Genes & Diseases (2019) 6, 78–87

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://ees.elsevier.com/gendis/default.asp

FULL LENGTH ARTICLE

# Integrated transcriptome interactome study of oncogenes and tumor suppressor genes in breast cancer

**G. Pranavathiyani, Raja Rajeswary Thanmalagan, Naorem Leimarembi Devi, Amouda Venkatesan***

*Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry 605014, India*

Received 28 May 2018; accepted 31 October 2018
Available online 20 November 2018

**KEYWORDS**
Breast cancer;
Differential gene expression;
Network analysis;
Oncogenes;
Tumor suppressor genes

**Abstract** Breast cancer is the leading cause for mortality among women worldwide. Dysregulation of oncogenes and tumor suppressor genes is the major reason for the cause of cancer. Understanding these genes will provide clues and insights about their regulatory mechanism and their interplay in cancer. In the present study, an attempt is made to compare the functional characteristics and interactions of oncogenes and tumor suppressor genes to understand their biological role. 431 breast cancer samples from seven publicly available microarray datasets were collected and analysed using GEO2R tool. The identified 416 differentially expressed genes were classified into five gene sets as oncogenes (OG), tumor suppressor genes (TSG), druggable genes, essential genes and other genes. The gene sets were subjected to various analysis such as enrichment analysis (*viz.*, GO, Pathways, Diseases and Drugs), network analysis, calculation of mutation frequencies and Guanine-Cytosine (GC) content. From the results, it was observed that the OG were having high GC content as well as high interactions than TSG. Moreover, the OG are found to have frequent mutations than TSG. The enrichment analysis results suggest that the oncogenes are involved in positive regulation of cellular protein metabolic process, macromolecule biosynthetic process and majorly in cell cycle and focal adhesion pathway in cancer. It was also found that these oncogenes are involved in other diseases such as skin diseases and viral infections. Collagenase, paclitaxel and docetaxel are some of the drugs found to be enriched for oncogenes.
Copyright © 2018, Chongqing Medical University. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

* Corresponding author. Centre for Bioinformatics, School of Life Sciences, Pondicherry University, R. V. Nagar, Kalapet, Pondicherry 605014, India.
*E-mail addresses:* pranavathiyani@gmail.com (G. Pranavathiyani), rajitbio@gmail.com (R.R. Thanmalagan), leimarembi@gmail.com (N. Leimarembi Devi), amouda@yahoo.com (A. Venkatesan).
Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2018.10.004
2352-3042/Copyright © 2018, Chongqing Medical University. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## Introduction

Breast cancer is the most common cancer among women worldwide and about two million cases and six lakh deaths were reported this year by the World Health Organization (WHO).[1,2] Nearly ninety percentage of the women diagnosed with breast cancer at the earliest stage can survive up to several years, compared to women who are diagnosed at later stages.[3] In this scenario, there is a requirement for the awareness among women and need for the early detection and diagnosis. Understanding cancer related genes could provide better insights of the pathogenesis and lead to the identification of targets for early detection. Breast cancer progression involves various genetic events — by activating the oncogenes or disrupting the regular functions of specific tumor suppressor genes.[4] Over expression and amplification of oncogenes often assist in the development of cancer and cooperatively regulate genetic and epigenetic changes. A number of studies have reported that the oncogenes ErbB2, PI3KCA, MYC and CCND1 are often found to be deregulated in breast cancer. Among the oncogenes, HER-2 activation is found in about 20% of all primary breast cancer cases.[4,5] On the other hand, tumor suppressor genes are the negative regulators

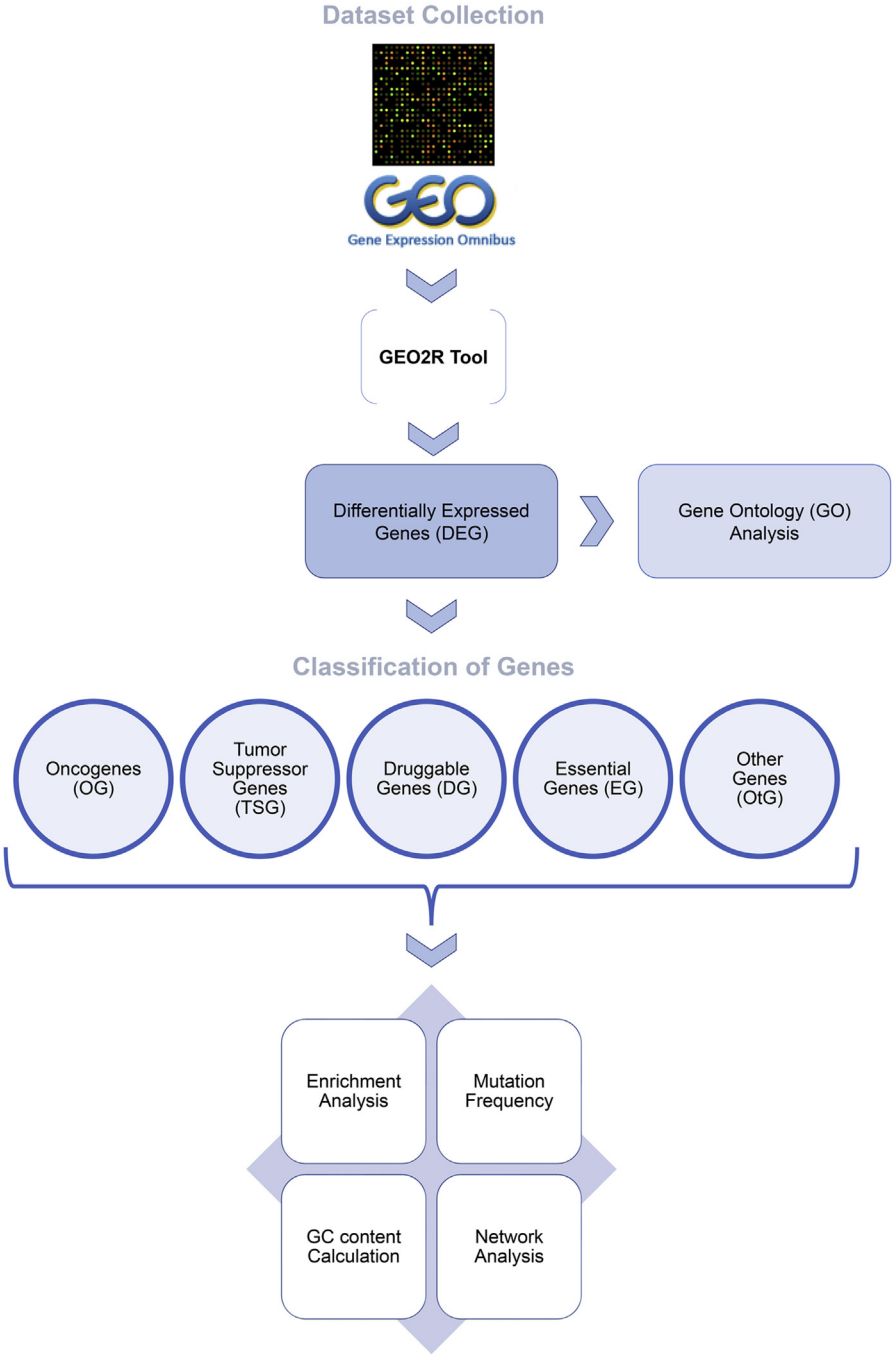


**Figure 1** Overview of methodology adopted in the present study.

**Table 1** Gene expression omnibus (GEO) datasets used in the present study with number of samples, platform information along with the number of identified differentially expressed genes (DEG).

| Sl. No. | Dataset accession | No. of samples | Platform | No. of DEG | |
|---|---|---|---|---|---|
| | | | | Upregulated genes | Downregulated genes |
| 1 | GSE45584 | 90 | GPL6480 | 98 | 124 |
| 2 | GSE45581 | 45 | GPL6480 | 707 | 1279 |
| 3 | GSE21422 | 19 | GPL570 | 1020 | 1155 |
| 4 | GSE6883 | 24 | GPL96 | 606 | 495 |
| 5 | GSE79058 | 76 | GPL19956 | 39 | 35 |
| 6 | GSE45827 | 155 | GPL570 | 2917 | 1118 |
| 7 | GSE1299 | 22 | GOL96 | 212 | 346 |

of cellular progression and growth which regulate the invasiveness and metastatic potential. Loss of function of these genes often lead to malignancy. Till date, there are several tumor suppressor genes reported to be associated with breast cancer - BRCA1, BRCA2, P53, PTEN, ATM and CHK2 are some among those.[6] Recent advancements in high throughput technologies like DNA microarrays, next generation sequencing had produced massive data which are analysed with modern approach to gain better insights and also to identify novel targets.[7,8] In the present study, various bioinformatics analysis has been carried out in order to understand mechanism and interaction of oncogenes and tumor suppressor genes in breast cancer. With available breast cancer microarray data, differentially expressed genes (DEG) were identified and classified into five gene sets. Further these gene sets were subjected to various *in silico* analysis to understand their characteristic role and mechanism based on evidences from publicly available experimental data. From the findings, it was observed that the oncogenes were having high mutation frequency rate and also enriched with guanine-cytosine content than the tumor suppressor genes. It was also found that the oncogenes are highly interconnected and the genes CDK1, FOS, CCNA2, MMP9, CDH1, CCNB1 and TOP2A were identified to be hubs. This exploration of oncogenes and tumor suppressor genes in breast cancer could aid cancer biology research for early diagnosis and treatment options.

## Materials and methods

### Dataset collection

Breast cancer microarray datasets were collected from Gene Expression Omnibus (GEO), a public functional genomics data repository of NCBI.[9] The criteria for selection of dataset is that, it must have samples of both healthy and breast cancer tissue with no drug treatment or any other illness. Comparison of normal versus cancer tissues will help in identification of genes that are deregulated.

### Identification of differentially expressed genes

The collected microarray datasets[10–15] were analysed individually by comparing as groups of breast cancer tissue versus healthy tissue as controls using GEO2R, a web-based tool for gene expression analysis. The tool is based on R packages GEOquery and limma for calculation of *p*-value, logFC, adjusted *p*-value, t-statistic and B-statistic. Differentially expressed genes (DEG) were filtered with the cut-off of $|\log FC| \geq \pm 2$ with $p$-value $< 0.05$ from each dataset. The genes which are found to be differentially expressed in more than one dataset was considered as breast cancer associated genes. The identified differentially expressed breast cancer genes were validated with available genes in the International Cancer Genome Consortium (ICGC) data portal.[16]

### Enrichment analysis

The enrichment analysis of the identified breast cancer associated genes were performed using WEB-based GEne SeT AnaLysis Toolkit (WebGestalt), an online software toolkit comprising information from various public resources for biological analysis.[17] The enrichment analysis such as Gene Ontology (GO), pathways, diseases and drugs were carried out with top 10 results as significant using hypergeometric test and Benjamini & Hochberg method.

### Gene set classification & analysis

The identified differentially expressed breast cancer genes were classified into five gene sets namely, oncogenes (OG), tumor suppressor genes (TSG), druggable genes (DG), essential genes (EG) and other genes (OtG). The classification of the gene sets was performed based on the mapping of DEG with various databases and resources. For OG, the collection of all oncogenes from Bushman lab (http://www.bushmanlab.org/links/genelists) was used and TSGene (https://bioinfo.uth.edu/TSGene/) database, a web resource for tumor suppressor genes was considered for TSG classification.[18] Cancer Gene Census list from the Catalogue Of Somatic Mutations In Cancer - COSMIC (https://cancer.sanger.ac.uk/cosmic) was also used for the classification of TSG and OG. The Drug Gene Interaction database (DGIdb), an online database (http://www.dgidb.org/) of drug–gene interactions and druggable genome data was used for the classification of druggable genes.[19] The essential genes of humans were collected from the database of essential genes (http://www.essentialgene.org/) and mapped to identify essential genes from DEG.[20] The genes which did not map to any of these resources were considered as other genes (OtG). The classified five

gene sets were subjected to enrichment analysis (KEGG pathways, GO, diseases and drugs) using WebGestalt to understand their biological role and properties.

### GC content and mutation frequency

The Guanine-Cytosine (GC) content percentage for each of the five classified gene sets were calculated using bioMart tool (https://www.ensembl.org/biomart) from Ensembl.[21] GC content is one of the fundamental features in a genome which is widely studied for methylation profiles and binding of transcription factors in regulating gene expression. The mutation frequency of each set of genes were calculated from The Cancer Genome Atlas (TCGA) breast cancer data.[22] The mutation frequency is the ratio of samples where the gene is mutated among the entire samples sequenced for a gene.

### Protein–protein interaction and cluster analysis

The protein–protein interaction (PPI) of the DEG was constructed using STRING (https://string-db.org/) database, an online biological database for known and predicted protein–protein interactions.[23] The network of interacting proteins was downloaded and visualized using Cytoscape v3.5.1, an open source software tool for visualizing molecular interactions.[24] The top 10 modules of highly interacting gene clusters among the DEG were found using MCODE plugin with default parameters.[25] For the classified five gene sets, the protein–protein interaction network was constructed and the network topological parameters such as degree, betweenness centrality, shortest path and closeness centrality were calculated using NetworkAnalyzer in Cytoscape.

## Results

### Identification of differentially expressed genes (DEG)

Oncogenes and Tumor suppressor genes are the major key players in the dysregulation of cancer pathways. To study their properties and biological mechanism of action in breast cancer, seven microarray datasets were analysed and the identified differentially expressed genes were subjected to various bioinformatics analysis. The methodology adopted in the present study is depicted in Fig. 1. The detailed information about the collected seven microarray datasets is given in Table 1 along with the number of identified up and down regulated genes. The gene expression analysis results of the seven microarray datasets with up and down regulated genes is provided as supplementary data.

### Enrichment analysis and classification of DEG

A total of 416 DEG were identified and subjected to enrichment analysis and the results are given in Fig. 2 and Table 2. From ICGC data portal it was observed that 99% of the genes were matched to breast cancer genes. The enrichment results include enriched GO, pathways, diseases and drugs with the number of genes involved. The GO enrichment result of the DEG is summarised in Fig. 2 which shows cell cycle, cell division, mitosis, cell cycle phase are the major biological processes these genes are involved and the molecular functions includes, protein binding, extracellular matrix structural constituent, ATP binding and microtubule motor activity. The DEG cellular component were mostly found to be in nucleus, extracellular matrix, kinetochore and chromosome. The overview of GO analysis results for the DEG with adjusted *p*-value is given in supplementary. The KEGG pathway enrichment results indicated that the DEG are involved in pathways such as cell cycle, ECM receptor interaction, focal adhesion, p53 signalling and small cell lung cancer pathway. The enriched disease terms for the DEG were cancer, neoplasm, breast diseases, carcinoma, adenocarcinoma, breast neoplasm and collagenase, paclitaxel, heparin, urokinase, progesterone, epirubicin, doxorubicin, alteplase, podofilox, zidovudine are the top 10 drugs found in drug enrichment analysis. The DEG were classified into five gene sets as 102 oncogenes (OG), 41 tumor suppressor genes (TSG), 77 druggable genes (DG), 243 essential genes (EG) and 112 other genes (OtG) with the help of existing resources discussed in the methods section. Some of the genes were

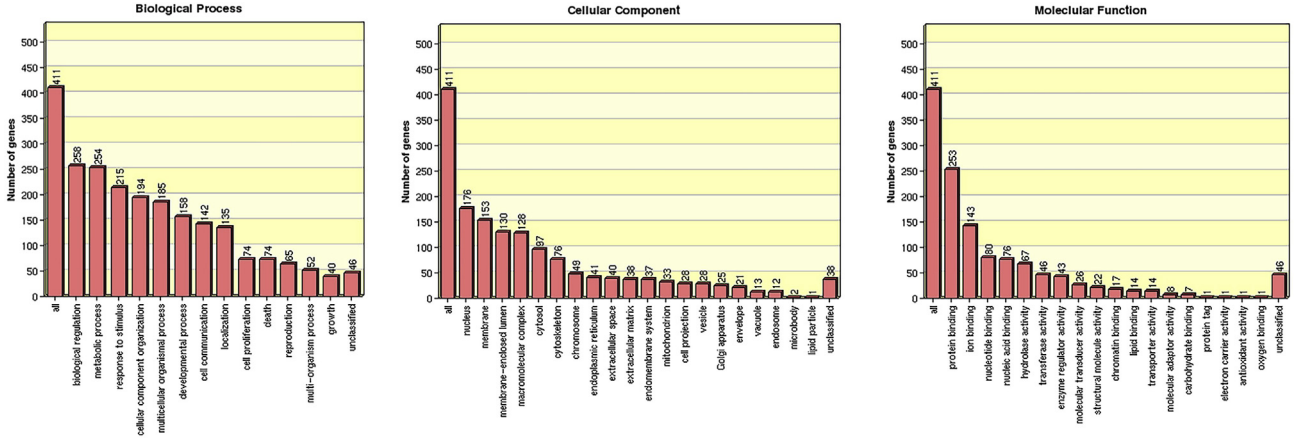


**Figure 2** Comprehensive bar chart of gene ontology (GO) results for differentially expressed genes showing number of genes in various biological process, cellular components and molecular function.

**Table 2** Enrichment analysis result of pathways, diseases and drugs for the identified differentially expressed genes.

| KEGG Pathway Enrichment | | Disease Enrichment | | Drug Enrichment | |
|---|---|---|---|---|---|
| Pathways | No. of Genes | Diseases | No. of Genes | Drugs | No. of Genes |
| Cell cycle | 27 | Cancer or viral infections | 86 | Collagenase | 26 |
| ECM-receptor interaction | 20 | Neoplasms | 72 | Paclitaxel | 15 |
| Pathways in cancer | 28 | Breast Diseases | 50 | Heparin | 17 |
| Focal adhesion | 23 | Breast Neoplasms | 51 | Urokinase | 11 |
| p53 signalling pathway | 12 | Neoplastic Processes | 52 | Progesterone | 12 |
| Small cell lung cancer | 12 | Carcinoma | 55 | Epirubicin | 7 |
| Amoebiasis | 13 | Neoplasm Invasiveness | 40 | Doxorubicin | 10 |
| Progesterone-mediated oocyte maturation | 11 | Neoplasm Metastasis | 40 | Alteplase | 9 |
| Toll-like receptor signalling pathway | 11 | Skin and Connective Tissue Diseases | 45 | Podofilox | 9 |
| Bladder cancer | 8 | Adenocarcinoma | 40 | Zidovudine | 6 |

found to be overlapped in multiple gene sets, the classified gene set list is given in supplementary. Further the classified five gene sets were subjected for enrichment analysis to understand their properties and role in cancer. It was not surprising that the same pathways, diseases and drugs of DEG were enriched for the gene sets as the genes had some overlaps. The enrichment results for the gene sets are given in supplementary.

## GC content and mutation frequency

The GC content percentages of all the classified gene sets were imported using bioMart tool and the average was considered for this study. The list of all genes and their GC percentages are given as supplementary. From the average percentage of all gene sets, it was observed that the druggable genes (DG) were having the highest GC content of 45.51% followed by oncogenes (OG) with 44.66%. The essential genes (EG) and tumor suppressor genes (TSG) were seen to have more or less similar GC content of 44.27% and 44% respectively. On comparison, the TSG is having genes with less GC content than the OG. The probability of change of a unit length of DNA with time is referred as rate of mutation. Of the five gene sets, OG was found to have the highest mutation frequency rate of about 1.19% and DG with 0.928%. The TSG and EG sets were having 0.746% and 0.753% average mutation frequency rates. The comparison of GC content and mutation frequency percentage for the five gene sets is given in Fig. 3.

## Analysis of PPI network and clusters

The protein—protein interaction (PPI) network of DEG was retrieved and analysed by considering proteins as nodes and their interactions as edges. The constructed PPI network comprises 412 nodes and 2628 edges interactions based on experiments, co-expression, text mining, neighbourhood, gene fusion and databases. To understand the interplay, network analysis was performed by calculating the network topological parameters such as degree, betweenness centrality, closeness centrality and shortest path distance. Each of these parameters determine the role and property of gene/protein in the network. The average degree of the DEG network was 12.8, showing each gene on average has about 12.8 interactions with others in the network and the average clustering coefficient is 0.605. To know the highly interconnected genes in the network, the DEG were subjected to cluster analysis and the top 10 modules of clusters

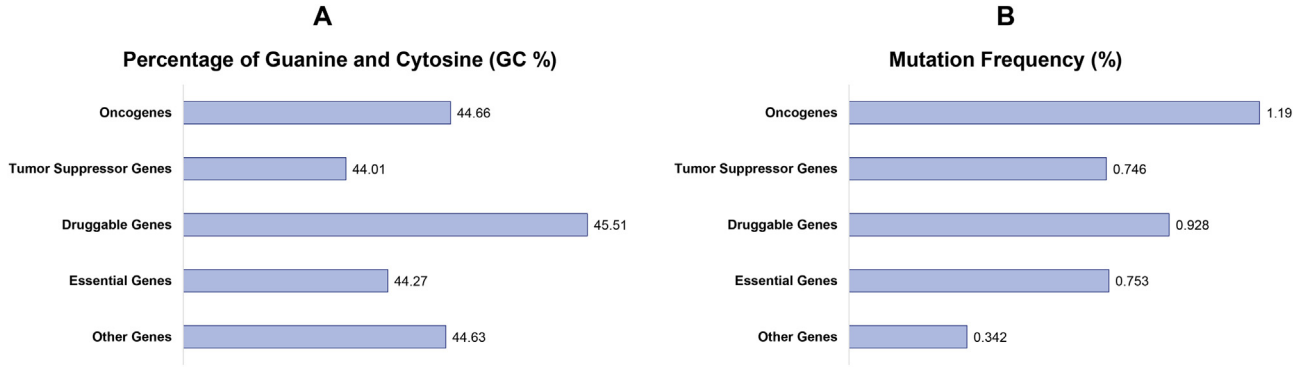


**Figure 3** Comparison of guanine-cytosine (GC) content percentage (A) and mutation frequency percentage (B) of the five gene sets.

**Table 3** Top 10 clusters of highly interconnected genes among the differentially expressed genes.

| Cluster | Score (Density × No. of Nodes) | No. of Nodes | No. of Edges | Node IDs |
|---|---|---|---|---|
| 1 | 29.444 | 37 | 530 | DLGAP5, AURKB, ECT2, CDCA5, RAD21, CASC5, ZWINT, KIF23, RACGAP1, CCNB2, MLF1IP, AURKA, TOP2A, CENPA, CDK1, NCAPG, NDC80, NEK2, KIF11, KIF4A, BUB1, PRC1, CDCA8, CCNA2, CDC20, BIRC5, BUB1B, MAD2L1, ZWILCH, CKAP5, KIF18A, KIF2C, CENPE, CENPF, CENPK, KIF20A, CCNB1 |
| 2 | 11 | 13 | 66 | ITGA6, COL11A1, ITGB6, COL5A2, COMP, COL1A2, COL12A1, COL5A1, COL4A6, ITGB4, COL10A1, COL3A1, ITGB1 |
| 3 | 6.1 | 21 | 61 | FOXM1, ANLN, MCM4, PLK4, MELK, CDC45, CEP55, NUF2, CDC6, PCNA, CKS2, MCM2, TACC3, TTK, CCNE2, KIFC1, UBE2C, SMC4, PBK, RRM2, NUSAP1 |
| 4 | 6.08 | 26 | 76 | POSTN, COL1A1, FPR3, RGS1, EZH2, MMP1, MYB, MYBL1, CXCR4, PIK3CA, TBL1XR1, THBS1, LAMA3, SPP1, LAMB3, CXCL11, CCR5, SERPINE1, CCL5, LAMC2, SDC1, FN1, CXCL9, TIMP3, RGS20, CXCL10 |
| 5 | 5.417 | 25 | 65 | VEGFA, EIF5A, HMOX1, OAS2, CUL2, SQLE, RET, CTSS, IFI30, OASL, FGFR3, SRGN, MMP9, IFI6, OAS3, HAPLN1, VCAN, MAPK13, IRF6, FOS, MMP11, PLAUR, MMP3, MAX, STAT1 |
| 6 | 4 | 4 | 6 | FEN1, EXO1, TRIP13, RAD51 |
| 7 | 4 | 4 | 6 | HIST1H2BD, HIST1H3H, HIST1H2BH, HIST1H2BK |
| 8 | 4 | 4 | 6 | KYNU, KMO, QPRT, TDO2 |
| 9 | 3.333 | 4 | 5 | DBF4, CDC7, CCNG2, CHEK1 |
| 10 | 3 | 3 | 3 | ISG15, RSAD2, DDX58 |

were taken into consideration, which is given in Table 3. The highest interacting cluster was found to have 37 nodes with 530 edges and the density score is 29.44. DLGAP5, AURKB, ECT2, CDCA5, RAD21, CASC5, ZWINT, KIF23, RACGAP1, CCNB2, MLF1IP, AURKA, TOP2A, CENPA, CDK1, NCAPG, NDC80, NEK2, KIF11, KIF4A, BUB1, PRC1, CDCA8, CCNA2, CDC20, BIRC5, BUB1B, MAD2L1, ZWILCH, CKAP5, KIF18A, KIF2C, CENPE, CENPF, CENPK, KIF20A, CCNB1 are the genes found in the top cluster with high interaction. It was observed that genes are majorly involved pathways related to cancer, cell cycle and p53 signalling. Furthermore, the PPI network for each of five classified gene sets were constructed and the network topological measures were calculated for the comparison of their network properties. The calculated network topologies of the five gene sets is provided in Table 4.

## Discussion

The exploration of relationship between OG, TSG and DG in breast cancer can provide basic understanding of the genetic and functional association of genes, which could aid in identification of disease markers. Breast cancer is a heterogeneous disease comprising various molecular interactions among different cell types.[26] These molecular and cellular interactions are the major drive for the expression of cancer related genes and progression of the disease. Recent explosion of biological big data with high throughput technologies has evolved the understanding of tumor progression in cancer and mere understanding the role of cancer genes is inexpedient, stressing the need for integrated analysis.[26,27] The present study was aimed to understand the integrated network of interactions among the differentially expressed cancer related genes, especially in focus to oncogenes and tumor suppressor genes. For the integrated approach, various bioinformatics analysis such as GO, pathways, disease, drugs and networks were utilised. From the gene expression profiles, the oncogenes were found to be majorly deregulated along with the tumor suppressor genes. Further, expression analysis of the identified differentially expressed OG and TSG in TCGA portal for breast cancer showed that only a certain number of genes were found to have significant variations/mutations related to breast cancer. The enrichment analysis of these genes gave an overview of possible dysregulated pathways and associated diseases according to the genes.

**Table 4** Calculated average network topological parameters for the five gene sets. OG: Oncogenes; TSG: Tumor Suppressor Genes; DG: Druggable Genes; EG: Essential Genes; OtG: Other Genes.

| Network topological parameters | OG | TSG | DG | EG | OtG |
|---|---|---|---|---|---|
| Degree | 11.1364 | 4.5334 | 7.1045 | 17.5 | 1.7074 |
| Average Shortest Path Length | 2.3984 | 2.4276 | 2.8151 | 2.6036 | 1.3701 |
| Betweenness Centrality | 0.0163 | 0.051 | 0.028 | 0.0077 | 0.1056 |
| Closeness Centrality | 0.4291 | 0.4272 | 0.3689 | 0.3963 | 0.798 |
| Clustering Coefficient | 0.444 | 0.2828 | 0.4779 | 0.4716 | 0.2651 |

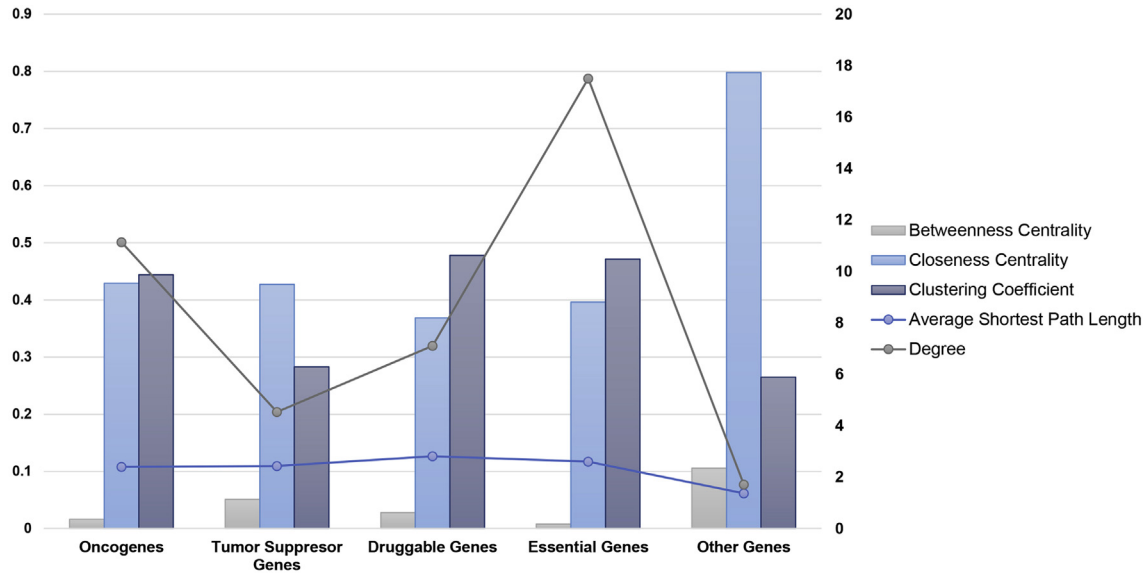


**Figure 4** Comparison of various network topological parameters of five sets of genes, where oncogenes are showing high centrality measures compared to tumor suppressor genes.

The role of OG and TSG in cancer and cell growth is mediated by various other factors, computing the GC content percentage for these genes will extend knowledge about the physicochemical properties and expression pattern. In mammalian cells, the genes with high GC content have several fold high expression levels compared to GC poor genes.[28] From the study it was observed that the OG were having higher GC content percentage than TSG, suggesting that OG might be evolutionarily conserved and are mediated by methylation of CpG islands in the genes. Genome wide study of GC content and methylation profiles gave evidences that expression level and methylation are correlated stating, high GC rich genes have possibilities of high methylation and suppression in gene expression.[29] The constant mutation rate of a genome is about 10-10/bp per generation.[30] In a cellular environment, several physiological and pathophysiological conditions can alter the rate of mutation drastically. Cancer cells are often believed to have high mutation rate, leading to resistance mechanism and progression. With the help of TCGA data, the mutation rates of the breast cancer associated genes were calculated, which clearly shows that OG have higher mutation frequency rate, compared to the other gene sets. The genes PIK3CA, GATA3 and CDHI were the top genes with high mutation frequency rate. PIK3CA was reported as one of the reasons for cervical cancer and oncogenic in nature.[31] GATA3 is a gene which encodes transcription factor binding proteins belonging to GATA family, involved in T-cell development, endothelial cell biology and also been reported in breast cancer.[32]

Proteins encoded by genes often interact physically/structurally to perform their functions, these interactions between proteins are referred as protein—protein interaction (PPI). By considering proteins as nodes and the interactions among them as edges, PPI networks can be studied with the help of network topological parameters to identify key proteins.[33,34] The present study was designed to understand the interplay of interactions among the gene sets and in particular to compare the network parameters of OG and TSG. Interestingly, these two gene sets had no difference in measures of their network properties except degree centrality. It must be noted that the classified number of OG and TSG from the DEG is 102 and 41 respectively. Degree centrality of a node is an important parameter which is measured by the number of interactions with other nodes in a network. In case of OG, the average degree was found to be 11.13 which is significantly higher than the degree of TSG, which is 4.53. On the other hand, the essential genes were having highest average degree of 17.5 and druggable genes with 7.10, which clearly indicates that the essential genes are major players in a network and are indispensable for an organism. The betweenness centrality is a measure to find number of shortest paths passing through a particular node, which possibly act as a intermediate for exchange of information. The TSG are found to have slightly higher betweenness centrality than OG, indicating their intermediate level in process like activation. Clustering coefficient (CC) of a node represents the interconnectivity to form triangular sub clusters in a network. The gene sets OG, EG and DG were found to have almost similar clustering coefficient of about 0.4, which shows their interconnectivity in the network is on average, yet a complete picture of network might provide more reasonable explanation. The CC of TSG and OtG were about 0.2, which is comparatively less than that of OG indicating the relatedness in a network. The path which costs least number of edges to reach from one node to other is called as shortest path in a network. On average, OG and TSG has the shortest path distance to other nodes of 2.39 and 2.42 respectively meaning they interact more closely and can be traversed to other genes in the network. Druggable genes (DG) are vital and can act as therapeutic target for a disease. The average degree of interaction of DG is 7.10 and with clustering coefficient 0.47 which clearly shows that these genes are significantly interacting to genes in the network to a certain extent (Fig. 4). From the DEG interactions, a subnetwork of OG and TSG was constructed to

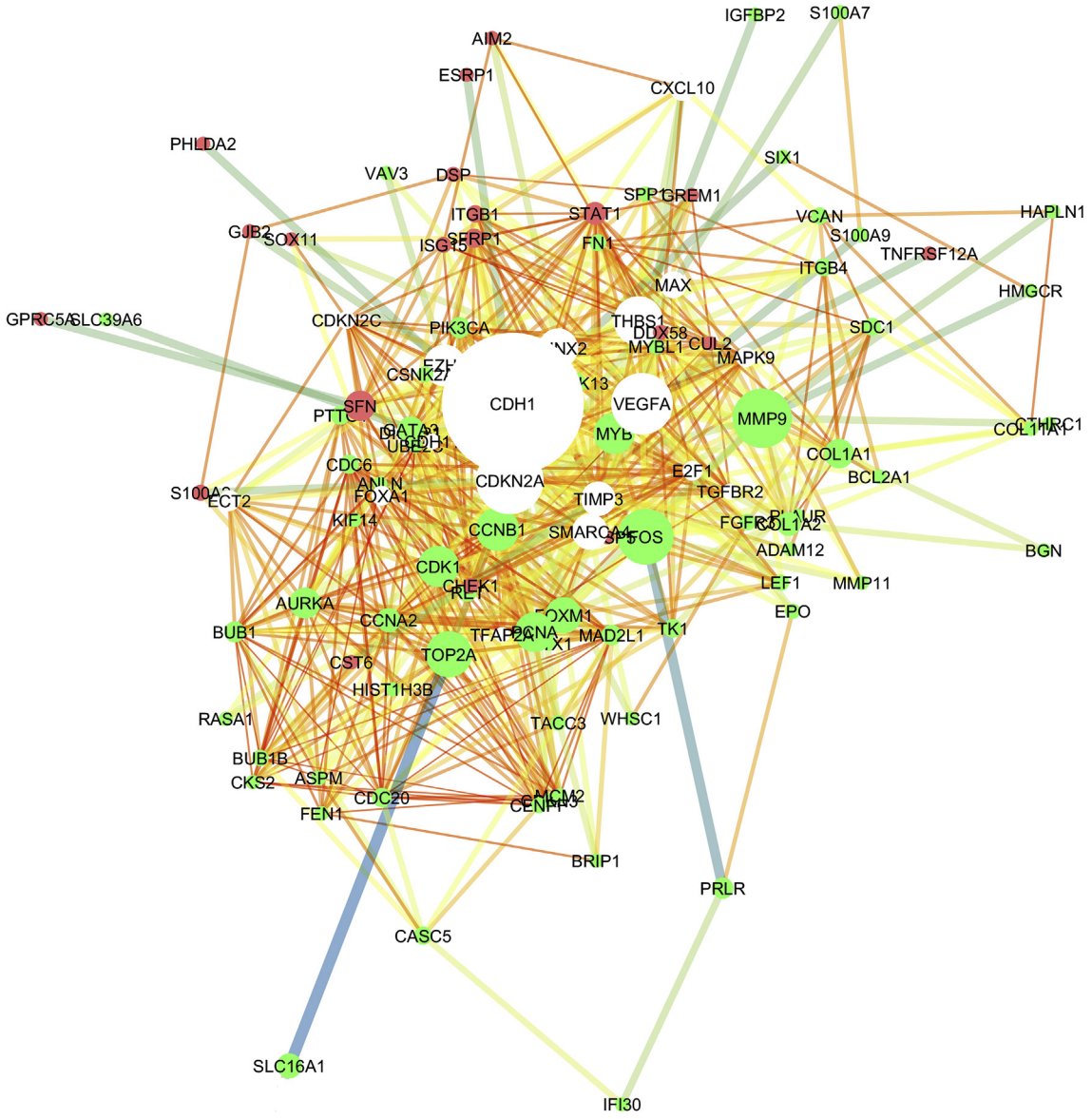


**Figure 5** Interaction network of oncogenes and tumor suppressor genes, where high degree nodes are represented with larger node size. The nodes coloured in green and red are oncogenes and tumor suppressor genes respectively.

look at the interplay in particular, which is depicted in Fig. 5 Analysing of this network showed that the tumor suppressor genes CDH1, CDKN2A and the oncogenes CDK1, CCNB1, FOS, TOP2A and AURKA were having high degree of interaction.

The TSG genes namely CDH1, CDKN2A, CUL2, E2F1, ITGB1, STAT1 and the oncogenes CDK1, E2F1, FGFR3, FN1, FOS, LEF1, MAX, MMP9, RET, TGFBR2 and VEGFA are reported to be involved in many of the major cancer pathways like WNT signalling pathway, focal adhesion pathway and p53 signalling pathway.[35–37] CDH1, a calcium dependant cell adhesion protein is involved in regulation of β-catenin, loss of this protein plays a critical role in cadherin-based adhesion and it also act as a co-activator of WNT signalling pathway to initiate differentiation process.[38] From the observations, it was noticed that the genes CDH1 and LEF1 have remote interactions suggesting that the lymphocyte enhance factor (LEF1) or the T-Cell factor with cadherin free β-catenin complex act as control to facilitate WNT mediated gene expression to promote chromatin remodelling, transcription initiation and elongation to evade apoptosis. The cell matrix proteins FN1, ITGB1 and FGFR3 are involved in focal adhesion and cytokine–cytokine interactions which were identified to be differentially expressed in breast cancer. FN1, an ECM glycoprotein is known to be a key regulator of breast cancer cell adhesion and migration by binding to interleukin in focal adhesion pathway.[39] The integrin subunit β1 (ITGB1) and FN1 are facilitating in the focal adhesion pathway and these genes mediate the ECM interaction by dysregulating the focal adhesion pathway which is observed in transition from ductal carcinoma in situ (DCIS) to invasive breast cancer.[40] The oncogene COL11A1, a collagen type XI alpha 1 is overexpressed in most of the samples in the study and is

also known to promote tumor aggressiveness.[41] Cyclin dependant kinase inhibitor 2A protein (CDKN2A) regulated the tumor suppression of p53 signalling pathway and loss of p16 decrease the ability to repair DNA damage.[42] It was also reported that the thrombospondin-1 (THBS1) is a negative regulator of new blood vessel formation and metastasis. SFN and CDK1 are noticed to downstream target genes for G2 phase arrest in cell cycle. Stratifin (SFN 14-3-3 ς) binds to CDK2 and CDK4 to arrest cell cycle in eukaryotes [43]and inactivation of SFN lead to carcinogenesis through p53 signalling.[42]

In conclusion, the present study reported insights on oncogenes and tumor suppressor genes in breast cancer using an integrated gene expression and interactome analysis. From the observations it was noticed that the dysregulated genes in breast cancer are involved in cell cycle, neoplasm related pathways and majorly involved in protein binding, nucleotide binding and regulating cellular functions. Oncogenes were found to have high GC content, mutation frequency and interactions when compared to the tumor suppressor genes. This study explored several key genes which are majorly involved in cancer progression and development, providing complementary clues for further experimental studies to validate and identify potential biomarkers.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Compliance with ethical standards

No ethical approval was required for this study.

## Consent for publication and competing interests

All of the authors declare that they have no competing interests. All authors have read and approved the final manuscript.

## Acknowledgement

Authors thank Centre for Bioinformatics, Pondicherry University for providing computational facility to carry out this work. Leimarembi Devi Naorem acknowledges a Senior Research Fellowship from the Council of Scientific & Industrial Research (CSIR).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2018.10.004.

## References

1. Anders CK, Johnson R, Litton J, Phillips M, Bleyer A. Breast cancer before age 40 years. *Semin Oncol*. 2009;36(3):237—249.
2. WHO. *Breast Cancer*; September 2018. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.
3. Caplan L. Delay in breast cancer: implications for stage at diagnosis and survival. *Front Public Health*. 2014;2:87.
4. Ingvarsson S. Molecular genetics of breast cancer progression. *Semin Cancer Biol*. 1999;9(4):277—288.
5. Lee EYHP, Muller WJ. Oncogenes and tumor suppressor genes. *Cold Spring Harb Perspect Biol*. 2010;2(10). a003236-a003236.
6. Osborne C, Wilson P, Tripathy D. Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications. *Oncologist*. 2004;9(4):361—377.
7. Oliver DJ, Nikolau B, Wurtele ES. Functional genomics: high-throughput mRNA, protein, and metabolite analyses. *Metab Eng*. 2002;4(1):98—106.
8. Jayapal M, Melendez AJ. DNA microarray technology for target identification and validation. *Clin Exp Pharmacol Physiol*. 2006;33(5-6):496—503.
9. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207—210.
10. Woodward WA, Krishnamurthy S, Yamauchi H, et al. Genomic and expression analysis of microdissected inflammatory breast cancer. *Breast Cancer Res Treat*. 2013;138(3):761—772.
11. Kretschmer C, Sterner-Kock A, Siedentopf F, Schoenegg W, Schlag PM, Kemmner W. Identification of early molecular markers for breast cancer. *Mol Cancer*. 2011;10(1):15.
12. Liu R, Wang X, Chen GY, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N Engl J Med*. 2007;356(3):217—226.
13. Norton N, Advani PP, Serie DJ, et al. Assessment of tumor heterogeneity, as evidenced by gene expression profiles, pathway activation, and gene copy number, in patients with multifocal invasive lobular breast tumors. *PLoS One*. 2016;11(4), e0153411.
14. Gruosso T, Mieulet V, Cardon M, et al. Chronic oxidative stress promotes H2AX protein degradation and enhances chemosensitivity in breast cancer patients. *EMBO Mol Med*. 2016;8(5):527—549.
15. Mecham BH, Klus GT, Strovel J, et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res*. 2004;32(9):e74.
16. Zhang J, Baran J, Cros A, et al. International cancer genome Consortium data portal—a one-stop shop for cancer genomics data. *Database*. 2011;2011. https://doi.org/10.1093/database/bar026.
17. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*. 2005;33(suppl_2):W741—W748.
18. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res*. 2016;44(D1):D1023—D1031.
19. Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug—gene interactions. *Nucleic Acids Res*. 2016;44(D1):D1036—D1044.
20. Luo H, Lin Y, Gao F, Zhang C-T, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res*. 2014;42(D1):D574—D580.
21. Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*. 2011;2011. https://doi.org/10.1093/database/bar030.
22. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The cancer genome Atlas Pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113—1120.
23. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein—protein association

networks, made broadly accessible. *Nucleic Acids Res*. 2017; 45(D1):D362—D368.
24. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498—2504.
25. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf*. 2003;4(1):2.
26. Polyak K. Heterogeneity in breast cancer. *J Clin Invest*. 2011; 121(10):3786.
27. He KY, Ge D, He MM. Big data analytics for genomic medicine. *Int J Mol Sci*. 2017;18(2). https://doi.org/10.3390/ijms18020412.
28. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLoS Biol*. 2006;4(6):e180.
29. Mugal CF, Arndt PF, Holm L, Ellegren H. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3 Genes Genomes Genetics*. 2015;5(3):441—447.
30. Balin SJ, Cascalho M. The rate of mutation of a single gene. *Nucleic Acids Res*. 2010;38(5):1575—1582.
31. Ma Y-Y, Wei S-J, Lin Y-C, et al. PIK3CA as an oncogene in cervical cancer. *Oncogene*. 2000;19(23):2739.
32. Takaku M, Grimm SA, Wade PA. *GATA3 in Breast Cancer: Tumor Suppressor or Oncogene*. 2015. https://doi.org/10.3727/105221615X14399878166113.
33. Ideker T, Sharan R. Protein networks in disease. *Genome Res*. 2008;18(4):644—652.
34. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101—113.
35. MacDonald BT, Tamai K, He X. Wnt/β-Catenin signaling: components, mechanisms, and diseases. *Dev Cell*. 2009;17(1): 9—26.
36. Gasco M, Shami S, Crook T. The p53 pathway in breast cancer. *Breast Cancer Res*. 2002;4(2). https://doi.org/10.1186/bcr426.
37. Kopnin BP. Targets of oncogenes and tumor suppressors: key for understanding basic mechanisms of carcinogenesis. *Biochemistry*. 2000;65(1):2—27.
38. Pećina-Šlaus N. Tumor suppressor gene E-cadherin and its role in normal and malignant cells. *Cancer Cell Int*. 2003;3(1):17.
39. Wozniak M. Focal adhesion regulation of cell behavior. *Biochim Biophys Acta BBA Mol Cell Res*. 2004. https://doi.org/10.1016/s0167-4889(04)00099-0.
40. Rizwan A, Cheng M, Bhujwalla ZM, Krishnamachary B, Jiang L, Glunde K. Breast cancer cell adhesome and degradome interact to drive metastasis. *NPJ Breast Cancer*. 2015;1:15017.
41. Wu Y-H, Chang T-H, Huang Y-F, Huang H-D, Chou C-Y. COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*. 2014;33(26):3432—3440.
42. Li J, Poi MJ, Tsai M-D. Regulatory mechanisms of tumor suppressor P16(INK4A) and their relevance to cancer. *Biochemistry*. 2011;50(25):5566—5582.
43. He G, Siddik ZH, Huang Z, et al. Induction of p21 by p53 following DNA damage inhibits both Cdk4 and Cdk2 activities. *Oncogene*. 2005;24(18):2929—2943.