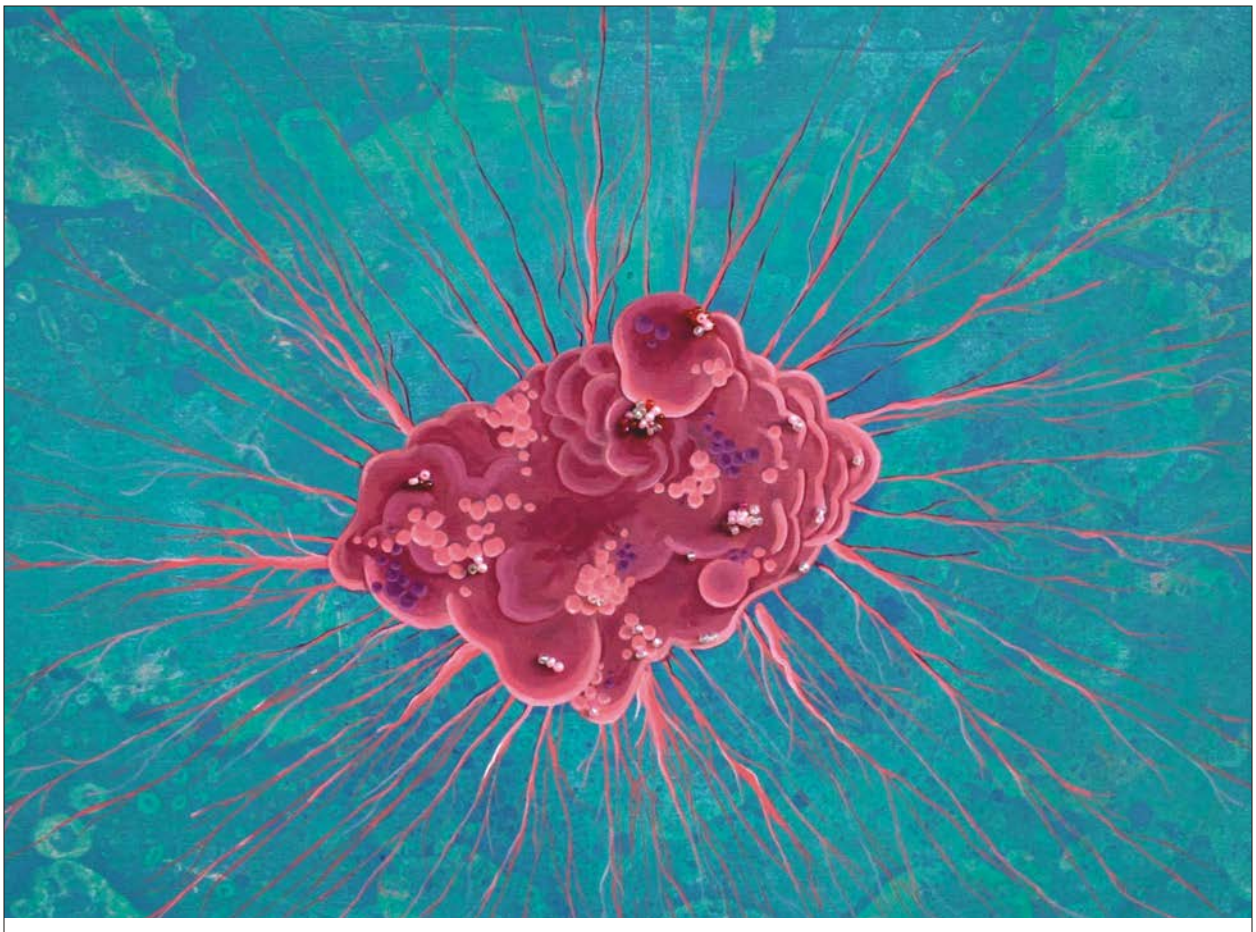

International Journal of Cancer Research

ISSN 1811-9727

Academic Journals Inc.

www.academicjournals.com

OPEN ACCESS International Journal of Cancer Research

ISSN 1811-9727
DOI: 10.3923/ijcr.2017.51.58

Research Article
Identification of Target Genes in Breast Cancer Pathway using Protein-Protein Interaction Network

Divya Bafna and Arnold Emerson Isaac

Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology University, Vellore, Tamil Nadu, India

Abstract

Background and Objective: The emergence of cancer genomics has expanded the background of protein-protein interaction network applications for identification of protein targets for drug development. The objective of the present study was to analyse breast cancer pathway genes to determine the potential drug targets. **Methodology:** In this study, breast cancer subnetwork was constructed from the pathway genes and potential targets in breast cancer pathway were identified using node or gene deletion analysis. The most popular centrality measures, such as betweenness and closeness centrality play a major role in the network robustness. Deleting the genes with the highest centrality values may result in significant destruction of the network. The significantly mutated values of the genes involved assists in selecting a precise target were determined using z-score. **Results:** On deleting the top 10 genes with highest betweenness centrality, significant ($p<0.05$) changes were observed in both shortest path length (L) and clustering coefficient (c) values as compared to breast cancer subnetwork. Out of these top 10 genes two of them had positive significant mutation values. **Conclusion:** These genes were identified to be NOTCH 1 (NOTCH family of proteins) and epidermal growth factor receptor (EGFR family) and therefore, these genes are possible target for drug therapy.

Key words: Breast cancer, closeness centrality, hubs, mutations, NOTCH proteins, EGFR family, gene expression

Citation: Divya Bafna and Arnold Emerson Isaac, 2017. Identification of target genes in breast cancer pathway using protein-protein interaction network. Int. J. Cancer Res., 13: 51-58.

Corresponding Author: Arnold Emerson Isaac, Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology University, Vellore, Tamil Nadu, India Tel: +(91)-9442309479

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Cancer is a group of diseases causing abnormal out of control cell growth which has the ability to spread to other parts of the body. Mutations in the growth regulating genes as well as abnormal changes in the genes may result in cancer. The term "breast cancer" refers to a malignant tumor that has developed from cells in the breast. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands or the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast. Over time, cancer cells can invade nearby healthy breast tissues and make their way into the underarm lymph nodes, small organs that filter out foreign substances in the body. If cancer cells get into the lymph nodes, then they have a pathway into other parts of the body. The most common type of breast cancer is ductal carcinoma, which begins in the cells of the ducts[1]. Ductal carcinoma in situ is a condition in which abnormal cells are found in the lining of the ducts but they haven't spread outside the duct. Breast cancer that has spread from where it began in the ducts or lobules to surrounding tissues is called invasive breast cancer. In inflammatory breast cancer, the breast looks red and swollen and feels warm because the cancer cells block the lymph vessels in the skin. Recently there has been several review articles on breast cancer which describes the updates in diagnosis and treatment[2-4].

All known or predicted protein interactions in an organism are summarized as a protein network[5]. Proteins' function and cellular localization of yeast can be uncovered using large scale properties of a protein-protein network[6]. Topological and dynamic features of protein networks can be inferred by systematic mapping of protein interactions. This can also be used to illuminate the mechanisms and development of diseases[7]. Protein-protein interaction network has also been used to predict the outcome of breast cancer patients and it was revealed that human interactome may work as an indicator of the same[8].

Protein-protein interaction network of human genome was constructed. Then the list of genes in breast cancer pathway from KEGG database was used to construct the breast cancer subnetwork. Network analyzer plug-in was used to analyse the network. The network was analyzed on the basis of clustering coefficient (C) values and shortest path length (L) values. To determine the robustness of the network, deletion analysis was performed for the subnetwork. The top 10 genes with the highest values of hubs, betweenness centrality and closeness centrality were used in the deletion analysis. Betweenness centrality was chosen as the primary parameter. The mutation count of the genes involved was obtained from dbSNP and the significant mutated genes were identified using the z-score value. The purpose of the study is to analyse the breast cancer pathways genes using protein-protein interaction network and to determine the potential target genes that are responsible for the disease progression.

MATERIALS AND METHODS

The human interactome network was obtained from freely accessible BioGRID on December, 2016 which is a unified database of physical and genetic interactions making it useful for analysis of protein network[9]. Cytoscape (3.4.0) was used to construct the human Interactome network. Cytoscape is open source software used to integrate biomolecular interaction networks with high-through put expression data and other molecular states into a unified conceptual framework[10]. The breast cancer pathway genes were obtained from KEGG disease which is a database containing higher order functional information linking disease genes, pathways, drugs and diagnostic markers[11]. There are 130 genes involved in the breast cancer pathway as shown in Table 1. To determine the genes involved in breast cancer pathway that undergo frequent mutations, the dbSNP database was used[12]. The gene mutations were normalized with the gene length by dividing the gene mutation count with its respective length.

Statistical analysis: The statistically ($p \leq 0.05$) significant mutated genes were evaluated using the z-score values of the gene mutation count[13]:

$$Z = \frac{X-\mu}{\sigma}$$

where, X is the normalized mutation count for the gene, μ is the average of the normalized score and σ is the corresponding standard deviation.

RESULTS

The human interactome was obtained from BioGRID database that contains 19634 nodes and 270970 edges as shown in Fig. 1a. The nodes are denoted by the colour blue while the edges are of colour grey. The grid layout was selected to represent the human protein-protein interaction network. To determine the breast cancer subnetwork, the breast cancer pathway entry was selected from the KEGG

Table 1: List of breast cancer pathway genes

Sr. No.	Gene_ID	z-scores	Hubs	Betweeness centrality values	Closeness centrality values
1	2260	-0.2439	11	1	1
2	2932	-1.18122	98	0.2042279	0.5
3	1499	-0.32057	112	0.19012231	0.51052632
4	4851	1.828162	24	0.12288291	0.45116279
5	1956	0.563471	126	0.10632966	0.46190476
6	4040	-0.27989	16	0.09675091	0.39271255
7	2099	-0.75115	160	0.09362488	0.48989899
8	7157	-0.8482	107	0.09011549	0.47087379
9	5594	-0.1030	64	0.08412381	0.47317073
10	5295	-0.86063	58	0.07200074	0.47087379
11	5925	-0.20628	159	0.06677725	0.42731278
12	207	-0.14534	55	0.06592063	0.48743719
13	672	-0.57661	128	0.05004408	0.45971564
14	8312	1.395835	97	0.04163865	0.42173913
15	6667	-0.60218	71	0.03883024	0.44090909
16	5894	-0.03990	86	0.03844442	0.41810345
17	5728	-1.37439	12	0.03841729	0.41276596
18	4609	0.38391	40	0.03456511	0.44907407
19	3265	-0.00410	34	0.03089781	0.39430894
20	89780	0.133484	4	0.03026763	0.28783383
21	2064	0.895364	41	0.02630644	0.42920354
22	673	-0.13547	27	0.02401321	0.37890625
23	54361	-0.62126	3	0.02295609	0.30599369
24	1869	-0.27843	75	0.02237904	0.41276596
25	5294	-0.00217	10	0.02199722	0.29846154
26	23493	0.001711	5	0.02061856	0.31493506
27	6464	0.064001	101	0.02009434	0.39754098
28	8321	-0.07588	3	0.01952232	0.34892086
29	5290	-0.10430	18	0.01904211	0.40756303
30	5296	0.238344	31	0.01662587	0.40248963
31	7474	-1.0599	4	0.01512285	0.29041916
32	3845	-1.50306	15	0.01341261	0.36603774
33	8202	0.272378	58	0.01286258	0.43111111
34	2885	-1.30842	167	0.01279437	0.37164751
35	208	-0.81479	8	0.01052089	0.41810345
36	3480	-0.41503	16	0.01022718	0.38188976
37	4854	1.259889	8	0.00890637	0.37022901
38	5293	0.175714	9	0.00865919	0.34642857
39	8503	-0.6279	11	0.00855573	0.37307692
40	1026	-0.47932	48	0.00834347	0.39917695
41	3725	-0.6430	69	0.00824302	0.42173913
42	2475	0.543097	46	0.00819799	0.41991342
43	5595	0.227172	38	0.00694461	0.39271255
44	5291	-0.6099	11	0.00520035	0.37451737
45	4853	0.260143	15	0.00504817	0.3540146
46	5605	0.73901	22	0.00484143	0.38188976
47	1452	-1.04560	6	0.00482327	0.388
48	595	-0.84931	75	0.00478243	0.38955823
49	1857	-0.15584	11	0.00214839	0.3605948
50	1856	1.510756	11	0.00195165	0.37890625
51	1855	2.497405	11	0.00175677	0.3540146
52	5604	-0.76292	52	0.0016313	0.35793358
53	8313	0.798976	8	0.00161208	0.37307692
54	369	0.123287	10	0.00156544	0.33916084
55	4041	2.618113	5	0.00133093	0.3003096
56	51176	-0.6172	20	0.00130531	0.37022901
57	8648	0.268686	63	0.0012728	0.36742424
58	4893	-1.38257	2	0.00111798	0.3003096
59	6654	-0.54994	49	0.00109809	0.33797909
60	324	0.215738	42	0.00106933	0.37307692
61	8325	0.416373	4	9.31E-04	0.28529412

Table 1: Continue

Sr. No.	Gene_ID	z-scores	Hubs	Betweeness centrality values	Closeness centrality values
62	2353	0.133465	47	9.30E-04	0.37743191
63	1019	-0.02093	84	7.23E-04	0.35661765
64	10023	-0.97280	3	6.29E-04	0.33916084
65	2535	-0.43728	2	4.30E-04	0.22769953
66	5241	-0.88271	23	4.13E-04	0.37743191
67	6655	0.251712	15	3.52E-04	0.35144928
68	2100	1.173758	30	3.41E-04	0.37307692
69	3479	-1.43875	2	1.50E-04	0.31290323
70	1021	-1.46893	20	1.24E-04	0.33797909
71	182	0.264526	10	4.30E-05	0.31493506
72	3714	4.69177	4	4.30E-05	0.31493506
73	28514	0.846624	6	4.30E-05	0.31493506
74	2246	-1.15782	4	0	0.6
75	2247	-1.28797	1	0	0.6
76	8074	-0.68754	2	0	0.6
77	675	-0.12108	17	0	0.34519573
78	7855	-0.84824	2	0	0.34519573
79	10297	0.838099	5	0	0.34035088
80	6934	-0.04408	6	0	0.33916084
81	83439	0.749097	1	0	0.33916084
82	10000	-0.2415	1	0	0.33448276
83	23401	-0.8432	1	0	0.33448276
84	6198	-1.08419	27	0	0.33219178
85	3815	-0.13159	14	0	0.32993197
86	1950	0.492781	9	0	0.32550336
87	4855	1.47888	1	0	0.32119205
88	25759	2.086843	2	0	0.31803279
89	53358	-0.48815	3	0	0.31803279
90	399694	1.568908	1	0	0.31699346
91	7482	-0.39402	1	0	0.31596091
92	23462	-0.17753	4	0	0.31391586
93	4791	0.468657	5	0	0.30696203
94	1870	-0.61543	6	0	0.3003096
95	1871	-0.81021	4	0	0.3003096
96	2324	1.222096	6	0	0.29305136
97	7471	-0.31088	2	0	0.28445748
98	6199	2.56192	4	0	0.27556818
99	3280	0.205859	1	0	0.24009901
100	7477	-0.40928	1	0	0.23486683
101	23533	0.253561	4	0	0.2304038
102	7473	-0.82629	1	0	0.22401848
103	2248	0.093074	0	0	0
104	2252	-1.29745	0	0	0
105	2253	0.550788	0	0	0
106	2254	-1.25689	2	0	0
107	2255	1.509493	0	0	0
108	2257	-1.31101	0	0	0
109	2258	-1.20381	0	0	0
110	6932	1.425963	0	0	0
111	7472	-0.17489	0	0	0
112	7478	-0.32501	0	0	0
113	7480	0.149664	0	0	0
114	7481	0.582552	0	0	0
115	7484	-0.78058	0	0	0
116	7976	-1.27452	0	0	0
117	8322	-0.91938	0	0	0
118	8323	0.161067	0	0	0

Table 1: Continue

Sr. No.	Gene_ID	z-scores	Hubs	Betweeness centrality values	Closeness centrality values
119	8324	0.045165	0	0	0
120	8326	1.074483	0	0	0
121	8600	-0.50446	0	0	0
122	8817	-0.68462	0	0	0
123	8823	-0.75683	0	0	0
124	9965	-0.96573	0	0	0
125	26291	0.729261	0	0	0
126	26508	-0.66019	0	0	0
127	51384	-0.20245	0	0	0
128	81029	0.183003	0	0	0
129	122011	0.164667	2	0	0
130	388585	-0.90447	0	0	0

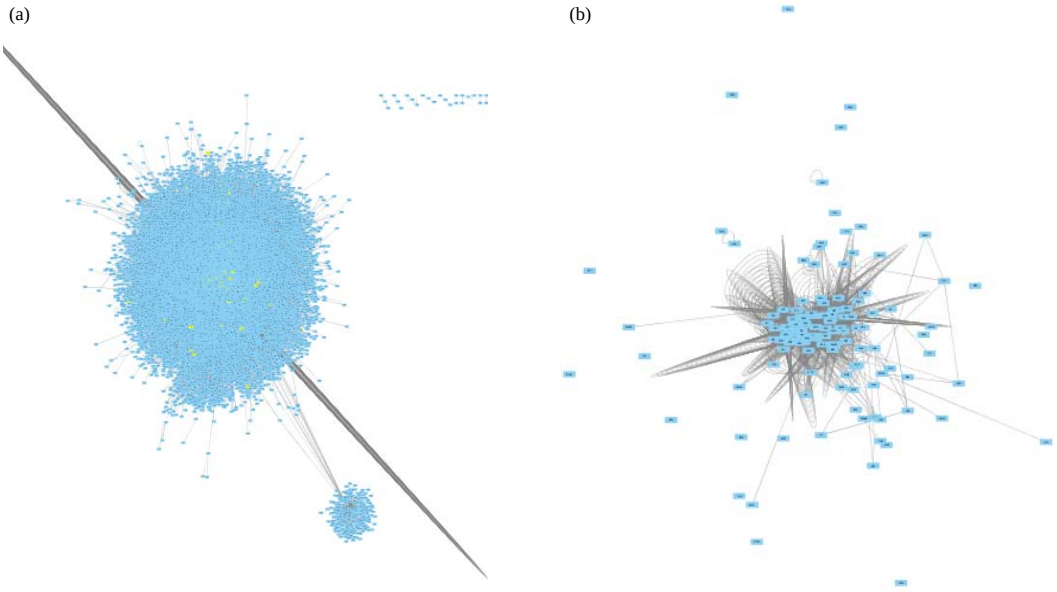

Fig. 1(a-b): (a) Normal human protein-protein interaction network and (b) Breast cancer protein-protein interaction sub network

database. The numbers of genes obtained were 130 and then Cytoscape version 3.4.0 was used to construct a subnetwork of breast cancer proteins as shown in the Fig. 1b. The breast cancer subnetwork contains 130 nodes and 1538 edges.

Basic network parameters were calculated using Network Analyzer. The shortest path length (L) was found to be 2.85 and the clustering coefficient (C) was found to be 0.286 for the breast cancer subnetwork. For deletion analysis, the highest 10 values of hubs, betweenness centrality and closeness centrality were considered. On deleting the genes with the highest number of hubs, the shortest path length was found to be 4.177 and the clustering coefficient was found to be 0.188. From the above result the value of L was found to be increased whereas the value of C was decreased. Therefore, the results suggest that the hubs play a major role in the robustness of the network. The highest 10 closeness centrality genes were deleted and the parameters were found to be 3.547 for L and 0.251 for C. These values were not much deviated from the breast cancer subnetwork parameters but the L value was found to be increased.

After deletion of the genes with the highest 10 values of betweenness centrality, the shortest path length was found to be 3.915 and the clustering coefficient was found to be 0.187. The value of C was found to lower than that of the breast cancer subnetwork analysis while L was higher than the breast cancer subnetwork. According to the Kolmogorov-Smirnov test, the difference in L and C were found to be statistically significant ($p \leq 0.05$). Then the

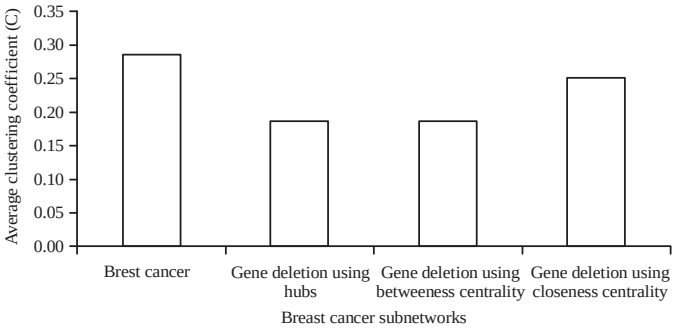

Fig. 2: Clustering coefficient values for breast cancer subnetworks, removal of 10 genes with the highest value of hubs, betweenness and closeness centrality. Deletion analysis indicates that the top 10 genes are essential in communication with in the breast cancer pathways

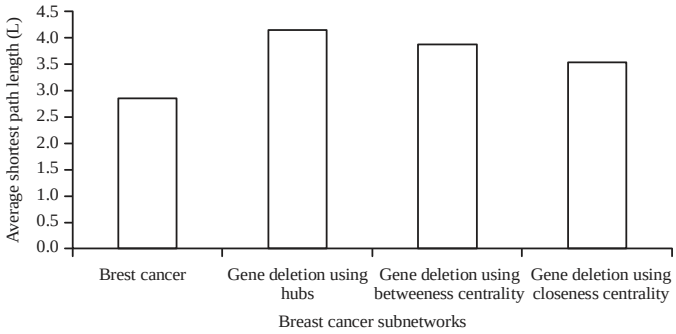

Fig. 3: Shortest path length values for breast cancer subnetworks, removal of 10 genes with the highest value of hubs, betweenness and closeness centrality

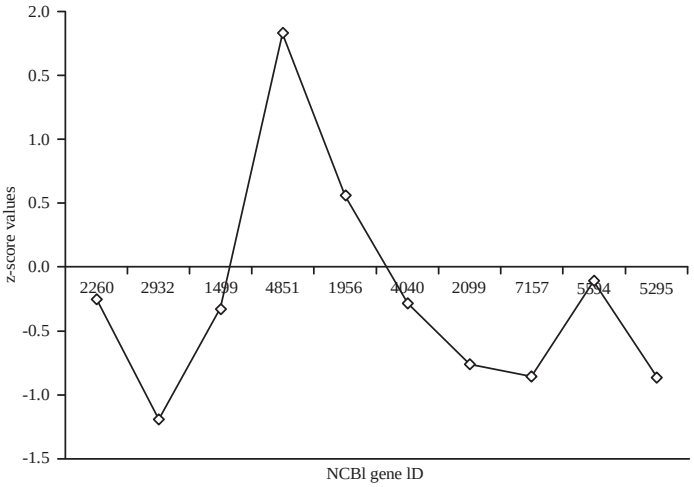

Fig. 4: Statistically significant mutation values for genes with highest betweenness centrality

betweenness centrality was chosen as the parameter to compare genes are the clustering coefficient (Fig. 2) was lower and the shortest path length (Fig. 3) was higher when the genes with top 10 betweenness centrality were deleted than that of top 10 genes with the highest closeness centrality deleted.

The SNPs for all the genes involved in the breast cancer subnetwork were obtained and normalized based on the gene length. The statistically significant central genes were determined using z-score function represents the top 10 genes with the highest betweenness centrality values and their respective z-score values represented in Fig. 4. Among

the 10 genes, the two genes with the positive z-score values were selected and their NCBI gene ID was found to be 4851 (NOTCH family of proteins) and 1956 (EGFR family).

DISCUSSION

The protein-protein interaction network of breast cancer proteins was constructed from KEGG pathway database. The network showed characters of scale free network and hierarchical network as well. Node deletion or gene deletion in a network may help to understand the robustness and attack tolerance of the network. When the largest 10 central genes were removed, the shortest path length increased as well as the clustering coefficient decreased. The deletion analysis indicates that cancer pathway related genes are highly robust to removal of central genes.

The study concluded that NOTCH-1 family of proteins and EGFR protein are the most effective target for drug development for the destruction of the breast cancer protein-protein interaction subnetwork. These genes have a high value of betweenness centrality. Betweenness centrality signifies the centrality of the protein in the network. It has been shown that highly connected vertices in protein interaction networks are often functionally important and the deletion of such vertices is related to lethality[14].

Hence, the two genes exhibit a relatively low value of clustering coefficient and a high value of shortest path length in their respective deletion analysis. For further precision of the proteins to be targeted out of the proteins with the top 10 highest betweenness centrality, positive significant mutation values are taken into account. It has also been argued that network analysis can be used in general to infer novel functions, to quantify positional importance of protein in a disease associated pathway[15]. The significant mutation value is useful in finding protein targets that themselves are susceptible to a high number of mutations per gene length. Thus the targets play central role in the destruction of the breast cancer network while possessing a value of significant mutation a well making them an ideal candidate for drug development.

It has been found that targeting NOTCH signalling maybe of therapeutic value in breast cancer as it is over expressed and highly activated[16]. A recent review suggested that NOTCH signalling pathways play a vital role development and progression of breast cancer[17]. The EGFR receptor has been extensively studied in breast cancer. The EGFR is known to be over expressed in triple negative breast cancer. Growth factors such as EGFR, c-kit or p53 mutation status and several proliferative mechanisms like mitogen-activated protein kinase (MAPK) and protein kinase components of the extracellular signal-regulated kinases (ERK) pathway have been indicated as possible determinants of sensitivity to chemotherapy in TNBC[7]. The TNBC is strongly associated with EGFR expression. Yet, the most benefit of tandem high-dose chemotherapy was shown among TNBCs but not in the small subgroup of EGFR-positive tumors indicating the need for additional targeted therapies in this fraction[7].

Complex network analyses in breast cancer disease were also used to identify the target genes. The gene regulatory network comprising transcription factors and target genes were used to identify the regulatory circuits governing breast cancer disease[18]. Even the structural changes in the gene regulatory network have been analyzed using network analysis of breast cancer genes[19]. The genes associated to the disease were determined using breast cancer networks' centrality[20]. Another study in cancer genomics, that describes the overview of pathway and network analysis techniques in tumour biology[21].

CONCLUSION

The breast cancer subnetwork was created from the genes that take part in cancer pathway and the mutations were mapped to their corresponding genes. Deletion analysis was carried out for the top 10 genes with highest hub values, betweenness centrality and closeness centrality values, respectively. Among the three network parameters, betweenness centrality values were found to have a major deviation in comparison with the breast cancer network. The genes with the top 10 betweenness centrality values were isolated out of which the genes with positive significant mutation values were selected. Further analysis shows that two of these have been reported as breast cancer associated genes. It is suggested that NOTCH family of proteins and EGFR family genes could be used as a target for drug development.

SIGNIFICANCE STATEMENTS

This study discovers the potential drug targets for breast cancer that can be beneficial for the cancer research community to develop new drugs for the treatment of cancer. This study will help the researchers to uncover the critical areas of determining the receptors that are crucial for the disease progression and also it will enhance the knowledge in target identification that many researchers were not able to explore. Thus a new theory on node deletion or gene deletion analysis for identifying the drug targets and

in combination of significantly mutated genes, may help to identify new drug targets for breast cancer disease.

ACKNOWLEDGMENT

The authors thank to the management of Vellore Institute of Technology for providing the computational facility required for this study.

REFERENCES

1. Mardekian, S.K., A. Bombonati and J.P. Palazzo, 2016. Ductal carcinoma in situ of the breast: The importance of morphologic and molecular interactions. Hum. Pathol., 49: 114-123.
2. Majidinia, M. and B. Yousefi, 2017. DNA repair and damage pathways in breast cancer development and therapy. DNA Repair, 54: 22-29.
3. Castaneda, S.A. and J. Strasser, 2017. Updates in the treatment of breast cancer with radiotherapy. Surg. Oncol. Clin., 26: 371-382.
4. Prabha, A.T. and D. Sekar, 2017. Deciphering the molecular signaling pathways in breast cancer pathogenesis and their role in diagnostic and treatment modalities. Gene Rep., 7: 1-17.
5. Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund and D. Heller *et al*., 2014. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucl. Acids Res., 43: D447-D452.
6. Yook, S.H., Z.N. Oltvai and A.L. Barabasi, 2004. Functional and topological characterization of protein interaction networks. Proteomics, 4: 928-942.
7. Rual, J.F., K. Venkatesan, H. Tong and T. Hirozane-Kishikawa, 2005. Towards a proteome-scale map of the human protein-protein interaction network. Nature, 437: 1173-1178.
8. Taylor, I.W., R. Linding, D. Warde-Farley, Y. Liu and C. Pesquita *et al*., 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat. Biotechnol., 27: 199-204.
9. Stark, C., B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, 2006. BioGRID: A general repository for interaction datasets. Nucl. Acids Res., 34: D535-D539.
10. Shannon, P., A. Markiel, O. Ozier, N.S. Baliga and J.T. Wang *et al*., 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res., 13: 2498-2504.
11. Kanehisa, M. and S. Goto, 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucl. Acids Res., 28: 27-30.
12. Smigielski, E.M., K. Sirotkin, M. Ward and S.T. Sherry, 2000. dbSNP: A database of single nucleotide polymorphisms. Nucl. Acids Res., 28: 352-355.
13. Del Sol, A., H. Fujihashi, D. Amoros and R. Nussinov, 2006. Residue centrality, functionally important residues and active site shape: Analysis of enzyme and non-enzyme families. Protein Sci., 15: 2120-2128.
14. Jeong, H., S.P. Mason, A.L. Barabasi and Z.N. Oltvai, 2001. Lethality and centrality in protein networks. Nature, 411: 41-42.
15. Jordan, F., T.P. Nguyen and W.C. Liu, 2012. Studying protein-protein interaction networks: A systems view on diseases. Briefings Funct. Genomics, 11: 497-504.
16. Zang, S., C.H. Ji, X. Qu, X. Dong and D. Ma *et al*., 2007. A study on NOTCH signaling in human breast cancer. Neoplasma, 54: 304-310.
17. Liu, J., J.X. Shen, X.F. Wen, Y.X. Guo and G.J. Zhang, 2016. Targeting NOTCH degradation system provides promise for breast cancer therapeutics. Crit. Rev. Oncol./Hematol., 104: 21-29.
18. Castro, M.A.A., I. de Santiago, T.M. Campbell, C. Vaughn and T.E. Hickey *et al*., 2016. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nat. Genet., 48: 12-21.
19. Parikh, A.P., R.E. Curtis, I. Kuhn, S. Becker-Weimann, M. Bissell, E.P. Xing and W. Wu, 2014. Network analysis of breast cancer progression and reversal using a tree-evolving network algorithm. PLoS Comput. Biol., Vol. 10.
20. De Anda-Jauregui, G., T.E. Velazquez-Caldelas, J. Espinal-Enriquez and E. Hernandez-Lemus, 2016. Transcriptional network architecture of breast cancer molecular subtypes. Front. Physiol., Vol. 7. 10.3389/fphys. 2016.00568.
21. Creixell, P., J. Reimand, S. Haider, G. Wu and T. Shibata *et al*., 2015. Pathway and network analysis of cancer genomes. Nat. Methods, 12: 615-621.