DOI: https://doi.org/10.54393/pbmj.v5i10.813

PAKISTAN BIOMEDICAL JOURNAL
https://www.pakistanbmj.com/journal/index.php/pbmj/index
Volume 5, Issue 10 (October 2022)

Original Article

Functional and Network Exploration of RNASeq Data of Breast Cancer

Tehreem Anwar[1*], Mirza Jawad ul Hasnain[2] and Vina Kanwal[2]

[1]Lahore College for Women University, Lahore, Pakistan
[2]Virtual University of Pakistan

ARTICLE INFO

Key Words:
Breast cancer, Bioinformatics, R software, Network analysis, Gene expression analysis

How to Cite:
Anwar, T. (2022). Functional and Network Exploration of RNA Seq data of Breast Cancer: Analysis of RNASeq data of breast cancer. Pakistan BioMedical Journal, 5(10).
https://doi.org/10.54393/pbmj.v5i10.813

***Corresponding Author:**
Tehreem Anwar
Lahore Medical Research Center, Lahore, Pakistan
Tehreemanwar33@gmail.com

Received Date: 11th October, 2022
Acceptance Date: 23rd October, 2022
Published Date: 31st October, 2022

ABSTRACT

This study comprised of RNASeq data analysis of breast cancer. It includes statistical, functional and network analysis by various bioinformatics tools. Breast cancer is the most frequent cancer in women and affects everyone, including the young and elderly, rich and poor, women and children. **Objective:** To explore dataset of breast cancer, network and functional wise. Although there is extensive research on breast cancer, in silico studies on this topic are very rare. **Methods:** The study makes use of GEO (Gene Expression Omnibus) database from where data was collected. The data obtained of Breast cancer samples was normalized for which R language was used (using Limma, RPKM values) which eventually gave differentially expressed genes which were mainly involved in causing this Breast cancer and up- and down-regulatory genes were found using logFC values. Then functional analysis of these up- and down-regulated genes was performed using David Software. Then network analysis was performed, which showed the co-relation between the genes in making this Breast cancer prevalent in patients. Finally, importance of our genes was studied by using cBioPortal database. **Results:** Six important and novel genes were identified as differentially expressing through R software. Functional and network analysis and their significance studied by cBioportal dictated several potential genes taking part in important cancer and other pathways paving way for further research. **Conclusions:** The pathways and candidate genes were selected based on high enrichment score and these genes and pathways play a significant role in breast cancer.

INTRODUCTION

Breast cancer is basically a heterogeneous group of neoplasms originated via epithelial cells, which line the milk ducts. Three major types of breast cancer are: Luminal tumors, Human Epidermal Growth Factor 2+ (HER2+) tumors and Basal like tumors [1]. Breas cancer is a matter of discussion and research in histology and clinical outcomes for a long span of time. Breast cancer is the most common type of cancer in females and a major cause of death also. 565,650 deaths from cancer occurred in the United States in 2008 [2]. In China in year 2005, The mortality rate was 70.7 thousand in which 1.2 thousand were of males and 69.5 thousand were of females [3]. Among the Asian population, Karachi reports the highest incidence of breast cancer [4]. Of all the key challenges that occur during breast cancer research, mapping of pathways that give rise to metastasis is one of them. It is significant to analyze gene expression profiles to identify markers which correlate with metastasis [5]. Clinical and pathological risk factors, such as patient age, tumor size, and steroid receptor status, are commonly used to assess the likelihood of metastasis development. Aggressive adjuvant therapy can be prescribed when metastasis is likely, which has widely led to significant decrease in breast cancer mortality rates [6]. The established risk factors of Breast cancer are linked to oestrogens. Risks are increased by several factors including early menarche, late menopause, and obesity in postmenopausal women. Childbearing somehow reduces risk, with a greater protection for early first birth and a larger number of births; illustrated as breastfeeding probably has a protective

DOI: https://doi.org/10.54393/pbmj.v5i10.813

effect [7]. Moreover, the oral contraceptives and hormonal therapy for menopause both tend to cause a little increase in breast-cancer risk. Alcohol also increases risk, whereas physical activity probably stays protective. Mutations in certain genes also massively enhance the risk to breast cancer, but these certainly account for a minority of cases [8]. Different bioinformatics tools were used in this study for comprehensive analysis of RNASeq dataset obtained of breast cancer from an online database. R Language is quite famous because it offers different platforms to different users. It is basically a programming language that requires input via a command line which may not be an easy thing for non-coders or beginners so there are several packages in R that can be utilized to complete the task for statistical analysis and data visualization [9]. STRING predicts interaction information as well as offer exclusively comprehensive attention and ease of access to both experimental information. The freshly restructured DAVID Bioinformatics Resources contains the DAVID Knowledgebase and five integrated, web-based functional annotation tool sets: the DAVID Gene Functional Classification Tool, the DAVID Functional Annotation Tool, the DAVID Gene ID Conversion Tool, the DAVID Gene Name Viewer and the DAVID NIAID Pathogen Genome Browser [10]. Cytoscape is a broadly used open-source software structure for net visualization and manipulation [11,12]. The objective of this study is to discover and explore transcriptome data of breast cancer by using next generation sequencing platform (i.e., RNASeq) which will be a road to success in further analysis of the fundamental genes and their networks of communications that would have been involved in development of cancer.

METHODS

GEO is an international public source that keeps records and spontaneously allocates microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. Accession number of dataset taken from GEO is GSE52036. This dataset studied the effect of regulation of Endogenous aryl hydrocarbon receptor (AHR) on expression of tumor necrosis factor target genes in MCF-7 cancer cells and had total of about 10 samples of MCF-7 Breast Cancer cells, 5 of which were AHR knockdown replicates and other 5 were controls. The data collected from GEO was in the form of raw reads against each gene in tabular .csv form. R software 3.4.4 was used on the dataset for the normalization and Differential Gene Expression of read count file for the comparison. The normalization of our dataset was performed via R package "Limma". The language holds immense popularity due to its user-friendly platforms for different users. Then differential gene expression analysis (DGE) analysis was performed using reads per kilo-base per million mapped reads (RPKM) package from R. The formula used for the purpose of finding the RPKM values is: raw.data [,2+i] / (raw.data$Length/1000) / (library.sizes[i]/1000000). Finally, edgeR package of R language was used to find upregulated and down regulated genes based on the LogFC values [13]. The functional analysis is usually performed by using DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool to figure out the gene annotations along with their performance enrichment score. To find the GO (Gene Ontology) terms and the appropriate pathways where the significant genes exist, our differentially Expressed gene was uploaded in the DAVID Tool. For constructing linkages of diverse genetic factor, to find out associations of genes and to know whether these genes play a vital role in Breast cancer or not, gene interactions were studied via STRING Tool and the network analysis was performed via Cytoscape Tool. The genes of importance founded by STRING and cytoscape was then selected and submitted to cBioPortal to find significance of reported genes.

RESULTS

Library size (the sum of all the read counts for each sample) for the dataset is given in Table 1. To normalize the dataset, first the zero counts were removed from data which may influence results. About 10.1% genes were with zero counts which is a big ratio to corrupt a data analysis. Then for further narrowing down the data, reads per million (RPM) analysis was performed because genes in sample with 0 value will give 0 RPM. Then RPKM values was found by using RPKM package of R language which gives us upregulated and down regulated genes (Table 2). After that we selected top 10 genes based on the LogFC values found by edgeR (Table 3) and compared their RPKM values (Figure 1). Only those genes are significant whose RPKM value is greater than 1.

Sample	Read Counts
AHR1	45364096
AHR2	32781187
AHR3	53206492
AHR4	54161455
AHR5	64111047
Control1	55142434
Control2	73188324
Control3	46699885
Control4	109927638
Control5	653292

Table 1: Library size of dataset

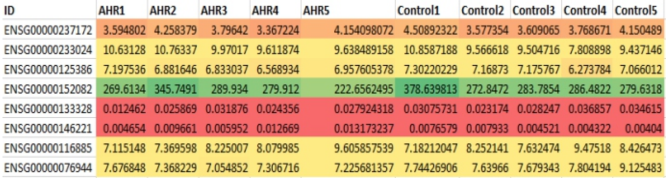
DOI: https://doi.org/10.54393/pbmj.v5i10.813

Sample	Genes
Downregulated (-1)	228
Not affected (0)	17860
Up regulated (1)	722

Table 2: Upregulation and downregulatory genes found by RPKM

LogFC	pValue	ID	GeneID	Gene Name
-9.086250033	8.98E-72	8607	ENSG00000237172	B3GNT9
-6.543974824	8.65E-32	15326	ENSG00000233024	AC126755.2
-6.499471573	4.95E-31	10776	ENSG00000125386	FAM193A
-5.538470452	4.04E-72	1679	ENSG00000152082	MZT2B
-4.983856717	1.79E-05	4479	ENSG00000185894	BPY2C
7.827358342	1.37E-170	13440	ENSG00000133328	HRASLS2
7.820179631	3.34E-170	15038	ENSG00000146221	TCTE1
7.308999452	3.54E-28	1847	ENSG00000116885	OSCP1
7.308999452	3.54E-28	13546	ENSG00000076944	STXBP2
7.281094657	1.10E-27	2744	ENSG00000116882	HAO2

Table 3: Top 10 genes selected by their LogFC values



ID	AHR1	AHR2	AHR3	AHR4	AHR5	Control1	Control2	Control3	Control4	Control5
ENSG00000237172	3.594802	4.258379	3.79642	3.367224	4.154098072	4.50892322	3.577354	3.609065	3.768671	4.150489
ENSG00000233024	10.63128	10.76337	9.97017	9.611874	9.638489158	10.8587188	9.566618	9.504716	7.808898	9.437146
ENSG00000125386	7.197536	6.881646	6.833037	6.568934	6.957605378	7.30220229	7.16873	7.175767	6.273784	7.066012
ENSG00000152082	269.6134	345.7491	289.934	279.912	222.6562495	378.639813	272.8472	283.7854	286.4822	279.6318
ENSG00000133328	0.012462	0.025869	0.031876	0.024356	0.027924318	0.03075731	0.023174	0.028247	0.036857	0.034615
ENSG00000146221	0.004654	0.009661	0.005952	0.012669	0.013173237	0.0076579	0.007933	0.004521	0.004322	0.00404
ENSG00000116885	7.115148	7.369598	8.225007	8.079985	9.605857539	7.18212047	8.252141	7.632474	9.47518	8.426473
ENSG00000076944	7.676848	7.368229	7.054852	7.306716	7.225681357	7.74426906	7.63966	7.679343	7.804194	9.125483

Figure 1: Heat Map of RPKM values of genes selected on the basis of LogFC values obtained by Differential Gene Expression Analysis by EdgeR on Raw Data. Values less than 1 are not significant.

Both up and down regulated genes were put into DAVID for functional analysis. Pathways and ontologies were selected based on enrichment score. Max enrichment score recorded as 1.43 and least enrichment score recorded as 0. Among 82 clusters found by DAVID, 7 had enrichment score greater than 1. Top 5 clusters were selected to study their correlated behavior in term of expressions. Up regulated genes showed more enrichment in dilated cardiomyopathy, 109. chemokine families and Tumor suppressor inhibition of ribosomal biogenesis. Down regulated genes showed more enrichment in pancreatic cancer pathways (Table 4).

Up Regulatory Genes	Downregulatory Genes
Oxidative stress induced gene expression via Nrf2	Dilated Cardiomyopathy
Pancreatic cancer	109.Chemokine Families
	Tumor Suppressor Arf Inhibits Ribosomal Biogenesis,

Table 4: Functional annotation of upregulated and downregulated genes

The genes that were majorly involved in pathways and form their network on STRING was submitted to find interactions between the genes. The network formed had 12 nodes and 6 edges (Figure 2). With clustering coefficient 0.611 and PPI enrichment value 0.00457.

Figure 3 depicts the network formed by our candidate genes by Cytoscape. Candidate genes are shown in black circle. The network contains 57 nodes, including 7 query genes and the 50 most frequently altered neighbor genes. The clusters can be seen in which our genes are involved. TTN, TNNT2, CACNB4 are in one cluster, SMAD2 and TGFB3 are in second cluster and ACTC1 is in third cluster. Arrow color represents the type of interaction. Blue color means "change the state of" and we can see a lot of blue interaction within our candidate genes and neighboring genes. Green color means "control expression of" and we can see here that SMAD2 controls expression of MYC gene

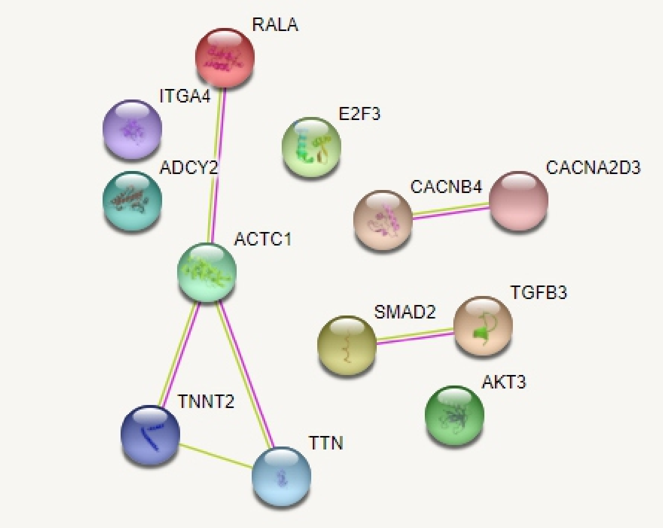


Figure 2: String Network with text mining and experimental interactions selected

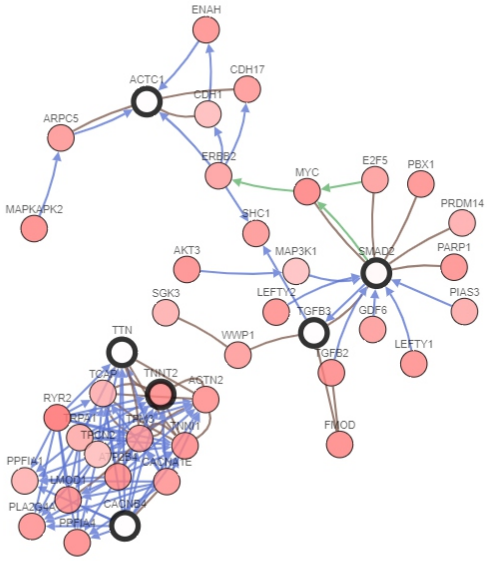


Figure 3: Network formed by candidate genes

Then the genes that were experimentally linked in this STRING network was took and submitted it to cBioPortal to find significance of reported genes. Oncoprint, cancer types summary, survival and network formed by these genes were observed. Oncoprint shows that our candidate genes altered in 587 (23%) of 2509 sequenced cases/patients (2509 total). Most alteration was shown in TNNT2 which is 20% after that RALA shows 1.9% and SMAD2 shows 1% alteration.

DISCUSSION

This study has basically focused on using bioinformatics techniques i.e., statistical computing and tools to analyze dataset. Booms et al., observed risk loci that are active in MCF-7 lines are significant target in breast cancer and their enhancement and inhibition can be of importance in progression of breast cancer [14]. The Breast cancer dataset (GSE5203) used in this study was formed by testing 10 samples of MCF7 breast cancer cells and extracted the wild type Control cDNA reads and mutant AHR reads. This dataset was originally for testing the expression of genes and the comparison of before and after expression. Very few studies have been done on the in-silico analysis of Breast cancer though much has been done at molecular level. Basha et al., devised a novel approach for prediction and analysis of breast cancer dataset. But they make use of random forest algorithm, neural networks and Root Mean Square Error (RMSE) with the accuracy of up to 0.98 [15]. Our depth statistical analysis of the RNASeq Breast Cancer using R packages Limma, RPKM and edgeR, pour a light of different genes' behaviors relative to their differential expression, time, functional significance, cell type and cell condition which in return is a major step contributing to molecular level understanding of gene behavior. This helped in generating a profile for the 10 samples (5-AHR, & 5-Control) of one group of patients. In this study gene expression analysis was performed through R of the above-mentioned dataset. This includes checking quality of data, normalization and find RPKM values to see if significant reads have aligned or not. RPKM values less than 1 means there are not significant hits of reads so we can't consider those genes. After that DEG analysis was performed by edgeR package to find up and down regulatory genes. There were 20930 genes in total. 950 of them was identified as differentially expressed in which 228 were down regulatory and 722 was up regulatory. Top 10 genes were selected that shows most expressional change before and after, based on logFC values. We then move back to our RPKM values to see if these genes are significant or not. We found that some genes have less than 1 RPKM values. We excluded them and got six genes that were, B3GNT9, AC126755.2, FAM193A, MZT2B, OSCP1, STXBP2. These are novel genes and not been discovered yet. These genes with most expression changes may be studied at molecular level to find new insights into breast cancer. Gene ontologies are important in finding out functional annotations of gene datasets and their analysis. These annotations help researchers in extracting meaningful and novel biological information from gene datasets [16]. DAVID is a tool that helps in performing high-throughput gene functional analysis. Functional analysis of regulatory genes obtained was performed by DAVID [17]. Pathways were Dilated cardiomyopathy DCM and pancreatic cancer. The genes involved in these pathways were ACTC1, ADCY2, CACNA2D3, CACNB4, ITGA4, TTN, TGFB3, TNNT2, AKT3, E2F3, RALA, SMAD2. The pathways were selected based on high enrichment score and these genes and pathways play a significant role in cancer. After that network analysis was performed. Functional interactions of protein provide critical information about cellular machinery. Their connectivity helps in understanding full picture of biological process. STRING database collect and integrate all information available on interaction of proteins with other elements and make computational predictions based on that information [18]. Genes majorly involved in pathways was taken and formed a STRING network to find interactions between them, only text mining and experimental interactions were selected. In network we can see that genes RALA, ACTC1, TNNT2, TTN, SMAD2, TGFB3, CACNB4, CACNNA2D3 are experimentally connected. Also, PPI enrichment score of this network shows that our proteins have more interactions among themselves than what would be expected for a random set of proteins of similar size, drawn from the genome. Such an enrichment indicates that the proteins are at least partially biologically connected, as a group. After that we find the network of our candidate genes by Cytoscape and see several significant interactions there. Our candidate genes were involved in changing states of several neighboring genes. The network also shows that SMAD2 controls the expression of MYC. Pathway analysis based on knowledge of biological processes has become important in research as they reduce complexity of grouping thousands of genes into hundreds of pathways for functional analysis. They also explain significance of study by annotating genes and proteins to active pathways in several diseases [19]. Biological processes associated with these genes are muscle filament sliding, secretion, tissue morphogenesis, cell surface receptor signaling pathway and anatomical structure morphogenesis. Only one molecular function is associated with this group of genes that is "type 1 transforming growth factor beta receptor binding". KEGG pathways associated with this group are: Dilated cardiomyopathy, Hypertrophic cardiomyopathy HCM, Adrenergic signaling in cardiomyocytes, pancreatic cancer, and cardiac muscle contraction. Liu et al., used cBioPortal for analysis of genetic alterations in breast cancer and difference in survival of patients with or without genetic mutations. Significant expression, interaction and correlation of genes was found in breast cancer [20]. Candidate genes retrieved from STRING was submitted to cBioPortal to find their significance. Oncoprint, cancer types summary, survival and network formed by these genes. In oncoprint, most alteration is shown in TNNT2 which is 20% after that RALA shows 1.9% and SMAD2 shows 1% alteration. Then we looked on overall survival and it was identified that cases without alterations in query lead to almost six months decrease in survival. Then we looked over at classification of breast cancer and types in which our genes were identified and the types were: Breast Mixed Ductal and Lobular Carcinoma, Breast Invasive Ductal Carcinoma, Phyllodes Tumor of the Breast, Breast Invasive Lobular Carcinoma and Invasive Breast Carcinoma. The pathways were selected based on high enrichment score and these genes and pathways play a significant role in cancer. The findings for this study as per its scope may be laid down as a foundational step for further analysis of the contrasts such as network construction and the functional analysis of newly discovered genes for Breast cancer etc.

CONCLUSIONS

This study has focused on using bioinformatics techniques i.e., statistical computing and tools to analyze a dataset of breast cancer. R software was used for making plots for quality checking, normalizing data, and finding differentially expressed genes. Top 10 genes were selected based on LogFC values, excluded those whose RPKM was less than 1 and six genes were found out that were B3GNT9, AC126755.2, FAM193A, MZT2B, OSCP1, STXBP2. These are novel genes and not been discovered yet. These genes with most expression changes may be studied at molecular level to find new insights into breast cancer. Functional analysis of these regulatory genes by DAVID revealed important pathways. Genes involved in these pathways were ACTC1, ADCY2, CACNA2D3, CACNB4, ITGA4, TTN, TGFB3, TNNT2, AKT3, E2F3, RALA, SMAD2. In network analysis, genes including RALA, ACTC1, TNNT2, TTN, SMAD2, TGFB3, CACNB4, CACNNA2D3 were observed to be significant and are also experimentally connected. Also, PPI enrichment score of this network shows that this group is biologically connected at least at some level. The pathways were selected based on high enrichment score and these genes and pathways play a significant role in cancer. The findings for this study as per its scope may be laid down as a foundational step for further analysis of the contrasts such as network construction and the functional analysis of newly discovered genes for Breast cancer etc

Conflicts of Interest

The authors declare no conflict of interest.

Source of Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Akram M and Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. Biological Research. 2017 Oct; 50(1):33. doi: 10.1186/s40659-017-0140-9

[2] Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. CA a Cancer Journal for Clinicians. 2008 Mar-Apr; 58(2):71-96. doi: 10.3322/CA.2007.0010

[3] Chen, W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA: A Cancer Journal for Clinicians. 2016 Mar; 66:115-132. doi: 10.3322/caac.

[4] Bhurgri Y, Bhurgri A, Hassan SH, Zaidi S, Rahim A, Sankaranarayanan R, et al. Cancer incidence in Karachi, Pakistan: First results from Karachi Cancer Registry. International Journal of Cancer. 2000 Feb; 85:325-329. doi: 10.1002/(sici)1097-0215(20000201)85:3<325::aid-ijc5>3.0.co;2-j.

[5] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Molecular Systems Biology. 2007 oct; 3:140. doi: 10.1038/msb4100180.

[6] Weigelt B, Peterse JL, Veer LJ. Breast cancer metastasis: markers and models. Nature Reviews Cancer. 2005 Aug; 5(8):591-602. doi: 10.1038/nrc1670.

[7] Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk Factors and Preventions of Breast Cancer. International Journal of Biological Sciences. 2017 Nov; 13(11):1387-1397. doi: 10.7150/ijbs.21635.

[8] Timothy JK, Pia KV, Banks E. Epidemiology of breast cancer. The lancet oncology, 2001 Mar; 2(3):63-140. doi: 10.1016/S1470-2045(00)00254-0.

[9] Tippmann S. Programming tools: Adventures with R. Nature. 2015 Jan; 517(7532):109-10. doi: 10.1038/517109a.

[10] Huang DW, Sherman BT, Tan Q, Kir J, Collins JR, Alvord WG, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Research. 2007 Jul; 35:W169-75. doi: 10.1186/gb-2007-8-9-r183

[11] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Research. 2003 Nov;

13(11):2498–2504. doi: 10.1101/gr.1239303.

[12] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nature Protocols. 2007 Jun; 2(10):2366–2382. doi: 10.1038/nprot.2007.324.

[13] Magoc T, Wood D, Salzberg' SL. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. Evolutionary Bioinformatics. 2013 Mar; 9:127-36. doi: 10.4137/EBO.S11250.

[14] Booms A, Coetzee GA, Pierce SE. MCF-7 as a Model for Functional Analysis of Breast Cancer Risk Variants. Cancer Epidemiology Biomarkers and Prevention. 2019 Oct; 28(10):1735-1745. doi: 10.1158/1055-9965.EPI-19-0066.

[15] Basha SM, Rajput D, Iyengar NC, Caytiles RD. A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R. International Journal of Grid and Distributed Computing. 2018 Feb; 11(2):41–54. doi: 10.14257/ijgdc.2018.11.2.05

[16] Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biology. 2004 Nov; 5(12):1-8. doi: 10.1186/gb-2004-5-12-r101

[17] Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, Liu D, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007 Nov; 8:426. doi: 10.1186/1471-2105-8-426

[18] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research. 2021 Jan; 49(D1): D605-12. doi: 10.1093/nar/gkaa1074.

[19] Du J, Li M, Yuan Z, Guo M, Song J, Xie X, et al. A decision analysis model for KEGG pathway analysis. BMC Bioinformatics. 2016 Oct; 17:407. doi: 10.1186/s12859-016-1285-1

[20] Liu X, Liu Y. Comprehensive Analysis of the Expression and Prognostic Significance of the CENP Family in Breast Cancer. International Journal of General Medicine. 2022 Mar; 15:3471-3482. doi: 10.2147/IJGM.S354200.