Chapter 6

Identification of Potential Prognostic Biomarkers for ESCC using Single-cell RNA Sequencing Data Analysis

6.1 Introduction

ESCC is one of the main histological subtypes of Esophageal cancer. Several risk factors are associated with ESCC, although there is no specific ESCC biomarker available, which ultimately results in inadequate treatment modalities for patients suffering from ESCC. The advancement of Next Generation Sequencing (NGS), has made it possible to identify genomic alterations and gene expression changes that are associated with cancer development and pathogenesis. scRNA-seq enables researchers to identify rare subpopulations of cancer cells that are involved with cancer progression which helps in developing targeted therapeutics for cancer.

Biomarkers are molecules that can be used to identify a particular cell type or state. These molecules are helpful in the investigation of cancer cells, for example, to determine how healthy a cell is. Biomarkers can be identified by studying the proteins or genes that are expressed by a cell. Single-cell RNA sequencing data anlaysis has already been established to be useful to identify biomarkers that are specific to a particular cell type or state. This helps improve the accuracy of biomarker identification. This study reveals that scRNA-seq data analysis is helpful in analyzing gene expression patterns and uncovering biological insights. By leveraging the scRNA-seq data, one can gain a better understanding of gene expression patterns in different cell types and use this knowledge to develop more effective treatments for diseases. We analyse the ESCC scRNA-seq dataset using (i) a consensus function of differential expression analysis methods and (ii) a novel measure called SNMRS and its application in clustering. We identify four potential genes that have been found to have a close association with ESCC development.

6.2 Related Work

A lot of ongoing work on developing pipelines for identifying biomarkers in single cell data. Seurat is one of the most popular tools for this task, and there are a number of different ways to use it to identify biomarkers. One common approach is to use Seurat [293] to cluster cells into groups based on their gene expression profiles, and then to identify biomarkers that are differentially expressed between the clusters. This approach can be used to identify biomarkers that are associated with a particular disease or condition, or that are associated with different stages of a disease. Another common approach is to use Seurat to identify subpopulations of cells based on their gene expression profiles. This can be used to identify biomarkers that are associated with a particular subpopulation of cells, or that are associated with different stages of a disease. There are also a number of different ways to use Seurat to identify biomarkers that are associated with survival or progression of a disease. Each of these approaches has its own advantages and disadvantages, and there is still a lot of work to be done in order to determine which approach is the best for identifying biomarkers in a particular dataset.

Recently, there has been a growing interest in using single-cell technologies to identify novel biomarkers. Single-cell RNA sequencing (scRNA-seq) is a powerful tool for identifying gene transcripts within individual cells and can be used to detect changes in gene expression that may be indicative of disease or may predict progression or outcome [1]. Biomarker identification in scRNA-seq data begins with the selection of a set of DEGs that are significantly associated with a given phenotype. Once the differential expressions of the genes have been identified, the next step is to perform functional enrichment analysis to identify markers that are associated with specific biological pathways. The analysis of scRNA-seq data is challenging because of replication noise, sparsity in transcripts, and outlier cell populations [294]. Seth et al. [295] discuss the application of dimensionality reduction and hierarchical clustering for cluster-specific potential biomarker discovery in scRNA-seq data. The author has created a Seurat object to store data and analysis together for the dataset and using Louvain algorithm they have detected the cell clusters and identified cluster-specific biomarkers using DEA tool called Model-based Analysis of Single-cell Transcriptomics (MAST) [296]. They reported five hub genes. Also they discovered cluster-specific frequent biomarkers, i.e. overlapping biomarkers from single-cell RNA sequencing data and performed KEGG pathway and Gene Ontology enrichment analysis of the cluster markers. In this work it is observed that the authors have not used an imputed matrix and multiple DEA tools to eliminate biased DEGs. Wang et al. [297] identify novel biomarkers for diabetic kidney disease by combining scRNA-seq and bulk RNA-seq data. The authors use the Seurat package and identify key biomarkers by the method of MCODE. Cui et al. [298] establish a framework for analyzing scRNA-seq data, including DEA, differential correlation analysis, network analysis (WGCNA), and differential network analysis. Algabri et al. [294] present a framework based on co-expression network analysis for scRNA-seq data. They used Seurats MAST method to perform a DEA and WGCNA package [78] for

[1] https://www.biomage.net/blog/biomarker-discovery-using-scrnaseq

GCN analysis. From the above discussion, it is evident that the Seurat framework with some variations has been successfully used in scRNA-seq gene data analysis, towards the identification of crucial genes for a given disease.

6.3 Materials and Methods

6.3.1 Datasets used

A benchmark scRNA-seq ESCC dataset of size 18938 Genes x 314 cells with accession number: GSE81812 downloaded from cancerSEA[2] has been used. scRNA-seq data obtained from KYSE-180 cells before exposure to fractionated irradiation serves as a control (0-Gy) and post-fractionated irradiation data are obtained after cumulative doses of 12-Gy and 30-Gy.

6.3.2 Proposed framework

Figure 6-1 depicts the framework used to (i) compare the behaviour of genes in different sample types and (ii) identify crucial genes associated with the disease ESCC. First, the dataset is pre-processed and Ensemble Ids present in the dataset are converted into Gene Symbols. Only distinct Symbols are kept and we removed "NA" symbols, doublet cells and inactive genes from the dataset. Metadata is collected and cell information are combined with the dataset. Next, with the filtered genes and cells, a Seurat object is created for subsequent downstream analysis. We also filter cells based on the percentage of mitochondrial genes present and perform quality control analysis to discard low-quality cells. Next, the global-scaling normalization method is applied for data normalization.

Highly variable genes are identified and performed scaling and dimensionality reduction. Next step deals with construction of nearest neighbour graph to find all markers. We identify the clusters using the Louvain algorithm that match the cell types of data. After clustering, differentially expressed genes are identified

[2] http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp

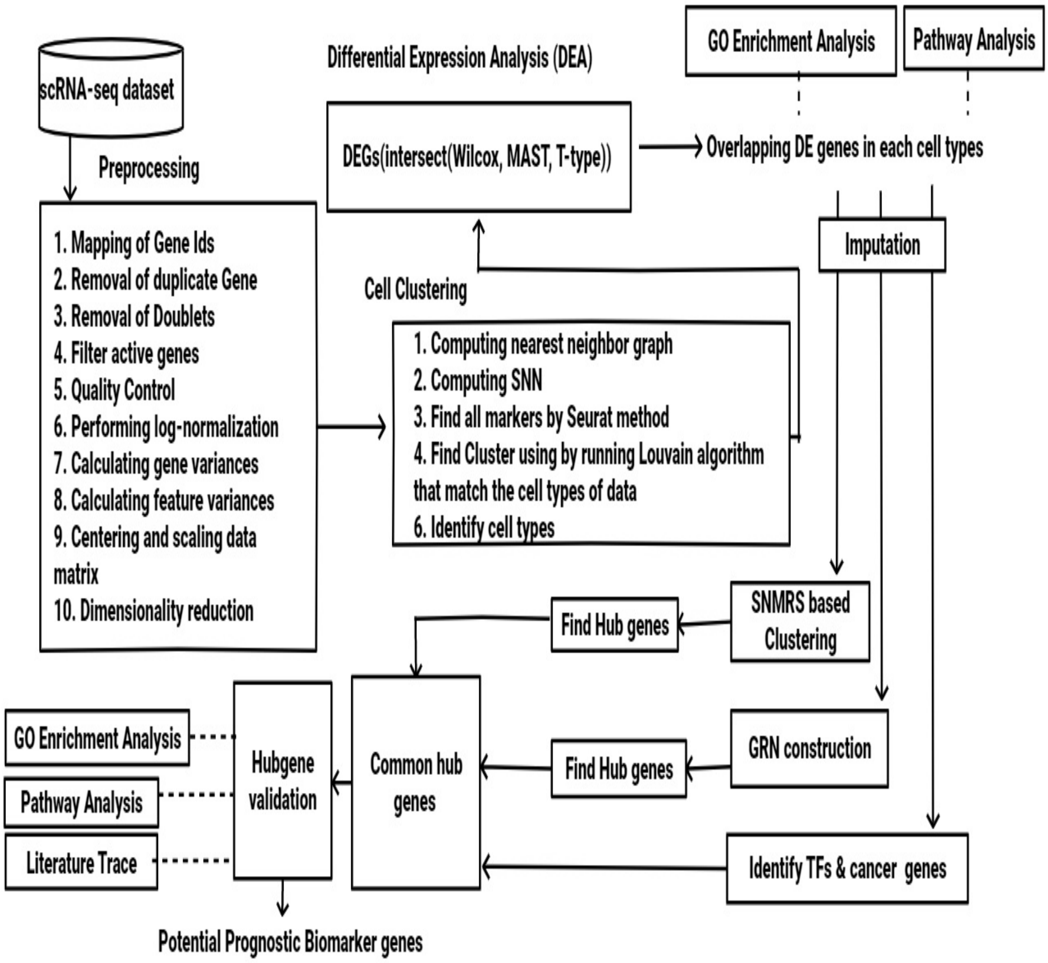

scRNA-seq dataset
Preprocessing
1. Mapping of Gene Ids
2. Removal of duplicate Gene
3. Removal of Doublets
4. Filter active genes
5. Quality Control
6. Performing log-normalization
7. Calculating gene variances
8. Calculating feature variances
9. Centering and scaling data matrix
10. Dimensionality reduction
Differential Expression Analysis (DEA)
DEGs(intersect(Wilcox, MAST, T-type))
Overlapping DE genes in each cell types
GO Enrichment Analysis
Pathway Analysis
Cell Clustering
1. Computing nearest neighbor graph
2. Computing SNN
3. Find all markers by Seurat method
4. Find Cluster using by running Louvain algorithm that match the cell types of data
6. Identify cell types
Imputation
SNMRS based Clustering
Find Hub genes
GRN construction
Find Hub genes
Identify TFs & cancer genes
Common hub genes
Hubgene validation
GO Enrichment Analysis
Pathway Analysis
Literature Trace
Potential Prognostic Biomarker genes

Figure 6-1: Framework of the proposed method.

for each sample type i.e. 0 Gy (Control), 12 Gy (ESCC) and 30 Gy (ESCC) individually using three different DEA methods i.e., Wilcox [299], MAST [296], and t-test [300]. MAST (Model-based Analysis of Single-cell Transcriptomics) [296] is a DEA method which can identify transcriptional changes and characterize the heterogeneity of the single cell data. It is based on a probabilistic model of gene expression that allows for the estimation of expression levels for each gene and the identification of differentially expressed genes. MAST can be used to compare expression profiles between different cells or conditions, or to identify genes that are differentially expressed in a particular cell or condition. T-test [300] is a general-purpose method, used to compare the mean expressions of genes across two cellular groups. scRNA-seq data are highly (positively) skewed, but the T-test DEA method is known to have a certain robustness against skewness.

The Wilcox method [299] is an NP method used to test whether a genes mean expressions across the two cell groups is significantly different or not. This method is based on the Mann-Whitney U statistic. The main idea of the Wilcox method is to compare the ranks of the expression values that come from the two cell groups. This rank-based test mostly ignores the magnitude of the expression of deviations of genes between the two cell groups. The following consensus method of DEAs is applied in our approach. A consensus of DEA is performed independently to avoid biases for each sample type. A set of DEGs considered as common DEGs for 0-Gy sample type w.r.t. a user-defined threshold α=logfc.threshold=0.25 (default) which are identified using mentioned three different methods. Similarly, common DEGs are obtained for 12-Gy and 30-Gy.

With these DEGs, Gene Regulatory Networks are constructed individually for each sample type and identify already established transcription factors and ESCC cancer genes. GO enrichment analysis and KEGG pathways analyses are carried out for all these selected DEGs and if found significantly enriched in GO and pathways then we consider them for downstream analysis. In order to reduce the effect of noise and dropout events, an imputation method SAVER[31] has been used in the datasets with DEGs for each sample type individually to impute the missing gene expression values. Co-expression networks are generated from imputed datasets of DEGs for each sample type by computing a measure of co-expression between pairs of genes. Here, SNMRS measure [301] is applied to compute co-expression between genes and extracted modules using the hierarchical clustering method for each sample type. Also, a list of hub genes is extracted from these modules based on degree information and considering those genes having degree values$\geq$mean(degree). From GRNs, a list of the top five hub genes or central genes is identified for each sub-network in different sample types and accordingly, a list of hub genes has been identified. Next, the common hub genes from GCN modules and GRN are matched with the transcription factors and if found valid, consider them as the crucial hub genes for ESCC. After biological validation and literature trace, the identified hub genes are further tested towards possible acceptance as the potential biomarker genes for ESCC.

Although the proposed work is done based on the Seurat framework, however, our approach differs from the framework in the following ways.

(1) During DEG identification it exploits three different techniques such as Wilcox, MAST, and t-Test and to eliminate the individual biases it applies —-operation on the findings of these three to obtain the final set of DEGs.

(2) Once the DEGs are identified the rest of downstream analysis towards validated prognostic biomarker identification is not included in the Seurat framework.

6.4 Results

In this section results obtained during our analysis are presented.

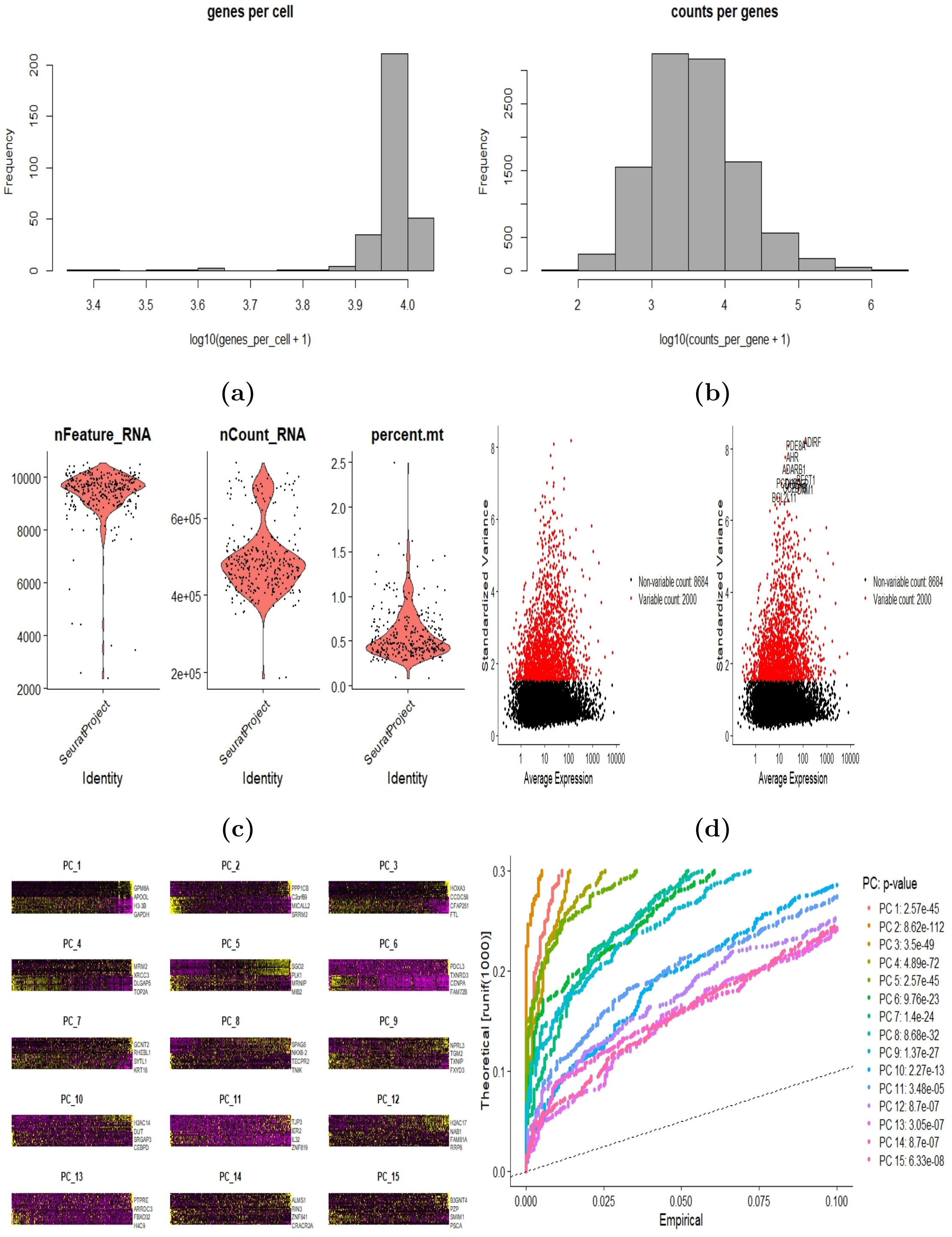
6.4.1 Preprocessing

Raw ESCC tpm matrix data of size 18938 genes and 314 cells is pre-processed for downstream analysis. First, ensemble Ids are mapped into gene symbols and removed 39 duplicate genes and 1 ”NA” gene. Next, doublet cells are removed and cell size is reduced to 309. The next step in the analysis is to identify the genes that are active or inactive based on their zFPKM values. Genes that have a median of zFPKM value>-3 are the only active genes and now 10684 selected filtered genes are present in the dataset. Since this data contains all the genes with non-doublet cells, we need to select those with median values greater than -3. The Seurat object is initialized with the raw (non-normalized) data. Now, the size of the data becomes 10684 genes x 309 cells. The percentage of mitochondrial genes present in each cell has been calculated and plot this in terms of a violin plot in Figure 6-2c. Cells having (1) greater than 200 genes are filtered out and (ii) cells having less than 2500 genes and keep only those cells that have mitochondrial percentage less than 5%. Now the size of the resulting dataset is 10684 genes x 255 cells. To compare the gene expression profiles across multiple cells we need to normalize the data. Next, performed quality control analysis to discard low-quality cells and the global-scaling normalization method is applied for data normalization. Highly variable genes are identified and performed scaling and dimensionality reduction. By finding a subset of genes that exhibit high cell-to-cell variation in the dataset (i.e., they are highly expressed in some cells, and lowly expressed in others) helps in principal component analysis to highlight biological signal in our ESCC dataset. Figure 6-2 visualizes measurement of library size, QC metrics as a violin plot, highly variable genes and DimHeatMap for PC1-PC15 respectively. Next, a linear transformation, also known as "scaling," is performed which is a typical pre-processing step before dimensionality reduction methods like PCA. Each gene's expression is changed via the ScaleData function, resulting in a mean expression across cells of 0 scales of each gene's expression so that the variation between cells is 1. In order to prevent highly expressed genes from dominating, this step assigns equal weight to downstream analyses.

6.4.2 Cell Clustering and finding DEGs

In the next step, the nearest neighbour graph and SNNs are computed to find all potential markers. The clusters are identified by running Louvain algorithm [302] that matches the cell types of data. Metadata is combined to the Seurat object to extract genes for each cell type. Visualization of clusters are presented in Figure 6-3.

After clustering, differentially expressed genes are identified for each sample type i.e. 0 Gy (Control), 12 Gy (ESCC) and 30 Gy (ESCC). A consensus of differential expression analysis method is performed to avoid biases. For this, three DEA methods i.e., Wilcox, MAST, and t-type are used. A total of 138 DEGs from 0-Gy, 289 DEGs from 12-Gy, and 763 DEGs from 30-Gy were obtained and venn diagrams are shown in Figure 6-4. SAVER is used to impute the missing gene expression values in the sub-matrix of DEGs. With these DEGs, Gene Regulatory



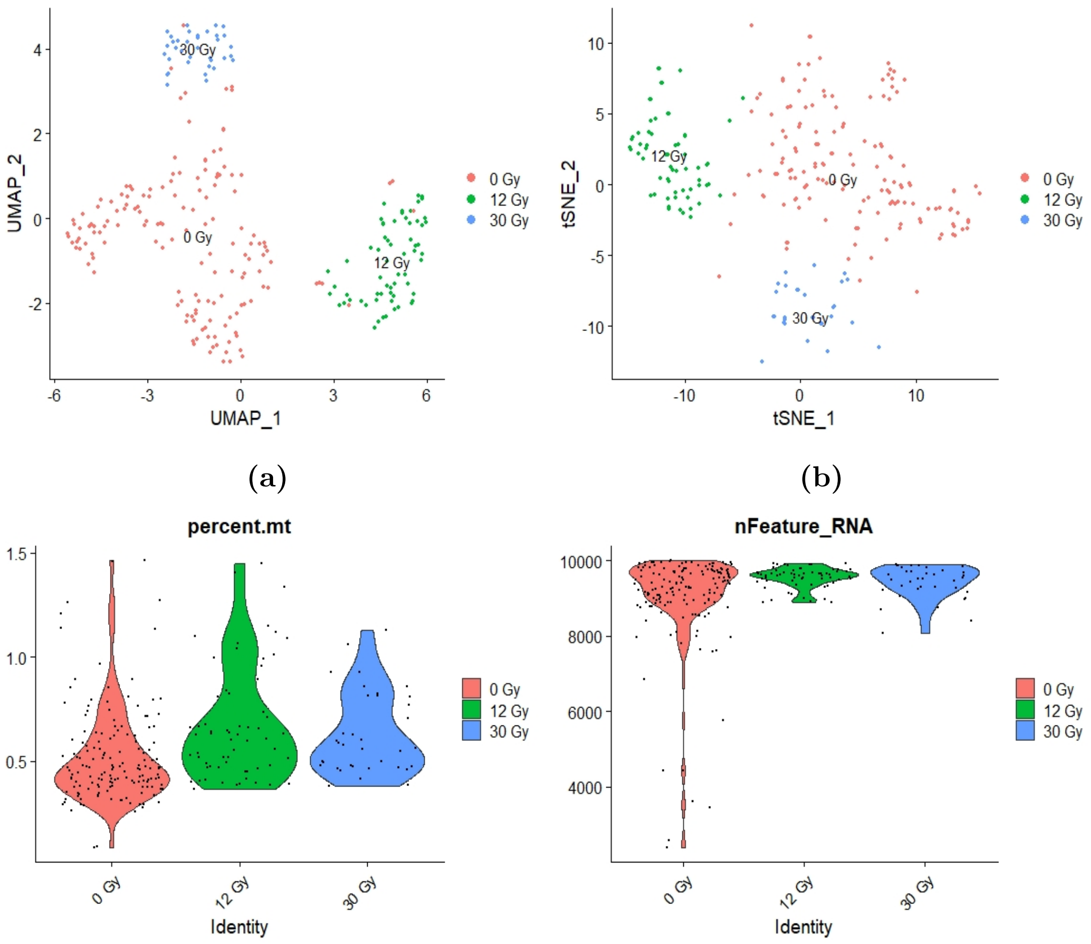
(a) (b)

(c) (d)

(e) (f)

Figure 6-2: Preprocessing of ESCC data. (a) Measurement of library size for genes per cell, (b) Measurement of library size for counts per genes, (c) Quality control, (d) Variable genes in ESCC dataset, (e) DimHeatMap for fifteen principal components, (f) Visualization of strong enrichment of features with low p-values.

Networks are constructed individually for each sample type and identified already established transcription factors and ESCC cancer genes. It has been found that



(a)

(b)

(c)

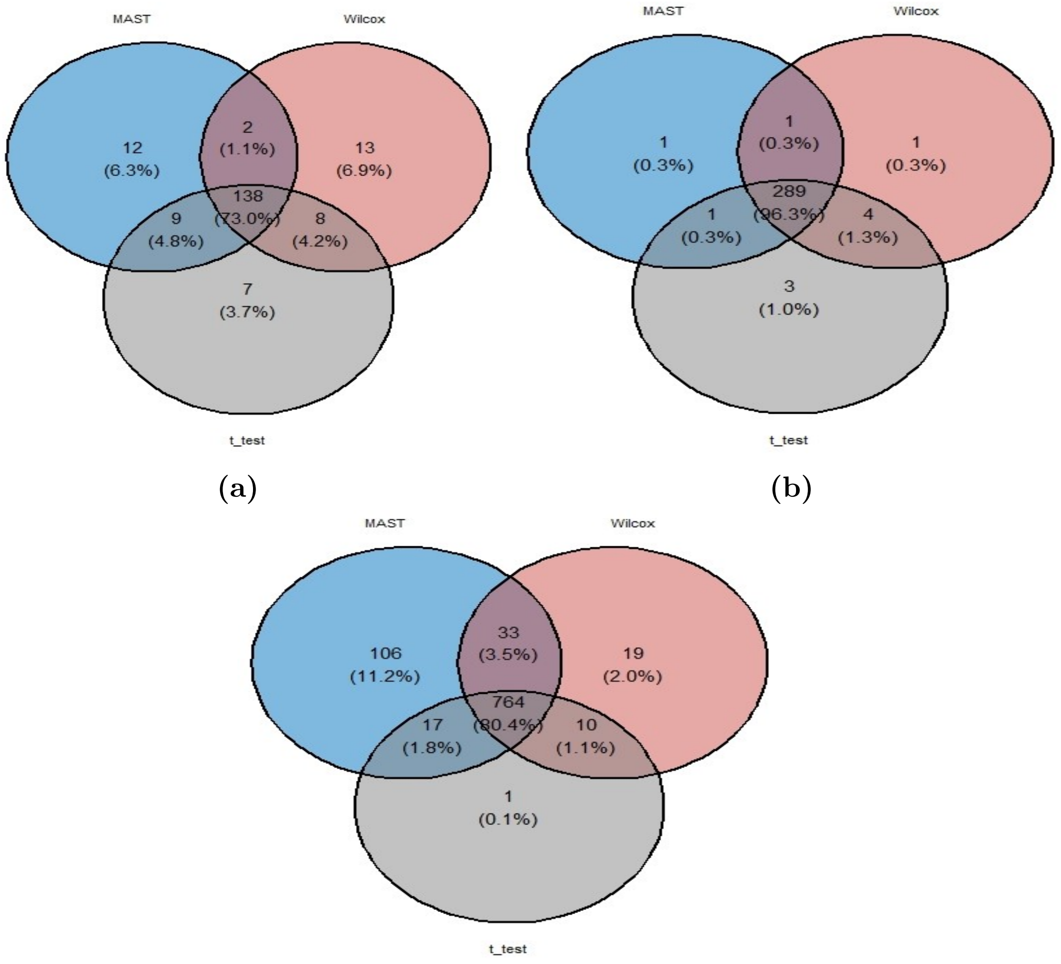
(d)

Figure 6-3: (a) UMAP clustering of the dataset, (b) t-SNE clustering of the dataset, (c) Genes distributed based on the percentage of mitochondria, (d) Gene clustering and distribution in all sample-type.

in the 30-Gy type, the number of identified transcription factors and ESCC cancer genes are comparatively more than the 0-Gy type. All the 1363 DEGs were significantly enriched in DNA replication, cell cycle, RNA transport, and oocyte meiosis pathways as shown in Figures 6-5 and 6-6. Information regarding the extracted differentially expressed genes (DEGs) can be found in Table 6.1.

6.4.3 Finding Hub gene

A hub gene can be defined as the central gene to a network which has comparatively more connectivity in the case of an undirected graph and in the case of a directed graph a node which has more outgoing edges. We know that a hub gene



(a)

(b)

(c)

Figure 6-4: Venn diagram of differentially expressed genes found using consensus method in sample (a) 0-Gy, (b) 12-Gy, (c) 30-Gy.

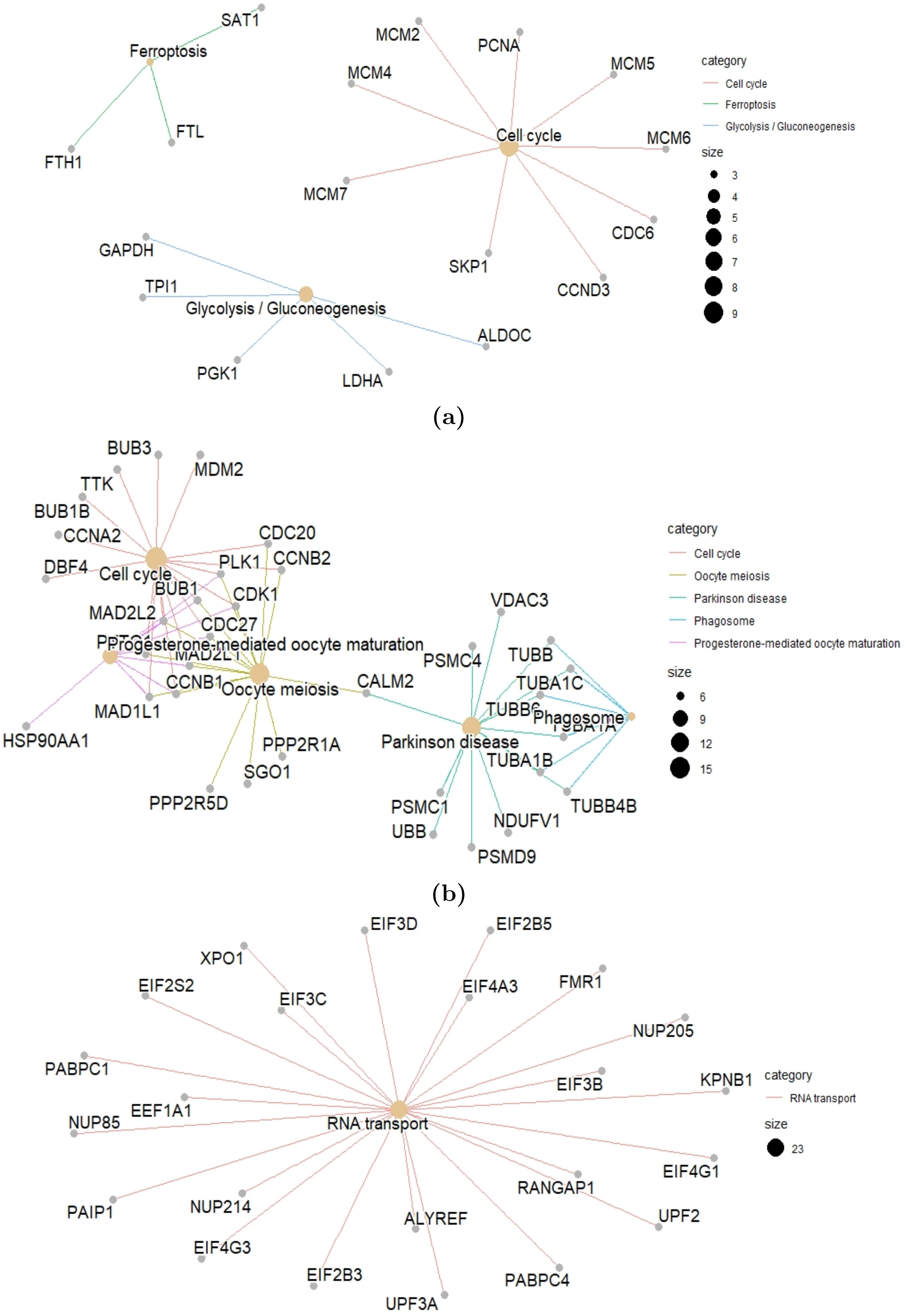
always has a significant role in disease progression and investigating these genes might provide a clue related to disease or dug discovery to researchers. Here, we aim to find hub genes from different angles such as from co-expression networks, gene regulatory networks, and transcription factor lists.

(a) Constructing gene co-expression networks (GCN): GCNs are generated from gene expression data by computing a measure of co-expression between pairs of genes. Here, SNMRS measure [301] is applied to compute co-expression between genes and found modules using the hierarchical clustering method for each sample type. The results obtained from the analysis of GCN are presented in Table 6.2. Also, a list of hub genes is extracted using degree and taking genes above mean(degree). For 0-Gy, TXN, LAMTOR5, B2M, PDIA3, SRSF7,

Table 6.1: DEG Details

Sample Type	**0 Gy**	**12 Gy**	**30 Gy**
DEGs count	138	289	763
No. of Modules	3	3	6
TF	MCM6, MCM5, PCNA, MCM2, CDC6, MCM4, CNBP, IER2, THAP7, RBBP4, MCM7, LITAF, PSMC3IP, UBE2N, TXN, HMGN4, APEX1, SUMO1, EPAS1, SQSTM1, CRABP2, FOS	DEPDC1, CCNA2, CENPF, HMGB2, PTTG1, DEPDC1B, CENPA, TRIP13, BRD8, HP1BP3, PHF19, EHF, TGIF1, MORF4L2, SMARCB1, DYNLL1, PTBP1, MIS18BP1, GTF2A2, HMGB3, SNAPC2, HMGN2, MDM2, PPP2R1A, TMEM175, DAZAP2, VPS72, ECSIT, PRMT5, DUS3L, PNRC2, RBBP7, CERS2	BRD4, HNRNPA3, RPL7A, PABPN1, SLC2A4RG, HDGF, DNAJC21, SFPQ, FUS, MTDH, CEBPB, SSRP1, PCBP2, IRX2, KMT2A, EIF3C, PHF14, JUND, MBD6, EP400, HNRNPAB, UBE3A, THAP4, CEBPZ, PRPF4B, CITED4, TRIM33, YBX1, PPP1R13L, CHAF1A, BRD7, BAZ1B, LRPPRC, YY1, TCERG1, HNRNPUL1, PHF20L1, SART1, MPHOSPH8, NFAT5, NR2F6, PTMA, PAWR, DVL1, INF2, NKX6-2, HNRNPD, SMARCA4, RBM6, BPTF, FUBP1, CCAR1, PDS5B, SAFB2, SMARCA5, SUGP2, CCDC88A, CHD2, GADD45A, PSIP1, CBX3, NSD2, SSBP4, MBD2, ZNRF1, FAM189B, TAF1D, FBXW7, SMARCC1, PCBP1, FMR1, TRIOBP, PCGF2, TRIM56, ZNF493, SUPT5H, ZNF638, PRDM2, GTF3C1, FOXD1, RNF166, SRRM1, ZMAT2, BCLAF1, ALYREF, KHSRP, NFATC4, CREBBP, HMX1, CDKN2A, MAPK3, MEIS3, PLXND1, CALR, HNRNPL, ZFAND3, SON, BAZ2A, CTDSPL, YAP1, LARP1, DNMT1, YBX3, RFXANK, TARBP1, LRRFIP1, CHD8, RUFY3, EHMT2, ZC3H11A, ASXL1, GTF2IRD1, ZC3H7A, HNRNPU, CBX1, RFC1, SSH3, RC3H2, KDM5A, UBR4, ZNF317, HES1, ZNF395, GON4L, EHMT1, WWOX, RSF1, TRIP4, GTF2IRD2, RIN3, CCDC85B, SAMD4B, SIRT7, BAZ1A, CHD7, CAPN15, TBL1XR1, TNRC6A, CHD9, CCT4, SF1, ZMYND8, ATF4, HMGXB4, GTF2I, TIAL1, TOP2B, MYSM1, TCF3, MNAT1, AGAP3, PRKDC, SPEN, TULP4, ANP32A, FLNA, ATM, ZRANB2, RERE, ZFP36L2, CUL3, WBP11, MORF4L1, SRSF9, AHCTF1, SMARCD1, NME2, CHD4, CIZ1, PKN1, HP1BP3, PSMC5, NR1H3, KDM1A, JMJD1C, EEF1A1, TCF25, MECP2, BRD2
Cancer Gene	MSH6, EPAS1, PPP1CB	BUB1B, TRIP13, GTSE1, SMARCB1, MKI67, LMNA, FANCD2, PPP2R1A, NUP107, PC, HSP90AA1, SS18	FUS, KMT2A, MAP2K2, RTN4, TRIM33, RPL5, DCTN1, COP1, RECQL4, SMARCA4, BPTF, FUBP1, AFDN, PDS5B, NF1, KTN1, CHD2, NSD2, FBXW7, PCBP1, XPO1, BCLAF1, CREBBP, CDKN2A, DNMT1, SETD5, ZC3H11A, ASXL1, DROSHA, SPTAN1, FGFR2, EIF4G1, MYH9, TRIO, PRKDC, SPEN, ATM, MKI67, ZFP36L2, CUL3, TPR, LRP5, ANKRD11, CHD4, HLA-B, EEF1A1

ACP1, SUMO1, KDELR2, PGK1, ITM2B, and CAPZA1 hub genes are identified from the brown module. Hub genes for turquoise module of 0-Gy are IER2, TPI1, EIF3I, SKP1, UBE2N, MRPL17, CD55, POLD4, FTH1, EBP, CTSB, OAZ2, PLD3, RHOC, LGALS3BP, DUSP1, SH3BGRL3, TMED10, CNBP, CFL1, GPX3, LGALS3, CD164, CNIH1, CST3, S100A14, TMBIM6, AKR1B1 PRDX2,

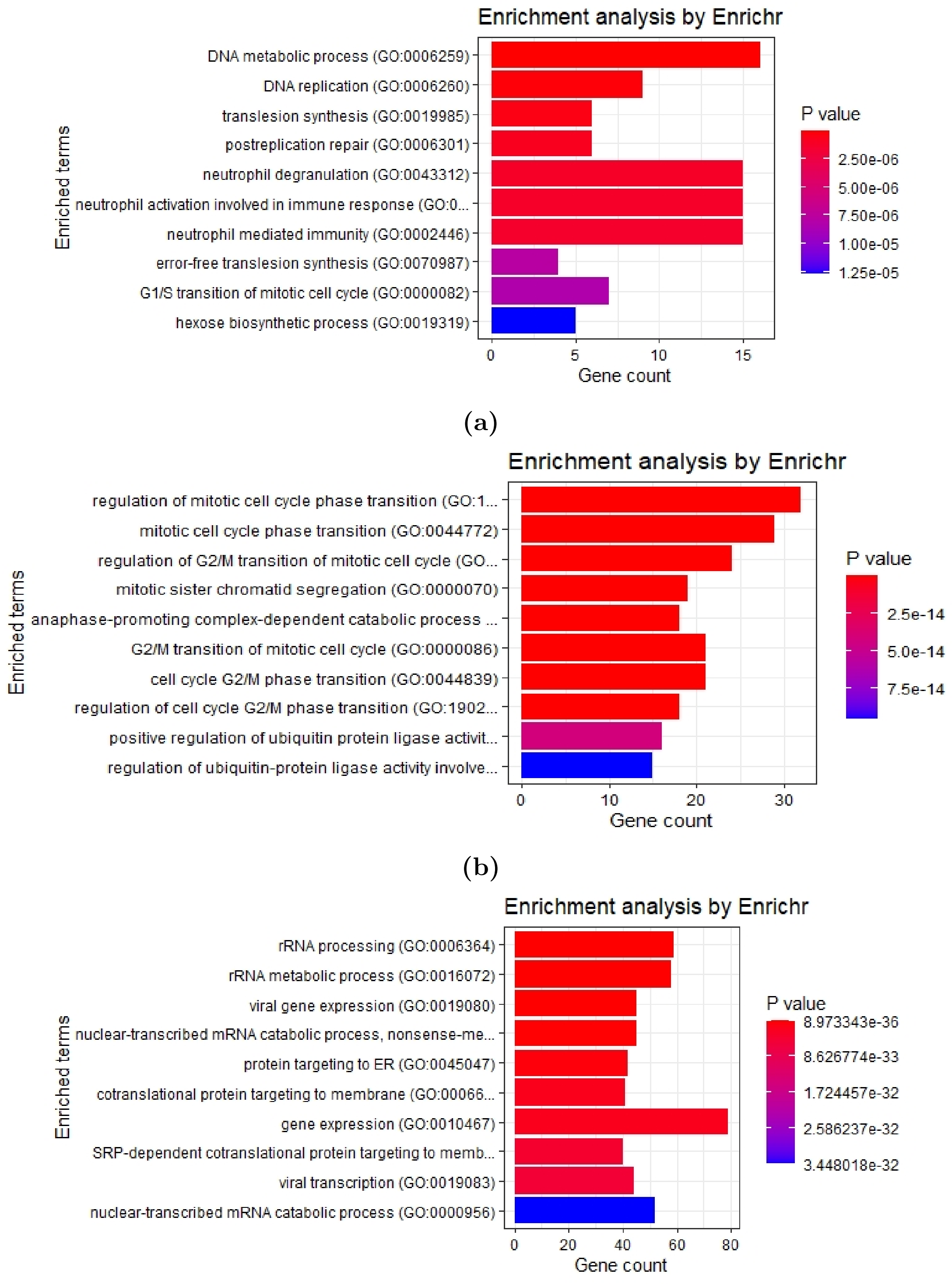


(a)

(b)

(c)

Figure 6-5: Network diagram for KEGG for all the DEGs in (a) O-Gy (b) 12-Gy (c) 30-Gy



(a)

(b)

(c)

Figure 6-6: Enrichment analysis of DEGs in (a) O-Gy (b) 12-Gy (c) 30-Gy.

TIMP1, ELOF1, FTL, GAPDH, and CSNK2B. Hub genes for blue module of 0-Gy are MCM4, MCM2, PCNA, CDC6, THAP7, RFC4, GINS2, PPIF, WDR76, DTL, FAM111B, LIPH, GMNN, RFC2, SNX8, and TM2D2. Hub genes identified for the turquoise module of 12-Gy are KRT8, FBLN1, MDH2, TAGLN2, and SPINT2. For blue module of 12-Gy CCNB1, CKS2, DEPDC1, FAM83D, TOP2A, PLK1, AURKB, NEK2, CCNA2, CENPF, TPX2, CDC20, SPAG5, HMMR, CDKN3, CDCA8, DLGAP5, and KPNA2 are detected as the hub genes. For the brown module of 12-Gy identified hub genes are UBB, EPS8, ATP5F1B, and MT1X. For blue module of 30-Gy RPL22, RPS25, RPL37A, for red module TOP2A, H2AX, MKI67, and PRC1, for turquoise module CHCHD10, LDLRAD2, BASP1, STUB1, and FAM83H, and for yellow module DDX56, and PRPF38B genes are identified as the hub genes. It has been observed that the number of modules extracted is comparatively more in the case of the 30-Gy sample type.

Table 6.2: Details of GCN Analysis

Type	**Module Name**	**Hub Gene from GCN Modules**
0-Gy	brown	TXN, LAMTOR5, B2M, PDIA3, SRSF7, ACP1, SUMO1, KDELR2, PGK1, ITM2B, CAPZA1
	turquoise	IER2, TPI1, EIF3I, SKP1, UBE2N, MRPL17, CD55, POLD4, FTH1, EBP, CTSB, OAZ2, PLD3, RHOC, LGALS3BP, DUSP1, SH3BGRL3, TMED10, CNBP, CFL1, GPX3, LGALS3, CD164, CNIH1, CST3, S100A14, TMBIM6, AKR1B1 PRDX2, TIMP1, ELOF1, FTL, GAPDH, CSNK2B
	blue	MCM4, MCM2, PCNA, CDC6, THAP7, RFC4, GINS2, PPIF, WDR76, DTL, FAM111B, LIPH, GMNN, RFC2, SNX8, TM2D2
12-Gy	turquoise	KRT8, FBLN1, MDH2, TAGLN2, SPINT2
	blue	CCNB1, CKS2, DEPDC1, FAM83D, TOP2A, PLK1, AURKB, NEK2, CCNA2, CENPF, TPX2, CDC20, SPAG5, HMMR, CDKN3, CDCA8, DLGAP5, KPNA2
	brown	UBB, EPS8, ATP5F1B, MT1X
30-Gy	blue	RPL22, RPS25, RPL37A
	brown	NA since degrees are equal
	green	NA since degrees are equal
	red	TOP2A, H2AX, MKI67, PRC1
	turquoise	CHCHD10, LDLRAD2, BASP1, STUB1, FAM83H
	yellow	DDX56, PRPF38B

(b) Finding Transcription Factors (TFs) and cancer genes: TFs are the key players in regulatory network interactions. Based on human transcriptional regulatory factors extracted from the HumanTFDB database[3] and TFCheckpoint, a

[3] bioinfo.life.hust.edu.cn/HumanTFDB/

total of 22, 33, and 179 TFs were detected in DEGs of 0-Gy, 12-Gy, and 30-Gy samples, respectively. Further, a list of ESCC cancer driver genes is downloaded from cBioportal [4] and intOGen[5] and found a total of 11, and 36 cancer genes in DEGs of 12-Gy, and 30-Gy samples, respectively.

(c) Constructing gene regulatory network (GRN): GRN is used to obtain information regarding the regulatory interactions between transcriptional regulators and their target genes. We used the significant DEGs as input to construct the regulatory network so as to observe the regulatory behaviour of the corresponding genes. GENIE3 [303] R package is applied for the prediction of gene regulatory networks from gene expression data and it uses the random forest or Extra-Trees approach. The resulting GRNs are in the form of an adjacency list with weighted directed edges from regulatory genes to other target genes. In Cytoscape tool [304], GRNs are plotted and by applying cytoHubba tool [305] a list of the top five hub genes or central genes which act as regulators are found for each sub-network in the different sample types. It is seen from the GRN (Figure 6-7) that PCNA, FOS, TXN, IER2, and MCM6 are the hub gene regulators extracted from sub-networks of 0-Gy GRN. CENPA, CENPF, BRD8, HMGB2, and TTK are the hub gene regulators extracted from sub-networks of 12-Gy GRN. NME2, HNRNPAB, SLC2A4RG, RPS15A, and ASXL1 are the hub gene regulators extracted from sub-networks of 30-Gy GRN.

(d) Selection of hub genes: From a different perspective, we found a good number of hub genes. Therefore, the common hub genes from GCN modules and GRN are matched with the transcription factors and if found then considered as the most crucial genes in ESCC. After biological validation and literature trace TXN, IER2, PCNA, and CENPF genes are considered as the prognostic biomarkers for ESCC.

(e) Biological analysis of hub genes: Biological analysis (Biological Process, Cellular Component and Molecular Function) are performed for each crucial gene

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

[5] https://www.intogen.org/

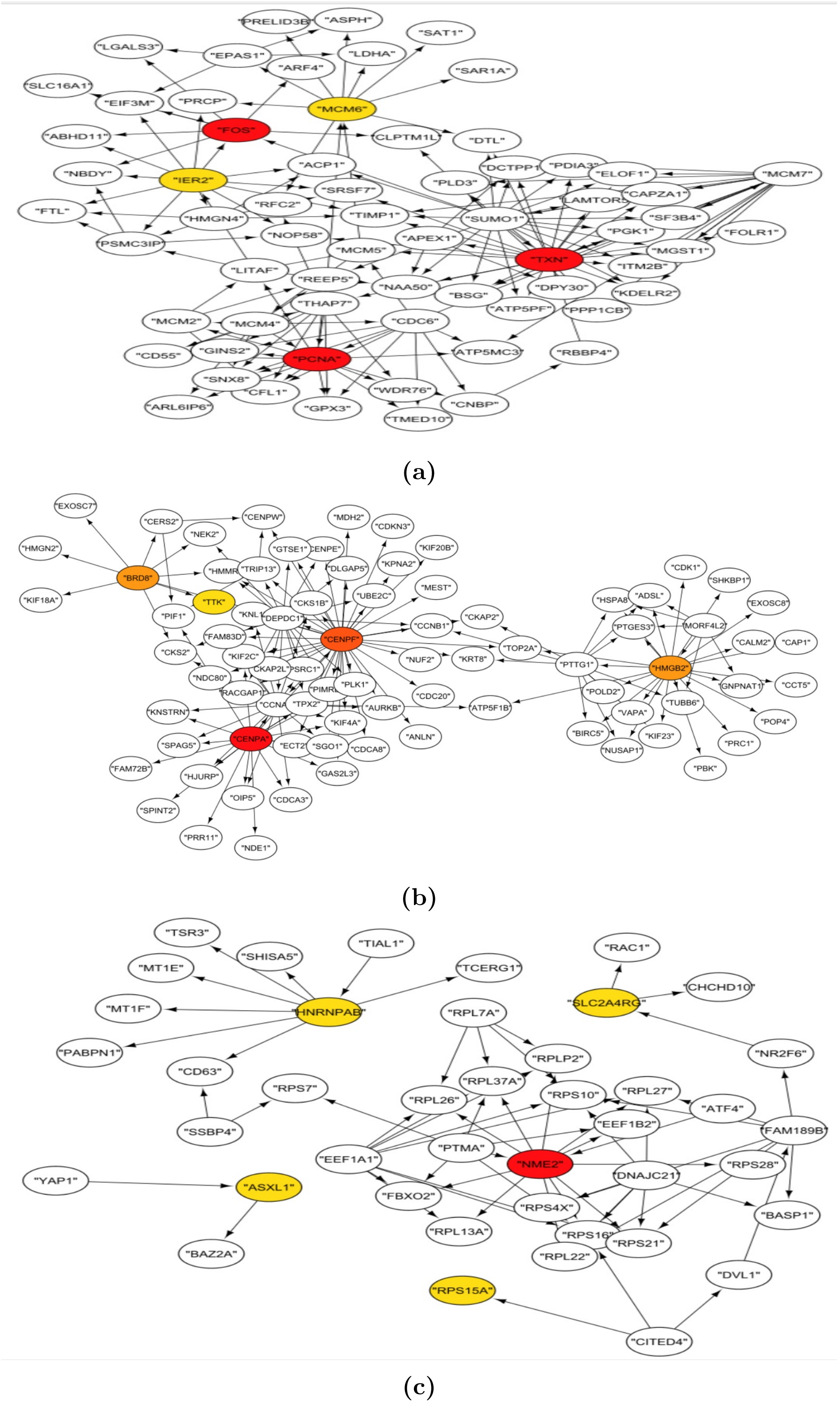


Figure 6-7: GRN and hub genes of their subnetworks obtained by cytoHubba for (a) 0-Gy (b) 12-Gy (c) 30-Gy

in DAVID[6] and reported in Table 6.3. The GO enrichment term nucleoplasm, Acetylation, and nucleus are associated with genes CENPF, PCNA, TXN, and IER2 genes. Protein C-terminus binding, chromatin binding, and centrosome GO terms are associated with CENPF and PCNA genes.

Table 6.3: Biological analysis of selected hub genes

Category	Term	PValue	Genes
GOTERM_CC_DIRECT	GO:0005654~nucleoplasm	7.1E-03	CENPF, PCNA, TXN, IER2
UP_KW_PTM	KW-0007~Acetylation	1.5E-02	CENPF, PCNA, TXN, IER2
GOTERM_CC_DIRECT	GO:0005634~nucleus	2.5E-02	CENPF, PCNA, TXN, IER2
GOTERM_MF_DIRECT	GO:0008022~protein C-terminus binding	3.3E-02	CENPF, PCNA
UP KW CELLULAR COMPONENT	KW-0539~Nucleus	3.3E-02	CENPF, PCNA, TXN, IER2
GOTERM_MF_DIRECT	GO:0003682~chromatin binding	7.3E-02	CENPF, PCNA
GOTERM_CC_DIRECT	GO:0005813~centrosome	7.9E-02	CENPF, PCNA

6.4.4 Literature Trace

From the literature of established wet lab results, the following observations have been made.

The protein TXN helps catalyze dithiol-disulfide exchange reactions and participates in various redox reactions [306]. Abdo et al. [306] found that TXN was expressed 2.07x less in EAC tumors compared to normal esophageal tissue. IER2 is also known to play a role in promoting tumour motility and metastasis and is a potential prognostic and therapeutic target [307]. Talukder et al.[308] performed Pearson correlation network link analysis for ESCC and found one of the most linked genes in the tumour network was IER2. PCNA is a nuclear protein that is known to be present in proliferating cells, including normal proliferating cells and cancer cells [309]. The expression level of PCNA has been found to be appreciably correlated with the TNM stage of ESCC patients. CENPF is found

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

upregulated in ESCC [310]. The expression of CENPE is associated with DNA methylation status in esophageal adenocarcinoma and is an independent predictor of unfavourable overall survival [311].

6.5 Discussion

Exploring the potential biomarkers underlying ESCC development would be of considerable benefit for prognosis prediction. In this study, differentially expressed genes are identified using a consensus model. Our experiment showed significant changes after receiving a 12Gy or 30Gy cumulative dose. We analyzed how the diversity of individual KYSE-180 cells changed in response to an FIR stimulus of 12Gy and 30Gy cumulative radiation doses. The detailed heterogeneity of ESCC cells after irradiation at the single-cell-transcriptome level is studied. This study revealed that gene expression of single cells was highly varied.

The scRNA-seq dataset used here is GSE81812. scRNA-seq technology counts the number of RNA molecules present in each cell of a given sample. The radiotherapy is one of the most useful method to destroy the cancer cells or to resist the cells against cancer. The most common treatment regimen used in the dataset ESCC consists of a standard fractionated irradiation, i.e., 1.8-12 Gy and 1.8-30 Gy. A total of 138 DEGs from 0-Gy, 289 DEGs from 12-Gy, and 763 DEGs from 30-Gy are obtained. It is found that in 30-Gy type the number of identified transcription factors and ESCC cancer genes are comparatively more than control type. All these 1363 DEGs are significantly enriched in cell cycle, DNA replication, RNA transport, and oocyte meiosis pathways. It is observed that number of modules extracted are comparatively more in case of 30-Gy sample type. Also, a list of hub genes are extracted using degree and taking genes above mean(degree). Again, in Cytoscape tool GRNs are plotted and by applying cytoHubba tool a list of top five hubgenes or central genes which act as regulators are found for each sub network in the different sample types. From different perspective, a good number of hub genes are found. Therefore, the common hub genes from GCN modules and GRN are matched with the transcription factors and if found then considered as the most crucial genes in ESCC. After biological validation and literature trace TXN, IER2, PCNA, and CENPF genes are considered as the most crucial genes associated with ESCC progression.

6.6 Conclusion

In this piece of work, an effective framework is introduced to find highly similar patterns containing genes with high biological relevance. In this work, a consensus model is used to identify differentially expressed genes with high biological relevance in ESCC development. This study shows significant changes in gene expression after receiving a 12-Gy or 30-Gy cumulative radiation dose. This study suggests that TXN, IER2, PCNA, and CENPF are potential prognostic biomarkers for ESCC diagnosis and can be considered as potential therapeutic targets for ESCC.