Original Article

Exploring Breast Cancer-Associated Genes: A Comprehensive Analysis and Competitive Endogenous RNA Network Construction

CharuMeena H[1], Sagaya Jansi R[1*], Aishwarya S[1], Balamurugan Shanmugaraj[2], Praveen Kumar Panthagani[3]

1. Department of Bioinformatics, Stella Maris College, Chennai, Tamil Nadu, India.
2. Department of Biotechnology, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.
3. Molecular Biology Department, Ampath Lab, Hyderabad, Telangana.

How to cite this article: **CharuMeena H, Sagaya Jansi R, Aishwarya S, Balamurugan Shanmugaraj, Praveen Kumar Panthagani. Exploring Breast Cancer-Associated Genes: A Comprehensive Analysis and Competitive Endogenous RNA Network Construction. Archives of Razi Institute. 2025;80(2):447-460.** DOI: 10.32592/ARI.2025.80.2.447

ABSTRACT

Breast cancer is a most common cancer that primarily affects women, in which cells become abnormal and multiply in an uncontrollable fashion. The etiology of these cancers is predominantly hereditary, with gene mutations and geographic indications being the predominant factors in most invasive breast cancer types. However, it is important to note that several other factors, including age, gender, ethnic background, and environmental influences, also contribute to the development of these diseases. Non-coding RNAs refer to a class of endogenous molecules that play a role in the development of various types of cancer. The objective of this research is to identify differentially expressed genes in cases of breast cancer. A series of analyses were conducted on the RNA-Seq data from the TCGA related to breast cancer. These analyses included both expression and survival studies. The objective of these analyses was to explore the gene expression of the samples and genes computationally through the use of R programming. The results obtained after each analysis were inferred both visually and logically. A total of 613 genes were identified as exhibiting differential expression among the samples, with 254 genes demonstrating increased expression and 359 genes exhibiting decreased expression. The differentially expressed genes obtained using the R package "TCGA Biolinks" were subsequently employed in the construction of the ceRNA network. A comprehensive analysis of the TCGA biolinks data set revealed the presence of aberrantly expressed long non-coding RNAs (lncRNAs), microRNAs (miRNAs), and messenger RNAs (mRNAs) in breast cancer samples. The analysis identified a total of 352, 183, and 254 cases, respectively, demonstrating significant disparities in gene expression patterns among different breast cancer samples. A study of 352 long non-coding ribonucleic acids (lncRNAs) revealed that two of these molecules, LINC00461 and MALAT1, exhibited particularly high levels of expression. These findings suggest that these two molecules may serve as more effective therapeutic biomarkers. Furthermore, the study identified a significant enrichment of microRNA target genes within the samples examined, suggesting a potential regulatory relationship between these molecules and their target genes. Consequently, this investigation has constructed competing endogenous RNA networks and has further elucidated the underlying biomarkers for breast cancer cohorts.

Keywords: Breast Cancer, miRNAs, lncRNAs, ceRNA network, R program.

Article Info:
Received: 21 January 2024
Accepted: 2 March 2024
Published: 30 April 2025

Corresponding Author's E-Mail:
sagayajansi@stellamariscollege.edu.in

1. Introduction

Cancer is a deadly disease that arises due to uncontrolled growth of body cells. The specific cause of cancer is impossible to predict in some cases; however, it is known that cancer cells can often be modulated by the external environment, intake of alcohol, tobacco, and so on. Furthermore, exposure to ultraviolet (UV) and other radiations has been identified as a contributing factor to cancer development (1). Proto-oncogenes are a group of genes that facilitate normal cellular growth. The proto-oncogenes, by definition, are subject to mutation, which can ultimately result in oncogene formation and, consequently, cancer. The activation of oncogenes is typically caused by acquired mutations, such as chromosome rearrangements and gene duplications. As Levine and Puzio-Kuter (2010) have demonstrated, mutations in tumor suppressor genes such as p53, p16, and p21 result in the initiation of uncontrolled cell proliferation (2). Breast cancer is among the most prevalent cancers affecting women in the United States, Asia, and Africa. Recent research suggests that one in eight women is at risk of developing breast cancer in her lifetime (3). The development of this neoplasm originates in breast cells, lobules (the glands that contain milk) and ducts (tubes that carry the milk from lobules to the nipples). The development of breast cancer is characterized by the unregulated growth of modified breast cells, lobules, or ducts (4). Breast cancer is categorized into two distinct classifications: invasive and non-invasive. Invasive breast cancer is characterized by its capacity to extend into the surrounding tissues, whereas noninvasive breast cancer remains confined to the ducts or lobules within the breast. The classification of breast cancer is typically divided into three distinct subtypes. The first of these is hormone receptor-positive breast cancer, which is characterized by the presence of oestrogenic and/or progestin receptors within the tumor. It is estimated that between 60% and 70% of breast cancers are of this type. In addition, it has been determined that HER2-positive cancers account for between 15% and 25% of all breast cancers. This specific type of cancer manifests the presence of HER2 receptors (iii). Triple-negative breast cancer is characterized by the absence of oestrogenic, progesterone and HER2 receptors. Approximately 15% of breast cancers are classified as triple negative. This particular type of breast cancer has been observed to be prevalent among young women (5). Non-coding RNAs (ncRNAs) are transcriptionally non-functional and constitute the majority of the transcriptome (98%). Long non-coding RNAs (lncRNAs) constitute a subset of non-coding RNAs (ncRNAs) which typically range in size from approximately 200 nucleotides. These elements play pivotal roles in gene regulatory networks, operating through cis or trans acting pathways at transcriptional, posttranscriptional, and epigenetic levels (6,7,8). Furthermore, microRNAs (miRNAs) have been identified as a group of small non-coding molecules of approximately 22 nucleotides in length. These proteins have the capacity to bind to the 3'UTR of target mRNA, thereby modulating gene expression. Anomalous expressions of microRNAs (miRNAs) have been observed to function in conjunction with long non-coding RNAs (lncRNAs). LncRNAs have been demonstrated to influence the expression of miRNAs by sequestering target miRNAs and participating in the regulation of mRNA expression (9) . The presence of microRNA response elements (MREs) has been demonstrated to result in the downregulation of target molecules due to the inhibition of protein synthesis. This phenomenon is prevented by the presence of competitive endogenous RNAs (ceRNAs), which compete for microRNAs (miRNAs) with shared microRNA response elements (MREs). The function of ceRNAs is to regulate other RNA transcripts by competing with them for shared microRNAs. The ability of ceRNAs to modulate a single mRNA or even multiple mRNAs is well documented (10,11), as is their capacity to influence the available miRNAs in the cell. The function of ceRNAs is to modulate the expression levels of microRNAs (miRNAs), with a consequent effect on the repression of the target genes of the latter, thus contributing to the development of cancer. A ceRNA network is typically defined as an interactive network comprising mRNAs, miRNAs and lncRNAs that are differentially expressed in a specific cell or cancer sample (12, 13). The present study has two objectives. Firstly, it seeks to utilize R programming to analyze the TCGA breast cancer data in order to predict therapeutic long non-coding RNA (lncRNA) targets for breast cancer. Secondly, it aims to construct a competitive endogenous RNA (ceRNA) network through the analysis of differentially expressed mRNAs and their interacting microRNAs and LncRNAs.

2. Materials and Methods

2.1. Data Recovery and Preprocess

As demonstrated in the study by Tomczak et al. (14) , the Genomic Data Commons (GDC) data portal contains over 1,097 cases of breast cancer from the TCGA database (https://portal.gdc.cancer.gov/). In the present study, ten breast cancer samples (five tumor samples and five healthy samples) from the TCGA-BRCA project were selected and retrieved from TCGA via the GDC data portal using their corresponding barcodes (Table 1). The selection of ten samples was made on the basis of the following criteria: a) Project =TCGA-BRCA,b) Data Category = Gene expression) Data Type = Gene expression quantification) Experimental Strategy = RNA-Seq,e) Platform = Illumina HiSeq. The TCGA biolinks package utilizes multiple functions to facilitate the execution of these analyses. In order to successfully download the TCGA samples from the GDC data portal, it was necessary to execute the R code with the GDC query and GDC download functions. This was achieved by specifying the sample barcodes.

Table 1. TCGA Barcodes of the samples selected for the study.

Tumor samples	*Healthy samples*
TCGA-A1-A0SD-01A-11R-A115-07	*TCGA-A7-A0CE-11A-21R-A089-07*
TCGA-A8-A06Y-01A-21R-A00Z-07	*TCGA-AC-A2FB-11A-13R-A17B-07*
TCGA-AC-A2FK-01A-12R-A180-07	*TCGA-BH-A0AU-11A-11R-A12P-07*
TCGA-A7-A13E-01A-11R-A12P-07	*TCGA-E2-A15I-11A-32R-A137-07*
TCGA-BH-A18P-01A-11R-A12D-07	*TCGA-GI-A2C9-11A-22R-A21T-07*

Furthermore, a heat map and a box plot were constructed for the same.

2.2. Data Processing and Differential Expression Analysis

The differential gene expression investigation encompasses functions such as normalization and data filtering. The analysis was conducted utilizing the TCGA Biolinks linked edgeR package. Within the framework of the edgeR package, a dispersion estimate is assigned to each gene. The raw p-values are then adjusted through the implementation of the False Discovery Rate (FDR) correction, a process which identifies the top differentially expressed genes. Subsequently, the output of the DEGs was filtered automatically using the absolute value of the LogFC greater than or equal to 1. The data was stored in the global environment of RStudio, which was subsequently utilised during the process of constructing the ceRNA network.

2.3. Functional Enrichment and Survival Analysis

Gene Ontology (GO) analysis was conducted to investigate the functional roles of the DE genes. The clinical data relevant to the specified samples were retrieved to facilitate the execution of survival analysis. TCGA Biolinks is a software program that can be used to create a survival plot. In addition, it can be used to conduct further Kaplan–Meier curve analysis. This analysis can be used to assess the correlation between the obtained RNA and survival time. The statistical significance of the results can also be evaluated.

2.4. Prediction of Differentially Methylated Regions and Analysis of Sample Mean

The differentially methylated CpG sites were identified and visualized using a volcano plot. Furthermore, a boxplot illustrating the mean DNA methylation levels per group was generated from the obtained data. A Starburst plot was utilised in the study of DNA methylation and gene expression in conjunction.

2.5. Principal Component and Oncoprint Analysis for DEgenes

The objective of this analysis was twofold: firstly, to reduce the number of dimensions of our gene set, and secondly, to visualize the results in a more effective manner. Furthermore, an Oncoprint barplot was constructed for the purpose of visualizing the various genomic alterations and mutations detected in the breast cancer samples.

2.6. Selecting DEmRNAs for ceRNA Network Analysis

The differentially expressed mRNAs were then used to construct the ceRNA network. In order to predict the interactive miRNAs and long non-coding RNAs (lncRNAs), the initial screening of these mRNAs was based on the criterion of |LogFC| and false discovery rate (FDR) value. The mRNAs with a value of |logFC|> 1.5 and FDR-value < 0.01 were retained. Following this filtration process, the number of mRNAs was reduced from 663 to 254.

2.7. Analysis of mRNAs-target miRNAs

The interacting microRNAs were then assessed using the miRSystem database (http://mirsystem.cgm.ntu.edu.tw/) with the DEmRNAs that had previously been filtered. A total of 183 microRNAs (miRNAs) were found to strongly interact with target mRNAs (15).

2.8. Prediction of miRNAs-target lncRNAs

The target long non-coding RNAs (lncRNAs) for the list of previously obtained microRNAs (miRNAs) were assessed using the miRcode database (http://www.mircode.org/). In the study by Jeggari et al. (16), the researchers exclusively retained those long non-coding RNAs (lncRNAs) which were shared among the predicted microRNAs.

2.9. Construction and Visualization of ceRNA Network

In light of the voluminous nature of the data, the top 20 mRNAs, along with their interacting microRNAs and long non-coding RNAs, were selected for the construction of the network. The data were then collated and the mRNA-miRNA-lncRNA network was generated and visualised using Cytoscape software (17). The complete workflow can be viewed in Figure 1.

3. Results

3. 1. Preprocessing and Preparation of Matrix of Gene Expression

The breast cancer data obtained for 10 samples from TCGA were first subjected to preprocessing prior to analysis. In this step, a matrix of gene expression was obtained, comprising more than 21,000 genes in the rows and TCGA sample barcodes in the columns. A heatmap illustrating the gene expression data in the samples and a boxplot following sample normalization were prepared (Figure 2). Each tile in the heat map represented the expression of a specific gene for a given sample, with the color scale indicating levels ranging from low to high. In this instance, the horizontal axis of the box plot is indicative of different samples, with the vertical axis representing expression value. Following the implementation of the normalization process, the black lines of the box plot were

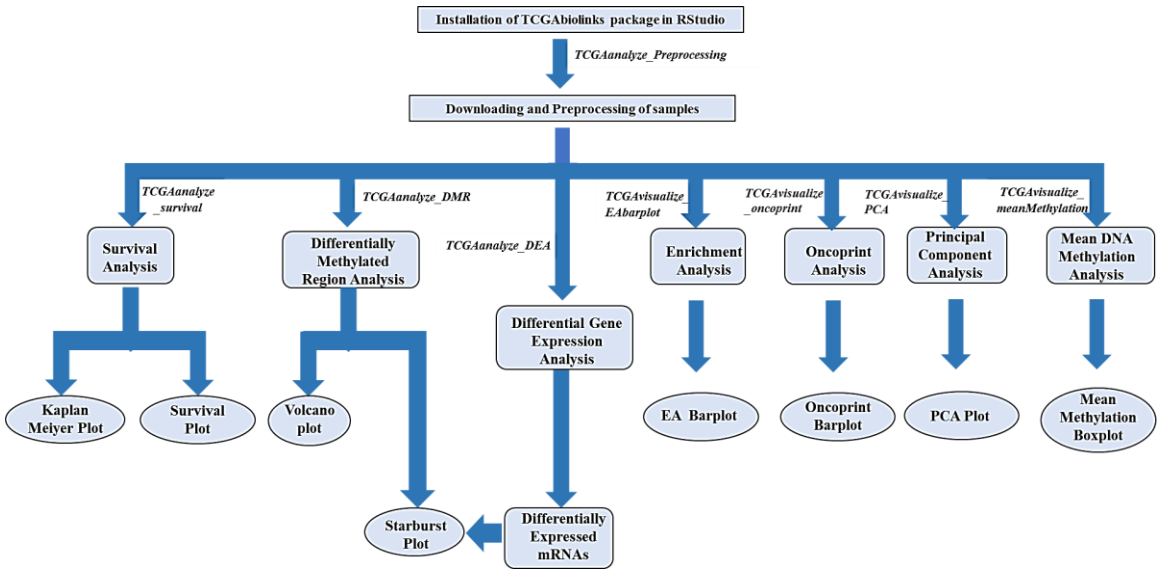

Figure 1. Workflow of Integrative biological analyses of breast cancer data using TCGAbiolinks package and ceRNA network construction.

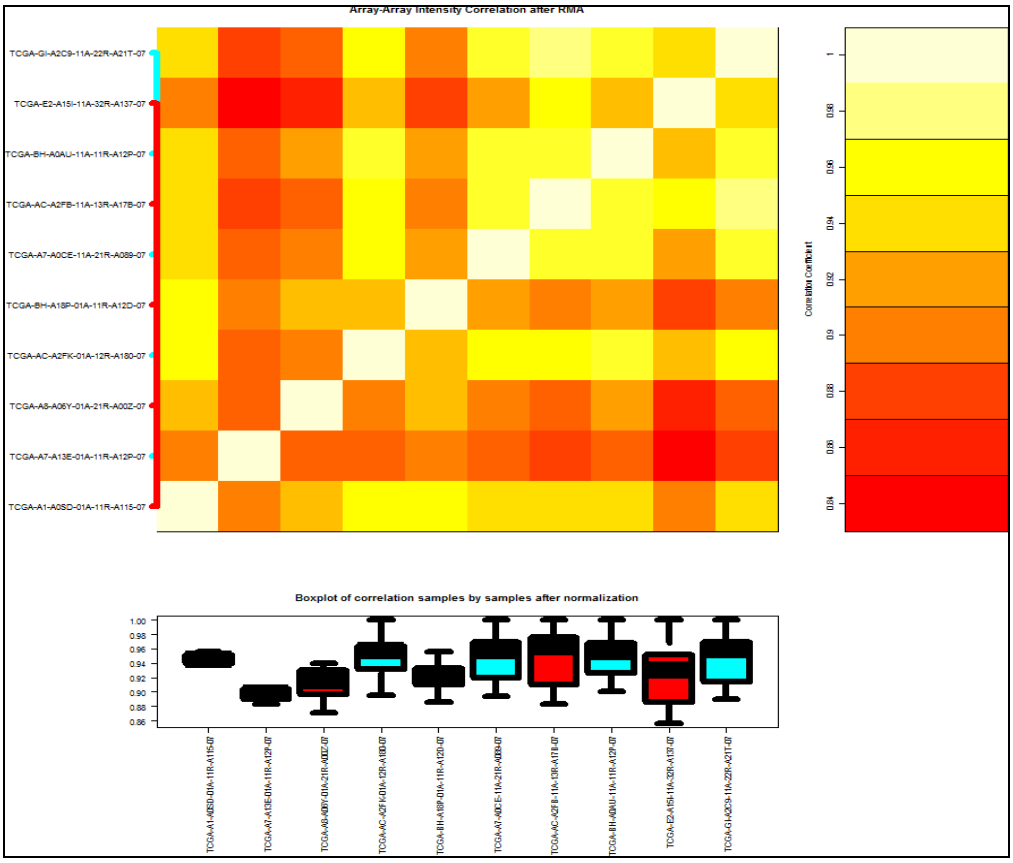

Figure 2. Heat map and Box plot for visual representation of breast cancer gene expression.

observed to be almost on a straight line, indicating a high level of normalization.

3.2. Differential Gene Expression Analysis

Among the total 21,022 genes, 613 genes were found to be differentially expressed with 254 genes upregulated and 359 genes downregulated (Table 2 and 3) after filtering them by criteria |log2FC| >1.5.

3.2. Enrichment Analysis

Using the given set of differentially expressed genes that are up-regulated under certain conditions, an enrichment analysis was performed to identify classes of genes or proteins that were over-represented, using annotations for that gene set. The bar plots in Figure 3 depict the involvement of these genes in biological processes, cellular components, molecular functions, and biological pathways derived from this analysis. The analysis mainly projected genes associated with nuclear division, mitotic influencers and cell cycle interactors closely associated with cell proliferation and division.

3.4. Survival Analysis

The survival probability of the 10 breast cancer samples was analyzed and the plot (Figure 4) was constructed using the respective clinical data of the samples. The samples whose vital status was alive showed a standard progression whereas the dead samples showed a decline in the survival probability. Survival probability was also analyzed using the Kaplan-Meier method which resulted in the list of survival genes in the samples (Table 4) and the plots for the respective genes for the given p-values were constructed and shown in Figure 5. The two sample cohorts are compared by a Kaplan-Meier survival plot, and the hazard ratio with 95% confidence intervals were also calculated.

3.5. Analysis of differentially methylated regions

The differentially methylated CpG sites among the samples were searched in this analysis. The beta-values were used (methylation values between 0.0 to 1.0) to compare two groups. First, the difference between the mean DNA methylation of each group for each probe was calculated and then the p-value was calculated utilizing Wilcoxon test adjusting by Benjamini-Hochberg method. The parameters were adjusted to require a minimum absolute beta-values difference of 0.2 and a p-value of < 0.01 (Table 5). A volcano plot with hypomethylated region as green dot (Figure 6) and the scale with x-axis showing differential mean methylation and y-axis showing its significance and FDR corrected -P values were obtained. This plot helps in identifying the differentially methylated CpG sites and to return the object with the calculus in throw Ranges.

3.6. Principal Component Analysis

To foresee the biological meanings, a PCA plot was constructed to visualize the top 200 differentially expressed genes from PC1 and PC2 in the DEGS list (Figure 7). By interpreting this matrix, the axes of maximal variance are noted called as: The Principal Components (PCs). They are fixed up in descending order of their contribution to the variance.

3.7. Sample Mean DNA Methylation Analysis

A mean DNA methylation boxplot in Figure 8 was created using the data and calculating the mean DNA methylation per group. To identify differentially methylated CpG sites, the method involves first computing the difference in mean DNA methylation between groups. Subsequently, differential expression between two groups is assessed using the Wilcoxon test, with adjustments made using the Benjamini-Hochberg method.

3.8. Starburst Plot

The starburst plot was processed to combine information from two volcano plots and was utilized to study the relationship between distribution of gene expression and DNA methylation levels differentially expressed genes between the low and high expression groups. It facilitates the comparison of these two variables, plotting the log10 (FDR-corrected P value) for DNA methylation (beta value) on the x-axis and gene expression on the y-axis for each gene. A black dashed line indicates the FDR-adjusted P value threshold of 0.01 (Figure 9) calculated using the Wilcoxon signed-rank test with the Benjamini-Hochberg adjustment method.

3.9. Oncoprint Analysis

Oncoprint bar plot was constructed to study the genomic alterations and mutations in the samples such as deletions, insertions, SNPs etc. (Figure 10). The SNPs were defined SNPs as maroon, insertions as yellow and deletions as purple. The upper barplot displays the number of genetic mutation per patient, while the barplot along the right shows the number of genetic mutations per gene. The oncoprint bar plot of the top 10 mutated genes showed major proportion of SNP'S and a few places with insertions.

3.10. Structure of Competitive Endogenous RNA network

The ceRNA network was constructed using the mRNA-miRNA-lncRNA interaction. First, the mRNAs that were differentially expressed (DGEs) were filtered. The filtered 254 mRNAs were used to retrieve the corresponding 153 miRNAs. Then the corresponding 352 lncRNAs for these miRNAs were retrieved which were collectively put together to form the network. The interaction network diagram of breast cancer related genes with the protein indicated the top-ranked PPI hub genes in which the connectivity mRNA (red spots)-miRNA (blue)-lncRNA (green) (Figure 11). The result of the analysis showed that the most frequently expressed lncRNAs were LINC00461 and MALAT1. Hence from the current study, it can be hypothesized that LINC00461 and MALAT1 shall be considered as promising therapeutic targets for Breast Cancer.

4. Discussion

Breast cancer is a severe life-threatening disease, stands as fifth leading cause of death in both developed and in developing countries. India stands first with 70,218 deaths in 2012 followed by China and US, reports Globocan,

Table 2. The 254 upregulated genes in the tumor samples.

FN1	LRRC15	FHDC1	SPAG5	BUB1	DLGAP5	CXADRP3	FAM64A	TDO2	HHIPL2
SLC7A2	RAG1AP1	A2ML1	KIF20A	KIF2C	HOXB13	HS6ST3	SKA3	FAM40B	GRM4
CRABP2	NELL2	B4GALNT3	KMO	PBK	ESM1	GTSE1	FBN2	PRR11	OIP5
CEACAM6	SOX12	CDK1	CST5	INHBA	SRPX	AURKB	HIST1H4H	DQX1	CENPI
PRLR	MBOAT2	PAQR4	GJB2	CEP55	HJURP	PYDC1	ESCO2	POLQ	C1orf230
MMP11	TPX2	CCNB1	EZH2	NCAPG	CDCA5	MUC13	CCNE2	HOXC11	FAM54A
CYP2B7P1	C19orf21	NUSAP1	DTL	CENPE	CASC5	NCAPH	DEPDC1	LHX9	ADAMTS19
UGT2B4	IGSF9	CST1	CDC20	E2F2	KIF14	IYD	SDS	CDC25C	NKX2-2
GPRC5A	CHRNA9	KIF11	ASF1B	BUB1B	NUF2	BIRC5	CDCA2	SPC25	ASCL1
NPNT	MMP9	CEACAM5	GLIS3	RTKN2	MYCN	SHCBP1	ELAVL2	XRCC2	C1orf135
SQLE	S100P	LINGO1	CCNA2	AURKA	BMP8A	TNFSF11	SLC30A8	ARX	DNAJC5B
BAMBI	ECT2	FGFR4	CCDC67	CCNB2	PKMYT1	WISP1	C3orf67	SGOL1	ZNF695
DIO2	ANLN	MMP13	DIO1	LOC84740	TTK	CDH2	CENPA	CLSPN	KLK4
COL11A1	CACNA1H	CBS	PVALB	MMP10	CA9	GAD1	NEURL	C6orf154	PPEF1
TOP2A	IQGAP3	EGLN3	UBE2C	KIF4A	NDC80	CKAP2L	SKA1	TRIM72	C6orf223
DHRS2	FOXM1	RRM2	KIAA1199	FAM83D	CGA	TROAP	CTAG2	PPAPDC1A	NEIL3
CENPF	MMP1	ZWINT	KIF23	CDC6	NXPH1	OTX1	E2F8	KRT75	LOC645323
KIAA1244	PCDHB2	MYBL2	NEK2	PPP2R2C	TLCD1	DIAPH3	GPRIN1	ULBP1	PAX2
SLC7A5	SBSN	TNNT1	WDR62	LOC387646	HMMR	F12	DEPDC1B	GAS2L3	AIM1L
EEF1A2	ASPM	SCGN	PLK1	UBE2T	EXO1	KIF18B	HAGHL	FAM72D	HAPLN1
COL10A1	KCNK1	KIFC1	HOXC13	ESPL1	DLX5	PHEX	LEFTY1	CLEC5A	POTEC
SCUBE3	SIX4	PAX7	CCDC3	FAM5C	DNA2	KIAA1211	IBSP	E2F7	PAPL
MKI67	TK1	GDF15	MELK	C20orf103	KIF15	PCDHA1	SULT1C3	CCDC87	NMU
COMP	AOC3	CNTNAP2	CYP2C8	KCNG1	HOTAIR	ART3	HSD17B6	SPOCD1	EPO
SYT13	DBNDD1	FUT2	SIM2	MCM10	CDCA3	HORMAD1	TFR2	GABRD	SALL4
SPC24	SYNGR3	ARHGAP11A	DKFZp686O24166						

Table 3. The 359 downregulated genes in the tumor samples.

TXNIP	SLIT3	GHR	AKR1C1	GGTA1	DENND2A	HBB	KLB	P2RY14	MARCO
DCN	C10orf116	ANK2	EBF1	IGFBP6	CIDEC	SLC19A3	HBA2	C6	AVPR2
GSN	PER1	GPD1	EEPD1	MMRN1	CREB5	TFPI2	HYAL1	SLC22A3	BCHE
FOS	DGAT2	LIMS2	SDPR	VSIG4	MRGPRF	PDZK1	USHBP1	PHYHIP	BMPER
DUSP1	FHL1	LIPE	HSPB6	GPR146	DTX1	NOVA1	ALDH1L1	PLA2G2A	EBF2
ZFP36	ACACB	KLF2	GNAI1	PPARG	CALB2	CXCL2	MSX1	F10	C2orf89
SERPINF1	PALMD	GYG2	CDKN1C	FAM107A	FAM89A	GPR109A	ABCC6	PCK1	ACSM5
CAV1	LPL	TF	G0S2	PDE3B	CD300LG	LGI4	FXYD1	NNAT	NAALAD2
GPX3	DPT	CDCA8	NRN1	TMOD1	ANGPT1	CA3	TCEAL7	MYEOV	SLC17A7
SORBS1	SEMA3G	GINS1	ADAMTSL4	DPP4	PKDCC	LYVE1	AQP7	TMEM88	DMGDH
CD36	S1PR1	ITGA7	BIN1	CNRIP1	ANGPTL1	GABRE	ATOH8	EMX2OS	NMUR1
ANGPTL2	PLIN1	ALPL	TMEM37	TUSC5	GRASP	MYOM1	ACVR1C	CNTFR	LRRC2
FZD4	ADAMTS5	GALNTL2	NPR1	GPR34	RSPO3	CD209	T6GALNAC	LGALS12	ABCD2
SOCS3	KIF26B	ISM1	CES1	MRC1	EBF3	KLF15	C1QTNF7	HSD11B1	FOXP2
PLIN4	ITIH5	CFD	AKR1C2	TBX15	KANK3	PCOLCE2	CCDC85A	IL6	FNDC5
LTBP4	GNG11	AKR1C3	LOC654342	MS4A4A	NR4A3	HSPB7	CITED1	IGSF1	CHST8
PDK4	CCDC69	C13orf15	APBB1IP	CDO1	TAL1	BHMT2	SLC2A4	LILRB5	FAM3D
MFAP4	KLF4	CYGB	CLDN5	RBP4	HRASLS5	TMEM100	CIDEA	MT1M	GPR64
ADH1B	ECM2	GYPC	EPDR1	PRUNE2	HSPB2	LEP	DES	BMP2	BANK1
TNXB	RARRES2	PAMR1	PDE2A	WISP2	SOX17	TMEM22	ERMN	TIMP4	TNNI3K
ENPP2	APCDD1	MMD	SNCG	C13orf33	ATP8B4	LRRN4CL	RDH5	RRAD	KLHL4
ADIPOQ	ALDH1A1	NDN	CLEC3B	FOLR2	GPIHBP1	PLAC9	MLXIPL	GPR109B	SCUBE1
FABP4	PTGER3	ABCC9	BTNL9	PRG4	KCNIP2	NIPSNAP3B	ADRB2	P2RY12	SLC25A18
AKAP12	CHRDL1	GIMAP5	IRAK3	STX11	NMB	LRRC70	SCN4A	MT1L	PYGM
GPAM	FOSB	TPPP3	HSPA12B	C14orf139	GPR133	TMEM132C	AGAP11	PTH1R	IQSEC3
IL17D	MAP1LC3C	ABCB5	ARHGAP36	AADAC	C14orf180	C4orf49	LY6D	LMX1A	TDRD10
MYOC	GFRA2	SPINK5	SLC14A2	KRT4	EMX2	C2CD4B	C8orf34	CCL23	C1QTNF9
SGK2	ITIH2	CALCR	TMEM195	C13orf36	RGS6	GPBAR1	GLYAT	SNTG2	FGF10
ASPA	CCBE1	CDR1	ADRA1A	SFTPA1	PKHD1L1	MGC45800	MRAP	FGB	FAM162B
AQP4	KCNC2	ELOVL3	FRMPD2	FGG	PLA2G5	LOC283392	CES4	CPA1	COL25A1
CYP11A1	PKD1L2	ANGPTL7	MMP27	CSN1S1	SLC1A7	ADH1A	CYP4F12	SCG3	APOB
ANKRD53	CPEB1	GPR97	PAPPA2	LALBA	HBA1	TNNT3	HSD17B13	GLRA3	XPNPEP2
PFKFB1	MPZ	FAM166B	HAS1	FGA	DNAH9	ADRB1	MYH1	MUC7	TNMD
HIF3A	AQPEP	KIAA0408	ODF3L1	XAGE1D	CCDC141	DMBT1	CSF3	NTS	DMRT2
HEPACAM	FAM180B	MYO16	CHGB	HEPN1	TRHDE	CRHBP	RASGEF1C	NPY5R	BMP3
PRRT4	CNKSR2	PRCD	IGFBP1	FRMD1	SFTPB	GYS2	CLEC4G	C12orf39	

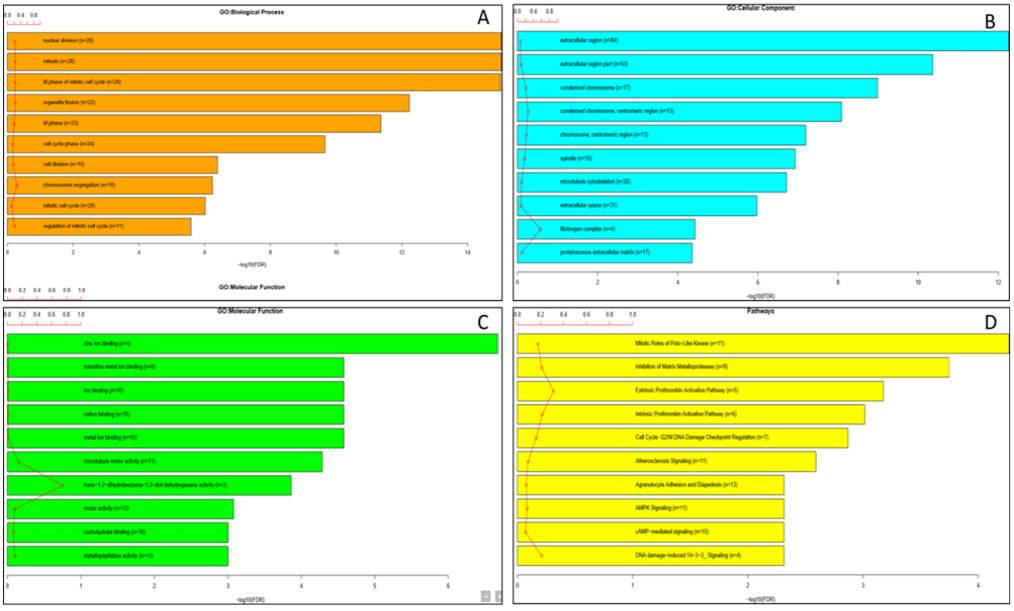

Figure 3. Gene Ontology Enrichment graphs in (A) Biological process, (B) Cellular components (C) Molecular functions (D) Molecular pathways.

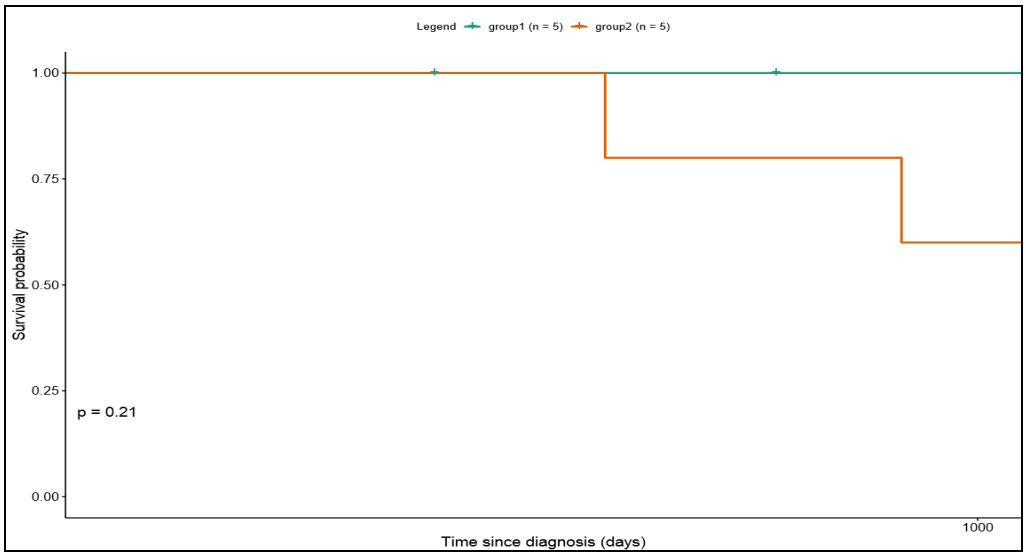

Figure 4. Survival plot representing the live (Green) and dead (Red) state of the samples.

Table 4. KM List of genes after survival analysis.

Genes	Pvalue	Group2 Deaths	Group2 Deaths with Top	Group2 Deaths with Down	Mean Group2 Top	Mean Group2 Down	Mean Group1
ABCF1|23	0.03895	2	2	0	12.39	11.468	12.007
A2M|2	0.05878	2	0	2	16.86	13.821	15.591
AADAC|13	0.05878	2	0	2	7.2325	0	2.9774
ABCA1|19	0.05878	2	0	2	12.771	10.652	11.644
ABCA2|20	0.1573	2	0	2	12.085	11.348	11.775
ABCA4|24	0.19854	2	0	2	8.685	4.9951	6.7635

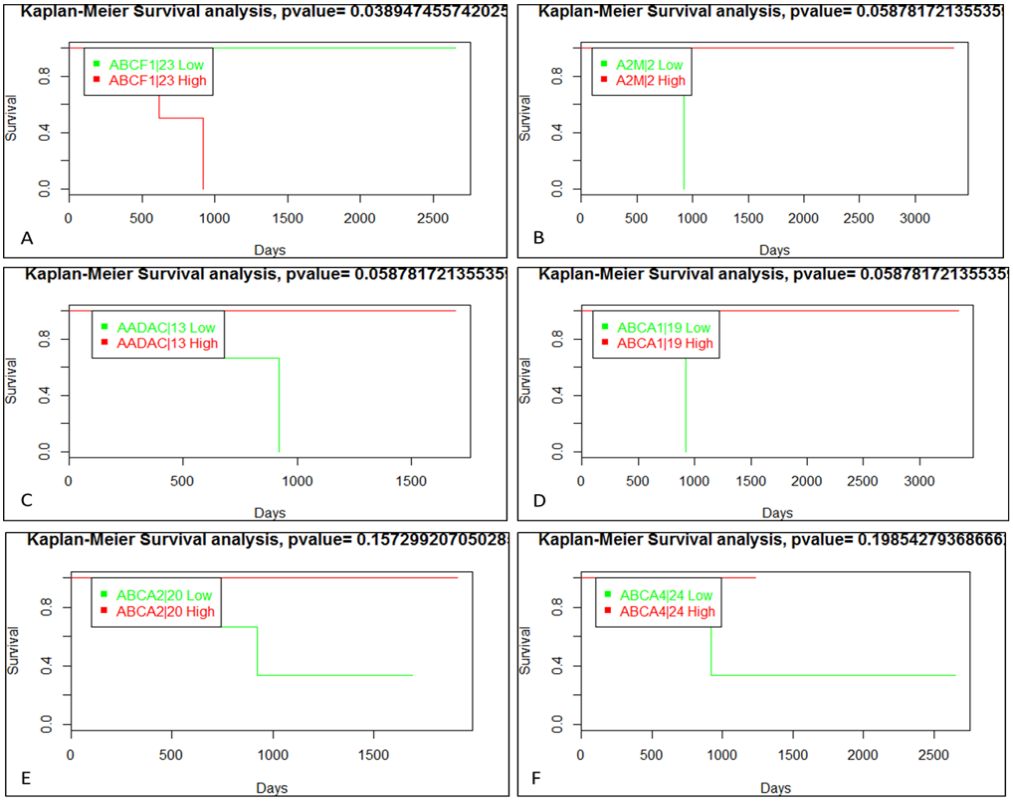

Figure 5. KM analysis of gene (A) ABCF1|23, (B) A2M|2, (C) AADAC|13, (D) ABCA1|19, (E) ABCA2|20 and (F) ABCA4|24.

Table 5. DMR results showing 13 out of 200 probes (Column highlighted shows the Hypermethylated and Hypomethylated probes).

feature _id	mean. group. 1	mean. group. 2	Diff mean. group.1. gr oup.2	p.value.g rou p.1.grou p.2	p. value. adj. gro up.1. group.2	status. group. 1. group.2	Diff mean.gro up.2. group.1	p.value. group.2. group.1	p.value. adj.grou p.2.group.1	status. group. 2. group.1
ID001	56612	5959.83	298.8017	0.85343	0.990319	Not Significant	-298.82	0.85343	0.990318884	Not Significant
ID002	5358.42	6447.374	1088.972	0.3935	0.977444	Not Significant	-1088.97	0.39305	0.977443609	Not Significant
ID003	6277.774	4941.862	-1335.91	0.24745	0.97744	Not Significant	1335.912	0.24745	1.977443609	Not Significant
ID004	6330.645	4827.311	-1503.33	0.27986	0.97744	Not Significant	1503.334	0.27986	1.977443609	Not Significant
ID005	5252.913	4950.991	-301.922	1	1	Not Significant	301.9225	1	1	Not Significant
ID006	4827.31	5966.27	1138.969	0.52885	0.97744	Not Significant	-1138.97	0.52885	0.977443609	Not Significant
ID007	5059.958	4393.714	-666.244	0.52885	0.97744	Not Significant	666.244	0.52885	0.977443609	Not Significant
ID008	4277.769	5860.694	1582.925	0.35268	0.97744	Not Significant	-1582.93	0.35268	0.977443609	Not Significant
ID177	6832.162	4391.746	-2440.42	0.00389	0.777241	Hypomethyl ated in group 2	2440.416	0.00389	0.777241335	Hypermethyla ted in group 1
ID009	5679.338	4807.155	-872.183	0.63053	0.977444	Not Significant	872.1829	0.63053	0.977443609	Not Significant
ID010	4451.933	2348.386	-2103.55	0.01469	0.977444	Not Significant	2103.547	0.01469	0.977443609	Not Significant
ID011	4679.282	5703.219	1023.937	0.52885	0.977444	Not Significant	-1023.94	0.52885	0.977443609	Not Significant
ID012	5218.846	5557.657	338.8109	0.9118	0.990319	Not Significant	-338.811	0.9118	0.990318836	Not Significant

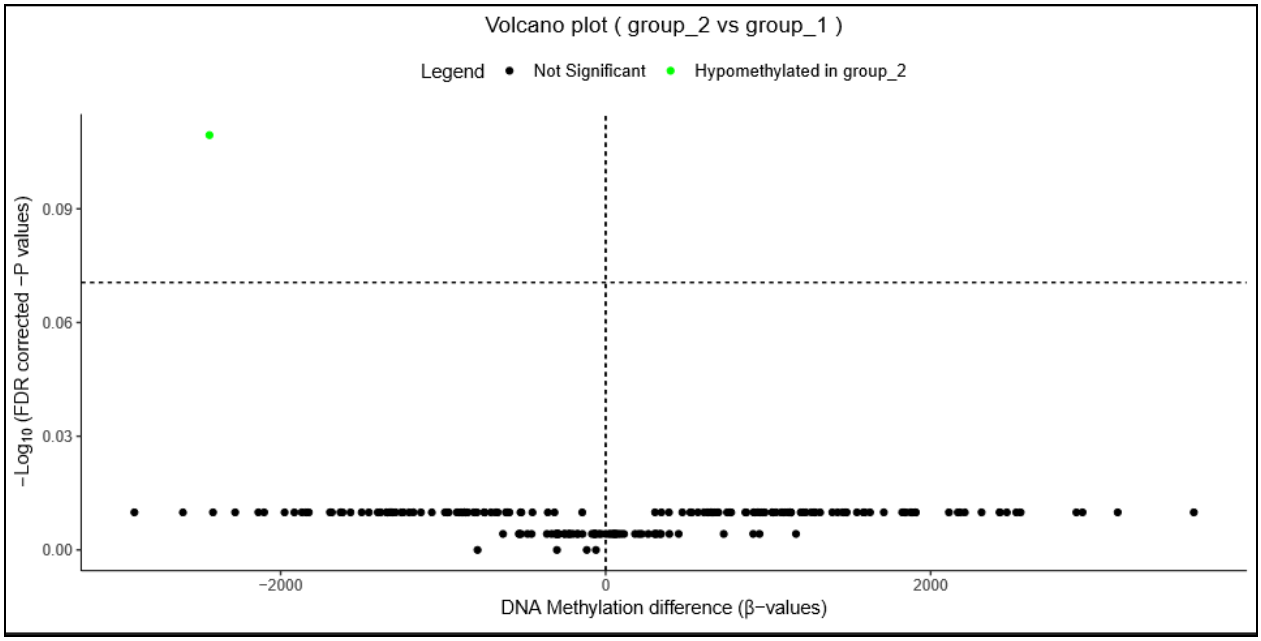

Figure 6. Volcano plot showing differentially methylated regions.

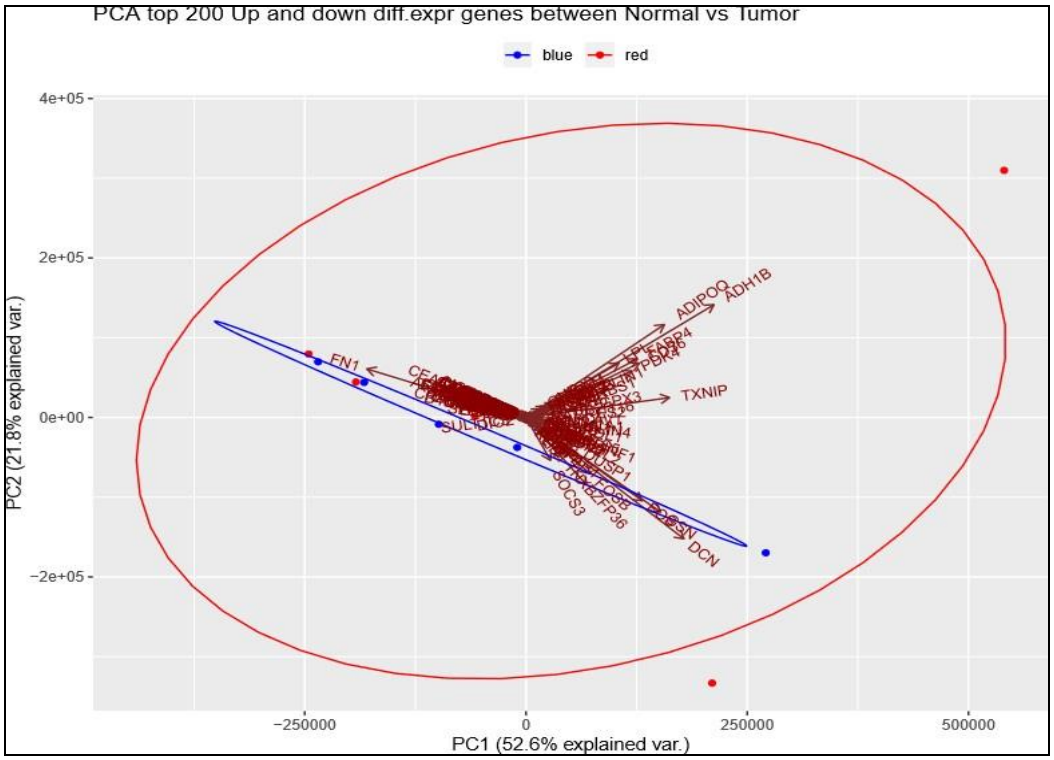

Figure 7. PCA plot for DEGs between Normal vs Tumor samples.

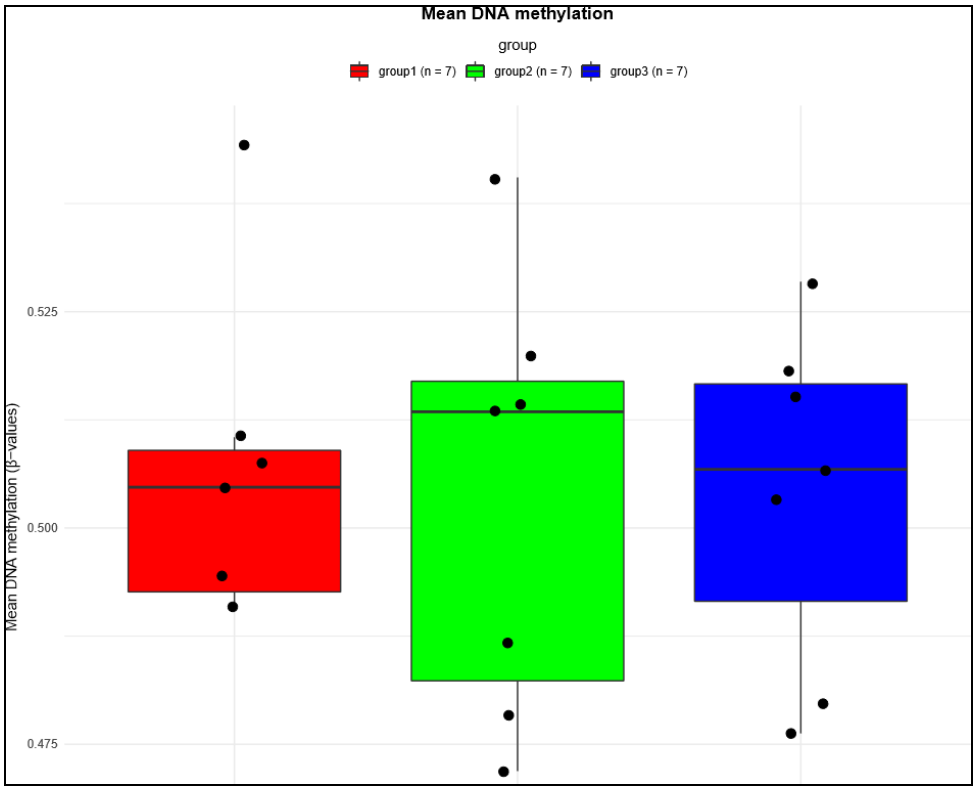

Figure 8. Boxplot showing Mean DNA Methylation for 3 groups.

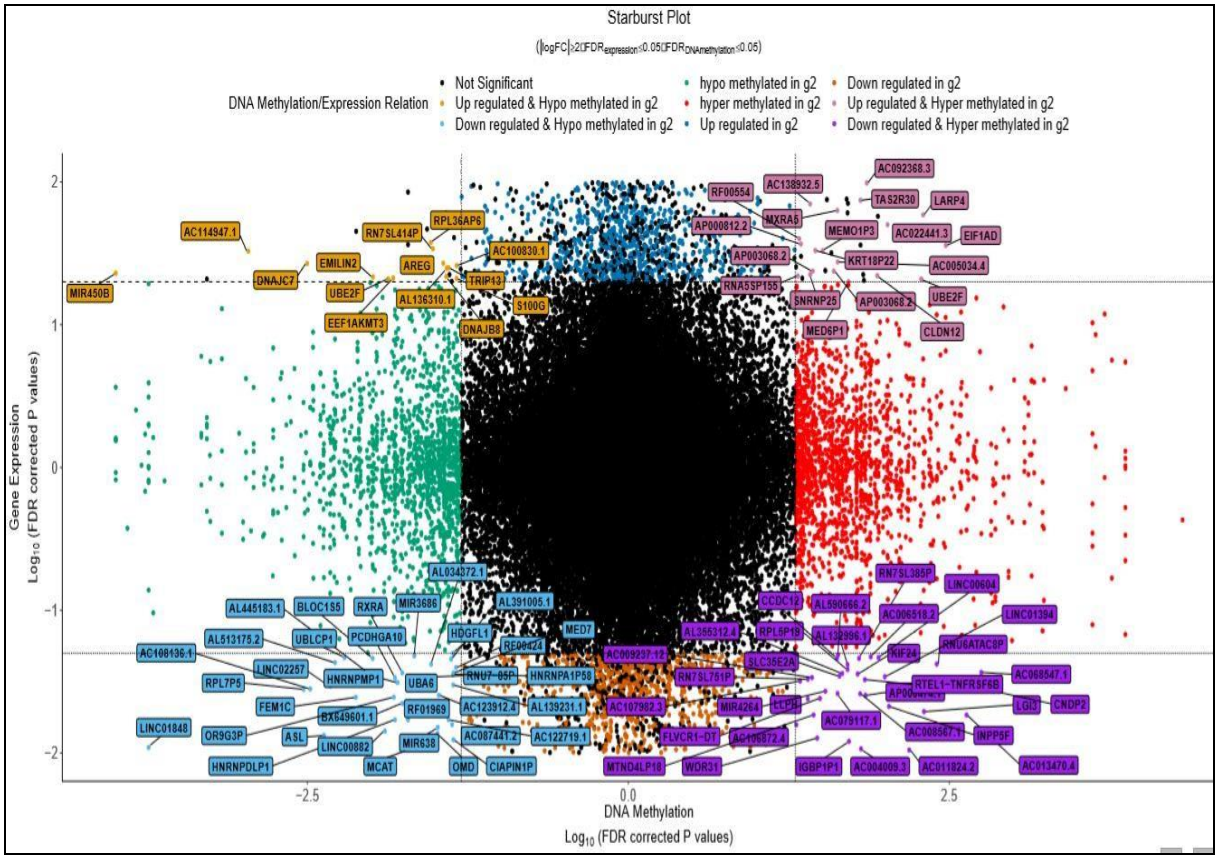

Figure 9. Starburst Plot showing DNA Methylation/Expression Relation.

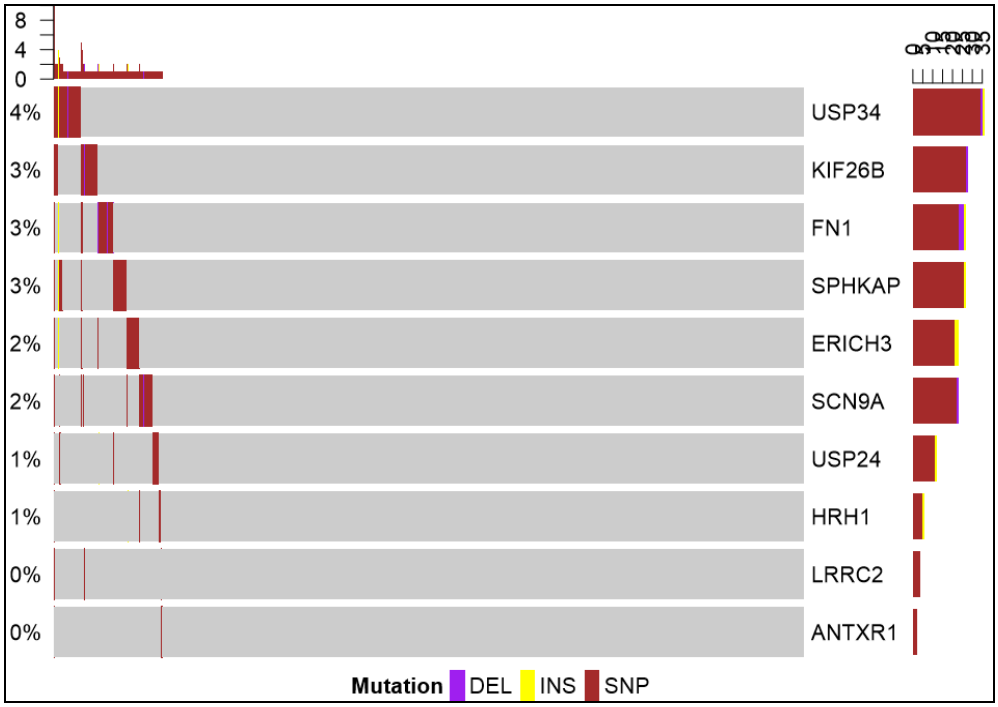

Figure 10. Oncoprint showing mutations across the top 10 mutated genes.

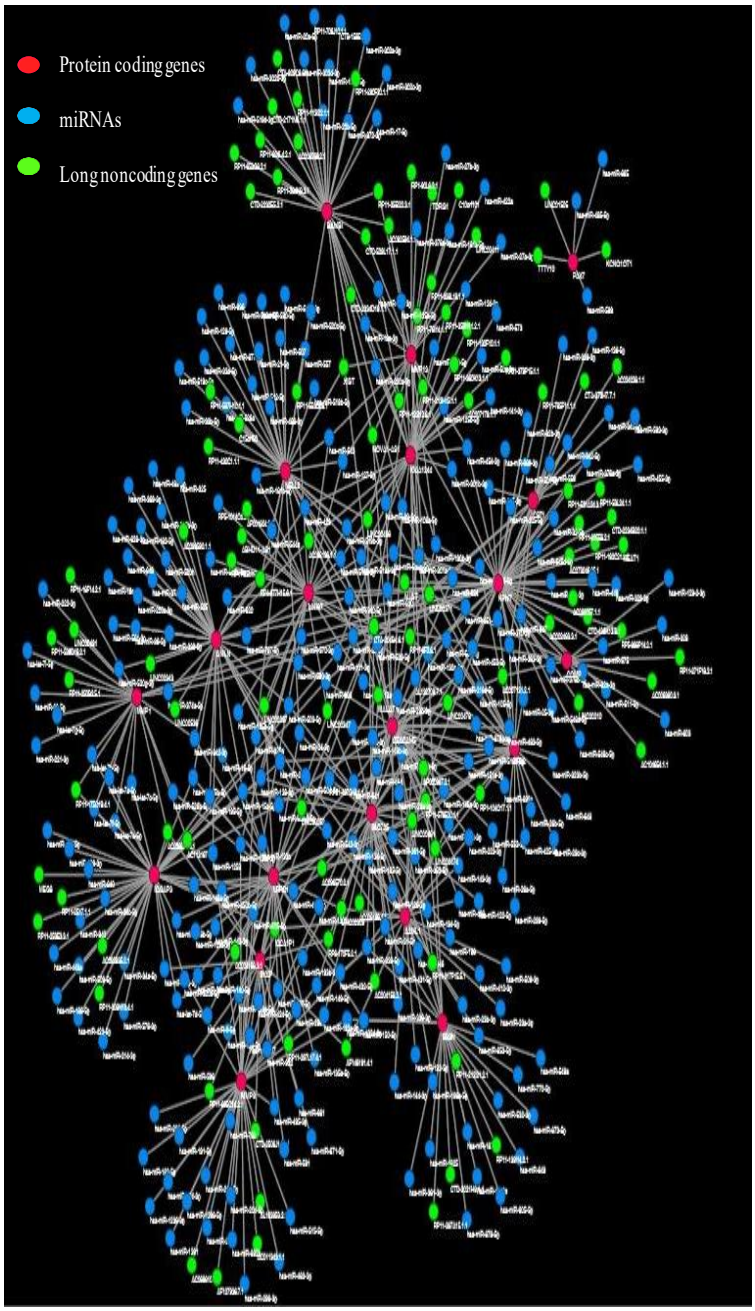

Figure 11. Competitive endogenous RNA network.

WHO (18). The past two decades has led to incredible advancement in our understanding of the breast cancer gene expression pattern, metabolomics and interactomics (19). Breast cancer represents a highly heterogeneous disease characterized by various distinct biological entities, each with specific pathological features and behaviors. Different breast tumor subtypes exhibit unique risk factors, histological characteristics, and responses to treatments (20) . Non-coding RNAs (ncRNAs) are transcriptionally nonfunctional as they lack protein coding function. They act as key regulators in cellular processes, such as gene expression, cell proliferation, differentiation, and apoptosis. The competitive endogenous RNA (ceRNA) is the pool of mRNAs, lncRNAs, and other non-coding RNAs sharing common MREs with miRNAs. The ceRNA regulation hypothesis has been demonstrated to play a significant role in cancer development (12) . The current research on integrated analysis of cancer data from TCGA database using computational methods including data preprocessing, normalization and other such analyses such as enrichment analysis, probability of survival, prediction of differentially methylated regions and differentially expressed genes in cancers has been previously reported in Pancreatic cancer and colorectal cancer (21) but unexplored in breast cancer. Identification of potential lncRNA biomarkers from integrated analysis of long non-coding RNA-associated ceRNA network was already proven to be successful in previous researches on cancers such as pancreatic cancer and ovarian cancer (10) and human lung adenocarcinoma (22). Differential gene expression analyses were performed with the sequence count data from TCGA (23) . The TCGA biolinks in R/Bioconductor package addresses the challenges such as TCGA data retrieval and integration with clinical data and other molecular data types like RNA and DNA methylation. It offers a guided workflow enabling users to query, download, and conduct integrative analyses of TCGA data. By combining methodologies from computer science and statistics, and incorporating techniques from previous TCGA marker studies, TCGA biolinks enhances reproducibility and facilitates integrative analysis. It leverages various Bioconductor packages to foster advancements and expedite novel discoveries (14, 24). This is an attempt to computationally analyze the TCGA cancer data and then evaluate the lncRNAs associations to construct ceRNA network in order to identify the therapeutic lncRNA biomarkers for breast cancer. In a previous study (25), a ceRNA network for breast cancer was analyzed with miRNAs but did not relate the miRNA expression. Whereas, Zhou et al., (26) performed a ceRNA network based on the miRNA–mRNA combinations. In 2018, Zhou et al. (27) constructed four BC-related ceRNA networks with lncRNAs, miRNAs and mRNA, but the work did not construct lncRNAs prognostic signatures. In our study, we have investigated the association of lncRNA in breast cancer thereby identify novel lncRNAs as therapeutic targets. Recent reports predict the combined frequency of differentially expressed lncRNAs with clinical variables of Colorectal Cancer (CRC). Two lncRNAs (LINC00400 and LINC00355) in ceRNA network has proven to show significant changes in multiple colorectal cancer pathological stages thereby acting as potential targets for CRC (13) In the present study, differentially expressed genes from the given samples were identified by applying the criteria thresholds to $|\log2FC| > 1.5$ and FDR-value < 0.01. The results identified 352 lncRNAs, 183 miRNAs and 254mRNAs with aberrant expression in breast cancer. Out of the total

352 lncRNAs, some lncRNAs were found to be commonly shared among the mRNAs. From these, LINC00461 and MALAT1 lncRNAs can be considered as targets for breast cancer as the interactions were common with mRNAs. In conclusion, we utilized breast cancer RNA-Seq data from TCGA and conducted comprehensive computational analyses using R programming. Visual and logical inference was drawn from the results of each analysis. A total of 613 genes were identified as differentially expressed among the samples, with 254 genes upregulated and 359 genes downregulated. These differentially expressed genes, identified using the TCGA biolinks package were subsequently used to construct a ceRNA network, bridging the initial and subsequent parts of the study. From the results using TCGA biolinks, 352 lncRNAs, 183 miRNAs and 254 mRNAs were found to show aberrant expression in the studied breast cancer samples. Notably, among the 352 lncRNAs, LINC00461 and MALAT1 emerged as consistently and prominently expressed lncRNAs and found that their regulated miRNA target genes are enriched in the samples.

Acknowledgment
The authors thank the management of Stella Maris College (Autonomous) and Karpagam Academy of Higher Education for their support and encouragement.

Authors' Contribution
Conceptualization, SJ; methodology, CH, SJ, AS, BS, PP; investigation, SJ, AS; resources, CH, SJ, AS; data curation, CH, SJ; writing— original draft preparation, CH, SJ; writing—review and editing SJ, BS, PP; visualization, CH, SJ; supervision, SJ. All authors have read and agreed to the published version of the manuscript.

Ethics
All data were acquired from the TCGA database. Hence, ethical committee assessment is not required for the study.

Conflict of Interest
The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Grant Support
Not Applicable.

Data Availability
The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

References

1. Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The lancet. 2005 Nov 19;366(9499):1784-93.
2. Levine AJ, Puzio-Kuter AM. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. Science. 2010 Dec 3;330(6009):1340-4.
3. Howlader NN, Noone AM, Krapcho ME, Miller D, Brest A, Yu ME, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS. SEER cancer statistics review, 1975–2016. National Cancer Institute. 2019 Apr 8;1.
4. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics, 2013. CA: a cancer journal for clinicians. 2014 Jan;64(1):52-62.
5. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical cancer research. 2005 Aug 15;11(16):5678-85.
6. Wang P, Li J, Zhao W, Shang C, Jiang X, Wang Y, Zhou B, Bao F, Qiao H. A novel LncRNA-miRNA-mRNA triple network identifies LncRNA RP11-363E7. 4 as an important regulator of miRNA and gene expression in gastric Cancer. Cellular Physiology and Biochemistry. 2018 Jul 26;47(3):1025-41.
7. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007 Jun 8;316(5830):1484-8.
8. Sana J, Faltejskova P, Svoboda M, Slaby O. Novel classes of non-coding RNAs and cancer. Journal of translational medicine. 2012 Dec;10:1-21.
9. Ye Y, Li SL, Wang SY. Construction and analysis of mRNA, miRNA, lncRNA, and TF regulatory networks reveal the key genes associated with prostate cancer. PloS one. 2018 Aug 23;13(8):e0198055.
10. Zhou M, Diao Z, Yue X, Chen Y, Zhao H, Cheng L, Sun J. Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. Oncotarget. 2016a Jul 28;7(35):56383.
11. Zhou M, Wang X, Shi H, Cheng L, Wang Z, Zhao H, Yang L, Sun J. Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer. Oncotarget. 2016b Feb 3;7(11):12598.
12. Fang XN, Yin M, Li H, Liang C, Xu C, Yang GW, Zhang HX. Comprehensive analysis of competitive endogenous RNAs network associated with head and

neck squamous cell carcinoma. Scientific Reports. 2018 Jul 12;8(1):10544.
13. Yuan W, Li X, Liu L, Wei C, Sun D, Peng S, Jiang L. Comprehensive analysis of lncRNA-associated ceRNA network in colorectal cancer. Biochemical and biophysical research communications. 2019 Jan 8;508(2):374-9.
14. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia. 2015 Jan 20;2015(1):68-77.
15. Lu TP, Lee CY, Tsai MH, Chiu YC, Hsiao CK, Lai LC, Chuang EY. miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. PLoS One. 2012;7(8):e42390.
16. Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics. 2012 Aug 1;28(15):2062-3.
17. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape automation: empowering workflow-based network analysis. Genome biology. 2019 Dec;20:1-5.
18. Donepudi MS, Kondapalli K, Amos SJ, Venkanteshan P. Breast cancer statistics and markers. Journal of cancer research and therapeutics. 2014 Jul 1;10(3):506-11.
19. Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. Biological research. 2017 Dec;50:1-23.
20. Dai X, Xiang L, Li T, Bai Z. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. Journal of cancer. 2016 Jun 23;7(10):1281.
21. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, Colaprico A. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell. 2018 Apr 5;173(2):338-54.
22. Sui J, Li YH, Zhang YQ, Li CY, Shen X, Yao WZ, Peng H, Hong WW, Yin LH, Pu YP, Liang GY. Integrated analysis of long non-coding RNA-associated ceRNA network reveals potential lncRNA biomarkers in human lung adenocarcinoma. International journal of oncology. 2016 Sep 30;49(5):2023-36.
23. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology 2010;11(10):R106.
24. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research. 2016 May 5;44(8):e71.
25. Chen J, Xu J, Li Y, Zhang J, Chen H, Lu J, Wang Z, Zhao X, Xu K, Li Y, Li X. Competing endogenous RNA network analysis identifies critical genes among the different breast cancer subtypes. Oncotarget. 2016 Dec 29;8(6):10171.
26. Zhou X, Liu J, Wang W. Construction and investigation of breast-cancer-specific ceRNA network based on the mRNA and miRNA expression data. IET systems biology. 2014 Jun;8(3):96-103.
27. Zhou S, Wang L, Yang Q, Liu H, Meng Q, Jiang L, Wang S, Jiang W. Systematical analysis of lncRNA–mRNA competing endogenous RNA network in breast cancer subtypes. Breast cancer research and treatment. 2018 Jun;169:267-75.