Chapter 4

Identifying Crucial Genes for ESCC using Differential Expression Analysis

4.1 Introduction

During Esophageal Cancer (EC), healthy cells start growing uncontrollably along the surface of the esophagus. Esophageal Squamous Cell Carcinoma (ESCC) and Adenoarcinoma (EAC) are the two main types of EC. ESCC develops from regular squamous cells, running along the surface of the esophagus. ESCC is aggressive and the most prevalent type of EC. It is ranked as the sixth leading cause of cancer death and found all over the world and especially in China and India [214]. It is reported that the survival rate of ESCC patient is declining from 20% to 4% in the advanced cases[215]. For all stages combined, survival is lowest for cancers of the esophagus (19%)[216]. Unfortunately, it has one of the lowest survival rates, since esophageal cancer is rarely diagnosed early. It occurs due to cigarette smoking and alcohol drinking [217] and this disease also has a very high mortality rate. For this cause, doctors and researchers are constantly seeking better methods of detection and diagnosis, as well as more effective treatments. Late diagnosis and few therapeutic options increase the mortality rate. Therefore, the identification of crucial genes which might be biomarkers for ESCC is very important. It also helps in the early detection of ESCC and development of new drugs target. The information carried out by a gene in terms of gene expression is used in the synthesis of a functional gene product. To investigate the progression of the disease through gene expression profile in ESCC disease, analysis of sequence and microarray data is the most important application of bioinformatics. Due to the low cost, sequencing technology has been rapidly growing, thus count data is also generating at a high speed. Differential Gene Expression Analysis (DEA) helps quantify the statistically significant change of gene expressions between two experimental conditions and Differentially Expressed Genes (DEGs) have biological significances. The RNA-Seq technology is found very useful for differential expression analysis.

Analysing of sequencing data is difficult comparative to microarray data due to its huge dimension. There are various techniques available which the help in analysing the count data. Searching of Differentially Expressed Genes (DEGs) from those large number of genes is a challenging task and again it also challenging to find the most significant genes which participate in causing particular deadly disease. Though a good number of useful Differential Expression Analysis (DEA) tools have been developed to identify DEGs, however, none of these tools can be considered effective for all cases. Hence, an ensemble approach has been chosen to help to improve the performance of significant DEGs identification. In addition, it was possible to obtain a large number of data sets related to such deadly diseases due to the growing advancement of sequencing technology. Every type of dataset, however, has its own specificity and limitations. Therefore, it may not be justified to experiment with a single type of dataset towards conclusive identification of a number of responsible genes for a given disease. In order to identify an unbiased set of biomarkers for a given disease, it is essential to conduct an integrative study using data sets generated by different technologies supported by an effective consensus function. Critical genes for ESCC using RNA-seq data have not been explored much till date.

In this chapter, a systematic analysis is performed based on publicly available ESCC microarray and RNA-seq data to identify interesting genes in ESCC. Significant differential behaviour of genes between ESCC and normal samples are analyzed with DEA tools and DEGs are identified. Topological behaviour is studied across the conditions followed by biological validation and existing literature evidences.

4.2 Background

DEA with DESeq2 [39] package involves several steps. DESeq2 [39] is based on the DESeq algorithm, which is a powerful tool for identifying DEGss. The DeSeq algorithm uses a statistical model to identify genes that are differentially expressed between two or more conditions. DESeq2 [39] models the raw count using normalization factors (size factors) to account for library depth variations. Further, it estimates the gene-wise dispersions which help to shrink these estimates resulting in more accurate estimates of dispersion to model the counts. Finally, the testing of hypothesis using the Wald test or Likelihood ratio test is performed by DESeq2 with the help of the negative binomial model. GCNs are constructed by using the correlation matrix where genes represent nodes and edges represent computed pairwise correlation between pairs of co-expressed genes. It is a systems biology method for describing the correlation patterns among genes across microarray or RNAseq samples. Most of the tools use pearson correlation to detect co-expression between samples. It identifies group of tightly correlated genes associated with biological processes. GRNs define how the genes are connected to interpret the biological insights among them[218]. For this network, transcription factors and target genes are needed. Transcription factors act as regulators and target genes are the remaining genes from the lists. The transcription factor has the different characteristics and it is identified based on the ability to bind to DNA and to recruit RNA polymerase/alter transcription of a gene[218].A network of protein-protein interactions (PPIs) is a platform which systematically recognises disease-related genes from the associations between proteins with similar functions. In analysis of gene enrichment, biological importance of a set of genes are examined and assigned functions and roles in biological processes from the previously studied records.

4.3 Proposed Methods

This section presents two approaches for crucial gene identification for ESCC and their validations.

4.3.1 Method I: A Generic DEA Approach:

RNA-seq data (SRA: SRP008496, GEO: GSE32424) used in this paper is downloaded from Recount2 https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP008496. It contains 12 clinical samples from ESCC from homosapiens and 58,037 genes. Among 12 clinical samples, seven samples are tumors (SRR349741, SRR349742, SRR349743, SRR349744, SRR349745, SRR349746, and SRR349747) and other five samples are non-tumors (SRR349748, SRR349749, SRR349750, SRR349751, and SRR349752). DEGs helps to find out which genes are crucial in the progression of ESCC. R [219] software platform is used for the downstream analysis of the dataset. DESeq2 [39] is a popular and widely used differential expression analysis tool available in R [219] which is used to identify the DEGs.

The conceptual framework for the identification of crucial genes of ESCC in progression RNA-seq data is shown in Figure 4-1. The raw count data (GSE32424) is pre-processed and analyzed by R language software. The differentially expressed genes are extracted using DESeq2 [39] from the processed data to obtain the up and down regulated genes. Up-regulated genes have the log fold change value greater than 0 and down-regulated genes have the log fold change value lower than 0. The top 10 Up-regulated and top 10 down-regulated DEGs are identified based on their *log fold change* value, say this is the *list1*, presented in Table 4.1-4.2. Fold change is a parameter for measuring change in the expression level of a gene during analysis of gene expression data. The DEGs with highest *log fold change* might not have the lowest adjusted P-value. The adjusted P-value can be defined as the smallest familywise significance at which a specific comparison is considered statistically significant as part of multiple comparative tests. A set of significant DEGs are identified with reference to an adjusted P-value cut-off lower than 0.05. Top 100 DEGs are selected from this set of genes, say *list2* shown in Table 4.3 and examined how many genes are mapped from *list1* with higher significance level. My aim is to consider the most significant genes based on adjusted P-value (*list2*) as well as *log fold change* value. Hence, the common genes from *list1* and *list2* are found out. Among the 20 genes from *list1*, 8 genes are considered because they are more significant from my research point of view i.e. these 8 genes are found common to both the lists. The rest of the genes are neglected because remaining genes from *list1* are randomly scattered in *list2* and there is a difference of significance level of selected 8 DEGs from the remaining genes. These eight genes are considered as the initial set of crucial genes and used for the downstream analysis. Co-expression, gene regulatory, and PPI networks are constructed with these initial set of genes and finally, gene enrichment and pathway analysis are performed to validate the set of suspected genes. The final set of genes are further studied in the existing literature to establish the roles of these genes in the progression of ESCC.

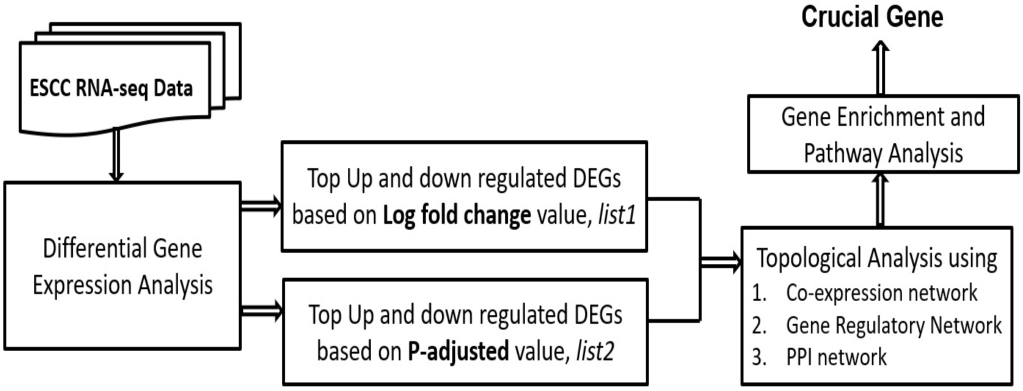

ESCC RNA-seq Data
Differential Gene Expression Analysis
Top Up and down regulated DEGs based on Log fold change value, list1
Top Up and down regulated DEGs based on P-adjusted value, list2
Topological Analysis using
1. Co-expression network
2. Gene Regulatory Network
3. PPI network
Gene Enrichment and Pathway Analysis
Crucial Gene

Figure 4-1: Conceptual Framework for identification of crucial genes for ESCC RNA-seq data

4.3.1.1 Experimental Results

Pre-filtering of GSE32424: The low counts are removed by considering the genes with at most 1 zeroes. Version numbers from raw GTEx (GENCODE) Gene IDs (Ensembl) are removed using ”cleanid” function available in R. Ensembl IDs are again mapped with Symbol IDs. Occurrences of ”NA” from the dataset are deleted and finally, a total of 12,903 transcripts are extracted. Dataset is not normalized since DESeq2 takes input in terms of raw value (data type integer). DEseq2[39] performs normalization during the execution process.

A. Identification of DEGs

DESeq2[39] is applied in the processed dataset with 12,903 transcripts and 12 samples to identify the differentially expressed genes between normal and ESCC samples. In this analysis, 3,474 (27%) up-regulated (log2FoldChange>0) and 3,906 (30%) down-regulated genes (log2FoldChange<0) are identified. Most recent studies consider cut off criteria for identifying DEGs at padj<.05 [220]. Also, 3,087 number of up-regulated and 3,439 down-regulated DEGs are extracted with reference to this threshold. Adjusted p-values are used to control the FDR, ensuring that the proportion of falsely identified significant genes is limited. A threshold of 0.05 is a conventional choice to strike a balance between sensitivity and specificity. The threshold of 0.05 is a widely accepted standard in scientific research. It provides a common ground for comparing results across different studies and allows for easier interpretation and communication of findings. It helps researchers identify genes that are statistically significantly differentially expressed while minimizing the likelihood of including too many false positives. Over time, the use of a 0.05 threshold has become a convention in statistical hypothesis testing. Consistency in the choice of thresholds facilitates reproducibility and comparisons across studies.

Table 4.1: Top 10 Up-regulated DEGs, ranked by *log2FoldChange* value

Serial No.	**Gene Name**
1	MAL
2	KRT4
3	KRT78
4	CLCA4
5	FAM25A
6	CAPN14
7	**FMO2**
8	**PRSS27**
9	CNFN
10	SPRR3

Table 4.2: Top 10 Down-regulated DEGs, ranked by *log2FoldChange* value

Serial No.	**Gene Name**
1	**FN1**
2	**COL1A1**
3	SPP1
4	**TNC**
5	**COL12A1**
6	**POSTN**
7	**VCAN**
8	LOC101927136
9	TRIM74
10	SPATA13

Here, top 10 DEGs (up-regulated and down-regulated) are extracted based on higher *logfoldchange* value (positive and negative) shown in Table 4.1- 4.2. Again, top 100 genes shown in Table 4.3 are also selected from DEGs list with a cutoff *padj* value lower than 0.05. The top 10 up and down-regulated genes (Table 4.1- 4.2) are matched with the DEGs of top 100 (Table 4.3) and common genes are marked in bold. Total eight common genes are identified such as FMO2, PRSS27, FN1, COL1A1, TNC, COL12A1, POSTN, and VCAN.

Table 4.3: Top 100 Up and Down-regulated DEGs, ranked by *padj* value, Bold gene: common gene

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
THY1	4829.780025	-6.346169979	0.345707208	-18.35706585	2.90E-75	1.98E-71
COL1A1	79691.08739	-8.413805944	0.458422037	-18.35384267	3.08E-75	1.98E-71
LAMC2	37275.82834	-5.542758525	0.308612186	-17.96027112	3.99E-72	1.71E-68
ADAMTS2	4920.146397	-6.061987268	0.341390392	-17.75676005	1.53E-70	4.92E-67
FN1	181649.5053	-8.222104419	0.474595349	-17.32445215	3.08E-67	7.93E-64
PIM1	42755.71584	3.854282258	0.22276014	17.30238745	4.51E-67	9.69E-64
COL12A1	47856.70431	-6.672382153	0.400244028	-16.67078506	2.14E-62	3.94E-59
COL5A2	18280.45918	-6.429870927	0.390642863	-16.4597169	7.14E-61	1.15E-57
ST3GAL4	8995.995796	3.954588218	0.241013217	16.40817989	1.67E-60	2.39E-57
CXCR2	4920.643462	6.072532962	0.376603248	16.12448376	1.72E-58	2.21E-55
RMND5B	13440.99001	3.456259672	0.219586365	15.7398647	8.06E-56	9.45E-53
POSTN	40952.11114	-7.747065835	0.497745713	-15.56430449	1.27E-54	1.37E-51
ADAM12	4826.287217	-6.15294879	0.397145909	-15.49291746	3.87E-54	3.84E-51
TRIP10	16056.88274	3.986371988	0.261954274	15.21781616	2.69E-52	2.48E-49
LUM	45988.43525	-4.951477482	0.329347359	-15.0342104	4.38E-51	3.77E-48
LAMA3	15311.98511	-3.625304579	0.242251039	-14.96507338	1.24E-50	1.00E-47
FMO2	29124.0302	7.062021632	0.47636551	14.82479627	1.01E-49	7.68E-47
LEXM	3256.659762	6.18141436	0.419963727	14.71892443	4.87E-49	3.49E-46
TMEM40	31612.7687	5.152150988	0.353738479	14.56485875	4.70E-48	3.19E-45
C6orf132	22551.56257	3.890906817	0.267420679	14.54976045	5.86E-48	3.78E-45
SPARC	118600.3026	-4.841257552	0.336428876	-14.3901368	5.97E-47	3.66E-44
RNF222	3565.500282	5.633596299	0.393950789	14.3002539	2.18E-46	1.28E-43
SERPINB1	123343.2718	4.770905492	0.334812385	14.24948926	4.52E-46	2.53E-43
COL4A1	14627.27515	-4.557388882	0.320552572	-14.21729003	7.16E-46	3.84E-43
RHCG	709479.5522	6.434233656	0.455121165	14.13740813	2.23E-45	1.11E-42
CSTB	1372242.639	5.511962738	0.389883505	14.13746073	2.23E-45	1.11E-42
TNC	64376.6442	-6.463542179	0.458039345	-14.11132525	3.23E-45	1.54E-42
TCP11L2	25709.84086	4.013446948	0.284902751	14.08707684	4.56E-45	2.10E-42
PRSS27	37234.17027	6.718938671	0.478985004	14.02745099	1.06E-44	4.55E-42
UBL3	36513.15223	4.229885666	0.301528583	14.02814164	1.05E-44	4.55E-42
COL4A2	20274.78159	-4.316236534	0.309005665	-13.96814694	2.44E-44	1.01E-41
CD276	3158.72067	-3.565675805	0.256654281	-13.89291384	6.99E-44	2.82E-41
COL5A1	11586.49807	-5.912263015	0.428061337	-13.81171927	2.17E-43	8.46E-41
ALDH9A1	19624.9843	2.965937691	0.215162691	13.78462816	3.15E-43	1.20E-40
VCAN	34011.33394	-6.485585193	0.473516046	-13.69665347	1.06E-42	3.92E-40
AIF1L	21917.32824	5.36849827	0.39481115	13.59763589	4.14E-42	1.48E-39
GRHL1	32056.55097	3.198778742	0.235710599	13.57078871	5.97E-42	2.08E-39
TMEM2	7160.360127	-2.507931141	0.184932304	-13.56134696	6.79E-42	2.30E-39
MXD1	69316.50753	3.966933834	0.295599565	13.41995828	4.62E-41	1.53E-38
KRT13	3444280.396	6.265108333	0.467551522	13.39982447	6.06E-41	1.95E-38
KAT2B	17491.4411	4.055043137	0.303202256	13.3740533	8.57E-41	2.70E-38
DPYSL3	17228.92587	-4.344984457	0.328673845	-13.21974511	6.75E-40	2.07E-37
GMDS	10089.58508	4.122008091	0.312083659	13.20802284	7.89E-40	2.36E-37
LAMC1	10221.65459	-3.775912692	0.286258235	-13.19058189	9.94E-40	2.91E-37
BICDL2	7343.039247	4.441014816	0.336907252	13.18171332	1.12E-39	3.20E-37
PTK6	30586.50244	4.762290492	0.361906168	13.15890945	1.51E-39	4.24E-37
PITX1	91905.99949	4.271973924	0.32961395	12.96053739	2.05E-38	5.62E-36
EHD3	22683.50204	3.836940993	0.297148233	12.91254859	3.82E-38	1.03E-35
CALU	11381.19778	-3.135307865	0.243172743	-12.89333592	4.91E-38	1.29E-35
CALD1	25305.63373	-3.577924467	0.277775088	-12.88065281	5.78E-38	1.49E-35
RANBP9	19274.4568	2.990187779	0.232873675	12.84038556	9.74E-38	2.46E-35
TMPRSS11E	118884.2098	5.929667306	0.461934614	12.83659445	1.02E-37	2.54E-35

Continued on next page

Table 4.3 – *Continued from previous page*

Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CDH11	6210.810544	-5.927628626	0.461864286	-12.83413506	1.06E-37	2.57E-35
ACTA2.AS1	6173.680836	-4.850089176	0.381201425	-12.72316644	4.40E-37	1.05E-34
SPINT1	29984.9069	3.555776328	0.279529745	12.72056511	4.55E-37	1.07E-34
GALE	8795.210953	3.443019304	0.270770858	12.7156199	4.84E-37	1.11E-34
AGFG2	14294.89654	4.333237683	0.342744779	12.6427533	1.23E-36	2.77E-34
FAM129B	95382.00311	3.37513708	0.267166487	12.63308552	1.39E-36	3.08E-34
BGN	7441.279285	-4.533584372	0.359019456	-12.62768436	1.49E-36	3.25E-34
AQP3	232059.326	6.040920179	0.480490879	12.57239304	3.00E-36	6.43E-34
MFSD5	9545.879161	2.265874497	0.18045964	12.55612888	3.68E-36	7.77E-34
ESPL1	5890.654444	3.345698311	0.267073883	12.52723883	5.30E-36	1.10E-33
HOPX	176753.4791	5.933176055	0.473852517	12.52114496	5.72E-36	1.17E-33
SPAG17	4679.285588	4.428817534	0.353773302	12.5188009	5.89E-36	1.19E-33
IL1RN	283189.9994	5.623433999	0.449487646	12.51076432	6.52E-36	1.29E-33
THBS1	33937.76103	-4.687751038	0.375422515	-12.4866007	8.83E-36	1.72E-33
ACPP	13258.13474	3.563615188	0.285412904	12.48582363	8.92E-36	1.72E-33
TGFBI	34296.5564	-4.134967061	0.332247269	-12.44545086	1.48E-35	2.81E-33
KRT7	166890.2056	8.313731856	0.671314957	12.3842494	3.18E-35	5.94E-33
GRPEL2.AS1	1660.483921	4.748874436	0.384009207	12.36656399	3.96E-35	7.30E-33
WDR26	28894.85354	2.48702274	0.202212381	12.29906266	9.16E-35	1.66E-32
CARHSP1	16441.46779	3.020013399	0.246829055	12.23524274	2.02E-34	3.61E-32
PDLIM3	2436.469936	-4.887648845	0.402268773	-12.15020695	5.72E-34	1.01E-31
LOC440434	3491.220659	2.52137897	0.207738486	12.13727421	6.70E-34	1.17E-31
N4BP3	3251.494204	4.048901462	0.334171113	12.11625216	8.66E-34	1.49E-31
ACADM	12716.78695	2.848985866	0.235192358	12.11342873	8.97E-34	1.52E-31
NCCRP1	68372.79468	6.268609098	0.51855541	12.08860033	1.21E-33	2.03E-31
COL6A3	37965.73171	-5.059818655	0.418759069	-12.08288735	1.30E-33	2.15E-31
EPHA2	22450.99902	3.342511137	0.276909142	12.07078651	1.51E-33	2.46E-31
ABLIM3	11188.47785	4.942711439	0.410193382	12.04971035	1.95E-33	3.14E-31
TAGLN	16206.02829	-4.786646146	0.397711983	-12.03545869	2.31E-33	3.68E-31
TMPRSS2	10933.52417	5.820889372	0.485062485	12.0002877	3.54E-33	5.57E-31
STN1	7850.852744	3.591563246	0.299431636	11.99460182	3.79E-33	5.89E-31
SEMA3C	3735.783579	-5.541321276	0.462375765	-11.98445442	4.29E-33	6.58E-31
SESN2	8749.522319	3.918387959	0.327071806	11.98020704	4.51E-33	6.84E-31
MMP2	14805.05584	-5.089611929	0.425399971	-11.96429777	5.47E-33	8.19E-31
SDCBP2	4725.015159	3.70781534	0.310018836	11.95996794	5.76E-33	8.53E-31
LYPD3	63034.79089	3.599396791	0.301606025	11.93410109	7.86E-33	1.15E-30
PMM1	9737.844398	3.5076909	0.295355445	11.87616803	1.57E-32	2.28E-30
FBN1	14840.78056	-4.957674005	0.41792637	-11.86255369	1.85E-32	2.65E-30
WDR66	2551.195043	-4.142504875	0.350739228	-11.8107829	3.43E-32	4.86E-30
SLC35C1	5801.432072	3.263112906	0.277059328	11.77766844	5.09E-32	7.13E-30
PDGFRB	7465.757153	-3.637579521	0.310365774	-11.72029852	1.00E-31	1.39E-29
GPR157	8360.437935	3.02043467	0.258529819	11.68311912	1.55E-31	2.13E-29
FKBP10	5449.581025	-5.418447488	0.46403527	-11.67680096	1.67E-31	2.27E-29
IL18	27793.41571	4.875775621	0.417985018	11.66495307	1.93E-31	2.58E-29
DDR2	4394.368264	-4.0064582	0.34450708	-11.62953808	2.92E-31	3.88E-29
DUSP5	39611.84658	5.126600523	0.441421291	11.61384969	3.50E-31	4.61E-29
ANKRD22	5112.400757	3.00641139	0.259231553	11.5973976	4.25E-31	5.53E-29
GIPC1	41835.05319	2.911257142	0.25107735	11.59506081	4.37E-31	5.63E-29

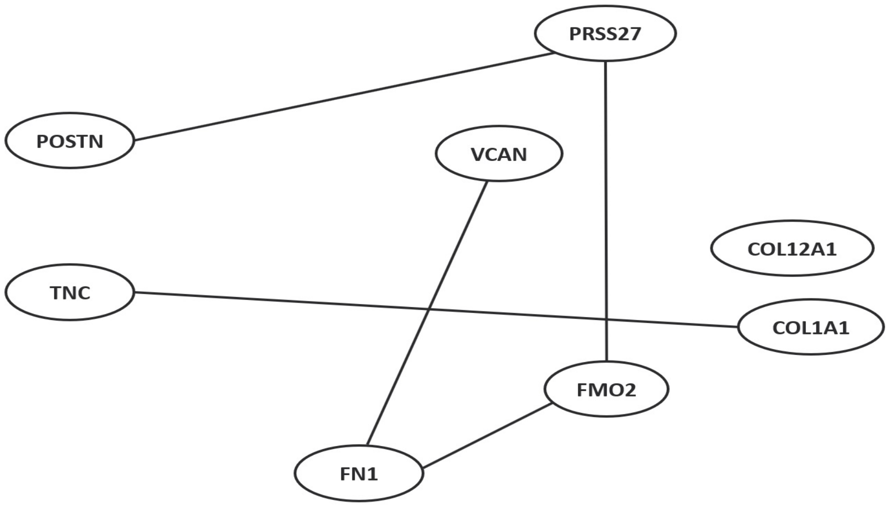


Figure 4-2: Co-expression network of the suspected eight genes across the states (i.e. Normal and disease). The total number of edge connectivity is 5.

B. Construction of Co-expression Network

In this study, two different co-expression networks are formed for normal and disease conditions as shown in Figure 4-2- 4-3. For these networks, only eight suspected DEGs are considered. In these networks, the genes are connected if the Pearsons correlation coefficient is greater than 0.01. Here, considered Pearsons correlation coefficient is greater than 0.01 because the size of our adjacency matrix is very small i.e. 8x8 and all the values with Pearsons correlation coefficient lower than and equal to 0.01 are 0. By comparing these two networks, it is observed that the connectivity of genes with their immediate neighbours in normal condition are lesser (edge connectivity is 5) than the connectivity of genes with their immediate neighbours in disease condition (edge connectivity is 13). When genes transmit from normal to disease conditions, their edge connectivity with neighbours increases and topological behaviour (network connectivities) are significantly different from each other across conditions.

C. Construction of Gene Regulatory Network (GRN)

Transcription factors among 52 genes are identified with TFcheckpoint[1] tool. Among the 52 genes (Genes from Table 4.1- 4.3), only three genes such as PITX1, GRHL1, and MXD1 are found as transcription factors. To analyze the different behaviour of GRN in normal and disease conditions, only edge connectivity among regulators (transcriptions factors) and target gene are shown

[1] http://www.tfcheckpoint.org/index.php/search

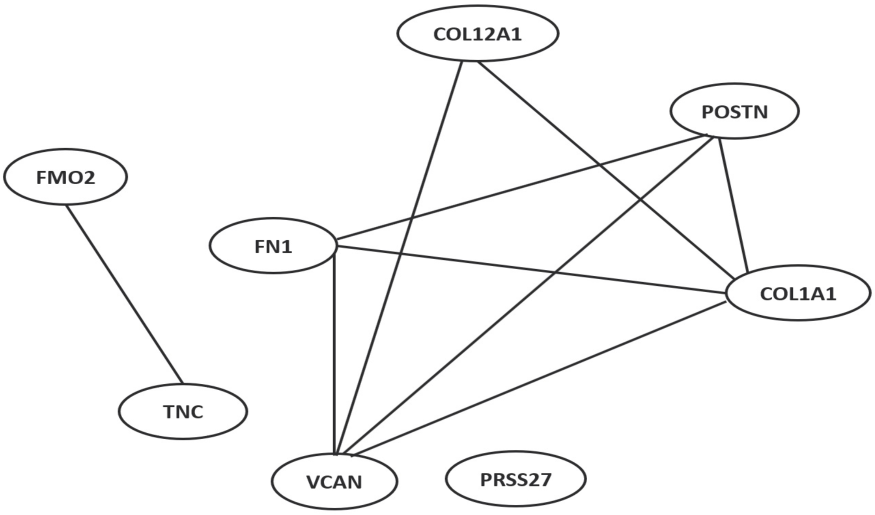


Figure 4-3: Co-expression network of the suspected eight genes across the disease condition. The total number of edge connectivity is 13.

in Table 4.4- 4.5. GENIE3[221] package available in R platform is used to construct GRN.

Table 4.4: Topological statistics of GRN among 3 regulatory gene and 8 suspected genes in normal state

RegulatoryGene	**TargetGene**	**Weight**
PITX1	TNC	0.718599097
GRHL1	COL12A1	0.530555566
GRHL1	POSTN	0.485545356
MXD1	FMO2	0.484461331
MXD1	FN1	0.469279134
MXD1	PRSS27	0.426127752
PITX1	POSTN	0.41000477
GRHL1	FMO2	0.409450798
MXD1	COL12A1	0.396999557
GRHL1	PRSS27	0.396835715
GRHL1	VCAN	0.395777311
GRHL1	COL1A1	0.380200849
MXD1	VCAN	0.353484894
GRHL1	FN1	0.339335059
MXD1	COL1A1	0.332040763
PITX1	COL1A1	0.287758388
PITX1	VCAN	0.250737794
PITX1	FN1	0.191385808
PITX1	PRSS27	0.177036534
GRHL1	TNC	0.145307643
MXD1	TNC	0.13609326
PITX1	FMO2	0.106087871
MXD1	POSTN	0.104449874
PITX1	COL12A1	0.072444878

Table 4.5: Topological statistics of GRN among 3 regulatory gene and 8 suspected genes in disease state

RegulatoryGene	TargetGene	Weight
PITX1	FN1	0.514260159
MXD1	TNC	0.511935998
PITX1	COL1A1	0.458977461
PITX1	FMO2	0.396016814
GRHL1	COL12A1	0.381068216
GRHL1	PRSS27	0.378760255
PITX1	COL12A1	0.37851948
GRHL1	POSTN	0.378347311
PITX1	VCAN	0.371907687
MXD1	POSTN	0.343706288
MXD1	PRSS27	0.320786888
GRHL1	VCAN	0.316487829
MXD1	FMO2	0.314507509
MXD1	VCAN	0.311604484
PITX1	PRSS27	0.300452857
MXD1	COL1A1	0.29652297
GRHL1	FMO2	0.289475676
GRHL1	TNC	0.285323713
PITX1	POSTN	0.277946401
GRHL1	FN1	0.250241668
GRHL1	COL1A1	0.244499569
MXD1	COL12A1	0.240412304
MXD1	FN1	0.235498173
PITX1	TNC	0.202740288

D. Construction of PPI Network

The STRING tool builds an interaction network of the DEGs FMO2, PRSS27, FN1, COL1A1, TNC, COL12A1, POSTN, and VCAN. This network is formed based on the evidences of known interactions among genes from curated databases, text mining, experiments, co-expression, neighbourhood, gene fusion, and co-occurrence with a confidence score of 0.5. Figure 4-4 describes the interaction among the eight DEGs. There are several hub proteins (FN1, POSTN, VCAN, COL1A1) in the network, which are proteins that interact with many other proteins. These hub proteins may play an important role in the development of ESCC. By adjusting the confidence score from 0.9 to 0.1, it has been observed that there is no association between the differentially expressed genes (DEGs) FMO2 and PRSS27 in the PPI network. The use of a threshold of 0.5 is a common practice in PPI network analysis. The threshold of 0.5 is used to filter out weak interactions from the network. This is because weak interactions are more likely to be noise or false positives.

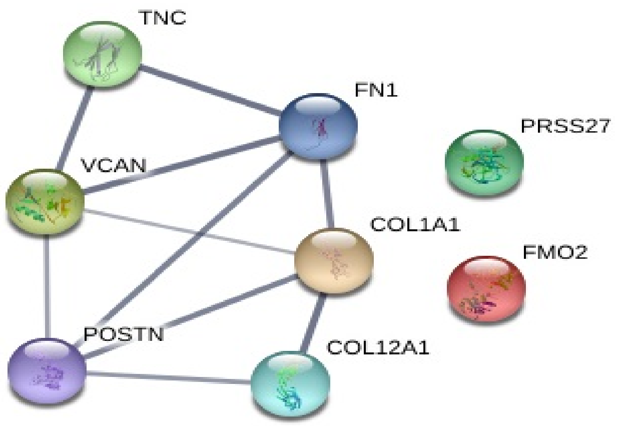


Figure 4-4: PPI network of the suspected eight genes

E. Gene Enrichment Analysis of Suspected Genes

The eight DEGs are imported into DAVID[2] to reveal the enrichment analysis of GO terms and KEGG pathway. Total 8 genes such as FMO2, PRSS27, FN1, COL1A1, TNC, COL12A1, POSTN, and VCAN are enriched in biological process, the cellular process and molecular functions. The enriched GO terms of these 8 DEGs are GO:0043062 extracellular structure organization (TNC, COL12A1, POSTN, and COL1A1), GO:0030198 extracellular matrix organization (COL12A1, POSTN, and COL1A1), GO:0007155 cell adhesion (TNC, COL12A1, POSTN, and FN1), GO:0022610 biological adhesion (TNC, COL12A1, POSTN, and FN1), GO:0001501 skeletal system development (COL12A1, POSTN, and COL1A1), and GO:0030199 collagen fibril organization (COL12A1 and COL1A1).

After the KEGG pathway analysis of the 8 DEGs, two pathways are found enriched such as hsa04512: ECM-receptor interaction (Shared by TNC, COL1A1, and FN1) and hsa04510: Focal adhesion (Shared by TNC, COL1A1, and FN1). The cutoff for adjusted p-value was set as less than 0.01 for the significantly enriched biological processes and the KEGG pathway analysis, adjusted p-value cut off is set as less than 0.05.

F. Discussion

Here, FN1, COL1A1, and TNC among 8 DEGs are identified as suspected genes for ESCC, because these genes are significantly differentially

[2] https://david-d.ncifcrf.gov/

expressed based on *adjusted P-value,* and *log fold change* values between normal and disease conditions. Further, these genes are found highly enriched in terms of GO terms and KEGG pathway analysis. From the observation (Figure 4-2- 4-3), it is seen that COL1A1 is only associated with TNC in normal condition but in disease condition, the degree of COL1A1 is 4. The degree of FN1 and TNC in normal condition are 2 and 1 and in disease condition, their degrees are 3 and 1 respectively. Again, from Table 4.4- 4.5, it is observed that PITX1 regulates TNC with the highest weight rank in normal condition but in tumor they are connected with the lowest weight rank. In tumor, the top 3 genes which are regulated by the transcription factors with the highest weight rank are FN1, TNC, and COL1A1. The PPI network also depicts the experimentally known interaction from several evidences among these genes, TNC, FN1, and COL1A1 or directly and indirectly connected with each other. Even when the confidence score is at its maximum in PPI network, i.e., 0.9, there is still a direct connection between COL1A1 and FN1, and between COL1A1 and POSTN. The strength of the interactions between the genes varies, with some genes having more interactions than others. This suggests that some genes may be more important than others in the development of ESCC. FN1, the metastasis marker of ESCC, was observed in the marginal cells of ESCC and was strongly expressed in the cytoplasm of the tumor cells[222]. FN1 was found overexpressed in ESCC and it activates ERK pathway which was experimented by Western blot test and RT-PCR analysis[222]. By up-regulation of FN1 and PDGFRB, SATB1 performs an oncogenic role in ESCC[223]. COL1A1 is reported as a crucial gene for ESCC[224]. TNC, an extracellular matrix protein, is associated with a poor prognosis of ESCC[225]. Given that the outcomes of this research align with prior studies highlighting the role of these genes in cancer progression and metastasis, our approach has proven successful in identifying crucial genes for esophageal squamous cell carcinoma (ESCC). It suggests that these genes could be potential biomarkers for early detection and development of new drugs for ESCC. Further research on these genes could provide insights into the underlying mechanisms of ESCC and lead to the development of new treatments for ESCC and the improvement of patient outcomes.

4.3.2 Method II: An Ensemble Approach

In this work, initially an independent differential expression analysis is conducted on both microarray and RNA-seq gene expression data to identify a set of significant DE genes for each type of data. A gene is considered as Differentially

Expressed (DE) if the observed difference or change in read counts or expression levels between two experimental conditions is statistically significant [48]. The significant changes of expression values for the corresponding gene(s) in normal and disease states signifies the hidden truth behind the selection of some interesting genes for subsequent downstream analysis. Recently, a good number of tools have been introduced for DE gene identification using microarray or RNA-seq data. However, my observations are: (i) tools developed for microarray data often do not work for RNA-seq data and (ii) a significant variation in terms of a number of genes exists between any pair of such tools. So, to identify DE genes, multiple DEA tools are used for each dataset type such as DESeq2 [39], edgeR [37], limma-voom [44], limma [46], SAM [226], EBAM [227] and use an appropriate consensus function towards generation of unbiased set of differentially expressed genes. Based on the selected DE genes, a co-expression network is constructed using WGCNA (a freely available R tool) [78] followed by a preservation analysis has been carried out to identify a set of low preserved modules across the states. The module preservation statistics quantify the preservation of within-module between a reference network and a test network [99]. A WGCNA R package, *Zsummary* statistics score is calculated to determine the module which is highly/lowly/moderately preserved in a normal state but not in disease state. Zsummary statistics score for the highly preserved modules is greater than 10 and also their Medianrank score is relatively low [103]; for the moderately preserved modules *Zsummary* score is between 2 and 10, and for lowly preserved *Zsummary* score is below 2. The 'hub' is the node with the highest degree, plays an important role in molecular mechanisms which may be highly significant in determining the outcome or phenotype of a disease of interest. Intra-modular connectivity of low preserved modules is used to find out the hub genes. Additionally, in STRING[3] tool, PPI networks are built for each low preserved module and hub genes are identified for the same. The genes belonging to the low preserved modules have been carefully investigated and assessed using based on both hub gene-centric and pathway enrichment analysis tools such as GeneAnalytics (https://ga.genecards.org) and David (https://david.ncifcrf.gov/). This experimental analysis reveals that SOX11, COL27A1, TOP3A, BAG6, CDC6, EZH2, COL7A1, G6PD, and AKR1C2 are a few prominent hub genes found responsible for ESCC.

This section presents the descriptions of both the microarray and RNA-seq data, their pre-processing and analysis, and validation using both topological and biological approaches.

[3] https://string-db.org/

4.3.2.1 Dataset description and pre-processing

For this experiment, we consider three ESCC dataset - SRP064894, GSE20347, and GSE23400 reported in Table 4.6. It is essential to remove genes with very low read counts in all samples before using DEA tools. The low read count instances are discrded using a user-defined threshold (here, it is 5). DESeq2, edgeR, limma-voom calculate individual library size factor by which expression values are normalized. For network construction, Trimmed Mean of M-value (TMM) normalization method available in edgeR and CPM function are used to obtain the normalized expression values of the RNA-seq dataset. TMM method is used for normalizing the read count before DEA. GSE20347 dataset was already normalized across all samples by the Robust Multiarray Average (RMA) algorithm implemented in Bioconductor in R (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20347). The TMM normalization function is used to normalize the dataset GSE23400. For all the three datasets, *goodSamplesGenes()* function is used which is available in WGCNA to investigate missing entries and zero-variance genes.

Table 4.6: Dataset Used. Note: T: Tumor, N:Normal

Dataset	Type	Link	Size	Sample(T)	Sample(N)
SRP064894	RNA-seq	Recount2[4]	58000x29	14	15
GSE20347	Microarray	GEO[5]	22278x34	17	17
GSE23400	Microarray	GEO[6]	22287x106	53	53

4.3.2.2 Proposed framework

In this study, microarray and RNA-seq datasets are used individually for the extraction of significant DEGs. More than one DEA tools are used individually for each dataset for the identification of unbiased DEGs by an appropriate consensus function and extracted final set of DEGs are considered as the input for the condition-specific co-expression analysis using WGCNA [78] in R. Module preservation analysis is conducted with the modules obtained from WGCNA co-expression analysis and some low preserved modules are detected for each datasets. After module preservation, hub genes are identified for each low preserved modules using intra-modular connectivity method in WGCNA. Additionally, PPI-network is constructed in STRING tool for the low preserved modules and hub genes are detected based on the highest connectivity. Identified low preserved modules and hub genes are further investigated topologically and biologically. Existing literature evidence is also considered while investigating the hub genes. These hub genes are termed as critical genes since they are detected from the experimental analysis and identified to be associated with ESCC disease by sufficient evidences. The conceptual framework of my method is shown in Figure 4-5.

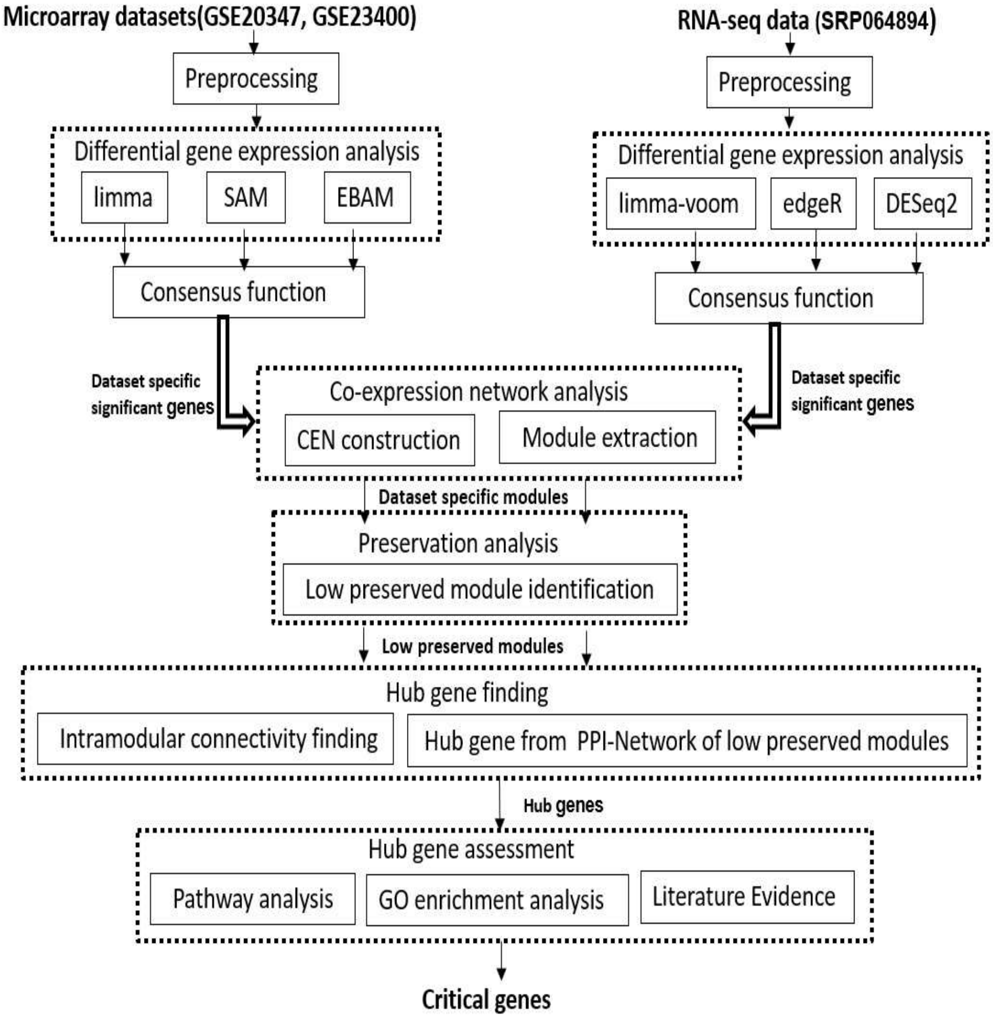


Figure 4-5: Conceptual framework to find critical genes from ESCC datasets.

In this work, an ensemble approach is used supported by an effective consensus function to find an unbiased set of DEGs by multiple DEG finding tools. These DEGs are considered as input for GCN construction and for subsequent downstream analysis. GCN analysis is carried out to extract modules from the co-expression network to perform module preservation analysis. The low preserved modules which are identified by module preservation analysis are further investigated using topological, pathway, GO enrichment analysis and in light of relevant literature evidence to identify an interesting set of biomarkers for the microarray data and RNA-seq data. Finally, interesting biomarkers are identified. This method is comprised of eight major steps which are stated below.

Algorithm 1: Ensemble based DEA to identify potential biomarkers.

```
Input: $D_1$: Microarray data, $D_2$: RNA-seq data; $\alpha$, $\beta$, $\gamma$, $k$ : user defined
    Thresholds
Output: A set of DEGs, Potential biomarkers
Pre-process $D_1$ and $D_2$ to obtain $D_1^{'}$, $D_2^{'}$.
Execute the following steps to obtain lists of DEGs for both $D_1^{'}$, $D_2^{'}$.
    1. For $D_1^{'}$, use tools limma, SAM, and EBAM to generate ${DGLM_i}^{\alpha}$ i.e. a
       set of DEGs considered w.r.t. a user defined threshold $\alpha$, where
       i={1,2,3}.
    2. For $D_2^{'}$, use tools DESeq2, edgeR and limma-voom to generate ${DGLR_i}^{\beta}$
       i.e. a set of DEGs considered w.r.t. a user defined threshold $\beta$,, where
       i={1,2,3}.
Obtain a common list of DEGs i.e. $S^1_{common}$ through consensus building.
Mathematically,
    1. Generate the common set of DEGs based on ${DGLM_i}^{\alpha}$.
           $${S^1}_{common} \leftarrow {DGLM_1}^{\alpha} \cap {DGLM_2}^{\alpha} \cap {DGLM_3}^{\alpha}$$
    2. Identify top-k other significant (with lower P-value) genes from $DGLM_i$
       w.r.t. a user defined threshold ($\gamma$) for filtering. Mathematically,
           $${DGLM_i}' \leftarrow {DGLM_i}^{\gamma} - {S^1}_{common}; i = 1, 2, 3$$
    3. Obtain the final set of DEGs for all the tools by taking common of
       $S^1_{common}$ and ${DGLM_i}^{'}$. Mathematically,
           $$DEGM \leftarrow {S^1}_{common} \cup {DGLM_1}' \cup {DGLM_2}' \cup {DGLM_3}'$$
    4. Repeat the steps (a) through (c) for RNA-seq data to obtain DEGR.
Perform the following steps for topological and presentation analysis on
DEGM and DEGR.
    1. Construct Co-Expression Networks (CENs) for both DEGM and DEGR.
    2. Extract sets of modules i.e. $M_{DEGM}$ and $M_{DEGR}$ for the CENs.
Perform preservation analysis to find set of low preserved modules say
    ${LPM_j}^{DEGM}$ and ${LPM_j}^{DEGR}$ (for j=1,2..) based on *Zsummary* score.
Calculate intra-modular connectivity for each low preserved modules of
    ${LPM_j}^{DEGM}$ and ${LPM_j}^{DEGR}$ to identify a sets of hub genes $G^{DEGM}$ and
    $G^{DEGR}$.
Validate each member gene of $G^{DEGM}$ and $G^{DEGR}$ to obtain the useful
    biomarker genes for both $D_1$ and $D_2$.
```

4.3.2.3 Complexity Analysis

The complexity involves various steps, including data processing, DEG identification, consensus building, and network analysis. The efficiency of the algorithm depends on factors such as the size of the datasets, the number of tools used, and the values chosen for the user-defined thresholds. In the integration of outputs from three differential expression analysis methods DESeq2, edgeR, and limma voom, computational complexity is notably influenced by the method with the highest computational demands. Among the three, the most computationally extensive method tends to dominate the overall complexity. The consensus building step includes both intersection and union operation. The computational complexity of an intersection operation is also typically O(min(m, n)), where m and n are the sizes of the two sets being intersected. The computational complexity of a union operation is generally O(m + n), where m and n are the sizes of the two sets being unioned. In general, network construction method ranges from O(n^2) to O(n^3) for most steps. However, the pre-processing and network visualization steps have lower complexity. Module identification and preservation steps involves (O(n^3 log n)) complexities.

4.3.2.4 Experimental Results

A. Differentially Expressed Gene (DEG) Identification

By considering the results obtained from the various DEA tools for both the datasets, a consensus is built and applied for each dataset to obtain two lists of DEGs and this consensus function are applied separately for both datasets.

A.1 Consensus findings for Microarray data: For microarray data, *limma, SAM* and *EBAM* are considered for DEG identification, and total 2916, 3594 and 7386 DEGs are identified respectively for a threshold limit $adjPvalue <$ 0.01 for *limma* and the false discovery rate $FDR <$ 0.01 for *EBAM* and *SAM*. From these genes 1471 DEGs are detected as common among three tools, denoted as ${S^1}_{common}$. For the second sets of DEGs, highly significant DEGs are considered. By considering $adjPvalue <$ 0.001 for limma and found 2125 DEGs. EBAM and SAM give 2288 and 4186 DEGs which are most significant by considering $FDR <$ 0.001. After excluding ${S^1}_{common}$ from each second set of DEGs, total 977, 3024 and 1129 DEGs are filtered out for limma, SAM and EBAM respectively. In this case, total of 3108 DEGs are reported as common. Finally, the final list of

DEGs is prepared by considering all these 3108 and $S^1{}_{common}$, total *4579* DEGs are identified for downstream analysis. limma results total 4282 up and 4099 down regulated genes. The DEA statistics are reported in Table 4.7. Venn diagram of DEGs found by three different methods for both microarray datasets at P-value cut off .01 are shown in Figure 4-6[a]-[b]. For the microarray dataset GSE23400, a total of 9362 DEGs are identified across 106 conditions and the DEA statistics are reported in Table 4.8.

Table 4.7: DEA statistics for GSE20347

P-value<	limma	SAM	EBAM	Common DEGs
0.01	2916	3594	7386	1471
0.001	2125	2288	4186	1084
$DEG_{0.001-1471}$(Excluding duplicate gene)	977	3024	1129	3108 (Union)

Table 4.8: DEA statistics for GSE23400

P-value<	limma	SAM	EBAM	Common DEGs
0.01	6655	7613	9927	4607
0.001	5044	8916	6754	3950
$DEG_{0.001-2272}$(Excluding duplicate genes)	591	4309	2955	4755 (Union)

A.2 Consensus findings for RNA-seq data: Similarly, for RNA-seq data, *DESeq2, edgeR* and *limma-voom* tools are applied and total *7722, 5337* and *2308* number of DEGs are obtained respectively for a threshold limit $Pvalue <$ 0.01. Total up and down regulated genes (up, down) detected by DESeq2, edgeR and limma-voom are (6087, 4113), (3592, 2383), and (3710, 3945). For the first set of common genes i.e. $S^1{}_{common}$, total of 2272 DEGs are found. The DEGs present in DESeq2 but not in $S^1{}_{common}$ and by choosing $Pvalue < 0.001$ is 2807. Similarly, for edgeR and limma-voom, 2198 and 223 DEGs are found respectively. Among 2807, 2198 and 223 DEGs, total of 3137 genes are identified as common. After that, the union between 3137 DEGs and $S^1{}_{common}$ are computed and finally, 5409 genes are detected as DEGs. After removing of missing entries and filtering of low read counts for co-expression network construction, *5165* genes are considered as the significant DEGs. Total 1807 DEGs are found identical for both the datasets. The DEA statistics for SRP064894 are reported in Table 4.9. Total up and down-regulated genes (up, down) identified by DESeq2, edgeR and limma-voom are (6087, 4113), (3592, 2383), and (3710, 3945) respectively. Venn diagram of DEGs identified by the three different methods for RNA-seq data is shown in Figure 4-7. Venn diagram of common genes resulting from the different DEA tools for the three datasets at P-value cut off .001 are shown in Figure 4-8a- 4-9.

Table 4.9: DEA statistics for SRP064894

P-value<	DESeq2	edgeR	limma-voom	Common DEGs
0.01	8254	5975	7655	2272
0.001	4448	3591	1440	1025
$DEG_{0.001-2272}$(Excluding duplicate genes)	2807	2198	223	3137 (Union)

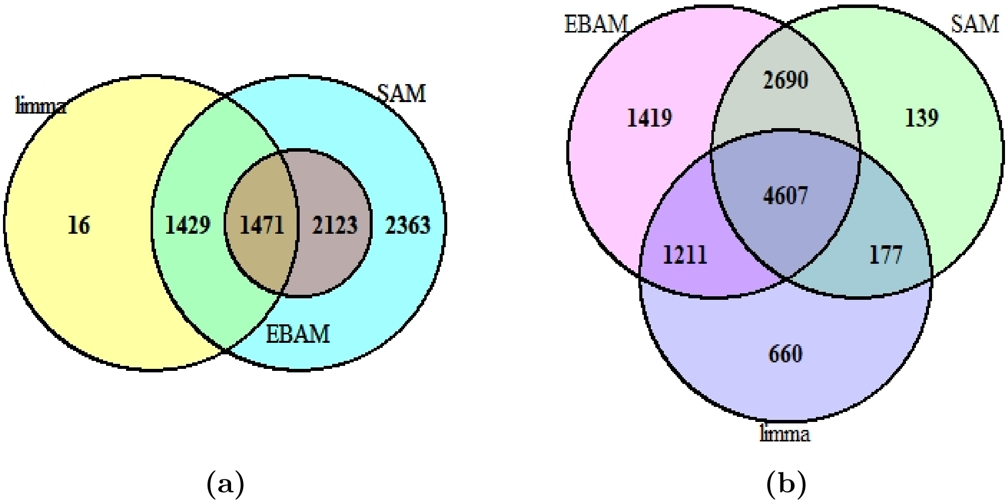


Figure 4-6: Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs identified by three methods at P-value cutoff .01 for (a) GSE20347 (b) GSE23400

Table 4.10: Percentage of common gene detected by DEA methods

Dataset	**DEA Method**	**%Common gene contributed**
RNA-seq data	DESeq2	58.5
	edgeR	54.04
	limma-voom	18.38
GSE20347	limma	40.93
	SAM	69.1
	EBAM	28.11
GSE23400	limma	40.82
	SAM	75.58
	EBAM	54.27

B. Weighted Co-expression Network Construction

For the downstream analysis, preprocessed datasets of DEGs are partitioned into normal and disease samples and scale-free GCN is built using WGCNA package available in R [219] for the two partitions individually for each dataset. An adjacency matrix between finally identified DEGs is computed with a soft threshold. Hierarchical clustering method is applied to find modules for each dataset, resulting in 27 different modules. A unique color is assigned to

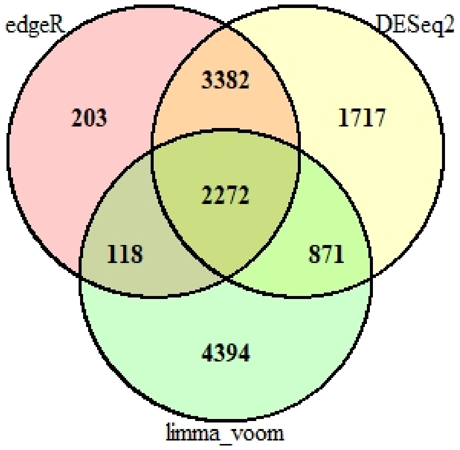
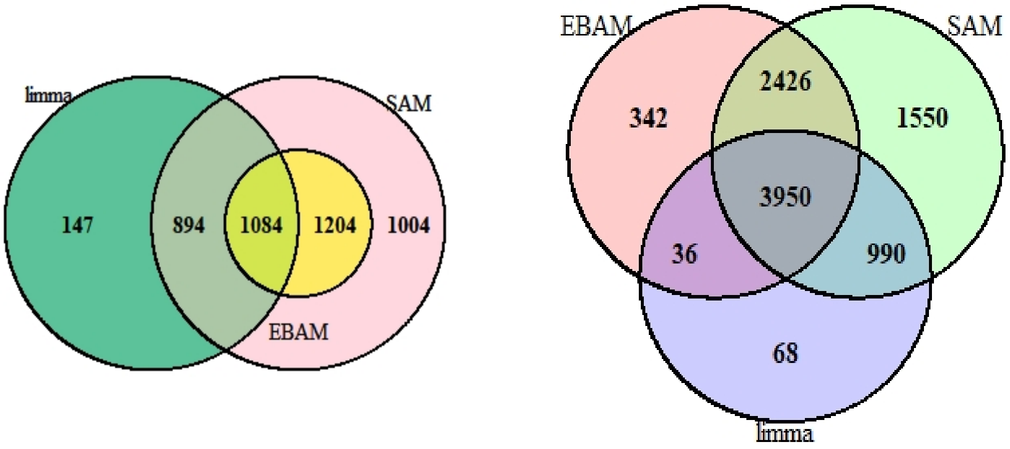


Figure 4-7: Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs identified by three methods for RNA-seq data SRP064894 P-value cutoff .01



(a) **(b)**

Figure 4-8: Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs at P-value cutoff .001 identified by three methods for microarray data (a) GSE20347 (b) GSE23400

each module, and the modules are further analyzed by identifying eigengenes for each module. Eignmodules are clustered using dissimilarity of module eigengenes and most related modules are combined (MEDissThres=0.25) (Figure 4-10[a]-[f] and Figure 4-11[a]-[f]). Eventually, 22 and 25 modules were derived from normal and disease datasets respectively for the GSE20347 dataset. For the another microarray dataset GSE23400, 9 modules are found for normal state and 17 modules are extracted for tumor state.

Similarly, for RNA-seq data, WGCNA is applied separately to normal and disease samples with 5165 genes, and the results of WGCNA are shown in Figure 4-

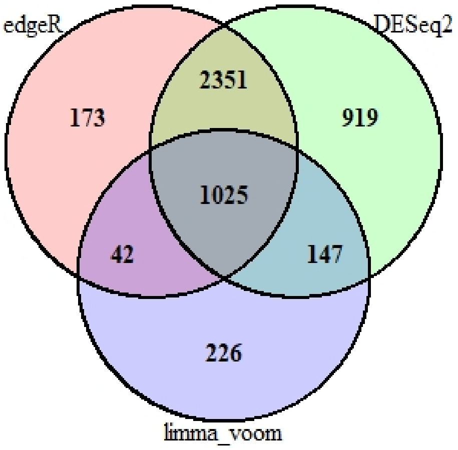


Figure 4-9: Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs at P-value cutoff .001 identified by three methods for RNA-seq data SRP064894

12[a] and Figure 4-12[d]. Finally, 14 control network modules and 21 disease network modules (Figure 4-12[a]-[f]) are extracted.

C. Module Preservation Analysis

In this study, module preservation is carried out for two cases, i.e. (a) normal to tumor, where, we consider test data as the normal state and reference data as tumor state. This analysis aims to check which module of normal state is lowly preserved in tumor state and (b) tumor to normal, where, test data is considered as the tumor state and reference data as normal state. It aims to check which module of tumor state is lowly preserved in normal state. My observations for both the cases are as follows.

(a) Normal to tumor: For GSE20347 microarray data as shown in Figure 4-13[a] [228], white module is detected as the low-preserved module with *Zsummary* score 1.1 and *MedianRank* 18. This module comprises a total of 41 genes. On the other hand, *Zsummary* for the steelblue module is 1.7 for the RNA-seq results, which is the lowest and *MedianRank* is 13 as shown in Figure 4-13[b] [228]. This module includes 55 genes in total and is recommended for further study. For GSE23400, purple module has been identified as the least preserved module with *Zsummary* score 7.9 and *MedianRank* 5 as shown in Figure 4-13[c]. Total of 102 genes are found in this module.

(b) Tumor to normal: Greenyellow module is detected as the least preserved module with *Zsummary* score 3.4 and *MedianRank* score 16 for the dataset

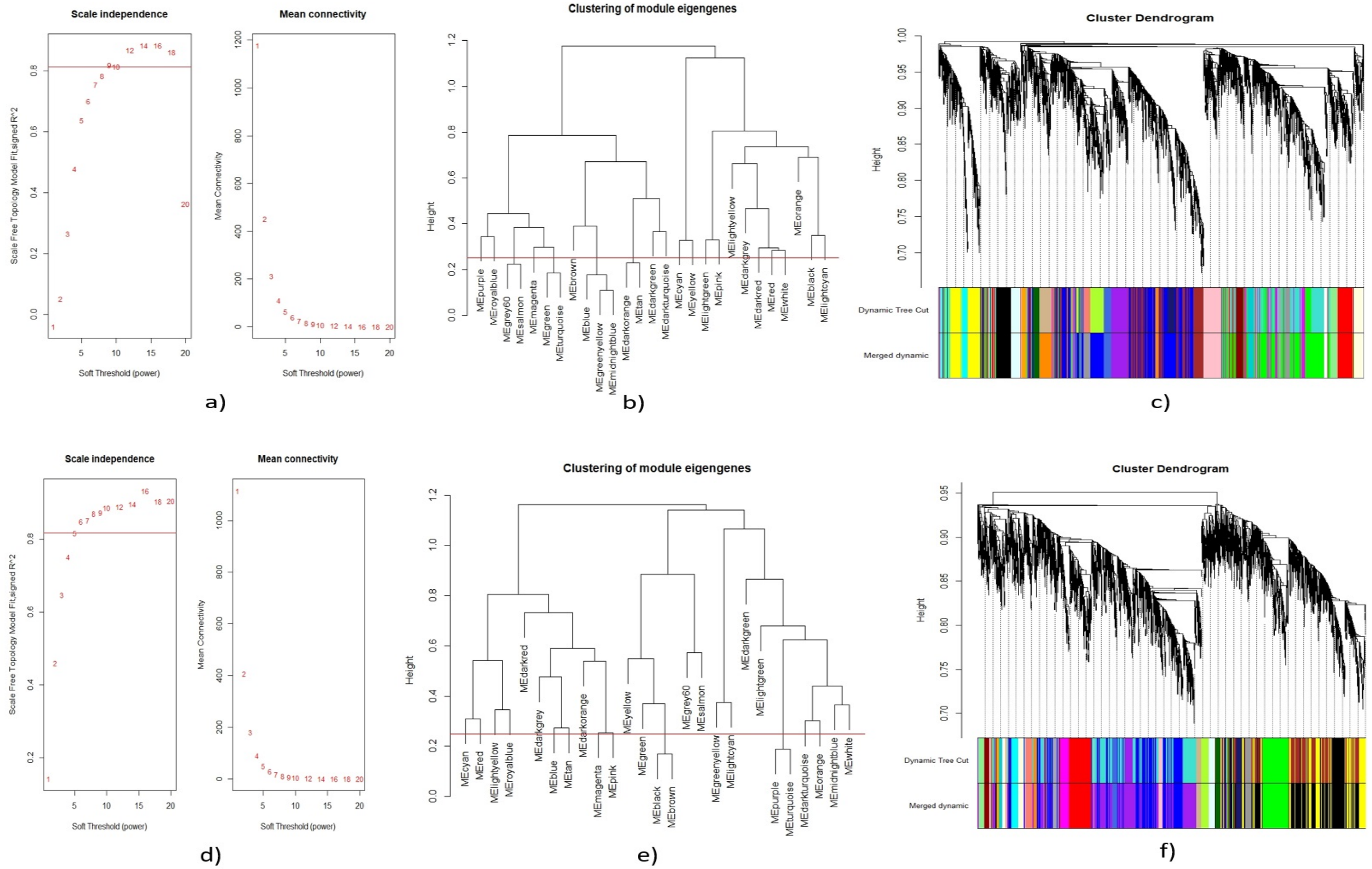


Figure 4-10: GCN analysis of DE genes for microarray data GSE20347 for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM).

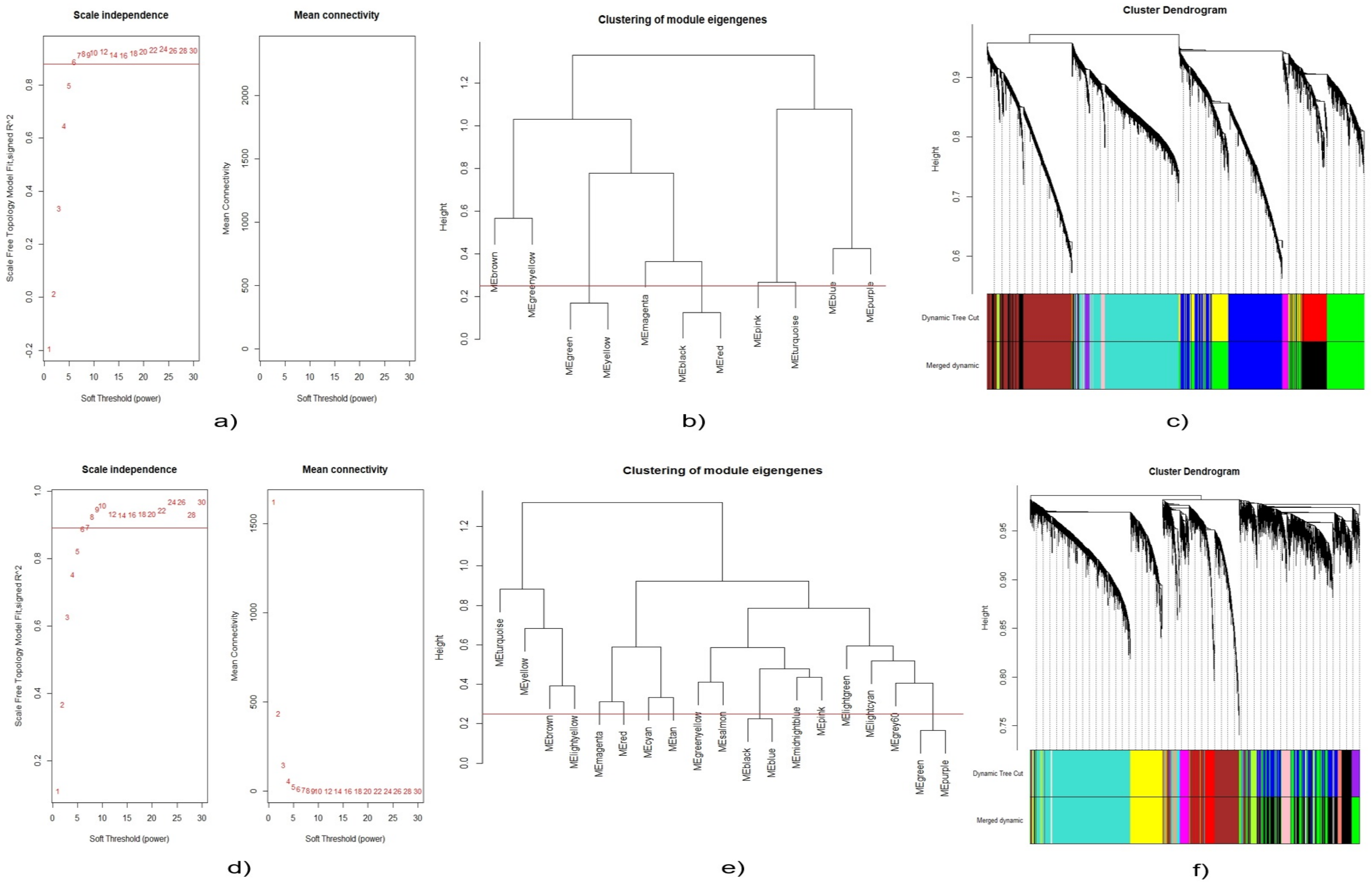


Figure 4-11: GCN analysis of DE genes for microarray data GSE23400 for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM).

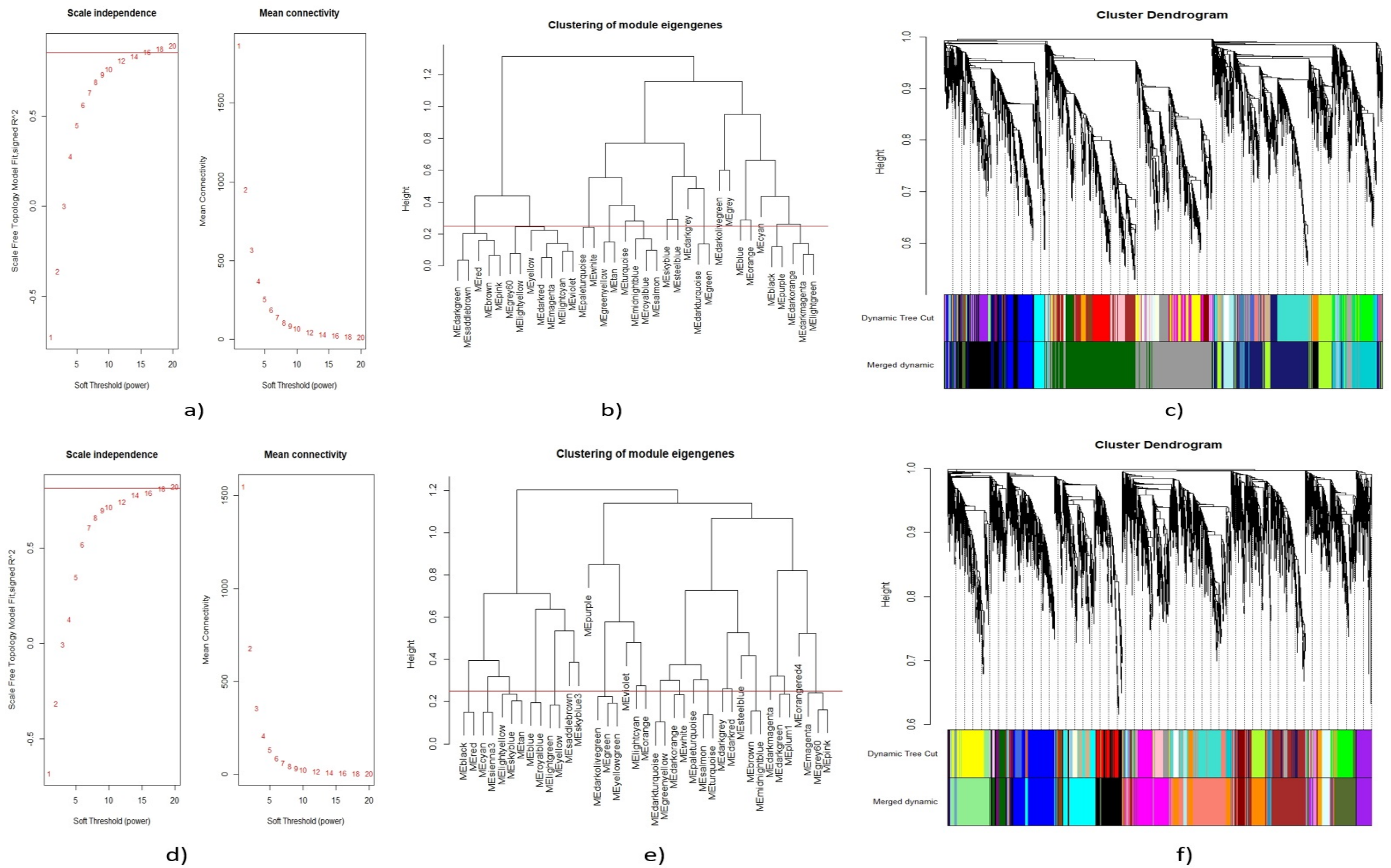


Figure 4-12: GCN analysis of DE genes for RNA-seq data for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM).

GSE23400 and is shown in Figure 4-13[d]. For the microarray data, darkgreen, lightcyan, grey60, orange, royalblue, and yellow modules (Figure 4-13[e]) have been detected as lowly preserved modules with 57, 75, 73, 44, 60, and 404 genes in each module. Their *Zsummary* scores are 0.32, 1.0, 1.3, 0.82, 1.9, and 1.6 and *MedianRank* scores are 26, 20, 24, 18, 12, and 22, respectively. In this analysis, only darkgreen module is considered as it is found that *MedianRank* score comparatively higher but *Zsummary* score is lower. For the RNA-seq data, only darkred module is found as the lowly preserved module with *Zsummary* score 1.9 and *MedianRank* score 22 and is shown in Figure 4-13[f].

D. Hub Gene Finding

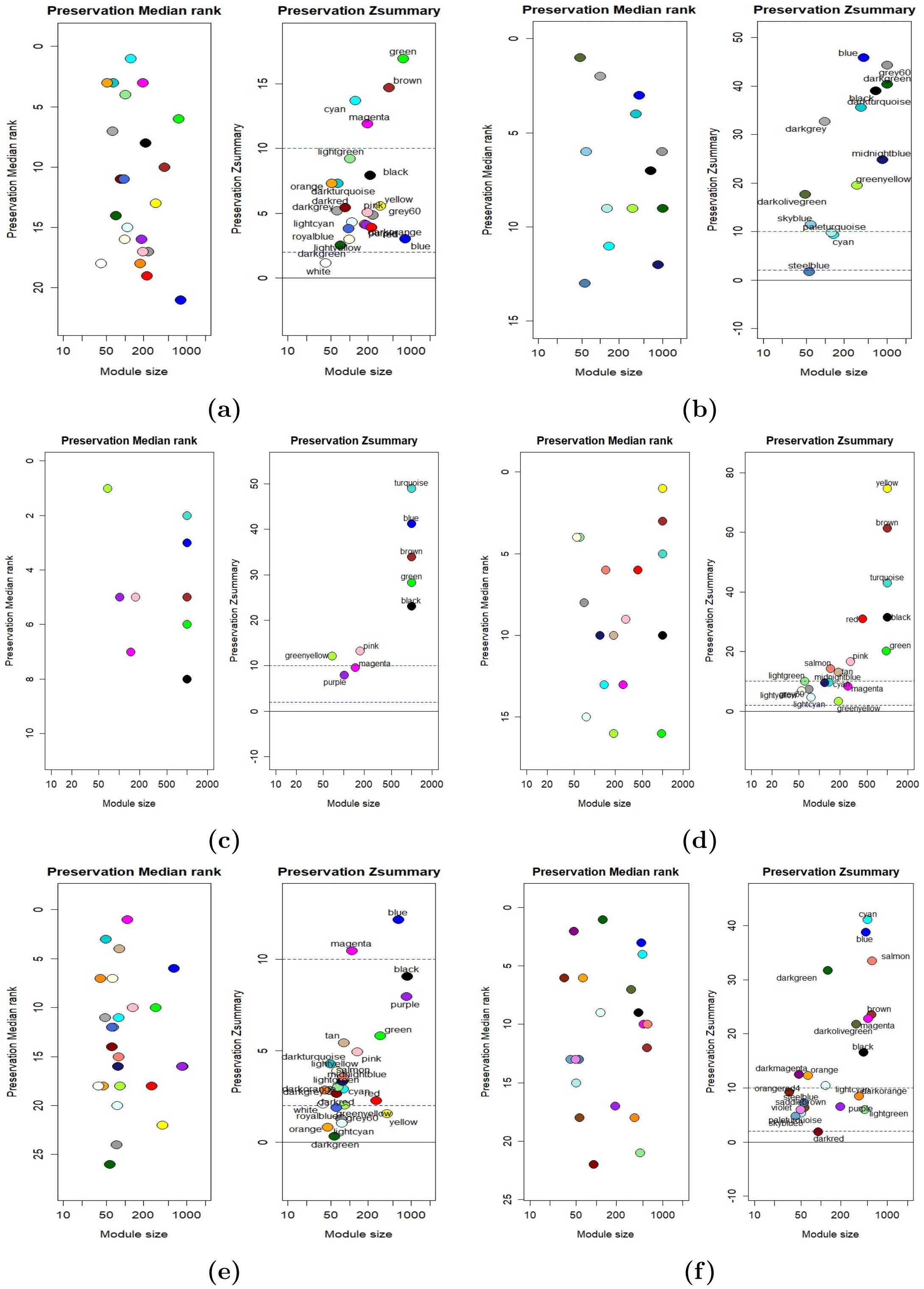
Hub genes play an important role in the study of biological networks. the low preserved modules are considered and studied their topological behaviour in terms of intra-modular connectivity for identification of the hub gene(s) and six genes are found prominant to be considerd as hub genes, such as - BAG6, COL27A1, SOX11, TOP3A, MROH7, and AKR1C2. The degrees are reported in terms of correlation weight since the network is signed. The weights of the hub genes identified by intra-modular connectivity using WGCNA are shown in Table 4.11. From each module with the highest intra-modular connectivity, these hub genes are found, looking at all genes in the expression data in R. Again, using STRING tool, PPI networks (shown in Figure 4-14-4-16)are constructed for each low preserved module and hub genes are identified based on degree and confidence score (Table 4.12). Among them CDC6, EZH2, COL7A1, ALB, PTH1R, and G6PD are considered as the hub genes. The interesting DE genes ”SOX11”, ”TOP3A”, ”COL27A1”, ”CDC6”, ”EZH2”, ”COL7A1”, ”ALB”, ”MROH7”, and ”G6PD” are found as up-regulated genes and ”BAG6”, ”PTH1R”, and ”AKR1C2” are identified as down-regulated genes in the ESCC datasets. From the website of cBioPortal[7], hub genes BAG6, COL27A1, SOX11, TOP3A, ALB and EZH2 are found to match Esophageal cancer mutated genes.

E. Finding Biological Significance of Low preserved modules and Hub Genes

From biological GO analysis and pathway analysis done in DAVID[8], GeneAn-

[7] https:/www.cbioportal.org/

[8] https://david.ncifcrf.gov



(a) (b) (c) (d) (e) (f)

Figure 4-13: Preservation analyses: Test-Normal, Reference-Tumor([a], [b], [c]); Test-Tumor, Reference-Normal([d], [e], [f]). [a] GSE20347: Here, the white module is the least preserved module. [b] RNA-seq data: Here, the steelblue module is the least preserved module. [c] GSE23400: Here, the purple module is the least preserved module. [d] GSE23400: Here, the greenyellow module is the least preserved module. [e] GSE20347: Here, the darkgreen module is the least preserved module. [f] RNA-seq data: Here, the dark red module is the least preserved module.

Table 4.11: Top hub genes identified by WGCNA in each non-preserved module

Normal to Tumor					
Steelblue		**White**		**Purple**	
Gene Name	**Weight**	**Gene Name**	**Weight**	**Gene Name**	**Weight**
COL27A1	13.5255198	**BAG6**	8.34646898	**MROH7**	24.8255450
MDC1	13.4520206	UBA1	8.22319123	REM1	24.5396723
ADAR	12.6529693	ADRM1	7.98908503	HOXA3	23.3803377
TENM4	11.640437	BOP1	7.97059197	EDA	23.3501984
PLCG2	11.4433501	FLAD1	7.94529456	GABRA5	23.2320597
Tumor to Normal					
Darkgreen		**Darkred**		**Greenyellow**	
Gene Name	**Weight**	**Gene Name**	**Weight**	**Gene Name**	**Weight**
SOX11	15.4369594	**TOP3A**	10.7344736	**AKR1C2**	26.3176096
ACAD8	13.5834851	RAG2	10.7104490	CYP4F11	24.5396723
TCF4	13.4515386	SCO1	8.53560949	TXNRD1	24.562100
CCNE2	12.7797402	LINC0181	8.44543105	GPX2	23.2497139
FST	12.6711707	XKR9	8.30333824	ADAM23	23.2166119

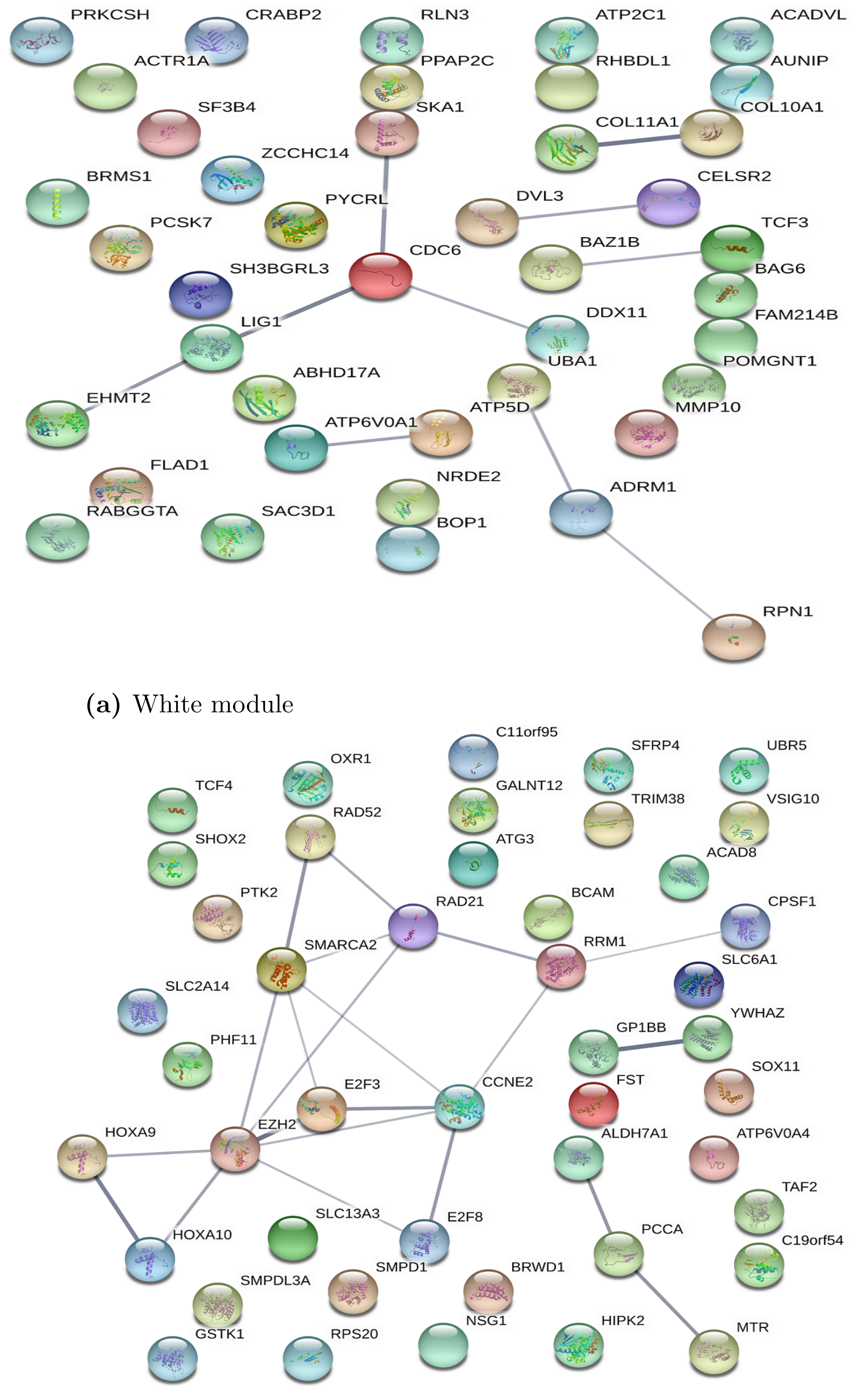
Table 4.12: Top hub genes identified from PPI network by STRING tool in each non-preserved module. Note: CG: Central Gene, CS: Confidence score

GSE20347					
White			**Darkgreen**		
Central gene	**Degree**	**CS**	**CG**	**Degree**	**CS**
CDC6	3	2.115	EZH2	7	4.21
LIG1	2	1.552	CCNE2	5	2.802
ADRM1	2	1.024	SMARCA2	5	2.602
SRP064894					
Steelblue			**Darkred**		
CG	**Degree**	**CS**	**CG**	**Degree**	**CS**
COL7A1	3	2.502	ALB	5	2.633
COL5A3	2	1.852	RBBP4	4	3.307
COL27A1	2	1.847	CD44	4	2.793
GSE23400					
Purple			**Greenyellow**		
CG	**Degree**	**CS**	**CG**	**Degree**	**CS**
PTH1R	4	3.136	G6PD	10	6.578
GPR20	4	3.121	TXNRD1	9	5.567
ADRB1	3	2.712	GCLC	8	5.348

alytics[9] and STRING[10] database, we can decipher the biological interpretation behind the low preserved modules and identified interesting genes. Percentage of enrichment in Biologiocal Process, Cellular Component and Molecular Function for each module are reported in Table 4.13. GO enrichment analysis performed in DAVID for selected four low preserved modules are presented in Table 4.14- 4.19

[9] https://ga.genecards.org

[10] https://string-db.org/

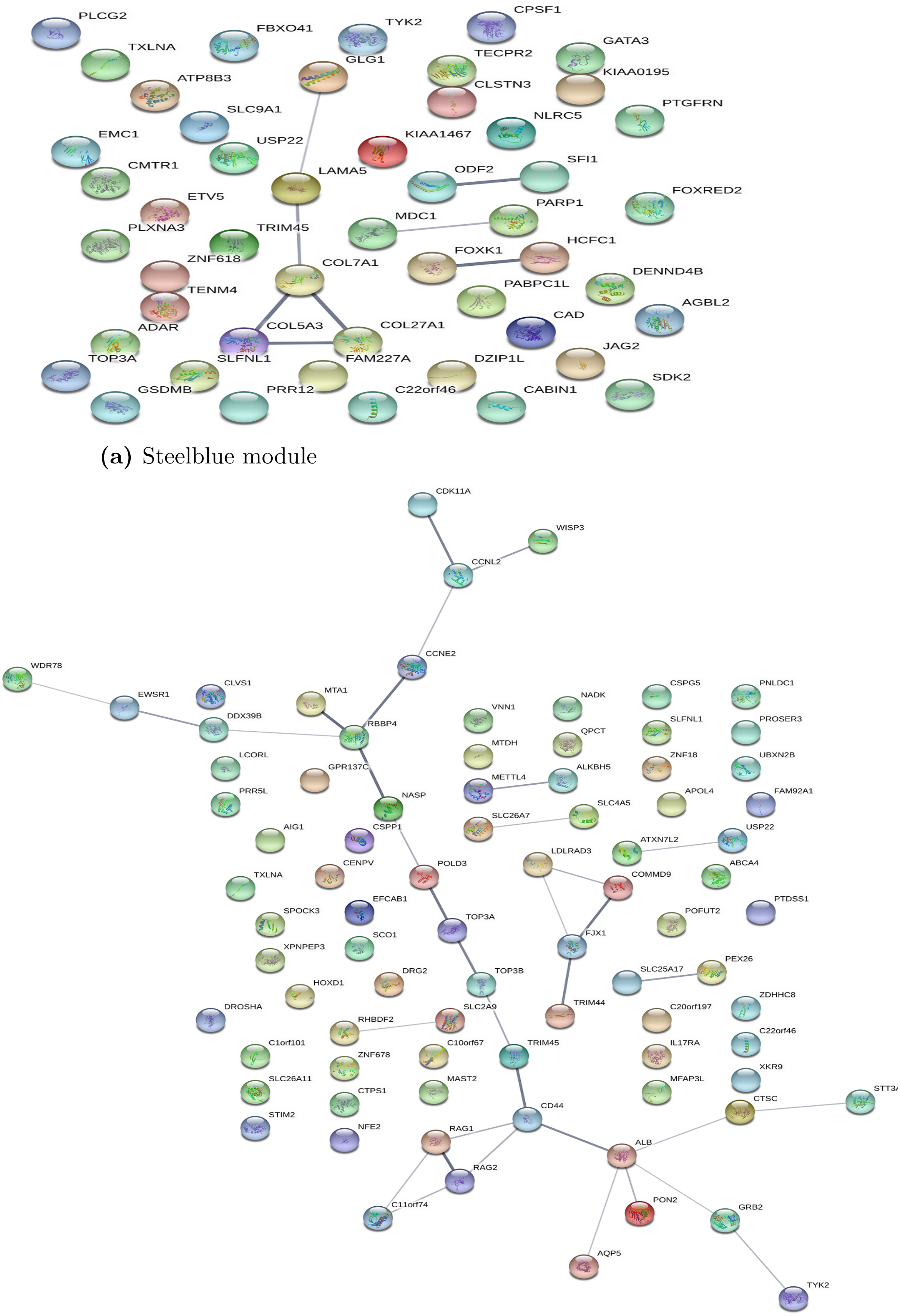


(a) White module

(b) Darkgreen module

Figure 4-14: PPI network of low preserved modules constructed in STRING

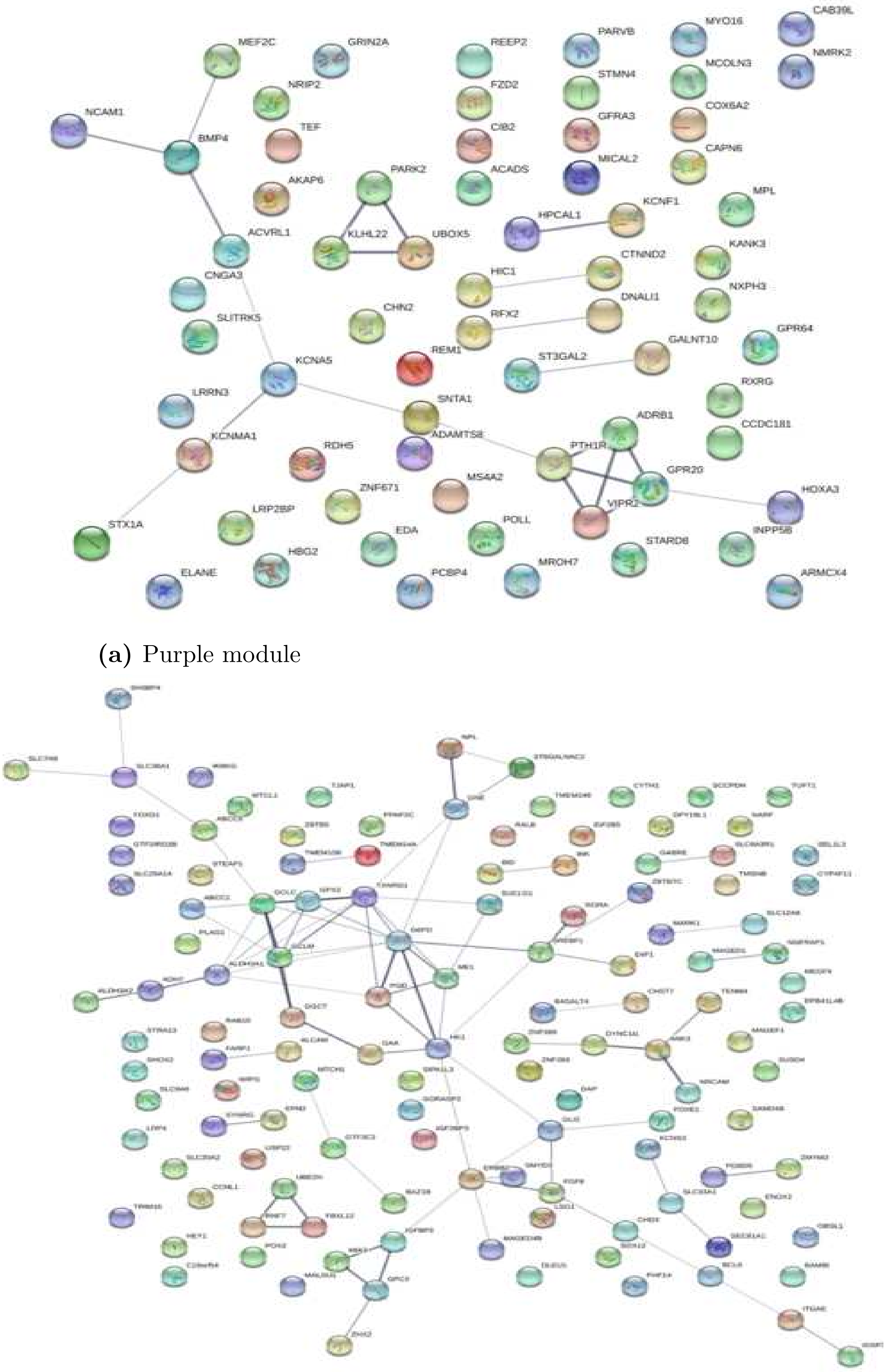
and pathway analysis of each modules performed in GeneAnalytics are shown in Table 4.20- 4.25.



(a) Steelblue module

(b) Darkred module

Figure 4-15: PPI network of low preserved modules constructed in STRING



(a) Purple module

(b) Greenyellow module

Figure 4-16: PPI network of low preserved modules constructed in STRING

Table 4.13: Enrichment analysis result (ordered) of low-preserved modules for microarray and RNAseq data. Abbreviations: MF-Molecular Function, BP-Biologiocal Process, CC-Cellular Component

Module	BP(%)	CC(%)	MF(%)
White	89.7	97.4	97.4
Darkgreen	84.8	84.8	84.8
Steelblue	65.2	63	73.9
Darkred	63.3	68.4	68.4
Purple	92.7	96.9	92.7
Greenyellow	92.4	96.8	93.7

Table 4.14: GO enrichment analysis for White module

Term	Count	PValue	Benjamini	FDR
GO:0051301~cell division	4	0.033358876	0.947156712	35.81981197
GO:0030574~collagen catabolic process	3	0.007416366	0.855639431	9.271735471
GO:0006464~cellular protein modification process	3	0.019075346	0.918222373	22.25565118
GO:0006974~cellular response to DNA damage stimulus	3	0.066139623	0.971512041	59.1160656
GO:0007067~mitotic nuclear division	3	0.089556965	0.982849095	70.66477686
GO:0007130~synaptonemal complex assembly	2	0.041694585	0.937229404	42.6892135

Table 4.15: GO enrichment analysis for Darkgreen module

Term	Count	PValue	Benjamini	FDR
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	9	0.001903736	0.313635295	2.618123812
GO:0006351~transcription, DNA-templated	9	0.086710366	0.86336057	71.71401757
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	8	0.00140119	0.425271886	1.933246712
GO:0045893~positive regulation of transcription, DNA-templated	5	0.033465441	0.775504547	37.74292552
GO:0007399~nervous system development	4	0.030660075	0.827469574	35.17937986
GO:0008284~positive regulation of cell proliferation	4	0.098819086	0.885035181	76.51077147

F. Literature evidence

Total twelve critical genes are identified which are mostly responsible for

Table 4.16: GO enrichment analysis for Steelblue module

Term	Count	PValue	Benjamini	FDR
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	6	0.088913564	0.917644755	73.07277067
GO:0030154~cell differentiation	5	0.025679591	0.844339402	30.68813707
GO:0030198~extracellular matrix organization	4	0.012055085	0.92583985	15.70849685
GO:0045087~innate immune response	4	0.087073841	0.938675315	72.29644071
GO:0030334~regulation of cell migration	3	0.014060187	0.78099409	18.0872015
GO:0016477~cell migration	3	0.065953567	0.961313134	61.76248877

Table 4.17: GO enrichment analysis for Darkred module

Term	Count	PValue	Benjamini	FDR
GO:0006260~DNA replication	4	0.024035991	0.85369122	28.73244597
GO:0051453~regulation of intracellular pH	3	0.00905313	0.972464292	11.89284408
GO:0042157~lipoprotein metabolic process	3	0.010050573	0.863990507	13.11956119
GO:0015701~bicarbonate transport	3	0.013318834	0.733950293	17.02889228
GO:0051726~regulation of cell cycle	3	0.087740883	0.973412997	72.15515345
GO:0002331~pre-B cell allelic exclusion	2	0.011922996	0.79387963	15.37967875

Table 4.18: GO enrichment analysis for purple module

Term	Count	PValue	Benjamini	FDR
GO:0007165~signal transduction	14	0.007265947	0.578813197	10.63857129
GO:0043547~positive regulation of GTPase activity	10	0.002794556	0.538947204	4.224599616
GO:0008285~negative regulation of cell proliferation	9	0.001106675	0.601101595	1.69340761
GO:0000165~MAPK cascade	6	0.0122013	0.639021456	17.25061942
GO:0007155~cell adhesion	6	0.093222288	0.868736987	77.89545793
GO:0007050~cell cycle arrest	5	0.006443268	0.591066444	9.489493088

ESCC. Out of these, the prominent nine genes such as COL27A1, SOX11, BAG6, TOP3A, CDC6, EZH2, COL7A1, G6PD, and AKR1C2 are highlighted below.

(a) ***SOX11:*** SRY-Box 11 gene is a transcriptional factor that is believed to be involved in the regulation of some important biological functions such as devel-

Table 4.19: GO enrichment analysis for greenyellow module

Term	Count	PValue	Benjamini	FDR
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	17	0.010650312	0.673019887	15.64218127
GO:0055114~oxidation-reduction process	16	1.88E-04	0.177973582	0.297780156
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	12	0.047608456	0.837773451	53.92612093
GO:0006810~transport	10	0.003271628	0.574842193	5.072775376
GO:0006357~regulation of transcription from RNA polymerase II promoter	10	0.014434356	0.717743868	20.62456118
GO:0045892~negative regulation of transcription, DNA-templated	10	0.029261442	0.755688131	37.61193737

opment, differentiation and cell-fate decision. Therefore, any dysregulation in the expression of SOX11 gene would lead to the development of cancer [229]. The expression SOX11 gene is found to be down-regulated in several cancers viz. gliomas, ovarian cancer, hematologic cancer and nasopharyngeal cancer [230]. SOX11 is found to be associated with ESCC cell growth [231] and it acts as a tumor suppressors in esophageal squamous cell carcinoma, hematopoietic malignancies, and gastric and liver cancers, respectively [232][233]. Interestingly, recent evidence provided that the gene expression of SOX11 is up-regulated in ESCC [234]. But exactly, how this gene is related to ESCC is still unclear. In this study, using intra-modular connectivity, SOX11 is identified as a hub gene, which is up-regulated in the ESCC dataset. However, no pathway was found to be associated with SOX11 gene in KEGG as well as in the reactome pathway database. Therefore, from GeneCards, it has been found that that major pathway involved with SOX11 gene is ERK/MAPK signaling pathway. The ERK signaling pathway plays a crucial role in various cellular processes that include cell development, cell proliferation, differentiation and survival. This regulatory signaling pathway is often found to be up-regulated in many types of human cancers [235]. Recent evidences have provided that ERK signaling is highly up-regulated in ESCC and its expression is negatively correlated with STAT1 transcription factor [236]. However, the molecular link between SOX11 and ERK pathway is yet to be known. Since SOX11 has a major role in proliferation , survival and development, there is a high probability that SOX11 is correlated to ERK signaling pathway. From the STRING database, another KEGG pathway is found - transcriptional misregulation in cancer (hsa05202) associated with SOX11.

Table 4.20: Pathway analysis for White module. (R: Reactome; N: NCBI Biosystem; K: KEGG; C: Cell signalling technology)

Name	Matched Genes	Sources
Wnt Signaling Pathway Netpath	TCF3, DVL3	N
Wnt / Hedgehog / Notch	BAG6, CELSR2, DVL3	C
CDK-mediated phosphorylation and removal of Cdc6	ADRM1, LIG1, RPN1, CDC6, UBA1, DVL3	K, R, B
Metabolism of proteins	ACADVL,ADRM1,POGNT1, RPN1, PRKCSH, DDX11, RABGGTA, UBA1	R
Unfolded Protein Response (UPR)	ACADVL, DDX11	R
Riboflavin metabolism	FLAD1	K
Degradation of the extracellular matrix	MMP10, COL10A1, COL11A1	R
WNT mediated activation of DVL	DVL3	R
Advanced glycosylation endproduct receptor signaling	PRKCSH	R
Mitotic Prometaphase	SKA1, DVL3	R
Signaling pathways regulating pluripotency of stem cells	TCF3, DVL3	K
Cell Cycle, Mitotic	LIG1, CDC6, SKA1, ACTR1A	R
Ectoderm Differentiation	CELSR2, TCF3	N
p38 signaling mediated by MAPKAP kinases	TCF3	N
Urea cycle and metabolism of amino groups	PYCR3	N
Protein processing in endoplasmic reticulum	RPN1, PRKCSH	K
Notch signaling pathway (KEGG)	EHMT2, DVL3	K, N
Mitochondrial Fatty Acid Beta-Oxidation	ACADVL	R
Signaling events mediated by PRL	RABGGTA	N
Mannose type O-glycan biosynthesis	POMGNT1	K

(b) ***COL27A1:*** Collagen Type XXVII Alpha 1 Chain is a type of fibrillar collagen and the components of extracellular matrix (ECM). However, its significance in cellular processes is poorly understood [237]. In this dataset, COL27A1 is identified as one of the up-regulated hub gene in ESCC. ECM plays an important role by regulating cell proliferation, cell degradation and remodeling. Therefore, aberrant changes in the gene expression of ECM components will lead to malignant transformation [238]. The identified KEGG pathway associated with COL27A1 is protein digestion and absorption which justifies the role of COL27A1 in ESCC. Further, COL27A1 has been found to be significantly upregulated in ESCC and plays a major role in extra cellular matrix organization [239].

(c) ***TOP3A:*** Topoisomerases are predominantly present in both prokary-

Table 4.21: Pathway analysis for Darkgreen module. (R: Reactome; N: NCBI Biosystem; K: KEGG; G: GeneGo; Q: Qiagen)

Name	Matched Genes	Sources
E2F transcription factor network	E2F3, RRM1, SMARCA2, CCNE2	N
Retinoblastoma (RB) in Cancer	E2F3, RRM1, SMARCA2, CCNE2	B
Valine, leucine and isoleucine degradation	ACAD8, ALDH7A1, PCCA	K, N, R
GP1b-IX-V activation signalling	GP1BB, YWHAZ	R
Cell cycle	E2F3, RAD21, YWHAZ, CCNE2	K, N
DNA damage	HIPK2, RAD21, RRM1, CCNE2, YWHAZ	C
Regulation of Wnt/B-catenin Signaling by Small Molecule Compounds	TCF4, SFRP4	N
Small cell lung cancer	E2F3, PTK2, CCNE2	K
Glioblastoma Multiforme	E2F3, TCF4, CCNE2	Q
Integrin alphaIIb beta3 signaling	GP1BB, PTK2	R
Diseases of metabolism	MTR, PCCA	R
Defective MTR causes methylmalonic aciduria and homocystinuria type cblG	MTR	R
Cellular senescence	HIPK2, E2F3, CCNE2	K
Transcription Ligand-dependent activation of the ESR1/SP pathway	SMARCA2, CCNE2	G
Cell cycle Cell cycle	E2F3, CCNE2	R, G
Lysine degradation	ALDH7A1, EZH2	K
Transcriptional misregulation in cancer	HOXA9, HOXA10, PTK2	K
Human Embryonic Stem Cell Pluripotency	E2F3, TCF4, CCNE2	Q
Regulation of TP53 Activity	HIPK2, CCNE2, YWHAZ, TAF2	R
Tryptophan metabolism	ALDH7A1, UBR5	K, N
Preimplantation Embryo, ERK Signaling	SOX11	GeneCard

otes and eukaryotes. It plays a major biological role in DNA topological and conformational changes. TOP3A belongs to family of type lA topoisomerase family, which is mainly associated with regulation of cell cycle checkpoints, DNA-repair mechanism and to maintain stability of genome [240]. Studies have reported the involvement of TOP3A gene in bladder cancer [241], lungs cancer and nasopharyngeal squamous cell carcinoma [241]. But the association of TOP3A with ESCC has not been studied yet. It is the first time TOP3A is identified as an significant gene in ESCC by module preservation analysis.

(d) ***BAG6:*** BCL2 Associated Athanogene 6 is associated with Discrete Subaortic Stenosis and Acute Diarrhea disease. Wnt/Hedgehog/Notch pathway is shared by BAG6 which plays key roles in embryogenesis and carcinogenesis.

Table 4.22: Pathway analysis for Steelblue module. (R: Reactome; N: NCBI Biosystem; K: KEGG; C: Cell signalling technology; RD: RandD Systems; PH: PharmGKB; G: GeneGo; Q: Qiagen)

Name	Matched Genes	Sources
Phospholipase-C	COL27A1, COL5A3, LAMA5, GLGG1, COL7A1, PLCG2	Q
Peginterferon alpha-2a/Peginterferon alpha-2b Pathway (Hepatocyte), Pharmacodynamics	ADAR, TYK2	Ph
Interleukin-4 and 13 signaling	LAMA5, GATA3, TYK2	R
Formation of editosomes by ADAR proteins	ADAR	R
Th2 Differentiation Pathway	JAG2, GATA3	RD
Degradation of the extracellular matrix	COL27A1, COL5A3, LAMA5, COL7A1	R
Deubiquitination	FOXK1, HCFC1, GATA3, USP22	R
Th17 cell differentiation	JAG2, GATA3, TYK2	K, Q, G
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	GATA3, CABIN1	N
NF-kappaB Signaling	GATA3, CABIN1, TYK2, PLCG2	C
AIF Pathway	PARP1	Q
Cytokine Signaling in Immune system	ADAR, TYK2, LAMA5, GATA3, TRIM45, TXLNA	R
Collagen chain trimerization	COL27A1, COL5A3, COL7A1	R, Q, K
Integrin Pathway	COL27A1, COL5A3, LAMA5, COL7A1, PLCG2	Q
Interferon gamma signaling	ADAR, TYK2, TRIM45	R
Th1 Differentiation Pathway	GATA3, TYK2	N, RD
IL12-mediated signaling events	ETV5, TYK2	G, N
UVA-Induced MAPK Signaling	COL27A1, PLCG2	Q
Immune response IFN alpha/beta signaling pathway	ADAR, TYK2	R, G
mRNA surveillance pathway	CPSF1, PABPC1L	K

This pathway acts as activators for epidermal growth factor receptor in ESCC [242]. BAG6 gene is found to be associated with lung cancer [243].

(e) ***CDC6:*** Cdc6's deregulated expression can play a role in human cancer oncogenic transformation or disease progression, and Cdc6 provides a good target for biotechnological strategies to inhibit cell proliferation [244]. Altered Cdc6 expression has been observed in malignant prostate cancer cells, but semi-quantitative PCR, microarray analysis, and Western blotting all revealed a decrease in CDC6 transcription and protein expression relative to normal prostate tissue [245]. CDC6 is under YB-1 control in cancer cells and evidence says that CDC6 expression plays an essential role in the YB-1-induced cell proliferation and

Table 4.23: Pathway analysis for Darkred module (R: Reactome; N: NCBI Biosystem; K: KEGG)

Name	Matched Genes	Sources
Resolution of D-loop Structures through Synthesis-Dependent Strand Annealing (SDSA)	POLD3, TOP3B, TOP3A	K, N, R
Defective SLCO1B3 causes hyperbilirubinemia, Rotor type (HBLRR)	ALB	R
Retinoblastoma (RB) in Cancer	RBBP4, POLD3, CCNE2	N
Interleukin-7 signaling	RAG1, RAG2	R
Primary immunodeficiency	RAG1, RAG2	K
Cytokine Signaling in Immune system	IL17RA, GRB2, RAG1, RAG2, CD44, TYK2, TRIM45, TXLNA	R
Interleukin-11 Signaling Pathway	GRB2, TYK2	N
FoxO signaling pathway	GRB2, RAG1, RAG2	K
RNA Polymerase I Promoter Escape	RBBP4, MTA1	R
Signaling events mediated by PTP1B	GRB2, TYK2	N
Hepatitis C and Hepatocellular Carcinoma	GRB2, CD44	N
Cell cycle Cell cycle (generic schema)	RBBP4, CCNE2	R
Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	ABCA4, ALB, AQP5, SLC4A5, SLC26A11, SLC26A7	R
Fanconi anemia pathway	TOP3B, TOP3A	K
IL6-mediated signaling events	GRB2, TYK2	N
Phase I biotransformations, non P450	PON2	N
DNA Damage Reversal	ALKBH5	R
Homologous DNA Pairing and Strand Exchange	POLD3, TOP3A	R
E2F transcription factor network	RBBP4, CCNE2	N
Validated targets of C-MYC transcriptional activation	MTDH, MTH1	N

Table 4.24: Pathway analysis for Purple module

Name	Matched Genes	Sources
Neuroactive ligand-receptor interaction	THRA, ADRB1, PTH1R, P2RX2, GABRA5, GRIN2A, VIPR2	KEGG
Basal cell carcinoma	BMP4, FZD2, APC	KEGG
Signaling pathways regulating pluripotency of stem cells	BMP4, FGFR1, FZD2, APC	KEGG
Pathways in cancer	BMP4, FGFR1, RXRG, FZD2, RARB, APC	KEGG

Table 4.25: Pathway analysis for Greenyellow module

Name	Matched Genes	Sources
Glutathione metabolism	GPX2, GCLC, G6PD, GGCT, PGD, GCLM	KEGG
Carbon metabolism	ME1, G6PD, SUCLG1, PGD, HK1	KEGG
Central carbon metabolism in cancer	FGFR2, G6PD, ERBB2, HK1	KEGG
Pancreatic cancer	ERBB2, IKBKG, RALB, IKBKB	KEGG
Glycolysis / Gluconeogenesis	HK1, ADH7, ALDH3A2, ALDH3A1	KEGG
Metabolic pathways	ME1, GCLC, GNE, SLC33A1, PGAP1, SUCLG1, PGD, HK1, ADH7, GCLM, ALDH3A2, PLPP2, ALDH3A1, NNT, G6PD, IDS, GAA, B4GALT4, PRODH	KEGG
Prostate cancer	FGFR2, ERBB2, IKBKG, IKBKB	KEGG
Biosynthesis of antibiotics	G6PD, SUCLG1, PGD, HK1, ALDH3A2, PRODH	KEGG
Amino sugar and nucleotide sugar metabolism	GNE, NPL, HK1	KEGG
Pathways in cancer	BID, FGFR2, FGF8, ERBB2, IKBKG, RALB, GLI2, IKBKB	KEGG

cell cycle G1/S. CDC6 is found as a novel therapeutic target for cancer radiosensitization [246]. In ESCC disease, CDC6 is found to be upregulated [247]. It is mainly associated with the regulation of cell cycle KEGG pathway [248].

(f) ***EZH2:*** Studies revealed that overexpression of EZH2 and H3k27me3 could act as a prospective marker for ESCC patients. By reducing EZH2 expression using siRNA or treatment of EZH2 inhibitor could result in cell growth inhibition and reduced tumor formation in various cancers [249]. EZH2 was found to be upregulated in ESCC patients in comparison to adjacent normal tissue. Deregulated EZH2 expression is significantly associated with large size, depth of invasion, presence of distant metastasis and shorter disease-free survival time [250]. Aberrant overexpression of EZH2 serves as a poor prognostic biomarker for ESCC patients [249]. EZH2 has been identified as a repressor of gene transcription and it is also reported to be associated with biological malignancy in several cancers [251].

(g) ***COL7A1:*** COL7A1 gene, is the major component of anchoring fibre in the basement membrane. From studies, it has been found that COL7A1 is upregulated in most of the ESCC patients [252]. Moreover, validation of differentially expressed COL7A1 mRNA by qRT PCR revealed that COL7A1 is significantly altered in the majority of the ESCC cases [253].

(h) ***G6PD:*** The expression of G6PD mRNA and protein has been found to be significantly upregulated in ESCC and it can be a novel predictor for the prognosis of the patients with ESCC [254]. Recent studies have shown that G6PD can act as an important regulator in ESCC development and progression by manipulating the STAT3 signaling pathway and can thus be an underlying molecular target for ESCC [255] patient therapy.

(i)***AKR1C2:*** The ARK1C2 gene is found to be altered in ESCC. It is indicated to be at least indirectly involved in esophageal carcinogenesis and is considered one of the most important molecular targets in the treatment of ESCC. This gene is also found to be altered in breast cancer and prostate cancer [256][257][258].

4.3.2.5 Discussion

Esophageal Squamous Cell Carcinoma (ESCC) is a highly aggressive form of esophageal cancer, especially common in regions like North-East India. An in-depth analysis is done to identify critical genes associated with ESCC using an ensemble approach, integrating data from multiple sources including microarray and RNA-seq datasets. The findings of this method shed light on several important aspects of ESCC pathogenesis and potential targets for precision treatment and prevention. In the analysis, nine genes (COL27A1, SOX11, BAG6, TOP3A, CDC6, EZH2, COL7A1, G6PD, and AKR1C2) are identified as significant candidates linked to ESCC. These genes play crucial roles in the progression of ESCC, either directly or indirectly. A notable part of our analysis involves examining the biological roles of these critical genes. They are found to be significantly involved in various aspects like Biological Process, Cellular Component, and Molecular Function, highlighting their importance in ESCC development. The supporting literature also strengthens the case for these genes as potential targets for further research and clinical use in ESCC and cancer in general. This analysis is not just about understanding the basics; it has real-world impacts. It gives us hope for developing new ways to diagnose and treat esophageal cancer (ESCC). Certain genes, like SOX11, COL27A1, and EZH2, could be focused on for personalized treatments in ESCC. This finding sets the stage for future research to uncover the detailed inner workings of ESCC at a molecular level. This, in turn, brings us closer to creating better ways to intervene and achieve better outcomes for patients.

Comparison with other competing methods

In terms of finding candidate genes and using DEA software, the performance of proposed method is compared with eight other competing methods. Table 4.26 shows the comparison. Unlike other methods, this method uses six different DEA tools to identify and unbiased set of DE genes by eliminating the biasness of the individual tools. Further, unlike most other tools, the proposed method validates the critical genes identified by this approach from multiple aspects such as topological, pathway-based and literature evidences.

4.4 Conclusion

In method I, RNA-seq data (GSE32424) for Esophageal Squamous Cell Carcinoma is analysed. DEA is performed to identify differentially expressed genes. Then, the topological behaviour of networks at normal and disease conditions are studied. Finally, gene enrichment analysis is done for validation of the identified genes. From differential expression analysis, topological analysis, and functional enrichment analysis from which set of genes are found to be associated with ESCC and these genes are considered as crucial genes for ESCC. These genes need further investigation to uncover their other activities while the disease is in progression. In this work, differentially expressed genes are extracted using only one method, but several popular methods are available to perform the same task and results different set of DEGs. Further, next work is based on an ensemble approach of DEA methods. Main goal is to develop an interesting framework to handle various sources of data at a time (e.g., microarray and RNA-seq) towards the identification of interesting biomarkers through consensus building for a given disease.

From the experimental study of Method II, four low preserved modules have been identified. Based on the topological and biological analysis of these modules it is found that several interesting genes found directly or indirectly associated with ESCC. Out of these, nine genes have been reported as the most significant based on their (i) topological weight to identify hub genes, (ii) biological enrichment analysis in terms of BP, CC, and MF, and (iii) literature evidences. These nine genes i.e. *COL27A1*, *SOX11*, *BAG6*, *TOP3A*, *CDC6*, *EZH2, COL7A1, G6PD,* and *AKR1C2* have been found to have significant enrichment either in ESCC or carcinogenesis in general. Based on the study of differential co-expression and differential expression analysis, these findings will make a significant contribution to future research aimed at characterizing the role of specific genes in ESCC pathogenesis and helping to improve diagnosis and treatment. These nine genes may provide potential targets for precision treatment and prevention of this deadly

Table 4.26: Comparison of existing methods to find critical genes in ESCC

Title	Year	Candidate genes	Used DEA tool
RNA-sequencing based identification of crucial genes for esophageal squamous cell carcinoma[224]	2015	CRISP2, CRISP3 and mucin 21	edgeR
Combination of meta-analysis and graph clustering to identify prognostic markers of ESCC[259]	2012	CXCL12, CYP2C9, TGM3, MAL, S100A9, EMP-1 and SPRR3	limma
Identification of crucial genes associated with esophageal squamous cell carcinoma by gene expression profile analysis[260]	2018	UBE2C, CDC20, MCM6, TFRC, TEAD4, RISP3, NELL2, PPP1R3C, MAL and CYP3A5	limma
Data mining of esophageal squamous cell carcinoma from The Cancer Genome Atlas database[261]	2018	TNFRSF10B and DDX18	edgeR
Comprehensive bioinformation analysis of methylated and differentially expressed genes in esophageal squamous cell carcinoma[220]	2019	NOTCH1, MMP9, IL8, IFNG, BCR, and SPP1	GEO2R tool
Ranking candidate genes of esophageal squamous cell carcinomas based on differentially expressed genes and the topological properties of the co-expression network[262]	2014	CRISP3, EREG, CXCR2, and CRNN	SAM
Identification of the Key Genes and Pathways in Esophageal Carcinoma[263]	2016	CDK4, CCT3, THSD4, SIM2, MYBL2, CENPF, CDCA3, and CDKN3	limma
The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma[264]	2019	COL1A1, COL10A1 and COL11A1	edgeR
Identifying Critical Genes in Esophageal Squamous CellCarcinoma using an Ensemble Approach	-	COL27A1, SOX11, BAG6, TOP3A, CDC6, EZH2, COL7A1, G6PD, and AKR1C2	edgeR, SAM, limma, limma-voom, DESeq2, EBAM

disease.

A gene co-expression similarity measure SNMRS is proposed to measure similarity between two genes and to handle different patterns of gene expression data. SNMRS is applied in finding network modules from weighted signed co-expression network towards identification of potential biomarkers.