Briefings in Bioinformatics, 22(4), 2021, 1–13

https://doi.org/10.1093/bib/bbaa268
Problem Solving Protocol

Sample-specific perturbation of gene interactions identifies breast cancer subtypes

Yuanyuan Chen, Yu Gu, Zixi Hu and Xiao Sun

Corresponding author: Xiao Sun, State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China. E-mail: xsun@seu.edu

Abstract

Breast cancer is a highly heterogeneous disease, and there are many forms of categorization for breast cancer based on gene expression profiles. Gene expression profiles are variables and may show differences if measured at different time points or under different conditions. In contrast, biological networks are relatively stable over time and under different conditions. In this study, we used a gene interaction network from a new point of view to explore the subtypes of breast cancer based on individual-specific edge perturbations measured by relative gene expression value. Our study reveals that there are four breast cancer subtypes based on gene interaction perturbations at the individual level. The new network-based subtypes of breast cancer show strong heterogeneity in prognosis, somatic mutations, phenotypic changes and enriched pathways. The network-based subtypes are closely related to the PAM50 subtypes and immunohistochemistry index. This work helps us to better understand the heterogeneity and mechanisms of breast cancer from a network perspective.

Key words: gene interaction network; individual-specific edge perturbations; network-based subtypes; breast cancer

Introduction

Breast cancer is a major public health issue mainly in women. Hundreds of thousands of women die of this disease each year. Breast cancer is a highly heterogeneous disease, and there are many forms of categorization, such as the classic grading system and the tumor–node–metastasis (TNM) classification. Among all methods, the immunohistochemistry (IHC) index has played a vital role in diagnostics, prognostics and therapy response prediction. The method based on histological classification and IHC-based marker selection has become the essential criterion that clinicians use for breast cancer typing [1]. Additionally, breast cancer samples are also classified on the basis of the expression of some biomarker genes. Microarray-based coding mRNA expression profiling has identified five 'intrinsic' subtypes (called PAM50 subtypes), including luminal-A, luminal-B, Human epidermal growth factor receptor 2 (HER2)-enriched, basal-like and normal-like [2, 3], which adds significant prognostic and predictive information to standard parameters for breast cancer patients. Recently, with advances in whole-transcriptome sequencing (RNA sequencing, RNA-seq), this classification has been refined to the identification of 12 breast tumor subgroups using the top 3662 variably expressed genes [4]. Meanwhile, some work has shown that long noncoding RNAs and microRNAs might play key roles in mammary tumor development

Yuanyuan Chen is an assistant professor at College of Science, Nanjing Agricultural University, Jiangsu, Nanjing, China, and a postdoc at State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.
Yu Gu is a PhD candidate at State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.
Zixi Hu is a PhD candidate at State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.
Xiao Sun is a professor at State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China.
Submitted: 27 July 2020; **Received (in revised form):** 9 September 2020

[5, 6], and consequently, breast cancer subgroups on the basis of lncRNAs have been studied in [7], which report its tight correlation compared with the PAM50-defined mRNA-based subtypes. There is strong heterogeneity not only in breast cancer but also in a single subtype of breast cancer. For example, the subtypes and corresponding treatment strategies of triple-negative breast cancer (TNBC) have been investigated in [8]. In addition, well-characterized and cancer-associated heterocellular signatures have been applied to reveal luminal-A breast cancer heterogeneity and to study differential therapeutic responses [9]. This analysis stratified the luminal-A breast cancer samples into five subtypes, with a majority of them belonging to one subtype (stem-like), which is enriched for stem and stromal cell gene signatures representing potential luminal progenitor origin. In the case of molecular typing, the TNBC and the normal-like subtypes have no unique biological markers for response to any specific drugs, which indicates that more precise tools are needed to improve its predictive, therapeutic and prognostic performance.

In terms of dynamics, the gene expression in a biological system is variable and may be different if measured at different time points or under different conditions, even for the same cell. In contrast, biological networks are relatively stable against time and condition [10, 11]. Macromolecular interaction networks can more reliably characterize the biological system or state of the tissue. Many network methods are based on biological pathways, which are focused on the inference of pathway activity by using pathway-specific genes [12, 13]. Application of the pathway approach in breast cancer research implicates the methodological means for the quantification of the pathway activity in individual tumors [14, 15]; hence, there are some pathway-targeted therapies in breast cancer. For example, pathway-targeted therapy by vascular endothelial growth factor signaling inhibitors may target the enhanced angiogenesis, proliferative signaling, invasion and metastatic properties of cancer cells [16]. Pathway-targeted therapies might confer a lower systemic risk of adverse effects by targeting only the specific disordered pathways [17].

To improve our understanding of breast cancer heterogeneity, we proposed a rank-based sample-specific gene interaction perturbation (named edge perturbation in the gene interaction network) method, where the gene interaction relations were derived from Reactome and other pathway and interaction databases [18]. Different from previous pathway-based approaches, this method utilizes not only the gene set information (nodes in pathways) but also and more importantly the interaction information (edges in pathways). The gene interactions in a biological network are overall stable in a particular type of normal human tissue but widely perturbed in diseased tissues [19, 20]. These perturbations in gene interactions (edge perturbations) in each sample can be measured by the change in the relative gene expression value. The edge perturbations at an individual level can be used to characterize the perturbation of the biological network for each sample efficiently. Then, an unsupervised cluster analysis of breast cancer based on the edge-perturbation matrix can be performed to reveal the heterogeneity among breast cancer patients (Figure 1). Our results suggest that the new network-based subtypes are significantly different in prognosis, somatic mutations, phenotypic changes (measured as scores in TCGA [21]) and enriched pathways. Moreover, our network-based subtypes correlated with the PAM50 subtypes and the IHC index. These findings will help us to understand the mechanisms of breast cancer carcinogenesis from an interactome perspective.

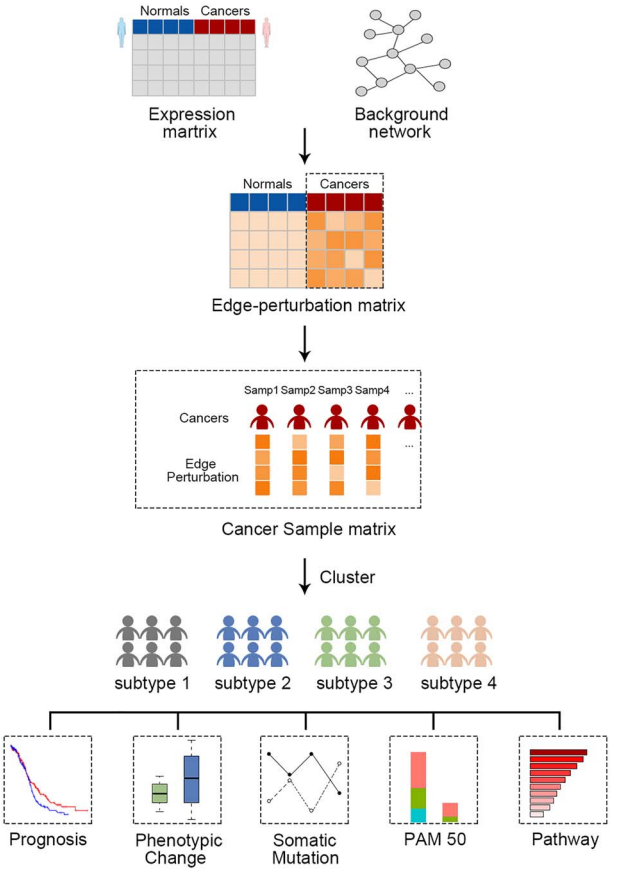

Figure 1. A framework identifying breast cancer subtypes. Gene interaction perturbations for each individual sample are measured by an edge-perturbation matrix, derived from the expression matrix and background network. The cancer sample matrix represents selected edge perturbations for breast cancer samples. Then, the breast cancer samples are clustered by using a partition edge-perturbation matrix to reveal new network-based subtypes. The identified subtypes are characterized from different aspects, including prognosis, phenotypic changes, somatic mutations, connection with PAM50 subtypes and enriched pathways.

Materials and methods

Data sources

Transcriptomics data

The gene interaction perturbation program takes RNA sequencing (RNA-seq) data and clinical data as input. We downloaded level three RNA-seq data in the form of FPKM and clinical data from TCGA data set https://portal.gdc.cancer.gov/ by the Data Transfer Tool. The expression data of 1093 breast cancer samples were assigned as the case group. For the control group, RNA-seq data of 290 normal breast tissues were obtained from the Genotype-Tissue Expression (GTEx) database (https://gtexportal.org/). GTEx, an auxiliary TCGA data mining project, is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression. The samples from GTEx were all from normal tissues but not tissues adjacent to carcinoma, as in TCGA, which can avoid confusion with the tumor tissues. The two data sets were both converted to TPM form with 33 562 genes in total, which were prepared for further analysis. In addition, somatic mutations in 127 significantly mutated genes of 772 breast cancer tumors in TCGA were obtained from the

mutational landscape study in [30]. The independent validation data were downloaded from GEO DataSets (GSE3494), which contains 251 expression profiles of breast cancers by microarray.

Background network

A pathway-based analysis is needed for projecting candidate genes onto protein functional relationship networks in the study. We intended to acquire a human functional protein interaction network derived from pathways. The Reactome Pathway Database is a great option [22]. The core unit of the Reactome data model is the reaction. Entities (nucleic acids, proteins, complexes, anticancer therapeutics and small molecules) participating in reactions form a network of biological interactions and are grouped into pathways. ReactomeFIPlugIn app can help us access the Reactome Functional Interaction network, a highly reliable, manually curated pathway-based network by extending curated pathways with noncurated sources of information, including protein–protein interactions, gene coexpression, protein domain interaction, Gene Ontology annotations and text-mined protein interactions, which cover over 50% of human proteins [18]. We downloaded all the gene interaction networks (224 in total) of Reactome pathways by using the app reactomeFIPlugIn in Cytoscape 3.5.1 [23]. The prepared interaction networks are in the form of three columns. The first column is the name of a pathway, and the other two columns are the coding genes connected in the pathway. All the networks were integrated into a large network as the background network with 169 710 edges in total (see the Supplementary Data).

Overview of the edge-perturbation-based approach

The motivation for our edge-perturbation-based approach stems from the following idea. The lesion extent, which reflects the physiological status of an individual with disease, can be measured by the perturbation degree of the background network. Furthermore, the perturbation of the background network is essentially the result of the gene interaction changes occurring in the network. The edge perturbation in the gene interaction network can ultimately characterize the interaction change between two genes. Hence, the overall edge perturbations of all gene pairs in the background network can be reasonably used to reveal the pathological condition of an individual with disease. To measure the perturbation in the whole background network at an individual sample level, there are three major steps in our method (see the flowchart in Figure 2): transfer the gene expression matrix into a gene expression rank matrix; calculate the delta rank matrix and construct the edge-perturbation matrix, which is used to measure the gene interaction perturbations for each sample. As an effective quantization of the sample-specific gene interaction perturbation, the edge-perturbation matrix will be converted to a cancer sample matrix that is used for subsequent clustering analysis.

Construction of the edge-perturbation matrix

First, each gene expression value was converted into its rank within each sample (the smallest expression value corresponds to the minimum rank, and the largest expression value corresponds to the maximum rank). As a result, the expression matrix was transformed to a rank matrix (denoted by R with element $r_{i,s}$, which represents the rank of gene g_i in sample s) by ranking all genes according the expression values in all samples. We then calculated the delta rank matrix whose rows and columns represented edges in the background network and samples, respectively. An element $\delta_{e,s}$ (delta rank) in the delta rank matrix was calculated by subtracting the ranks of the two genes connected by an edge ($\boldsymbol{e}$) in the background network (Equation 1).

$$\delta_{e,s} = r_{i,s} - r_{j,s} \tag{1}$$

where genes g_i and g_j are connected by edge e in the background network.

Gene–gene interaction profiling shows high conservation in normal samples, and there are few interaction perturbations [24]. The within-sample delta ranks of gene pairs are highly stable among samples under normal conditions but are often widely disrupted after certain treatments, such as gene knockdown, gene transfection, drug treatment and tissue canceration [19]. Therefore, we could assume that the background network system is stable across all normal samples. We ranked genes according to their mean gene expression value among normal samples and similarly calculated the delta rank as the benchmark delta rank vector with elements denoted by $\overline{\delta}_e$, where e is an edge in the background network. This vector measures the mean relative ranks of gene pairs in all normal samples. Each sample should be compared with the benchmark vector, and the corresponding difference represents the gene interaction perturbations on the sample. Upon subtracting the benchmark delta rank vector from the delta rank of each sample, we finally obtained the edge-perturbation matrix Δ with element Δ_{es} (Equation 2). For an edge e in the background network and an individual sample $\boldsymbol{s}$,

$$\Delta_{e,s} = \delta_{e,s} - \overline{\delta}_e. \tag{2}$$

The edge-perturbation matrix can measure the sample-specific interaction perturbation in the same whole background network effectively. Each column of the edge-perturbation matrix represents the gene interaction perturbations for an individual sample, i.e. the sample-specific perturbation of the gene interaction.

Discovery and validation of the network-based subtypes

We selected the clustering features based on two aspects: the ability of the selected features to distinguish breast cancer samples from normal samples easily and that they can also maintain heterogeneity within breast cancer samples. First, we calculated the difference between breast cancer samples and normal samples for each edge in the edge-perturbation matrix by using the Kruskal–Wallis test. The top 30 000 significantly different edges (approximately 20%) were selected. Next, the SDs of the edge perturbations of all breast cancer samples were calculated. We also selected the top 30 000 edges with high SDs. Then, the cancer sample matrix could be obtained by selecting the edges in the intersection of the above two sets with 30 000 edges over all cancer samples in the edge-perturbation matrix, which would be used for clustering analysis. The columns of the cancer sample matrix represent cancer samples, and the values in a column are the perturbation degree on each feature edge for an individual cancer sample.

Furthermore, we extrapolated the network-based subtypes for the TCGA breast cancer samples using the consensus clustering method [25], which was performed with the R package *ConsensusClusterPlus* by subsampling a proportion of items and features from the cancer sample matrix.

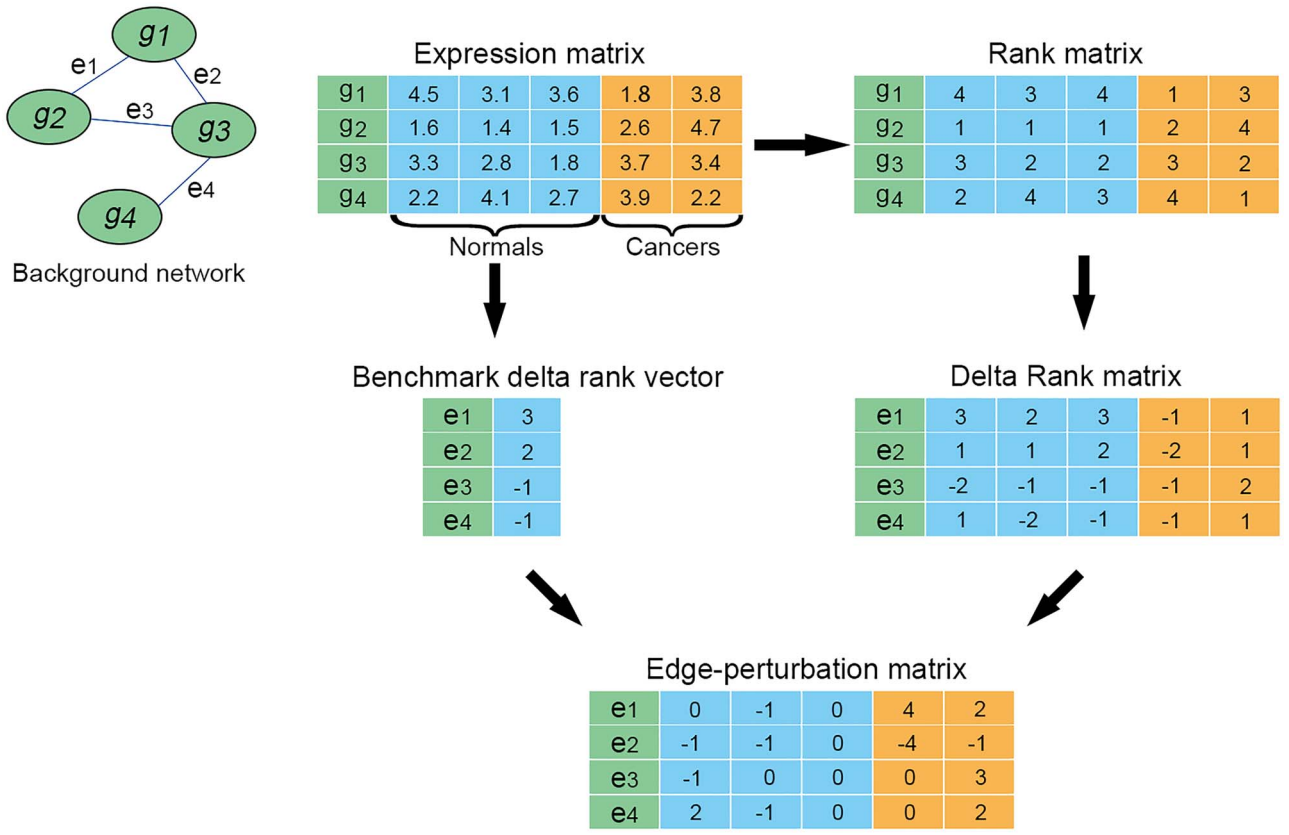

Figure 2. Flowchart of the edge-perturbation-based method. The background network consists of four genes and four edges. There were three normal samples (blue) and two cancer samples (orange). A rank matrix was obtained by ranking the genes according to the expression value of each sample. The rank matrix was converted to a delta rank matrix with four rows and five columns representing edges and samples, respectively. The benchmark delta rank vector was calculated as the delta rank of the mean expression value in normal samples. The edge-perturbation matrix was obtained by subtracting the benchmark delta rank vector from the delta rank matrix.

To confirm the clusters of breast cancer samples in TCGA based on the edge perturbations, we independently applied the same analysis procedure on the validation data set. In addition, we used in-group proportion (IGP) [26] to measure the cluster consistency. IGP can be used to evaluate the reproducibility of the clusters derived from two independent data sets by providing a quantitative value to measure the similarity between the clusters. IGP will be 100% if the clusters are identical between two data sets and will be 0% conversely. Because the expression data generation methods are different in these two data sets, specifically, one based on RNA-seq and the other based on microarray, the edge-perturbation values were normalized to Z-score prior to the IGP analysis. The IGP analysis and the prediction of PAM50 subtypes were performed by using R packages *clusterRepro* and *genefu*, respectively.

Identifying subtype-specific pathways

The cancer sample matrix was normalized by the Z-score method, which scaled the mean of each row (corresponding to feature edge) to zero and variance to one. First, the rows of the matrix were clustered using hierarchical clustering based on the complete linkage method with the cluster number set to 100, and clusters containing more than 30 edges were retained. We then computed the mean values of perturbation for each edge in each subtype through Z-scores. For each subtype, we counted the percentage of edges whose absolute value of the average perturbation was greater than 0.5 in each retained cluster. A cluster with a percentage greater than 70% was regarded as a perturbed cluster for this subtype. All edges in all of the perturbed clusters for each subtype constituted the subtype-specific networks. All genes involved in each subtype-specific network were used for pathway enrichment analysis by Metascape (http://metascape.org). The KEGG and Reactome pathways with a *P*-value less than 0.01 were retained. Finally, the subtype-specific pathways were identified.

Results

The constructed networks

The initial background network from the Reactome database was composed of 169 710 edges and 7360 genes in total. After filtering out genes that were not in the expression data, the background network was decreased to having 161 276 edges and 7074 genes and was then used to calculate the edge-perturbation matrix. Both the initial background network and the filtered network used in this study are scale free, which means that the fraction of nodes with degrees follows a power law distribution. Supplementary Figure S1**A** and **B** illustrate the degree distributions of the two networks, and the determination coefficients R^2 are 0.701 and 0.687, respectively. Here, R^2 is used to measure the fitting level of the power law curve. The better the curve fitting level is, the closer R^2 is to 1. Both the degree distribution figures

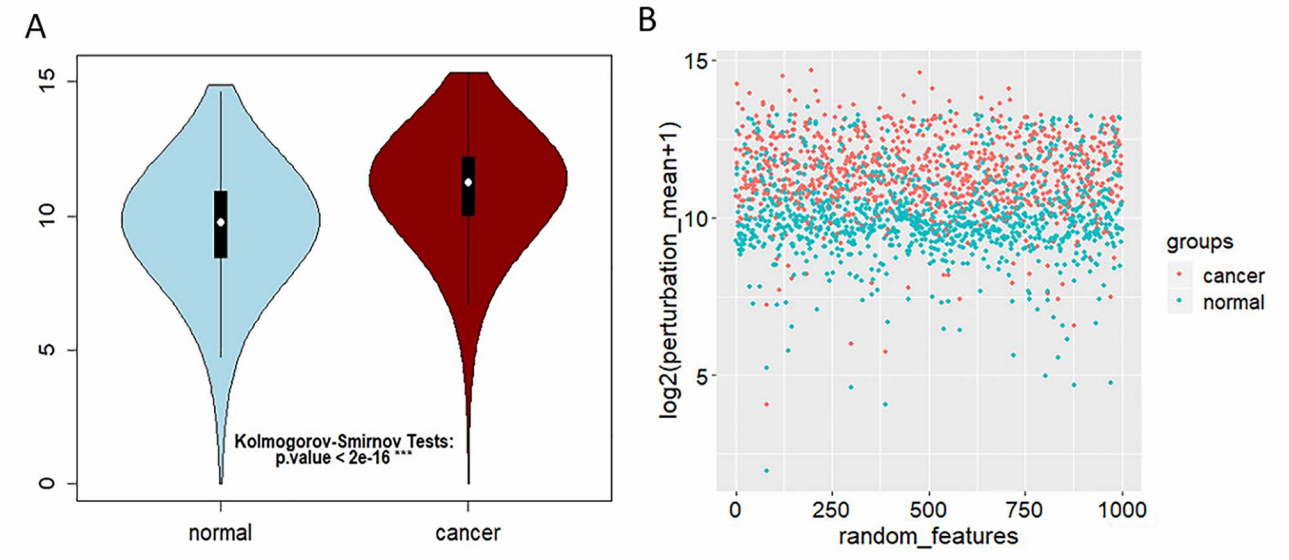

Figure 3. Perturbation of gene interactions in normal and breast cancer tissues. (**A**) Distribution of log2-transformed edge perturbations in both normal and cancer samples. Violin plots show the distributions of the edge perturbations of 1000 randomly selected edges in the edge-perturbation matrix in both the cancer and normal groups. The distributions in these two groups were significantly different, as assessed by the Kolmogorov–Smirnov test. (**B**) The scatterplot for the log2-transformed mean of the edge perturbations in the 1000 randomly selected edges in both normal (blue points) and breast cancer (red points) tissues. The edge perturbations of normal samples are much denser and less than those of cancer samples.

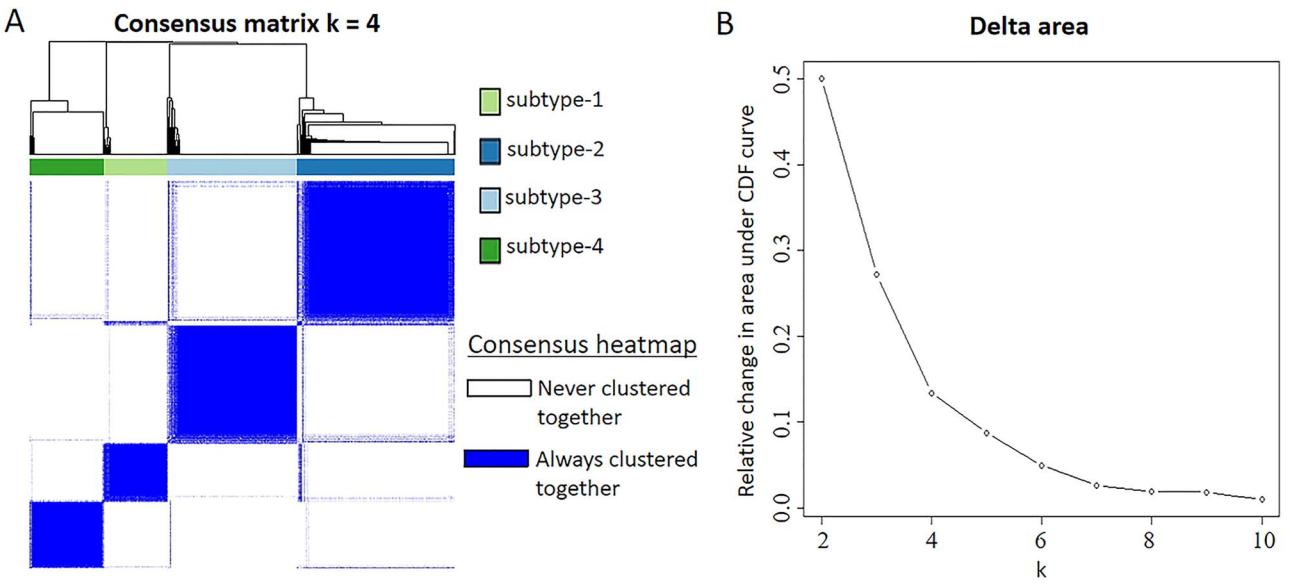

Figure 4. Unsupervised consensus clustering of network-based subtypes. (**A**) Consensus matrix heatmap of the chosen optimal cluster number ($k = 4$) for the 1093 TCGA breast cancer samples. The rows and columns represent patient samples, and consensus matrix values range from 0 in white (meaning that patients are never clustered together) to 1 in dark blue (meaning that patients are always clustered together). (**B**) The delta area plot for k changed from 2 to 10. The vertical axis is the relative change in the area under the CDF curves when the cluster number varies from k to $k+1$. The range of k changed from 2 to 10, and the optimal $k = 4$.

and the determination coefficients show that the networks used in this study were all scale free.

Stable gene interaction in normal breast tissues

Both 290 normal samples from GTEx and 1093 breast cancer samples from TCGA were used to evaluate the stability of the edge perturbation in normal samples and variability in cancer samples, as well as the difference between them. The edge-perturbation matrix with 161 276 rows was constructed by the edge-perturbation-based method (see the Materials and methods section for details).

Zero center normalization was performed on the delta rank matrix by Equation (2), which was used to deduce the edge-perturbation matrix. The edge-perturbation matrix can measure the sample-specific perturbation in the same background network effectively. For a given gene pair, the greater the absolute value in the edge-perturbation matrix is, the greater the perturbation is. The mean absolute magnitude of the edge perturbations in normal samples was 1692.3, and cancer samples

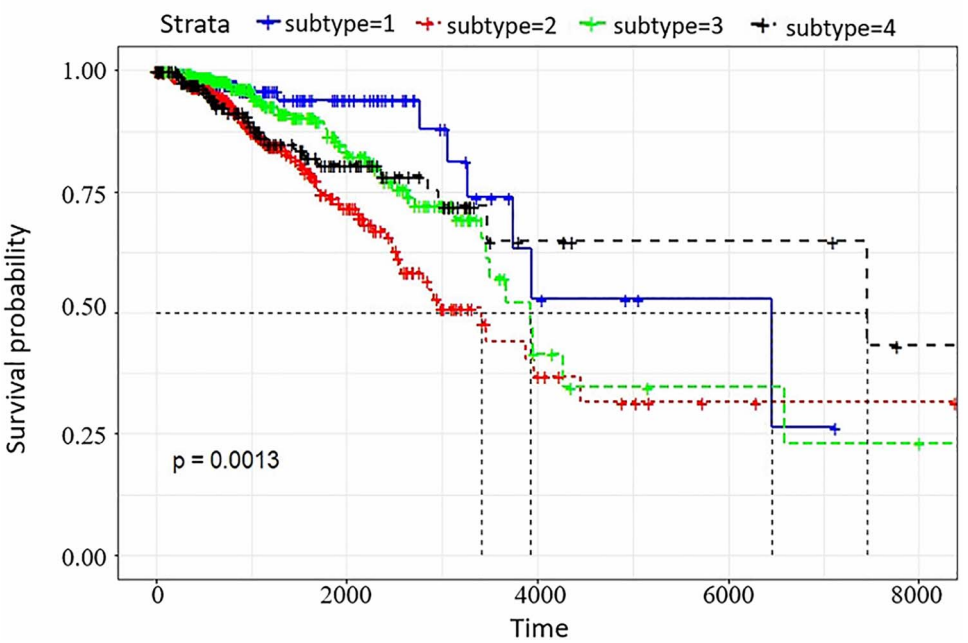

Figure 5. Survival curves of the network-based subtypes. Kaplan–Meier plot of survival for the four network-based subtypes in 1087 breast cancer samples from TCGA with prognosis information. The horizontal axis represents the survival time (days), and the vertical axis is the probability of survival. The log-rank test was used to assess the statistical significance of the differences in prognosis among the four network-based subtypes.

doubled as expected. Furthermore, 90.4% of all 161 276 gene pairs showed more dispersion in cancer samples than in normal samples by comparing the sum of the edge-perturbation degrees. In addition, we selected 1000 features randomly from all the gene interaction features, and the Kolmogorov–Smirnov test was performed. The edge-perturbation distributions of the 1000 selected features in the normal and cancer groups were significantly different, with $P < 2e{-}16$. We plotted the distribution of edge-perturbation amplitude as $\log_2(|\Delta_{es}| + 1)$for both normal and cancer samples, as shown in Figure 3**A**. To clearly show the difference in the edge-perturbation distribution between normal and cancer samples, the mean edge-perturbation amplitude, as well as a similar log2 transformation of the 1000 selected features, was plotted in a scatter plot, as shown in Figure 3**B**. The edge perturbation of normal samples (blue points) is much denser and less than that of cancer samples (red points). These two plots reveal that the edge perturbations of normal samples are more stable, whereas a wider variation exists in cancer samples, making it possible to find the heterogeneity in breast cancer samples through the edge-perturbation matrix of all samples.

Network-based subtypes

The edge-perturbation matrix was converted to a cancer sample matrix, which was used for the clustering of breast cancer samples. The rows of the cancer sample matrix are 1911 edges. These edges form a network with 1461 genes, which was visualized in Supplementary Figure S2, and the corresponding determination coefficient R^2 is 0.739, which means that it is also a scale-free network.

Consensus clustering was performed using the *ConcensusClusterPlus* package in R [25] to explore the subgroups of breast tumors based on the cancer sample matrix. Consensus matrix heatmaps and delta area plots, which can be found in Figure 4**A** and B, respectively, were drawn to determine k, the optimal number of clusters. The consensus matrix is a better visualization tool to help assess the clustering number. The matrix is arranged so that the samples belonging to the same cluster are adjacent to each other. A color gradient of 0–1 is used, with dark blue corresponding to a consensus score of 1 and white corresponding to a consensus score of 0. The color-coded heatmap corresponding to the consensus matrix obtained by applying consensus clustering to these cases is shown in Figure 4**A**. The heatmap represents the consensus for $k = 4$ and accordingly displays a well-defined four-block structure. The four blocks are almost disjoint in the heatmap, which means that the four clusters are distinguishably clustered. The delta area plot in Figure 4**B** shows the relative change in the area under the cumulative distribution function (CDF) curve comparing k and $k - 1$ (k ranges from 2 to 10). The k at which there is no appreciable increase in consensus can be considered as an optimal cluster number. The four-cluster solution corresponded to the number with no large increase (approximately 0.1). Thus, the optimal cluster number was set to 4. Of the 1093 breast cancer samples analyzed in this study, 162 were subtype-1, 407 were subtype-2, 334 were subtype-3 and 190 were subtype-4. Next, we used the four network-based subtypes mentioned above for further analysis.

We independently applied the same analysis procedure on the validation data set to confirm the clustering consistency with the TCGA cohort. In total, 1536 features overlapped with the 1911 feature edges in the TCGA cohort were used to perform unsupervised consensus clustering. Interestingly, we observed that the samples in the validation cohort were also clustered into four optimal clusters (Supplementary Figure S3), which is very similar to that identified in the TCGA data set (Figure 4). The IGP values are 88.3, 76.7, 95.6 and 86.2% for subtype-1, subtype-2, subtype-3 and subtype-4, respectively, indicating that all subtypes show high consistency between the two data sets. This suggests that these four network-based subtypes are robust across different data sets of breast cancer.

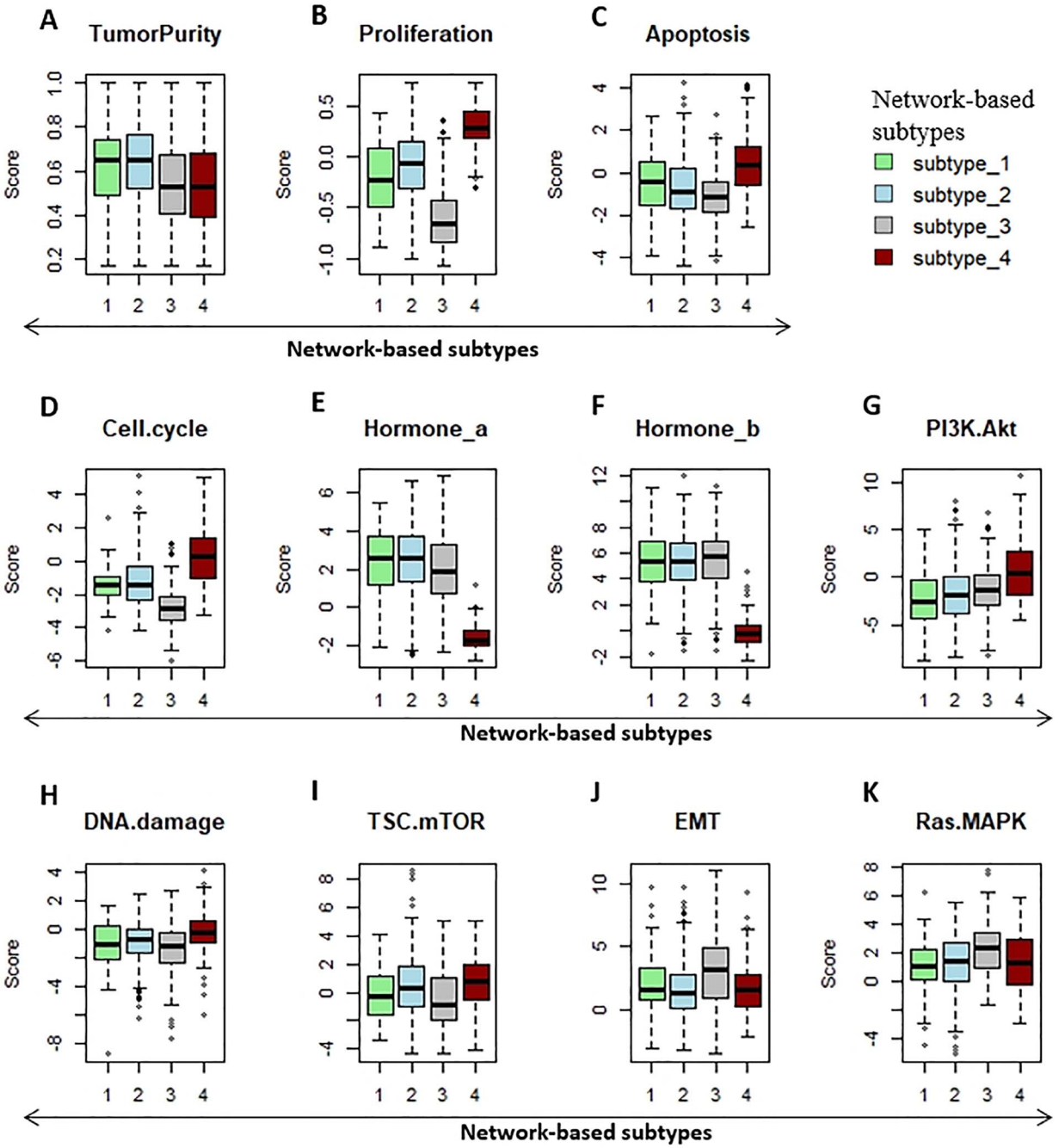

Figure 6. Phenotype heterogeneity among the network-based subtypes. Boxplots show differences in (**A**) tumor purity, (**B**) proliferation, (**C**) apoptosis, (**D**) cell cycle, (**E**) hormone_a, (**F**) hormone_b, (**G**) PI3K/AKT, (**H**) DNA damage response, I TSC-mTOR, J EMT and K Ras/MAPK scores from TCGA among network-based subtypes. The data from (**A**) were derived from ABSOLUTE in [27], which infers tumor purity and malignant cell ploidy directly from the analysis of somatic DNA alterations. The data from **B–K** were from RPPA data-based scores published by TCGA. The Kruskal–Wallis test was performed to calculate the *P*-value, and those associations with *P*-value < 0.01 were considered significant. EMT = epithelial–mesenchymal transition.

Heterogeneity among network-based subtypes

Prognosis

We compared the prognosis differences among the network-based subtypes. Kaplan–Meier survival analysis indicated that the differences in survival among the subtypes were significant ($P = 0.0013$, Figure 5). Subtype-2 has the worst prognosis compared with other subtypes, whereas subtype-1 portends a more favorable prognosis with a 5-year survival probability above 85%. In addition, the survival curves of five PAM50 subtypes are shown in Supplementary Figure S4. The Kaplan–Meier survival analysis indicates that the differences in survival among PAM50

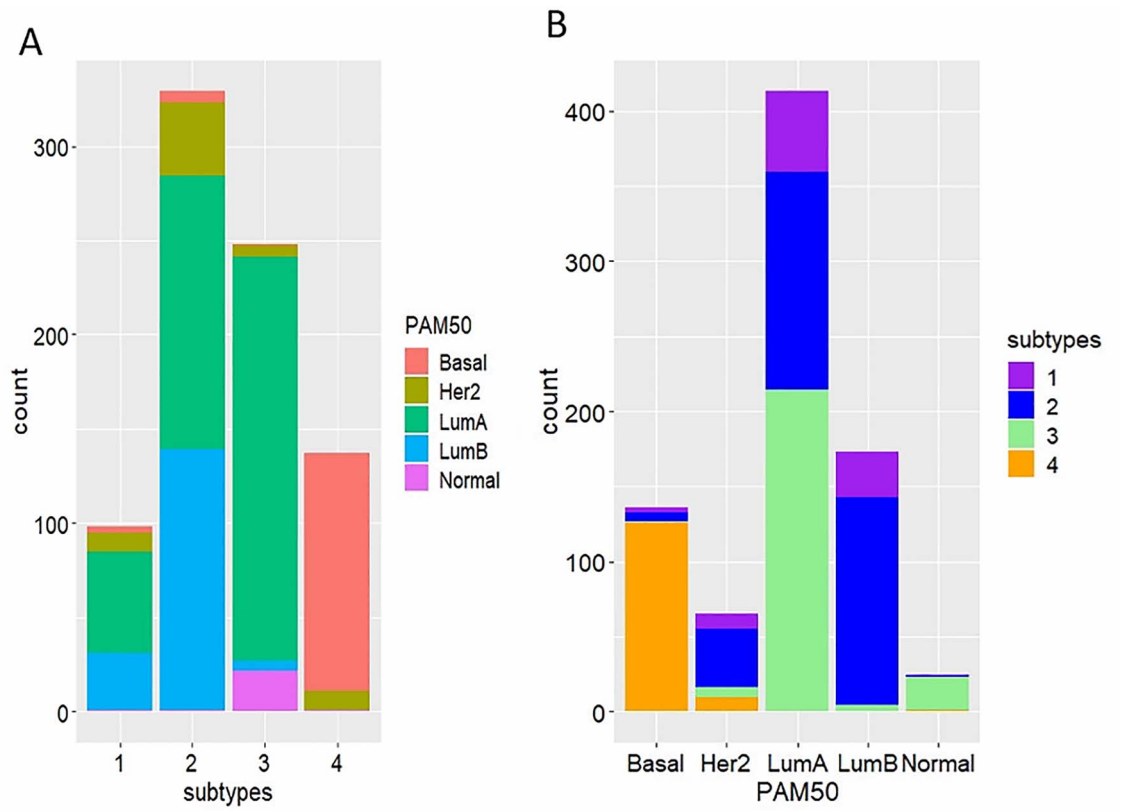

Figure 7. Comparison of the network-based subtypes and their PAM50 subtypes. (**A**) The distribution of the PAM50 subtypes in each of network-based subtype. (**B**) The distribution of the network-based subtypes in each of PAM50 subtype.

subtypes are significant ($P = 0.04 < 0.05$) by log-rank test, whereas the differences in survival among our four subtypes are more significant.

Phenotypic heterogeneity

The tumor purity scores in Figure 6A were derived from the computational method (ABSOLUTE) in [27, 28], which infers tumor purity and malignant cell ploidy directly from the analysis of somatic DNA alterations. ABSOLUTE can detect subclonal heterogeneity and somatic homozygosity and calculate statistical sensitivity to reveal specific aberrations. Our analysis shows that the tumor purity scores are significantly higher in both subtype-1 and subtype-2 than in subtype-3 and subtype-4.

Next, we sought to investigate whether phenotypic changes show differences among our network-based subtypes in breast cancer tumors (Figure 6B–K). The pathway scores, which are protein expression signatures of pathway activity, associated with tumor lineage (Figure 6B–K) were from a reverse-phase protein microarray (RPPA) as published by TCGA [21]. Our analysis implies that the pathway scores for proliferation, apoptosis, cell cycle, PI3K/Akt signaling, DNA damage response and TSC-mTOR were significantly higher in subtype-4 than in other subtypes. However, the pathway scores for hormone-a, hormone-b (representing signatures associated with hormone receptors [29]) and Ras.MAPK (Ras GTPase /MAP kinase signaling) were significantly lower in subtype-4. On the other hand, the pathway scores for proliferation, apoptosis, cell cycle and TSC-mTOR were significantly lower in subtype-3. All phenotypes from the TCGA data set were significantly associated with the network-based subtypes except for the receptor tyrosine kinase scores (Supplementary Table S1). These results suggest that the network-based subtypes show differences in most breast cancer-associated phenotypes.

Connection with PAM50, the IHC index and the TNM stage

The PAM50 subtypes, known as 'intrinsic' subtypes of breast cancer (including basal-like, luminal-A, luminal-B, HER2-enriched and normal-like), have been identified and intensively studied [2]. There were close relationships between our four network-based subtypes and the PAM50 subtypes. Specifically, basal-like tumors made up a significant share (96.2%) of subtype-4 tumors and most subtype-3 tumors were luminal-A (more than 86%). Subtype-2 was a mixed subtype mainly including luminal-A, luminal-B and normal-like (accounted for 44.1%, 41.9% and 11.9%, respectively). Subtype-1 mainly contained luminal-A and luminal-B (55% and 31%). Conversely, samples in luminal-B and Her2 were mainly from subtype-2, and the ratios of subtype-2 were 80% and 60%, respectively. Luminal-A is a mixed subtype that manly includes subtype-1, subtype-2 and subtype-3 (Figure 7).

Interestingly, we found a similar relationship between the network-based subtypes and the PAM50 subtypes in the validation dataset (Supplementary Figure S5). Similar to the relationship in the TCGA data set, subtype-1 mainly contained luminal-A and luminal-B (41.67% and 36.67%), and samples in subtype-2 were mainly from luminal-A and luminal-B (16.28% and 60.47%). Most subtype-3 tumors (74.44%) were luminal-A (Supplementary Figure S5A). Conversely, subtype-4 tumors made up a significant share (92%) of basal-like tumors, and most luminal-A tumors were subtype-3 (65.69%). Luminal-B tumors mainly contained

Table 1. Contingency table for the network-based subtypes and the IHC indexes

Clinical factors	Subtype-1 ($n=162$)	Subtype-2 ($n=407$)	Subtype-3 ($n=334$)	Subtype-4 ($n=190$)	P-value
ER status	($n=94$)	($n=301$)	($n=238$)	($n=130$)	<2.2e−16
ER+	84	271	223	11	
ER−	10	30	15	119	
PR status	($n=94$)	($n=301$)	($n=236$)	($n=129$)	<2.2e−16
PR+	70	233	203	6	
PR−	24	68	33	123	
HER2 status	($n=67$)	($n=203$)	($n=169$)	($n=91$)	2.90E−10
HER2+	16	76	15	13	
HER2−	51	127	154	78	

Note. The figures in this table are either the number of patients or the P-value of Pearson chi-square test.

subtype-1 (31.43%) and subtype-2 (37.14%) tumors, and normal-like tumors mainly (81.25%) contained subtype-3 tumors (Supplementary Figure S5B).

Furthermore, we explored the relationship between the network-based subtypes and IHC indexes, including estrogen receptor (ER), progesterone receptor (PR) and HER2 receptor (HER2) status, by performing Pearson chi-square test (Table 1). ER status, PR status and HER2 status were all significantly different among the network-based subtypes ($P < 0.001$). Of 589 ER-positive tumors analyzed in this study (Table 1), 46% were subtype-2 and 38% were subtype-3. Conversely, of 174 ER-negative tumors, 68% were subtype-4 and 17% were subtype-2. The majority of PR-positive tumors were subtype-2 (46%) and subtype-3 (40%). The majority of PR-negative tumors were subtype-4 (50%) and subtype-2 (27%). The majority of HER2-positive tumors were subtype-2 (63%). In addition, we found that our network-based subtype-4 tended to be ER-negative, PR-negative and HER2-negative, which corresponded to TNBC. Both subtype-1 and subtype-3 tended to be ER-positive, PR-positive and HER2-negative. Though the network-based subtypes have different distributions in ER, PR and HER2 statuses, these two classifications of breast cancers have a close relation.

For the stage factor, we used the cutoff defined in [30] to distinguish advanced breast from early breast cancer: the TNM classification, with stages i and ii defined as early cancer and stages iii, iv and x defined as advanced cancer. These two groups were significantly different among the network-based subtypes by Fisher test ($P = 0.004155$) in Supplementary Table S3.

Connection with somatic mutations

The progressive accumulation of somatic mutations over time in crucial oncogenes or tumor-suppressor genes has been implicated in many cancer types [31]. Recently, the somatic mutation statuses of 127 genes have been shown to have significant effects on breast cancer survival [32]. With the identification of network-based subtypes using the edge-perturbation-based method, the question arises as to whether the somatic mutations occurring in cancer driver genes are significantly different among these subtypes. To answer this question, for each gene, we calculated the mutation ratios for each subtype based on the mutation status of each sample. Then, the mutation ratios were multiplied by 100 to construct a one-dimensional contingency table. A chi-square goodness-of-fit test was performed on the contingency table to see whether the mutation probabilities were significantly different with respect to the network-based subtypes. Genes with a $P < 0.05$ are shown in Supplementary Table S2. Figure 8 shows the mutation ratios of these genes in the four network-based subtypes. *TP53* is a tumor suppressor transcription factor with paramount clinical value because of its ability to regulate cell division by keeping cells from growing and dividing (proliferating) at an excessive rate or in an uncontrolled way. *TP53* is related to tumor progression, metastatic potential, early relapse and response to chemotherapy and ultimately has an impact on prognosis and survival [33–36]. Our analysis found that the mutation ratios of *TP53* were the most significantly different among the four network-based subtypes. In addition, the mutation ratios of other genes, such as *PIK3CA*, *CDH1*, *Gata3*, *MAP3k1*, *BRCA1* and *MAP2k4*, were significantly different. Mutations in *PIK3CA*, *Gata3* and *CDH1* are associated with the invasion and metastasis of breast cancer. *MAP3k1* is a member of the family of mitogen-activated protein kinases that regulates the apoptosis, survival, migration and differentiation of cells. As a tumor suppressor gene, tumors with *BRCA1* mutations have a higher risk of developing breast cancer. These results demonstrate that our network-based subtypes are linked to mutations in these cancer-related genes.

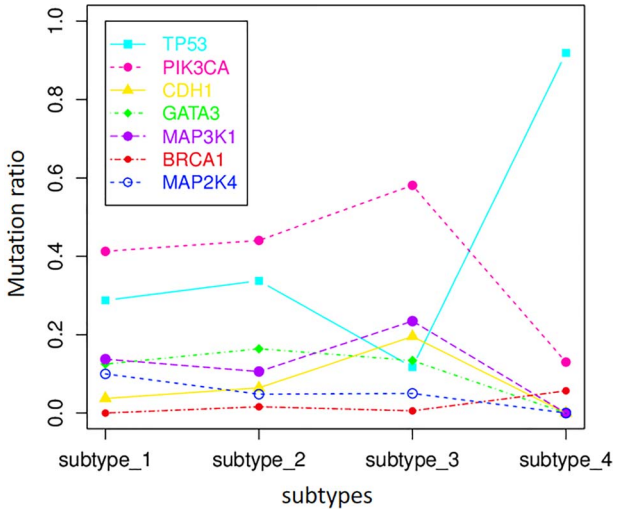

Figure 8. The differences in somatic mutations among the network-based subtypes. The genes with significantly different mutation ratios were obtained by the chi-square goodness-of-fit test (P-value <0.05). The vertical axis represents the mutation ratios of the selected genes among the four network-based subtypes.

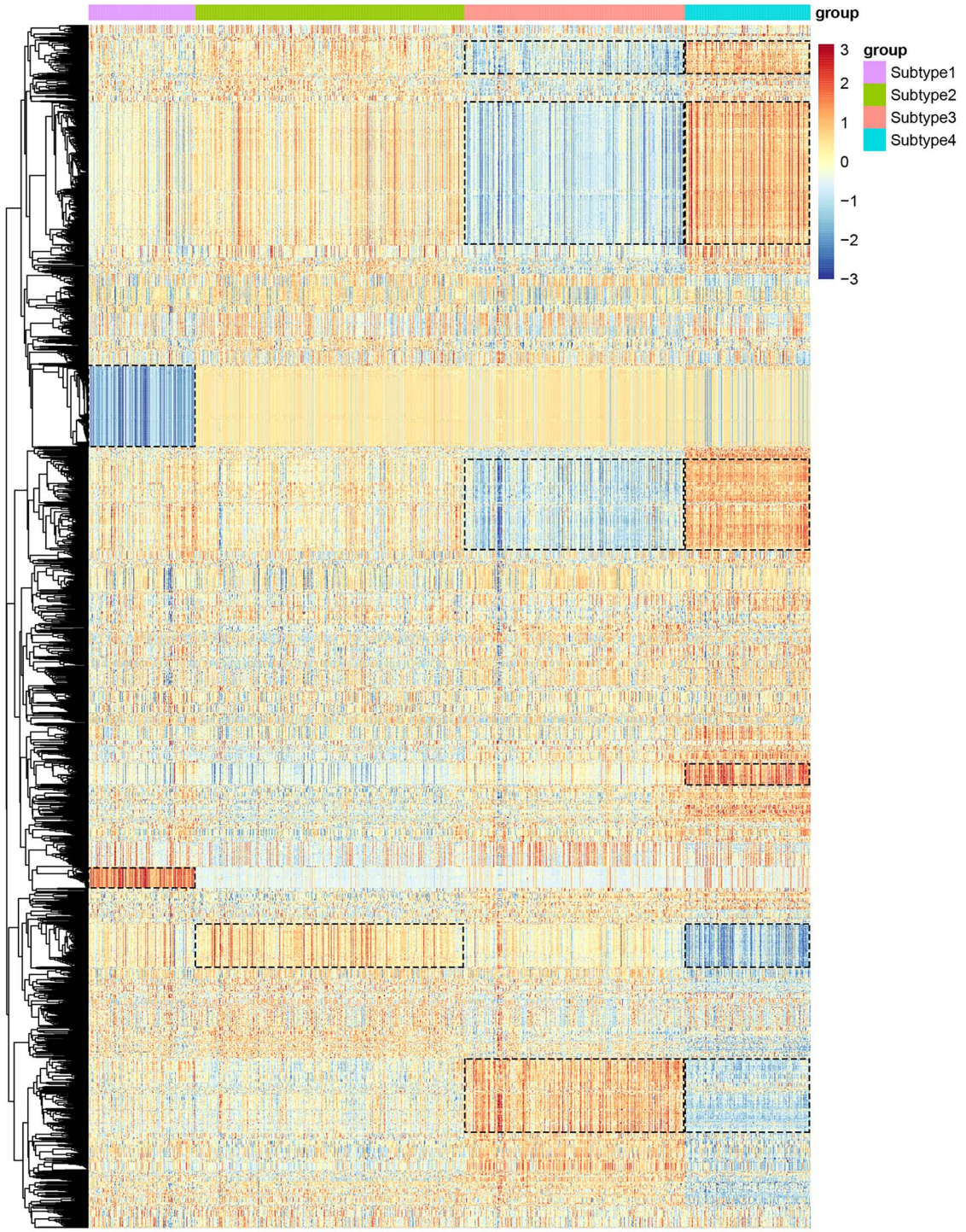

Figure 9. Clustering of the edge features of the cancer sample matrix. Every row in the matrix corresponds to one of 1911 edges; every column corresponds to one of 1093 breast cancer samples. Each row was Z-score normalized. The color bars at the top indicate the network-based subtypes. The blue color represents a negative perturbation, and the red color represents a positive perturbation. The black dotted boxes represent the identified blocks in the four subtypes.

Subtype-specific pathways

The heat map of edge clustering is shown in Figure 9. There are two, one, four and six perturbed clusters in subtype-1, subtype-2, subtype-3 and subtype-4, respectively. These perturbed clusters can form red or blue blocks in the corresponding subtypes in the heatmap, representing positive or negative perturbation patterns, respectively. It is rather remarkable that all the blocks in subtype-3 were also identified simultaneously in subtype-4 but in the opposite perturbation direction. For example, the first, second and third blocks are all blue in subtype-3 but

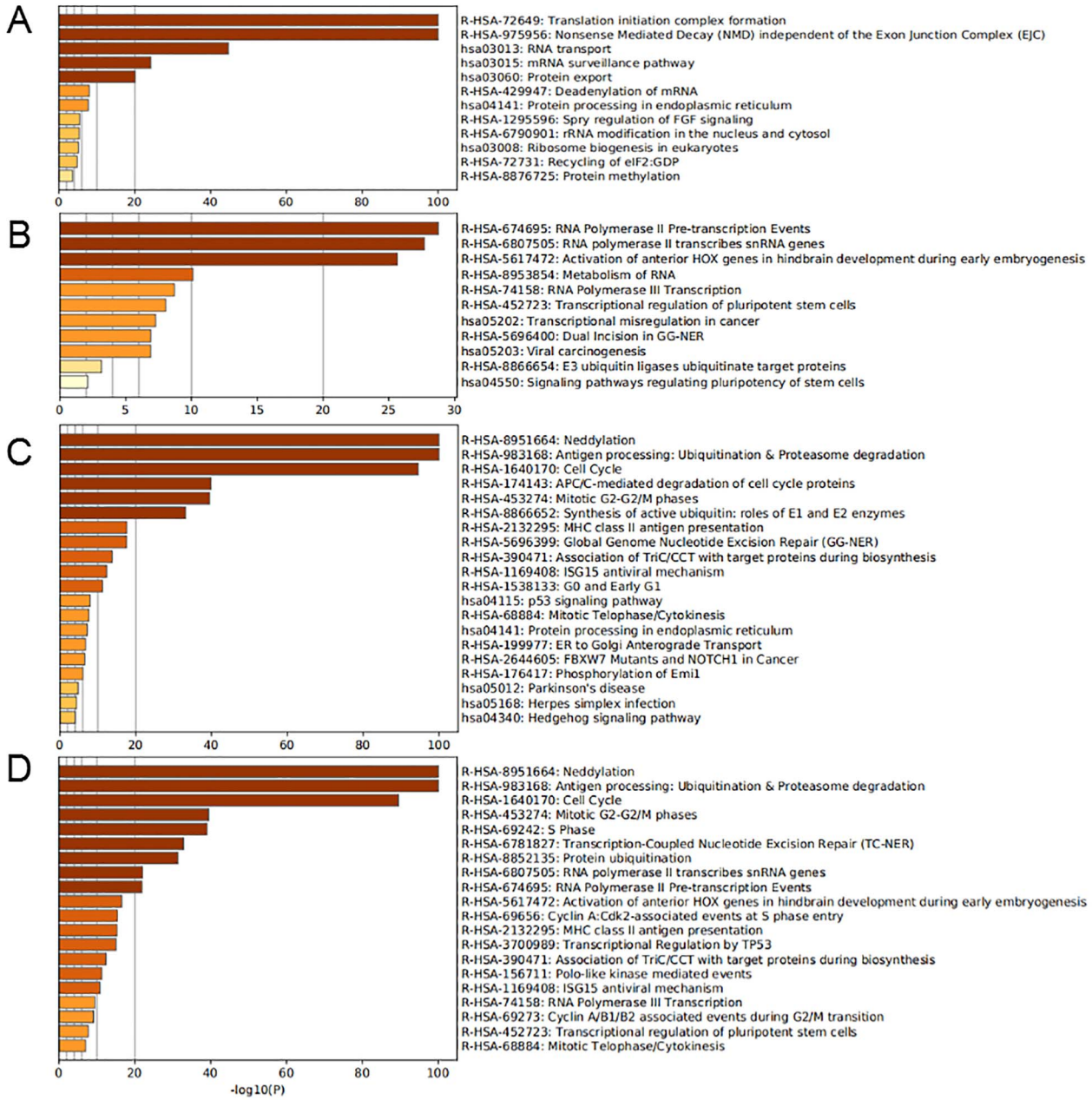

Figure 10. Subtype-specific pathways. (**A**–**D**) are pathways enriched in subtype-1, subtype-2, subtype-3 and subtype-4, respectively. The horizontal axis represents the negative log (base 10) of the *P*-value.

are red in subtype-4. The fourth block is red in subytpe-3 but blue in subtype-4. Therefore, the four blocks in subtype-3 can be used to distinguish subtype-3 and subtype-4 directly. In view of the findings above, it is reasonable to assume that the occurrences of subtype-3 and subtype-4 are due to the same disorder mechanism but in different directions. Furthermore, the expression of the ZBTB16 gene, which has the highest degree in the subtype-specific networks, was significantly different in subtype-3 and subtype-4 by the Wilcoxon rank-sum test (Supplementary Figure S6). Mainly involved in pathways including antigen processing, ubiquitination and proteasome degradation and neddylation, ZBTB16 is likely to be a substrate-recognition component of the E3 ubiquitin–protein ligase complex, which mediates the ubiquitination and subsequent proteasomal degradation of target proteins. In addition, it is interesting to note that subtype-3 tends to be ER-positive (94%) and PR-positive (86%), whereas subtype-4 tends to be ER-negative (92%) and PR-negative (95%).

The subtype-specific pathways can be found in Figure 10. Most pathways enriched in subtype-1 and subtype-2 are related to genetic information processing, such as translation initiation complex formation, RNA transport, mRNA surveillance pathway, protein export, RNA polymerase II transcribes snRNA genes, metabolism of RNA and so on. In addition, spry regulation of Fibroblast growth factors (FGF) signaling enriched in subtype-1 is closely associated with breast cancer. The E3 ubiquitin ligases ubiquitinate target proteins, viral carcinogenesis and transcriptional misregulation in cancer pathways are enriched

in subtype-2. Neddylation, which has been shown to be closely related to the activation state and ER-expression in breast cancer [37], is one of the enriched pathways in both subtype-3 and subtype-4. Some immune-related pathways are also enriched in subtype-3, such as antigen processing: ubiquitination and proteasome degradation, synthesis of active ubiquitin: roles of E1 and E2 enzymes and major histocompatibility complex (MHC) class II antigen presentation. In addition, the P53 signaling pathway, ER to Golgi anterograde transport and FBXW7 mutants and NOTCH1 in cancer are also enriched in subtype-3. In addition to immune and cancer-related pathways such as neddylation, antigen processing: ubiquitination and proteasome degradation, MHC class II antigen presentation and cell cycle, transcriptional regulation by TP53 is also enriched in subtype-4.

Discussion

To avoid the instability of transcript analysis, we used a relatively stable gene interaction network to explore the subtypes of breast cancer from a new point of view in this study. For this purpose, we developed a sample-specific gene interaction perturbation method based on relative gene expression profiles. We identified four network-based subtypes based on gene interaction perturbations at the individual level, revealing the substantial heterogeneity reflected in the interactome in breast cancer patients. The core biological characteristics of each subtype are uniform, but the heterogeneity among the subtypes lies in many aspects, including prognosis, phenotypes and somatic mutations. Kaplan–Meier survival analysis showed that the differences in survival among the network-based subtypes were significant ($P = 0.0013$). The new network-based subtypes of breast cancer are closely related to the PAM50 subtypes and IHC index. The phenotypic variations measured by pathway scores showed differences among the network-based subtypes in breast cancer tumors. Furthermore, the ratios of somatic mutations occurring in cancer driver genes were significantly different among the network-based subtypes.

Gene expression profiles are variables and may show differences if measured at different time points or under different conditions, so that the subtypes based on expression data are not stable, while the network-based subtypes should be more stable and reliable. In addition, the network-based subtype system reveals the fact that every molecular is not isolated but interacting with each other to perform function, and it also shows that there is a possibility to investigate the mechanism of breast cancer from an interaction perspective.

Many studies have shown that network-based (or pathway-based) features are more robust and effective than single-gene features. The advantages of network-based methods have been well documented and accepted in the analysis of noisy high-throughput data. However, most of these methods merely utilize the gene sets in a network (or pathway) but ignore the interactions among genes. Therefore, these methods can only be called gene set-based methods, not real network-based (or pathway-based) methods. Different from the usual pathway-based method, we made better use of the gene interaction relations in the background network to explore new subtypes of breast cancer. Specifically, perturbations in gene interactions measured by the relative gene expression value were used to represent the perturbation of the gene interaction network. The perturbation of the network can be used to reflect the lesion extent of an individual with disease, which was innovatively measured by the edge perturbations in our study. Another highlight of our study is the individual-specific analysis of the gene interaction network. The precision medicine philosophy advocates for an individual treatment plan that targets the unique characteristics of the tumor. Therefore, it is important to focus on the unique pattern shown in the individual tumor sample to identify the most promising treatment strategy for the patient. Our individual-specific edge perturbation analysis of breast cancer will promote the development of precision medicine.

Although our trial was carried out on breast cancer, the application of our method should not be confined to this single cancer type—the edge-perturbation method can be applied broadly to any given cancer samples, as long as there are corresponding normal samples that can be used to establish homeostasis. Thus, it acts as a form of formidable resource that can unravel the biological system changes that happen to in a single patient. Therefore, this method is an ideal tool for personalized or precision oncology, which represents one potential research direction of future development.

Availability and implementation

The edge-perturbation-based method introduced in this study has been implemented in R and is available at https://github.com/Marscolono/SSPGI.git

Key Points

- To avoid the instability of transcript analysis, we used a relatively stable gene interaction network to explore the subtypes of breast cancer from a new point of view.
- The network-based subtypes of breast cancer were explored by using individual-specific edge perturbations measured by the relative gene expression value.
- The new network-based subtypes of breast cancer show strong heterogeneity in prognosis, somatic mutations, phenotypic changes and enriched pathways.
- The biomarker edges of the network-based subtype were identified to have similar perturbation patterns.

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *BRIBIO* online.

Acknowledgments

We thank Wenlong Ming for his helpful input, and we thank very much for the reviewers' useful suggestions.

Funding

The National Natural Science Foundation of China (81830053, 61972084); the China Postdoctoral Science Foundation (2019M651658).

References

1. Blows FM, Driver KE, Schmidt MK, *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.

2. Parker JS, Mullins M, Cheang MCU, *et al*. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009;**27**:1160–7.
3. Renault A-L, Mebirouk N, Fuhrmann L, *et al*. Morphology and genomic hallmarks of breast tumours developed by ATM deleterious variant carriers. *Breast Cancer Res* 2018;**20**:28–8.
4. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.
5. Gupta RA, Shah N, Wang KC, *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;**464**:1071–6.
6. Silva JM, Boczek NJ, Berres MW, *et al*. LSINCT5 is over expressed in breast and ovarian cancer and affects cellular proliferation. *RNA Biol* 2011;**8**:496–505.
7. Van Grembergen O, Bizet M, de Bony EJ, *et al*. Portraying breast cancers with long noncoding RNAs. *Sci Adv* 2016;**2**:e1600220.
8. Jiang Y-Z, Ma D, Suo C, *et al*. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell* 2019;**35**:428, e425–40.
9. Poudel P, Nyamundanda G, Patil Y, *et al*. Heterocellular gene signatures reveal luminal-a breast cancer heterogeneity and differential therapeutic responses. *npj Breast Cancer* 2019;**5**:21.
10. Dai H, Li L, Zeng T, *et al*. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res* 2019;**47**:e62–2.
11. Liu X, Wang Y, Ji H, *et al*. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* 2016;**44**:e164–4.
12. Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One* 2007;**2**:e425–5.
13. Guo Z, Zhang T, Li X, *et al*. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* 2005;**6**:58–8.
14. Joshi H, Bhanot G, Børresen-Dale AL, *et al*. Potential tumorigenic programs associated with TP53 mutation status reveal role of VEGF pathway. *Br J Cancer* 2012;**107**:1722–8.
15. Joshi H, Nord SH, Frigessi A, *et al*. Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics* 2012;**13**:199–9.
16. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74.
17. Goeman JJ, van de Geer SA, de Kort F, *et al*. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**:93–9.
18. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* 2010;**11**:R53.
19. Li X, Cai H, Wang X, *et al*. A rank-based algorithm of differential expression analysis for small cell line data with statistical control. *Brief Bioinform* 2019;**20**:482–91.
20. Wang H, Sun Q, Zhao W, *et al*. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* 2014;**31**:62–8.
21. Ciriello G, Gatza ML, Beck AH, *et al*. Comprehensive molecular portraits of invasive lobular breast Cancer. *Cell* 2015;**163**:506–19.
22. D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. In: Wu CH, Chen C (eds). *Bioinformatics for Comparative Proteomics*. Totowa, NJ: Humana Press, 2011, 49–61.
23. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**: 2498–504.
24. Sahni N, Yi S, Taipale M, *et al*. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 2015;**161**:647–60.
25. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics (Oxford, England)* 2010;**26**:1572–3.
26. Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? *Biostatistics* 2007;**8**(1):9–31. doi: 10.1093/biostatistics/kxj029.
27. Carter SL, Cibulskis K, Helman E, *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21.
28. Akbani R, Ng PKS, Werner HMJ, *et al*. A pan-cancer proteomic perspective on the Cancer Genome Atlas. *Nat Commun* 2014;**5**:3887–7.
29. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science* 2015;**349**:1483.
30. Sobin LH, Gospodarowicz MK, Wittekind C. *TNM Classification of Malignant Tumours*, 7th edn. New York: Wiley-Blackwell, 2009, 310.
31. Kandoth C, McLellan MD, Vandin F, *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 2013;**502**:333.
32. Norberg T, Klaar S, Kärf G, *et al*. Increased p53 mutation frequency during tumor progression— results from a Breast Cancer Cohort. *Cancer Res* 2001;**61**:8317.
33. D'Assoro AB, Leontovich A, Amato A, *et al*. Abrogation of p53 function leads to metastatic transcriptome networks that typify tumor progression in human breast cancer xenografts. *Int J Oncol* 2010;**37**:1167–76.
34. Aas T, Børresen A-L, Geisler S, *et al*. Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. *Nat Med* 1996;**2**: 811–4.
35. Bertheau P, Turpin E, Rickman DS, *et al*. Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen. *PLoS Med* 2007; **4**:e90.
36. Takahashi S, Moriya T, Ishida T, *et al*. Prediction of breast cancer prognosis by gene expression profile of TP53 status. *Cancer Sci* 2008;**99**:324–32.
37. Jia X, Li C, Li L, *et al*. Neddylation inactivation facilitates FOXO3a nuclear export to suppress estrogen receptor transcription and improve fulvestrant sensitivity. *Clin Cancer Res* 2019;**25**:3658–72.