Chapter IV

Understanding the gallbladder cancer transcriptome and identifying potential genes, processes, and pathways associated with GBC pathogenesis

4.1 Introduction

The delayed diagnosis is one of the critical reasons associated with poor survival outcomes for GBC patients with a five-year overall survival rate of less than 5% [1]. Ultrasound and computed tomography methods are generally used for screening of GB mass [2]. Even after chemo and radiotherapy, GBC has the lowest survival rates of all BTCs [3-4]. Gemcitabine, the first-line approved treatment has minimal effects on patients with locally advanced or metastatic GBC [5], and no molecular-based treatment options specific to GBC have yet been approved. With the advent of precision medicine, transcriptome sequencing is becoming more widely used in cancer research. It is effective in determining the underlying cause of pathological origin, identification of biomarkers, and unraveling novel mechanisms in cancer pathogenesis [6]. Transcriptomic studies using next-generation sequencing methods have identified mutations or altered gene expressions in signaling pathways that could lead to the progression of GBC. For example, a recent study found recurrent ELF3 and WNT pathway changes in Indian and Korean GBC patients [7]. TP53 mutations are the most common among these driver mutations in GB carcinogenesis, while KRAS mutations are the most common in other types of BTCs. Exome and RNAseq data analysis revealed alternation in the KEAP1/NFE2L2 and WNT pathways. Furthermore, the exome and RNAseq data have revealed repetitive changes in the KEAP1/NFE2L2 and WNT pathways [8–10]. Another study from China used single-cell RNA sequencing to determine intra-tumoral heterogeneity and microenvironment in metastatic gallbladder tumors [11]. Despite numerous studies on identifying new targets for GBC diagnosis and prediction of clinical outcomes, diagnostic and therapeutic progress has been slow and the current state of knowledge about genetic and molecular mechanisms in GBC is still limited.

This chapter unravels the transcriptomic patterns, biological functions, and pathways in GBC by employing a comprehensive systems-level approach. It involves the analysis of both publicly available and in-house generated GBC transcriptomic datasets [**Figure 4.1**], aiming to identify genes and biological pathways potentially associated with GBC development and pathogenesis. Three specific case studies (case studies 1-3) utilizing publicly accessible RNAseq datasets were undertaken to identify crucial molecular signatures implicated in GBC. Case study 4 involves analysis of the transcriptome dataset obtained from GBC patients originating from Assam, where the highest prevalence of this cancer from the north-eastern region is reported.

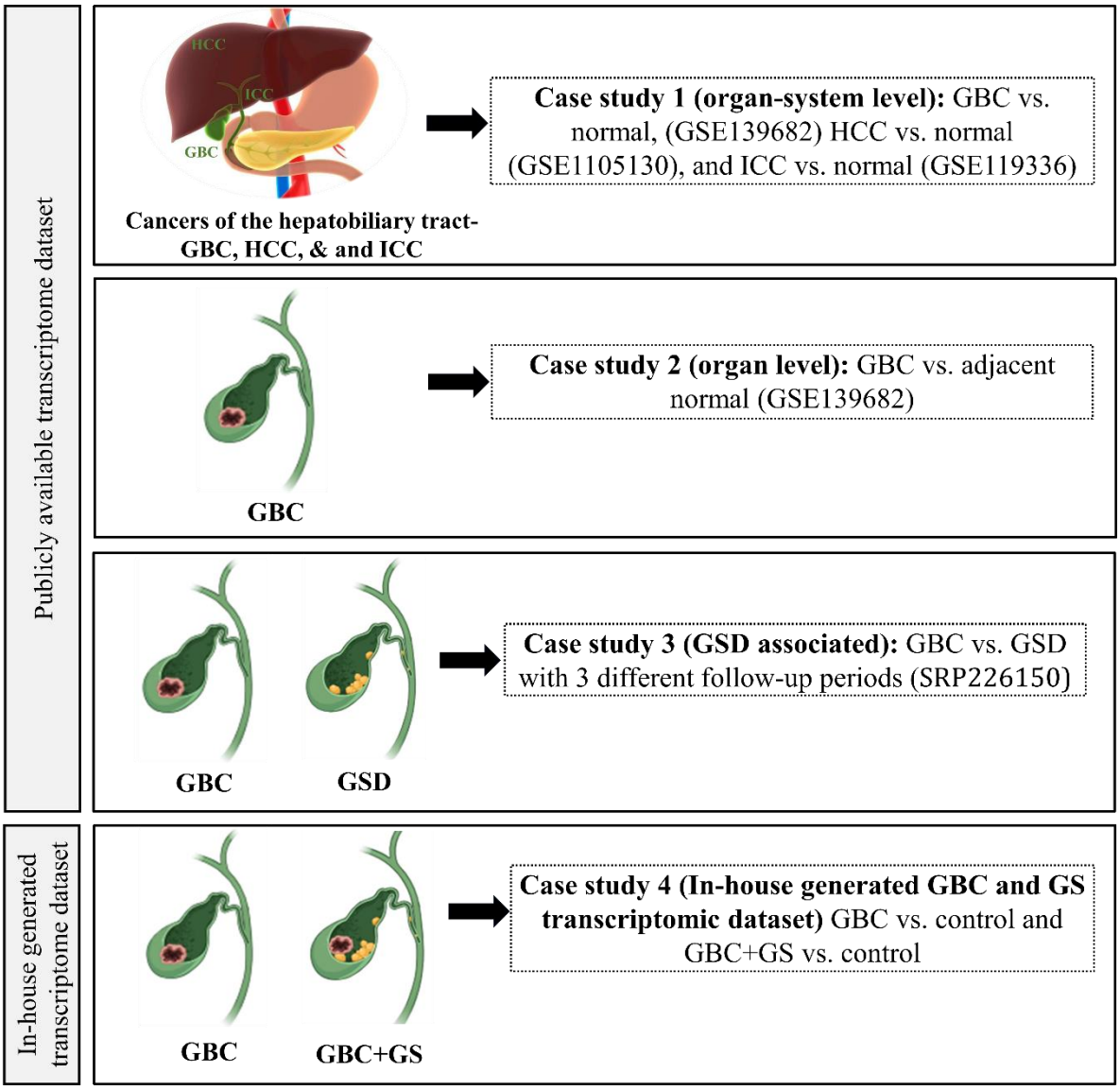


Figure 4.1: Schematic representation of the details of four different case studies performed to identify molecular signatures and pathways in GBC.

4.2 Results

The public and in-house generated GBC transcriptome datasets were analyzed as individual case studies (case studies 1 to 4) to identify crucial genes and biological pathways that drive GBC pathogenesis and development.

Case study 1 (organ-system level)

Comparative transcriptomic data analysis of three aggressive cancers of the hepatobiliary system

4.2.1 Comparative transcriptomic data analysis of hepatobiliary cancers (HBCs)- GBC, hepatocellular carcinoma (HCC), and intrahepatic cholangiocarcinoma (ICC).

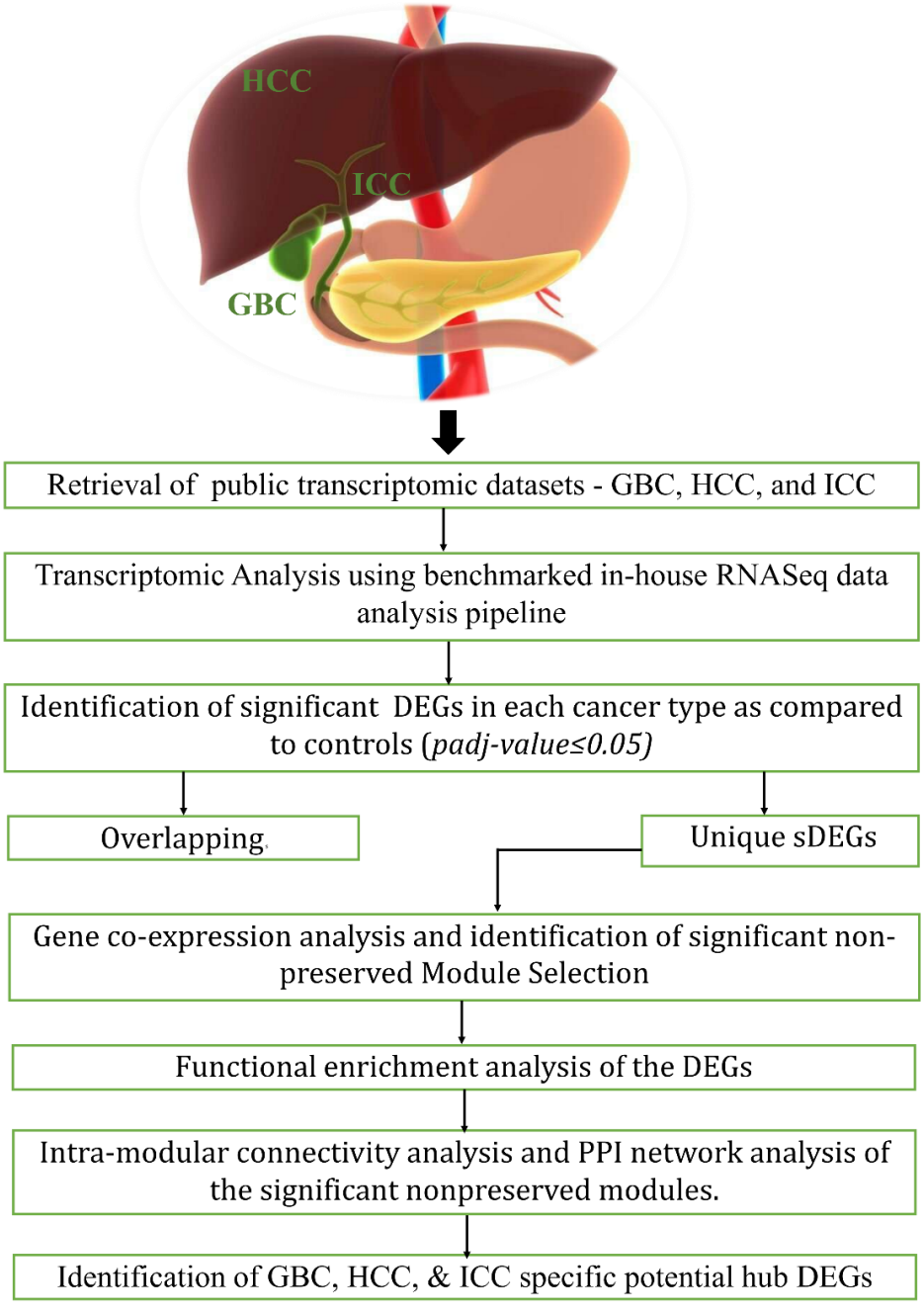


Figure 4.2: Outline of the workflow for identification of potential overlapping and unique DEGs in three aggressive cancers of the hepatobiliary tract (GBC, HBC, and ICC).

4.2.1.1 Dataset information

To obtain the relevant transcriptomic datasets on GBC, HCC, and ICC for this case study, a comprehensive search was conducted on the National Center for Biotechnology Information (NCBI) - Gene Expression Omnibus (GEO) and European Nucleotide Archive (ENA) database. Three paired-end RNAseq datasets – GSE139682 for GBC, GSE105130 for HCC, and GSE119336 for ICC were selected. The selected datasets were downloaded from the ENA database. The dataset information has been summarized in **Table 4.1**.

Table 4.1: The sample information of the RNAseq datasets considered for case study 1

Accession ID	Cancer Type	Sample Size	Tumor Sample	Tumor adjacent normal Sample
GSE139682	GBC	20	10	10
GSE105130	HCC	52	25	27
GSE119336	ICC	34	16	18

4.2.1.2 Identification of unique and shared significant DEGs in GBC, HCC, and ICC

For identifying the shared and unique DEGs among GBC, HCC & and ICC, differential gene expression analysis was carried out using two algorithms- DESeq2 and EdgeR. For each cancer dataset, the consensus of both algorithms was taken for screening the significant DEGs [**Figure 4.3 A**]. A total of 256 shared DEGs were identified among the three cancers, and 561, 2005, and 2580 unique DEGs have been identified in GBC, HCC, and ICC, respectively [**Figure 4.3 B**]. The Hierarchical clustering of the shared DEGs showed distinct gene expression patterns in each cancer, particularly in GBC. The overlapping DEGs in GBC were significantly downregulated as compared to HCC and ICC, suggesting that GBC exhibits a distinct trend of gene expression pattern [**Figure 4.3 C**]. The shared DEGs identified among GBC, HCC, and ICC are found to be linked with processes involved in cell cycle regulation such as mitotic spindle organization, mitotic sister chromatid segregation, microtubule cytoskeleton organization, mitotic cytokinesis, spindle assembly checkpoint, etc. [**Figure 4.3 D**].

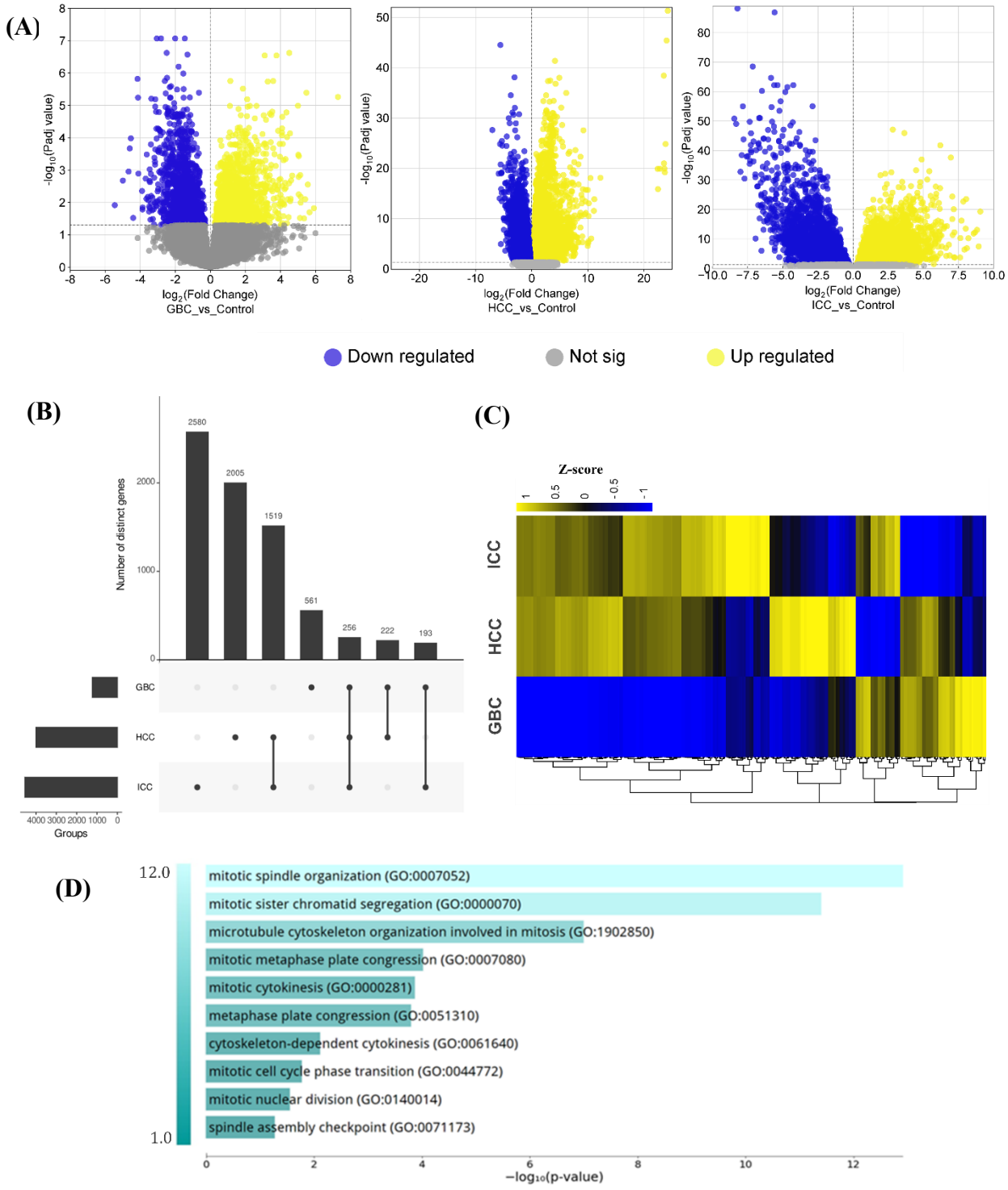


Figure 4.3: Identification of overlapping gene expression signatures among GBC, HCC, and ICC. **(A)** Volcano plot representing the identification of significant DEGs in GBC, HCC, and ICC. **(B)** Upset plot showing the number of shared and unique DEGs in GBC, HCC, and ICC. **(C)** The heatmap depicts the hierarchical clustering of the 256 shared DEGs between the three cancers of HBCs. **(D)** Bar plot representing the top ten significant biological processes associated with the shared DEGs between GBC, HCC, and ICC.

Downregulation of significant DEGs is higher in GBC as compared to HCC, and ICC [**Figure 4.4 A**]. Clustering analysis on gene expression data for the top 500 unique DEGs for GBC, HCC, and ICC using a complete linkage hierarchical approach identified distinct clusters for each cancer type as compared to normal samples [**Figure 4.4 B**]. These distinctions in terms of co-expressed groups of genes are a clear indication that there exists a unique set of molecular signatures for each cancer. These observations indicate that despite belonging to a common organ system, the progression of GBC, HCC, and ICC results from differential patterns of gene expression.

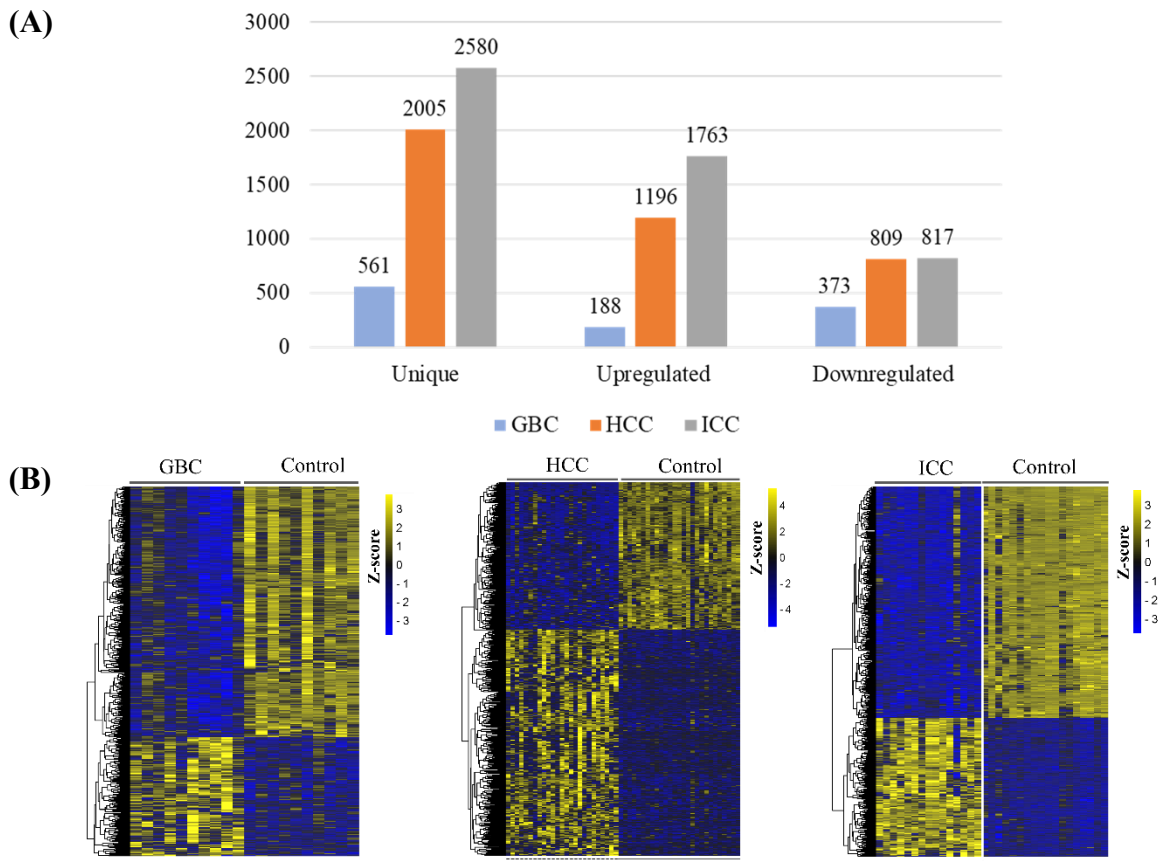
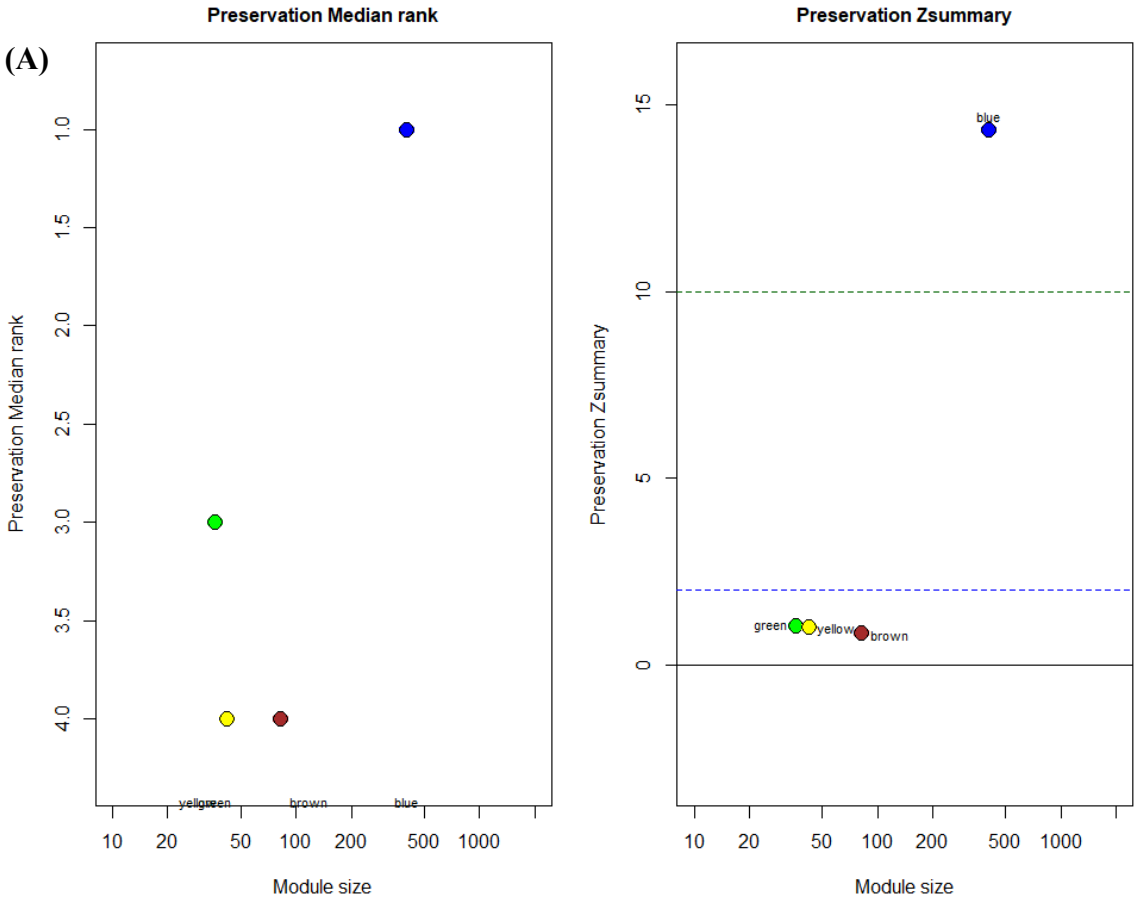


Figure 4.4: Identification of unique gene expression profiles in GBC, HCC, and ICC. (A) Bar plot showing the number of unique DEGs identified in each cancer type. **(B)** Complete linkage hierarchical clustering analysis of the expression profile of the top 500 unique DEGs identified in GBC, HCC, and ICC. The heatmap scale represents the Z-score of the gene expression counts.

4.2.1.3 Identification of module eigengene and detection of non-preserved modules

For the construction of gene co-expression network (GCN) for each type of hepatobiliary cancer, the unique significant DEGs (561, 2005, and 2580) in GBC, HCC, and ICC were used to construct GCNs for each cancer type using WGCNA package [**Figure 4.5**] in R (version 4.0.0) and an adjacency matrix was built by raising the Pearson's correlation coefficient of each pair of DEGs to the soft thresholding power β for generating a scale-free co-expression network. The soft-thresholding power β for (i) GBC and control was 12 and 9, respectively, for both (ii) HCC and control were 12, and for (iii) ICC and control were 14 and 12, respectively.



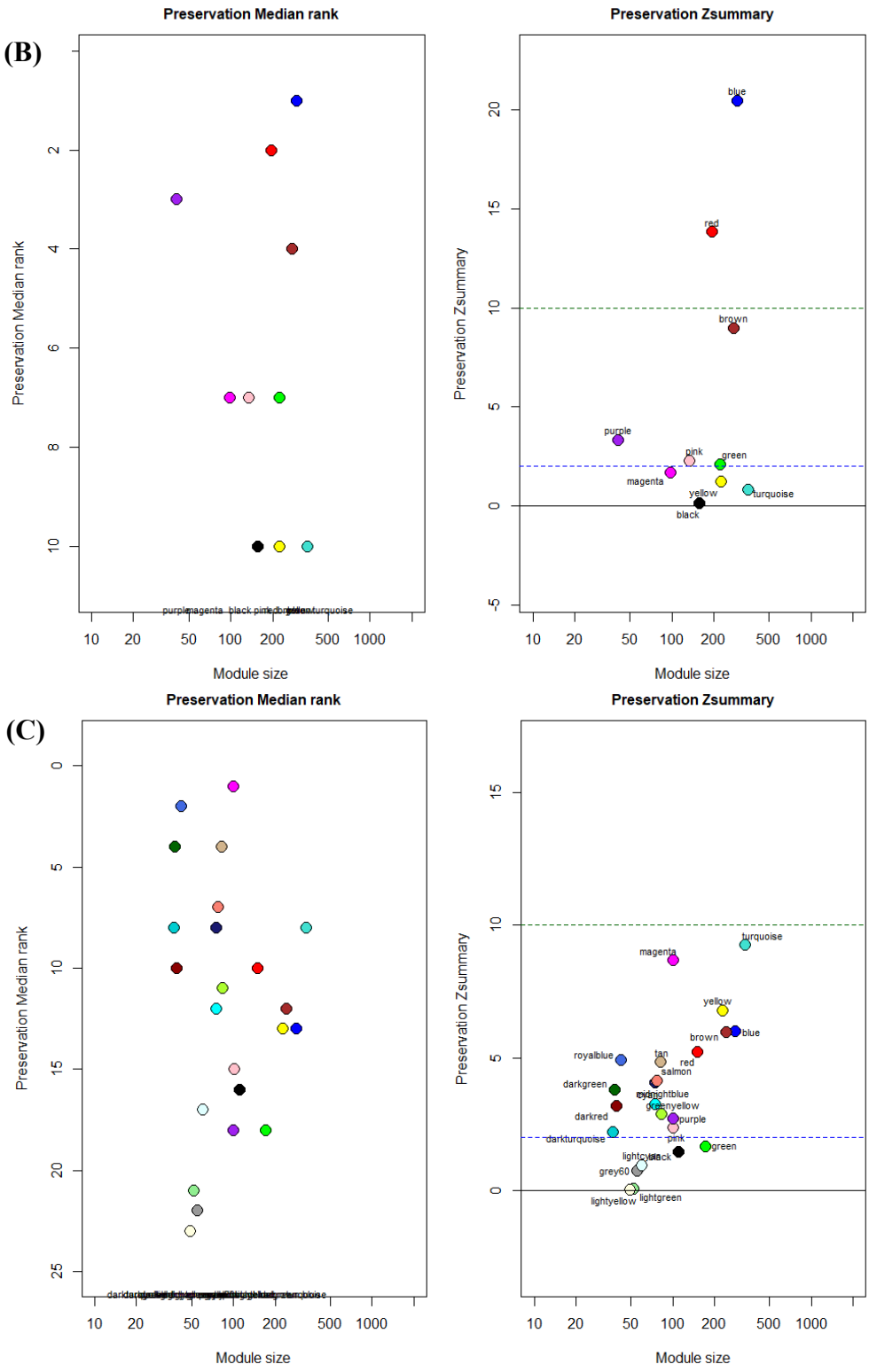


Figure 4.5: Construction of gene co-expression networks and identification of nonpreserved modules. Preservation analysis of modules was done based on *Z-summary* and *medianRank*. Modules whose topological properties changed in a cancer network compared to normal networks are termed non-preserved modules. The significant nonpreserved modules

were identified in GBC, HCC, and ICC respectively based on *Zsummary* and medianRank preservation.

The black (82 DEGs), brown (147 DEGs), and lightyellow (49 DEGs) modules were considered for the construction of coexpression networks [**Table 4.2**]. The *Zsummary* threshold of 10 and above indicates strong preservation, a threshold above 2 for low to moderate preservation, and below 2 indicates poor or no preservation. Similarly, the *medianRank* modules close to zero indicate a high degree of preservation, whereas, the *medianRank* of modules above 10 is highly nonpreserved. The *Zsummary* and *MedianRank* statistics help in the identification of nonpreserved modules that represent gene sets that are specific to certain biological contexts or states. Investigating the biological functions and pathways associated with nonpreserved modules can provide insights into condition-specific processes or responses.

Table 4.2: The *Zsummary* and *medianRank* preservation (pres) values of the significant nonpreserved modules identified in GBC, HCC, & ICC networks. The module colors in bold were considered for downstream analysis.

Cancer type	Nonpreserved modules	***medianRank*.pres**	***Zsummary*.pres**
GBC	**brown**	4	0.87
	green	3	1.1
	yellow	4	1
HCC	**black**	10	0.12
	green	4	0.16
	magenta	7	1.7
	turquoise	10	0.83
	yellow	10	1.2
ICC	black	16	1.4
	green	18	1.7
	grey60	22	0.73
	lightcyan	17	0.96
	lightgreen	21	0.065
	lightyellow	23	0.02

4.2.1.4 Pathway enrichment analysis of DEGs identified from significant nonpreserved modules.

The pathway analysis of the selected nonpreserved modules [**Table 4.3**] from each cancer showed that the gene sets in the brown module identified in GBC are enriched in MAPK signaling pathways whereas; the gene sets in black and lightyellow nonpreserved modules

identified in HCC and ICC are largely associated with immune response signaling pathways such as Rap, Ras and PI3K-Akt signaling pathways. The nonpreserved module lightyellow identified from the ICC coexpression network is significantly associated with ECM-receptor interaction, inflammatory mediators, and bile acids biosynthesis pathways.

Table 4.3: The top five significantly enriched KEGG pathways associated with nonpreserved modules identified from the GBC, HCC, and ICC networks.

Pathways	Counts	DEGs	Padj values
GBC			
MAPK signaling pathway	4	*EFNA1, MECOM, ERBB2, MAP3K13*	0.033
HCC			
Rap1 signaling pathway	6	*FGF22, VEGFD, PDGFC, PARD6G, FGF14, FGF21*	5.43E-4
Regulation of actin cytoskeleton	5	*FGF22, ACTN2, PDGFC, FGF14, FGF21*	0.004
Ras signaling pathway	5	*FGF22, VEGFD, PDGFC, FGF14, FGF21*	0.006
PI3K-Akt signaling pathways	5	*FGF22, VEGFD, PDGFC, FGF14, FGF21*	0.025
ICC			
ECM-receptor interaction	4	*THBS2, ITGB1, LAMB1, ITGAV*	0.002
PI3K-Akt signaling pathway	5	*THBS2, ITGB1, LAMB1, KDR, ITGAV*	0.022
inflammatory mediator regulation of TRP channels	3	*IL1R1, ITPR3, TRPM8*	0.038
Primary bile acid biosynthesis	1	*CYP7B1*	0.040
ECM-receptor interaction	4	*THBS2, ITGB1, LAMB1, ITGAV*	0.002

4.2.1.5 Identification of hub genes from the nonpreserved module clusters

The genes having a large number of interacting nodes in the non-preserved modules were recognized as hub genes using intra-modular connectivity analysis. The top five genes from each cancer type, having the highest correlation weight were considered as hubs. The weight of the potential candidate genes identified is indicated in [**Table 4.4].** The hub genes identified

for GBC, HCC, and ICC through intra-modular connectivity analysis are *MAP3K13*, *AC069287.1*, and *TRPC1*.

Table 4.4: List of top five hub DEGs identified from nonpreserved modules of GBC, HCC, and ICC through intramodular connectivity analysis.

GBC		HCC		ICC	
Gene	**Weight**	**Gene**	**Weight**	**Gene**	**Weight**
MAP3K13	6.76	*AC069287.1*	2.24	*TRPC1*	1.21
HYLS1	5.71	*SEPTIN14P19*	2.16	*SLC45A1*	0.99
KIAA0895	5.67	*SEPTIN14P5*	1.75	*PPP2R3A*	0.92
AC130456.2	5.51	*DCDC1*	1.59	*BACE1*	0.90
GPR39	5.48	*CICP6*	1.25	*SNPH*	0.90

Furthermore; PPI networks were constructed using the genes of the non-preserved modules for predicting the interactions of genes at the protein level from the STRING database (version 11) with a confidence score > 0.70. The genes having the highest degree were identified as hubs and interactive PPI networks generated using DEGs of the nonpreserved modules are represented in **Figure 4.6**. The hub genes for GBC, HCC, and ICC are *ERBB2*, *ACTN2*, and *GLI1*, respectively, and have been tabulated in **Table 4.5**. It was observed that the degree of interaction of the hub mRNAs identified in GBC is much higher than HCC and ICC.

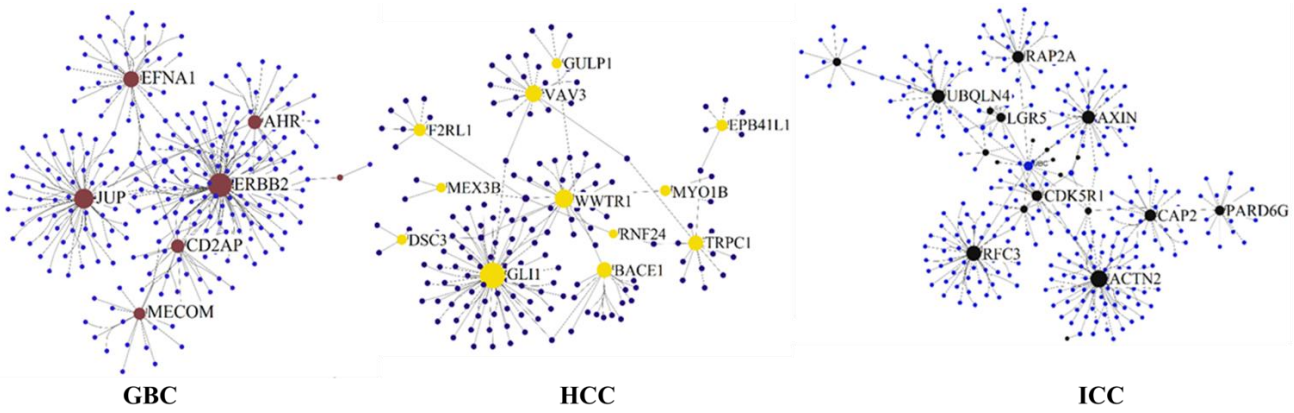


Figure 4.6: Construction of PPI networks with gene sets identified from nonpreserved modules of GBC, HCC, and ICC. The large brown, black, and yellow nodes represent the hub DEGs in GBC, HCC, and ICC respectively based on degree centrality. Significant PPIs were filtered using a combined score > 0.70.

Table 4.5: List of top five hub DEGs identified from nonpreserved modules of GBC, HCC, and ICC through PPI network analysis.

GBC		HCC		ICC	
Gene	**Degree**	**Gene**	**Degree**	**Gene**	**Degree**
ERBB2	119	*ACTN2*	71	*GLI1*	69
JUP	67	*RFC3*	49	*WWTR1*	26
EFNA1	37	*AXIN2*	35	*VAV3*	19
AHR	25	*UBQLN4*	34	*BACE1*	15
CD2AP	23	*RAP2A*	26	*TRPC1*	14

Case study 1 revealed that Overlapping DEGs among GBC, HCC, & ICC are majorly involved in cell cycle processes. DEGs in GBC are significantly downregulated compared to HCC and ICC Each of the cancers of the HBS is associated with distinct biological processes and cellular pathways (GBC – cellular processes, HCC – immune signaling pathways, ICC – metabolic pathways). Each of the cancers exhibits unique expression patterns despite being a part of the same organ system (HBS). The biological pathways primarily contributing to cancer progression vary in each cancer type providing scope for targeted therapy, particularly for GBC as there are no specific targeted therapies available for GBC patients. As compared to HCC and ICC, GBC shows a distinct trend of gene expression patterns and is highly downregulated. The hub genes identified in GBC- *MAP3K13* and *ERBB2* are associated with signal transduction processes that regulate cell proliferation, differentiation, and apoptosis.

To further understand, how the differential gene expression varies between GBC samples and GBC adjacent normal samples, case study 2 was performed to identify differential gene expression patterns and molecular signatures in GBC compared to normal samples.

Case study 2 (organ level)

Transcriptome data analysis and identification of molecular signatures in GBC compared to adjacent normal

4.2.2 Transcriptome data analysis and identification of differential gene expression signatures in GBC samples compared to adjacent normal samples.

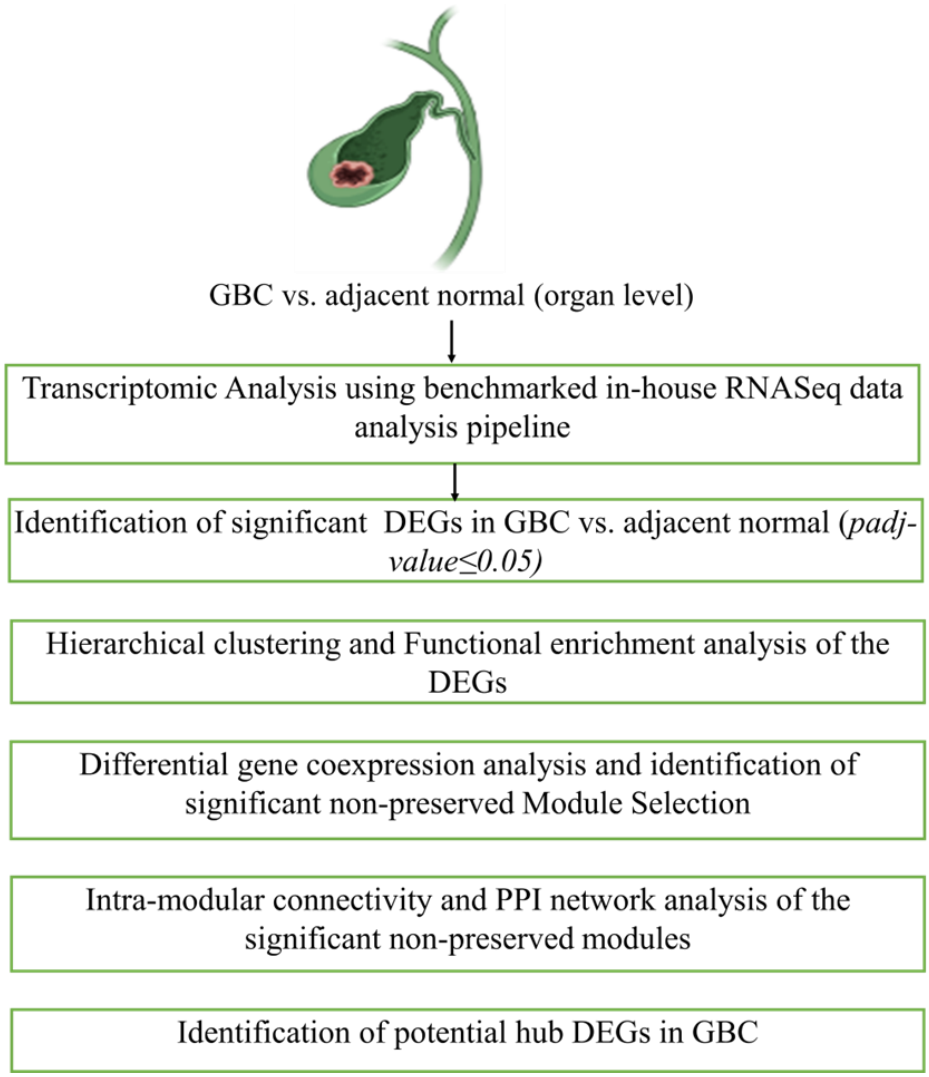


Figure 4.7: Outline of the workflow for identification of potential DEGs in GBC compared to normal.

4.2.2.1: Dataset Information

The dataset (Accession ID: GSE139682) used for this study was retrieved from the ENA database. It is a paired-end transcriptomic dataset comprising 10 GBC and 10 tumor-matched adjacent normal (control) tissue samples. The samples used for RNAseq were obtained from GBC patients without chemo/radiotherapy, and all the tissue samples were collected through resection surgery.

4.2.2.2 Identification of differential gene expression profile in GBC compared to normal samples

Transcriptomic data analysis has been carried out to identify differentially expressed molecular signatures in GBC compared to adjacent normal samples (considered as control). By taking *Padj* value ≤ 0.05, 2980 significant DEGs were identified in GBC as compared to normal. Out of 2980 DEGs, 1425 and 1555 DEGs were found to be upregulated and downregulated respectively. Hierarchical clustering analysis of the significant DEGs showed that the GBC and the adjacent control samples exhibit distinct differential gene expression [**Figure 4.8 A**], which indicates that the identified DEGs act differentially in GBC samples and normal samples. The significant DEGs identified in GBC are significantly associated with the cell cycle regulation and signal transduction processes. This suggests that genes related to cell cycle progression and checkpoint regulatory proteins might be crucial for GBC development [**Figure 4.8 B**].

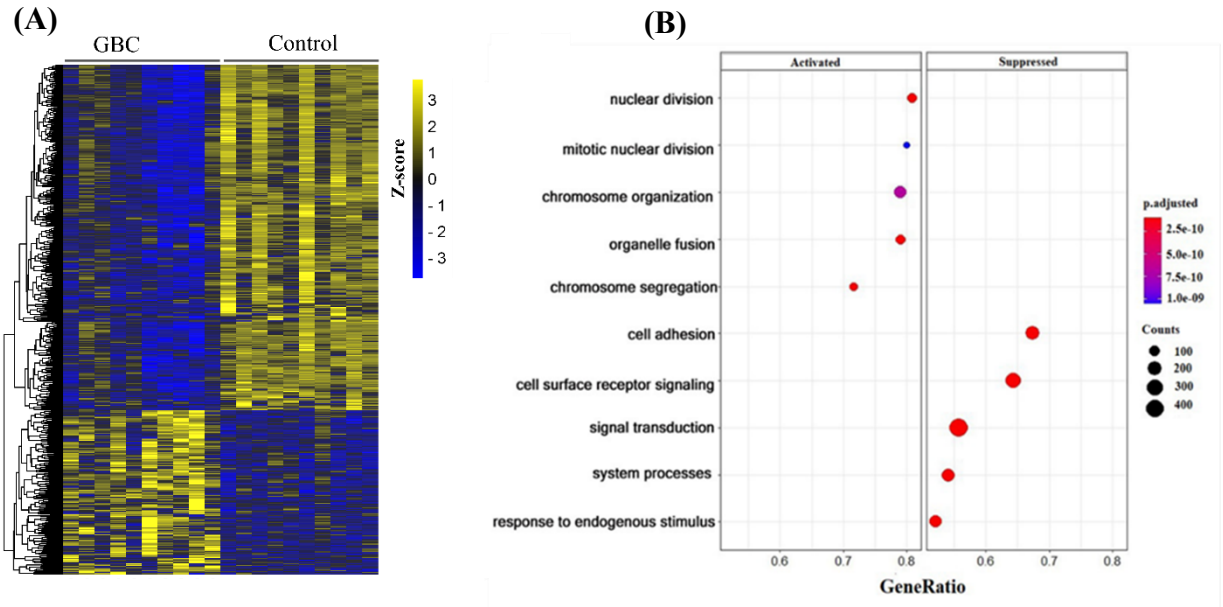


Figure 4.8: Differential gene expression in GBC as compared to control. (A) Gene-wise complete linkage hierarchical clustering heatmap of DEGs identified in GBC compared to control. **(B)** Identification of the top ten enriched biological processes associated with the significant DEGs. The x-axis represents the enrichment ratio between the number of DEGs and all UniGenes enriched in particular GO terms. The size of the dot represents the number of DEGs assigned to the particular GO term and the color of the dot represents the *Padj*. The left panel of the dot plot represents terms/pathways upregulated in GBC and the right panel represents terms downregulated in GBC as compared to the control.

4.2.2.3 Differential gene coexpression network analysis and identification of nonpreserved modules from GBC and normal coexpression network

The differential gene coexpression network analysis was carried out by taking the normalized gene expression counts of significant DEGs identified in GBC as compared to normal. The differential co-expression networks for GBC and control conditions were constructed separately using the WGCNA package in R. The soft-thresholding power β considered for GBC and control are 18 and 20 respectively. Subsequently, a hierarchical clustering approach identified the module clusters containing highly connected gene sets. A total of 18 and 20 modules were identified in control and GBC conditions respectively. The cluster dendogram containing modules and heatmap plot for control and GBC network is given in **Figure 4.9**.

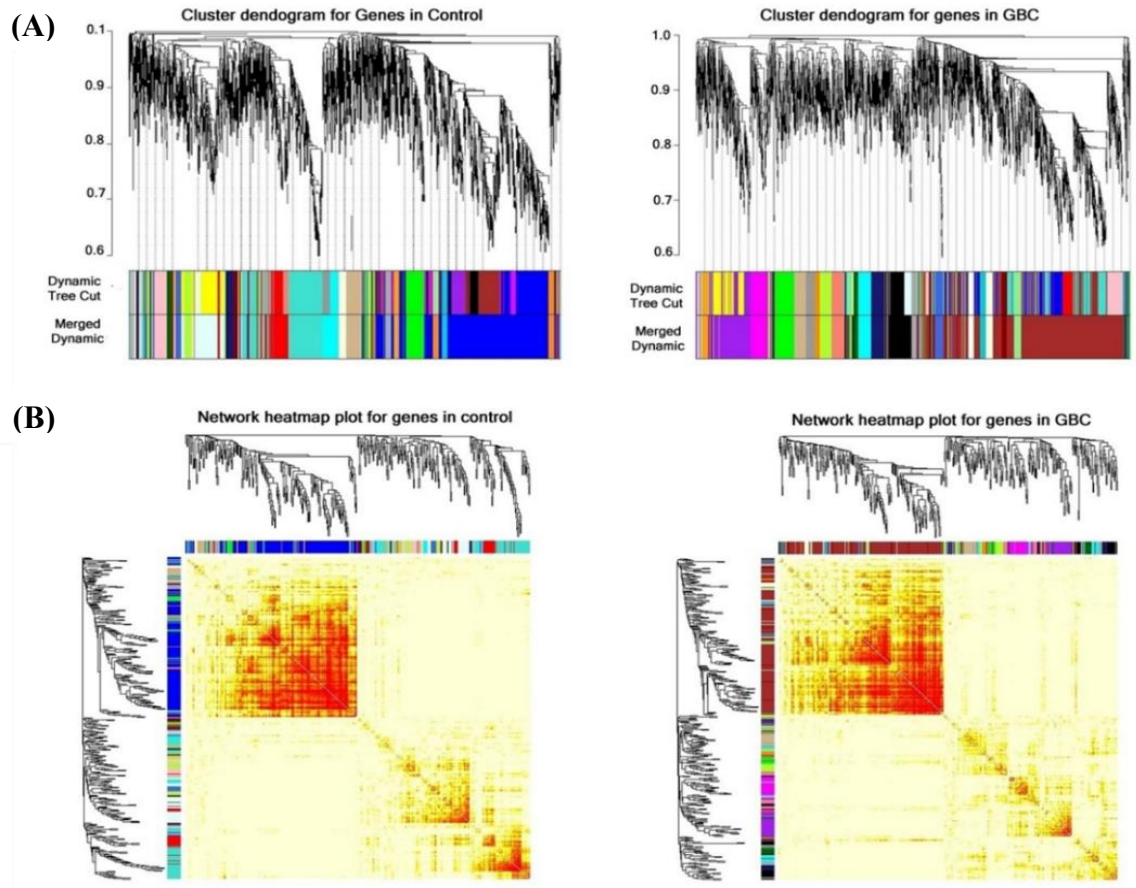
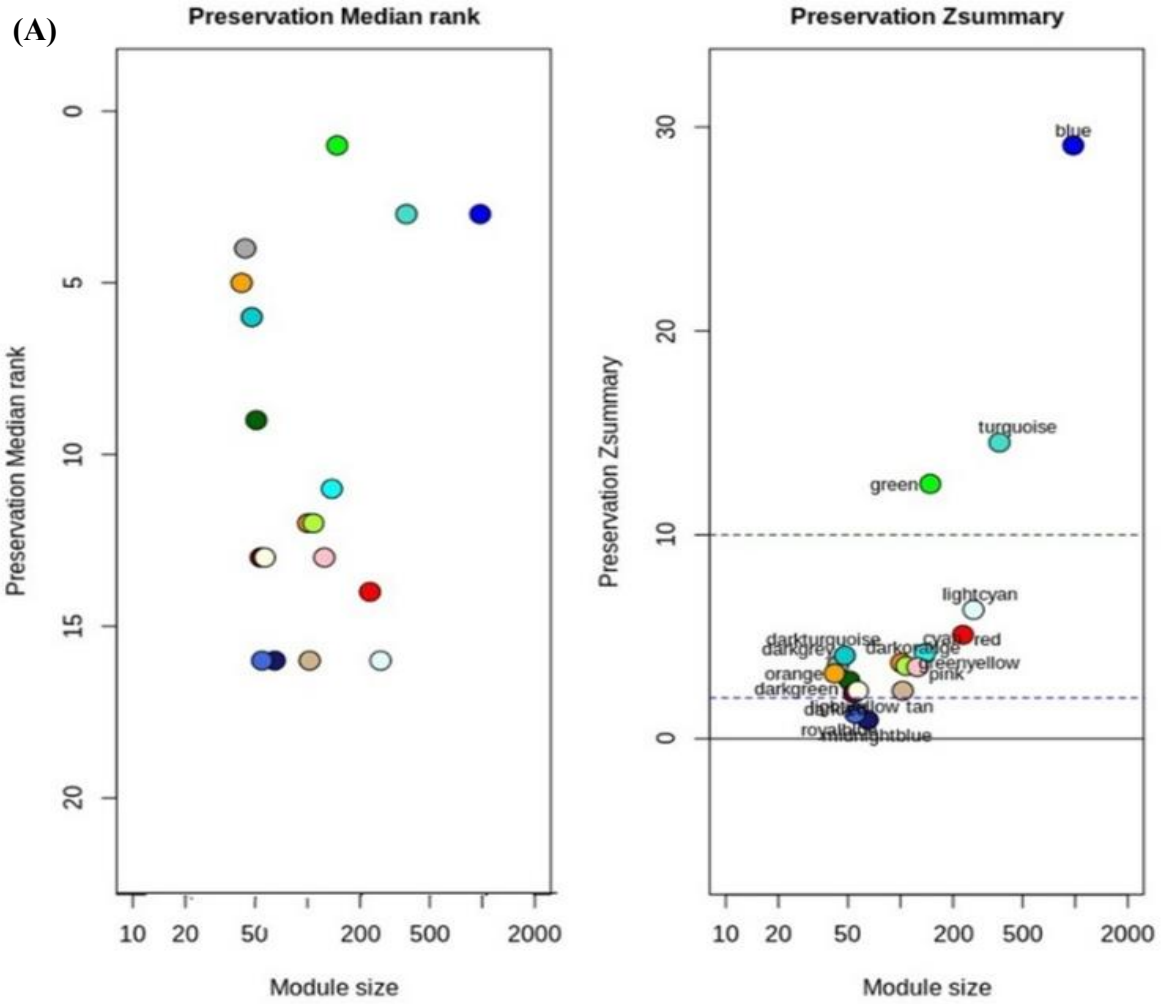


Figure 4.9: Construction of differential gene co-expression networks in GBC and control samples. (A) Hierarchical clustering dendogram of DEGs based on dissimilarity measure (the 1-TOM) matrix. The co-expressed modules identified in the GBC and control network are represented by different colors. **(B)** Clustering network heatmaps of co-expressed modules identified in GBC and control co-expression networks representing the pairwise correlation of of the gene sets in the modules across the samples.

The module preservation analysis from the differential co-expression network was performed to identify non-preserved modules in control and GBC conditions using statistical measures- Z-*summary* and *medianRank*. Here, the module preservation analysis was performed by the following approaches: (i) GBC vs. control, where the cancer data was considered as the test data and the reference data was the control data; (ii) Control vs. GBC, in which the control data was considered as the test data and the GBC data was the reference data. The non-preserved modules can give insights into distinct molecular signatures in GBC modules compared to those of control modules. In GBC to control module preservation analysis, salmon, tan, and grey60, and for control to GBC preservation analysis midnightblue and royalblue were considered for further downstream analysis [**Figure 4.10**].

(A)



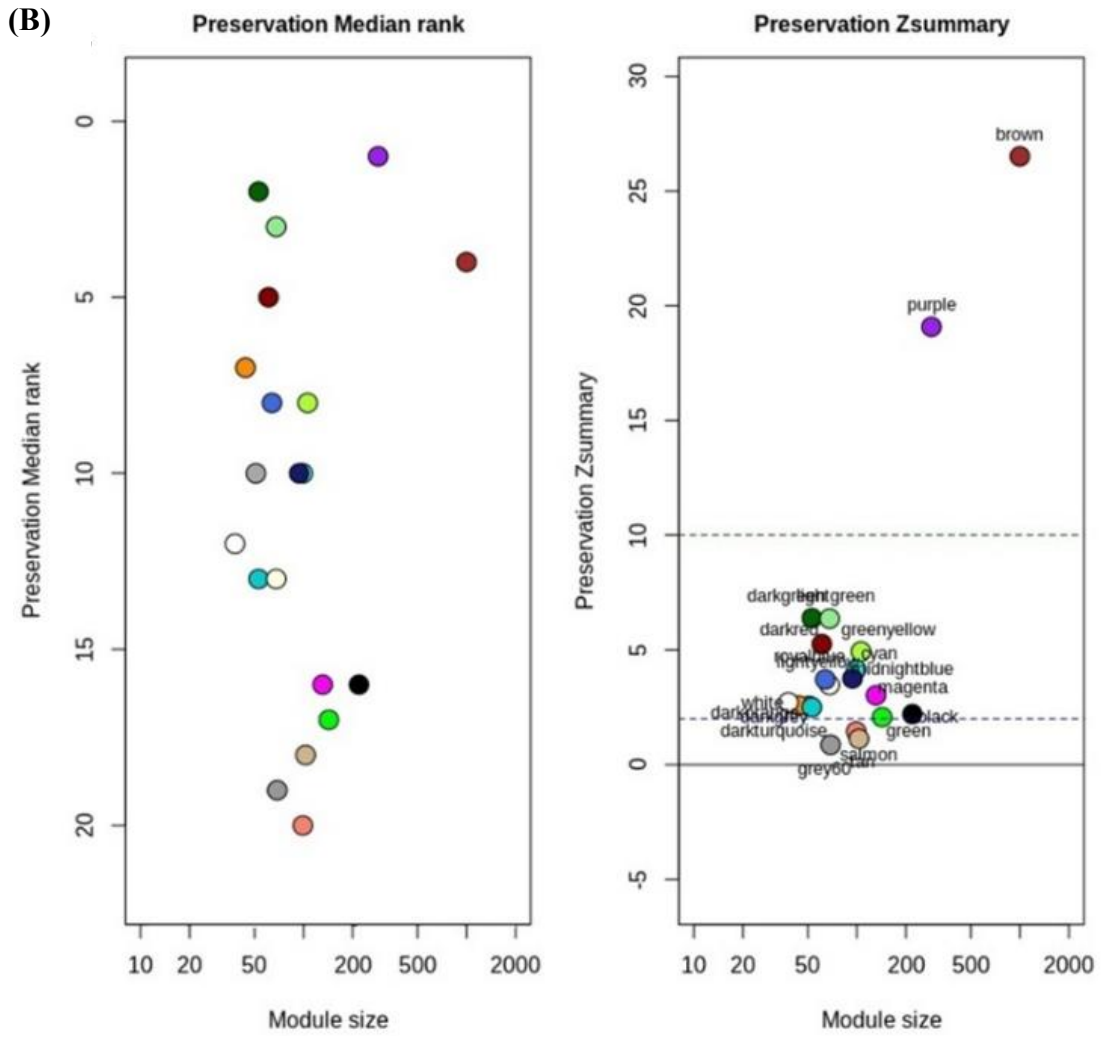


Figure 4.10: Identification of nonpreserved modules from GBC and control networks based on *Z-summary* and *medianRank*. (A) Identification of nonpreserved modules in the control condition. The modules in midnightblue and royalblue colors are identified as non-preserved. **(B)** Identification of modules in the GBC condition, where the modules in tan, salmon, and grey60 color are identified as non-preserved.

4.2.2.4 Functional annotation and pathway associated with genes of the non-preserved modules

The functional enrichment analysis was carried out with the gene sets present in the nonpreserved modules. The statistical significance of p-value <0.05 was considered for identifying the significantly enriched biological processes and pathways related to GBC pathogenesis. The functional annotation analysis of the nonpreserved modules identified from the control network is associated with the regulation of the DNA replication process and cell cycle glycerolipid and phosphatidyl signaling pathways. Whereas; the genes identified from

the nonpreserved modules from the GBC network are linked to developmental processes such as cell fate commitments, neuron migration, mitotic cell cycle processes, and cancer-related pathways including p53 signaling, Wnt signaling, small cell lung cancer and steroid biosynthesis pathways. The top 5 significantly enriched biological processes and pathways were tabulated in **Table 4.6** and **Table 4.7** respectively.

Table 4.6: The Top five significant biological processes associated with DEGs identified from nonpreserved modules of GBC and control networks.

Nonpreserved Modules	Biological Processes (GO Terms)	Counts	Genes	Padj Value
midnight blue	intracellular signal transduction	4	*NEK11, DGKB, GUCY1B2, NUDT4*	0.039
royalblue	negative regulation of DNA replication	2	*S100A11, CDC6*	0.039
salmon	dorsal/ventral axis specification	2	*PAX6, RGS20*	0.037
	neuron migration	3	*CELSR3, PAX6, PTK2*	0.042
	negative regulation of keratinocyte proliferation	2	*CTSV, EPPK1*	0.043
	interphase of mitotic cell cycle	4	*CCNB2, CCNE2, E2F5, TAF2*	0.038
grey60	Negative regulation of translation	2	*FXF1, EIFAK1*	0.008
	Cell fate commitment	3	*FOXA1, HOXA11, TFAP2C*	0.013
	Developmental growth		*HOXA11, TFAP2C, PLAC1*	0.018
tan	planar cell polarity pathway involved in axon guidance	2	*VANGL2, RYK*	0.012
	Epidermal cell differentiation	2	*OVOL2, SPINK5*	0.016
	Negative regulation of serine-type endopeptidase activity	2	*SPINK1, SPINK5*	0.028

Table 4.7: The Top five significant KEGG pathways associated with DEGs identified from nonpreserved modules of GBC and control networks.

Nonpreserved Modules	Pathways	Counts	Genes	P-adj values
midnightblue	Glycerolipid metabolism	2	*HLA-DMA*	0.006
	Toxoplasmosis	2	*DGKB, LIPC*	0.022
	Apoptosis	2	*CTSV, CASP12*	0.031
	Glycosaminoglycan degradation	1	*HYAL4*	0.038
royalblue	Phosphatidyl inositol signaling	2	*PIK3CB, ITPKA*	0.007
	Cell cycle	2	*MCM3, CDC6*	0.020
salmon	Cell cycle	3	*E2F5, CCNB2, CCNE2*	0.045
	Small cell lung cancer	2	*CCNE2, PTK2*	0.038
	P53 signaling	2	*CCNE2, CCEB2*	0.024
	Ubiquine biosynthesis	1	*COQ2T*	0.036
grey60	Thiamine metabolism	1	*ALPPL2*	0.026
	Necroptosis	2	*H2AW, RNF103-CHMP3*	0.029
	Alcoholism	2	*H2AW, H2BO1*	0.035
	Histidine metabolism	1	*ALDH3B2*	0.038
tan	Wnt signaling pathway	2	*VANGL2, RYK*	0.032
	Steroid biosynthesis	1	*CYP24A1*	0.033

4.2.2.5 Hub gene identification from non-preserved modules

The genes having a high degree of connectivity or high correlations in significant modules were regarded as hub genes. The nonpreserved modules identified from both GBC and control network were considered and the topological measure with respect to intra-modular connectivity was determined for identification of hub genes. A total of five genes have been considered as potential candidates in terms of the correlation weight (degree) of the genes [**Table 4.8**]. The genes identified with the highest intra-modular connectivity from each module (hub gene) are *AL009178.3* (novel transcript), *ADAM18* (ADAM Metallopeptidase domain 18), MAPK (Mitogen-Activated Protein Kinase 15), *L3MBTL1* (Lethal 3 malignant brain tumor-like protein 1) and *ALPPL2* (Alkaline phosphatase, placental-like 2). Moreover; it was observed that the weight of the genes identified from nonpreserved modules of the GBC network is much higher as compared to the control network, suggesting the potential role of these hub genes in GBC pathogenesis.

Table 4.8: List of top five hub DEGs identified in GBC from nonpreserved modules using intra-modular connectivity analysis. The DEGs with the highest interaction (weight) with the other DEGs in the modules are considered hubs.

Control to GBC				GBC to Control					
Midnightblue		Royalblue		Salmon		Tan		Grey60	
Gene	Weight	Gene	Weight	Gene	Weight	Gene	Weight	Gene	Weight
AL009178.3	2.87	*ADAM18*	2.87	*MAPK15*	12.51	*L3MBTL1*	11.38	*ALPPL2*	8.87
SPATC1	2.71	*CNTN4*	2.71	*TRAPPC9*	12.08	*ZNF337-AS1*	10.47	*PATE4*	8.01
CTSV	2.70	*NUP62CL*	2.70	*OPLAH*	11.72	*AC099661.1*	9.64	*AP00842.3*	7.85
AL353746.1	2.64	*QTRT2*	2.64	*OTUD6B*	10.97	*AC240565*	8.68	*GPATCH1*	7.80
AL360270.1	2.61	*LINC01517*	2.61	*TAF2*	10.72	*C1QTNF*	8.14	*AP000977.1*	7.06

Subsequently, PPI networks were constructed with the genes identified from each of the non-preserved modules [**Figure 4.11**]. The hub genes were selected based on the degree of centrality of the PPI networks. The genes with the highest degree of interaction are *BIRC7* (Baculoviral IAP repeat-containing protein 7), *CCNB2* (Cyclin B2), *CDC6* (Cell division cycle 6), *L3MBTL1*, and *WDR88* (WD repeat domain 88). All the identified hub DEGs were found to be upregulated in GBC, compared to the controls. This indicates that the upregulation of these hub genes might be involved in GBC development and progression.

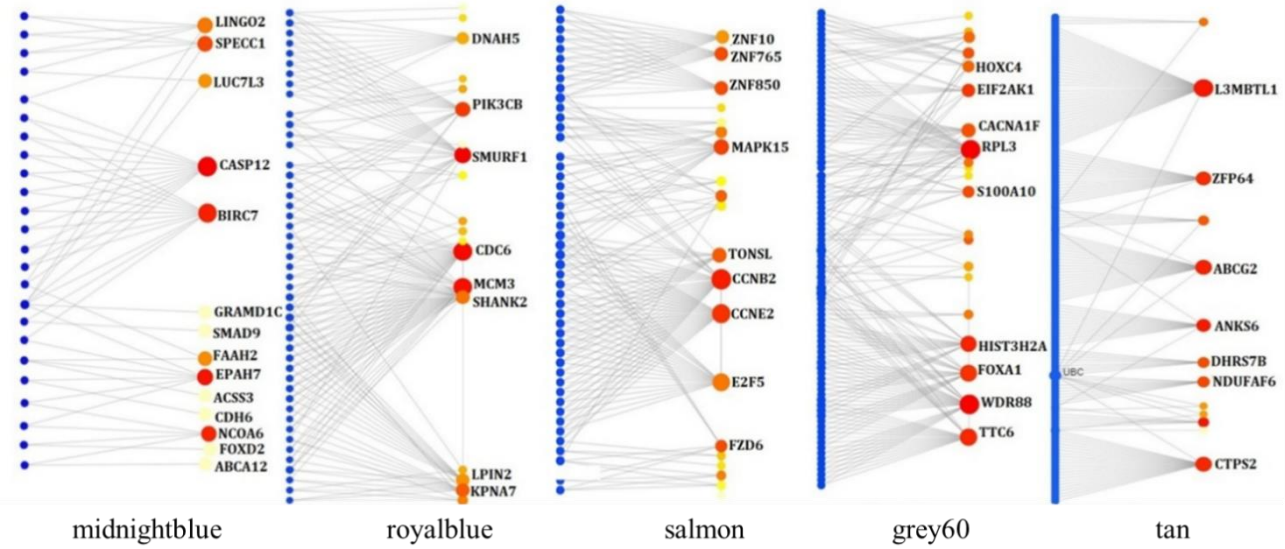


Figure 4.11: PPI network analysis of the significant non-preserved modules identified in GBC and control gene co-expression networks. Construction of the PPI networks with the DEGs identified from the non-preserved modules of the control network (midnightblue and

royal blue modules) and GBC network (salmon, tan, and grey60 modules). The small blue circles represent the proteins and the large red node represents the genes in the modules.

Table 4.9: List of top five hub DEGs identified in GBC from nonpreserved modules using PPI network analysis.

Control to GBC				**GBC to Control**					
Midnightblue		**Royalblue**		**Salmon**		**Tan**		**Grey60**	
Gene	**Degree**	**Gene**	**Degree**	**Gene**	**Degree**	**Gene**	**Degree**	**Gene**	**Degree**
CDC6	30	*BIRC7*	12	*CCNB2*	30	*L3MBTL1*	34	*WDR88*	31
MCM3	28	*CASP12*	11	*CCNE2*	25	*ABCG2*	23	*RPL3*	31
SMURF1	22	*UBC*	6	*E2F5*	22	*CTPS2*	12	*TIC6*	23
PIK3AB	17	*EPH7*	6	*MAPK15*	15	*SIM2*	7	*FOXA1*	22
SHANK2	14	*LINGO2*	5	*TONSL*	12	*PTP4A1*	7	*HIST3H2A*	18

The hub genes [**Table 4.9**] with the highest degree of centrality identified in control to the GBC network are *CDC6* (cell division cycle 6) and *BIRC7* (baculoviral IAP repeat containing 7). In the GBC to control network, the DEGs identified from nonpreserved modules with the highest degree of centrality are *CCNB2* (cyclin B2), *L3MBTL1* (Lethal (3) malignant brain-tumor like protein 1) and *WDR88* (WD repeat domain 88). These hub genes are associated directly and indirectly with cell cycle regulatory proteins. Moreover, it was observed that the degree of hub genes identified in GBC to control network is much higher as compared to the degree of hub genes in control to GBC network, which suggests that the hub genes in GBC to control network have high interconnectedness with the other genes in the network and might be involved in dysregulation of key regulatory genes and results in GBC development.

From this case study, it was found that the DEGs identified in GBC as compared to adjacent control samples are highly downregulated. The upregulated DEGs are significantly associated with cell cycle processes whereas; the downregulated DEGs are involved in signal transduction pathways. Moreover, the hub genes identified through differential gene co-expression network analysis followed by PPI analysis were directly or indirectly associated with components of the cell cycle system, apoptotic regulation, and cell-cell adhesion process. Similarly, in this case study, it was observed that genes related to cellular processes such as cell cycle, cell adhesion, and apoptosis were essential and significant in the pathogenesis and progression of GBC

For a comprehensive understanding of the association between GBC and GSD in terms of gene expression, case study 3 was performed to identify the differential gene expression patterns in GBC compared to GSD across three distinct follow-up periods.

Case study 3 (Association of GSD in GBC pathogenesis)

Comparative transcriptomic data analysis of GBC and GSD with different follow-up periods

4.2.3 Transcriptomic data analysis on GBC compared to GSD groups with different follow-up periods.

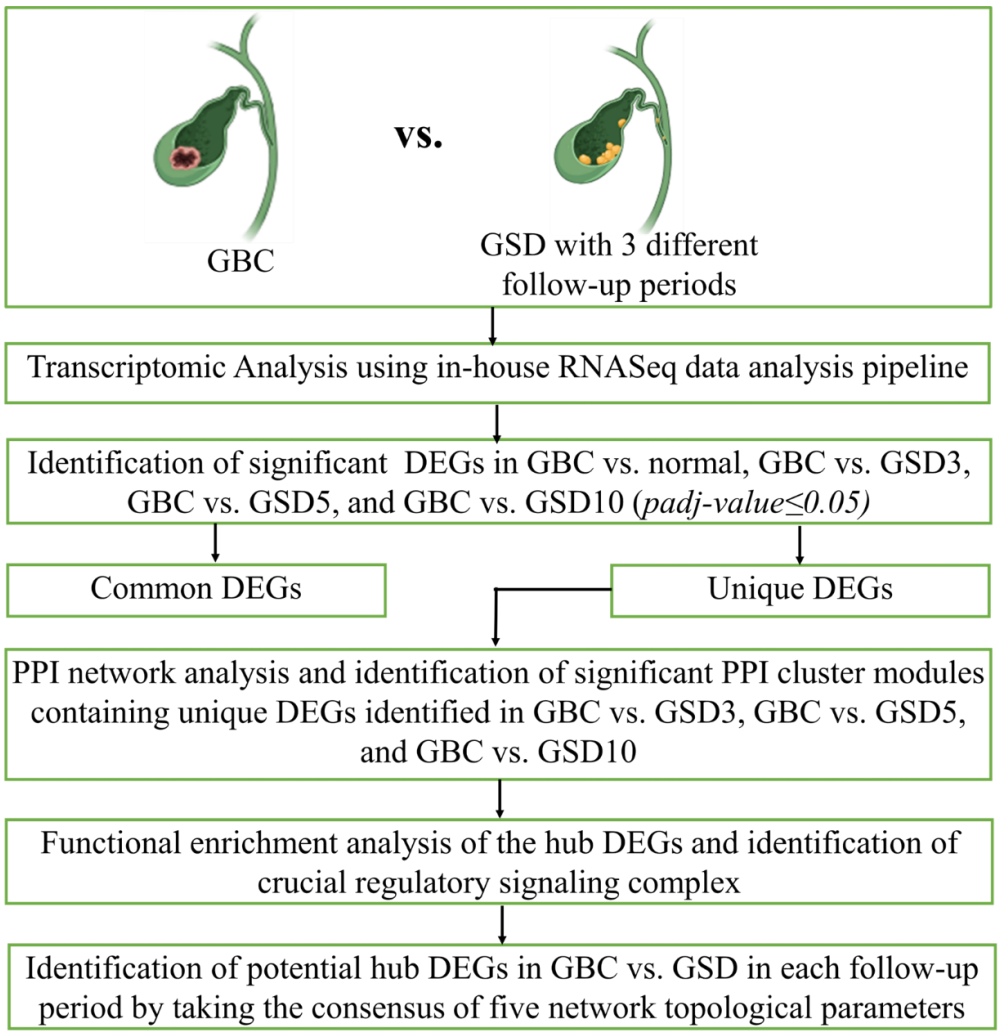


Figure 4.12: Outline of the workflow for identification of potential DEGs in GBC compared to GSD with three different follow-up periods.

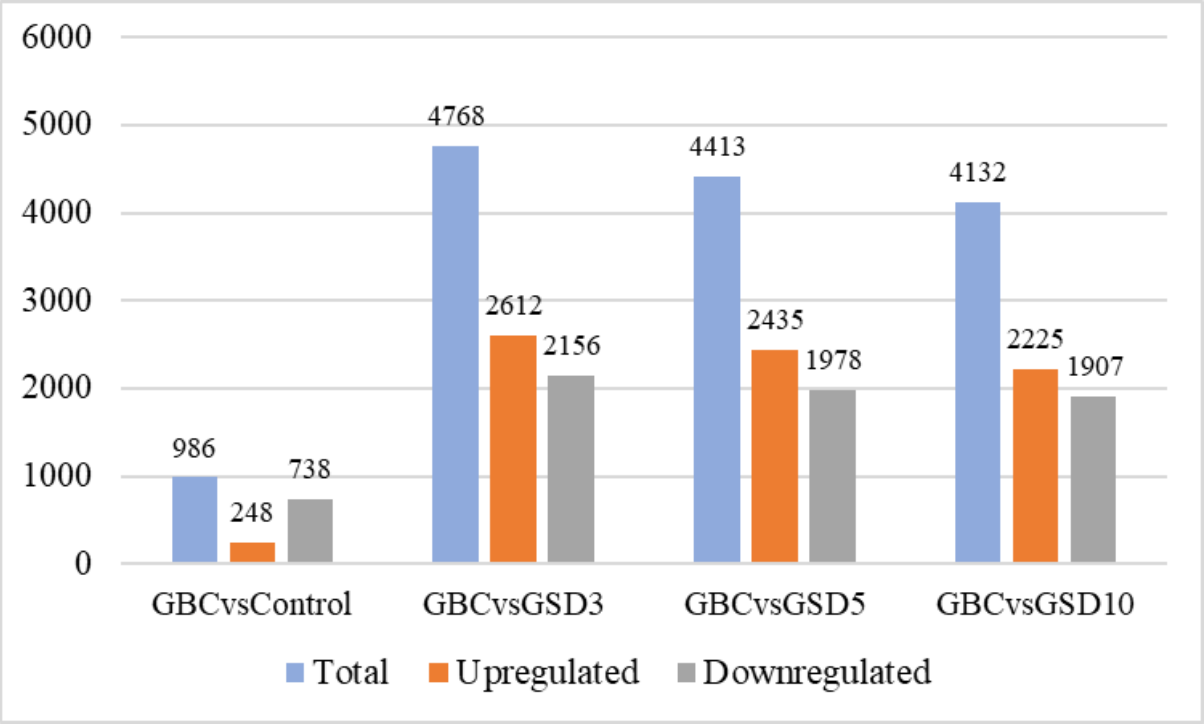
4.2.3.1 Dataset information and retrieval

The RNAseq dataset (SRP226150) comprising GBC and GSD samples was retrieved from the ENA database in Sequence Read Archive (SRA) format. The dataset contained a total of fifty samples obtained through surgical resection, which includes ten GBC tissues, ten adjacent normal tissue samples, and thirty GSD tissue samples from three different follow-up periods of 1-3 years (GSD3), 5-10 years (GSD5) and more than 10 years (GSD10). The Illumina HiSeq 2500 platform was used to generate the paired-end reads of these fifty samples.

4.2.3.2 Identification of differential molecular signatures in GBC compared to GSD with three different follow-up periods

The differential gene expression analysis identified significant DEGs taking p-adjusted value < 0.05 and log2FoldChange | 1 | in GBC vs. adjacent normal, GBC vs. GSD3, GBC vs. GSD5, and GBC vs. GSD10. The total DEGs including the total number of upregulated and downregulated DEGs have been represented in **Figure 4.13 A**. The number of downregulated DEGs identified in GBC vs. adjacent normal is higher as compared to downregulated DEGs. The hierarchical clustering analysis also showed that the expression pattern of DEGs in GBC is under-expressed as compared to normal samples **[Figure 4.13 B]**. The top 10 identified pathways [**Figure 4.13 C**] associated with significant DEGs in GBC vs. adjacent control were enriched in important cell signaling pathways such as cAMP signaling, AMPK signaling pathway, PPAR signaling, and Adipocytokine signaling pathways. These pathways are significantly associated with DEGs that are downregulated or suppressed in GBC compared to normal samples. In our previous case studies also, we have observed a pattern of high downregulation in GBC samples as compared to normal.

(A)



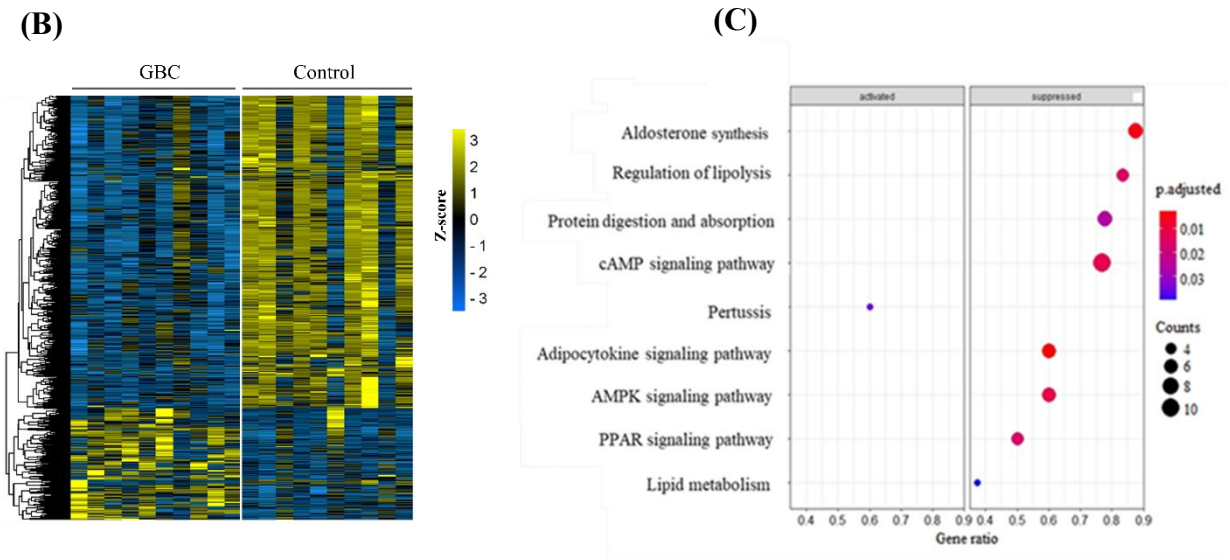


Figure 4.13: Identification of differential gene expression and pathways in GBC compared to normal. (A). Bar plot showing the number of total, upregulated, and downregulated DEGs identified in GBC as compared to adjacent normal samples. **(B)** Gene-wise hierarchical clustering of significant DEGs identified in GBC samples as compared to control samples. **(C)** Dot plot showing top ten significantly enriched pathways associated with DEGs identified in GBC vs. Normal.

4.2.3.3 Overlapping and unique DEGs in GBC compared to GSD with different follow-up periods.

To identify the unique and overlapping signatures in GBC vs. each GSD stage, we overlapped the DEGs identified in GBC vs. GSD3; GBC vs. GSD5, and GBC vs. GSD10 using venny 2.1. A total of 3102 overlapping and 824, 499, and 446 unique DEGs were identified in GBC vs. GSD3, GBC vs. GSD5, and GBC vs. GSD10 respectively [**Figure 4.14 A**]. The overlapping signatures are significantly involved in cell cycle processes [**Figure 4.14 B**]. The hierarchical clustering analysis of the top 500 significant unique DEGs revealed that there is significant variation in the expression pattern of DEGs identified in each GSD follow-up period as compared to GBC [**Figure 4.14 C**], which suggests that there might be a gradual change in gene expression patterns in each GSD stage that might result into a wide array of pathological spectrum and ultimately contributes to GBC pathogenesis and development.

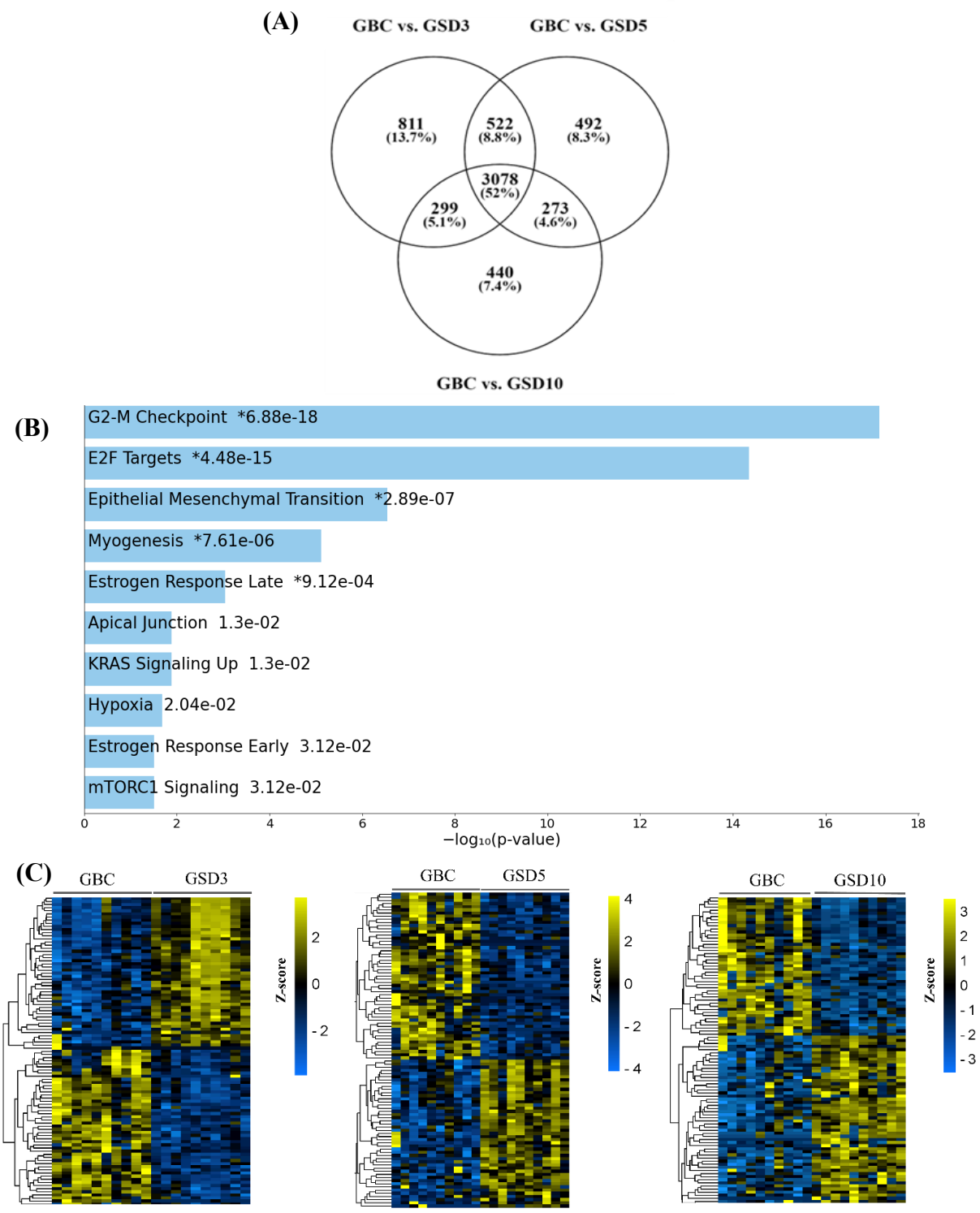


Figure 4.14: Identification of shared and unique differential gene expression profiles in GBC compared to GSD with three different follow-up periods. (A) Venn diagram showing the overlapping DEGs between GBC vs. GSD with 3 different follow-up periods. **(B)** Top 10 enriched pathways associated with the overlapping DEGs. **(C)** Complete linkage gene-wise hierarchical clustering of unique DEGs identified in GBC compared to GSD with 3 different follow-up periods.

4.2.3.4 Protein-protein interaction analysis and hub gene identification

The unique DEGs identified in GBC compared to that of GSD with different follow-up periods (GSD3, GSD5, and GSD10) were further considered to build a PPI network using the STRING database online tool (http://string-db.org/) to identify a panel of crucial gene sets in GBC. Only the queried DEGs were used to build the PPI networks with an effective binding score> 0.4. The interactive PPI networks were analyzed and visualized by Cytoscape v3.8.2 software.

CytoHubba, a cytoscape plugin was used to screen the hub DEGs from the PPI networks generated using DEGs identified from GBC vs. adjacent normal, GBC vs. GSD3, GBC vs. GSD5, and GBC vs. GSD10. The five topological parameters were considered to identify the predicted hub DEGs, which are maximum clique centrality (MCC), maximum neighbourhood component (MNC), Degree, edge percolated component (EPC), and Betweenness. The top-ranked 20 DEGs calculated from these five algorithms were considered. Finally, the predicted hub DEGs calculated from each topological parameter were further intersected for the identification of consensus-significant hub DEGs from the PPI networks. The list of significant hub genes is tabulated in **Table 4.10**.

Table 4.10: List of significant hub DEGs identified in GBC compared to GSD with three different follow-up periods through PPI network analysis.

DEGs	Degree	Betweenness	EPC	MNC	MCC
Hub DEGs identified from GBC vs. GSD3					
THBS1	34	5383.29	42.33	14	5803
TPT1	34	5573.66	39.02	15	4.79E+08
SERPINH1	32	2591.35	41.99	11	1.20E+05
COL1A1	28	2877.11	41.35	14	1.26E+06
Hub DEGs identified from GBC vs. GSD5					
GABRG2	22	2044.83	54.29	4	17
KIF5A	20	2916.83	54.35	7	17
HBEGF	10	1398.66	49.77	2	5
GJA1	10	2870.76	49.11	2	5
CX3CR1	10	1339.86	44.75	2	5
GRM1	8	873.93	51.76	2	4
GJA5	8	2180.73	48.38	2	4
HEY2	8	1620.96	48.99	2	4
Hub DEGs identified from GBC vs. GSD10					
GAPDH	42	8564.12	94.395	8	56
EGR2	22	3920.83	94.352	9	22
LCK	30	2634.42	94.395	14	905

NR4A1	16	1357.98	94.352	7	15
CCR7	24	965.62	94.395	11	787
IL17A	26	916.12	94.395	11	80
CD3E	26	876.45	94.395	12	789
IKZF1	28	773.78	94.395	13	949

The significant hub genes identified from PPI networks are mostly found to be downregulated. The hub DEGs identified in each group through PPI network analysis are involved in specific signaling components [**Figure 4.15**]. For instance, the hub DEGs identified in GBC vs. GSD3 are linked to extracellular matrix (ECM) signaling and components of focal adhesion such as integrins, whereas, the hub genes identified from GBC vs. GSD5 group are found to be linked to gap junction pathways. The majority of the hub DEGs identified in the GBC vs. GSD10 group are involved in TCR signaling and PI3K signaling pathways. These observations indicate that GSD progresses to GBC through the dysregulations of multiple signal transduction pathways at different stages with distinct pathological spectra.

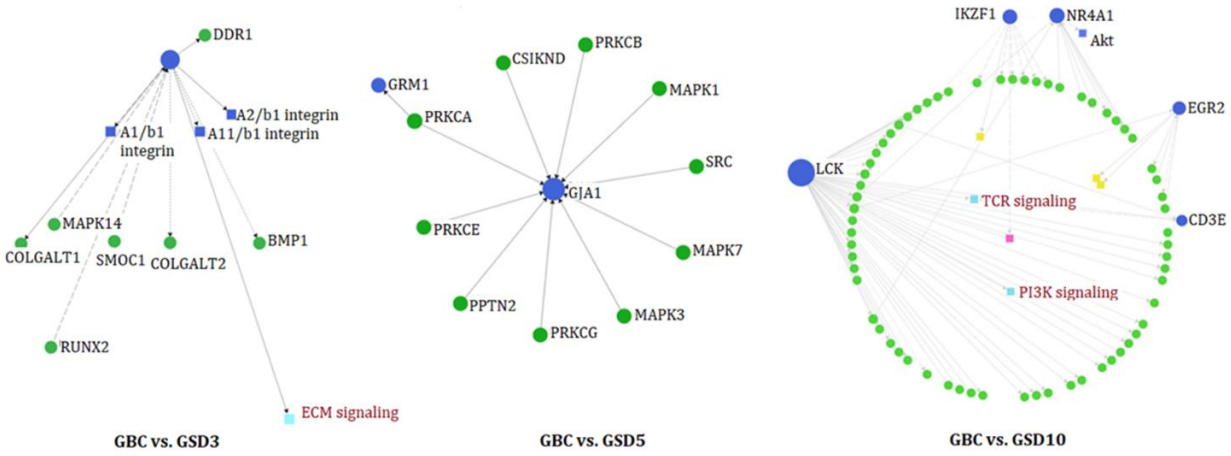


Figure 4.15: Signaling network showing the key signaling pathways and complexes associated with the hub DEGs identified in GBC. The green node represents associated proteins; light blue nodes indicate signaling pathways; circled blue nodes represent hub genes; squared blue nodes represent signaling complex and the yellow nodes indicate protein families.

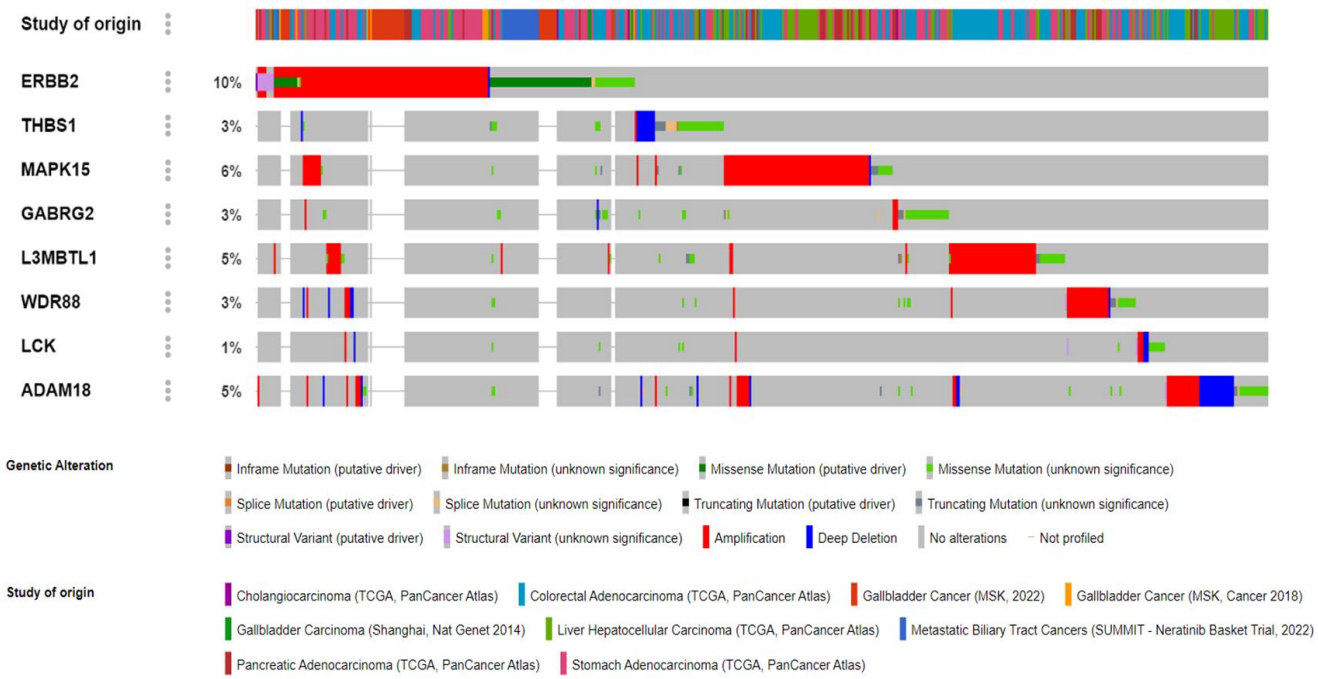
In case studies 2 and 3, the identified hub DEGs were found to be involved in pathways related to cell cycle and cell adhesion. However, case study 3 revealed that the hub DEGs in GBC compared to GSD in each follow-up period were associated with distinct processes and signaling pathways across each follow-up period including cell adhesion signaling (GSD3), MAPK signaling (GSD5) and immune response signaling (GSD10), which suggests that GSD

progress to GBC through the dysregulations of multiple signal transduction pathways at different stages (initiation-progression-metastasis) with distinct pathological spectrum.

4.2.3.5 In-silico validation of selected hub genes for each case study from TCGA datasets

The potential eight hub genes including *ERBB2*, *THBS1*, *MAPK15*, *GABRG2*, *L3MBTL1*, WDR88, LCK (lymphocyte-specific protein tyrosine kinase) and ADAM18 were further considered for in-silico validation using TCGA datasets of gastrointestinal cancers. The oncoprint tool of the cBioPortal database revealed that 28% (570/2041) of patients have genetic alteration that includes amplification, deletions, structural variations, and several other mutational types [**Figure 4.16 A**]. Furthermore, the expression level [**Figure 4.16 B**] of these eight altered DEGs in four gastrointestinal cancers in the TCGA database – pancreatic adenocarcinoma (TCGA-PAAD), colon adenocarcinoma (TCGA-COAD), liver hepatocellular carcinoma (TCGA- LIHC), and stomach adenocarcinoma (TCGA-STAD) were evaluated and it was found that the expression of the hub genes correlates with the level of hub gene expression in other gastrointestinal cancers. This implies that the identified hub DEGs play crucial roles in tumorigenesis, particularly in gastrointestinal cancers.

(A)



(B)

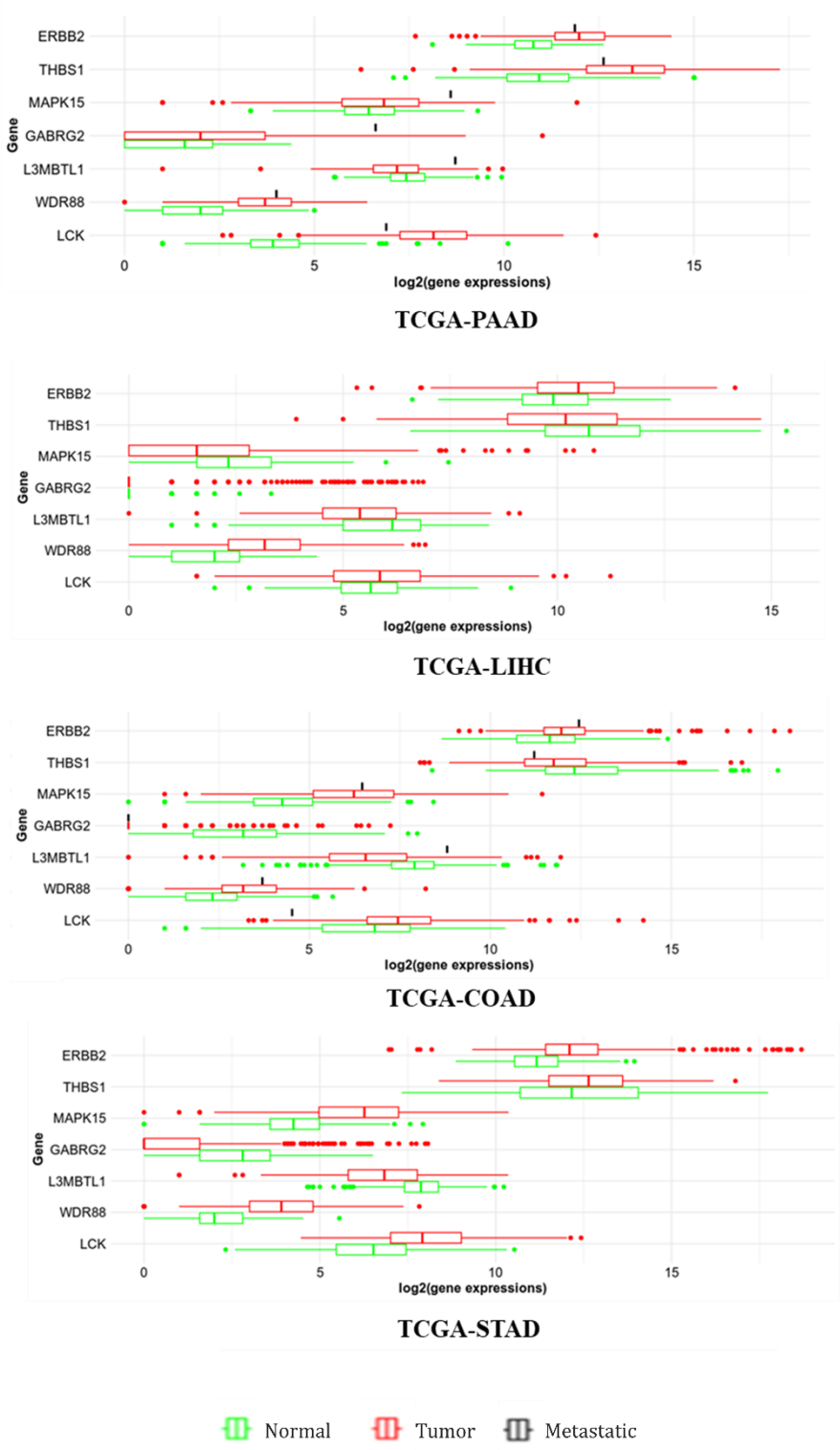


Figure 4.16: Validation of the expression and genetic alteration of the potential DEGs identified from case studies involving analysis of public transcriptomic datasets (case studies 1, 2, and 3). (A) The genetic alteration associated with hub DEGs identified in GBC. **(B)** Boxplot showing the gene expression level of the selected hub DEGs in tumor samples compared to normal from four TCGA gastrointestinal cancer datasets.

Case Study 4

Inhouse transcriptome data generation, analysis, and identification of differential molecular signatures in GBC and GBC+GS groups

4.2.4 Transcriptome sequencing and data analysis of GBC and GSD patients from Assam, North-East India.

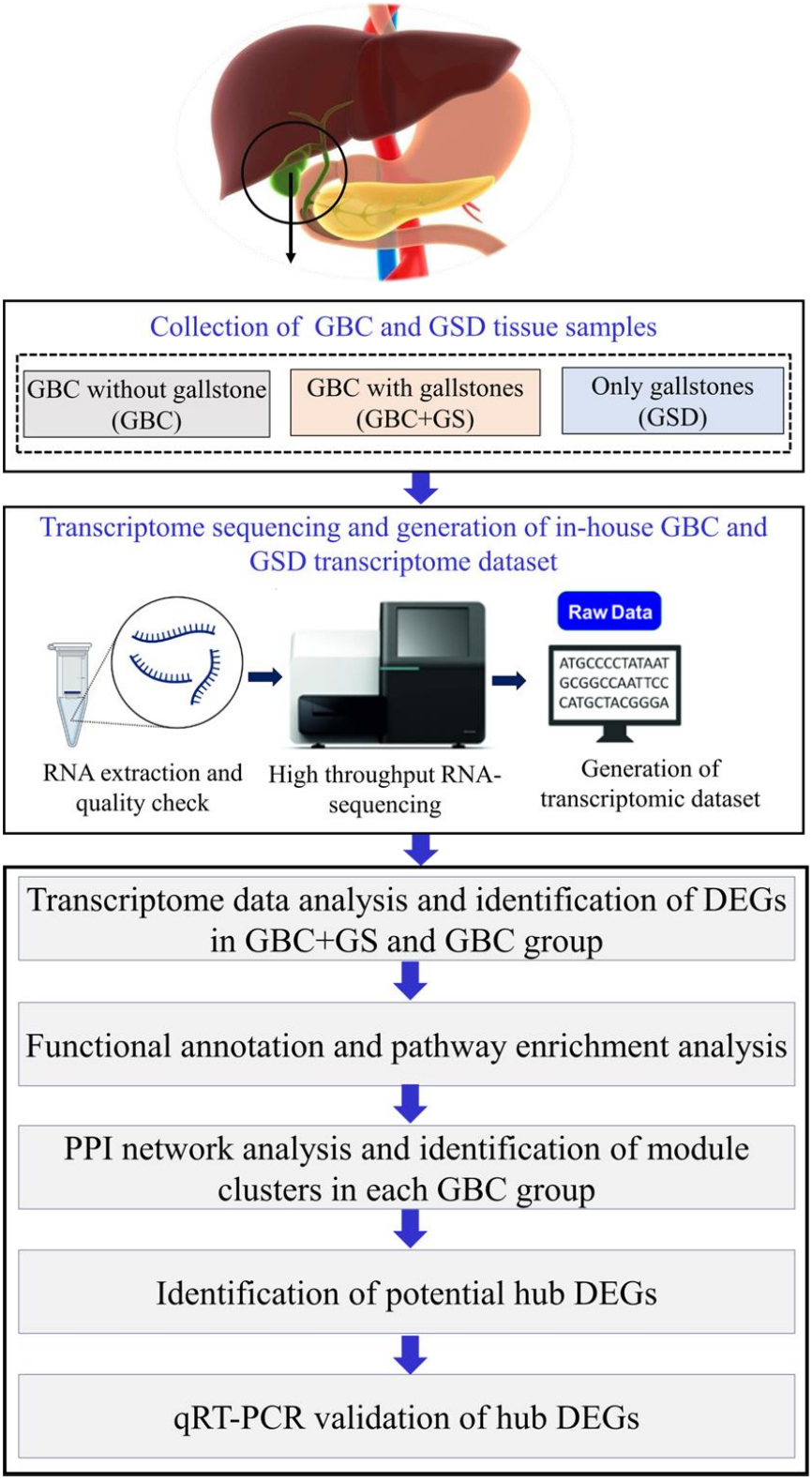


Figure 4.17: Schematic outline of the overall workflow carried out for identification of crucial molecular signatures in GBC from in-house generated GBC and GS transcriptomic dataset**.**

4.2.4.1 Collection of GBC and GS tissue samples

A total of fourteen tissue samples (n=14) were collected, including eight GBC (n=8) tissues and six GS (n=6) tissues from patients admitted at B. Borooah Cancer Institute and Swagat Super Speciality Hospital, Assam respectively. The tissue samples from GBC patients are categorized into two subgroups: the GBC+GS group and the GBC group. In this study, the GS samples without any malignancies were considered as control. The GBC and GS tissue samples were collected through radical cholecystectomy, ultrasonography (USG) guided biopsy, and simple cholecystectomy. All tissue samples included in this study were obtained with the approval of the ethical committee/institutional review board at BBCI and Tezpur University, and informed consent was taken from all the patients enrolled in this study. Out of fourteen, six were males and eight were females. Most of the GBC patients had jaundice and abdominal pain at the onset of the disease. All the patients had adenocarcinoma histopathology. Among the eight GBC patients, five had well-differentiated adenocarcinoma, two had moderately differentiated adenocarcinoma, and one had poorly differentiated adenocarcinoma. There was no family history of gallbladder diseases present in all the cases. The details of clinical samples considered for transcriptome sequencing are provided in **Table 4.11.**

Table 4.11: Clinical information of the GBC and GS tissue samples collected through surgical resection and USG-guided biopsy.

Sl. No.	Age	Sex	Study groups	Surgical procedure
P1	53	M	GS	Lap. Cholecystectomy
P2	42	F	GS	Lap. Cholecystectomy
P3	45	F	GS	Lap. Cholecystectomy
P4	35	M	GS	Lap. Cholecystectomy
P5	33	F	GS	Lap. Cholecystectomy
P6	44	F	GS	Lap. Cholecystectomy
P7	62	M	GBC+GS	Radical Cholecystectomy
P8	53	F	GBC+GS	Radical Cholecystectomy
P9	45	F	GBC+GS	Radical Cholecystectomy
P10	42	F	GBC+GS	USG guided biopsy
P11	43	F	GBC	USG guided biopsy
P12	45	M	GBC	USG guided biopsy
P13	52	F	GBC	Radical cholecystectomy
P14	35	M	GBC	USG guided biopsy

4.2.4.2 Transcriptome sequencing and generation of GBC transcriptome dataset.

For transcriptome sequencing, the total RNA was extracted from the tissue samples, and the quality of the total RNA was checked. The total RNA samples with good RNA integrity number (RIN) were considered for library preparation followed by transcriptome sequencing. The quality of the RNAseq libraries was checked and all the samples passed the quality check. **Figure 4.18** represents the visualization of the quality of RNAseq libraries along with their RIN values.

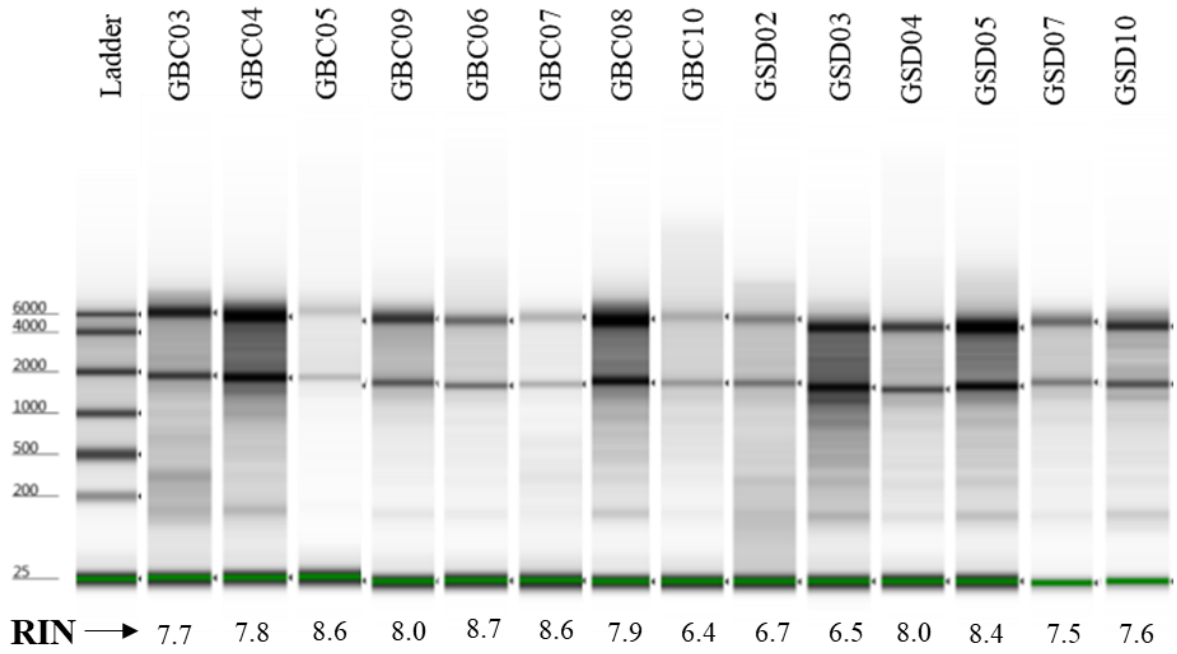


Figure 4.18: Visualization of the RNAseq libraries prepared from each sample in the TapeStation system

Transcriptome sequencing analysis of each of the GBC and GS tissue samples was performed using the Illumina Novaseq 6000 platform through 150 bp paired-end sequencing chemistry. After the removal of adapter sequences and low-quality and contaminating reads, the mapping of reads with the reference genome was performed. The clean reads were successfully aligned to the *Homo sapiens* reference genome with an average of 94.15% alignment rate for the fourteen RNAseq libraries [**Table 4.12**]. The FastQ files of the 14 samples have been submitted to the NCBI-SRA database (BioProject id: **PRJNA1015034**).

Table 4.12: Summary of reads generated from GBC and GS tissue samples generated through transcriptome sequencing.

Samples	Total clean reads	Number of total mapped reads
GBC03	47,251,785	44,585,632 (94.36%)
GBC04	59,602,954	55,907,612 (93.80%)
GBC05	53,878,134	51,365,067 (95.33%)
GBC09	56,162,639	45,301,606 (91.71%)
GBC06	81,341,286	77,422,757 (95.18%)
GBC07	66,659,907	64,023,334 (96.04%)
GBC08	62,655,913	60,202,025 (94.08%)
GBC10	63,556,245	60,305,378 (94.89%)
GS02	49,395,148	45,301,606 (91.71%)
GS03	61,061,770	57,786.768 (94.64%)
GS04	66,747,617	64,210,035 (96.20%)
GS05	71,683,338	65,441,963 (91.29%)
GS07	65,314,468	62,502,052 (95.69%)
GS10	45,565,898	42,496,764 (93.27%)

4.2.4.3 Identification of differential gene expression profile of GBC+GS and GBC patients from Assam

For the identification of potential DEGs associated with two distinct GBC subgroups (GBC+GS and GBC), transcriptome sequencing analysis was performed on four GBC+GS, four GBC, and six GS tissue samples using paired-end sequencing chemistry. In this study, the non-malignant GS samples were considered as controls. To identify crucial molecular signatures in GBC, filtration of lowly expressed mRNAs followed by differential gene expression (DGE) analysis was carried out in GBC+GS vs. control and GBC vs. control separately. DEG analysis identified 936 and 862 significant DEGs (Padj<0.05) in GBC+GS and GBC respectively [**Figure 4.19 A**]. Complete linkage hierarchical clustering revealed distinct expression patterns in GBC+GS and GBC groups [**Figure 4.19 B**]. The number of downregulated DEGs in both the GBC groups is much higher as compared to upregulated DEGs.

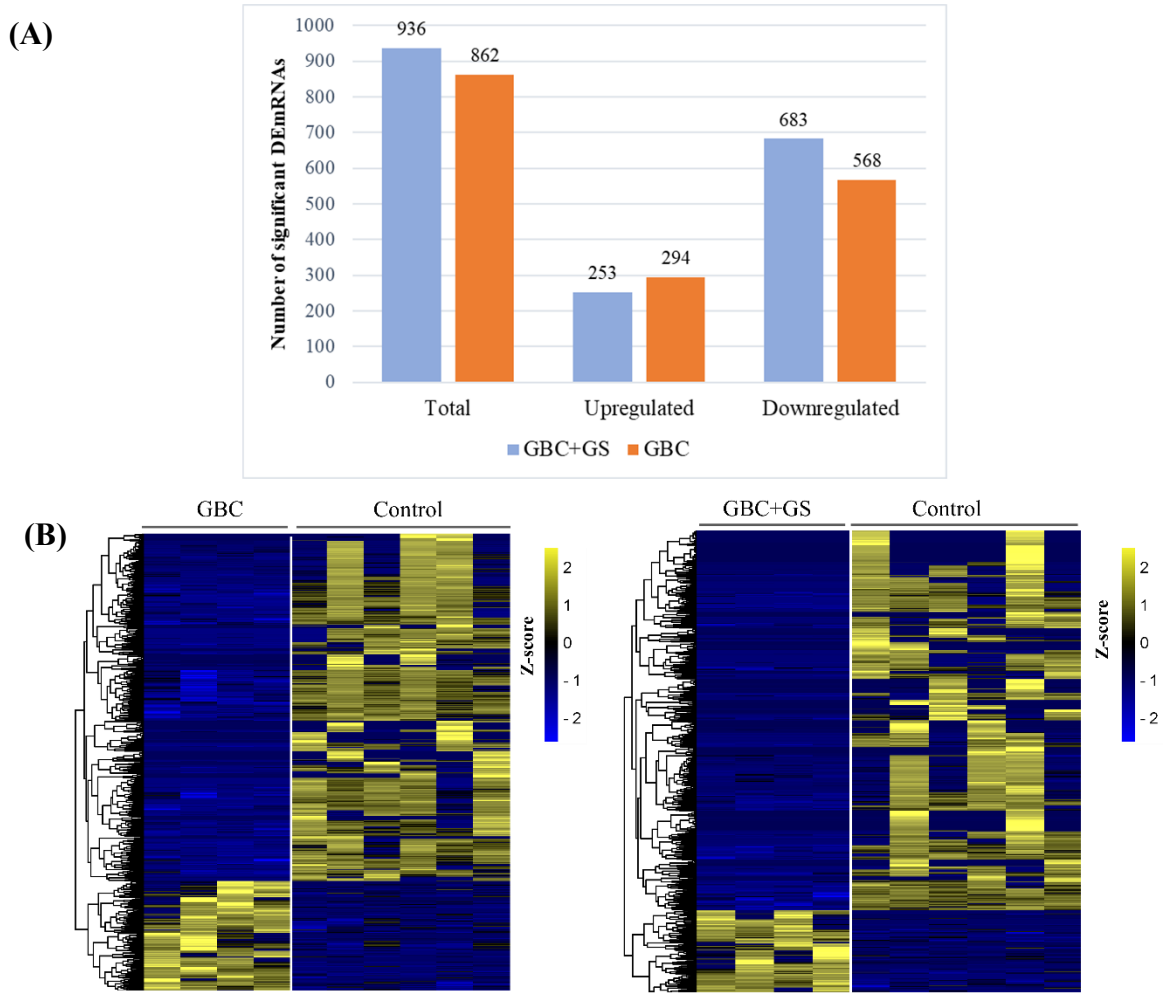


Figure 4.19: Identification of significant DEGs in GBC and GBC+GS group as compared to control. (A) Bar plot showing the number of DEGs identified in GBC+GS and GBC groups. **(B)** Hierarchical clustering analysis showing the expression profile of DEGs identified in the GBC and GBC+GS group compared to control (GS). The rows represent the gene expression counts and the columns represent the samples.

4.2.4.4 Identification of differentially expressed shared DEGs between GBC and GBC+GS groups.

The significant DEGs identified from the GBC+GS and GBC groups were further overlapped to identify the shared and unique DEGs among the two groups. A total of 188 overlapping/shared DEGs were identified between GBC+GS and GBC [**Figure 4.20 A**]. Interestingly, the trend of the expression patterns of the common DEGs is highly distinct among the two GBC groups [**Figure 4.20 B**]. The top five significant pathways found to be associated with the shared 188 DEGs are steroid hormone biosynthesis, bile secretion, IL-17 signaling,

retinol metabolism, and xenobiotics metabolism [**Figure 4.20 C**]. Retinol, bile secretion steroid hormones, and xenobiotics are crucial metabolic pathways associated with the functions of the liver and gallbladder. Therefore, the identified altered common gene sets associated with these pathways might play an important role in both GBC+GS and GBC development and progression.

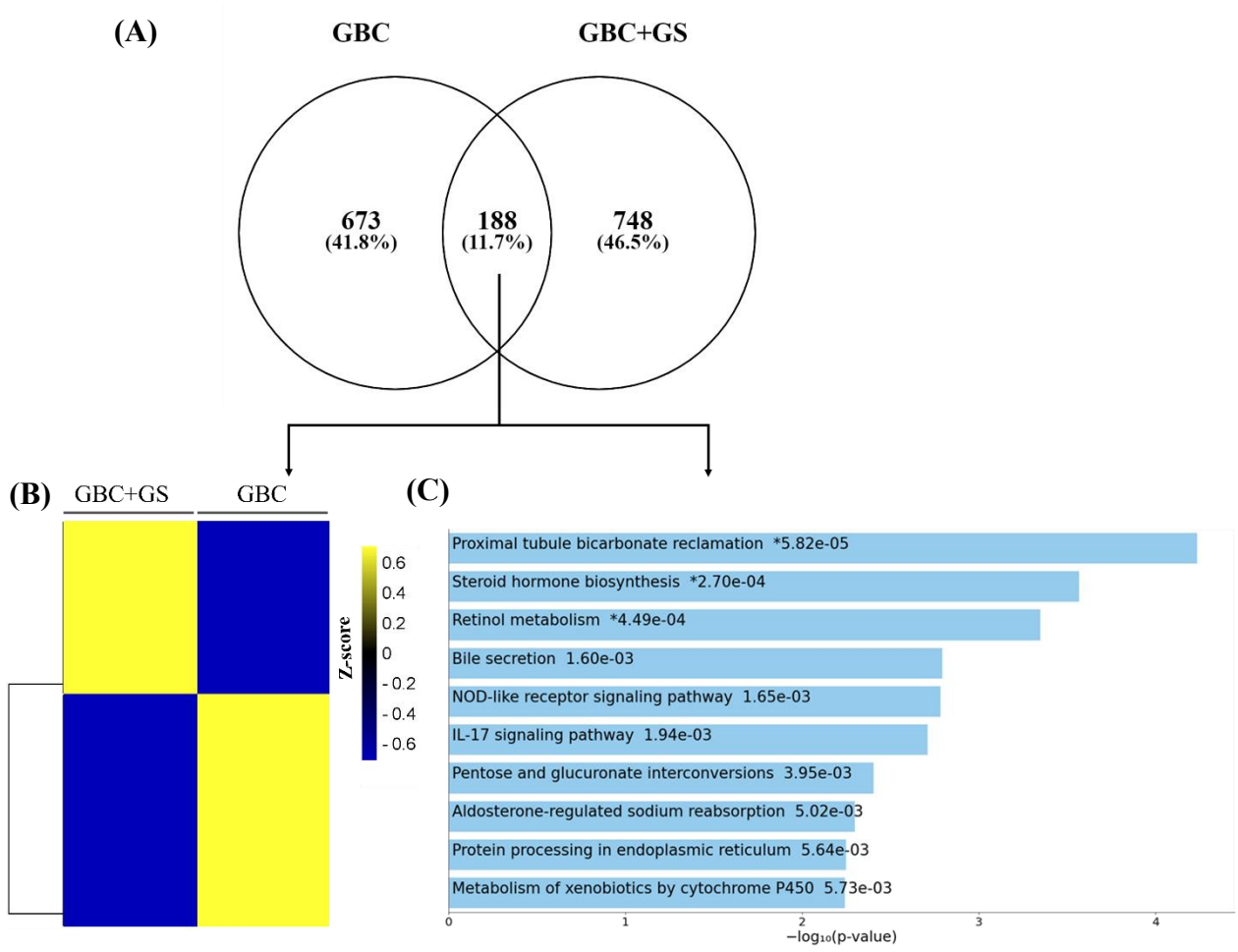


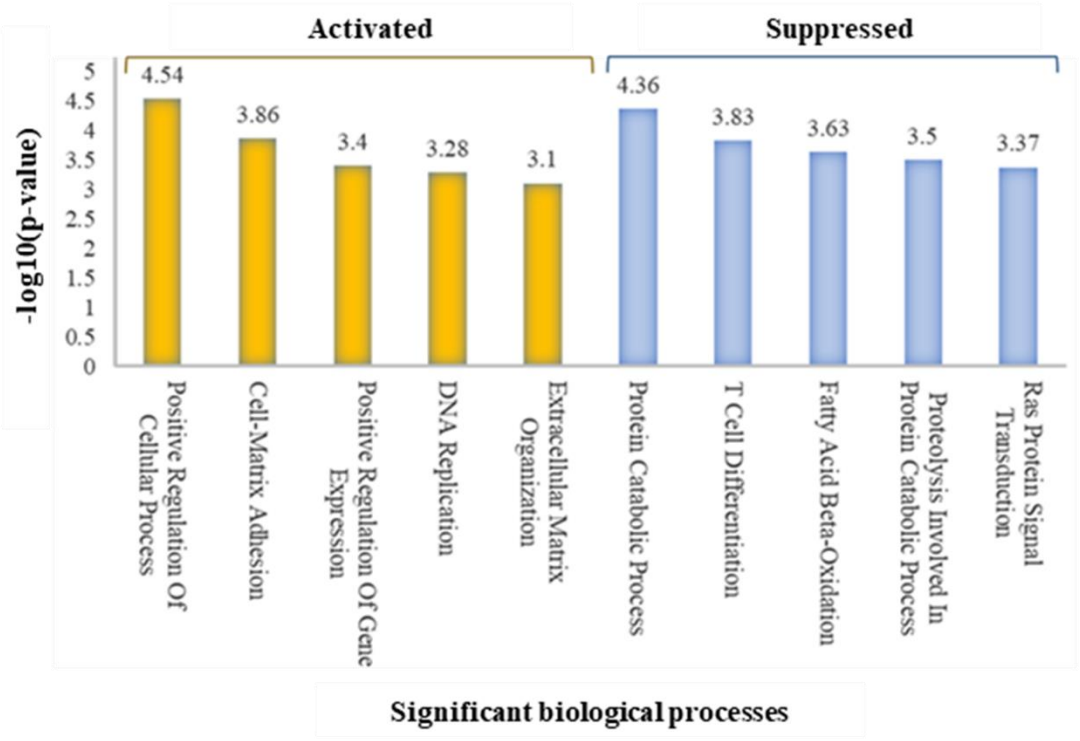
Figure 4.20: Identification of shared DEGs and pathways between GBC and GBC+GS groups. (A) The Venn diagram represents 188 shared DEGs between the two GBC groups. **(B)** Hierarchical clustering analysis showing heatmap of the expression level of the shared DEGs. **(C)** Bar plot representing the top ten significant pathways associated with the shared DEGs identified in both GBC and GBC+GS groups.

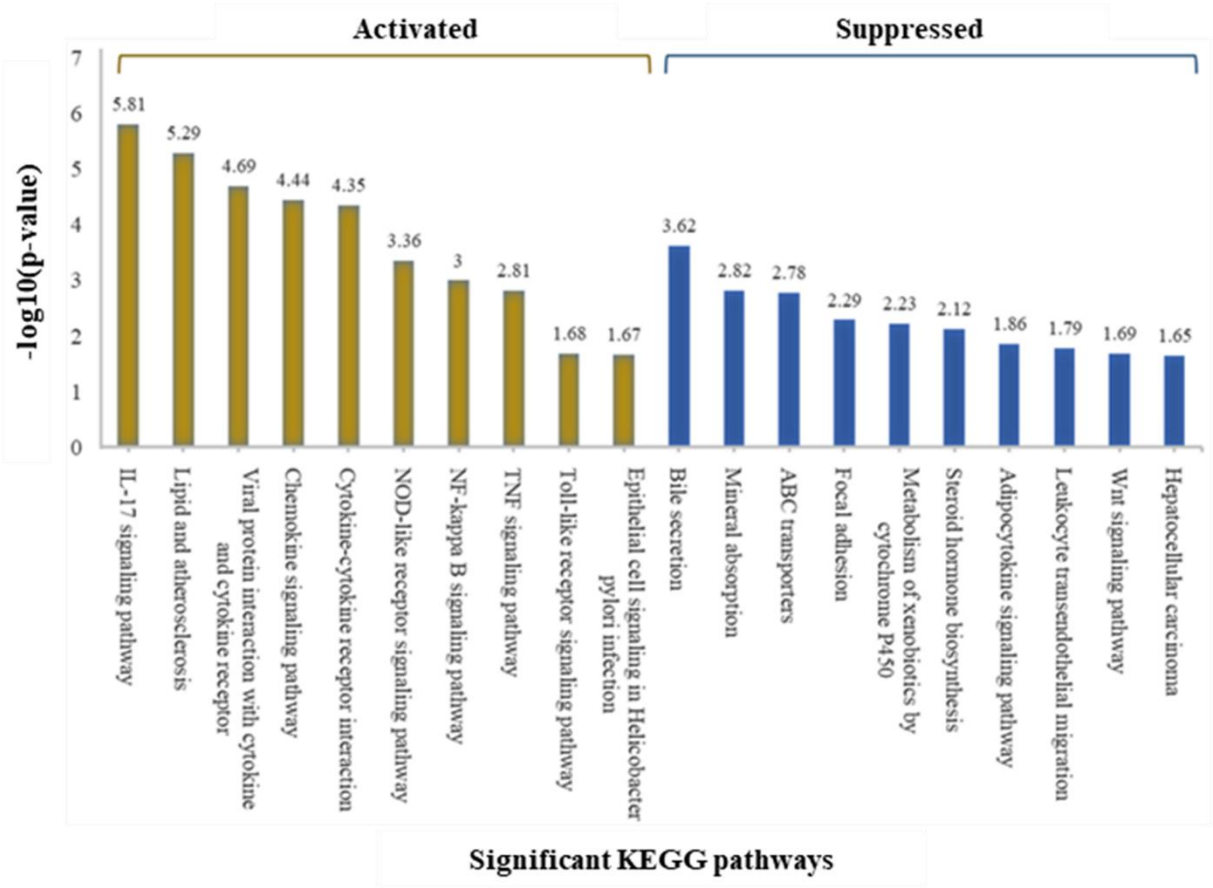
4.2.4.5 Functional annotation and pathway enrichment analysis of the DEGs identified in GBC+GS and GBC groups.

To better understand the role and functions of the DEGs identified in the GBC+GS and GBC groups, functional enrichment analysis was carried out to identify the top 10 significant biological processes (BP) and pathways associated with GBC pathogenesis. The analysis of the in-house generated transcriptome dataset on two subgroups of GBC patients showed that two GBC groups are associated with distinct pathological processes and pathways. Significant DEGs identified in the GBC+GS group are enriched in cell cycle processes and immune response-related pathways including IL-17 signaling, cytokine and chemokine, signaling, TNF signaling, Toll-like receptor signaling, and NF-kappa B signaling pathways [**Figure 4.21 A**]. These pathways are significantly activated and are intricately involved in inflammatory processes, suggesting that these pathways might play an important role in the progression of GBC from inflamed gallbladder with gallstones.

The DEGs identified in GBC revealed an association of distinct and diverse functions during pathogenesis [**Figure 4.21 B**]. Biological processes such as smooth muscle contraction, regulation of cell-matrix adhesion, and focal adhesion are significantly suppressed whereas, inflammatory responses, apoptotic processes, response to chemokine, and cell chemotaxis are highly activated. The upregulated DEGs are significantly associated with signal transduction processes such as AMPK signaling, PD-L1 signaling, NF-kappa B signaling, IL-17 signaling, and insulin signaling pathways. Whereas; focal adhesion, proteoglycans in cancer, salmonella infection, and T-cell receptor signaling were identified to be the significantly suppressed pathological pathways associated with the GBC patients without gallstone conditions.

Overall, the functional enrichment analysis revealed the involvement of distinct biological processes and pathways in the pathogenesis of GBC and GBC+GS cases. The DEGs identified in both groups are implicated in multiple pathways. The alteration of these diverse regulatory pathways promotes the aggressiveness of GBC and can contribute to the poor survival rates of GBC patients.


Activated
Suppressed
-log10(p-value)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
4.54
3.86
3.4
3.28
3.1
4.36
3.83
3.63
3.5
3.37
Positive Regulation Of Cellular Process
Cell-Matrix Adhesion
Positive Regulation Of Gene Expression
DNA Replication
Extracellular Matrix Organization
Protein Catabolic Process
T Cell Differentiation
Fatty Acid Beta-Oxidation
Proteolysis Involved In Protein Catabolic Process
Ras Protein Signal Transduction
Significant biological processes


Activated
Suppressed
-log10(p-value)
0
1
2
3
4
5
6
7
5.81
5.29
4.69
4.44
4.35
3.36
3
2.81
1.68
1.67
3.62
2.82
2.78
2.29
2.23
2.12
1.86
1.79
1.69
1.65
IL-17 signaling pathway
Lipid and atherosclerosis
Viral protein interaction with cytokine and cytokine receptor
Chemokine signaling pathway
Cytokine-cytokine receptor interaction
NOD-like receptor signaling pathway
NF-kappa B signaling pathway
TNF signaling pathway
Toll-like receptor signaling pathway
Epithelial cell signaling in Helicobacter pylori infection
Bile secretion
Mineral absorption
ABC transporters
Focal adhesion
Metabolism of xenobiotics by cytochrome P450
Steroid hormone biosynthesis
Adipocytokine signaling pathway
Leukocyte transendothelial migration
Wnt signaling pathway
Hepatocellular carcinoma
Significant KEGG pathways

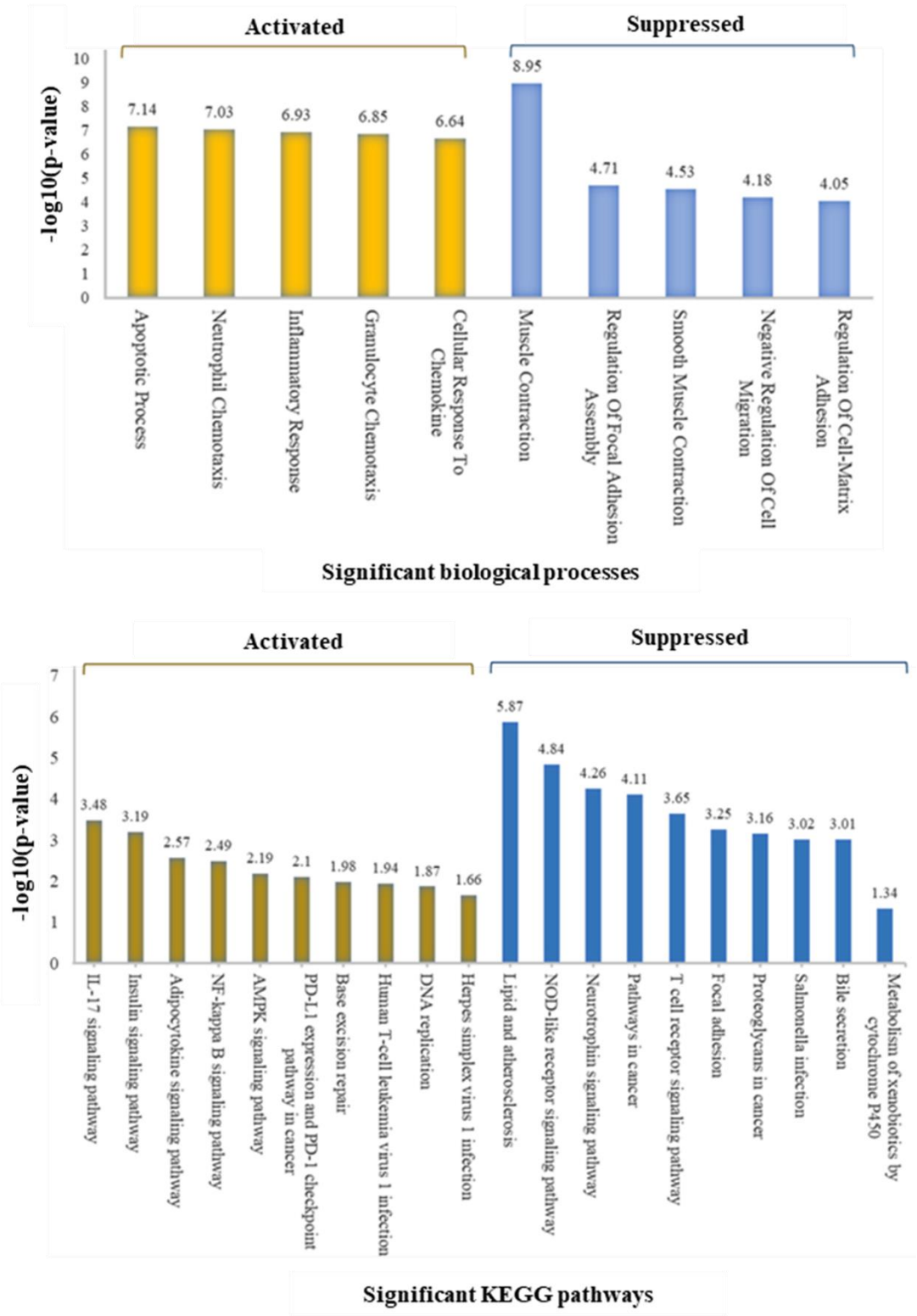


Figure 4.21: **Functional enrichment of significant DEmRNAs.** Bar plots representing the top ten significantly enriched biological processes and pathways associated with the upregulated (activated) and downregulated (suppressed) DEmRNAs in GBC+GS **(A)** and GBC **(B)** groups. The x-axis and y-axis represent the significantly enriched processes and pathways and p-values (log_{10} transformed) respectively.

4.2.4.6 Construction of PPI network modules and identification of hub genes in GBC+GS and GBC group

The significant DEGs identified in the GBC+GS and GBC groups were used to construct the PPI networks using the STRING database. Only the queried DEGs were considered to build the PPI network with a binding score > 0.50. The binding score measure the significance of the the PPI interaction between two proteins. From the whole PPI network; a significant functional module cluster with the highest cluster score was extracted using the MCODE algorithm [**Figure 4.22**]. The module cluster identified from the GBC PPI network comprises 19 DEGs with a cluster score of 6.56, the module cluster identified from the GBC+GS PPI network includes 17 DEGs and a cluster score of 7.62 and the PPI module with common DEGs contains 13 DEGs with a cluster score of 4.0

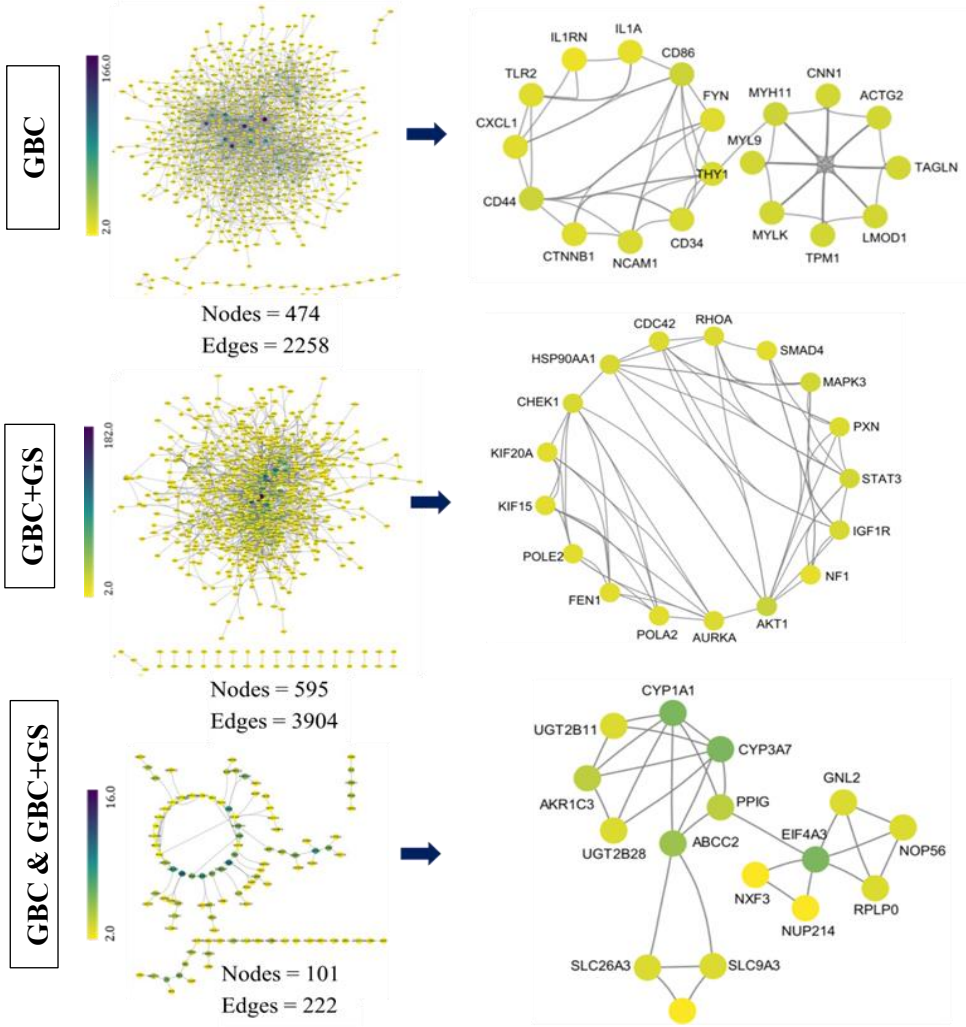


Figure 4.22: Construction of PPI networks and identification of significant module clusters from the whole PPI network. The color scale represents the degree of centrality of the interconnected nodes.

Furthermore; CytoHubba, a Cytoscape plugin was used to screen the hub DEGs from the PPI cluster module identified in each group. The five network topological measures- MCC, MNC, Degree, closeness, and Betweeness were considered for screening hub DEGs. Finally, the top 20 predicted hub DEGs calculated from each topological measure were intersected for the identification of significant hub DEGs from the PPI networks. A total of 8, 9, and 3 hub DEGs have been identified from GBC, GBC+GS, and overlapping groups [**Table 4.13**]. Interestingly, all the identified hub genes except *CXCL1*, *CD86* (GBC), and *STAT3* (GBC+GS) are significantly downregulated in GBC groups compared to control, indicating that downregulation caused by mutation or epigenetic methylation of DEGs is a key characteristic of GBC pathogenesis.

Table: 4.13: List of hub DEGs identified through PPI network analysis by taking the consensus of five network topology measures.

Hub DEGs	Degree	MCC	MNC	Betweenness	Closeness
GBC group					
MYH11	16	5041	7	154	12.08
CD86	16	60	8	68.09	11.83
LMOD1	14	5040	7	0	10.36
MYL9	14	5040	7	0	10.36
NCAM1	12	96	6	4.28	10.66
THY1	12	96	6	4.28	10.66
CXCL1	10	18	5	13.21	9.58
TLR2	10	18	5	13.21	9.58
GBC+GS group					
AKT1	22	1588	11	90.48	22
STAT3	18	1584	9	3.48	11.66
MAPK3	18	1584	9	3.48	11.66
SMAD4	12	144	6	0.78	10.16
HSP90AA1	16	1442	8	26.28	16
IGF1R	16	840	8	2.7	16
CDC42	14	1440	7	0.28	14
AURKA	14	722	7	35	14
CHEK1	16	724	8	66	14
Shared hub genes identified among the GBC+GS and GBC group					
E1F43	6	9	3	82	8.5
CYP1A1	6	18	6	24.66	8.16
CYP3A7	6	18	6	24.66	8.16
AKRIC3	4	12	4	0.66	6.58
PPIG	4	7	3	72	8

4.2.4.7 Identification of the crucial pathological pathways associated with gene sets in PPI network modules

The significant modules identified through PPI network analysis with the DEGs identified in GBC+GS and GBC were further analyzed to identify key biological pathways associated with the DEGs [**Figure 4.23**]. The pathways analysis of the gene sets identified in the module cluster revealed that the gene set module cluster in the GBC group is significantly linked to cell adhesion signaling such as vascular smooth muscle contraction, focal adhesion, leukocyte transendothelial migration, and cell adhesion molecules pathways. However, oncogenic signaling pathways including Ras signaling, Neurotrophin signaling, cancer proteoglycans, and pancreatic cancer are highly enriched. Interestingly, the gene set identified in the PPI module cluster constructed from the overlapping DEGs between the two groups is significantly associated with metabolic pathways which mainly include xenobiotics metabolism, retinol metabolism, bile secretion, and steroid hormone metabolism.

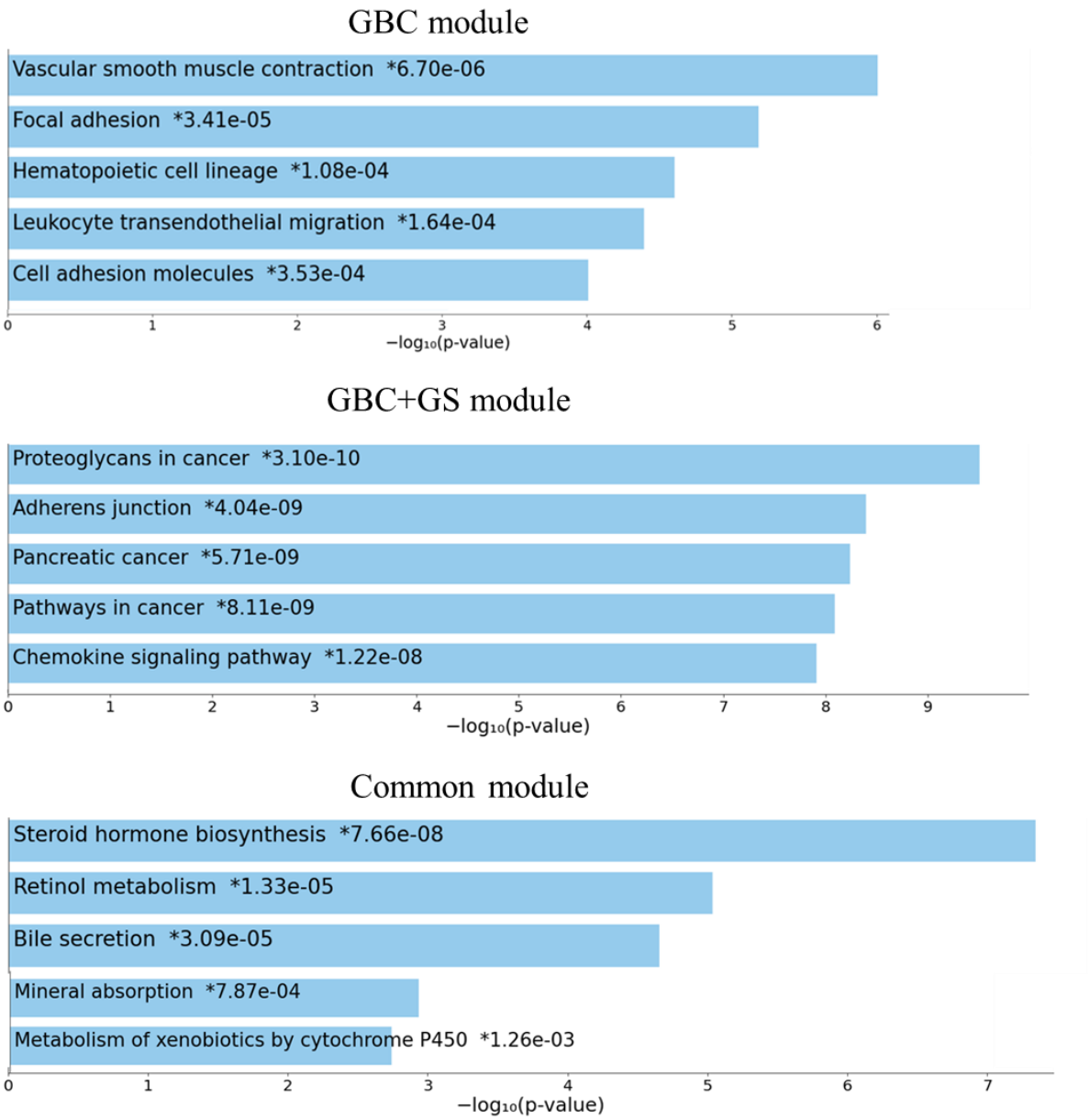


Figure 4.23: Bar plot showing a list of the top five significant pathways associated with the gene sets identified from PPI module clusters. The x-axis and y-axis represent the pathways and p-values respectively.

4.2.4.8 Validation of the selected hub DEGS through qRT-PCR.

The expression level of two significant hub DEGs- Leiomodin 1 (*LMOD1*) and *SMAD4* identified in the GBC and GBC+GS groups were further validated through quantitative Real-time (qRT)-PCR analysis and it was found that the expression level of *LMOD1* and *SMAD4* identified through RNAseq data analysis correlates with that of qPCR data analysis [**Figure 4.24]**.

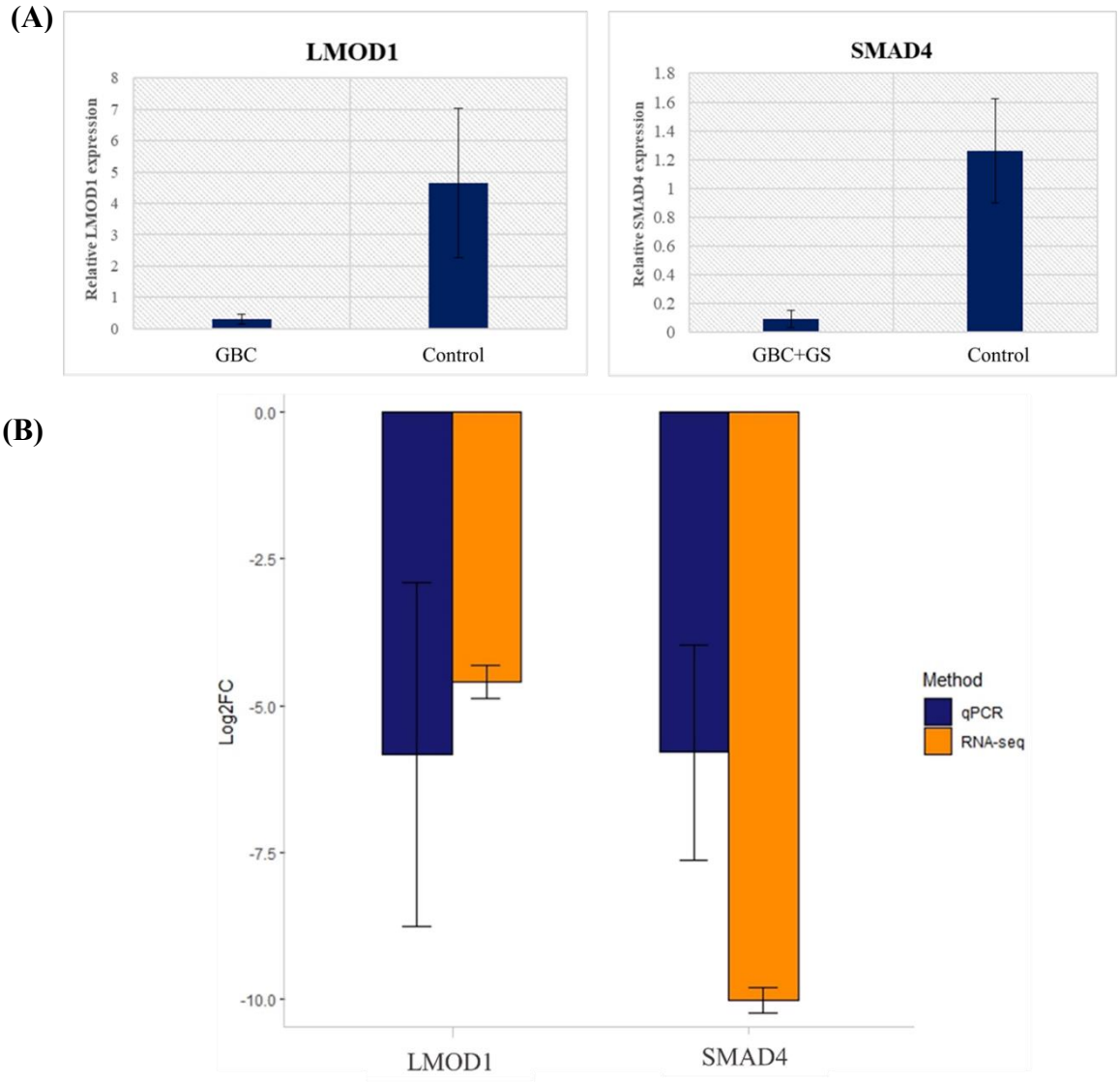


Figure 4.24: qRT-PCR validation of the hub DEGs identified in the GBC and GBC+GS group. (A) Bar plot representing the relative expression of *LMOD1* and *SMAD4* identified using qRT-PCR data analysis in GBC compared to control. **(B)** Bar plot showing the gene expression level (log2FoldChange) of *LMOD1* and *SMAD4* identified through RNAseq and qRT-PCR.

In case study 4, it was found that the differential expression pattern and associated pathways are distinct in gallstone-associated GBC and de-novo GBC development. Cell-adhesion-based molecules and cell cycle regulators are identified as key players in the GBC and GBC+GS groups respectively. *SMAD4* and *LMOD1* have been identified as crucial molecules that act as a tumor suppressor in GBC.

4.3 Discussions

GBC is one of the most fatal malignancies of the biliary tract system, ranking sixth among other gastrointestinal cancers. GBC etiology is multifactorial and it progresses mainly through metaplasia to dysplasia carcinoma sequence. GS is a major risk factor, often incidentally discovered in cases. Despite this, the precise molecular mechanism that drives GBC pathogenesis and progression remains unclear, creating a gap in the development of targeted therapies specific to GBC patients. Therefore, an integrative systems-level approach was adopted to identify the molecular underpinnings, aiming to identify crucial genes, processes, and pathways linked with GBC development.

Analyzing public and in-house GBC transcriptome datasets revealed that the majority of the DEGs are notably downregulated in GBC compared to controls. The identified hub DEGs from public GBC datasets are significantly associated with cell cycle signaling pathways. The cell cycle, governing cell division and growth, involves a complex network of proteins like CDKs and CDKIs that regulate genome integrity. Dysregulation in these components is observed in various malignancies, contributing to malignant transformation and therapy resistance [12-13]. Dysregulations in cell cycle components, including uncontrolled proliferation due to altered cyclin-dependent kinase (CDK) function or diminished CDK Inhibitor (CDKI) activity, are well-established in cancer development [14-15]. Genes involved in DNA replication and mitosis are pivotal in influencing cell cycle phase transitions, making them prime targets for mechanism-based therapies [16]. However, the initiation of GBC pathogenesis is not solely tied to cell cycle components. Carcinogenesis is also linked to dysfunctions in cell death machinery and cellular interactions, which interact with and augment cell cycle defects [17-18]. The identification of hub DEGs through differential gene co-expression networks and protein-protein interaction analysis revealed direct or indirect associations of potential hub DEGs with cell cycle components, apoptotic regulation, and cell-cell adhesion. These connections lead to uncontrolled cell proliferation, culminating in full-scale gallbladder malignancy.

The analysis of the in-house generated transcriptome dataset on two subgroups of GBC patients (GBC & GBC+GS) in comparison to the control (GSD) showed that the significant DEGs are strongly linked to pathways involving cell adhesion, including apical junctions, EMT, and immune signaling pathways such as TNF and JAK/STAT3 pathways. Hub DEmRNAs including *LMOD1*, *MYH11*, *THY1*, *MYL9*, and *NCAM1*—integral to cell adhesion

processes—showed substantial downregulation in GBC as compared to the control. Whereas, immune response-related DEGs such as *CXCL1* and *CD86* exhibited significant overexpression, while TLR2 showed marked downregulation in GBC. *CXCL1*, an inflammatory chemokine, has a significant role in tumor growth regulation and metastasis. Chemokines, like *CXCL1*, influence various aspects of tumors, from growth to metastasis, by regulating angiogenesis and tumor leukocyte interactions. They achieve this by binding to specific receptors, eliciting cellular responses like migration and adhesion [19,20]. *TLR2*, known for inducing cell adhesion, endothelial barrier disruption, and migration in cancer cells [21], also demonstrated decreased expression in GBC. The majority of hub DEGs identified in the GBC group participate in either cell adhesion or cell-chemotaxis signaling pathways. Cell adhesion molecules (CAMs), pivotal in cell-to-cell and cell-to-ECM interactions, impact malignant cell adhesion properties crucial for cancer development and progression [22]. Modifications in CAM expression or function influence not only cell adhesion but also signaling pathways involved in tumor invasion and metastasis [23]. Despite challenges in delineating a singular mechanism, CAMs' clinical potential as prognostic biomarkers or therapeutic targets in malignancies is a focus of extensive research [24–26]. The metastatic spread remains one of the most life-threatening pathological events for cancer patients, particularly for asymptomatic and aggressive cancers like GBC. In recent years, major efforts have been taken to unravel the molecular mechanism involved in metastasis, which includes tumor cell detachment, invasion, dissemination, and extravasation [27] Each of these steps is associated with the unique molecular program in which the modulation of the CAMs, and cytoskeletal properties of the disseminating tumor cells play essential roles. The primary mediators of CAMs induce EMT, which is one of the most crucial processes in cancer cell migration, invasion, and metastasis [27-28].

Among the eight hub genes identified in the GBC group, *LMOD1* was considered for validation through qRT-PCR analysis and it was found that the expression level of *LMOD1* identified through RNAseq data analysis correlates with that of qPCR data. The expression of *LMOD1* is significantly downregulated in GBC compared to control. *LMOD1*, with putative transmembrane properties and repetitive blocks, was initially discovered as a 64 kDa autoantigen in Hashimoto's thyroiditis patients [29]. A recent study identified *LMOD1* as a target gene specifically in smooth muscle cells, associated with the serum response factor/myocardial protein. *LMOD1* has been reported to play a critical role in the diagnosis and prediction of survival outcomes in early-onset colorectal cancer [30] and in distinguishing

between low-grade and high-grade prostate cancer cases [31]. Additionally, *LMOD1* was found to control the FAK-Akt/mTOR pathway and induce EMT in gastric cancer cells [32]. However, the specific role of *LMOD1* in GBC remains unreported.

The hub DEGs identified in the GBC+GS group were found to be associated with pathways governing diverse signal transduction processes and cell cycle regulatory pathways. From the nine identified hub DEGs in the GBC+GS group, the expression of *SMAD4* was considered for validation through qRT-PCR. The expression of *SMAD4* was found to be highly downregulated in GBC in both RNAseq data and qRT-PCR analysis. Originally known as *DPC4* (deleted in pancreatic carcinoma, locus 4), *SMAD4* was first recognized as TSG in pancreatic cancer by Harn et al., in 1996 [33]. *SMAD4* belongs to the Smad family which plays important roles in mediating the TGF-β signaling pathway that curbs epithelial cell development. It is a critical tumor suppressor and a key component of the transforming growth factor-β (TGF-β) pathway. The role of TGF-β in regulating the initiation, progression, and prognosis outcome of human malignancies is well-established but complex. It involves controlling the autonomous, local, and systemic cellular responses [34–36]. TGF-β plays a dual role in carcinogenesis. Firstly, it functions as a tumor suppressor during the initiation and early stages of a tumor by impeding proliferation and promoting apoptosis. Secondly, in an advanced stage of cancer progression, increased production of TGF-β contributes to tumor invasion, angiogenesis, and immune system evasion [37]. Alterations in the *SMAD4* gene have been observed across various cancers, including those affecting the biliary tract system [38–41]. Interestingly, *SMAD4* has not been identified in the GBC group, which implies that dysregulation in the *SMAD4* gene could be a contributing factor that promotes gallbladder carcinogenesis from gallstones.

Furthermore; the overlapping DEGs identified in the PPI module are significantly downregulated in both GBC groups and were found to be enriched in xenobiotics metabolism. Xenobiotic metabolism enzymes (XMEs) are crucial agents involved in processing foreign substances within the body. Studies demonstrated that *CYP450* can activate proto-oncogenes by oxidizing them, converting them into reactive compounds that irreversibly interact with cellular components [42]. Cancer cells sustain drug resistance by altering P450 expression, making these enzymes potential targets for anticancer therapy. Dysregulations in the xenobiotic metabolic pathways have been found to have significant impacts on drug-related toxicity and susceptibility to cancer [43]. The overlapping hub DEGs- cytochrome P450 family 1 subfamily

A member 1 (*CYP1A1*), cytochrome P450 3A4 (*CYP3A7*), and Aldo-keto reductase family 1 member C3 (*AKR1C3*), identified across both GBC+GS and GBC groups, are strongly associated with xenobiotic metabolism. Studies have shown that *CYP17* rs743572 is an important candidate gene in GBC in India as well as in the Chinese population [44]. Additionally, studies on steroid hormone synthesis, metabolism, and transport in a Chinese population revealed that the *CYP1A1* variant genotype (rs2606345) significantly influences GBC susceptibility [45].

4.4 Summary

Overall, this chapter unravels the differential gene expression signatures and associated biological pathways in GBC pathogenesis. Both public, as well as in-house generated transcriptomic datasets have been analyzed to identify crucial genes that can act as potential diagnostic and therapeutic biomarkers. The transcriptomic data analysis of the publicly available GBC dataset identified potential cell cycle regulatory genes whereas, the inhouse generated transcriptome dataset from GBC+GS and GBC patients from Assam revealed that the two GBC groups show distinct differential gene expression patterns and biological processes and pathways [**Figure 4.25**]. This is the first comparative study on two subgroups of GBC which showed that gallstones-associated and gallstone-independent gallbladder carcinogenesis progresses and is associated with distinct pathological pathways.

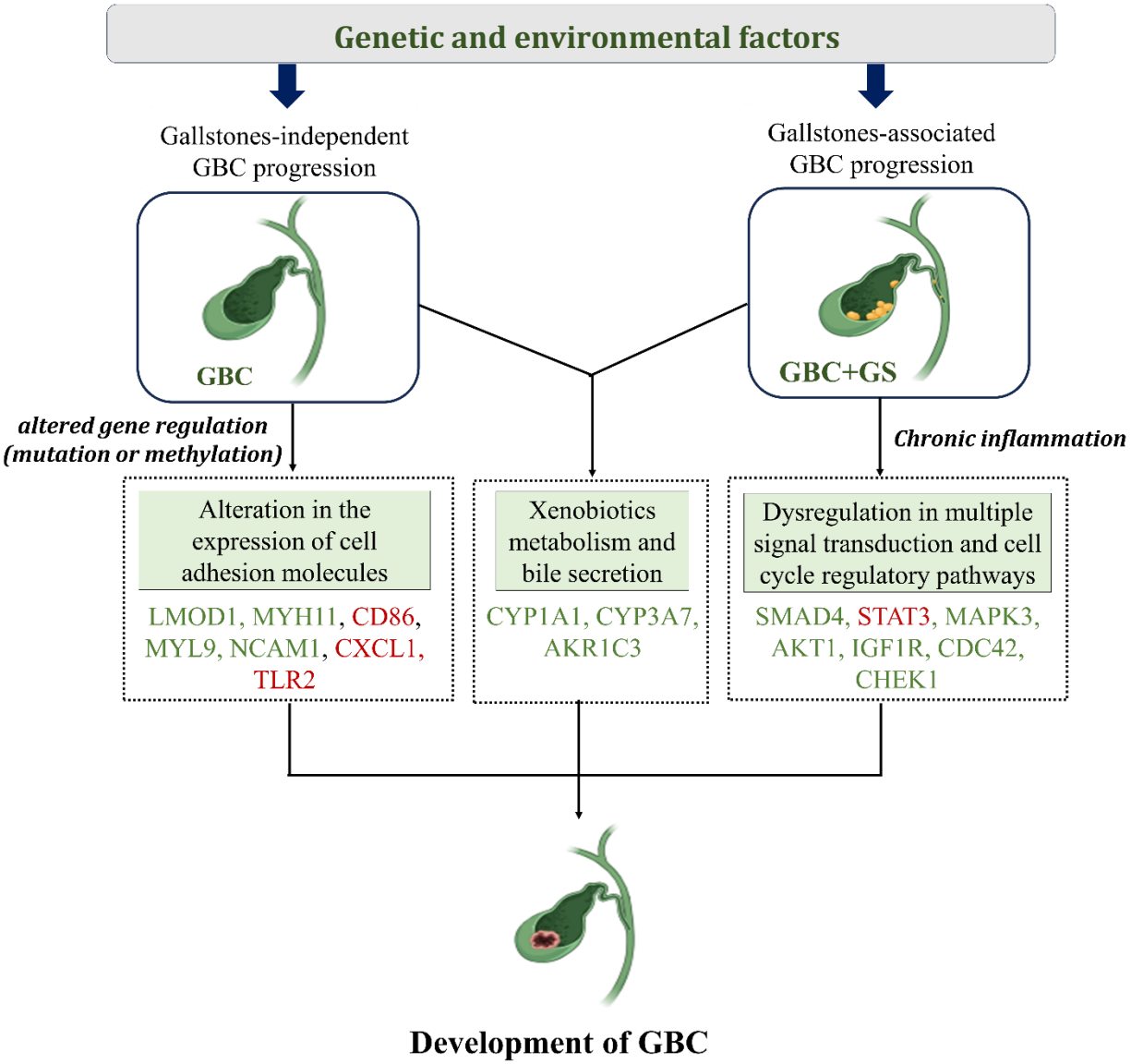


Figure 4.25: Schematic diagram illustrating gallstones-associated and gallstones-independent GBC development progresses through distinct molecular pathogenesis involving distinct gene sets and diverse biological pathways. The upregulated and downregulated genes are highlighted in red and green font respectively.

Bibliography

[1] Mao, W., Deng, F., Wang, D., Gao, L., & Shi, X. Treatment of advanced gallbladder cancer: a SEER-based study. Cancer medicine, 9(1), 141-150, 2020.

[2] Chantarojanasiri, T., Hirooka, Y., Kawashima, H., Ohno, E., Kongkam, P., & Goto, H. The role of endoscopic ultrasound in the diagnosis of gallbladder diseases. Journal of Medical Ultrasonics, 44, 63-70, 2017.

[3] Hundal, R., & Shaffer, E. A. Gallbladder cancer: epidemiology and outcome. *Clinical epidemiology*, 99-109, 2014.

[4] Sharma, A., Sharma, K. L., Gupta, A., Yadav, A., & Kumar, A. Gallbladder cancer epidemiology, pathogenesis and molecular genetics: Recent update. *World journal of gastroenterology*, 23(22): 3978, 2017.

[5] Xu, S., Zhan, M., Jiang, C., He, M., Yang, L., Shen, H., ... & Wang, J. Genome-wide CRISPR screen identifies ELP5 as a determinant of gemcitabine sensitivity in gallbladder cancer. *Nature communications*, 10(1): 5492, 2019.

[6] Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. Current achievements and applications of transcriptomics in personalized cancer medicine. *International Journal of Molecular Sciences*, 22(3): 1422, 2021.

[7] Pandey, A., Stawiski, E. W., Durinck, S., Gowda, H., Goldstein, L. D., Barbhuiya, M. A., et.al. Integrated genomic analysis reveals mutated ELF3 as a potential gallbladder cancer vaccine candidate. *Nature communications*, 11(1): 4225, 2020.

[8] Yang, P., Javle, M., Pang, F., Zhao, W., Abdel-Wahab, R., Chen, X., et.al. Somatic genetic aberrations in gallbladder cancer: comparison between Chinese and US patients. *Hepatobiliary surgery and nutrition*, 8(6): 604, 2019.

[9] Nakamura, H., Arai, Y., Totoki, Y., Shirota, T., Elzawahry, A., Kato, M., et.al. Genomic spectra of biliary tract cancer. *Nature genetics*, 47(9): 1003-1010, 2015.

[10] Li, M., Zhang, Z., Li, X., Ye, J., Wu, X., Tan, Z., et.al. Whole-exome and targeted gene sequencing of gallbladder carcinoma identifies recurrent mutations in the ErbB pathway. *Nature genetic*s, 46(8): 872-876, 2014.

[11] Zhang, Y., You, W. H., Li, X., Wang, P., Sha, B., Liang, Y., et.al. Single-cell RNAseq reveals transcriptional landscape and intratumor heterogenicity in gallbladder cancer liver metastasis microenvironment. *Annals of Translational Medicine*, 9(10), 2021.

[12] Otto, T., & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nature Reviews Cancer*, 17(2): 93-115, 2017.

[13] Schwartz, G. K., & Shah, M. A. Targeting the cell cycle: a new approach to cancer therapy. *Journal of clinical oncology*, 23(36): 9408-9421, 2005.

[14] Malumbres, M., & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nature reviews cancer*, 9(3): 153-166, 2009.

[15] Williams, G. H., & Stoeber, K. The cell cycle and cancer. *The Journal of pathology*, 226(2): 352-364, 2012.

[16] Carnero, A. Targeting the cell cycle for cancer therapy. *British journal of cancer,* 87(2): 129-133, 2002.

[17] Williams, G. H., & Stoeber, K. Cell cycle markers in clinical oncology. *Current opinion in cell biology*, 19(6): 672-679, 2007.

[18] Bartek, J., Lukas, J., & Bartkova, J. Perspective: defects in cell cycle control and cancer. *The Journal of pathology*, 187(1): 95-99, 1999.

[19] Chen, X., Chen, R., Jin, R., & Huang, Z. The role of CXCL chemokine family in the development and progression of gastric cancer. *International journal of clinical and experimental pathology*, 13(3): 484, 2020.

[20] Łukaszewicz-Zając, M., Pączek, S., Mroczko, P., & Kulczyńska-Przybik, A. The significance of CXCL1 and CXCL8 as well as their specific receptors in colorectal cancer. *Cancer Management and Research*, 8435-8443, 2020.

[21] McCoy, M. G., Nascimento, D. W., Veleeparambil, M., Murtazina, R., Gao, D., Tkachenko, S., et.al. Endothelial TLR2 promotes proangiogenic immune cell recruitment and tumor angiogenesis. *Science signaling*, 14(666), eabc5371, 2021.

[22] Makrilia, N., Kollias, A., Manolopoulos, L., & Syrigos, K. Cell adhesion molecules: role and clinical significance in cancer. *Cancer investigation*, 27(10): 1023-1037, 2009.

[23] Cavallaro, U. G. O., & Christofori, G. Multitasking in tumor progression: signaling functions of cell adhesion molecules. *Annals of the New York Academy of Sciences*, 1014(1): 58-66, 2004.

[24] Syrigos, K. N., & Karayiannakis, A. J. Adhesion molecules as targets for the treatment of neoplastic diseases. *Current pharmaceutical design*, 12(22): 2849-2861, 2006.

[25] Syrigos, K. N., Katirtzoglou, N., Kotteas, E., & Harrington, K. Adhesion molecules in lung cancer: implications in the pathogenesis and management. *Current pharmaceutical design*, 14(22): 2173-2183, 2008.

[26] Syrigos, K. N., Harrington, K. J., & Pignatelli, M. Role of adhesion molecules in bladder cancer: an important part of the jigsaw. *Urology*, 53(2): 428-434, 1999.

[27] Yilmaz, M., & Christofori, G. EMT, the cytoskeleton, and cancer cell invasion. *Cancer and Metastasis Reviews*, 28: 15-33, 2009.

[28] Le Bras, G. F., Taubenslag, K. J., & Andl, C. D. The regulation of cell-cell adhesion during epithelial-mesenchymal transition, motility and tumor progression. *Cell adhesion & migration*, 6(4): 365-373, 2012.

[29] Nanda, V., & Miano, J. M. Leiomodin 1, a new serum response factor-dependent target gene expressed preferentially in differentiated smooth muscle cells. *Journal of Biological Chemistry*, 287(4): 2459-2467, 2012.

[30] Zhao, B., Baloch, Z., Ma, Y., Wan, Z., Huo, Y., Li, F., & Zhao, Y. Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. *Cancer control*, 26(1), 2019.

[31] Kawahara, R., Recuero, S., Nogueira, F. C., Domont, G. B., Leite, K. R., Srougi, M., et.al. Tissue proteome signatures associated with five grades of prostate cancer and benign prostatic hyperplasia. *Proteomics*, 19(21-22): 1900174, 2019.

[32] Tan, Y., Chen, Q., Pan, S., An, W., Xu, H., Xing, Y., & Zhang, J. LMOD1, an oncogene associated with Lauren classification, regulates the metastasis of gastric cancer cells through the FAK-AKT/mTOR pathway. *BMC cancer*, 22(1): 474, 2022.

[33] Hahn, S. A., Schutte, M., Shamsul Hoque, A. T. M., Moskaluk, C. A., Da Costa, L. T., Rozenblum, E., et.al. DPC4, a candidate tumor suppressor gene at human chromosome 18q21. 1. *Science*, 271(5247): 350-353, 1996.

[34] Massagué, J., Blain, S. W., & Lo, R. S. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell*, 103(2): 295-309, 2000.

[35] Shi, Y., Hata, A., Lo, R. S., Massagué, J., & Pavletich, N. P. A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature*, 388(6637): 87-93, 1997.

[36] Massagué, J., Blain, S. W., & Lo, R. S. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell*, 103(2): 295-309, 2000.

[37] Yang, G., & Yang, X. Smad4-mediated TGF-β signaling in tumorigenesis. *International journal of biological sciences,* 6(1): 1, 2010.

[38] Xia, X., Wu, W., Huang, C., Cen, G., Jiang, T., Cao, J., et.al. SMAD4 and its role in pancreatic cancer. *Tumor Biology*, 36:111-119, 2015.

[39] Yazumi, S., Ko, K., Watanabe, N., Shinohara, H., Yoshikawa, K., Chiba, T., & Takahashi, R. Disrupted transforming growth factor-β signaling and deregulated growth in human biliary tract cancer cells. *International journal of cancer*, 86(6): 782-789, 2000.

[40] Ullah, I., Sun, W., Tang, L., & Feng, J. Roles of Smads family and alternative splicing variants of Smad4 in different cancers. *Journal of Cancer*, 9(21): 4018, 2018.

[41] Royce, S. G., Alsop, K., Haydon, A., Mead, L., Smith, L. D., Tesoriero, A. A., et.al. The role of SMAD4 in early-onset colorectal cancer. *Colorectal Disease*, 12(3): 213-219, 2010.

[42] Zanger, U. M., & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics*, 138(1): 103-141, 2013.

[43] McFadyen, M. C., Melvin, W. T., & Murray, G. I. Cytochrome P 450 enzymes: Novel options for cancer therapeutics. *Molecular cancer therapeutics*, 3(3): 363-371, 2004.

[44] Rai, R., Sharma, K. L., Misra, S., Kumar, A., & Mittal, B. CYP17 polymorphism (rs743572) is associated with increased risk of gallbladder cancer in tobacco users. *Tumor Biology*, 35: 6531-6537, 2014.

[45] Park, S. K., Andreotti, G., Sakoda, L. C., Gao, Y. T., Rashid, A., Chen, J., et.al. Variants in hormone-related genes and the risk of biliary tract cancers and stones: a population-based study in China. *Carcinogenesis*, 30(4): 606-614, 2009.