Journal of Breast Cancer

J Breast Cancer 2017 September; 20(3): 240-245 https://doi.org/10.4048/jbc.2017.20.3.240

Check for updates

ORIGINAL ARTICLE

Gene Regulatory Network Analysis for Triple-Negative Breast Neoplasms by Using Gene Expression Data

Hee Chan Jung, Sung Hwan Kim[1], Jeong Hoon Lee[2], Ju Han Kim[2], Sung Won Han[3]

Department of Internal Medicine, Eulji University College of Medicine, Seoul; [1]Department of Statistics, Keimyung University, Daegu; [2]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul; [3]Division of Fusion Data Analytics Laboratory, School of Industrial Management Engineering, Korea University, Seoul, Korea

Purpose: To better identify the physiology of triple-negative breast neoplasm (TNBN), we analyzed the TNBN gene regulatory network using gene expression data. **Methods:** We collected TNBN gene expression data from The Cancer Genome Atlas to construct a TNBN gene regulatory network using least absolute shrinkage and selection operator regression. In addition, we constructed a triple-positive breast neoplasm (TPBN) network for comparison. Furthermore, survival analysis based on gene expression levels and differentially expressed gene (DEG) analysis were carried out to support and compare the network analysis results, respectively. **Results:** The TNBN gene regulatory network, which followed a power-law distribution, had 10,237 vertices and 17,773 edges, with an average vertex-to-vertex distance of 8.6. The genes *ZDHHC20* and *RAPGEF6* were identified by centrality analysis to be important vertices. However, in the DEG analysis, we could not find meaningful fold changes in *ZDHHC20* and *RAPGEF6* between the TPBN and TNBN gene expression data. In the multivariate survival analysis, the hazard ratio for *ZDHHC20* and *RAPGEF6* was 1.677 (1.192–2.357) and 1.676 (1.222–2.299), respectively. **Conclusion:** Our TNBN gene regulatory network was a scale-free one, which means that the network would be easily destroyed if the hub vertices were attacked. Thus, it is important to identify the hub vertices in the network analysis. In the TNBN gene regulatory network, *ZDHHC20* and *RAPGEF6* were found to be oncogenes. Further study of these genes could help to reveal a novel method for treating TNBN in the future.

Key Words: *Genes, Oncogenes, Triple negative breast neoplasms*

INTRODUCTION

Breast cancer is a serious disease among women and has become increasingly prevalent worldwide [1]. Triple-negative breast neoplasm (TNBN) account for 15% to 20% of breast cancers, and is intractable to treatment owing to its poor prognosis and high recurrence rate [2,3]. Over the years, a great deal of effort has been expended to enhance the efficacy of TNBN treatments using the angiogenesis inhibitors bevacizumab and paclitaxel; however, this remains only in the developmental stage [4]. Therefore, it is very important to better understand the physiology of TNBN. There are many methods available to identify and understand the physiology of cancers from the gene viewpoint, such as differentially expressed gene (DEG) analysis, and gene clustering and classification. However, these methods have limitations in identifying gene-gene interactions and connections. In addition, gene clustering and classification do not detect important genes in formed clusters.

Given these limitations, we attempted to construct a TNBN gene regulatory network using gene expression data. Previously, de Matos Simoes and Emmert-Streib [5] proved the utility of gene expression data for constructing a gene regulatory network of breast cancer via the BC3Net method, and found significant pathways enriched for the cell cycle and immune response [6]. However, in contrast to their method, we used conditional independence graphs with least absolute shrinkage and selection operator (LASSO) regression to exclude falsely detected gene regulatory networks. By doing this, we created a more precise network to identify gene-gene interactions and hub genes. In addition, we used triple-positive breast neoplasm (TPBN) gene expression data to compare with the TNBN gene regulatory network data, although TPBN is not a definitive entity of breast cancer.

Correspondence to: **Sung Won Han**
Division of Fusion Data Analytics Laboratory, School of Industrial Management Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-3384, Fax: +82-2-929-5888
E-mail: swhan@korea.ac.kr

Received: January 20, 2017 **Accepted:** July 31, 2017

http://ejbc.kr | pISSN 1738-6756
eISSN 2092-9900

METHODS

We retrieved RNA-Seq expression data for breast cancer from The Cancer Genome Atlas (TCGA) [7]. By definition, in terms of immunohistochemistry, TNBN is negative for the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), whereas TPBN is positive for all three receptors [7].

Data characteristics

Of the 1,088 patients with breast cancer logged in TCGA, 115 (10.6%) had TNBN and 97 (8.9%) had TPBN (Table 1). The mean age of the patients with TNBN was 54.73 years (range, 42.94–66.52 years), being statistically significantly lower than that of the patients with TPBN, which was 59.98 years (range, 45.88–74.08 years) ($p=0.004$, Student t-test). Of the combined TNBN and TPBN groups, 190 patients were alive and 22 patients were deceased at the time of study. The proportions of pathologic tumor stages were similar in both groups, with stage II being the most frequent followed by stages III and I. Most patients did not receive neoadjuvant chemotherapy, but they also did not have a surgical margin status. Differences in neoadjuvant chemotherapy and surgical margin status were not statistically significant between the two groups ($p>0.05$, generalized Fisher exact test). RNA-Seq V2 expression levels (log2-transformed and normalized RNA-Seq by expectation–maximization values) were retrieved from the TCGA portal.

Table 1. Demographics of the triple-positive and triple-negative breast neoplasm patients

Characteristics	Triple-negative (n=115) No. (%)	Triple-positive (n=97) No. (%)	All (n=212) No. (%)
Stage			
I	19 (16.5)	9 (9.3)	28 (13.2)
II	72 (62.6)	56 (57.8)	128 (60.4)
III	19 (16.5)	30 (30.9)	49 (23.1)
IV	2 (1.8)	1 (1.0)	3 (1.4)
NA	3 (2.6)	1 (1.0)	4 (1.9)
Age (yr)*	54.73 (42.94–66.52)	59.98 (45.88–74.08)	57.09 (43.98–70.20)
Status			
Alive	103 (89.6)	87 (89.6)	190 (89.6)
Dead	12 (10.4)	10 (10.4)	22 (10.4)
NAC			
Yes	0	4 (4.1)	4 (1.9)
No	114 (99.1)	93 (95.9)	207 (97.6)
NA	1 (0.9)	0	1 (0.5)
Margin status			
Positive	3 (2.6)	4 (4.2)	7 (3.3)
Negative	102 (88.7)	81 (83.5)	183 (86.3)
Close	4 (3.5)	1 (1.0)	5 (2.4)
NA	6 (5.2)	11 (11.3)	17 (8.0)
Race			
White	67 (58.3)	63 (64.9)	130 (61.3)
Black	32 (27.9)	8 (8.3)	40 (18.9)
Asian	8 (6.9)	5 (5.2)	13 (6.1)
NA	8 (6.9)	21 (21.6)	29 (13.7)

NAC=neoadjuvant chemotherapy; NA=not available.
*Median (range).

Statistical analyses

The statistical analyses involved a two-stage analytical scheme: (1) regression-based network inference and (2) post hoc analysis. In the first stage, we estimated probabilistic neighbors (typically called a conditional independence graph) on the basis of gene expression in the triple-positive and triple-negative patients, respectively. We used LASSO regression to estimate the probabilistic neighbors, applying the optimal penalty parameter to control the probability of including falsely estimated neighbors [8,9]. The LASSO-based approach estimates a network by finding probabilistic neighbors around each node, and is computationally efficient and requires only a small amount of memory in computing systems. Thus, this approach is very applicable to such high-dimensional data. The estimated neighbors indicate functional interactions between genes. After that, we calculated the degree of each gene (called the hub gene) and the number of neighboring genes around the hub genes. Subsequently, we sorted the hub genes by degree from large to small, and performed a post hoc analysis to understand their biological functions. In the second stage, both univariate and multivariate Cox proportional hazard models were used to assess survival rates related to hub-gene expression. The age, pathologic tumor stage, and ER, PR, and HER2 status were used as covariates to adjust the univariate factor in the survival model. In addition, DEGs between TNBN and TPBN were selected by performing empirical Bayes moderated t-statistics on the log2-transformed RNA-Seq data, with cutoff thresholds of Bonferroni corrected p-values of <0.05 and log fold changes of $>|1|$, using the Bioconductor "limma" R-package (http://bioconductor.org/packages/release/bioc/html/limma.html) [10,11]. For the network statistical analyses, NodeXL version 1.0.1.361 (The Social Media Research Foundation, Belmont, USA) was used.

RESULTS

In the TNBN gene regulatory network, a total of 10,237 vertices and 17,773 edges were observed. The graph density was 0.0003, the maximum vertex-to-vertex distance was 28, and the average vertex-to-vertex distance was 8.6, which means that if one were to go through eight vertices, all would be connected. Statistical results for the TPBN gene regulatory

network were similar (Table 2).

In the network centrality analysis, the TNBN gene regulatory network revealed the genes *RAPGEF6*, *GTF2A1*, and *ASXL2* to have the highest hub vertex degree with 38 edges (Tables 3-5). In addition, we conducted a hub vertex analysis with the betweenness centrality and eigenvector centrality. *ZDHHC20* had the highest value (2593718.407) for the betweenness centrality, whereas *ASXL2* had the highest value (0.019) for the eigenvector centrality.

For the network clustering analysis, we used the Clauset-Newman-Moore algorithm to divide the network into groups. In the TNBN gene regulatory network, there were 352 groups showing a modularity value of 0.825, and the largest group contained 1,241 vertices and 2,421 edges. On the other hand, there were 456 groups showing a modularity value of 0.820 in

Table 2. Gene regulatory network statistics of triple-negative and triple-positive breast neoplasms

Network statistics	TNBN	TPBN
Vertices	10,237	8,930
Total edges	17,773	15,223
Maximum geodesic distance (diameter)	28	29
Average geodesic distance	8.635109	8.649705
Graph density	0.000339225	0.000381835

TNBN = triple-negative breast neoplasm; TPBN = triple-positive breast neoplasm.

Table 3. Degree centrality results for triple-negative and triple-positive breast neoplasms

TNBN		TPBN	
Gene	Value	Gene	Value
ASXL2	38	*CLOCK*	39
GTF2A1	38	*REST*	32
RAPGEF6	38	*ATP5D*	30
ZDHHC20	35	*TGFBR2*	28
CCNT1	33	*ASXL2*	27
PDGFRB	32	*STRN*	26
REST	31	*RBM27*	25
TAOK1	31	*RIF1*	25
RIF1	30	*CCNT1*	25
ATP5D	30	*ZEB2*	25

TNBN = triple-negative breast neoplasm; TPBN = triple-positive breast neoplasm.

Table 4. Betweenness centrality results for triple-negative and triple-positive breast neoplasms

TNBN		TPBN	
Gene	Value	Gene	Value
ZDHHC20	2593718.407	*CLOCK*	1256280.673
RAPGEF6	2147541.505	*GYPC*	938225.682
ZNF192	1835667.961	*HIC1*	922622.728
GTF2A1	1468223.374	*REST*	890485.116
RIF1	1437912.482	*ATP5D*	812258.220
ATP5D	1369994.673	*FAM108A1*	794508.449
ASXL2	1366586.056	*GTF2A1*	773790.744
REST	1339309.745	*TRAPPC5*	772846.398
CCNT1	1198033.135	*ASXL2*	759382.681
GMCL1	1040959.614	*CCDC12*	686844.284

TNBN = triple-negative breast neoplasm; TPBN = triple-positive breast neoplasm.

Table 5. Eigenvector centrality results for triple-negative and triple-positive breast neoplasms

TNBN		TPBN	
Gene	Value	Gene	Value
ASXL2	0.019	*CLOCK*	0.016
GTF2A1	0.017	*CCNT1*	0.013
REST	0.016	*REST*	0.012
CCNT1	0.014	*ASXL2*	0.010
ZDHHC20	0.014	*STRN*	0.009
UHMK1	0.013	*NCOA2*	0.009
RAPGEF6	0.012	*UHMK1*	0.008
NCOA2	0.010	*EXOC6B*	0.007
LMTK2	0.009	*RC3H2*	0.006
TAOK1	0.009	*SHPRH*	0.006

TNBN = triple-negative breast neoplasm; TPBN = triple-positive breast neoplasm.

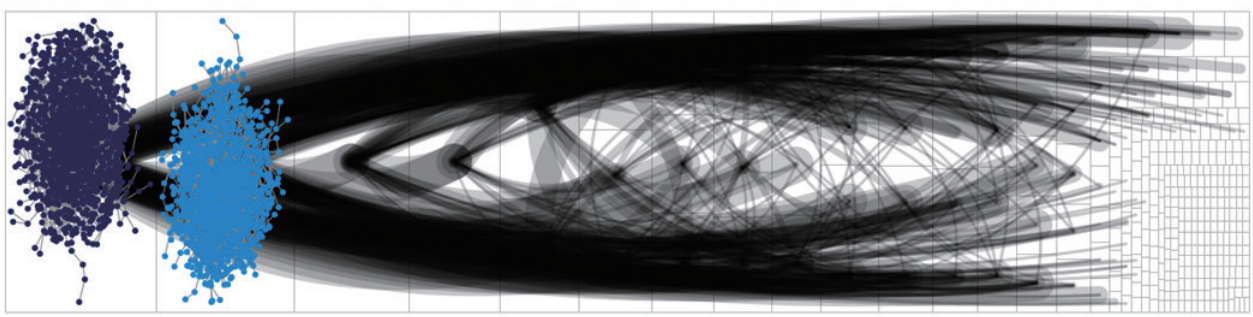

Figure 1. Cluster analysis of the triple-negative breast neoplasm gene regulatory network using the Clauset-Newman-Moore algorithm. The largest group (blue) and the second largest group (sky-blue) are connected the most frequently.

https://doi.org/10.4048/jbc.2017.20.3.240

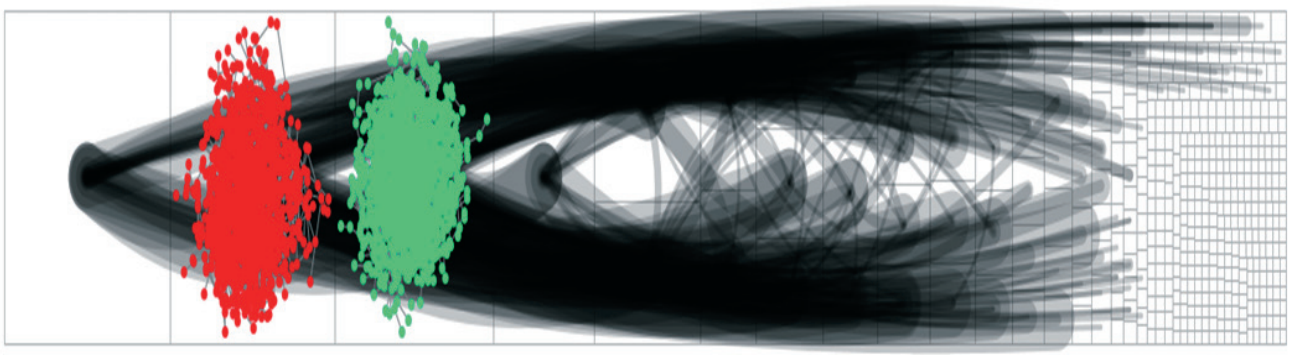

Figure 2. Cluster analysis of the triple-positive breast neoplasm gene regulatory network using the Clauset-Newman-Moore algorithm. The second largest group (red) and the third largest group (green) are connected the most frequently.

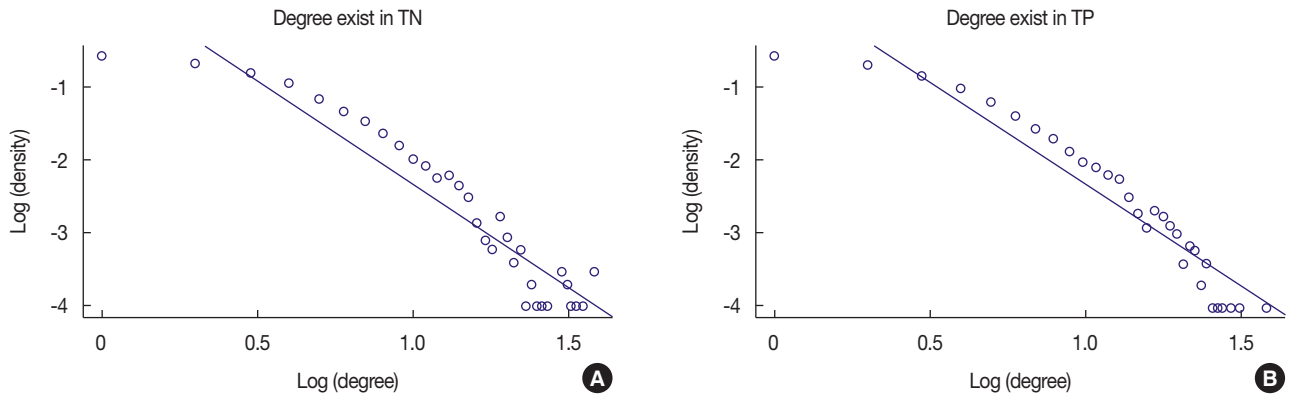

Figure 3. Regression analysis of the observed vertex degree and density values. (A) Regression analysis of degree exist in TN has slope -2.823, adjusted R^2 0.882, and $p<0.001$ which satisfy the power-law distribution. (B) Regression analysis of degree exist in TP has slope -2.727, adjusted R^2 0.897, and $p<0.001$ which satisfy the power-law distribution.
Degree exist in TN = triple-negative breast neoplasm group; Degree exist in TP = triple-positive breast neoplasm group.

Table 6. Cox regression based on clinical variables and hub genes

Clinical variable	Univariate		Multivariate	
	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value
Age at diagnosis (yr)	1.250 (1.081–1.445)	0.002	1.423 (1.137–1.782)	0.002
Stage				
I	Reference		Reference	
II	1.516 (0.800–2.871)	0.201	1.225 (0.523–2.868)	0.640
III	2.645 (1.349–5.187)	0.004	2.792 (1.115–6.987)	0.028
IV	4.737 (1.990–11.289)	<0.001	5.431 (1.443–20.433)	0.012
ER				
Positive	Reference		Reference	
Negative	1.582 (1.034–2.420)	0.034	1.205 (0.437–3.323)	0.718
PR				
Positive	Reference		Reference	
Negative	1.674 (1.119–2.505)	0.012	1.831(0.687–4.879)	0.226
HER2				
Positive	Reference		Reference	
Negative	0.313 (0.170–0.576)	<0.001	0.465 (0.240–0.898)	0.022
CLOCK	1.523 (1.238–1.874)	<0.001	1.779 (1.297–2.440)	<0.001
RAPGEF6	1.184 (1.000–1.402)	0.050	1.508 (1.108–2.053)	0.009
ZDHHC20	1.148 (0.974–1.352)	0.100	1.565 (1.179–2.079)	0.002

HR = hazard ratio; CI = confidence interval; ER = estrogen receptor; PR = progesterone receptor; HER2 = human epidermal growth factor receptor 2.

the TPBN gene regulatory network, with the largest group containing 1,153 vertices and 2,377 edges. In the TNBN gene regulatory network, the largest and second largest groups were connected most frequently (Figure 1), whereas the second and third largest groups were connected most frequently in the TPBN gene regulatory network (Figure 2).

It is known that gene regulatory networks in nature generally satisfy the power law. The distributions of vertex degrees were expected to follow the power law precisely, as defined by $P(k) \sim k^{-r}$, where r is an exponential factor. Using log-transformed values, we performed a regression analysis on the observed vertex degree and density values of the two gene regulatory networks (Figure 3, Supplementary Table 1, available online). The slopes in Supplementary Table 1 are the estimated -r values. We noted that the results adequately satisfied the power-law distribution.

In addition, to confirm the subtype-specific clinical relevance, the subset of patients was selected by the status of ER, PR, and HER2 for analysis by Cox proportional hazard regression. Survival analyses were performed on three genes (*CLOCK*, *RAPGEF6*, and *ZDHHC20*), and the hazard ratio and *p*-value are shown in Supplementary Table 2 (available online). The analysis on the three genes revealed that the HER2-negative, ER-positive, and PR-positive groups had meaningful hazard ratios in both univariate and multivariate analyses. In the multivariate survival analysis, the survival rate tended to decrease with a higher expression of *CLOCK*, *ZDHHC20*, and *RAPGEF6*, with hazard ratios of 1.76, 1.54, and 1.51, respectively (Table 6).

DISCUSSION

In network analysis, the most important feature is the hub vertex distribution. The TNBN and TPBN gene regulatory networks both showed a scale-free characteristic. This means that, unlike a random network, the TNBN gene regulatory network could be easily destroyed if the hub vertices were attacked. Among the TNBN gene regulatory network hub vertices, the most interesting genes were *RAPGEF6* and *ZDHHC20*. Because these genes are cancer related, they were consistently observed in TNBNs only in the centrality analysis. Draper and Smith [12] have reported that *ZDHHC20* was associated with cellular transformation and cell proliferation, but to the best of our knowledge, its relationship to breast cancer has thus far not been fully elucidated. In this situation, *ZDHHC20* may be a targetable hub vertex in TNBNs. In addition, *RAPGEF6* is known to convert GDP into GTP in the Ras-related proteins Rap1 and Rap2, which are cell-junction related proteins [13]. Activated Rap1 interacts with JamA, Bag3, Afadin, Riam, and RapL to regulate cadherin and integrin, which are connected to the cell junction and extracellular matrix [14]. Thus, if we want to identify changes in cell-to-cell interactions in TNBNs, it is essential to study *RAPGEF6*.

CLOCK was at the top of all three centrality analyses of TPBN. Interestingly, in TNBNs, *CLOCK* had 5 degrees, with an eigenvector and betweenness centrality values of 0.003 and 215085.373, respectively. These results show that even though TNBN may not be affected by hormonal dysregulation, its oncogenic property may affect its genesis. The result that higher expression of *CLOCK*, *ZDHHC20* and *RAPGEG6* related to lower survival rate also support the importance of the *CLOCK*, *ZDHHC20*, and *RAPGEF6* genes in each network group. In addition, the *ASXL2*, *CCNT1*, and *NCOA2* genes were also frequently observed in the centrality analysis in both groups. These genes are well-known in tumorigenesis [15-17]. Thus, by conducting a thorough network analysis, we can find not only well-known genes but also genes that are not as well known in cancers. In fact, in the DEG analysis, except for *RAPGEF6*, we could find no other genes showing a meaningful *p*-value (Bonferroni) between the two groups (Supplementary Table 3, available online). In addition, we compared these genes using 100 normal and 1,084 cancer samples. The genes *CCNT1* and *ASXL2* showed a *p*-value of less than 0.001 (Bonferroni), whereas the other genes showed a *p*-value of 1. Therefore, the DEG analysis using cancer and normal samples suggests that the important genes found from the network analysis cannot be found in the DEG analysis (Supplementary Table 4, available online).

Through network analysis, we have attempted to understand the physiology of TNBNs. The TNBN and TPBN gene regulatory networks showed similar network statistics, with both having similar network densities, diameters, average vertex-to-vertex distance values, and scale-free network characteristics. However, the TNBN gene regulatory network was less clustered than the TPBN gene regulatory network, albeit showing a similar modularity. In addition, the hub vertices were different in both groups. Although we could not conduct specific analyses on each cluster in the TNBN gene regulatory network, we were able to find some oncogenes through the centrality analyses.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin

https://doi.org/10.4048/jbc.2017.20.3.240

2016;66:7-30.
2. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med 2010;7:e1000279.
3. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med 2010;363:1938-48.
4. Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med 2007;357:2666-76.
5. de Matos Simoes R, Emmert-Streib F. Bagging statistical network inference from large-scale gene expression data. PLoS One 2012;7:e33624.
6. Emmert-Streib F, de Matos Simoes R, Mullan P, Haibe-Kains B, Dehmer M. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. Front Genet 2014;5:15.
7. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113-20.
8. Han SW, Chen G, Cheon MS, Zhong H. Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. J Am Stat Assoc 2016;111:1004-19.
9. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Stat 2006;34:1436-62.
10. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
11. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p.397-420.
12. Draper JM, Smith CD. DHHC20: a human palmitoyl acyltransferase that causes cellular transformation. Mol Membr Biol 2010;27:123-36.
13. Gao X, Satoh T, Liao Y, Song C, Hu CD, Kariya KK, et al. Identification and characterization of RA-GEF-2, a Rap guanine nucleotide exchange factor that serves as a downstream target of M-Ras. J Biol Chem 2001; 276:42219-25.
14. Iwasaki M, Tanaka R, Hishiya A, Homma S, Reed JC, Takayama S. BAG3 directly associates with guanine nucleotide exchange factor of Rap1, PDZGEF2, and regulates cell adhesion. Biochem Biophys Res Commun 2010;400:413-8.
15. Moiola C, De Luca P, Gardner K, Vazquez E, De Siervi A. Cyclin T1 overexpression induces malignant transformation and tumor growth. Cell Cycle 2010;9:3119-26.
16. Munteanu AI. Genetic alterations in nuclear receptor coactivators in breast cancer [dissertation]. [Los Angeles, USA]: University of Southern California; 2010.
17. Park UH, Kang MR, Kim EJ, Kwon YS, Hur W, Yoon SK, et al. ASXL2 promotes proliferation of breast cancer cells by linking ER alpha to histone methylation. Oncogene 2016;35:3742-52.