both the quality of connections and their closeness.

Tab. 6.2: Comparison of the four of the seven measures employed by scDiffCoAM. Comparison of the other three measures are done in Table 5.2.

Measure	Pros	Cons
Alpha Centrality [53]	• Alpha centrality allows for a more customized measurement of centrality by assigning different weights to nodes or biasing the diffusion process towards specific nodes. • The alpha value can be changed to determine the degree of personalization and impact imparted by nodes, allowing for a more fine-grained investigation of network dynamics. • It can assist in addressing biases or information bubbles that may exist in networks. • It supplements previous centrality measures and provides a more thorough insight of network structures and dynamics.	• By embracing personalization, alpha centrality incorporates subjectivity into the centrality assessment. • The correctness and relevance of the personalization element substantially influence the effectiveness of alpha centrality. • Calculating alpha centrality can be time-consuming, especially for large-scale networks. • Small changes in network topology or connection can have a large impact on centrality rankings in alpha centrality.
Average Distance [128]	• It provides a global perspective on the network's efficiency in terms of transmitting information or resources throughout the network by analyzing the average path length. • Comparing average distances between networks or over time can aid in assessing structural changes, identifying similarities and differences, and assessing the impact of network alterations or interventions. • It can help to guide the design or optimization of routing algorithms by identifying the shortest paths or bottlenecks in the network.	• The average distance measure provides a global perspective on network efficiency but may overlook local variances or unique pathways inside the network. • The average distance measure is susceptible to outliers or extreme values in the network. • The average distance measure considers all edges or links in the network to be equal and does not take into account the weights or strengths associated with the connections. • It does not take into consideration dynamic changes, temporal dynamics, or the network's evolution over time.
Barycenter Centrality [712]	• Barycenter centrality considers a node's proximity to other nodes in the network and is valuable in finding nodes with significant connections and influence inside the network. • It can assist in identifying nodes that serve as bridges or links between different communities. • It can be useful in determining a network's resilience or vulnerability to node or connection failures.	• Barycenter centrality may not correctly reflect the true centrality or effect of nodes in networks with irregular or skewed distributions. • It makes the assumption that shorter distances always equal greater centrality, which may not be true in networks with various paths or where specialised routes are significant. • It gives a proximity-based measure that does not take into account other factors that may influence node centrality.
Decay Centrality [277]	• Decay centrality recognizes that the influence or interaction between nodes decreases as their distance rises. • It enables the decay function or weight assignment to be customized depending on specific preferences or domain expertise. • In networks where long-distance connections are becoming less important, decay centrality can help to offset any bias or overemphasis on nodes connected by shorter channels. • High decay centrality nodes are more likely to have strong connections and influence over neighboring nodes, indicating their ability to control or affect information flow within a particular radius.	• The decay function chosen can have a major impact on the centrality results. • It does not have a broadly accepted definition or technique. • It focuses exclusively on the distance decay effect while ignoring other factors that may influence node importance or centrality. • Because of the complexity of the decay function and the variable weights provided to edges or pathways, interpreting decay centrality data can be difficult.

6.3.2 Seurat

Seurat [219] [3] is an open source R (Section 2.2.1) package developed by Satija et al.[608] that provides for well structured organization of scRNA-seq data in the form of objects called the Seurat Objects along with an efficient set of methods for their processing. It also provides for maintaining the gene and cell related data and meta data existing in the input data sets as well as the derived meta data obtained from preprocessing for later use, in the respective Seurat objects. The well structured organization, provision for derived meta data, and ready availability of methods frequently used in computational biology make Seurat a very very useful tool for researchers in the field.

6.3.3 High Dimensional WGCNA (hdWGCNA)

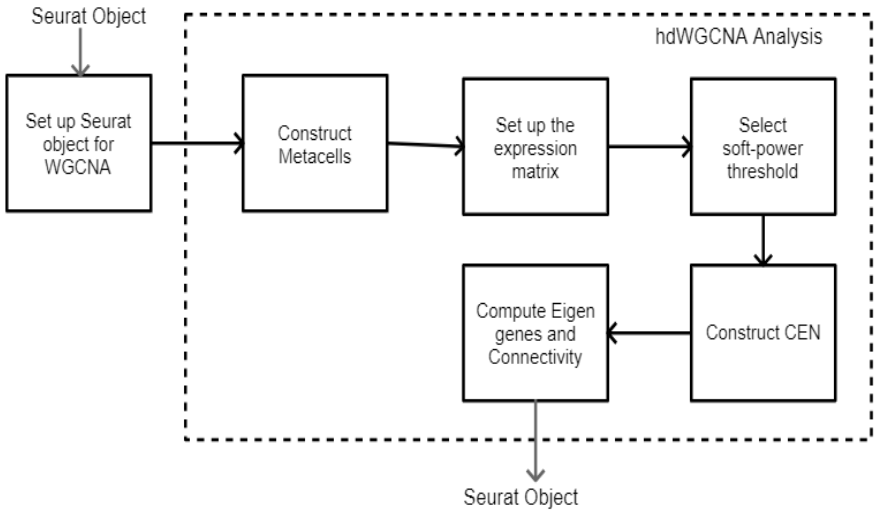


Fig. 6.1: Steps involved in WGCNA analysis for high dimensional data using hdWGCNA

Morabito et al. [510] recognize the immense complexity of biological systems with multi-scale hierarchies of functional units based on tightly-regulated interactions among organs, organisms, molecules, and cells. Further, they state that regardless of the existence of experimental methods that enable transcriptome-wide measurements across millions of cells, most omnipresent bioinformatic tools do not support systems-level analysis. Thus, Morabito et al.[510] present High Dimensional WGCNA (hdWGCNA), a comprehensive framework for analyzing co-expression networks in high-dimensional transcriptomics data such as single-cell and spatial RNA-seq. hdWGCNA provides built-in functions for a) Network inference, b) Gene module identification c) Functional gene enrichment analysis, d) Statistical tests for network reproducibility, and e) Data vi-

[3] https://satijalab.org/seurat/

sualization. hdWGCNA is further capable of performing isoform-level network analysis using long-read single-cell data.

The following are the most advantageous characteristics of hdWGCNA:

- hdWGCNA does not require prior knowledge or databases, making it a purely unsupervised approach.
- The co-expression information computed by hdWGCNA can be easily retrieved from the Seurat object to facilitate custom downstream analyses beyond the hdWGCNA package.
- hdWGCNA allows for comparisons between experimental groups via differential module eigengene testing and module preservation analysis.
- The CENs inferred by hdWGCNA are highly reproducible in unseen datasets, indicating that this is a robust methodology.

hdWGCNA is available as an R package[4] for performing weighted gene co-expression network analysis (WGCNA) [327] in high dimensional transcriptomics data such as scRNA-seq or spatial transcriptomics. hdWGCNA requires data formatted as Seurat [608] objects. Fig. 6.1 describes, in brief, the steps involved in WGCNA analysis for high dimensional data using hdWGCNA [510, 509]. Firstly, before running hdWGCNA the Seurat objects are set up for the operation. Setting up Seurat objects is then followed by the first step of running the hdWGCNA pipeline which is the construction of metacells from the scRNA-Seq dataset. In a nutshell, metacells are an aggregation of small groups of similar cells from the same biological sample of origin. Identification of these groups of similar cells is achieved through the k-Nearest Neighbors (KNN) algorithm followed by the computation of summed expression of these cells which finally results in a metacell gene expression matrix. Next step is the specification of the expression matrix that will be used for further network analysis. As in the case of WGCNA[327], this step is very important for hdWGCNA. hdWGCNA infers the co-expression relationship among genes through the construction of a gene-gene correlation adjacency matrix. Removal of weak connections while retaining the strong connections entails the reduction of the amount of noise in the matrix by raising the correlations to a power. This enhances the critical choice of soft power threshold. It is essential that the network has a scale-free topology. Construction of CEN starts off by optionally filtering genes and samples with too many missing entries or zero variance in at least one set. Module detection leaves

[4] https://smorabit.github.io/hdWGCNA/index.html

out the filtered genes. Genes are pre-clustered into blocks and for each block of genes, hdWGCNA constructs the network and topological overlap matrix (TOM) [574]. Using average linkage hierarchical clustering genes are clusters with the aim to identify modules. Processing of each block is then followed by checking, reassigning, and merging of modules based on kMEs (i.e., correlation with module eigengene). Computation of module connectivity involves the calculation of pairwise correlations between genes and module eigengenes.

6.4 ScDiffCoAM: A Complete Framework To Identify Potential ESCC Biomarkers Using ScRNA-Seq Data Analysis

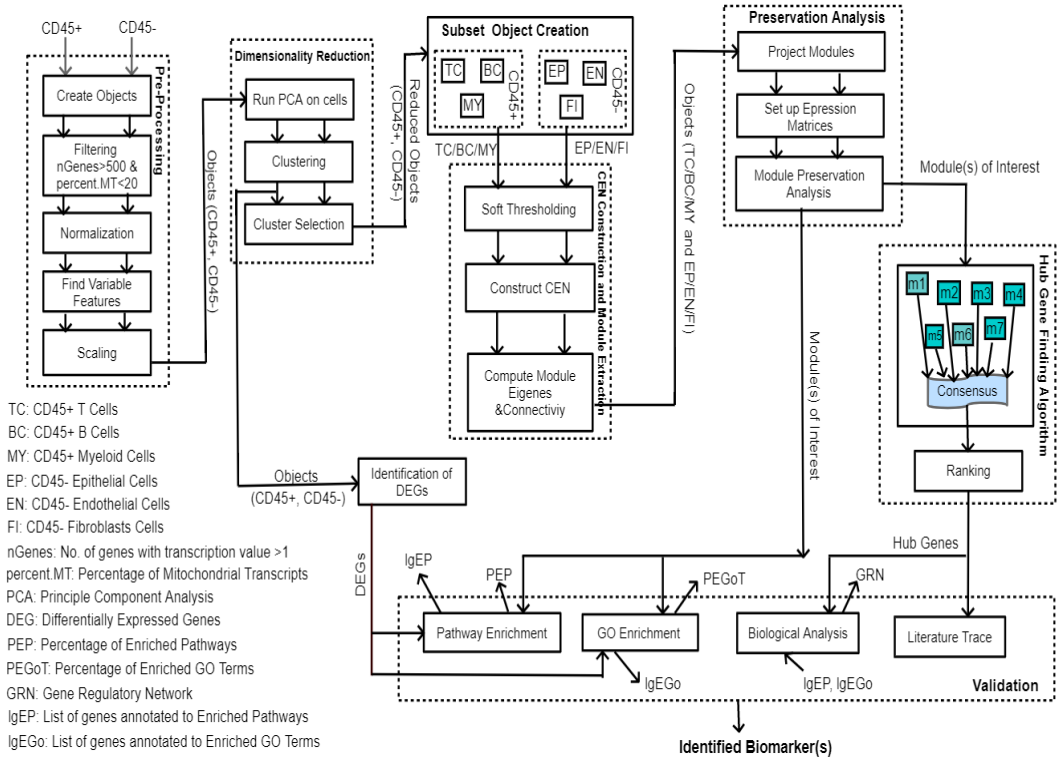


Fig. 6.2: Proposed framework for DCA on scRNA-Seq Dataset, scDiffCoAM

Our scDiffCoAM, scDiffCoAM, closely follows the conventional DCA pipeline with the aim to identify biomarkers that includes the following: 1) CEN construction 2) module extraction 3) identification of modules of interest (MoIs) 4) hub-gene detection, and 5) identification of biomarker(s). In Fig. 6.2 we depict the proposed framework, scDiffCoAM, for DCA on scRNA-Seq data. The datasets for DCA on microarray or bulk RNA-Seq are separated into two categories based on conditions, normal and disease. We observed transcriptional changes in gene-gene interactions under normal and disease conditions using DCA on such datasets. DCA on scRNA, on the other hand, may

enable an improved comprehension of the interaction of intrinsic cellular processes under two different conditions. We construct CENs for each cell type under each condition because our focus is primarily on interactions at the cellular level.

6.4.1 Pre-processing

The initial input to our framework, scDiffCoAM, consists of two datasets that represent the two different conditions considered for DCA. We have immune(CD45+) and non-immune(CD45-) conditions for the scRNA-Seq dataset GSE160269, as discussed in Section 2.6.3 and Table 2.2. ScRNA-Seq data are incredibly large when compared to bulk RNA-Seq and microarray data. Because of this, downstream analysis is very computationally-intensive. The basic pipeline we follow to pre-process scRNA-Seq data is discussed in detail in Section 2.7.3. It is preferred to create objects or data structures that streamline data administration and other related analyses. The following information is contained in these objects designed to facilitate downstream analysis: (a) the original count data, and (b) the data used for quality filtering, pre-processing, and other testing including meta-information such as gene counts for each sample, mitochondrial RNA content, etc. As a result, the pre-processing module first takes two datasets as input and generates two related objects, d_1 and d_2, for condition 1 (control) and condition 2 (disease), respectively. These objects are filtered, normalized, and scaled; the pipeline steps for each of these operations are covered in Section 2.7.3 in more detail. By applying quality filtering, we eliminate genes whose expression was only found in 0.1% of the cells as well as cells with a low gene content or a high mitochondrial content. In the pre-processing unit, the count data is normalized and scaled with the intention of facilitating additional analysis, such as WGCNA and other statistical tests.

6.4.2 Dimensionality Reduction

We identify the variable features of each object, facilitating principal component analysis (PCA) [546, 292], an effective statistical method for reducing the dimensionality of a sizable dataset. It is possible to identify features (samples) that are outliers on a 'mean variability plot' as variable features, which makes PCA and dimensionality reduction in the subsequent steps of the framework easier. We start by identifying variable features for each object (d_1, d_2) before using PCA (Section 2.1.6). Employing PCA with previously identified variable features as input thus results in the identification of

the principal components (PCs). On each object, clustering is used, with the PCs serving as the initial cluster pivots. The process of choosing the largest clusters yields the dimensionally reduced objects d_1' and d_2'.

6.4.3 Partitioning into cell types

We divide/partition the two reduced objects, d_1' and d_2', which we now refer to as condition-objects, into their corresponding cell-type objects to enable the cellular-level analysis of the gene-gene interactions. This results in the creation of multiple subgroups of objects for each condition-object that correspond to different cell types that constitute that condition-object. There are m different cell types, for instance, for the condition-type object d_1' (for condition 1). In order to correspond to m cell types of object d_1', we create m cell-type objects, $a_1, a_2, \ldots a_m$. Similar to this, we generate n cell-type objects, $b_1, b_2, \ldots b_n$, for condition-type object d_2'.

6.4.4 CEN Construction and Module Extraction

The creation of Co-expressed networks (CENs) corresponding to each condition is a critical step in DCA. The conditions are further divided into cell types in scRNA-Seq. As a result, it is imminent to perform subsequent DCA using CENs which correspond to each cell type. Thus, m and n CENs are constructed for each cell-type object that corresponds to the condition-type objects d_1' and d_2', respectively. The choice of power, referred to as the soft threshold, is required for CEN construction. For the purpose of calculating an adjacency matrix and corresponding Topological Overlap Matrix (TOM), co-expression similarity is raised to this power. The selection of soft thresholding power is based on the approximate scale-free topology criterion. It is unavoidable that CENs are not constructed from the cell-type object expression matrix as a whole for the construction of CENs because there are too many missing or zero entries. The CENs are instead constructed by identifying them as modules. Therefore, for every cell-type object such as $a_1, a_2, \ldots, a_m$ or $b_1, b_2, \ldots, b_n$, blocks of genes constitute the modules, and the CEN for that cell-type object consists of all such modules.

To identify modules of interest (MoI), we employ module preservation analysis (Section 2.1.9) We define an 'module of interest' (MoI), as a module that is not highly preserved because the majority of its connections are not retained [329]. We perform preservation analysis on each pair (a_j, b_j) where a_i correspond to m cell-type objects in

condition-type object d_1', and b_j correspond to n cell-type objects in condition-type object d_2' in order to detect MOIs. For example in a pair (a_1, b_1) , if a module x in cell-type object a_1 is not highly preserved (Section 2.1.9) in majority of modules in b_j then x is an MoI. For each pair (a_1, b_2) we perform preservation analysis such that we analyze which modules in a_1 does not retain most of its connections (not highly preserved as disscussed in Section 2.1.9)in b_1 and also analyze which modules in b_1 are not highly preserved in a_1.

6.4.5 Hub-gene Finding

Hub genes, which are thought to be significant in gene-gene networks because of their high interconnectedness with a large number of neighbouring nodes, can initially be considered potential biomarkers. We utilize a hub-gene finding algorithm [592] (Algorithm 1) variation that was developed employing centrality measures. The CENs are constructed by identifying sizable groups of genes as modules. For scRNA-Seq data, all nodes (genes) have the same degree, unlike for microarray data or bulk RNA-Seq data. As a result, the CBDCEM's [592] degree [171], betweenness [170], pageRank [652], and katz [302] centrality measures are proven to be useless in these situations. As they were proven to be more effective in our networks than in CBDCEM [592], we experimented with alpha centrality [53], average distance [128], barycenter centrality [712], and decay centrality [277].

In essence, we compute each chosen measure, namely alpha centrality [53], average distance [128], barycenter centrality [712], closeness centrality [39], decay centrality [277], eigenvector centrality [519], and radiality [766], on all genes present in the module for each MoI. The genes are then sorted according to the calculated value after that. It is significant that the measure determines whether the sorting is ascending or descending. Each measure's $top;k$ genes are given the value 1, while the rest are given the value 0. A gene is regarded as a hub gene in its associated MoI if it ranks in the $top;k$ of at least 4 out of 7 (majority) measurements. Here, $top;k$ is determined as follows with the goal of finding K hub-genes:

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

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

6.4.6 Identification of DEGs

In order to establish the biological significance of the critical genes identified by the hub-gene finding unit of the framework, lists of genes annotated to enriched GO keywords (lgEGo) and lists of genes annotated to enriched pathways (lgEP) are essential. We identify a set of DEGs for each non-reduced condition-type object d_1 and d_2 . The validation unit takes as input every gene that is identified as a DEG in d_1 or d_2.

6.4.7 Validation

The validation unit of the framework validates both modules in general and hub-genes in specific. A module is GO enriched and pathway enriched if at least one enriched GO term and enriched pathway is present the module. Gene Ontology (GO) enrichment analysis (Section 2.4.1.1) and pathway enrichment analysis (Section 2.4.1.2) are used to validate MoIs identified by the preservation analysis unit. All detected MoIs are used as input in the validation unit's GO enrichment and pathway enrichment analysis sub-unit of the framework. These subunits calculate the percentage of enriched GO terms (PEGoT) for each MoI across all three GO databases (BP: Biological Process, CC: Cellular Component, and MF: Molecular Function) as well as the percentage of enriched pathways (PEP) in KEGG.

We further validate each hub-gene detected by the framework so as to establish them as potential biomarkers. We assess the acceptability of hub-genes as potential biomarkers based on the following criterion.

a We examine the pertinent literature related to that disease with respect to the genes identified as crucial to support the claim in order to support the direct or indirect relationship of the identified hub-genes as potential biomarkers with the disease of interest.

b We perform pathway enrichment and GO enrichment analysis to determine the biological relevance of the identified hub-genes to the dataset and to comprehend how they interact with one another within a network.

c We identify the transcription factors (TFs) among the list of identified hub-genes and employ gene regulatory networks (GRN) to analyze their regulatory behavior in order to examine the association patterns between the target genes (TG) and the cor-

responding transcription factors (TF) as well as regulatory behavior among the list of identified hub-genes.

We initially find lgEGo and lgEP with $p - value = 0.05$ for the validation of the hub-genes identified by scDiffCoAM. The GO enrichment and pathway enrichment sub-units utilize the DEGs discovered by the framework's identification of DEGs unit as input. Two lists—lgEP, and lgEGo—are the outcomes. The hub-gene list, lgEGo, and lgEP are provided to the biological analysis unit in order to validate the hub-genes identified by the hub-gene finding unit of the framework. The biological analysis unit finds hub-genes that are annotated to enriched pathways and enriched GO terms. In other words, the hub-genes that are present in lgEGo and lgEP are identified by the biological analysis unit. In order to determine how these hub-genes behave in terms of regulation within the network, this unit further identifies hub-genes that are TFs and constructs GRN (Section 2.4.2). The validation unit of the framework's literature trace sub-unit finds hub-genes that have published literature traces confirming their status as biomarkers for ESCC or other SCCs closely associated with ESCC.

6.5 Experimental Results

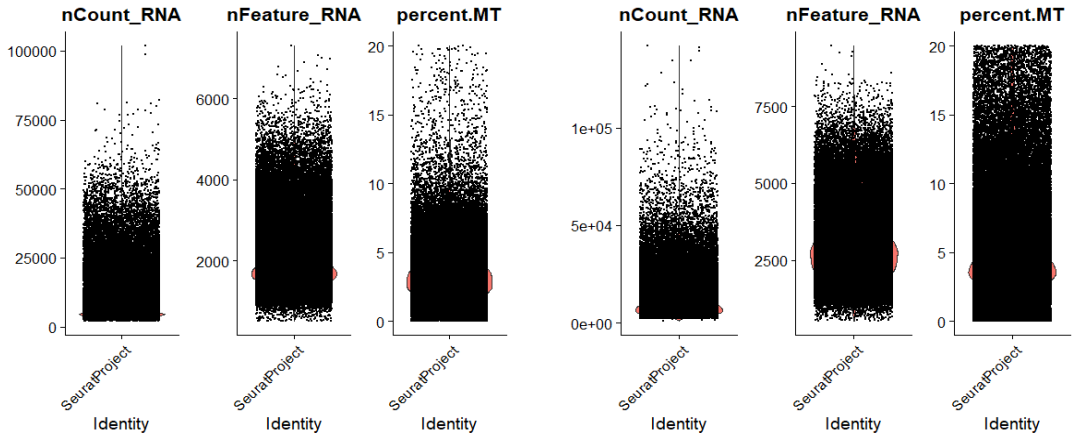
Our main area of interest is ESCC, a single cell RNA-Seq dataset, GSE160269 was used to validate our proposed framework scDiffCoAM. The detailed specifications for the dataset are provided in Table 2.2 and Section 2.6.3. Zhang et al.[877] analyzed 208,659 single cell transcriptomes in ESCC and obtained samples from four adjacent normal tissue and sixty ESCC tumors and samples from 60 individuals. The immune (CD45+) or non-immune (CD45-) cells were obtained through the CD45-FITC staining of single cell suspension. CD45+ immune cells has 3 cell types namely, Tcells (TC), Bcells (BC), and Myeloid (MY) while CD 45- non-immune cells have 5 cell types namely, Epithelial, Endothelial, Fibroblasts (FI), Pericytes (PE), and Fibroblastic Reticular Cells (FRC). DELL workstation running Windows 10 Pro for workstations with a 3.70GHz Intel(R) Xeon(R) W-2145 CPU and 64 GB of RAM serves as the test platform.In the R programming environment (Section 2.2.1), we carry out the experiments.

6.5.1 Pre-processing

As mentioned in Section 6.4.1, the input to the presprocessing unit are two datasets CD45+ (immune) and non-immune(CD45-) datasets (Table 2.2 and Section 2.6.3). We

employ the Seurat package [608] (version 4.1.1) [5] to construct two Seurat objects, CD45+ and CD45-, for quality filtering as well as subsequent downstream analysis. We intend to simplify the maintenance of the original count data and computation of meta-information, such as gene counts for each sample, mitochondrial content, etc. useful in quality filtering, by using Seurat to create the two objects. Furthermore, all information pertaining to the implementation of each test on a Seurat object can be stored in the same object and easily accessed thus aiding in the implementation of other statistical tests and analyses.

CD45+ and CD45- condition-type Seurat objects are of sizes $15,175 \times 1,11,028$ and $17,012 \times 97,631$, respectively. According to Zhang et al. [877], we achieve quality filtering by eliminating genes whose expressions were found in less than 0.1 percent of all cells and eliminating cells with gene counts below 500 or mitochondrial RNA contents above 20% (Figures 6.3a and 6.3b). In this step, the number of CD45+ Seurat cells (columns) is reduced by one, resulting in a dataset that is $15,175 \times 1,11,027$, while the number of CD45- cells is left unchanged at $17,012 \times 97,631$. Based on average expression and dispersion level thresholds, genes with highly variable expression, or to put it another way, outlier genes in a 'mean variable plot' were chosen [877]. Data scaling is accomplished by regressing normalized expression levels against the sum of UMI counts and the amount of mitochondrial RNA present in each cell for each gene using a linear model.



(a) CD45+ *(b)* CD45-

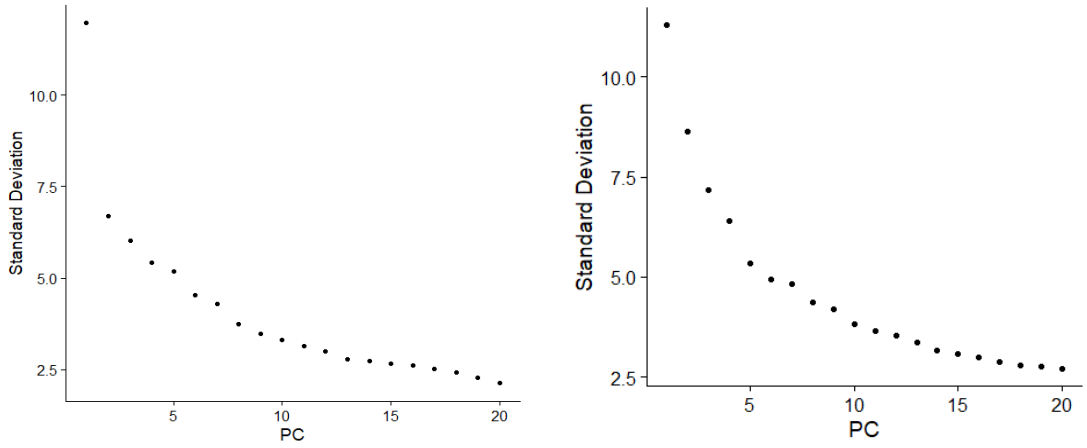
Fig. 6.3: Violin Plots for (a) CD45+ and (b) CD45-. Here, nCoun_RNA= no. of UMIs per cell, nFeature_RNA=no. of genes detected per cell and percent.MT= Percentage of mitochondrial RNA content.

[5] https://satijalab.org/seurat/

6.5.2 Dimensionality Reduction

We initially perform PCA (Section 2.1.6) on CD45+ and CD45- Seurat objects in order to their reduce dimensions. In order to identify PCs, PCA incorporates the variable features obtained in the pre-processing unit. The PCA elbow plots for CD45+ and CD45- are shown in Fig. 6.4a and Fig.6.4b, respectively. By assessing the minimum value of the following, we determined where PCA begins to elbow [6].

1. the point at which PCs cumulatively contribute 90% of the standard deviation (sd) but only contribute 5% of sd individually.
2. the point at which the fluctuation in percentage between two consecutive PCs is less than 0.1%.



(a) CD45+ *(b)* CD45-

Fig. 6.4: Elbow Plots for (a) CD45+ and (b) CD45-.

These values are respectively PC 45 and PC 20 for CD45+ and CD45-, and respectively PC 42 and PC 14. As a result, we established PC 20 and PC 14 as the minimal values for CD45+ and CD45-, respectively. We use Shared Nearest Neighbour (SNN) to perform graph-based Louvain clustering on 20 (CD45+) (Fig. 6.4a)and 14 (CD45-) (Fig. 6.4b) principal components (PCs), and then we find the clusters. As a result, 20 and 25 clusters are detected in CD45+ and CD45-, respectively. According to the cluster results, the size of the clusters significantly decreases in the seventh (7^{th}) and eighth (8^{th}) clusters for CD45+ and CD45-, respectively. As a result, we select a subset of the first six (CD45+) and first seven (CD45-) clusters in order to decrease the number of cells. This lowers CD45+ to $15,175 \times 74,588$ and CD45- to $17,012 \times 62,484$. We refer to these reduced condition-type Seurat objects as $CD45'+$ and $CD45'-$ Tcell reduces from $69,278$ to $53,694$, Bcell decreases from $22,477$ to $12,021$, and Myeloid decreases

[6] https://hbctraining.github.io/scRNA-seq/lessons/elbow_plot_metric.html

from 19,273 to 8,873. Epithelial decreases from 44,730 to 20,092, endothelial from 11,267 to 6,63, fibroblast from 37,213 to 35,803, pericytes from 3,102 to 8, and FRC from 1,319 to 218 (Table 2.2). With Bonferroni p-value correction [49], we identify cluster-specific markers to detect DEGs. We particularly used the MAST method [165] [7], which uses a hurdle model customized for scRNA-seq data to identify DEGs between two groups of cells. 5,321 and 7,292 genes, correspondingly, were identified as markers (DEGs) for CD45+ and CD45-.

6.5.3 Partitioning into cell types

As mentioned earlier in Section 6.4.3, we primarily focus on transcription changes in gene-gene interaction at cellular level. As such we subset each reduced Seurat object into their respective cell types. We create cell-type Seurat objects, Tcell (TC), Bcell (BC), and Myeloid (MY) Seurat objects from CD45+ condition-type Seurat object with 53,694, 12,021, and 8,873 cells, respectively. Similarly, Epithelial (EP), Endothelial (EN), and Fibroblast (FI) cell-type Seurat objects of sizes 20,092, 6,63, 35,803, respectively are created from conditon-type Seurat object CD45-. Here, it is noteworthy that we were unable to create subsets for Pericytes and FRC due to their smaller size as compared to the other cell types.

6.5.4 CEN Construction and Module Extraction

Following definitions are useful in understanding the subsequent discussion.

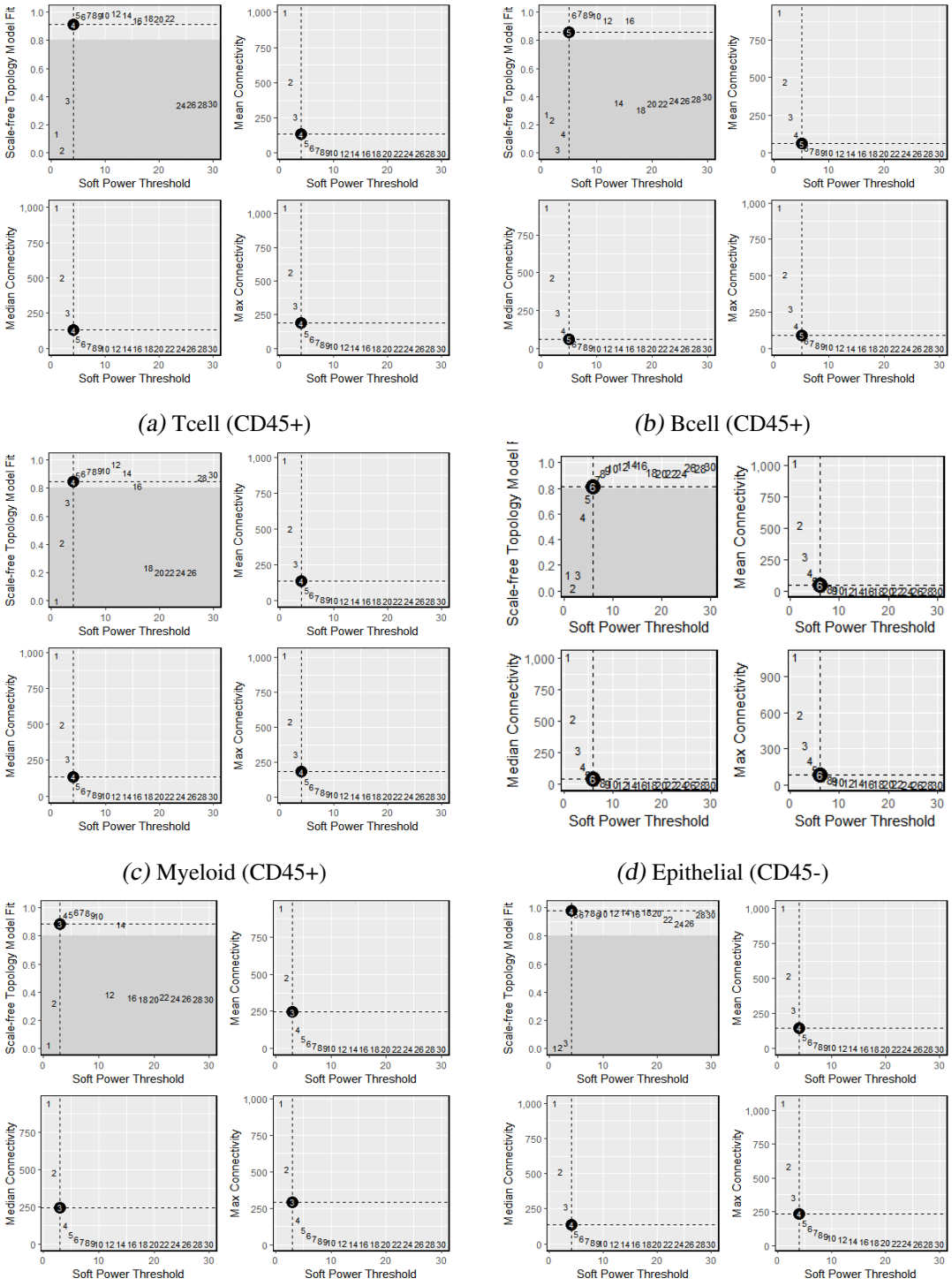
Definition 6.5.1 (CEN)**.** A CEN can be defined as a graph, $G(V,E)$, where V represents the set of genes in a Seurat object and E represents the set of associations among the genes in terms of their expression similarity.

Definition 6.5.2 (Module)**.** A module is a subset of genes, $M \subset G$ in a Seurat object, where there exists high coherence or homogeneity among the genes in terms of associations or expression similarities.

It is not feasible to implement WGCNA on sc-RNA-Seq because of its inherent limitations. The application of hdWGCNA makes it possible to generate CENs and conduct further analyses on highly dimensional data. Furthermore, the treatment of sc-RNA-Seq data as Seurat objects is plausible thanks to hdWCNA. In Section 6.3.3, we

[7] https://github.com/RGLab/MAST

discussed every step involved in CEN construction employing hdWGCNA. The initial stage in building a CEN is to put up all six Seurat objects for the WGCNA, then build metacells, and finally set up the expression matrices.



(a) Tcell (CD45+)

(b) Bcell (CD45+)

(c) Myeloid (CD45+)

(d) Epithelial (CD45-)

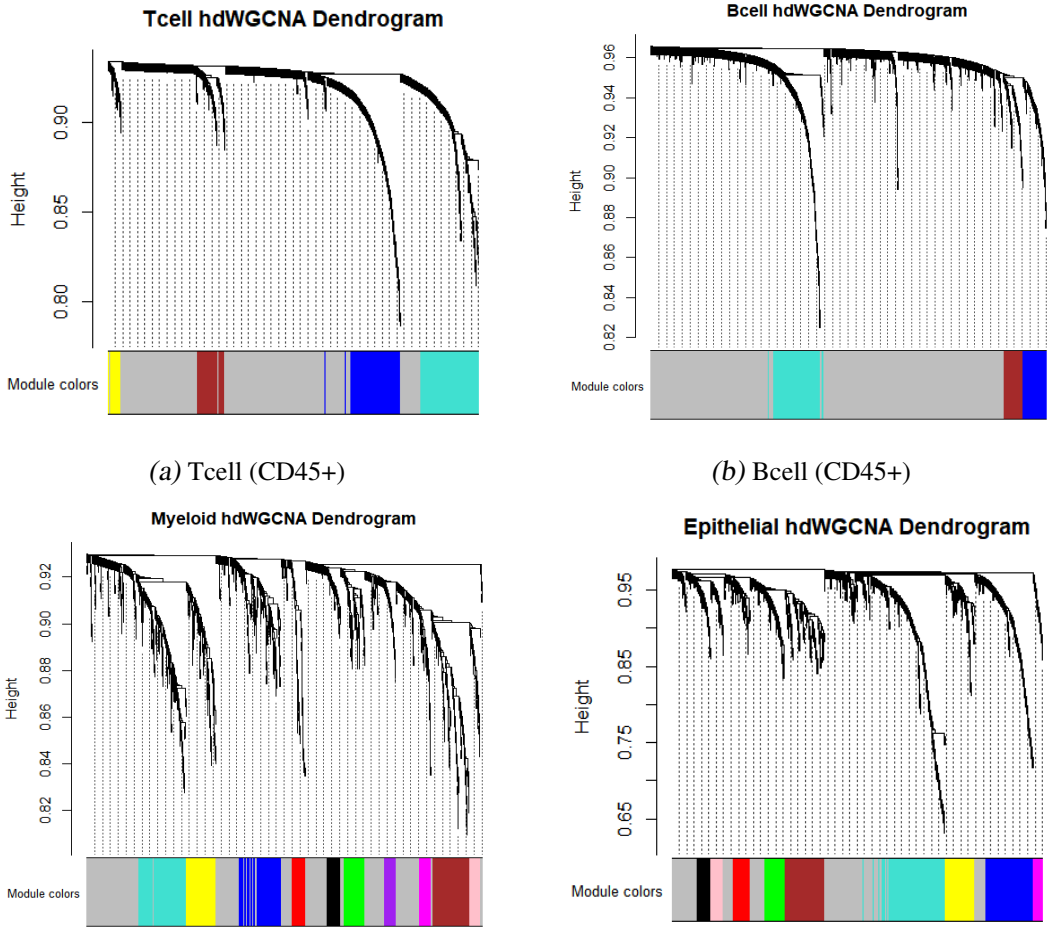
(e) Endothelial (CD45-)

(f) Fibroblast (CD45-)

Fig. 6.5: Soft Thresholds for CD45+ cell types a) Tcell and b) Bcell c) Myeloid are 4, 5 iand 4, respectively, and for CD45- cell type d) Epithelial, e)Endothelial, and f) Fibroblast are 6, 3, and 4, respectively.

Soft power thresholds are chosen prior to CEN construction. Soft power thresholds for

the CD45+ cell types Tcell, Bcell, and Myeloid are four (Fig. 6.5a), five (Fig. 6.5b), and four Fig. 6.5c), respectively. Soft power thresholds for CD45-cell types Epithelial, Endothelial, and Fibroblast are six (Fig. 6.5d), three (Fig. 6.5e), and four (Fig. 6.5f), respectively. We construct the CEN using multiple gene blocks as modules, which is covered in Section 6.3.3. Next, we compute the module eigens and connectivity, which leads to the merging of the modules. For CD45+ cell types, the Fig. 6.6a, Fig. 6.6b, and Fig. 6.6c, respectively, represent all modules found in Tcell, Bcell, and Myeloid. The dendrograms for Epithelial, Endothelial, and Fibroblast are shown in Fig. 6.6d, Fig. 6.7a, and Fig. 6.7b, respectively. The computation of module eigens and module connectivity for all six Seurat objects comes after CEN construction.



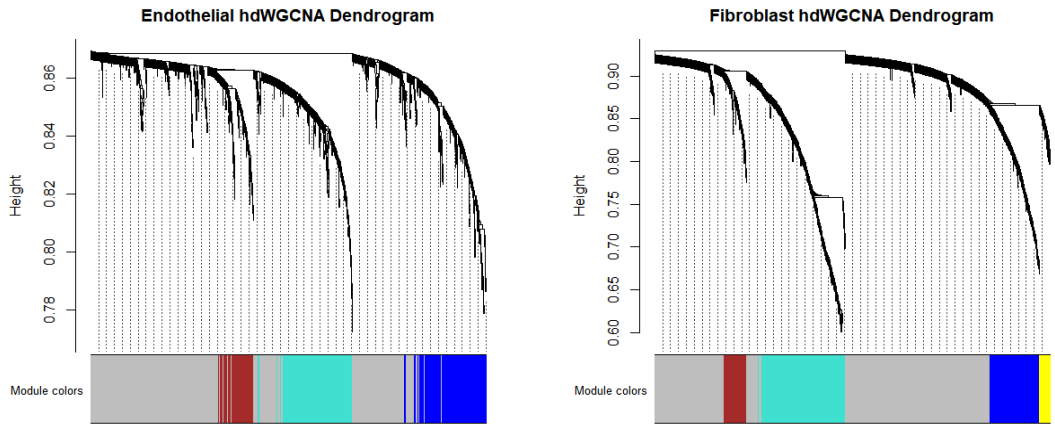
(a) Tcell (CD45+)

(b) Bcell (CD45+)

(c) Myeloid (CD45+)

(d) Epithelial (CD45-)

Fig. 6.6: Dendrograms for the CD45+ cell types a) Tcell, b) Bcell, and c) Myeloid with 4, 3, and 10 modules, respectively and the CD45- cell type d) Epithelial with 9 modules.



(a) Endothelial (CD45-)

(b) Fibroblast (CD45-)

Fig. 6.7: Dendrograms for the CD45- cell types a) Endothelial and b) Fibroblast with 3 and 4 modules, respectively.

6.5.5 Preservation Analysis

The concept of preservation analysis is discussed in detail in Section 2.1.9. In scDiffCoAM, the retainment of associations of modules from one Seurat object in another Seurat object is analyzed using module preservation analysis. For instance, let's say we wish to discover the modules found in the Tcell Seurat object that have the most connections retained in the Epithelial Seurat object. Here, we refer to the Seurat object that contains the modules for preservation analysis as the reference Seurat object and the Seurat object that analyses the preservation of the modules as the query Seurat object. In other words, selecting a Seurat object pair from CD45+ and CD45- is required for module preservation analysis. By alternating between each object in a pair as a reference and a query Seurat object, we analyze each object pair for preservation. The following are the steps for performing a module preservation analysis using the hdWGCNA programme [510, 509] [8].

1. Modules of the reference Seurat object are projected on the query Seurat object.
2. The expression matrices for both reference and query Seurat objects are constructed.
3. An adequate number of permutations are chosen in the module preservation analysis. (We've chosen 250 in this case).

The table 6.3 provides the $Z_{summary}$ statistics (Section 2.1.10) for the preservation analysis of all modules in instances of cell types as reference Seurat objects compared to corresponding query Seurat objects. All modules in each cell type (Seurat object) high-

[8] https://smorabit.github.io/hdWGCNA/index.html

lighted in the **bolded** and blue are MoIs and thus taken into account for subsequent downstream analysis.

Definition 6.5.3 (Module of Interest (MOI))**.** A module, i.e. a subset of genes is defined as 'module of interest', if (i) its $size \geq 100$, and (ii) it is not highly preserved or non-preserved ($Z_{summary} < 2$) [329] or moderately preserved ($2 \leq Z_{summary} \leq 10$) [329] in at least 2 out of 3 corresponding query Seurat objects.

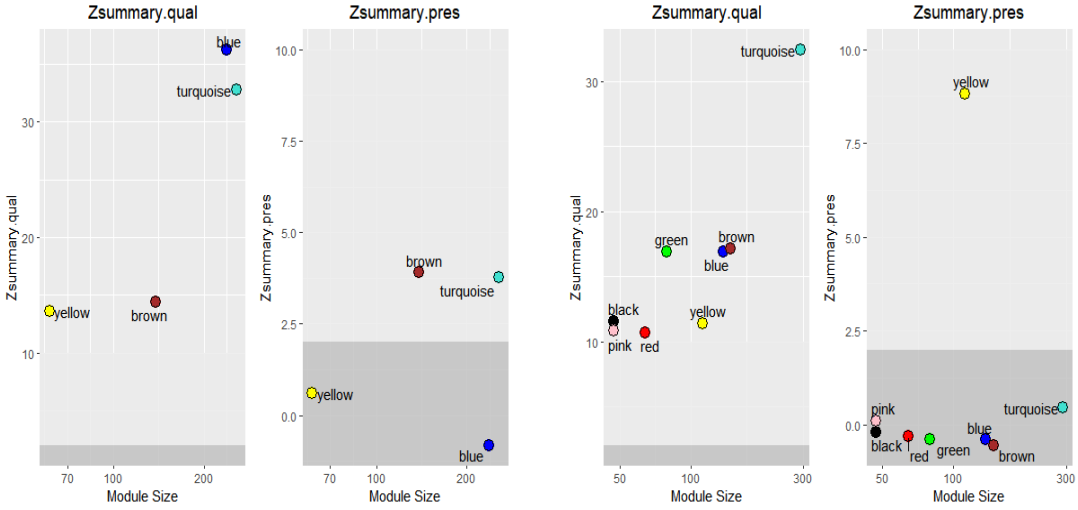
Despite having a size of 159 (genes) but a $Z_summary \leq 10$ (at least moderately preserved) in the Bcell query Seurat object only, the module *yellow* in Epithelial is not regarded as a MoI. However, module *green* in Epithelial is non-preserved ($Z_summary < 2$) [329] in all of the three corresponding query Seurat objects and is therefore disqualified as a MoI because it is $size = 79$ (i.e., $size < 100$).

Tab. 6.3: Preservation Analysis (Z_{summary}) of CD45+ modules in CD45- dataset and vice versa. Rows represent the Reference Seurat object while columns represent the query Seurat object. $Z_{summary}$ values ≥ 10 in modules of $Size \geq 100$ are highlighted in *italics* .Modules with $Size \geq 100$ and atleast moderately preserved (i.e, $Z_{\text{summary}} \leq 10$), highlighted in *italics*, in atleast two (out of three) corresponding test/query Seurat object are considered for subsequent downstream analysis and highlighted in blue and **bolded**. Here, TC: Tcell, BC: Bcell, MY: Myeloid, EP: Epithelial, EN: Endothelial and FI: Fibroblasts.

	Ref	Module	Size	EP	EN	FI
CD45+	TC	*yellow*	61	0.597	0.864	0.174
		brown	151	3.916	4.573	3.702
		blue	276	-0.830	0.402	0.170
		turquoise	300	3.767	5.365	1.452
	BC	*brown*	87	2.799	4.961	4.018
		blue	108	0.996	1.754	0.889
		turquoise	244	0.615	0.715	0.530
	MY	*magenta*	54	2.033	0.781	1.696
		pink	54	2.364	0.045	0.747
		purple	54	5.256	5.316	4.973
		black	65	28.490	6.287	6.560
		green	102	2.459	2.532	0.986
		yellow	145	7.270	8.406	7.575
		brown	178	1.807	1.802	3.210
		blue	189	3.150	3.760	2.856
		turquoise	228	4.245	4.791	5.099

	Ref	Module	Size	TC	BC	MY
CD45-	EP	*black*	47	-0.645	-0.161	-0.012
		pink	47	-0.450	-0.536	-0.271
		red	64	-0.659	-0.257	-0.045
		green	79	0.344	1.554	1.749
		yellow	159	10.038	7.540	10.951
		brown	212	-0.962	-0.152	1.143
		blue	253	-0.568	-0.127	3.063
		turquoise	336	1.923	7.553	15.771
	EN	***brown***	152	5.810	5.173	7.033
		blue	313	3.539	5.144	5.515
		turquoise	330	0.127	6.629	1.66
	FI	*yellow*	52	2.314	2.144	5.073
		brown	114	-0.099	-0.499	1.004
		blue	250	1.125	1.533	1.392
		turquoise	424	1.121	0.795	2.196

Fig. 6.8a, Fig. 6.8c, and Fig. 6.9a show the $Z_{summary}$ plots for Tcell in Epithelial, Endothelial, and Fibroblast, respectively. Similar preservation plots for Bcell modules in Epithelial, Endothelial, and Fibroblast are shown in Fig. 6.9c, Fig. 6.9e, and Fig. 6.10a, respectively, while plots for Myeloid modules in Epithelial, Endothelial, and Fibroblast are shown in Figures 6.10c, Fig. 6.10e, and Fig. 6.11a. On the other hand, Fig. 6.8b, Fig. 6.9d and Fig. 6.10d shows the $Z_{summary}$ statistics for Epithelial modules in Tcell, Bcell, and Myeloid, Fig. 6.8d, Fig. 6.9f and Fig. 6.10f shows the plots for Endothelial modules in Tcell, Bcell, and Myeloid and Fig. 6.9b, Fig. 6.10b and Fig. 6.11b shows the plots for Fibroblast modules in Tcell, Bcell, and Myeloid, respectively.



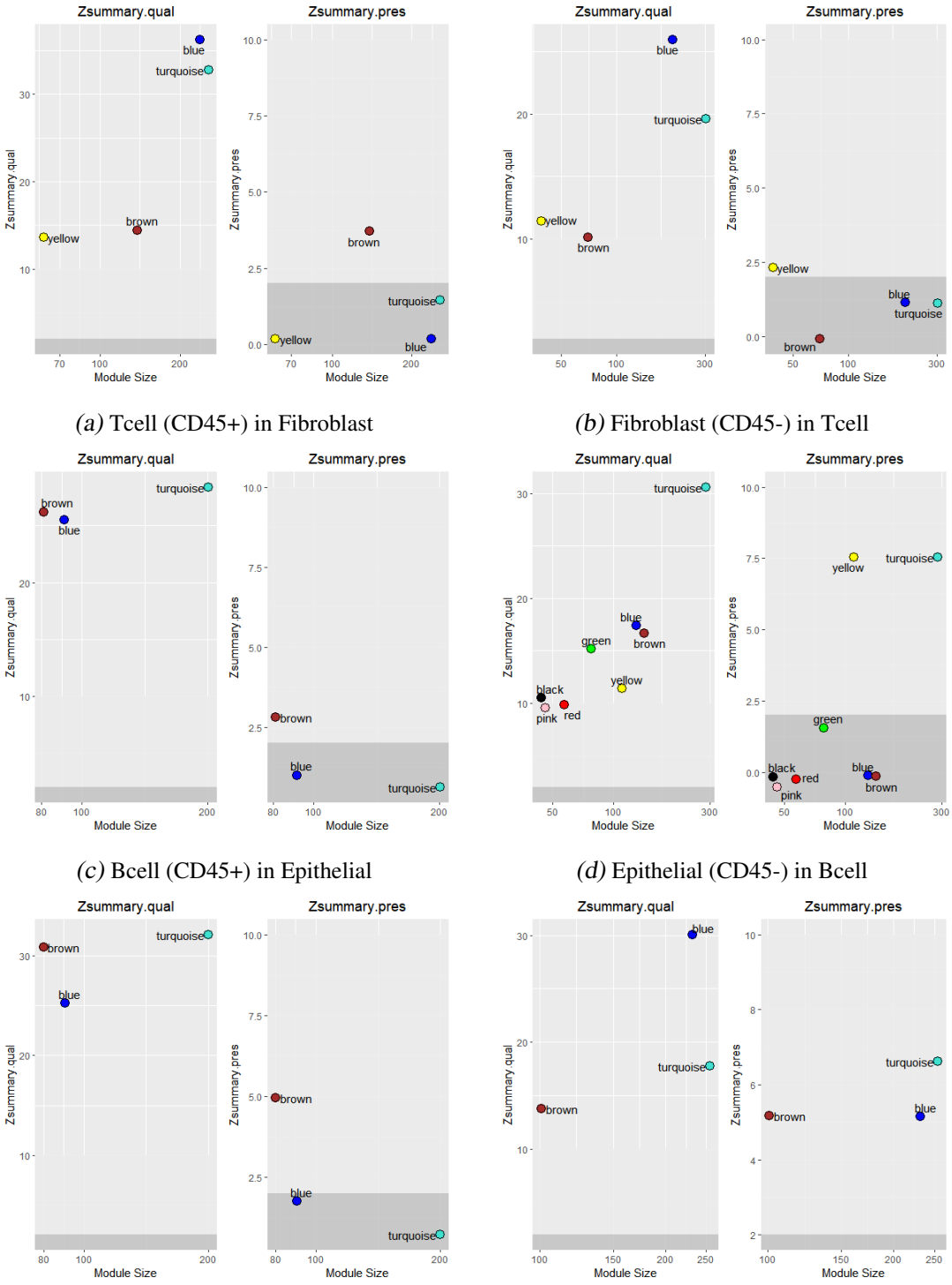
(a) Tcell (CD45+) in Epithelial

(b) Epithelial (CD45-) in Tcell

(c) Tcell (CD45+) in Endothelial

(d) Endothelial (CD45-) in Tcell

Fig. 6.8: $Z_{summary}$ plot for a) Tcell (CD45+) in Epithelial (CD45-) , b) Epithelial (CD45-) in Tcell (CD45+), c) Tcell (CD45+) in Endothelial (CD45-), and d) Endothelial (CD45-) in Tcell (CD45+)



(a) Tcell (CD45+) in Fibroblast

(b) Fibroblast (CD45-) in Tcell

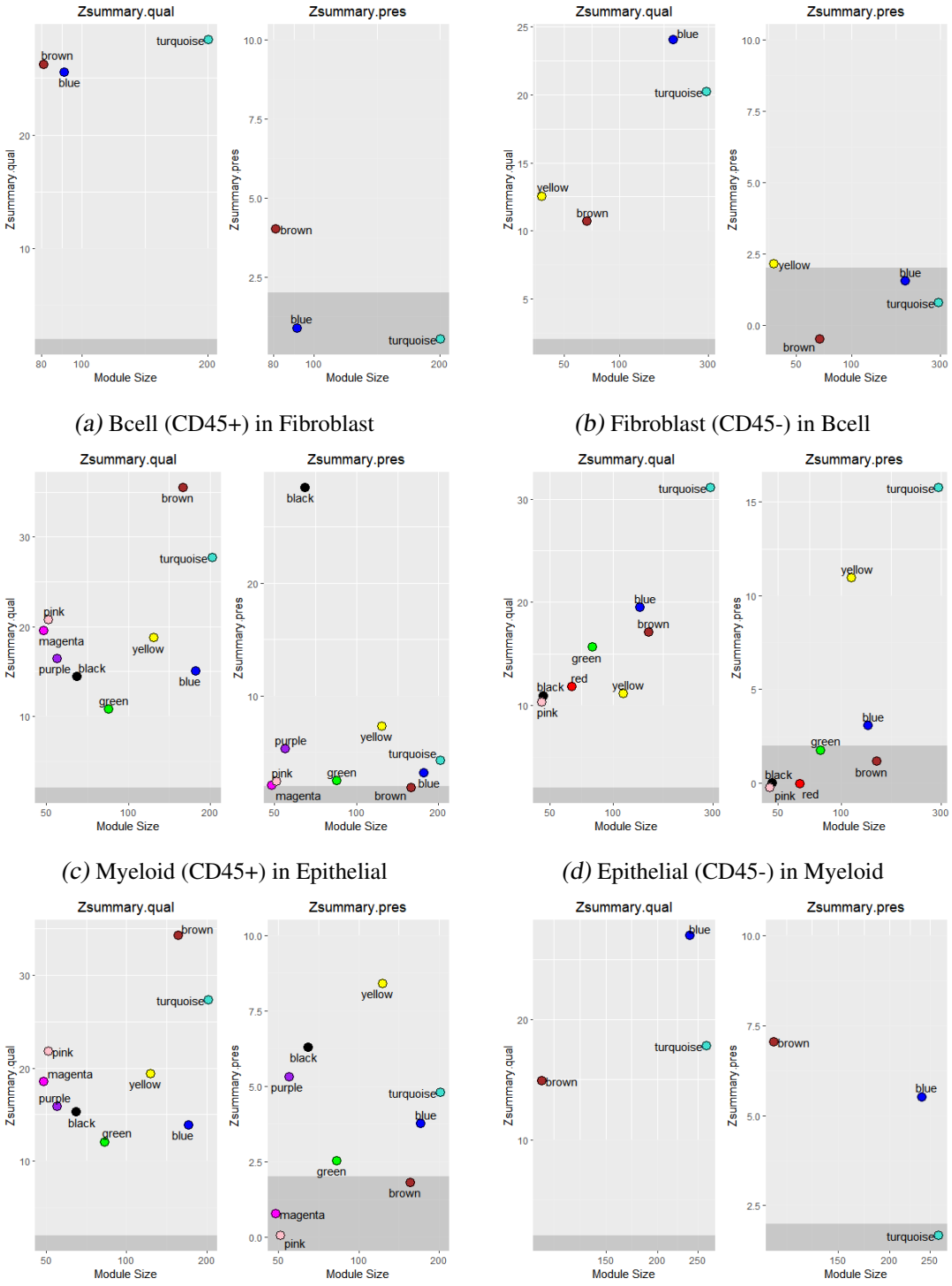
(c) Bcell (CD45+) in Epithelial

(d) Epithelial (CD45-) in Bcell

(e) Bcell (CD45+) in Endothelial

(f) Endothelial (CD45-) in Bcell

Fig. 6.9: $Z_{summary}$ plot for a) Tcell (CD45+) in Fibroblast (CD45-) , b) Fibroblast (CD45-) in Tcell (CD45+), c) Bcell (CD45+) in Epithelial (CD45-), and d) Epithelial (CD45-) in Bcell (CD45+), e) Bcell (CD45+) in Endothelial (CD45-), and f) Endothelial (CD45-) in Bcell (CD45+)



(a) Bcell (CD45+) in Fibroblast

(b) Fibroblast (CD45-) in Bcell

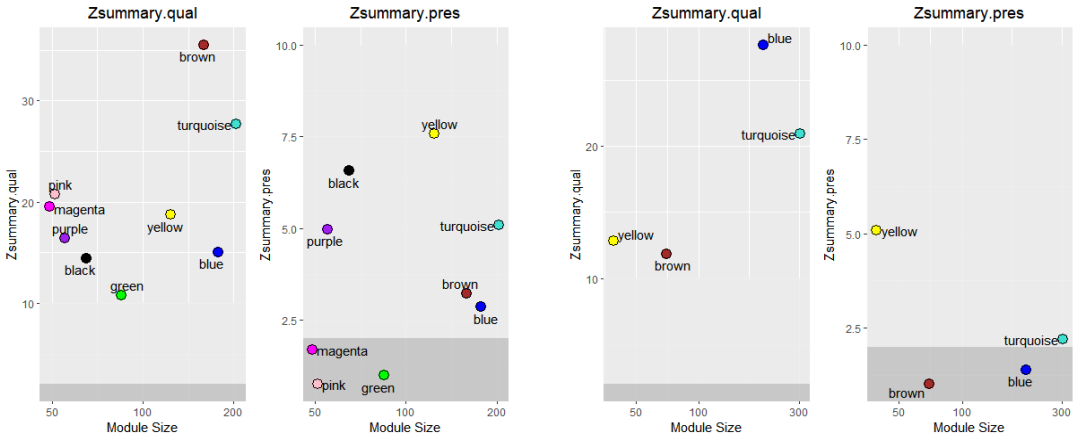
(c) Myeloid (CD45+) in Epithelial

(d) Epithelial (CD45-) in Myeloid

(e) Myeloid (CD45+) in Endothelial

(f) Endothelial (CD45-) in Myeloid

Fig. 6.10: $Z_{summary}$ plot for a) Bcell (CD45+) in Fibroblast (CD45-) , b) Fibroblast (CD45-) in Bcell (CD45+), c) Myeloid (CD45+) in Epithelial (CD45-), and d) Epithelial (CD45-) in Myeloid (CD45+), e) Myeloid (CD45+) in Endothelial (CD45-), and f) Endothelial (CD45-) in Myeloid (CD45+)



(a) Myeloid (CD45+) in Fibroblast *(b)* Fibroblast (CD45-) in Myeloid

Fig. 6.11: $Z_{summary}$ plot for a) Myeloid (CD45+) in Fibroblast (CD45-) and b) Fibroblast (CD45-) in Myeloid (CD45+).

6.5.6 Hub Gene Finding

As previously noted, we employed the hub-gene finding algorithm described in CBDCEM [592] in detail. But instead of using degree [171], betweenness [170], pageRank [652], and katz [302] centralities, we replace them with alpha [53], average distance [128], barycenter [712], and decay [277]. We found that the substituted measures are ineffectual through repeated trials in which we incorporate the original method suggested by CBDCEM [592] into scDiffCoAM. We have seen that the degree of the nodes has a significant impact on degree [171], betweenness [170], and katz centrality [302]. However, because hdWGCNA constructs networks on blocks of genes, the network modules that are discovered are highly connected, and every node (gene) in the module has the exact same degree. As a result, all genes have zero values for the centralities described before. To find the centrality measures that worked well for our research, we used the CINNA R package [31] [9]. An R package called CINNA [31] for network science centrality analysis is helpful for compiling, contrasting, assessing, and visualizing various centrality measurements. In the past, we designated three, two, and five modules in CD45+ cell types, Tcell, Bcell, and Myeloid as MoIs. In each CD45- cell type, epithelial, endothelial, and fibroblast, three modules are MoIs. With $K = 20$ and the goal of identifying 20 hub-genes in each module, these nineteen MoIs are taken as input into the hub-gene finding unit. There are approximately 380 hub-genes detected. It is worth noting that many hub-genes are identified in both CD45+ and CD45-cell types in across

[9] https://cran.r-project.org/web/packages/CINNA/vignettes/CINNA.html

multiple modules. Table 6.4 summarizes all the hub-genes identified using the CBD-CEM [592] hub-gene finding algorithm across all six cell types.

Tab. 6.4: Top 20 hub genes for each extracted MoI in CD45+ and CD45- datasets using our hub-gene finding algorithm. Hub genes with strong literature evidence of association to ESCC are marked in Red while hub genes with evidence of association with five other SCCs, HNSCC, LaSCC, LSCC, OSCC, and OSCC are marked in Blue.

	Cell Type	Module	Hub genes
CD45+	TC	*blue*	*ALDH1A2, ATF5, CCNB2, CDH1, , HBEGF, IFI30, IGHG1, IGKC, LAMP3, MERTK, MPP1, NEURL3, TCF4, TNFAIP6, ASAH1, ELF3, IGHM, PPT1, TRAV4, VASH2*
		brown	*DTL, ADM, ARL5B, BAMBI, DLX2, FEZ1, GLA, HEY1, HIST1H2AG, KCNQ1OT1, MNDA, PROK2, PSTPIP2, RAB3A, RP11-61J19.5, RRAD, TPD52, TRAM2, VASN, IL15*
		turquoise	*NLRP3, SLAMF8, APOC2, B3GNT7, CD68, COL3A1, COL6A3, FLT1, FUT7, ITGAX, KIAA0101, KRT17, LIPA, MCM7, MYL9, NCAPG, POGLUT1, RAD51AP1, SGPL1, SLC12A8*
	BC	*blue*	*BCAS4, BCAT1, CCR1, CD81, CDCA7, GATM, GPR137B, HCK, HIST3H2A, IGKC, IQCG, SINGLEC6, TRAM2, ZBED2, CYFIP1, IL5RA, PGD, PTAFR, RP11-731F5.1, ZNF296*
		turquoise	*ABCA1, CD1D, CFP, IFITM3, IGHJ4, IGLC3, MXD1, PLK2, ABCB9, ACP2, FAM64A, GCHFR, HK2, HLA-DQB2, IGHV2-70.1, IGHV3-43, IGLV3-1, KLHDC8B, NDUFAF6, TUBB3*
	MY	*blue*	*ADAMTS2, CAV1, CCR4, CD1B, COL6A3, DBN1, FAM3C, GATA3, IGLV2-14, PCSK1N, PTPN13, SH2D4A, TRBV12-3, A4GALT, ALDOC, B9D1, GLDN, MMP10, UCHL1, ZNF385A*
		brown	*TBXAS1, CTSZ, LINC00996, RARRES1, CUL9, ERLEC1, FCGR1B, GFRA2, HMOX1, HVCN1, MGLL, PGM2L1, PILRA, PTFAR, PTTG1IP, SDSL, SGPL1, SUCNR1, VMO1, ZC2HC1A*
		green	*LMO2, LY9, RP11-62414.2, TRDV2, ACY3, AFF3, FLI21408, GCSAM, GMDS, GPR18, ILF3-AS1, KIFC1, MID1IP1, POGLUT1, RP11-350N15.5, TNFRSF13C, TRAV16, TRAV9-2, YWHAH*
		turquoise	*ANGPTL4, BEST1, C11orf96, CEACAM3, DOCK4, EMILIN2, HOXA5, IGFBP7, JAML, OSM, P2RX1, PEA15, SLC8A1, BACH2, CCND2, CEMIP, GPR31, NEURL3, RHCG, SPRED2*
		yellow	*C4orf46, CCR2, CH25H, HCAR2, KIR2DL4, LGALS3, LINC00309, NRN1, NUB1, PLBD1, RASGEF1B, RNU12, RP11-386I14.4, RP11-467L13.7, RP11-598F7.3, RP11-796E2.4, SPON2, TIFA, PDCD1, PRF1*

Continued on next page

	Cell Type	Module	Hub genes
CD45-	*EP*	*blue*	*FGF5, GOS2, MMP9, TMEM45A, ANPEP, APCDD1, ARHGAP15, C10orf10, CH25H, CLDN11, CLEC3B, COL14A1, COL9A1, HAS2, IL24, LEPR, MSC, PAMR1,RGS5,SRGN*
		brown	*AKR1B10, SLC7A11, CAPN14, G6PD, RAB27B, SRXN1, TRIM7, DAPL1, ENTPD3, GPRC5D, GSN, GSTA4, KIAA1324, KRT18, MATN2, MCF2L2, MTSS1, RGS2, SAT1, TSPAN13*
		turquoise	*AMTN, AREG, C12orf75, CEACAMI9, COL4A1, CRABP2, CRIP1, CST6, EGFL7, IFI27, IFIT3, IL1B, KRT6A, LAMP3, MED24, NEFM, NEURL3, S100A1, SERPINE2, TGM2*
	EN	*blue*	*ABCC9, AOX1, EPHA2, EREG, GUCY1A2, KANK4, NDRG1, OMD, RCL1, SELE, SLIT3, SOX7, ZG16B, ANK2, C11orf96, C3, EGFR, PLA2G2A, PROX1, MSC*
		brown	*ABCA4, AVPI1, C2orf40, CBR1, COL14A1, CYP3A5, DSC2, FABP4, FXYD1, GJB6, HEYL, HHIP, IDO1, KRT19, LRRC17, NOTCH3, RELN, RP11-277P12.20, TNS4*
		turquoise	*ACHE, CCL5, CSTA, DSC3, FCER1G, GINS2, IQCG, ITGA3, LIPG, NFE2L3, PRSS3, SEMA3B, SLC6A8, ABCC5, ESM1, FXYD3, HMMR, MEOX1, NUSAP1, UBE2T*
	FI	*blue*	*EPYC, LXN, TSPAN13, CMPK2, DUSP2, ECT2, FADD, FAM84A, FMO1 GBP5, GGH, HILPDA, KRT14, KRT18, LAYN, MMP19, MMP7, NAA20, RGS2, SGK1*
		brown	*AARD, CLIC5, CST2, CXCL14, ENPP2, ETV5, GRAMD3, HGF, KERA, PLXDC1, PTGR1, TNN, WFDC1, ECEL1, LYPD5, MCM5, PRR15, RNF183, SCNN1D, SHANK2*
		turquoise	*CA12, CLIC3, FAM46A, PAX9, ABCC1, ATF5, CDC6, CDKN2B, CITED2, CLDN4, CMTM5, EGFL8, GDF15, HTRA3, INPP1, MCM3, NFKBID, PON3, PTGER4, PVT1*

6.5.7 Identification of DEGs

We identify DEGs from the cell clusters deected in non-reduced CD45+ and CD45- condition-type Seurat objects, as shown in Fig. 6.2. Prior to relevant cluster selection, the dimensionality reduction unit generates twenty and twenty-five clusters in CD45+ and CD45-, respectively. To find DEGs, we use the 'MAST' differential expression testing that is already built into Seurat [608]. With a $p-value \leq 0.05$, 13,955 genes have been identified to be DEGs in CD45+. It is notable that many genes are identified as DEGs in several clusters. As a result, CD45+ has been found to have 5,321 distinct genes in twenty clusters. Similarly CD45-, which contains twenty-five clusters, 24,175

genes are recognized as DEGs with $p-value \leq 0.05$, 7,292 of which are distinct DEGs. We perform GO enrichment (Section 2.4.1.1) and pathway enrichment (Section 2.4.1.2) analysis on these 5,321 (CD45+) and 7,292(CD45-) DEGs resulting in list of genes annotated to enriched GO terms (lgEGo in Fig. 6.2) and pathways (lgEP in Fig. 6.2) with $p-value \leq 0.05$ each for CD45+ and CD45-.

6.6 Validation

Multiple approaches are used to validate our results. First and foremost, we confirm that the MoIs found by our approach are biologically significant and substantially enriched. Through functional enrichment analysis (Section 2.4.1), we achieve this. Only highly enriched MoIs are evaluated for further research since they are biologically relevant. Every hub-gene of the biologically significant MoIs is regarded as a potential biomarker candidate gene (BCG) (Definition 6.6.1). To further confirm the biological significance of these BCGs, we make use of Regulatory Behaviour Network analysis (Section 2.4.2). Additionally, we trace the research that shows the BCGs to be potential biomarkers for ESCC and five other SCCs connected to ESCC. We identify the potential biomarkers by using our proposed biomarker criteria, which are covered in Section 2.5.

Definition 6.6.1 (BCG)**.** A gene g_i is defined as a Biomarker Candidate Gene (BCG) if it is identified as a hub-gene in a given MoI extracted by scDiffCoAM.

First, GO enrichment followed by pathway enrichment analysis is used to validate all nineteen MoIs that the preservation analysis unit discovered across all six cell types. Second, enrichment analysis, biological analysis, and the presence of prior literature evidence are used to validate all hub-genes found in each MoI.

6.6.1 Enrichment Analysis of Modules

All nineteen MoIs are analyzed for GO enrichment and pathway enrichment as part of the validation process. All enrichment analyses are carried out using the widely known and open-source bioinformatics tool DAVID [628, 253] [10]. The percentage of genes in each module that have annotations in the corresponding GO and KEGG databases is summarized in Table 6.5. A module's biological significance is confirmed

[10] https://david.ncifcrf.gov/home.jsp

by the presence of at least one enriched KEGG pathway and one enriched GO term. In the nineteen MoIs, we have demonstrated that $> 50\%$ of the genes are annotated to enriched pathways, whereas $> 90\%$ are annotated to enriched GO terms. All nineteen MoIs are therefore biologically significant.

Tab. 6.5: Percentage of genes in each MoI that are annotated in the GO databases (BP: Biological Processes, CC: Cellular components or MF: Molecular function) and KEGG pathways. Three CD45+ cell types TC: Tcell, BC: Bcell, MY: Myeloid, and three CD45- cell types EP: Epithelial, EN: Endothelial, and FI: Fibroblasts

Cell Type	Module	Size	BP (%)	CC (%)	MF (%)	KEGG (%)	Cell Type	Module	Size	BP (%)	CC (%)	MF (%)	KEGG (%)
TC	*brown*	151	96.9	98.4	96.9	56.7	EP	*brown*	212	95.8	96.5	97.2	62.0
	blue	276	95.2	97.4	96.0	59.0		*blue*	253	97.0	99.2	95.5	49.6
	turquoise	300	93.9	96.8	96.0	56.7		*turquoise*	330	96.0	97.5	94.6	53.3
BC	*blue*	108	93.9	98.0	99.0	52.0	EN	*brown*	152	91.9	94.6	89.9	50.7
	turquoise	244	96.8	98.2	95.0	50.5		*blue*	313	97.3	98.7	96.0	57.8
MY	*green*	102	95.5	97.8	92.1	43.8		*turquoise*	330	95.9	97.5	94.6	52.2
	yellow	145	93.0	93.0	93.8	62.8	FI	*brown*	114	94.6	94.6	93.8	51.8
	brown	178	96.0	97.7	95.4	64.9		*blue*	250	93.2	98.3	96.2	54.7
	blue	189	95.7	98.9	97.3	61.3		*turquoise*	424	97.0	98.0	96.8	55.6
	turquoise	228	96.8	97.7	96.8	55.3							

6.6.2 Biological Analysis

We validate the biological relevance of the hub-genes identified through GO and pathway enrichment analysis (Sections 2.4.1.1 and 2.4.1.2). These lists, lgEP and lgEGo (Fig. 6.2) are input to the biological analysis component of validation unit. Based on the hub-genes identified biological analysis componenet extraxcts the hub-genes from lgEGo and lgEP. Table 6.6 and Table 6.7 summarize the hub-genes in modules of CD45+ Seurat objects (Tcell, Bcell, and Myeloid) and CD45- (Epithelial, Endothelial, and Fibroblast), respectively that are annotated to top 20 KEGG pathways. Table 6.8 and Table 6.9 summarize the hub-genes that are annotated to top 10, top 3 and top 3 enriched GO terms in GO_BP, GO_CC and GO_MF databases, respectively in the corresponding CD45+ and CD45- cell types, respectively.

Tab. 6.6: Summary of hub-genes detected by scDiffCoAM that have been annotated to the Top 20 KEGG enriched pathways in the CD45+ cell types

KEGG Pathways	T Cell	B Cell	Myeloid
hsa03010:Ribosome			
hsa05012:Parkinson disease		*TUBB3*	*UCHL1*
hsa05020:Prion disease		*TUBB3*	*CAV1*
hsa05169:Epstein-Barr virus infection			*CCND2*
hsa05171:Coronavirus disease - COVID-19	*NLRP3, TNFRSF1A, HBEGF*		
hsa05166:Human T-cell leukemia virus 1 infection	*CCNB2, IL15, TNFRSF1A*		*CCND2, TNFRSF13C*
hsa05014:Amyotrophic lateral sclerosis	*TNFRSF1A*	*TUBB3*	
hsa04145:Phagosome		*TUBB3*	
hsa00190:Oxidative phosphorylation			
hsa05016:Huntington disease		*TUBB3*	
hsa04640:Hematopoietic cell lineage		*CD1D*	
hsa04380:Osteoclast differentiation	*TNFRSF1A*		
hsa04932:Non-alcoholic fatty liver disease	*TNFRSF1A*		
hsa04141:Protein processing in endoplasmic reticulum	*SSR4*		*ERLEC1*
hsa05323:Rheumatoid arthritis	*FLT1, IL15*		
hsa04659:Th17 cell differentiation			*GATA3*
hsa04142:Lysosome	*LIPA, LAMP3, ASAH1, MCOLN1, PPT1, CD68, GLA*	*ACP2*	*CTSZ*
hsa03040:Spliceosome			
hsa03050:Proteasome			
hsa04064:NF-kappa B signaling pathway	*TNFRSF1A*		*TNFRSF13C*

Tab. 6.7: Summary of hub-genes detected by scDiffCoAM that have been annotated to the Top 20 KEGG enriched pathways in CD45- cell types

KEGG Pathways	Epithelial	Endothelial	Fibroblast
hsa05012:Parkinson disease			
hsa03010:Ribosome			
hsa05010:Alzheimer disease			FADD
hsa05208:Chemical carcinogenesis - reactive oxygen species	*GSTA4*	*CBR1, EGFR*	
hsa04510:Focal adhesion	*COL4A1, COL9A1, RAC2*	*ITGA3, EGFR, RELN*	
hsa05020:Prion disease	*RAC2*		
hsa05016:Huntington disease	*TGM2*		
hsa04932:Non-alcoholic fatty liver disease	*LEPR*		
hsa05415:Diabetic cardiomyopathy	*PDK4, G6PD, RAC2*		
hsa05014:Amyotrophic lateral sclerosis			
hsa05200:Pathways in cancer	*RAC2, GSTA4, FGF5, COL4A1*	*EGFR, HEYL, NOTCH3, ITGA3*	*FADD, PTGER4, CDKN2B*
hsa03040:Spliceosome			
hsa05165:Human papillomavirus infection	*COL9A1, COL4A1, OASL*	*EGFR, RELN, HEYL, NOTCH3, ITGA3*	*FADD, PTGER4*
hsa05418:Fluid shear stress and atherosclerosis	*RAC2, GSTA4*	*SELE*	
hsa05022:Pathways of neurodegeneration - multiple diseases			*FADD*
hsa04218:Cellular senescence			*CDKN2B*
hsa04141:Protein processing in endoplasmic reticulum			
hsa05171:Coronavirus disease - COVID-19	*CSF3*	*C3, EGFR*	
hsa05205:Proteoglycans in cancer		*ANK2, EGFR*	
hsa03050:Proteasome			

Tab. 6.8: Summary of hub-genes detected by scDiffCoAM that have been annotated to the top enriched GO terms in the three GO databases for CD45+ cell types

	GO Term	T Cell	B Cell	Myeloid
GO_BP	GO:0002181 cytoplasmic translation			
	GO:0006915 apoptotic process	*GSN, NLRP3, TNFRSF1A*	*CDCA7*	*PRF1, PEA15, ZNF385A, P2RX1, PDCD1*
	GO:0006955 immune response	*IL15, IGKC*	*IFITM3, CD1D, CFP, CCR1, HLA-DQB2, PTAFR, IGKC, IGLV3-1*	*VPREB3, CCR4, CCR2, PTAFR, OSM, IGLV2-14*
	GO:0051301 cell division	*NCAPG, CCNB2*	*CDCA7*	*KIFC1, CCND2*
	GO:0006412 translation			
	GO:0006954 inflammatory response	*LIPA, NLRP3, TNFAIP6, ADM, IL15, TNFRSF1A*	*CCR1, HCK, PTAFR*	*CCR4, CCR2, MGLL, PTAFR*
	GO:0043066 negative regulation of apoptotic process	*ATF5, PPT1*	*PLK2, HCK*	*CCND2*
	GO:0000398 mRNA splicing			
	GO:0007049 cell cycle	*FANCI, MCM7*		
	GO:0009615 response to virus		*IFITM3*	*GATA3*
GO_CC	GO:0005829 cytosol	*ATF5, DHFR, GMNN, SRM, IL15, PPT1, GSN, FANCI, MCM7, PSTPIP2, NLRP3, RBKS, LIPA, NCAPG, DTL, IFI30, CCNB2, MNDA*	*CDCA7, SIGLEC6, BCAT1, MXD1, PLK2, GCHFR, PGD, HCK, HK2, CYFIP1*	*PLBD1, PGM2L1, TBXAS1, MID1IP1, LGALS3, ACY3, SDSL, HMOX1, PEA15, MGLL, SPRED2, TIFA, PRF1, CCR2, DBN1, CCND2, AFF3, YWHAH, BACH2, NUB1, BEST1, CUL9, PTPN13, UCHL1, B9D1, DOCK4, GMDS*
	GO:0016020 membrane	*NCAPG, SLC16A3, ITGAX, TNFRSF1A, MCOLN1, CD68, CCNB2, PPT1, CDH1, FANCI, MCM7, PSTPIP2, MPP1*	*PTAFR, SIGLEC6, ACP2, CD81, GPR137B, HK2*	*PTAFR, PRF1, LGALS3, PTTG1IP, CCND2, KIR2DL4, CAV1, A4GALT, HMOX1, KIFC1, MGLL, SLC8A1, BEST1, B9D1, DOCK4, SPRED2*
	GO:0005654 nucleoplasm	*LIPA, ATF5, GMNN, DTL, MCOLN1, IL15, RAD51AP1, FANCI, MCM7, MNDA*	*CDCA7, MXD1, GCHFR*	*GATA3, JAML, LGALS3, PTTG1IP, CCND2, AFF3, ZNF385A, BACH2, NUB1, HMOX1, PEA15, MGLL, SLC8A1, PTPN13, UCHL1, LMO2*

Continued on next page

	GO Term	T Cell	B Cell	Myeloid
GO_MF	GO:0005515 protein binding	*ARL5B, TNFRSF1A, SRM, TNFAIP6, FANCI, MCM7, RBKS, SLC16A3, DTL, MERTK, TPRA1, CCNB2, HBEGF, MNDA, B3GNT7, ATF5, GMNN, ITGAX, IL15, TCF4, GLA, PPT1, GSN, CLCF1, IGHG1, MPP1, NLRP3, NCAPG, FLT1, MCOLN1, IFI30, SLC12A8, SLAMF8, CD68, CDH1, TPD52, RAD51AP1, SGPL1, TRAM2*	*PTAFR, CFP, PLK2, IFITM3, GPR137B, CDCA7, SIGLEC6, ACP2, MXD1, GCHFR, HCK, HK2, ZNF296, ABCA1, CD1D, CCR1, CD81, CYFIP1, TUBB3, TRAM2*	*PTAFR, PDCD1, LGALS3, FAM3C, TIFA, SPON2, PTTG1IP, PILRA, DBN1, CCND2, KIR2DL4, YWHAH, DOCK4, GMDS, GATA3, MID1IP1, ACY3, OSM, CAV1, HMOX1, PEA15, MGLL, SLC8A1, EMILIN2, P2RX1, LMO2, SPRED2, PRF1, CTSZ, CCR4, CCR2, ERLEC1, BACH2, NUB1, CUL9, IGFBP7, GCSAM, PTPN13, SGPL1, UCHL1, B9D1*
	GO:0003723 RNA binding	*SLC16A3, RAD51AP1*		*LGALS3, ZNF385A*
	GO:0003735 structural constituent of ribosome			

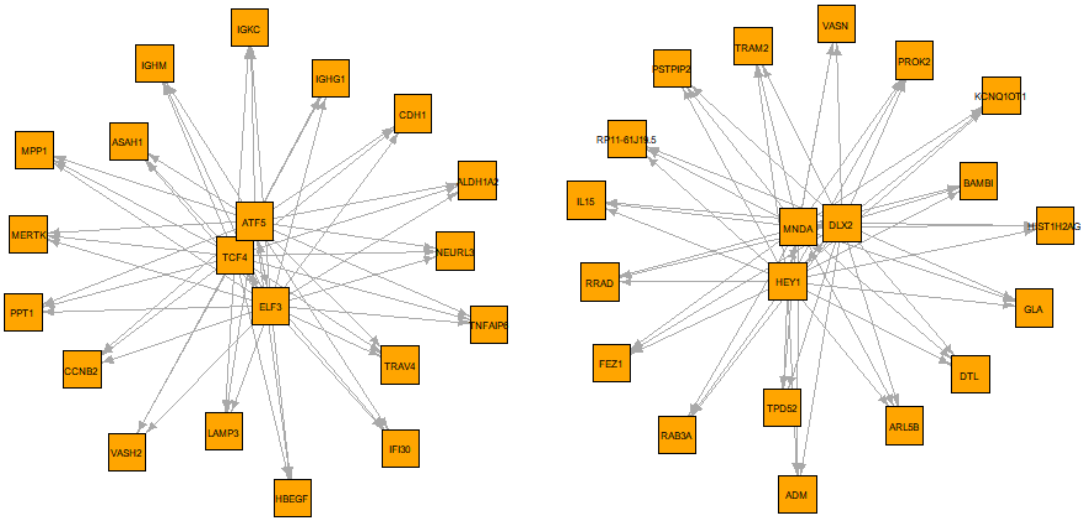
Tab. 6.9: Summary of hub-genes detected by scDiffCoAM that have been annotated to the top enriched GO terms in the three GO databases for CD45- cell types

	GO Term	Epithelial	Endothelial	Fibroblast
GO_BP	GO:0002181 cytoplasmic translation			
	GO:0006412 translation		*EGFR*	
	GO:0006915 apoptotic process	*IFI27, IL24, GSN, SRGN*	*TNS4, GJB6*	*BEX2, FADD, SGK1*
	GO:0043066 negative regulation of apoptotic process	*IFIT3, KRT18, TGM2*	*EGFR*	*ATF5, CITED2, KRT18*
	GO:0016477 cell migration		*EPHA2*	*GPC4*
	GO:0001525 angiogenesis	*SAT1, ANPEP, LEPR, ROBO4, EGFL7*	*ESM1*	*MMP19, PLXDC1, APLN*
	GO:0051301 cell division		*CDC6, MCM5*	
	GO:0045893 positive regulation of transcription	*SOX7, MED24, SOST*	*SOX7, EGFR*	*ATF5, CITED2*
	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	*CSF3*	*EGFR, HEYL, NOTCH3, TOP2A, MEOX1, PROX1*	*ATF5, PAX9, FADD, CITED2, ETV5*
	GO:0007049 cell cycle	*KRT18, RGS2*	*AVPI1*	*BEX2, CDKN2B, KRT18, RGS2, MCM3, MCM5*
GO_CC	GO:0005829 cytosol	*CH25H, CRABP2, RAC2, TGM2, GSN, IFIT3, KRT5, GSTA4, OASL, RBP5, SAT1, S100A1, RGS5, RGS2, G6PD, SERPINE2, KRT18, KRT16, KRT6B, KRT6A, ARHGAP15*	*ACTA2, GJB6, HMMR, CBR1, PROX1, CSTA, IQCG, IDO1, FABP4, DSC2, ANK2, NDRG1, NOTCH3, TNS4, KRT19, NAPRT, AOX1*	*ATF5, ECT2, GGH, SHANK2, CDKN2B, MCM5, TACSTD2, INPP1, RGS2, NAA20, CDC6, FADD, HILPDA, SGK1, KRT18, KRT14*
	GO:0070062 extracellular exosome	*RBP5, CRABP2, ANPEP, RAB27B, RAC2, TGM2, G6PD, PROCR, PRSS23, GSN, CLEC3B, KRT5, ROBO4, CST6, KRT18, KRT16, KRT6B, KRT6A*	*C3, FXYD3, FABP4, SLC38A1, DSC2, ITGA3, OMD, SCNN1A, NDRG1, ACTA2, PLA2G2A, ZG16B, CBR1, KRT19, NAPRT, AOX1, IQCG*	*HLA-DRB5, PTGR1, TACSTD2, GPC4, GGH, MMP7, CLIC3, GDF15, KRT18, KRT14, ABCC1, PON3*
	GO:0005654 nucleoplasm	*CRABP2, MED24, OASL, S100A1, OLFM2, MSC, SOX7*	*UBE2T, TOP2A, PROX1, CSTA, GINS2, HEYL, MSC, RCL1, NOTCH3, SOX7*	*ATF5, DUSP2, ECT2, ETV5, DENND2C, MCM3, MCM5, CITED2, CDC6, PAX9, HILPDA, SGK1, CMPK2*

Continued on next page

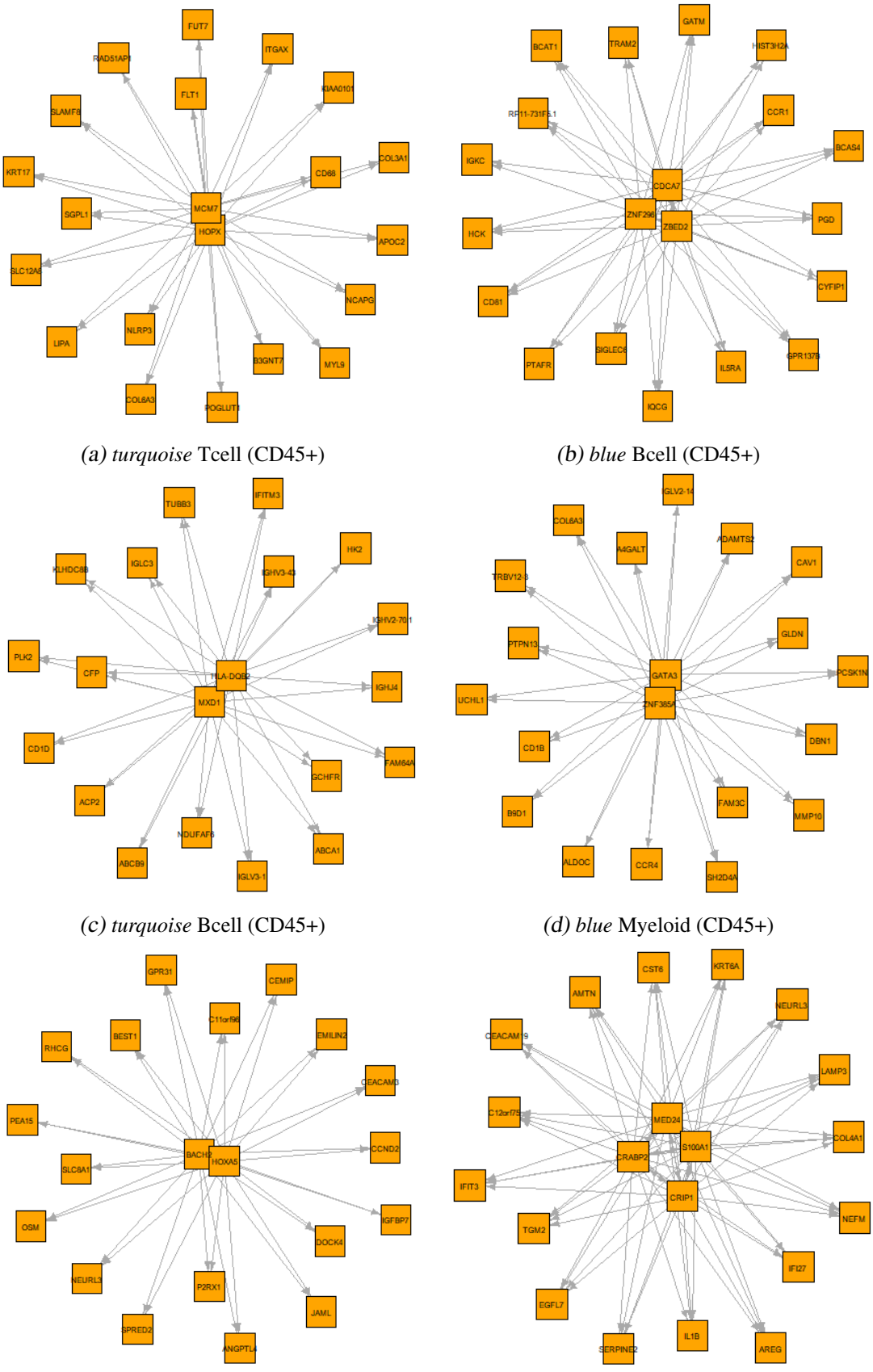
	GO Term	Epithelial	Endothelial	Fibroblast
GO_MF	GO:0005515 protein binding	*CH25H, TNFRSF4, RAC2, TGM2, DAPL1, G0S2, IFIT3, PDK4, APCDD1, MTSS1, MATN2, GSTA4, OASL, RBP5, AREG, ENTPD3, COL9A1, PCSK7, PROCR, MSC, SOST, NPW, KRT18, KRT16, KRT6B, KRT6A, NFKBID, CXCL6, EGFL7, CRABP2, MED24, RAB27B, IL24, GSN, CEACAM19, KLHDC7B, KRT5, ROBO4, CLDN11, CST6, SRGN, AMTN, SRPX, HAS2, SAT1, LEPR, S100A1, OLFM2, IFI27, TSPAN7, RGS5, RGS2, G6PD*	*RASL11A, PTX3, GJB6, HMMR, MEOX1, AVPI1, GINS2, SELE, LXN, C3, HEYL, DSC2, EPHA2, NDRG1, MSC, RCL1, ZG16B, KRT19, NUSAP1, NAPRT, ACHE, EGFR, ITGA3, SCNN1A, UBE2T, CYP3A5, DCUN1D5, TOP2A, PROX1, CSTA, GUCY1A2, TMEM97, SLC38A1, NFE2L3, ANK2*	*LYPD5, DUSP2, HTRA3, ECT2, TNFAIP6, ETV5, SHANK2, CDKN2B, INSIG1, MCM3, MCM5, LXN, FGFBP2, CXCL14, AARD, MMP7, NAA20, BEX2, FADD, CLIC3, SGK1, KRT18, KRT14, NFKBID, EGFL8, CMTM5, ATF5, CCDC102B, SYNM, GGH, RAET1L, CITED2, ZFAND2A, PTGER4, CLDN4, PTGR1, PLXDC1, TACSTD2, INPP1, GPC4, RGS2*
	GO:0003723 RNA binding	*COL14A1, IFIT3, KRT18, OASL*	*COL14A1, TOP2A, NUSAP1*	*KRT18*
	GO:0045296 cadherin binding		*EPHA2, EGFR, NOTCH3, NDRG1*	

Transcription Factors (TF) have remarkable diversity as well potency as drivers of cell transformation. Deregulation of TFs is a pervasive theme across many forms of human cancer, justifying the continued pursuit of TFs as potential biomarkers [45]. We observe that in CD45+, five, five and seven hub-genes detected by scDiffCoAM in Tcell, Bcell, and Myeloid, respectively are TFs. Similarly, 6, 8 and 9 hub-genes in CD45- cell types Epithelial, Endothelial, and Fibroblast, respectively are TFs. It is noteworthy however that *ATF5* TF is a hub-gene in both Tcell(CD45+) and fibroblast (CD45-). On the other hand , TF *MSC* is a hub-gene in epithelial (CD45-) and endothelial (CD45-) while TF *NFKBID* is a hub-gene in both epithelial (CD45-) and fibroblast (CD45-). Regulatory behaviors exhibited by these 20 TFs in their respective modules establish their biological relevance. With the aim to achieve comprehensive visualization, we extracted a manageable subset of hub-genes from the MoIs for these 20 TFs that are also hub-genes detected by scDiffCoAM. We construct a Gene Regulatory Network (RN) (Fig. 6.12a-6.15) with these hub-genes and associated TFs so as to observe the regulatory behavior of the corresponding genes. The resulting RN is in the form of an adjacency list with weighted directed edges from TFs to other target genes (TGs). In module *blue* (Tcell) (Fig. 6.12a), three hub-genes *ATF5,TCF4* and *ELF3* are TFs.



(a) blue Tcell (CD45+) *(b) brown* Tcell (CD45+)

Fig. 6.12: GRN for modules a) *blue*, and b) *brown* in CD45+ cell type, Tcell



(a) turquoise Tcell (CD45+)

(b) blue Bcell (CD45+)

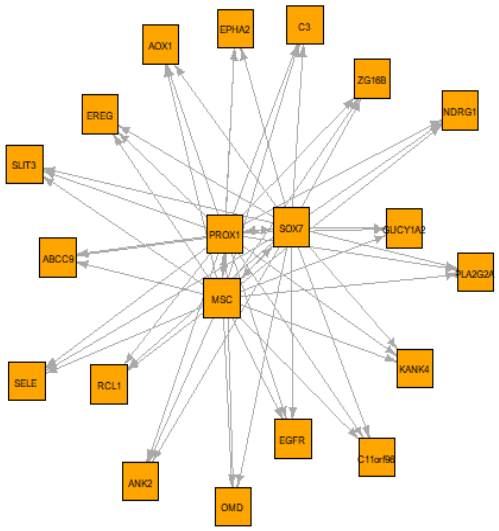
(c) turquoise Bcell (CD45+)

(d) blue Myeloid (CD45+)

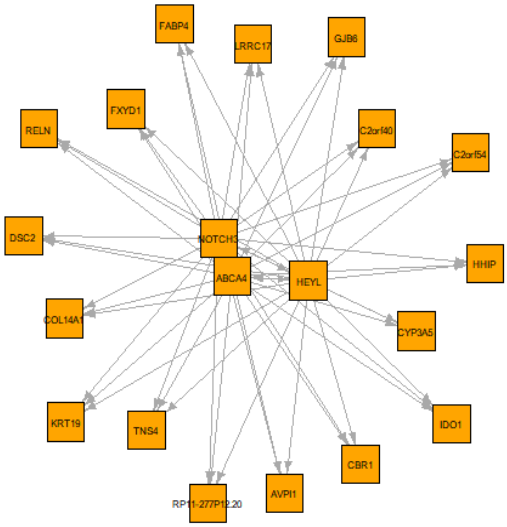
(e) turquoise Myeloid (CD45+)

(f) turquoise Epithelial (CD45-)

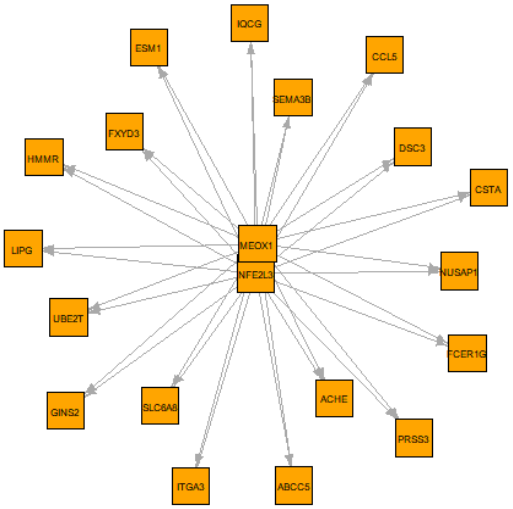
Fig. 6.13: GRN for modules a) *turquoise* in Tcell, and b) *blue* and c) *turquoise* in Bcell, and d) *blue* and e) *turquoise* in Myeloid. GRN for module f) *turquoise* in Epthelial.



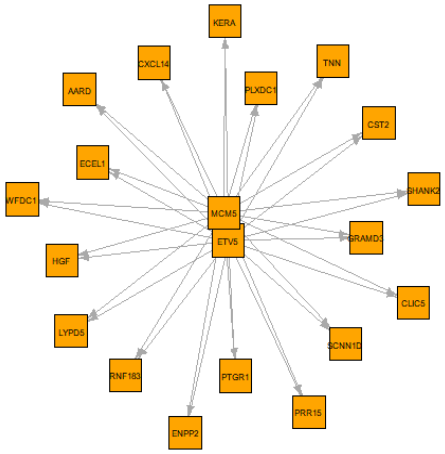
(a) blue Endothelial (CD45-)



(b) brown Endothelial (CD45-)



(c) turquoise Endothelial (CD45-)



(d) brown Fibroblast (CD45-)

Fig. 6.14: GRN for modules a) *blue*, b) *brown* and c) *turquoise* in Endothelial, and d) *brown* in Fibroblast.

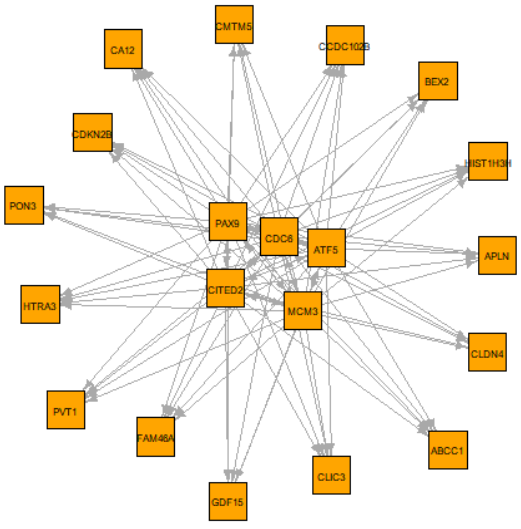


Fig. 6.15: GRN for module *turquoise* in Fibroblast

6.6.3 Literature Trace

Through tracing existing literature we perceive the association of the hub-genes detected by scDiffCoAM to ESCC. Furthermore, we also take into consideration the association with five other Squamous cell carcinomas (SCCs) namely, Oral SCC, Lung SCC, Tongue SCC, Head and Neck SCC, and Laryngeal SCC with the assumption that genes established as potential biomarkers in these SCCs might also be potential biomarkers for ESCC in specific.

- He et. al [225] found ATF5 to be upregulated in ESCC and their findings suggest that inhibition of ATF5 activity can be anti-tumorigenic.
- Yi et al.[826] found that Angiopoietin-like protein 4 (ANGPTL4) upregulation may play an important role in ESCC development, and serum ANGPTL4 level may be a potential tumor marker for ESCC diagnosis and prognosis. Shibata et al.[633] found that ANGPTL4 may potentially affect the prognosis of ESCC due to its role in metastasis through lymphovascular invasion.
- Yang et. al [817] identifies CCNB2 as one of 10 hub genes that might function as novel biomarkers for ESCC.
- Ando et al.[26] found that the ESCC patients with positive staining for caveolin-1 (CAV1) had significantly shorter survival than those with negative staining and thus CAV1 is a potential prognostic marker of ESCC. According to Kato et al., [301], over-expression of CAV1 is associated with lymph node metastasis and a worse prognosis after surgery in ESCC. Jia et al.[283] found that down-regulation of stromal CAV1 expression in ESCC had high malignant potential and suggests that it could be a powerful prognostic marker for patients with ESCC.
- Studies by Wu et al. [777] find that the chemokine (C-C motif) ligand 5 (CCL5) autocrine loop may promote ESCC progression. Results presented by Liu et al. [421] indicate that CCL5 plays a role in patient survival by serving as the key chemokines to recruit CD8(+) T lymphocytes into ESCC tissue.
- Li et al.[351] found that the overexpression of Cell Division Cycle Associated 7 (CDCA7) promoted proliferation, colony formation, and cell cycle in ESCC cells. Li et al.[350] states that CDCA7 might be a new therapeutic target in the suppression of metastasis and invasion of ESCC.
- According to Ishiguro et. al [270], decreased expression of CpG island hypermethy-

lation of E-cadherin (CDH1) in the cell membranes of cancer cells is associated with poor survival of patients with esophageal cancer. Lee et. al [338] in their study suggests that hypermethylation of CDH1 genes may be significantly associated with a recurrence-associated prognosis in stage I ESCC.

- Ghobadi et al.[188] present an association of a novel genetic variant in CDKN2B gene with the clinical outcome of patients with ESCC.
- Sung et al.[659] indicate that claudin-4 (CLDN4) expression is deregulated in ESCC, implying its potential use as a prognostic biomarker in ESCC. Lin et al.[399] suggest CLDN4 as a prognostic and CCRT response indicator for ESCC patients.
- Li et al. [359] identifies Cellular retinoic acid-binding protein 2 (CRABP2) as a suppressor factor that is expected to be a potential prognosis marker for esophageal squamous cell carcinoma. Yang et al. [815] further demonstrate that CRABP2 acted as a tumor suppressor in ESCC carcinogenesis by significantly inhibiting cell growth, inducing cell apoptosis, and blocking cell metastasis both in vitro and in vivo.
- According to Shiba et al., [632], relatively high levels of cysteine protease inhibitor A (CSTA) expression in tumors were correlated with tumor progression and advanced cancer stage in ESCC.
- Data presented by Guo et al. [205] suggests that for ESCC patients with low-level chemokine (CXC motif) ligand 14 (CXCL14), increasing CXCL14 expression combined with inhibition of SRC or EGFR might be a promising therapeutic strategy.
- According to Fang et al. [153], desmocollin 2 (DSC2) is involved in the transformation and development of esophageal tumors, and its expression level and intracellular localization may serve as a predictor for patient outcomes. Fang et al. [154] suggest that miR-25-mediated down-regulation of DSC2 promotes ESCC cell aggressiveness through redistributing adherens junctions and activating beta-catenin signaling.
- According to Sun et al. [653], epithelial cell transformation sequence 2 (ECT2) could regulate the expression of VEGF and MMP9 to inhibit cells proliferation, invasion, migration, and tumor development through the RhoA-ERK signaling pathway.
- Moghbeli et al.[505] illustrate the oncogenic function of epidermal growth factor receptor (EGFR) in the development of ESCC through advanced stages.
- Miyazaki et al. [504] found that Ephrin receptor A2 (EphA2) overexpression is related to a poor degree of tumor differentiation and lymph node metastasis in ESCC. Syed et al. [660] found that knockdown of EPHA2 in ESCC cell line TE8 resulted in a

significant decrease in cell proliferation and invasion.

- Li et al. [354] establish that silencing Endothelial cell-specific molecule 1 (ESM1) suppressed the proliferation, migration, and invasion of KYSE150 and KYSE510 cells. Zhu et al. [924] found that Overexpression of FXYD-3 in the cytoplasm may play an important role in the tumorigenesis and development in the human ESCC.
- According to Sun et al. [655], E26 transformation-specific (ETS) variant 5 (ETV5) promoted metastasis of ESCC.
- Zhu et al.[922] found that family with sequence similarity 3, member C (FAM3C) expression was dramatically increased in ESCC and might serve as a valuable prognostic indicator for ESCC patients after surgery.
- According to Iwabu et al., [273], fibroblast growth factor 5 (FGF5) methylation is a sensitive marker of ESCC to definitive chemoradiotherapy.
- Chi et al.[97] showed that GATA-binding protein 3 (GATA3) positivity is associated with poor prognosis in ESCC.
- Urakawa et al. [702] found that recombinant human growth differentiation factor 15 (GDF15) promotes cell proliferation and the phosphorylation of both Akt and Erk1/2 in ESCC cell lines in vitro. According to Okamoto et al. [532], GDF15 promotes ESCC progression by increasing cellular proliferation, migration, and invasion.
- Zhou et al. [914] suggest that GINS2 acts as an ESCC promoter and can be a novel diagnostic and prognostic marker.
- Wang et al. [745] employed Cox multivariate assay to demonstrate that glucose-6-phosphate dehydrogenase (G6PD) was an independent prognostic factor for the patients with ESCC. Furthermore, Wang et al. [746] suggest that G6PD may function as an important regulator in the development and progression of ESCC by manipulating STAT3 signaling pathway.
- Results by Ren et al. [577] suggest that serum hepatocyte growth factor (HGF) may be a useful biomarker of tumor progression and a valuable independent prognostic factor in patients with ESCC. The results presented by Xu et al. [796] indicate that the frequent overexpression of HGF proteins, secreted by esophageal epithelium and stromal fibroblasts, promoted the progression of ESCC. Takada et al. [661] indicate that HGF is significantly increased in ESCC and suggests the same as a useful biomarker.
- According to Zhang et al.[865] knockdown of homeobox A5 (HOXA5) suppressed the proliferation and metastasis partly by interfering with Wnt/β-catenin signaling

pathway in ESCC cells.

- Li et al.[373] increased insulin-like growth factor binding protein 7 (IGFBP7) may accelerate ESCC progression by promoting the expression of TGFβ1, α-SMA, and collagen I by activating the TGFβ1/SMAD signaling pathway.
- Jiao et al. [286] indicate that chemotherapy could promote tumor Indoleamine 2,3-dioxygenase (IDO1) expression, and the increased tumor IDO1 expression after neoadjuvant therapy predicted poor pathologic response and prognosis in ESCC.
- Jia et al.[282] demonstrates that interferon-induced transmembrane protein 3 (IFITM3) expression has a close relationship with prognosis in ESCC patients.
- Huang et al.[259] demonstrated that serum IGFBP7 is a potential biomarker in the early detection of ESCC.
- Che et al. [79] found that Interleukin-1 beta (IL-1B) is significantly linked to poor prognosis for patients with esophageal cancer and may be a promising molecular target for therapeutic intervention for ESCC.
- Du et al. [145] suggest integrin subunit α3 (ITGA3) a potential therapeutic target for the treatment of ESCC as they demonstrate that its knockdown suppressed cell proliferation, invasion, migration, and autophagy in ECA109 and TE1 cells.
- According to Cheng et al. [95], KIAA0101 is emerging as a meaningful marker for poor prognosis in EC, such as early recurrence and short survival.
- Results presented by Imai et al.[266] suggest that kinesin family member C1 (KIFC1) plays an important role in ESCC pathogenesis.
- According to Liu et al. [442], Keratin 17 (KRT17) upregulation in ESCC cells not only promoted cell proliferation but also increased invasion and metastasis. Haye et al. [224] established that KRT17 is a negative prognostic biomarker for the most common subtype of esophageal cancer.
- In their study Liao et. al [392] suggests that epithelial Lysosomal-associated membrane protein 3 (LAMP3) expression is an independent prognostic biomarker for ESCC. Furthermore, Huang et. al [254] identifies the role of LAMP3 in promoting cellular motility and metastasis in ESCC.
- According to the findings of Qiu et al. [563], mini-chromosome maintenance complex component 7 (MCM7) activates the AKT1/mTOR signaling pathway leading to the promotion of colony formation and migration of ESCC cells as well as tumor cell proliferation. Zhong et al. [906] state that MCM7 may serve as effective prognostic

factors and could also be used as biomarkers for predicting various clinical outcomes of ESCC in the Chinese population. According to et al. Ahn et. al [10], MCM7 expression is associated with the invasiveness of ESCC.

- Zhou et. al. [909] Serum autoantibody levels of matrix metalloproteinase-7 (MMP-7) may be a good diagnostic biomarker for esophageal squamous cell carcinoma. Malik et al. [484] state that the determination of the matrix metalloproteinase-7 (MMP-7) genotype may provide a useful genetic marker in predicting high-risk individuals for the development of ESCC. Data presented by Miao et al. [496] illustrates that overexpression of MMP-7 may be a suitable diagnostic biomarker for ESCC.
- Zeng et al. [850] overexpression of matrix metalloproteinase-9 (MMP-9) may be a potential independent prognosis factor of ESCC patients in Asia. Li et al. [378] MMP-9 may play important roles in ESCC carcinogenesis.
- Xie et al. [783] demonstrate that Metastasis suppressor-1 (MTSS1) expression in ESCC cells significantly influenced the aggressiveness of the esophageal cancer cells, by reducing their cellular migration and in vitro invasiveness.
- The study by Du et al.[144] suggests that MYC-associated factor X dimerization protein 1 (MXD1) is a crucial prognostic factor in ESCC patients.
- Wang et al. [725] MYL9 expression might be a promising prognostic marker and therapeutic target in ESCC.
- Ueki et al. [700] establish that (N-myc downstream regulated gene-1) NDRG1 plays a pivotal role in tumor progression and development of chemo-resistance in patients with ESCC undergoing neoadjuvant chemotherapy. Ando et al. [25] suggest that up-regulation of NDRG1 mRNA expression levels could be a good candidate for prognosis markers in ESCC. Ai et al. [11] indicate the pro-oncogenic role of NDRG1 in ESCC whereby it modulates tumor progression.
- Chen et al. [81] find that Nuclear factor, erythroid 2 like 3 (NFE2L3) affects the radiosensitivity of ESCC cells through IL-6 transcription and IL-6/STAT3 signaling pathway making it a putative target to regulate ESCC cell radiosensitivity.
- According to Yu et. al [837], the NLR pyrin family domain containing 3 (NLRP3) inflammasome is upregulated in human ESCC tissues and promotes ESCC progression. Findings by Zhou et al. [908] indicate that Alpha-1 Type III Collagen (COL3A1) confers a poor prognosis and malignant phenotype in ESCC, potentially representing a novel biomarker and/or providing a new curative target for ESCC.

- According to Matsuura et al., [493], NOTCH3 may serve as a novel biomarker to predict better clinical outcomes in ESCC patients. Pramanik et al.[555] indicate that the NOTCH3 H score is an independent predictor of survival in ESCC.
- Guan et al. [200] establish that suppression of nucleolar spindle-associated protein 1 (NUSAP1) inhibited cellular proliferation and invasion, and induced cell cycle arrest and apoptosis in vitro.
- Tan et al.[666] identifies paired box 9 (PAX9) as an independent prognostic factor for the surgical treatment of ESCC and a possible predictor of radiation sensitivity.
- Ren et al. [575] suggest that phospholipase A2 group IIA (PLA2G2A) may serve as a useful marker for the prognostic evaluation of ESCC patients. Zhai et al. [851] show that in patients with ESCC, PLA2G2A overexpression and PLA2G2A co-expression with COX-2 is significantly correlated with the advanced stage.
- Yokobori et al. [829] suggest that high expression of prospero homeobox 1 (PROX1) in ESCC could be used as an indicator of poor prognosis and as such it is a promising candidate molecular target for ESCC treatment.
- Li et al. [362] suggest that plasmacytoma variant translocation 1 (PVT1) promotes ESCC progression via functioning as a molecular sponge for miR-203 and LASP1. Similarly, Hu et al. [248] establish that PVT1 promoted ESCC progression via the miR-128/ZEB1/E-cadherin axis. According to Li et al.[345], up-regulated PVT1 can induce ESCC tumorigenesis by regulating the cell cycle and Wnt signaling pathway.
- Through multivariate Cox regression analyses, Yu et al. [834] validates that RAB27B expression is an independent prognostic factor for unfavorable overall survival in ESCC.
- Hu et al. [252] results demonstrated that RAD51-associated protein 1 (RAD51AP1) silencing significantly inhibited cell proliferation and invasion in ESCC, thereby highlighting its potential as a novel target for ESCC treatment.
- Ming et al.[498] supports the notion that RHCG is a novel tumor suppressor gene that plays an important role in the development and progression of ESCC.
- Findings by Zhang et al. [868] suggest that serpin family E member 2 (SERPINE2) promotes tumor metastasis by activating BMP4 and could serve as a potential therapeutic target for clinical intervention in ESCC.
- Tang et al. [673] suggest that semaphorin 3B (SEMA3B) is an important tumor-suppressor gene in the malignant progression of ESCC, as well as a valuable prog-

nostic marker for ESCC patients. Dong et al. [142] suggests SEMA3B as tumor suppressors and may serve as potential targets for antitumor therapy.

- As in the case of NDRG1, Ueki [700] found that serum-and glucocorticoid-regulated kinase 1 (SGK1) also plays a pivotal role in tumor progression and development of chemo-resistance in patients with ESCC.
- Zhu et al. [923] provides evidence that elevated serum SRGN has prognostic significance in ESCC patients, and sheds light on the molecular mechanism by which elevated circulating serglycin (SRGN) in cancer patients might promote cancer progression.
- He et. al [226] suggested that deregulation of T cell transcription factor-4 (TCF4) isoform may contribute to the tumorigenesis of ESCC.
- Yu et al.[839] identify that the expression level of tubulin beta 3 class III (TUBB3) and 4 other genes is closely associated with the clinical characteristics of patients with ESCC. Gong et al.[194] show that TUBB3 negative expression prior to treatment and pCR may indicate a better prognosis for stage II and III ESCC patients.
- According to et al., ubiquitin-conjugating enzyme E2 T (UBE2T) is involved in the development of ESCC, and gene signatures derived from UBE2T-associated genes are predictive of prognosis in ESCC.
- Wang et al.[724] highlight that ubiquitin carboxyl-terminal esterase L1 (UCH-L1) expression significantly increased with the progression of ESCC, implying the importance of UCH-L1 as a potential biomarker in cancer diagnosis and treatment.
- Ninomiya et al. [523] suggest that high Vasohibin-2 (VASH2) expression may be novel independent predictors of a poor prognosis in patients with ESCC. Furthermore, according to [799], high plasma concentrations were associated with poor clinical outcomes for both VASH1 and VASH2.

From all hub-genes detected in nineteen MoIs, we first identify the hub-genes that have previous literature traces of association with ESCC and five other previously mentioned SCCs. In Table 6.10 we summarize all hub-genes with literature trace to all six SCCs and can be termed as candidates for ESCC potential biomarkers. This is then followed by the establishment of the biological relevance of these candidates. Table 6.10 summarizes the literature evidence associated with hub genes (candidates) and corresponding GO databases they are annotated to, the associated enriched pathways as well as whether they exhibit regulatory behavior (TF).

Tab. 6.10: Summary of potential biomarkers candidates identified by scDiffCoAM. Here, All 3 under GO databases imply all three databases, BP, CC, and MF.

Dataset	Hub-Gene	GO Database	Enriched Pathway(s)	TF?	SCC Literature Evidence
Tcell *blue*	ALDH1A2	None	None	No	HNSCC [611, 240]
	ATF5	All 3	None	Yes (Fig. 6.12a)	ESCC[225]
	CCNB2	All 3	hsa05166,hsa04110,hsa04218,hsa04115,hsa05170,hsa04068, and hsa04114	No	ESCC[817]
	CDH1	All 3	hsa05200,hsa05216,hsa04514,hsa05100,hsa05219,hsa04520, and hsa05213	No	ESCC[270, 338]
	IGHG1	All 3	None	No	TSCC[905]
	LAMP3	BP,CC	hsa04142	No	ESCC[392, 254]
	MERTK	All 3	None	No	HNSCC[714]
	TCF4	All 3	None	Yes (Fig. 6.12a)	ESCC[271, 226]
	ELF3	None	None	Yes (Fig. 6.12a)	OSCC[3]
	PPT1	All 3	hsa04142,hsa01212	No	OSCC[462]
	VASH2	None	None	No	ESCC[523, 799], LSCC [428]
Tcell *brown*	DLX2	None	None	Yes(Fig. 6.12b)	LSCC[257], HNSCC[579]
	HEY1	None	None	Yes(Fig. 6.12b)	HNSCC[579]
	TPD52	All 3	None	No	OSCC[5]
	TRAM2	CC,MF	None	No	OSCC[175]
Tcell *turquoise*	NLRP3	All 3	hsa05171,hsa05132,hsa05164,hsa05130,hsa05135,hsa05417, hsa05131,hsa04625,hsa04621,hsa05133,hsa04217	No	ESCC[837], OSCC[161, 721]
	CD68	All 3	hsa04142	No	ESCC[747]
	COL3A1	None	None		ESCC [908], HNSCC[624]
	FLT1	All 3	hsa05323,hsa05202,hsa04010, and hsa04066	No	HNSCC[709]
	KIAA0101	None	None	No	ESCC[95]
	KRT17	None	None	No	ESCC[442, 224]
	MCM7	All 3	hsa04110, and hsa03030	Yes(Fig. 6.13a)	ESCC[563, 906, 10], OSCC[157]
	MYL9	None	None	No	ESCC[725]
	RAD51AP1	All 3	None	No	ESCC[252]
Bcell *blue*	BCAT1	All 3	hsa00270, and hsa01230	No	HNSCC[723]
	CDCA7	All 3	None	Yes(Fig. 6.13b)	ESCC[359, 350]

Continued on next page

Dataset	Hub-Gene	GO Database	Enriched Pathway(s)	TF?	SCC Literature Evidence
	TRAM2	CC,MF	None	No	OSCC [175]
Bcell *turquoise*	IFITM3	All 3	None	No	ESCC[282], HNSCC[353], OSCC[177]
	MXD1	All 3	None	Yes(Fig. 6.13c)	ESCC[144]
	TUBB3	All 3	hsa05012,hsa05020,hsa05014,hsa04145,hsa05016,hsa05132, and 3 others	No	ESCC[839, 194], OSCC[536], HNSCC[315]
Myeloid *blue*	CAV1	All 3	hsa05020,hsa05416,hsa04144,hsa05418,hsa05100, and hsa05205	No	ESCC[26, 301, 283]
	CCR4	All 3	hsa05167,hsa05203,hsa04062,hsa04061, and hsa04060	No	HNSCC[883, 495], TSCC[728]
	FAM3C	All 3	None	No	ESCC[922], OSCC[769]
	GATA3	All 3	hsa04659,hsa04658, and hsa05321	Yes (Fig. 6.13d)	ESCC [97]
	PTPN13	All 3	hsa04210	No	HNSCC[524]
	MMP10	None	None	No	ESCC[198, 415]
	UCHL1	All 3	hsa05012, and hsa05022	No	ESCC[724], HNSCC[856]
Myeloid *green*	KIFC1	All 3	None	No	ESCC[266]
Myeloid *turquoise*	ANGPTL4	None	None	No	ESCC[826, 633], OSCC[669], HNSCC[395]
	HOXA5	None	None	Yes(Fig. 6.13e)	ESCC[865], OSCC[585]
	IGFBP7	All 3	None	No	ESCC[373, 259]
	RHCG	None	None	No	ESCC[498]
Myeloid *yellow*	SPON2	All 3	None	No	Larygeal SCC[520]
	PDCD1	All 3	hsa04660,hsa05235, and hsa04514	No	HNSCC[191]
	PRF1	All 3	hsa04210,hsa05416,hsa05330,hsa05332,hsa04940,hsa04650, and hsa05320	No	HNSCC[151]
Epithelial *brown*	AKR1B10	All 3	hsa01100	No	OSCC[314], LaSCC[423]
	SLC7A11	None	None	No	HNSCC[234], LaSCC[479]
	G6PD	All 3	hsa05415,hsa00480,hsa01100,hsa05230, and hsa01200	No	ESCC [745, 746]
	RAB27B	All 3	None	No	ESCC[834]
	MTSS1	All 3	None	No	ESCC[783], HNSCC[124], LSCC[303]
	RGS2	All 3	hsa04022	No	OSCC[400], TSCC [4]
Epithelial *turquoise*	CRABP2	All 3	None	Yes(Fig. 6.13f)	ESCC[359, 815], HNSCC[61]
	IFI27	All 3	None	No	OSCC[722]
	IFIT3	All 3	None	No	OSCC[552]

Continued on next page

Dataset	Hub-Gene	GO Database	Enriched Pathway(s)	TF?	SCC Literature Evidence
Epithelial *turquoise*	IL-1B	None	None	No	ESCC [79], OSCC[336]
	LAMP3	BP,CC	hsa04142	No	ESCC [392, 254]
	SERPINE2	All 3	No		ESCC[868]
	TGM2	All 3	hsa05016	No	LSCC[73]
Epithelial *blue*	FGF5	BP,CC	hsa05200,hsa04151,hsa04810,hsa04015,hsa04010,hsa04014, and 3 others	No	ESCC[273]
	MMP9	None	None	No	ESCC[850, 378], OSCC[547, 117]
	CLDN11	All 3	hsa05130,hsa04530,hsa05160, and hsa04670	No	LaSCC[625]
	CLEC3B	All 3	None	No	LSCC[654]
	IL24	All 3	None	No	HNSCC[562]
	RGS5	All 3	None	No	TSCC[4]
	SRGN	All 3	None	No	ESCC[923]
Endothelial *blue*	EPHA2	All 3	hsa04151,hsa04360,hsa04015,hsa04010, and hsa04014	No	ESCC[504, 660],HNSCC[437, 582], LSCC[668, 155]
	NDRG1	All 3	None	No	ESCC[700, 25, 11], OSCC [125, 143]
	EGFR	All 3	hsa05208,hsa05200,hsa05205,hsa05215,hsa05225,hsa04510, and 34 others	No	ESCC[505]
	PLA2G2A	All 3	hsa01100, and hsa04014	No	ESCC[575, 851]
	PROX1	All 3	None	Yes (Fig. 6.14a)	ESCC [829], OSCC [606, 587]
Endothelial *brown*	DSC2	All 3	hsa05412	No	ESCC[153, 154]
	FABP4	All 3	None	No	OSCC[337]
	IDO1	All 3	hsa01100,hsa00380, and hsa01240	No	ESCC[286], HNSCC[147], OSCC[646]
	KRT19	All 3	hsa04915	No	LSCC[844]
	NOTCH3	All 3	hsa05200,hsa05224,hsa01522,hsa04371,hsa04658,hsa04330, and hsa04919	Yes (Fig. 6.14b)	ESCC[493, 555]
Endothelial *turquoise*	CCL5	None	None	No	ESCC[777, 421]
	CSTA	All 3	None	No	ESCC[632]
	DSC3	All 3	None	No	LSCC[113, 508]
	GINS2	All 3	None	No	ESCC[914]
	ITGA3	All 3	hsa04510,hsa05200,hsa05165,hsa04151,hsa05222,hsa04810, and 2 others	No	ESCC[145], OSCC[517, 69], HNSCC[156]
	NFE2L3	All 3	None	Yes (Fig. 6.14c)	ESCC[81]
	SEMA3B	All 3	hsa04360	No	ESCC[673, 142]

Continued on next page

Dataset	Hub-Gene	GO Database	Enriched Pathway(s)	TF?	SCC Literature Evidence
Endothelial *turquoise*	ESM1	All 3	None	No	ESCC[354], HNSCC[42, 789]
	FXYD3	All 3	None	No	ESCC[924], LaSCC[116]
	HMMR	All 3	hsa04512	No	HNSCC[455]
	NUSAP1	All 3	None	No	ESCC[200], OSCC[531]
	UBE2T	All 3	None	No	ESCC [743]
Fibroblast *blue*	DUSP2	All 3	hsa04010	No	HNSCC[522]
	ECT2	All 3	None	No	ESCC[653, 236], OSCC[276, 549], LSCC[236], LaSCC[917]
	FADD	All 3	hsa05010,hsa05200,hsa05165,hsa05022,hsa05132,hsa04210, and 16 others	No	OSCC[99], HNSCC [195, 572]
	MMP7	All 3	hsa05166, and hsa04310	No	ESCC[909, 484, 496], TSCC[843]
	RGS2	All 3	hsa04022	No	OSCC [400], TSCC [4]
	SGK1	All 3	hsa04151,hsa04068, and hsa04150	No	ESCC[700]
Fibroblast *brown*	CXCL14	All 3	hsa04062	No	ESCC[205],OSCC[588, 540],HNSCC[317, 380]
	ETV5	All 3	hsa05215, and hsa05202	Yes(Fig. 6.14d)	ESCC[655]
	HGF	None	None	No	ESCC[577, 796, 661]
	MCM5	All 3	hsa04110, and hsa03030	Yes(Fig. 6.14d)	OSCC[836, 216], LaSCC[525]
Fibroblast *turquoise*	PAX9	All 3	None	Yes(Fig. 6.15)	ESCC[666], OSCC[46]
	ATF5	All 3	None	Yes(Fig. 6.15)	ESCC [225]
	CDC6	All 3	hsa04110	Yes(Fig. 6.15)	OSCC[157], TSCC[158]
	CDKN2B	All 3	hsa05200,hsa04218,hsa05166,hsa05222,hsa05203,hsa04110, and 3 others	No	ESCC[188], OSCC[589]
	CLDN4	All 3	hsa05130,hsa04530,hsa05160, and hsa04670	No	ESCC[659, 399], OSCC[127]
	CMTM5	All 3	None	No	OSCC[859]
	GDF15	All 3	None	No	ESCC[702, 532], OSCC[470, 802]
	HTRA3	All 3	None	No	OSCC[511]
	MCM3	All 3	hsa04110,hsa03030	Yes(Fig. 6.15)	OSCC[707, 580]
	PON3	All 3	None	No	OSCC[916]
	PVT1	None	None	No	ESCC[362, 248, 345],OSCC[369],HNSCC[833]

Tab. 6.11: Summary of potential ESCC biomarkers identified by scDiffCoAM using the biomarker criteria (Section 2.5).

	Cell Type	Case 1	Case 2	Case 3	Case 4
CD45+	TC	*MCM7*	*CCNB2, CDH1, LAMP3, NLRP3*	*ATF5, TCF4, RAD51AP1*	
	BC		*TUBB3*	*CDCA7, IFITM3, MXD1*	
	MY	*GATA3*	*CAV1, UCHL1*	*FAM3C,KIFC1, IGFBP7*	
CD45-	EP		*G6PD, LAMP3, FGF5*	*RAB27B, MTSS1, CRABP2, SERPINE2, SRGN*	
	EN	*NOTCH3*	*EPHA2, EGFR, DSC2, IDO1, ITGA3, SEMA3B*	*NDRG1, PROX1, CSTA, GINS2, NFE2L3, ESM1, FXYD3, NUSAP1, UBE2T*	
	FI	*ETV5*	*MMP7, SGK1, CXCL14, CDKN2B, CLDN4*	*ECT2, PAX9, ATF5, GDF15*	*MCM5, CDC6, MCM3*

All hub-genes that belong to Cases 1 and 2 can be considered potential biomarkers for ESCC as discussed in the biomarker criteria (Section 2.5). This is because aside from the existing literature evidence of association to ESCC itself, these hub-genes are biologically relevant as they are annotated to highly enriched GO terms and pathways. Table 6.11 summarizes the cases of all hub-genes (candidates) that has literature trace of association to the ESCC and the other five SCCs fall under. Four hub-genes *MCM7, GATA3, NOTCH3* and *ETV5* fall under case 1 and thus are potential biomarkers for ESCC. These four hub-genes are also TFs and their corresponding GRNs are shown in fig 6.13a (MCM7), fig 6.13d (GATA3), fig 6.14b (NOTCH3), fig 6.14d (ETV5). Even though twenty hub-genes, *CCNB2, CDH1, LAMP3, NLRP3, TUBB3, CAV1, UCHL1, G6PD, FGF5, EPHA2, EGFR, DSC2, IDO1, ITGA3, SEMA3B, MMP7, SGK1, CXCL14, CDKN2B*, and *CLDN4*, do not exhibit regulatory behavior, they are biologically relevant due to their annotation to enriched GO terms and enriched pathways as well as associated to ESCC and other five SCCs

Twenty six hub-genes, *ATF5, TCF4, RAD51AP1, CDCA7, IFITM3, MXD1, FAM3C, KIFC1, IGFBP7, RAB27B, MTSS1, CRABP2, SERPINE2, SRGN, NDRG1, PROX1, CSTA, GINS2, NFE2L3, ESM1, FXYD3, NUSAP1, UBE2T, ECT2, PAX9* and *GDF15* fall under Case 3. Although there exists strong literature on their association with the ESCC and the other five SCCs, none of them have enriched pathways even though many of them are TFs (*ATF5, TCF4, CDCA7, MXD1, CRABP2, PROX1, NFE2L3* and *PAX9*).

Thus, they can be said to be probable potential biomarkers for ESCC but require further in-depth analysis. Three hub-genes , *MCM5, CDC6*, and *MCM3* fall under case 4. These three hub-genes exhibit regulatory behavior and are annotated highly enriched GO terms and pathways establishing their biological relevance. However, they do not have literature evidence of association to ESCC but are associated with the five previously mentioned SCCs. Thus, these hub-genes require further in-depth analysis to be potential biomarkers of ESCC.

Finally, we conclude that twenty-four hub-genes, *MCM7, GATA3, NOTCH3, ETV5, CCNB2, CDH1, LAMP3, NLRP3, TUBB3, CAV1, UCHL1, G6PD, FGF5, EPHA2, EGFR, DSC2, IDO1, ITGA3, SEMA3B, MMP7, SGK1, CXCL14, CDKN2B*, and *CLDN4*, are identified by scDiffCoAM as potential biomarkers for ESCC. Furthermore, twenty-six hub-genes, *ATF5, TCF4, RAD51AP1, CDCA7, IFITM3, MXD1, FAM3C, KIFC1, IGFBP7, RAB27B, MTSS1, CRABP2, SERPINE2, SRGN, NDRG1, PROX1, CSTA, GINS2, NFE2L3, ESM1, FXYD3, NUSAP1, UBE2T, ECT2, PAX9* and *GDF15* have moderate evidence of association to ESCC and requires further in-depth analysis but can be considered probable potential biomarkers for ESCC.

6.7 Discussion

We contrast our method with four other widely used hub-gene finding methods. Two frequently used hub-gene discovery methods, Weighted Gene Score (WGS) and p-value Cut Off (PCO), were proposed by Das et al. [120] in their work Differential Hub Gene Analysis (DHGA). In WGCNA[327], intramodular connectivity (IMC), which determines how connected nodes are to other nodes inside the same module. In in hdWGCNA [509, 510], using eigengene-based connectivity, also known as kME (HWH), of each gene, hub-genes are computed. We give a brief comparison of these four hub-gene discovery techniques with scDiffCoAM. It is unfair to compare these four hub-gene finding techniques with scDiffCoAM. As a result, we use the pipeline below to compare our hub-gene discovery method WGS, PCO, IMC, and HWH.

- We consider the MOIs identified by the framework while preserving the entire pipeline in scDiffCoAM, from pre-processing to preservation analysis.
- On all nineteen MOIs, we apply the other four methods and identify the corresponding lists of the top 20 hub-genes.
- We identify the hub-genes associated with the following cancers: a) ESCC, b) HN-

SCC, c) LaSCC, d) LSCC, e) OSCC, and f) TSCC based on literature evidence for each of the other four methods.

Table 6.12 summarizes the hub-genes detected by a) our hub-gene finding algorithm [592], b) WGS: DHGA[120] weighted gene-score, c) PCO: DHGA[120] p-value Cut Off d) IMC: WGCNA [327] Intramodular-Connectivity and e) HWH: hdWGCNA [510, 509] kME hub-gene finding algorithm. All hub-genes with existing literature associating them as potential biomarkers for ESCC itself (as well as the other five SCCs) are highlighted in red.

From the analysis summarized in Table6.12, we make the following observation. Except for module *brown* in myeloid cell type, our method can detect at least one hub-gene with association to ESCC and other five SCCs in the form of existing literature. Our method can detect at least one hub-gene that has also been suggested in the literature as a potential biomarker for ESCC in most MoIs. The exceptions are modules *brown* in Tcell, *brown*, and *yellow* in myeloid. The other four methods on the hand were able to detect at least one such hub-gene. With the exception of a few modules, most modules extracted by scDiffCoAM include hub-genes that were not detected by the other four methods for that particular module. For the other four methods, in many modules, a few hub-genes are commonly detected by all methods. For example in *blue* in Tcell, *FOXP3* is detected by WGS, IMC and HWH. However, none of the hub-genes detected by our method was detected by WGS, PCO, IMC, or HWH for that module. It is noteworthy that some hub-genes detected by our method in one module may be detected by our method or by another method but in a different module. For example, *LAMP3* detected by our method in *blue* (Tcell) is further by our method in *turquoise* (Epithelial) and by IMC in *turquoise* (Endothelial).

Tab. 6.12: Summary of potential biomarkers detected by scDiffCoAM, WGS: DHGA [120] Weighted Gene Score, PCO:DHGA [120] p-value Cut Off, IMC: WGCNA [327] Intramodular-Connectivity and HWH: hdWGCNA [510, 509] kME hub-gene finding algorithm with strong literature evidence of relation to ESCC (marked in Red), HNSCC, LaSCC, LSCC, TSCC, and OSCC.

	Module	Our Method	WGS	PCO	IMC	HWH
Tcell	*blue*	ALDH1A2[611, 240], ATF5[225], CCNB2[817], CDH1[270, 338], IGHG1[905], LAMP3[392, 254], MERTK[714], TCF4[271, 226], ELF3[3], PPT1[462], VASH2[523, 799]	CTSB[801, 818], PSAP[544], NCF2[560], FOXP3[798, 823, 642], TIMP1[471], GPX1[871, 339]	CSTB[792],CTLA4[878], CTSB[801, 818], EHMT1[201]	FOXP3[798, 823, 642], CTLA4[878], IL2RA[321], TBC1D4[454], GADD45A[269]	TBC1D4[454], CTLA4[878], IL2RA[321], FOXP3[798, 823, 642]
	brown	DLX2[257, 579], HEY1[579], KCNQ1OT1[419], TPD52[5]	IER3[781], ICAM1[691], EST2[363, 371], SGK1[700], MXD1[144], ATF3[785]	EGR1[693], HMGA1[687], ICAM1[691], KLF10[824], MARCKSL1[901]	RHOB[7], HSPA6[613], ATF3[785], IER3[781], MXD1[144], ICAM1[691]	KLF10[824], ETS2[363, 371], ICAM1[691], ITGA5[907], IER3[781],MXD1[144], EGR1[693],ATF3[785]
	turquoise	NLRP3 [837, 161, 721], CD68[747], COL3A1 [908, 624], FLT1[709], KIAA0101[95], KRT17[442, 224], MCM7[563, 906, 10, 157], MYL9[725], RAD51AP1 [252]	CD68[747], HLA-DQA1[622], CCL3[138], HLA-DQB1[246], IFITM3[282, 353, 177], PRF1[151]	ANXA2[372, 474, 477], CCL3[138], CCL4[398], CKS2[310, 178]	PTTG1[160, 272, 857], SOX4[757, 830], LAG3[886, 132], IFI6[440],, PRF1[151],FAM3C[922, 769]	ANXA2[372, 474, 477], IFI44L[535], LAP3[874], CCL3[138], LAG3[886, 132],ISG15[840] , IFI6[440]
Bcell	*blue*	BCAT1[723], CDCA7[359, 350], TRAM2[175]	CD68[747], RAB31[370], NCF2 [560], TIMP1[471], ANXA2[372, 474, 477]	CD68[747], BASP1[568], RAB31 [370]	CD70[126], ANXA2[372, 474, 477], RAB31 [370], NEK6[821], CD68[747]	CPNE5[701], NEK6[821], ANXA2[372, 474, 477], RAB31[370]
	turquoise	IFITM3[282, 353, 177], MXD1[144], FAM64A[897], TUBB3[839, 194, 536, 315]	HLA-DQB1[246], HLA-DQA1[622], TNFSF13[733], SOD2[474, 926], LAIR1[810], IFITM3[282, 353, 177]	CCR6[420, 727], GPX1[871, 339] , HLA-DQA1[622], HLA-DQB1[246]	SATB1[640, 695]	SATB1[640, 695], HLA-DQA1[622]
Myeloid	*blue*	CAV1[26, 301, 283], CCR4[883, 495, 728], FAM3C[922, 769], GATA3[97], PTPN13[524], ALDOC[381], MMP10[198, 415], UCHL1[724, 856]	NCF2 [560], SPP1[794], FBP1[232], SDC2[258], ANXA2[372, 474, 477], CSTB[792], ACTN1[784, 734]	ANXA2[372, 474, 477], CSTB[792], FAM3C[922, 769], GATA3[97], PTTG1[160, 272, 857]	SPP1[794], CSTB[792], SDC2[258], MMP12[485, 209], SPARC[90, 194], CHI3L1[256, 787]	ACTN1[784], ANXA2[372, 474, 477], TGM2[73],NCF2 [560], SDC2[258], FBP1[232], SPP1[794], CSTB[792]

	Module	Our Method	WGS	PCO	IMC	HWH
	green	KIFC1[266]	HLA-DQB1[246], HLA-DQA1[622], DAPK1[717], AXL[242, 858]	HLA-DQB1[246], HLA-DQA1[622], SPINT2[846], TCF4 [271, 226]	DAPK1[717], CCND1[316], AXL[242, 858], TCF4 [271, 226]	NDRG2[631, 665], DAPK1[717], SPINT2[846], AXL[242, 858], HLA-DQB1[246], HLA-DQA1[622]
	brown	None	CD68[747], FCGR2A[453, 118], TMEM176A[750], CTSB [801, 818]	CD68[747], FCGR2A[453, 118], CTSB [801, 818], MAFB[873], TMEM176A[750]	DAB2[29, 739], IDH1[502, 88], MERTK[714]	DAB2[29, 739], CD68[747]
	turquoise	ANGPTL4[826, 633, 669, 395], HOXA5[865, 585], IGFBP7[373, 259], RHCG[498]	IL1B[79, 336], IER3[781], CXCL8[?], SOD2[474, 926], EREG [751, 430], TNFAIP2[?], ICAM1[691], IL1RN[?]	ICAM1[691], IER3[781], IL1RN[?], EST2[363, 371], IL1B[79, 336], SOD2[474, 926], CXCL8[?], TIMP1[471]	IL1RN[?], EREG [751, 430], IL1B[79, 336], SOD2[474, 926], ICAM1[691], MXD1[144]	CXCL8[?], MXD1[144], ICAM1[691], IER3[781], , SOD2[474, 926], EREG [751, 430], IL1B[79, 336]
	yellow	SPON2[520], PDCD1[191], PRF1[151]	IFITM3[282, 353, 177]	IFITM3[282, 353, 177], ISG15[840] , IFI6[440]	IFITM3[282, 353, 177], ISG15[840], IFI6[440], GBP1[357], CD38[630, 140], LAP3[874], IFIT3[552], IFIT1[552], CXCL10[421], IFI27[722]	CXCL9[70], IDO1[286, 147, 646], LAP3[874], GBP1[357], CXCL10[421]
Epithelial	*brown*	AKR1B10 [314, 423], SLC7A11 [234, 479], G6PD [745, 746], RAB27B [834], MTSS1 [783, 124, 303], RGS2 [400, 4]	GPX2[342], KRT19[844], ELF3 [3], CLDN4[659, 399], ALDH3A1[565], EPCAM[491, 490], CLDN7[703, 409], CSTA[632]	CSTA[632], GPX2[342], KRT19[844], CBR1[847], CLDN4[659, 399], CLDN7[703, 409], ELF3 [3]	CDH1 [270, 338], ELF3 [3], CLDN4[659, 399], HES1[663], TGIF1[318]	CDH1 [270, 338], CLDN4[659, 399], ELF3 [3], GBP6[?], GPX2[342]
	turquoise	CRABP2 [359, 815, 61], IFI27 [722], IFIT3 [552], IL-1B [79, 336], LAMP3 [392, 254], SERPINE2 [868], TGM2[73]	UBE2C[364, 537, 290, 426, 820], KRT17[442, 224], TPX2[416, 244], BIRC5 [618, 690], CDKN3[835, 424], PTTG1[160, 272, 857]	CTSB [801, 818], HMGB3[180], IGFBP7[373, 259], LY6K[528, 854], MYL9 [725]	TPX2[416, 244], UBE2C[364, 537, 290], CDKN3[835, 424], RRM2[736, 274], TACC3[262]	BIRC5 [618, 690], CDKN3[835, 424], MCM7 [563, 906, 10, 157], RRM2[736, 274] PTTG1[160, 272, 857], TPX2[416, 244], UBE2C[364, 537, 290],
	blue	FGF5[273], MMP9 [850, 378, 547, 117], CLDN11[625], CLEC3B[654], IL24 [562], RGS5[4], SRGN [923]	COL3A1[908], SPARC[90, 194], MFAP4[214], CTHRC1[747, 335]	COL11A1[299], COL3A1[908]	TMEM176A[750], COL3A1[908], FAP[720, 772]	CTHRC1[747, 335], LOX[866], FSTL1[331], LOXL2[925]

	Module	Our Method	WGS	PCO	IMC	HWH
Endothelial	*blue*	EPHA2[504, 660, 668, 155, 437, 582], EREG [751, 430], NDRG1 [700, 25, 11, 125, 143], EGFR [505], PLA2G2A [575, 851], PROX1 [829, 606, 587]	FBLN1[804], MFAP4[214], SCARA5[293, 435], GPX3[919, 233, 405], CLEC3B[654], MFAP5[791, 557], CHRDL1[774], SFRP2[217, 429]	ADIRF[206], ALDH3A1[565], ANXA1[211, 210], CAV1[26, 301, 283]	SULF1[412], ANXA3[183], SERPINB1[692], IL6[902, 639, 78], PIM1[778, 365, 417]	CLEC3B[654], ANXA3[183], SULF1[412]
	brown	DSC2[153, 154], FABP4[337], IDO1[286, 147, 646], KRT19[844], NOTCH3[493, 555]	CXCL12[698, 607], GIMAP7[704], FLT1[709], IFI27[722], SRGN [923]	SRGN [923], SPINT2[846], GIMAP7[704], FLT1[709]	CBR1[847], CXCL12[698, 607], DSC2[153, 154], GBP1[357], IFI27[722], IFI6[440]	IFI6[440], ISG15[840], SERPINE2 [868], IFI27[722], SRGN [923]
	turquoise	CCL5[777, 421], CSTA[632], DSC3[113, 508], GINS2[914], ITGA3[145, 517, 69, 156], NFE2L3[81], SEMA3B[673, 142], ESM1[354, 42, 789], FXYD3[924, 116], HMMR[455], NUSAP1[200, 531], UBE2T[743]	SPARC[90, 194], KRT17 [442, 224], FXYD3[924, 116]	BGN[921, 894], CCND1[316], CKS2[310, 178]	FSTL1[331], SPARC[90, 194], IL32[480, 658, 202], LAMP3 [392, 254], FSCN1[60, 92, 386], FAP[720, 772]	PCDH17[222, 59], FSTL1[331], IGFBP7[373, 259], SPARC[90, 194]
Fibroblast	*blue*	DUSP2[522], ECT2[653, 236, 276, 549, 236, 917], FADD[99, 195, 572], GBP5[427], MMP7[909, 484, 496, 843], RGS2[400, 4], SGK1[700]	COL3A1 [908, 624], SPARC[90, 194], BGN[921, 894], CTSK[344], CTHRC1[747, 335]	BGN[921, 894], COL3A1 [908, 624], CTHRC1[747, 335], FAP[720, 772], KRT17 [442, 224]	COL3A1 [908, 624], CTSB[801, 818], BGN[921, 894], CTHRC1[747, 335], WISP1[862, 516], FAP[720, 772]	SDC1 [768, 742], BGN[921, 894], CTHRC1[747, 335], SPARC[90, 194], COL3A1 [908, 624]
	brown	CXCL14[205, 588, 540, 317, 380], ETV5[655], HGF[577, 796, 661], MCM5[836, 216, 525]	IGFBP7[373, 259], TIMP1[471], SCN7A[842], CBR1[847], TPD52 [5]	CXCL14[205, 588, 540, 317, 380], IGFBP7[373, 259], MCM7 [563, 906, 10, 157], TIMP1[471]	SCN7A[842], GPC3[24, 33], IGFBP7[373, 259], CBR1[847]	IGFBP7[373, 259], SCN7A[842]
	turquoise	PAX9[666, 46], ATF5[225], CDC6[157, 158], CDKN2B[188, 589], CLDN4[659, 399, 127], CMTM5[859], GDF15[702, 532, 470, 802], HTRA3[511], MCM3[707, 580], PON3[916], PVT1[362, 248, 345, 369, 833]	MFAP4[214], FBLN1[804], SCARA5[293, 435], GPX3[919, 233, 405], TNXB[812], CLEC3B[654]	ADIRF[206], AKR1C1[366, 72], AKR1C2[888, 366], ANXA1[211, 210], CD74[307]	TNXB[812], CD34[486] FHL1[281, 576], GPX3[919, 233, 405], SCARA5[293, 435], AKR1C2[888, 366]	GPX3[919, 233, 405], FHL1[281, 576], FBLN1[804], TNXB[812], IGFBP7[373, 259]

Unlike the other four methods, our method can identify 44 unique hub-genes with reference to existing literature that establishes them as potential biomarkers for ESCC. These hub-genes, *ALDH1A2, ATF5, CCNB2, VASH2, NLRP3, KIAA0101, RAD51AP1, CDCA7, TUBB3, MMP10, UCHL1, KIFC1, ANGPTL4, HOXA5, RHCG, SLC7A11, G6PD, RAB27B, MTSS1, CRABP2, FGF5, MMP9, EPHA2, NDRG1, EGFR, PLA2G2A, PROX1, NOTCH3, CCL5, GINS2, ITGA3, NFE2L3, SEMA3B, ESM1, NUSAP1, UBE2T, ECT2, MMP7, ETV5, HGF, PAX9, CDKN2B, GDF15* and *PVT1*, were detected by our method but not by other four methods. Fourteen hub-genes, *CCNB2, NLRP3, TUBB3, UCHL1, G6PD, FGF5, MMP7, ETV5, CDKN2B, EPHA2, EGFR, NOTCH3, ITGA3*, and *SEMA3B*, are among the twenty-four hub-genes further validated by scDiffCoAM as potential biomarkers for ESCC (Table 6.11). Like the other four methods, our method also identifies and validates ten hub-genes such as potential biomarkers, *CDH1, LAMP3, MCM7, CAV1, GATA3, SGK1, CXCL14, CLDN4, DSC2* and *IDO1*, validated by scDiffCoAM as potential ESCC biomarkers are also detected by other four methods in other modules across cell types.

We analyze the performance of our method against each of the four methods in terms of two parameters: a) Quantity, which measures the number of potential biomarkers identified by a method for the six previously mentioned categories of SCC in general, and b) Quality, which measures the number of potential biomarkers identified by a method for ESCC in particular. A method's overall performance should be favorable for both parameters. We provide a complete assessment of our method's performance in comparison to WGS, PCO, IMC, and HWH in all cell types in terms of two parameters, *quality*, and *quantity*, in Fig. 6.16. In terms of both *quantity* and *quality*, our method outperforms other methods.

6.7.1 Comparison with WGS, PCO,IMC, and HWH

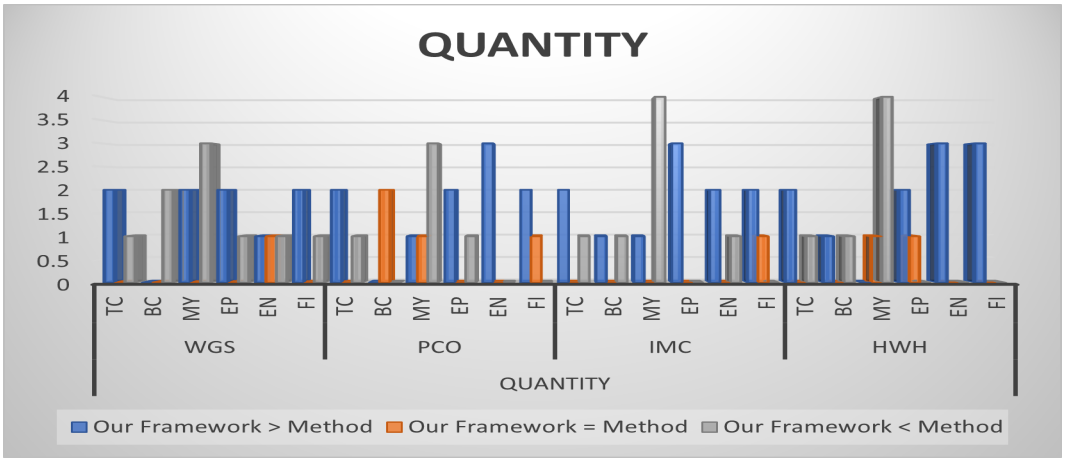
Comparing the performance of our proposed framework, scDiffCoAM with the existing schemes we can make the following observations:

- In nine and eight modules, respectively, scDiffCoAM outperforms WGS in terms of quantity and quality, while performing similarly to WGS in one module in terms of both quantity and quality.
- In ten and eight modules, respectively, scDiffCoAM outperforms PCO in terms of quantity and quality, and in four modules, scDiffCoAM performs similarly to PCO in

terms of quantity and quality.

- In eleven and nine modules, respectively, scDiffCoAM outperforms IMC in terms of quantity and quality, while for one and three modules, respectively, scDiffCoAM performs similarly to IMC in terms of quantity and quality.

(a) Quality



(b) Quantity

Fig. 6.16: Summary of performances of scDiffCoAM vs. three other methods. We compare these methods on MoIs in various cell types. Here, WGS: DHGA [120] Weighted Gene Score, PCO:DHGA [120] p-value Cut Off, IMC: WGCNA [327] Intramodular-connectivity and HWH: hdWGCNA [510, 509] kME hub-gene finding algorithm. *Quantity* measures the number of potential biomarkers identified by a method for the six previously mentioned categories of SCC in general and *Quality* measures the number of potential biomarkers identified by a method for ESCC, in particular.

- In most modules, scDiffCoAM performs better than HWH. In eleven and nine modules, respectively, scDiffCoAM outperforms HWH in terms of quantity and quality, while in two modules, scDiffCoAM outperforms HWH in terms of both quantity and quality. HWH outperforms scDiffCoAM in six and eight modules, respectively, in terms of quantity and quality.
- When compared to the other four methods, the performance of scDiffCoAM in Tcell

also stays consistent. In two of the three modules, scDiffCoAM outperforms WGS, PCO, IMC, and HWH in terms of both quantity and quality.

- In all modules extracted in Bcell, WGS outperforms scDiffCoAM in terms of quantity and quality. However, PCO performs similarly to scDiffCoAM in terms of quantity while performing similarly or better in terms of quality. In one module, scDiffCoAM outperforms IMC in terms of both quality and quantity, while IMC outperforms scDiffCoAM in the other. While HWH performs better in some modules and worse in others when in terms of quantity, scDiffCoAM outperforms or performs similarly with HWH in terms of quality.
- In terms of quantity and quality, all four methods perform better than scDiffCoAM in the majority of modules. It is to be noted that, in contrast to the majority of other cell types where we discovered two to four modules with varied sizes, in myeloid cells we detected nine modules, the majority of which were less than 200, with *turquoise* being the only exception.
- In most cases, scDiffCoAM outperforms all four methods in terms of quantity. In terms of quality, WGS performs better than or on par with scDiffCoAM, scDiffCoAM performs better than or similarly to PCO and IMC in every case, and scDiffCoAM performs better than or similarly to HWH in two out of the three modules. Even though epithelial detects more modules (eight) than other cell types, similar to myeloid, all MoIs are sizeable as compared to myeloid.
- In terms of quality, scDiffCoAM outperforms WGS and PCO in all modules while outperforming IMC and HWH in the majority of modules (two out of three). In terms of quantity, scDiffCoAM outperforms PCO and HWH in each of the three modules and outperforms IMC in the majority of the modules. In two out of three modules, WGS performs better or similarly to scDiffCoAM.
- In all three modules, scDiffCoAM outperforms HWH in terms of quantity, and in two of the three modules, it outperforms PCO and IMC. scDiffCoAM outperforms WGS in the majority of modules, although WGS performs better than scDiffCoAM in one module. In most modules (two out of three) scDiffCoAM outperforms WGS, IMC, and HWH in terms of quality, with the exception of one module where WGS, IMC, and HWH perform better than scDiffCoAM. On the other hand, in one module PCO performs better than scDiffCoAM, and in another module scDiffCoAM outperforms PCO.

6.7.2 Biomarker Ranking

We present a ranking method for all potential biomarkers identified by all our frameworks. By ranking these biomarkers, we aim to identify the genes that are most likely to play a crucial role. Our biomarker criteria discussed in Section 2.5 biologically validates the BCGs as potential biomarkers for ESCC based on fulfillment of minimum requirements. The minimum requirement for a BCG to be identified as a potential biomarker for ESCC are (Biomarker Citeria discussed in Section 2.5): a) at least one literature that establishes the BCG as an ESCC biomarker, b) annotated to at leat one enriched pathway, c) annotated at least one enriched GO term in two out of three GO databases (BP, CC, and MF). Taking this minimum requirement as the basis we add further significance to the identified potential biomarker. To rank all potential biomarkers for ESCC identified by the four frameworks we score them as follows.

(a) For every additional literature evidence found that associates that gene with ESCC, add 1.

(b) If enriched pathway the gene is annotated to a cancer pathway, we add 1 to the score. If the gene is annotated to more than 5 enriched pathways we add 2 while we add 3 when they are annotated to more than 10 enriched pathways.

(c) If the gene is annotated to an enriched GO term in all three GO databases as opposed to minimum requirement of two, we add 1 to the score.

(d) If the gene exhibits regulatory behavior in a GRN,i.e, it is a TF, we add 1 to the score.

(e) A gene detected as a potential biomarker by more than framework are genes very relevant to ESCC as they are significant enough to be detected by multiple analysis that target varying behavior of a gene. If a gene is detected by more than one framework, we add the number of frameworks to the score.

Table 6.13 gives a summary of all potential biomarker rankings. We have not included the genes that has a score < 3. Following are the observations made after ranking the seventy six potential biomarkers for ESCC identified by all four frameworks.

- Two genes PSAT1, and SEL1L only qualify the the minimum requirement with a score of 0 and thus are not considered significant.
- Thirty one genes score 1 as they are annotated to GO enriched terms in all three GO databases (BP, CC, and MF) and as most highly ranked genes with the exception of

FGF5 do fulfil this requirement these genes cannot be considered significant.

Tab. 6.13: Ranking all potential biomarkers for ESCC identified by all four proposed frameworks

Gene	Literature	Cancer Pathways (CPs)	GO databases	TF?	> *Frame*	Score
CAV1	+2 [26, 301, 283]	+1 (1 CP)	+1 (All)	No	+3 (3 FWs)	7
MCM7	+2 [563, 102, 906]	NULL	+1 (All)	+1 (Fig. 4.13b, 5.19a, 6.21a)	+3 (3 FWs)	7
E2F1	+1 [146, 361]	+ 3 (13 CPs)	+1 (All)	+1 (Fig. 3.17b)	No	6
KPNA2	+1 [475, 596]	+1 (1 CP)	+1 (All)	No	+2 (2 FWs)	5
DGKA	[76]	+1 (1 CP)	+1 (All)	No	+2 (2 FWs)	4
EPHA2	+1 [504, 660]	NULL	+1 (All)	No	+2 (2 FWs)	4
EGFR	[505]	+3 (17 CPs)	+1 (All)	No	No	4
HIF1A	+1 [619, 251]	+1 (2 CPs)	+1 (All)	+1 (Fig. 4.16b)	No	4
HSF1	+1 [694, 396]	+1 (1 CPs)	+1 (All)	+1 (Fig. 5.18b)	No	4
NOTCH3	+1 [493, 555]	+1 (2 CPs)	+1 (All)	+1 (Fig. 6.24b)	No	4
SEMA3B	+1 [673, 142]	NULL	+1 (All)	No	+2 (2 FWs)	4
CDH1	+1 [270, 338]	+1 (4 CPs)	+1 (All)	No	No	3
CTTN	+1 [460, 243]	+1 (1 CP)	+1 (All)	No	No	3
ETV5	[655]	+1 (2 CPs)	+1 (All)	+1 (Fig. 6.25b)	No	3
FGF5	[273]	+1 (5 CPs)	(BP,CC)	No	+2 (2 FWs)	3
GSK3B	+1 [52, 182]	+1 (5 CPs)	+1 (All)	No	No	3
G6PD	+1 [745, 746]	+1 (1 CPs)	+1 (All)	No	No	3
HMGA2	[538]	+1 (1 CP)	+1 (All)	+1 (Fig. 3.18b)	No	3
MMP7	+2 [909, 484, 496]	NULL	+1 (All)	No	No	3
PML	[825]	+1 (1 CPs)	+1 (All)	+1 (Fig. 5.20b)	No	3
STAT1	[884]	+1 (1 CPs)	+1 (All)	+1 (Fig. 4.15b)	No	3
TGFA	[384]	+2 (7 CPs)	+1 (All)	No	No	3
VEGFC	+1 [670, 305]	+1 (1 CPs)	+1 (All)	No	No	3

- Two genes CAV1 and the TF MCM7 are the highest potential biomarkers for ESCC detected by three of our frameworks. While most potential biomarkers with high scores such as KPNA2, DGKA, and EPHA2 are identified by more than one framework, the gene E2F1 ranks high as they are annotated to thirteen cancer pathways and exhibits regulatory behavior in a GRN.

- Similarly, the gene EGFR are annotated to seventeen cancer pathways and thus ranks high with a score of 4. On the other hand, HIF1A, HSF1, and NOTCH3 are highly ranked with a score of 4 because they exhibit regulatory behavior, are annotated to cancer pathways and have two literature evidence that associates them with ESCC. SEMA3B also has a score 4 as it is identified by two frameworks as potential biomarker.

6.8 Chapter Summary

The scDiffCoAM framework for differential co-expression analysis (DCA) on single cell RNA-seq data has been proven to be successful in extracting biologically relevant modules as well as discovering interesting hub-genes. We validated our framework, scDiffCoAM, on the scRNA-Seq ESCC dataset, GSE160269, which includes eight cell types. DCA has been performed by the framework on six of the eight cell types, three immune (CD45+), and three non-immune (CD45-). It can extract nineteen biologically significant modules, i.e., 'modules of interest' (MoI). The proposed framework is proven to be efficient in identifying potential biomarkers after further investigation of these nineteen MoIs. On scRNA-Seq data, the hub-gene finding method described in CBDCEM by Saikia et al. [592] is found to be effective when the choice of the seven measures is made based on the network properties. Twenty-four hub-genes have been identified to be potential biomarkers for ESCC by scDiffCoAM with strong evidence of association.

In most cases, scDiffCoAM performs better than or similarly to the four other hub-gene finding methods, which include weighted gene score [120], p-value cutoff [120], WGCNA [327] intra-modular connectivity, and hdWGCNA [510, 509] kME score. Furthermore, the framework can identify forty-four distinct potential biomarkers that none of the other four methods that were considered could. Fourteen of the twenty-four potential biomarkers found and verified by scDiffCoAM were among these forty-four hub-genes. Ten of the remaining twenty-four hub-genes that were identified and verified by scDiffCoAM as potential biomarkers are also detected by the other four methods in different modules across cell types.

Next chapter is the final chapter of the thesis that summarizes the concluding remarks for all four contributions of our work. Furthermore, we summarize few of the shortcomings observed in each framework and further suggest future directions for improvements of the same.