Chapter III

Materials and Methodology

3.1 Materials

3.1.1 Reagents and kits

Molecular reagents or kits	Manufacturer	Purpose
cDNA synthesis kit	Takara	Synthesis of cDNA for qPCR
Kapa mRNA Hyperprep Kit	Roche	Preparation of RNAseq libraries
Molecular grade ethanol	Merck	Purification and precipitation.
Qubit quantification kit	Invitrogen	Quantification of RNA
RNA isolation kit	Qiagen	Extraction of RNA from GBC and GSD tissue samples
RNA later	Invitrogen	Preservation of RNA in biological samples
SYBR green master mix	Applied Biosystems	Quantification of the amount of DNA in the qPCR

3.1.2 Computational Tools and Databases

Tools	Purpose
Basic Local Alignment Search Tool (BLAST)	Identification of novel lncRNA transcripts
ClusterProfiler & EnrichR	Functional enrichment analysis of differentially expressed genes.
Coding Potential Calculator 2 (CPC2)	Analysis of the coding potential of a transcript
Cytohubba	Topological analysis of PPI cluster modules
Cytoscape	Analysis and visualization of interactive networks
DESeq2	Identification of differentially expressed genes
Fast Quality Control (FASTQC)	Quality Control analysis of NGS data
Fast Preprocessing (FASTP)	Pre-processing of FASTQ reads
featureCounts	Quantification of mapped reads
GffCompare	Filtration of the transcripts into different class codes
Hierarchical indexing for spliced alignment of transcripts (HISAT2)	Mapping of processed reads with the reference genome
Molecular Detection Complex (MCODE)	Generation of PPI cluster modules

MEME Suite	Scanning of *Cis*-regulatory Modules (CRRs) for putative TFBS sites
miRanda	Prediction of lncRNA-miRNA and miRNA-mRNA interactions
Regulatory Sequence Analysis Tool (RSAT)	1-kb Upstream sequence retrieval of DEGs
Stringtie	Assembly of transcripts
Weighted Gene Co-expression Network Analysis (WGCNA)	Gene co-expression network analysis
Databases	**Purpose**
Catalogue of Inferred Sequence Binding Preferences (CIS-BP) database	Retrieval of TF motif PWMs
cBioPortal database	Visualization and analysis of large-scale cancer genomic datasets
Ensembl database	Retrieval of human reference genome and annotation files
European Nucleotide Archive (ENA) database	Retrieval of public RNAseq datasets
Kyoto Encyclopedia of Genes and Genomes (KEGG database)	Pathway enrichment analysis
Molecular Signature Hallmark Database (MsigDB)	Pathway enrichment analysis
STRING	Construction of protein-protein interaction (PPI) networks

3.1.3 In-house Computational Pipelines

Pipeline	Link to Repository
End-To-End Novel LncRNA analysis pipeline (ETENLNC)	https://github.com/EvolOMICS-TU/ETENLNC
Transcriptional Regulatory Network (TRN) construction and analysis pipeline (TF-TG)	https://github.com/EvolOMICS-TU/TF-TG
End-to-end and Beyond RNAseq analysis pipeline (ETENBR)	https://github.com/EvolOMICS-TU/-RNAseq-pipeline-Public

3.2 Methodology

The overall methodology of the study includes the integration of clinical, experimental, and computational approaches to generate transcriptomic datasets from GBC and GSD clinical tissue samples to identify crucial molecular signatures associated with GBC development. The workflow of the methodology is presented in **Figure 3.1**.

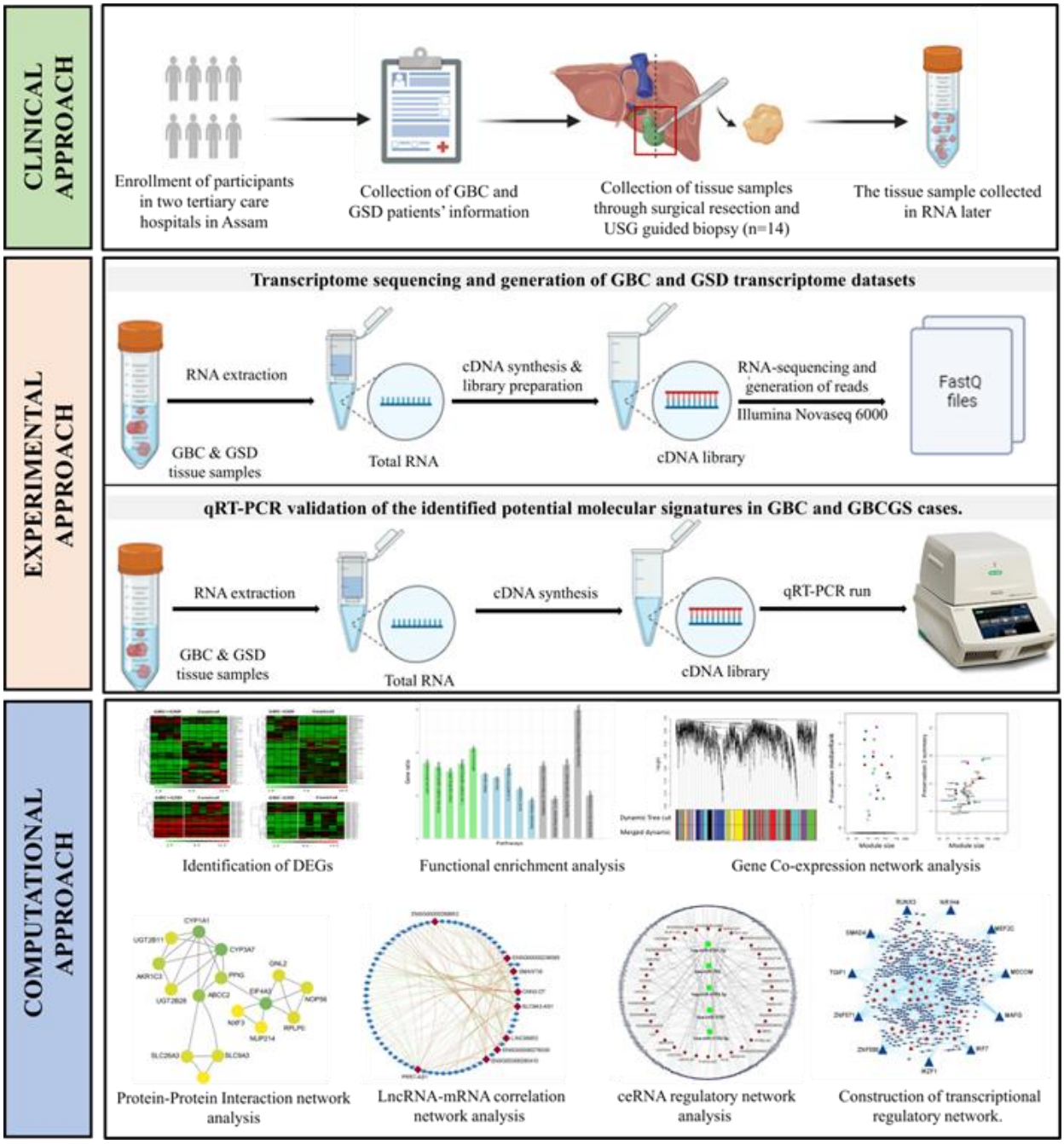

Figure 3.1: Illustrative representation of the overall methodologies employed to identify molecular signatures associated with GBC pathogenesis.

A. CLINICAL APPROACH

3.2.1 Study design and participants

The current experiment was planned as a hospital-based case-control study in Assam. A total of 14 tissue samples were collected, consisting of GBC tissues (n=8) and GSD (n=6) tissues. The GBC tissues were collected through surgical resection and ultrasonography (USG) guided biopsy from patients admitted at Dr. B. Borooah Cancer Institute (BBCI), Guwahati. GSD tissues were collected from individuals, who have undergone laparoscopic cholecystectomy for gallstone condition at Swagat Super Speciality Surgical Hospital (SSSSH), Guwahati, Assam [**Figure 3.2**]. Here, the GSD samples were considered as control. The study was ethically approved vide letter- *BBCI-TMC/Misc-01/MEC/254/2021* and *DoRD/TUEC/PROP/2022/06*. Clinically and histo-pathologically confirmed GBC and GSD cases were included in the study and all the participants voluntarily participated. The following were the exclusion and inclusion criteria considered for the study.

Inclusion Criteria

- Clinically and histo-pathologically confirmed cases of GBC and GSD
- GBC patients below the age limit of 86 years
- Both male and female patients

Exclusion Criteria

- Paediatric age group patients with GBC
- Patients who underwent radiation and chemotherapeutic treatment or any surgical interventions before cholecystectomy.
- Patients who were not willing to give consent.

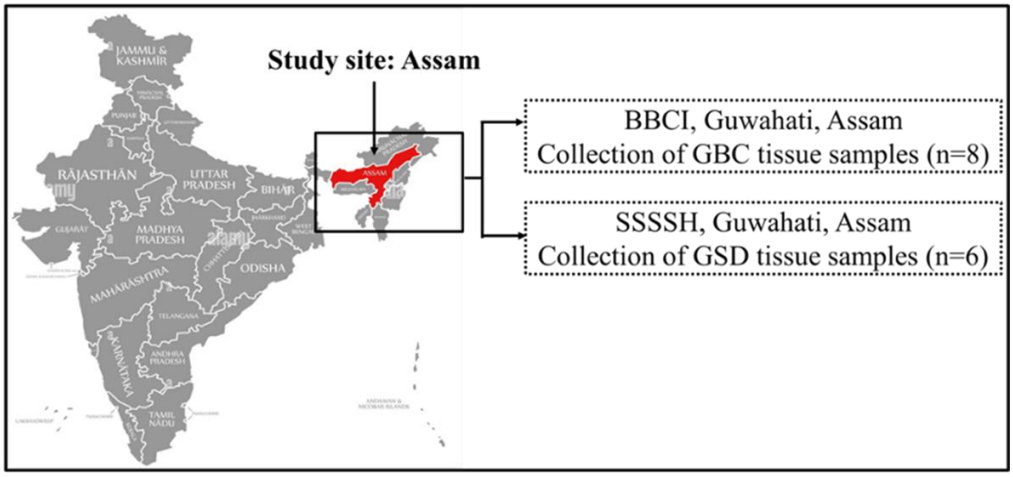

Figure 3.2: Collection of GBC and GSD clinical tissue samples from the population of Assam.

B. EXPERIMENTAL APPROACH

3.2.2 Transcriptome sequencing

Total RNA was isolated from approximately 25-30mg tissue samples collected from GBC patients and individuals who had GSD using RNA Easy Mini kit (Qiagen), according to the manufacturer's instructions. The quality of total RNA was checked using Agilent RNA 6000 Nano chip in 2100 Bioanalyzer (Agilent) and quantitation was performed by NanoDrop spectrophotometer (Thermo Scientific) followed by the fluorometric method in Qubit fluorometer (Invitrogen). Total RNA samples with high RNA integrity number (RIN) were selected for library preparation using Kapa mRNA Hyperprep Kit (Roche). The quality of RNAseq libraries was checked using high-sensitivity D1000 ScreenTape in Agilent 2200 TapeStation system (Agilent) and final library quantification was done using Real-Time PCR (QuantStudio 7 Flex). Paired-end (PE) 2 x 100 bp sequencing of these libraries was performed in Novaseq 6000 (Illumina).

3.2.3 Validation through quantitative Real Time (qRT)-PCR analysis

The expression of hub DEGs and DElncRNAs has been validated using SYBR green-based assays (Applied Biosystems, USA) on a Real-Time PCR System (Applied Biosystems, USA). For qRT-PCR analysis, the total RNA of the GBC+GSD, GBC-GSD, and control tissues was extracted using an RNA Easy Mini kit (Qiagen) and quantified using a Nanodrop spectrophotometer. The complementary DNA (cDNA) synthesis was done using the PrimeScript II 1st strand cDNA synthesis kit (Takara). The qRT-PCR was carried out in 20 µL of reaction volume consisting of 1X SYBR green PCR master mix, 0.4 µM of forward (F) and reverse (R) primer, 100ng of cDNA, and water to adjust the reaction volume. The cycling condition considered for the qRT-PCR is provided in **Table 3.1**. The RN18S1 was used as an internal control for the normalization of expression levels. All the primer sequences used in the qRT-PCR reaction are listed in **Table 3.2**. The Δ^{ct} and $2^{\Delta\Delta ct}$ methods were used to analyze the qRT-PCR data, where Δ^{ct} and $2^{\Delta\Delta ct}$ represent the sample's expression and relative expression of the target genes respectively

.

Table 3.1: Cycling conditions for Real-Time PCR.

Steps	Temperature	Time	Cycles
Uracil-DNA-Glycosylase (UDG) activation	50˚C	2 minutes	1
Activation of DNA polymerase	95˚C	2 minutes	
Denaturation	95˚C	15 seconds	40
Annealing	variable	30 seconds	
Extension	72˚C	1 minute	

Table 3.2: List of primer sequences of the hub DEGs and DElncRNA used in qRT-PCR.

Genes	Tm	Sequences (5´-3´)
LMOD1	Forward- 58.5, Reverse- 60.1	Forward: GAAGAACTCCCGTGACCAGCTA
		Reverse: AGCCTGGTCCTACTGAAGCAGT
KLF15	Forward- 57.2, Reverse-55.6	Forward: TATCACATGCTGCCCTCACC
		Reverse: GAAGTCCAAGATGCTGTCCTG
DIO3OS	Forward- 58.9, Reverse- 56.3	Forward: AGGCCCAGCCCAATAGGAA
		Reverse: GGCCCAAGAAACAGCAACA
SMAD4	Forward- 58.5, Reverse- 56.3	Forward: TGCCTCACCACCAAAACGG
		Reverse: CCAAACAAAAGCGATCTCCTCC
MECOM	Forward- 54.8, Reverse- 58.1	Forward: TATCCACGAAGAACGGCAATATC
		Reverse: CATGGAAACTTTTGGTGATCTGC
LINC00852	Forward- 57.0, Reverse- 55.3	Forward: CGTTGCCTACAGTCAAGTCAGT
		Reverse: GCCATGGTTCCCTTACTGATAC
MSTRG.16633.1	Forward- 56.1, Reverse- 56.3	Forward: TGTTTTGAAAGGAGCTGGGC
		Reward: CCTCATCGTCAGCTACACCT
MSTRG.53675.1	Forward- 55.6, Reverse-56.2	Forward: CTTTTCATCCAGCAGCACCT
		Reverse: CCAAATCTGCCTTCACCTGG
RN18S1	Forward- 55.3, Reverse- 56.0	Forward: GGAGTATGGTTGCAAAGCTGA
		Reverse: ATCTGTCAATCCTGTCCGTGT

C. COMPUTATIONAL APPROACHES

3.2.4 Retrieval of publicly available transcriptomic datasets

To obtain the relevant publicly available transcriptome datasets on GBC, GSD, Hepatocellular carcinoma (HCC), and intrahepatic cholangiocarcinoma (ICC), a comprehensive search was conducted on the ENA database [1] using the following criteria:

1. Study type: RNA Expression profiling by high throughput sequencing,
2. Attribute name: Tissue, and
3. Organism: *Homo sapiens*

The datasets containing (i) paired-end data, (ii) both case-control samples, (iii) sample size ≥ 20, and (iv) information about the sequencing platform used as well as the experimental protocol were considered for transcriptomic data analysis.

3.2.5 Transcriptome data analysis and generation of raw expression counts of transcribed genes.

Transcriptome/RNAseq data analysis involves four major steps:

1. Quality check of raw FastQ reads,
2. Pre-processing of reads,
3. Mapping of reads to the reference genome, and finally
4. Quantification of aligned reads.

The quality check (QC) of raw reads was performed using the FASTQC tool and the pre-processing of the reads was carried out using FASTP [2]. The pre-processing of reads using FASTP is based on the following criteria: (1) removal of adapters, (2) removal of bad quality reads with phred score threshold of 32, (3) removal of reads shorter than 10bp, and (4) filtration of reads with overall and per nucleotide phred thresholds. To eliminate reads that are less than 10 nucleotides, the default value for "read length" was set to 10, while the "phred quality" score threshold was set to "32" for bases. After that, another round of QC using FASTQC was carried out to identify and report any anomalous or nonconforming regions in the processed reads. The processed reads were aligned with the ensemble [3] reference human genome *Homo sapiens* (GRCh38) using Hisat2 [4]. The aligned or mapped reads were then quantified using the

featureCounts tool [5] to obtain the gene expression profile of each sample as a single count matrix file.

3.2.5.1 Identification of annotated and novel lncRNAs using in-house developed end-to-end novel lncRNA (ETENLNC) identification pipeline.

For the identification of novel lncRNAs, the generated mapped reads after pre-processing (as referred to in 3.2.4) were assembled using the Stringtie tool [6]. Assembly of mapped reads is performed in two steps: (i) The first step takes in BAM files to assemble full-length transcripts from several splice variants and generates a GTF file containing assembled transcripts, and (ii) The second step takes the transcripts that have been assembled from each of the GTF files and combines them into a single, non-redundant GTF file. The novel transcripts, by default, are labeled with a 'MSTRG' ID. To identify annotated and novel lncRNA from RNAseq datasets, genomic location-based filtration, and specific lncRNA sequence filtration steps were applied to obtain novel lncRNAs [**Table 3.3**].

Table 3.3: Filtration steps to identify novel lncRNAs from RNAseq datasets

(i) Genomic filtration	
Genomic-based filtration has been carried out through the isolation of putative classes of lncRNA transcripts using the GFFCompare tool [7]. To categorize the assembled transcripts according to their genomic locations, they are compared to a reference GTF. Class codes are assigned to the transcripts based on the coordinates of their genomic locations. Four classes of putative lncRNAs were selected, which include:	
Class code 'i'	It comprises fully intronic transcripts that arise from the intron sequences of the genome.
Class code 'o'	Class code 'o' includes overlapping transcripts arising from exonic-intronic overlaps.
Class code 'x'	It contains Natural Antisense Transcripts (NATs) arising from the exonic overlaps on the opposite strand
Class code 'u'	It comprises unknown transcripts, not found in the reference annotation provided to GFFCompare, and may contain novel unannotated transcripts and intergenic transcripts.
(ii) Advanced filtration	

The 'advanced filtration' of transcripts based on lncRNA properties was performed using customized scripts that carry out the following sub-filters.	
Length filter	RNA transcripts with sizes greater than 200 nucleotides are considered as lncRNAs. Using this filter, transcripts with sequence lengths greater than 200 nucleotides were selected.
Exon filter	The length-filtered transcripts were then filtered based on their exon number. Transcripts having an exon number greater than 2 were selected for further filtering.
ORF filter	LncRNA transcripts do not code for proteins and therefore have a smaller ORF, reported up to 300bps. ORF filter selects transcripts having ORFs below 300bps.

3.2.5.2 Coding Potential Analysis (CPA)

Filtered transcripts are further analyzed to determine their coding potential using CPC2 [8]. CPA calculates the coding potential of the identified filtered transcripts using a classifier based on Support Vector Machines (SVM). CPC2 classifies the transcripts as either coding or noncoding based on sequence properties like the isoelectric point (pI), FICKETT score, and ORF features. Transcripts labeled as "noncoding" are chosen for additional filtering.

3.2.5.3 BLASTn analysis

To identify novel lncRNA transcripts, the annotated or existing lncRNA transcripts had to be removed from the obtained set. This was done using a nucleotide BLAST (Basic Local Alignment Search Tool) search against known/annotated lncRNAs using the NCBI BLAST toolkit [9,10]. The parameters "-e-value" and "-word_size" are set to 0.001 and 7, respectively, to minimize errors and limit the match threshold of hits to seven nucleotides while running BLASTn. Transcripts with a BLAST score of greater than 95% were eliminated from the filtered transcripts. The filtered lncRNAs obtained after the BLASTn run were subjected to DEA for the identification of known and novel DElncRNAs. The overall methodology for the identification of novel lncRNAs from the transcriptomic dataset is presented in **Figure 3.3**.

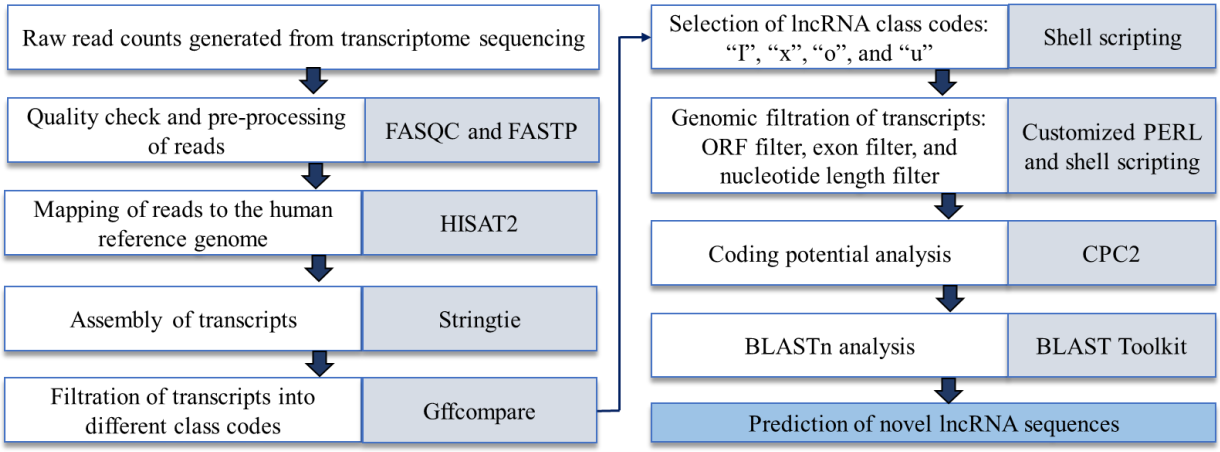

Figure 3.3: Schematic workflow for identification of predicted novel lncRNAs from the transcriptomic dataset.

3.2.6 Identification of significant differentially expressed genes and lncRNA from transcriptomic datasets.

The differential analysis of raw count data obtained through RNAseq, for evidence of expression changes across experimental conditions, is an important step in comparative high-throughput sequencing studies. The pre-processing steps such as low read count removal, normalization, and transformation of the raw count data are essential for downstream analysis because the technical variability or differences in the RNAseq library size and library composition can falsely reflect on differential gene expression [11]. Here, the DESeq2 [12] method was used to generate differentially expressed genes (DEGs) in GBC samples as compared to the control samples, and the consensus of both DESeq2 and EdgeR is used for case study 1. The DESeq2 method calculates the ratio of each read count to the geometric (logarithmic) mean of all read counts for a given gene across all samples. The Trimmed Mean of M-values chooses a reference sample and calculates fold changes and absolute expression levels relative to that sample. A count matrix K, comprising one row for each gene i and one column for each sample j, serves as the starting point for a DESeq2 analysis. The number of sequencing reads that have been conclusively mapped to a gene in a sample is indicated by the matrix entries K_{ij}. For each gene, a generalized linear model (GLM) [13] with a logarithmic link is used. The log2 fold change represents the estimated effect size, indicating the apparent alteration in gene or transcript expression between comparison and control groups. This value, expressed on a logarithmic scale with a base of 2, indicates increased expression (upregulation) with positive values and decreased expression (downregulation) with negative values. An absolute cut-off value >10 across all the samples for each gene was used for obtaining DEGs.

Applying *p-adjusted value (padj)* ≤ 0.05, the list of significant DEGs was obtained from DESeq2 and was considered for downstream analysis.

3.2.7 Functional enrichment and pathway analysis

To identify underlying biological processes and pathways associated with DEGs obtained through transcriptomic data analysis, functional enrichment analysis was performed. It is a widely used method of identifying biological functions that are over-represented or under-represented among a list of genes with respect to reference background. Functional enrichment analysis uses statistical methods such as Fisher's Exact Test to identify significantly enriched biological processes and pathways. Functional enrichment and pathway analysis were carried out using clusterProfiler [14] and enrichR [15] (R packages). Kyoto Encyclopedia of Genes and Genomes (KEGG) [16] and Molecular Signature Database (MsigDB) [17] were used to identify significant biological pathways associated with identified DEGs. A count number > 2 and *p*-value threshold < 0.05 were considered as selection criteria for enriched biological processes and pathways associated with GBC.

3.2.8 Gene coexpression network (GCN) analysis

Gene-gene correlation networks are being used increasingly in bioinformatic studies. Correlation networks facilitate network-based gene screening methods that can be used to identify putative biomarkers or therapeutic targets. For instance, WGCNA is a systems biology approach that describes the patterns of gene correlation obtained in RNAseq/microarray samples. WGCNA has been successfully applied in several biological studies, including the study of brain imaging data, cancer, yeast genetics, and mice genetics. For this study, the expression values of the significant DEGs identified through DEA were used as input to build the gene co-expression network using the WGCNA (R package) [18].

The main steps of WGCNA include – (i) Calculation of pairwise gene similarity using Pearson's correlations to create an adjacency matrix (Adj_{ij}), (ii) construction of a scale-free co-expression network from the adjacency matrix by taking the β value (soft thresholding parameter) into consideration, (iii) identifying the modules by deriving Topological Overlap Matrix (TOM) from adjacency matrix, (iv) identification of non-preserved modules using *Z-summary* and *medianRank* statistics and (v) intramodular connectivity analysis of the non-preserved modules to identify hub genes.

3.2.8.1 Pearson's correlations analysis for identification of co-expressed gene modules

Pearson's correlations analysis for each gene pair was calculated using the adjacency function of the WGCNA package to construct the adjacency matrix. Then, the adjacency matrix was used to create a scale-free co-expression network based on a soft-thresholding parameter βeta (β) to enrich strong correlations between gene pairs [19]. The function *TOMsimilarity* was utilized to transform the adjacency matrix that was computed into a TOM. Then, using the *flashClust* function for module identification, hierarchical clustering was carried out using this topological overlap matrix as an input. Finally, using the R package dynamicTreeCut, the network modules for the cancer and control samples were found, with a minimum module size (minClusterSize) = 30 and a minimum sensitivity (*deepSplit*) = 2 for the gene dendrogram.

3.2.8.2 Module Preservation Analysis

Module preservation was carried out to identify the non-preserved module between the cancer and the control network. The statistics behind module preservation is to calculate gene preservation within a module by comparing a reference network (control) with a test network (cancer) [18]. It was assumed that the genes found in non-preserved modules of the cancer network might be involved in the pathological process as compared to the control network. The module preservation analysis was performed using the WGCNA function module to determine the weight and connectivity of genes within the module of the cancer and control network. Based on degree and connectivity, the preservation analysis statistics- *Z-summary* and *medianRank* gave the overall significance of the preservation of a module. The *Z-summary preservation* < 2 indicates no preservation, $2 \leq$ *Z-summary* ≤ 10 suggests weak to moderate preservation, and *Z-summary preservation* > 10 implies strong preservation [20].

3.2.8.3 Intramodular Connectivity Analysis

In network biology, the connectivity between nodes (genes) is generally considered a degree. In this study, the intramodular connectivity approach was used for the screening of hub genes within non-preserved modules. The intramodular connectivity measures the degree of each gene within a module. The criteria used for this study were to calculate the connectivity from the whole network (*kTotal*) and the connectivity within modules (*kWithin*). This measure of connectivity is useful to determine the biologically significant modules by calculating the degree of nodes within modules. The intramodular connectivity approach helps in screening regulatory changes in gene expressions [19,20].

3.2.9 PPI network analysis

The STRING database [21] was used to construct the PPI networks with the significant DEGs identified in each case study. DEGs in the PPI networks were represented as nodes and the interactions between the DEGs were represented by edges. The PPI interactions from STRING comprise both direct (physical) and indirect (functional) interactions. To quantify the interaction confidence, a score is assigned to each edge in the network. The PPI networks were analyzed using Cytoscape version 3.9 [22]

For the identification of significant module clusters from the whole PPI networks, the MCODE tool [23] was used. The local density of each node in the network was taken into consideration to compute the module cluster score. For every cluster network, MCODE parameters included Node Score Cutoff = 0.2, K-Core = 2, and Threshold = 2. The MCODE cluster scores ≥ 4 and the number of nodes > 4 were set as cutoff criteria for obtaining significant PPI network modules.

Furthermore, Cytohubba [24], a Cytoscape plugin was used for the identification of hub DEGs from the PPI module obtained through MCODE. Each node ranking method is associated with function F which assigns a numerical value to each node v. The ranking of a node u is greater than that of another node v if the score of u (i.e. $F(u)$) is greater than that of v (*i.e.* $F(v)$). Cytohubba implements 11 node ranking methods which are divided into two classes: global and local node ranking methods. A local rank method only takes into account the relationship between a node and its immediate neighbours when calculating its score within a network; in contrast, the global method looks at the relationship between the node and the entire network. In this study, three local node ranking methods- degree, maximal clique centrality (MCC), and maximum neighbourhood component (MNC) and two global node ranking methods- betweenness and closeness were used. The predicted hub DEGs identified from each of the node ranking methods were further intersected for the identification of consensus-significant hub DEGs from the PPI modules [24].

3.2.10 ceRNA regulatory network analysis

The mRNA–miRNA–lncRNA regulatory relationship was predicted based on the ceRNA theory, i.e., mRNAs, and lncRNAs compete with each other for binding with MREs to regulate gene expression [25]. To identify potential ceRNA interactions between known/novel DElncRNAs, DEmRNAs and, all human miRNAs were used for the prediction. Putative

DElncRNA-miRNA and DEG-miRNA interactions were predicted using miRanda [26,27]. The miRanda algorithm is comparable to the Smith-Waterman algorithm [28]. However; the miRanda algorithm scores are based on complementarity of nucleotides (A=U or GC) rather than building alignments based on matching nucleotides (A-A or U-U, for example). Importantly, the scoring matrix used for this analysis allows G=U 'wobble' pairs, which are important for the accurate detection of RNA: RNA duplexes [29]. Complementarity parameters at individual alignment positions are +5 for G≡C, +5 for A=U, +2 for G=U, and -3 for all other nucleotide pairs. The ceRNA regulatory network was constructed based on the hub DElncRNAs and DEnLncRNAs identified through coexpression analysis in two GBC groups. The ceRNA network was visualized using Cytoscape version 3.9 software.

3.2.11 Prediction of RNA secondary structures

The secondary structures of the novel lncRNA were predicted using RNAfold. This tool uses a thermodynamic energy-based principle to predict the RNA secondary structure by taking the novel RNA sequences as an input. It determines the stability of RNA structures by computing their minimum free energy (MFE) and selecting the predicted RNA structure with the lowest free energy [30].

3.2.12 Transcriptional Regulatory Network (TRN) analysis for identification of potential regulatory TFs in GBC

Transcription factors (TFs) are known to be crucial regulators in the transcription process which regulates the overall gene expression by binding to the start site of the promoter region [31]. Transcriptional regulatory networks are real-world biological networks. This is because real-world networks exhibit scale-free properties and functional relevance. Scale-free networks are characterized by a few highly connected nodes (hubs) and many nodes with relatively few connections [32]. To construct the TF-TG regulatory networks, the 1-kb upstream FASTA sequence of the significant DEGs and DElncRNAs identified in case study 2 and case study 4 were extracted from RSAT [33]. The experimentally determined position weight matrix (PWM) of the identified DE-TFs was obtained from the cisBP database [34]. A PWM is a mathematical model that gives the binding specificity of a TF and is used to scan the upstream sequences of DEGs for determining the TF-TG interactions [**Figure 3.4**] [35]. This approach focuses on analyzing the cis-regulatory targets and will identify only the DEGs and DElncRNAs regulated directly by TFs through the transcription factor binding site (TFBS). The PWM scanning was

carried out using the FIMO tool of the MEME suite [36]. A p-value threshold of 10^{-4} was considered to obtain the significant TF-TG interactions.

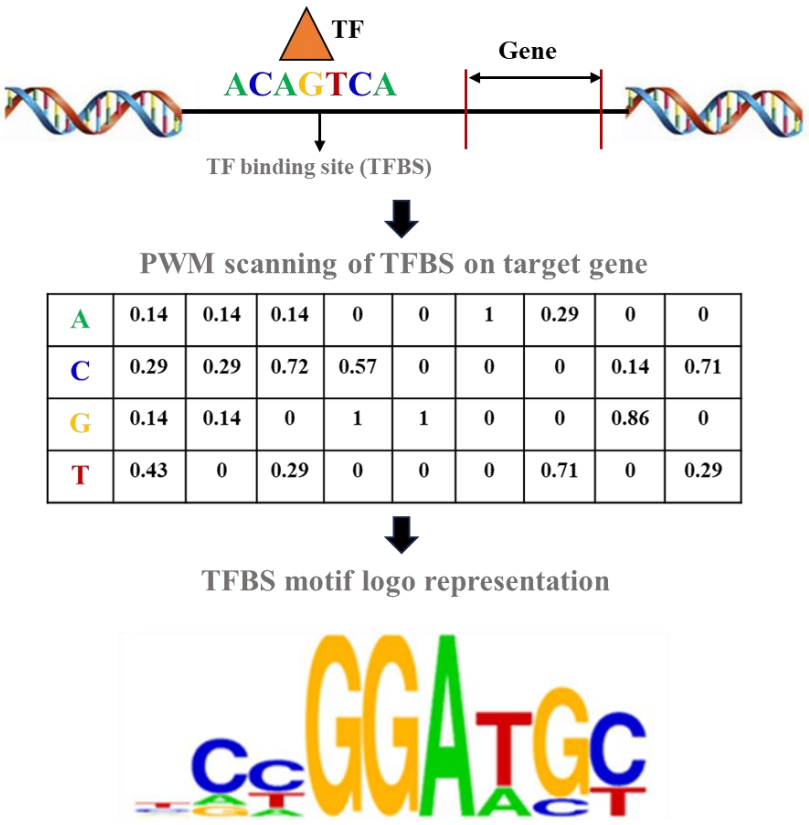

Figure 3.4: Schematic representation of the TF binding to target gene through PWM scanning. It is a scoring matrix for representing TF binding motifs. It represents a matrix of N rows and four columns, in which the matrix score of each base at each position is described.

Furthermore; the expression correlation analysis of DEGs and differentially expressed transcription factors (DETF) and the DElncRNAs and DETFs was performed to identify the significant DETFs-DEGs and DETFs-DElncRNAs correlated pairs. The Pearson's correlation coefficient (PCC) was estimated using a customized R-script and pairs exhibiting PCC of 0.9 (positive correlation) and -0.9 (negative correlation) with *p*-values less than 0.01 were selected. The final TF-TG and TF-lncRNA interaction network was constructed by taking the consensus of the common pairs identified through correlation analysis and PWM-scanning analysis. Finally, TRNs were constructed and visualized in the form of an interactive network using Cytoscape. Based on the degree centrality, the top ten significant TFs were identified as hubs for further analysis.

3.2.13 Calculation of EMT scores

Epithelial-mesenchymal transitions (EMTs) are complex cellular processes that play crucial roles in cancer metastasis and are largely associated with poor survival of cancer patients. The large-scale transcriptomic data associated with EMT has enabled the development of different EMT scoring metrics that calculate the extent of EMT in cancer [39]. For our study, three different scoring metrics – 76Gs, MLR, and KS were used to quantify EMT scores for each sample separately.

3.2.14 Cross-validation of the expression of DEGs and lncRNA identified through transcriptomic data analysis.

The TNM plotter tool was used for validation of the expression of hub genes in TCGA datasets identified from public and in-house generated transcriptomic datasets. TNM plotter is a web server that gives a customizable interactive analysis of the gene expression based on TCGA datasets [37]. For the validation of hub genes, the fold change value > 2 was considered. Furthermore; the genetic alterations such as mutations and copy number alterations linked with the potential hub genes were identified using the cBioPortal database [38] (https://www.cbioportal.org/). The results generated from cBioPortal were visualized as OncoPrint.

Bibliography

[1] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., et.al. The European nucleotide archive. *Nucleic acids research*, 39(suppl_1): D28-D31, 2010.

[2] Chen, S., Zhou, Y., Chen, Y., & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17): i884-i890, 2018.

[3] Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., et.al. Ensembl 2022. *Nucleic acids research*, 50(D1): D988-D995, 2022.

[4] Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37(8): 907-915, 2019.

[5] Liao, Y., Smyth, G. K., & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7): 923-930, 2014.

[6] Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNAseq reads. *Nature biotechnology*, 33(3): 290-295, 2015.

[7] Pertea, G., & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research*, 9, 2020.

[8] Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., & Gao, G. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, 45(W1): W12-W16, 2017.

[9] Chen, Y., Ye, W., Zhang, Y., & Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic acids research*, 43(16): 7762-7768, 2015.

[10] Ye, J., McGinnis, S., & Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2): W6-W9, 2006.

[11] Evans, C., Hardin, J., & Stoebel, D. M. Selecting between-sample RNAseq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5): 776-792, 2018.

[12] Love, M. I., Huber, W., & Anders, S. Moderated estimation of fold change and dispersion for RNAseq data with DESeq2. *Genome biology*, 15(12): 1-21, 2014.

[13] Cox, D. R., Hinkley, D. V., Rubin, D., & Silverman, B. W. (Eds.). *Monographs on statistics and applied probability.* London: Chapman & Hall, 1984.

[14] Yu, G., Wang, L. G., Han, Y., & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5): 284-287, 2012.

[15] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et.al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1): W90-W97, 2016.

[16] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1): D353-D361, 2017.

[17] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12): 1739-1740, 2011.

[18] Langfelder, P., & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1): 1-13, 2008.

[19] Zhang, B., & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1): 2005.

[20] Langfelder, P. Signed vs. Unsigned Topological Overlap Matrix Technical Report, 2013.

[21] Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1): 258-261, 2003.

[22] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et.al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11): 2498-2504, 2003.

[23] Bader, G. D., & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1): 1-27, 2003.

[24] Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC systems biology*, 8(4): 1-7, 2014.

[25] Tay, Y., Rinn, J., & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483): 344-352, 2014.

[26] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. Human microRNA targets. *PLoS biology*, 2(11): e363, 2004.

[27] Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. MicroRNA targets in Drosophila. *Genome biology*, 4: 1-27, 2003.

[28] Smite, T. F., & Waterman, M. S. Identification of common molecular subsequences. *Repr. from J. Mol. Biol. J. Mol. Bwl*, 147(147): 195-197, 1981.

[29] Wuchty, S., Fontana, W., Hofacker, I. L., & Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers: Original Research on Biomolecules*, 49(2): 145-165, 1999.

[30] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. ViennaRNA Package 2.0. Algorithms for molecular biology, 6: 1-14, 2011.

[31] Maston, G. A., Evans, S. K., & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7: 29-59, 2006.

[32] Barabási, A. L., & Albert, R. Emergence of scaling in random networks. *science*, *286*(5439): 509-512, 1999.

[33] Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R. S., Defrance, M., Vervisch, E., et.al. RSAT: regulatory sequence analysis tools. *Nucleic acids research*, 36(suppl_2): W119-W127, 2008.

[34] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et.al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6): 1431-1443, 2014.

[35] Aerts, S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Current topics in developmental biology*, 98: 121-145, 2012.

[36] Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. The MEME suite. *Nucleic acids research*, 43(W1): W39-W49, 2015.

[37] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et.al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113-1120, 2013.

[38] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et.al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269): pl1-pl1, 2013.

[39] Chakraborty, P., George, J. T., Woodward, W. A., Levine, H., & Jolly, M. K. Gene expression profiles of inflammatory breast cancer reveal high heterogeneity across the epithelial-hybrid-mesenchymal spectrum. *Translational oncology*, 14(4): 2021.