Feature genes in metastatic breast cancer identified by MetaDE and SVM classifier methods

YOULIN TUO[1], NING AN[2] and MING ZHANG[2]

Departments of [1]Breast Surgery and [2]Oncology, Sichuan Provincial People's Hospital, Sichuan Academy of Medical Sciences, School of Clinical Medicine of University of Electronic Science and Technology of China, Chengdu, Sichuan 610000, P.R. China

Received August 18, 2017; Accepted December 1, 2017

DOI: 10.3892/mmr.2018.8398

Abstract. The aim of the present study was to investigate the feature genes in metastatic breast cancer samples. A total of 5 expression profiles of metastatic breast cancer samples were downloaded from the Gene Expression Omnibus database, which were then analyzed using the MetaQC and MetaDE packages in R language. The feature genes between metastasis and non-metastasis samples were screened under the threshold of P<0.05. Based on the protein-protein interactions (PPIs) in the Biological General Repository for Interaction Datasets, Human Protein Reference Database and Biomolecular Interaction Network Database, the PPI network of the feature genes was constructed. The feature genes identified by topological characteristics were then used for support vector machine (SVM) classifier training and verification. The accuracy of the SVM classifier was then evaluated using another independent dataset from The Cancer Genome Atlas database. Finally, function and pathway enrichment analyses for genes in the SVM classifier were performed. A total of 541 feature genes were identified between metastatic and non-metastatic samples. The top 10 genes with the highest betweenness centrality values in the PPI network of feature genes were *Nuclear RNA Export Factor 1*, cyclin-dependent kinase 2 (*CDK2*), myelocytomatosis proto-oncogene protein (*MYC*), *Cullin 5*, *SHC Adaptor Protein 1*, *Clathrin heavy chain*, *Nucleolin*, *WD repeat domain 1*, *proteasome 26S subunit non-ATPase 2* and *telomeric repeat binding factor 2*. The cyclin-dependent kinase inhibitor 1A (*CDKN1A*), E2F transcription factor 1 (*E2F1*), and *MYC* interacted with *CDK2*. The SVM classifier constructed by the top 30 feature genes was able to distinguish metastatic samples from non-metastatic samples [correct rate, specificity, positive predictive value and negative predictive value >0.89; sensitivity >0.84; area under the receiver operating characteristic curve (AUROC) >0.96]. The verification of the SVM classifier in an independent dataset (35 metastatic samples and 143 non-metastatic samples) revealed an accuracy of 94.38% and AUROC of 0.958. Cell cycle associated functions and pathways were the most significant terms of the 30 feature genes. A SVM classifier was constructed to assess the possibility of breast cancer metastasis, which presented high accuracy in several independent datasets. *CDK2*, *CDKN1A*, *E2F1* and *MYC* were indicated as the potential feature genes in metastatic breast cancer.

Introduction

Breast cancer is one of the most commonly diagnosed types of cancer, accounting for one-third of cancer cases in the USA (1). The survival rate of breast cancer has improved steadily with the development of early diagnosis and adjuvant therapy; however, the overall survival of patients with metastatic disease still remains poor (2). It has been estimated that >90% of breast cancer mortalities are associated with tumor metastasis (3,4).

Metastasis is associated with poor patient prognosis and an acceleration of the carcinoma progress (5). Brain, bone, lungs and liver are the most frequently targeted organs for breast cancer metastasis, and the tumor microenvironment is considered to be a critical regulator for the metastatic process (6). Comprehensive understanding of metastasis progression is very important for identifying novel therapeutic strategies to prevent metastatic disease.

The MetaOmics software in R language is comprised of the MetaDE, MetaQC and MetaPath packages. The MetaDE package primarily contains 12 state-of-the-art genomic meta-analysis methods to detect differentially expressed genes (7). The MetaQC package is the quantitative and objective tool for the determination of the inclusion/exclusion criteria for meta-analysis (8). The MetaDE and MetaQC packages have been intensively utilized for data digging from microarray profiles. Fc fragment of immunoglobulin G binding protein, for example, has been reported as a candidate

Correspondence to: Dr Ming Zhang, Department of Oncology, Sichuan Provincial People's Hospital, Sichuan Academy of Medical Sciences, School of Clinical Medicine of University of Electronic Science and Technology of China, 32 West Section 2, 1 Ring Road, Chengdu, Sichuan 610000, P.R. China
E-mail: zhangming1123456@outlook.com

Key words: breast cancer, metastasis, protein-protein interactions, feature gene, support vector machine classifier

metastasis-associated gene using the integrated method of MetaDE and survival analysis (9).

As an effective classifier for identification, the support vector machine (SVM) classifier is well suited for signature modeling (10). Guyon *et al* (11) applied the SVM classifier to select feature genes from DNA microarrays, and the selected genes were proved to exhibit a greater classification performance. Fan *et al* (10) demonstrated that the SVM classifier for feature gene selection was able to speed up the classification process and the generalization performance.

In the present study, several microarray profiles of breast cancer samples (including metastatic and non-metastatic samples) were downloaded to investigate the feature genes in metastatic samples. A SVM classifier was constructed to identify feature genes, which was validated by another independent gene expression dataset from The Cancer Genome Atlas (TCGA) database.

Materials and methods

Processing of microarray data. Expression profiles matching the search terms of 'breast cancer', 'homo sapiens' and 'metastasis' in the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) database were screened on 22nd April 2016. The profiles were selected using the following filtering criteria: i) The data was gene expression microarray data; ii) data was collected from cancerous tissue samples or cancerous-metastasis samples; iii) and the metastatic statuses of the samples were clearly recorded.

A total of 5 microarray profiles were retrieved from the GEO database (Table I). The GSE46928, GSE43837, GSE46826, GSE39494 and GSE29431 profiles had a total of 52, 38, 27, 10 and 31 samples, respectively; these in turn included 11, 19, 21, 5 and 13 metastatic samples, respectively.

For GSE46928, GSE43837 and GSE29431 datasets based on the Affymetrix platform (Affymetrix; Thermo Fisher Scientific, Inc., Waltham, MA, USA), the raw data were used to perform background correction via Affymetrix microarray software Affy version 1.42.3 (https://bioconductor.org/packages/release/bioc/html/affy.html) in R version 3.1.0, and normalization via the quantiles method (12).

For GSE46826 and GSE39494 datasets based on the Agilent platform (Agilent Technologies, Inc., Santa Clara, CA, USA), the gene names in the microarray data were identified according to Agilent platform. Then, the average values were used as the expression levels of genes corresponding to multiple probes. The Limma package 3.22.1 (13) (https://bioconductor.org/packages/release/bioc/html/limma.html) was used for the normalization of these data.

Screening of feature genes. All of the selected datasets were merged to form a novel dataset for the screening of feature genes using MetaDE.ES in the MetaDE package 1.0.5 (14). Firstly, principal component analysis and standardized mean rank methods in the MetaQC package (8) were applied to ensure quality control (QC) within the novel datasets from the different profiles. In this process, the following parameters were used: Internal QC, external QC, accuracy QC (AQCg), precision of AQCg, consistency QC (CQCg) and precision of CQCg. Tests for heterogeneity were then performed to determinate the gene expression differentiations among the different datasets; Qpval >0.05 and $tau^2=0$ were used as the criteria for homogenous genes. Finally, the differentially expressed genes (DEGs) between metastatic samples and non-metastatic samples in the dataset were identified under the threshold of $P<0.05$, which were considered as feature genes in the following analysis.

Construction of the protein-protein interaction (PPI) network. The interactions between human genes in the Biological General Repository for Interaction Datasets (thebiogrid.org/, BioGRID Version 3.4.154 Released) (15), Human Protein Reference Database (www.hprd.org/, HPRD Release 9) (16) and Biomolecular Interaction Network Database (BIND 2.0) (17) were downloaded. The screened feature genes were then subjected to the downloaded interactions to obtain the PPI network, which was visualized using Cytoscape 3.6.0 software (18).

The degree (the connection with other genes) and the betweenness centrality (BC) value of feature genes in the network were calculated. The following formula was used for calculating BC:

$$C_B(v) = \sum_{t \neq v \neq u \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is the shortest path between s and t, and $\sigma_{st}(v)$ is the node numbers in the path of σ_{st}. A high BC value indicates a high degree of feature genes in the network.

Establishment of the SVM classifier. Feature genes were ranked according to their BC values, and those that were present in the most qualified samples were collected as the training dataset for the establishment of the SVM classifier. The remaining feature genes were used as the verification datasets for the classifier. The feature genes in the SVM classifier were used to perform the two-way clustering of samples and expression levels. The clustering results were visualized using a heatmap (19). The aim of the constructed SVM classifier was to distinguish whether the cancer had metastasized by analyzing the primary cancer samples.

A set of microarray data from breast cancer samples (https://cancergenome.nih.gov/) was downloaded from TCGA (tcga-data.nci.nih.gov/docs/publications/tcga/) for further clarification. In total, 597 samples were included in the dataset, among which 178 samples had clinical information regarding metastasis status, follow-up time and the clinical outcomes. There were 35 metastatic samples and 143 non-metastatic samples.

Function and pathway enrichment. Fisher's test was utilized with the 'runHyperKEGG' and 'runHyperGO' functions of the Easy Microarray Data Analysis package 1.4.4 (20) for the function and pathway enrichment of feature genes. $P<0.05$ was set as the cut-off criterion.

Results

Feature gene selection. The QC results of all 5 microarray profiles are displayed in Fig. 1 and Table II; the results

Table I. Basic information of downloaded microarray data.

GEO accession	Chip	Probe number	Total sample number	Non-metastasis samples	Metastasis samples
GSE46928	HG-U133A	22,283	52	41	11
GSE43837	U133_X3P	61,360	38	19	19
GSE46826	Agilent-021924	62,977	28	6	22
GSE39494	Agilent-014850	41,000	10	5	5
GSE29431	HG-U133_Plus_2	54,675	31	18	13

GEO, Gene Expression Omnibus.

Table II. Results of quality control parameters and standardized mean rank.

Microarray profile	IQC	EQC	CQCg	CQCp	AQCg	AQCp	SMR
GSE46928	4.91	4.78	93.87	148.67	153.83	56.44	2.42
GSE43837	5.12	5.00	52.41	101.36	184.06	39.30	1.57
GSE46826	4.56	4.22	68.15	146.58	106.19	29.43	4.83
GSE39494	2.16	2.92	21.58	64.14	46.61	33.90	7.17
GSE29431	3.19	4.16	43.66	89.52	113.24	31.16	3.36

QC, quality control; IQC, internal QC; EQC, external QC; AQCg, accuracy QC; AQCp, precision of AQCg; CQCg, consistency QC; CQCp, precision of CQCg; SMR, standardized mean rank.

indicated there was good quality within all datasets. Next, using the MetaDE package, 541 feature genes were identified and the top 10 were ranked by their P-values; these included, *non-SMC condensing I complex subunit H*, *small nuclear ribonucleoprotein U11/U12 subunit 25*, *cellular retinoic acid binding protein 2*, *guanosine triphosphate binding protein 2*, *homer scaffolding protein 2*, *family with sequence similarity 64 member A*, WD repeat domain (*WDR*) *45, dual specificity tyrosine phosphorylation regulated kinase 4, chromosome 12 open reading frame 10 and H2A histone family member Z* (Table III).

PPI network of feature genes. The PPI network of feature genes was comprised of 307 nodes (feature genes) and 586 lines (interactions; Fig. 2). There were 220 nodes (shown in green) that exhibited higher expression levels in metastatic samples, as well as 87 nodes (shown in purple) that exhibited lower expression levels in metastatic samples when compared to non-metastatic samples. As shown in Fig. 3, 168 genes exhibited a log (degree) of 0-1 and only 5 genes exhibited a log (degree) of >3 in the network. In addition, the top 30 genes with the highest BC values were listed in Table IV. The top 10 feature genes were Nuclear RNA Export Factor 1 (*NXF1*), cyclin-dependent kinase 2 (*CDK2*), myelocytomatosis proto-oncogene protein (*MYC*), Cullin 5 (*CUL5*), SHC Adaptor Protein 1 (*SHC1*), Clathrin heavy chain (*CLTC*), Nucleollin (*NCL*), *WDR1*, proteasome 26S subunit, non-ATPase 2 (*PSMD2*), telomeric repeat binding factor 2 (*TERF2*; Table IV). Among these feature genes the CDK inhibitor 1A (*CDKN1A*), E2F transcription factor 1 (*E2F1*) and *MYC* interacted with *CDK2*.

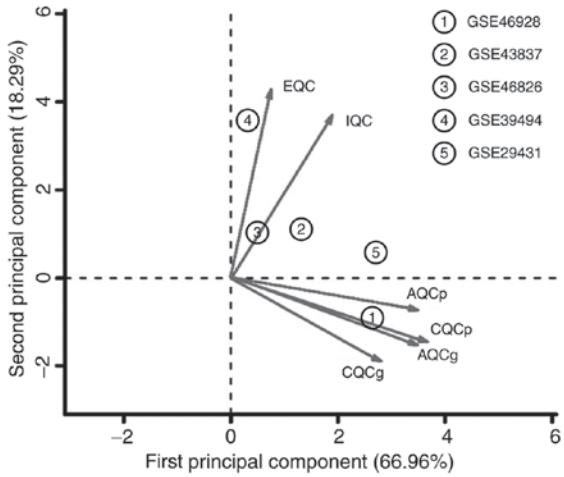

Figure 1. Quality control results of the merged datasets from 5 microarray profiles (marked as 1-5) obtained via MetaQC analysis. The first principal component is presented on the x-axis, while the second principal component is shown on the y-axis. QC, quality control; IQC, internal QC; EQC, external QC; AQCg, accuracy QC; AQCp, precision of AQCg; CQCg, consistency QC; CQCp, precision of CQCg.

SVM classifier. Feature genes ranked with BC values were picked at 10 intervals from the top 10 to the top 50, for the construction of the SVM classifier. The dataset GSE46928 with the largest sample size was used as the training dataset. As shown in Fig. 4A, the accuracy of the SVM classifier improved with the increasing number of genes and the accuracy stabilized at 100% once the top 30 genes were selected. The SVM classifier constructed by the top 30 feature genes was able to distinguish metastatic samples from the non-metastatic

Table III. Top 10 feature genes selected using the MetaDE package.

Gene	P-value	Q	Qp	tau^2	Exp
NCAPH	$4.17x10^{-5}$	1.4919	0.8281	0	1
SNRNP25	$1.20x10^{-4}$	3.8687	0.4241	0	1
CRABP2	$1.55x10^{-4}$	0.5088	0.9726	0	1
GTPBP2	$3.51x10^{-4}$	0.4245	0.9804	0	1
HOMER2	$3.74x10^{-4}$	3.4071	0.4921	0	1
FAM64A	$3.93x10^{-4}$	2.5196	0.6411	0	1
WDR45	$4.34x10^{-4}$	2.5287	0.6395	0	1
DYRK4	$4.61x10^{-4}$	1.4036	0.8436	0	1
C12orf10	$4.92x10^{-4}$	2.7885	0.5938	0	1
H2AFZ	$5.19x10^{-4}$	3.0197	0.5545	0	1

NCAPH, non-SMC condensing I complex subunit H; SNRNP35, small nuclear ribonucleoprotein U11/U12 subunit 25; CRABP2, cellular retinoic acid binding protein 2; GTPBP2, guanosine triphosphate binding protein 2; HOMER2, homer scaffolding protein 2; FAM64A, family with sequence similarity 64 member A; WDR45, WD repeat domain 45; DYRK4, dual specificity tyrosine phosphorylation regulated kinase 4; C12orf10, chromosome 12 open reading frame 10; H2AFZ, H2A histone family member Z.

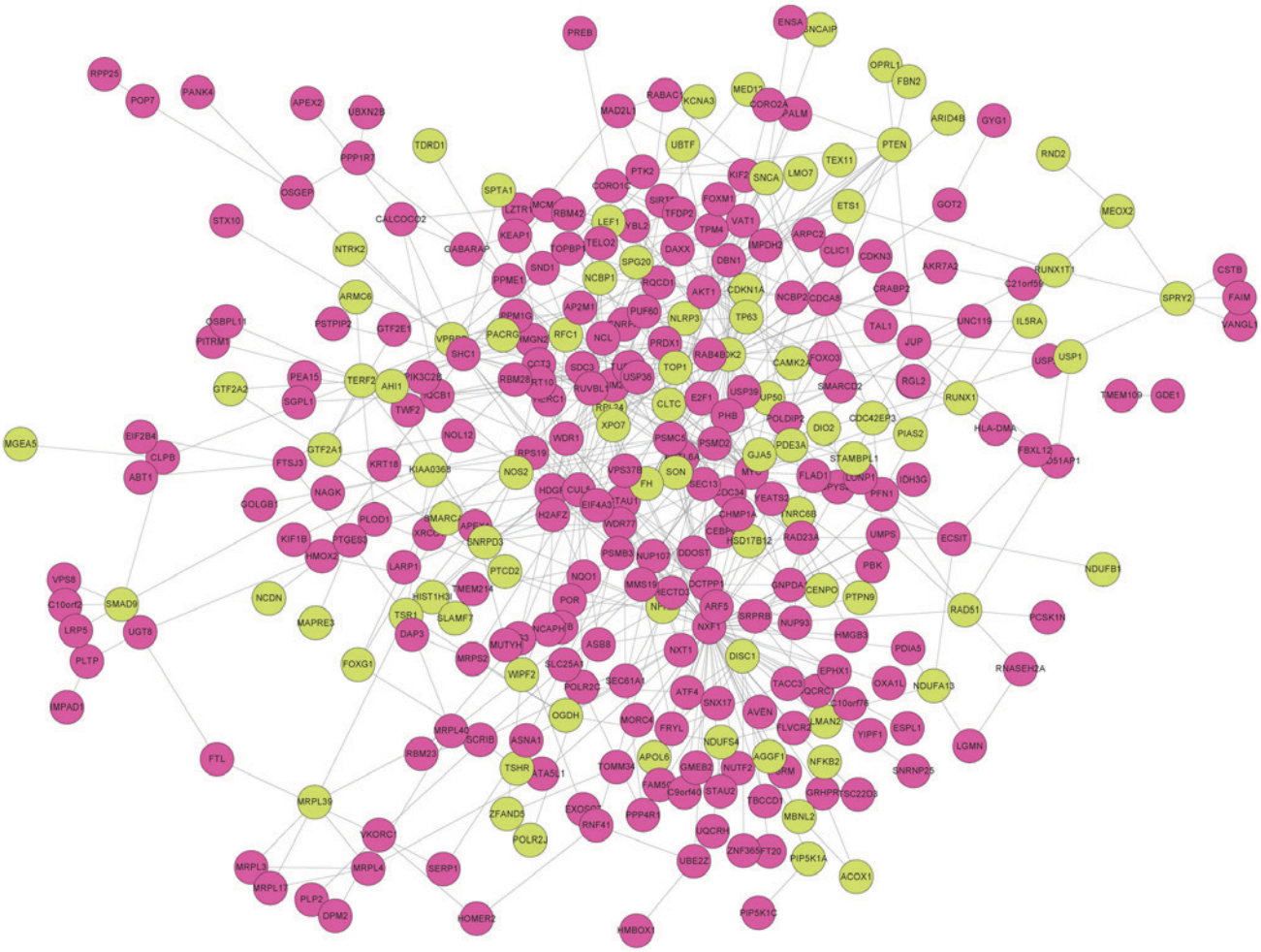

Figure 2. Protein-protein interaction network of feature genes. Green nodes are the genes that exhibited higher expression in metastatic samples, while the purple nodes are those that exhibited lower expression in metastatic samples when compared with non-metastatic samples.

samples with high accuracy (Fig. 4B). The selected 30 genes were considered to be the critical biomarkers for metastatic breast cancer, and included protein kinase B serine/threonine kinase 1 (*AKT1*), *CDKN1A*, ETS proto-oncogene 1 transcription factor (*ETS1*), runt related transcription factor 1 (*RUNX1*), RUNX1 translocation partner 1 (*RUNX1T1*), nitric oxide

Table IV. Top 30 feature genes with the highest betweeness centrality in the protein-protein interaction network.

Gene	BC	EXP	Degree	P-value	Q	Qp	tau^2
NXF1	0.3864	1	66	$3.43x10^{-2}$	3.7163	0.4458	0
CDK2	0.2047	0	44	$3.33x10^{-2}$	2.2882	0.6829	0
MYC	0.1382	1	27	$4.91x10^{-2}$	3.4827	0.4805	0
CUL5	0.1006	1	21	$2.86x10^{-2}$	3.0080	0.5565	0
SHC1	0.0974	1	16	$1.60x10^{-2}$	1.1518	0.8860	0
CLTC	0.0783	0	20	$2.66x10^{-2}$	2.8154	0.5892	0
NCL	0.0568	1	15	$9.12x10^{-4}$	1.3121	0.8593	0
WDR1	0.0532	1	8	$8.49x10^{-3}$	2.5722	0.6318	0
PSMD2	0.0476	1	13	$8.31x10^{-4}$	3.4061	0.4923	0
TERF2	0.0460	0	11	$1.65x10^{-2}$	0.3161	0.9888	0
RUVBL1	0.0450	1	13	$2.51x10^{-2}$	0.8904	0.9259	0
PRDX1	0.0394	1	10	$4.09x10^{-2}$	2.0057	0.7347	0
PTEN	0.0334	0	12	$1.99x10^{-3}$	3.5056	0.4770	0
HDGF	0.0313	1	10	$3.93x10^{-2}$	3.4475	0.4859	0
RUNX1T1	0.0291	0	4	$2.88x10^{-2}$	0.2956	0.9901	0
IQCB1	0.0283	1	12	$1.20x10^{-3}$	0.7995	0.9385	0
AKT1	0.0273	1	15	$3.26x10^{-3}$	2.0318	0.7299	0
APEX1	0.0268	1	6	$1.09x10^{-2}$	1.8543	0.7625	0
TSR1	0.0263	0	7	$2.06x10^{-2}$	2.2661	0.6870	0
TUBB2A	0.0258	1	9	$1.18x10^{-2}$	3.4922	0.4791	0
ETS1	0.0257	0	5	$4.11x10^{-3}$	3.2520	0.5166	0
PSMC5	0.0249	1	11	$1.85x10^{-2}$	2.7803	0.5952	0
RUNX1	0.0248	0	4	$4.45x10^{-2}$	2.3257	0.6761	0
SMAD9	0.0242	0	6	$3.52x10^{-2}$	1.3518	0.8525	0
STAU1	0.0239	1	14	$1.33x10^{-2}$	1.7706	0.7779	0
DBN1	0.0235	1	13	$2.31x10^{-3}$	2.1547	0.7073	0
SNCA	0.0229	0	10	$2.51x10^{-2}$	2.9088	0.5732	0
CDKN1A	0.0226	0	12	$1.48x10^{-2}$	3.7775	0.4369	0
SLC25A1	0.0223	1	2	$2.22x10^{-2}$	1.1438	0.8873	0
NOS2	0.0222	0	9	$4.71x10^{-2}$	1.0560	0.9012	0

EXP is the expression value ratio of genes between metastastic samples and non-metastastic samples, while values of 1 represent high expression in metastastic samples and values of 0 represent high expression in non-metastastic samples. BC, betweeness centrality.

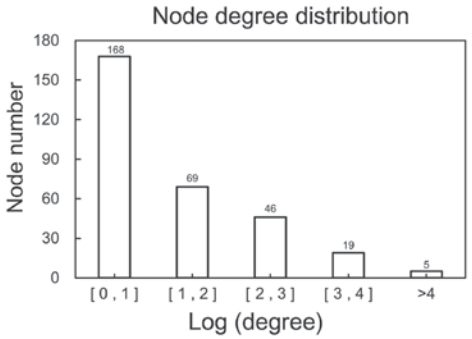

Figure 3. Distribution of node degrees in the protein-protein interaction network of feature genes. The x-axis is the log (degree) value and the y-axis is the corresponding node numbers to the degree.

synthase 2 (*NOS2*), *MYC*, phosphatase and tensin homolog (*PTEN*) and *CDK2*. Clustering analysis of these 30 feature genes and the samples in GSE46928 demonstrated that these genes have significantly different expression levels between the metastatic and non-metastatic samples (Fig. 5).

The classification efficacy of the constructed classifier was also tested on the other 4 microarray datasets (Fig. 6). All samples in GSE39494 (Fig. 6B) and GSE46826 (Fig. 6D) were correctly distinguished, and only 3 samples in GSE29431 (Fig. 6A) and 4 samples in GSE43837 (Fig. 6C) were misclassified. Overall, the SVM classifier displayed good performance in terms of distinguishing between metastatic and non-metastatic samples. The correct rate, specificity, positive predictive value (PPV) and negative predictive value (NPV) were >0.89, sensitivity was >0.84 and the area under the receiver operating characteristic curve (AUROC) was >0.96 (Table V).

An independent dataset of breast cancer samples was downloaded from the TCGA database to test the classification effect of the constructed classifier (Fig. 7). The results revealed

Table V. Classification effect evaluation of the support vector machine classifier.

Dataset	Number of samples	Correct rate	Sensitivity	Specificity	PPV	NPV	AUROC
GSE29431	31	1	1	1	1	1	1
GSE39494	10	0.903	0.846	0.944	0.917	0.895	0.975
GSE43837	38	1	1	1	1	1	1
GSE46826	28	0.895	0.895	0.895	0.895	0.895	0.965

PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristic curve.

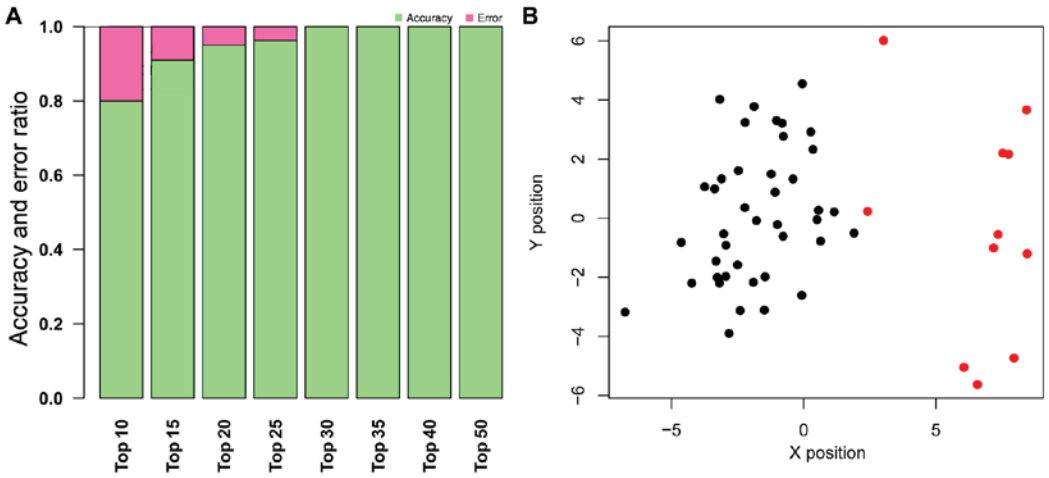

Figure 4. Accuracy and efficacy of the support vector machine classifier. (A) The accuracy and error ratio of the classifier at different gene numbers (top 10 to top 50). (B) The classification efficacy of the classifier constructed using the top 30 genes for samples in the GSE46928 dataset. Non-metastatic samples are marked in black and the metastatic samples are marked in red.

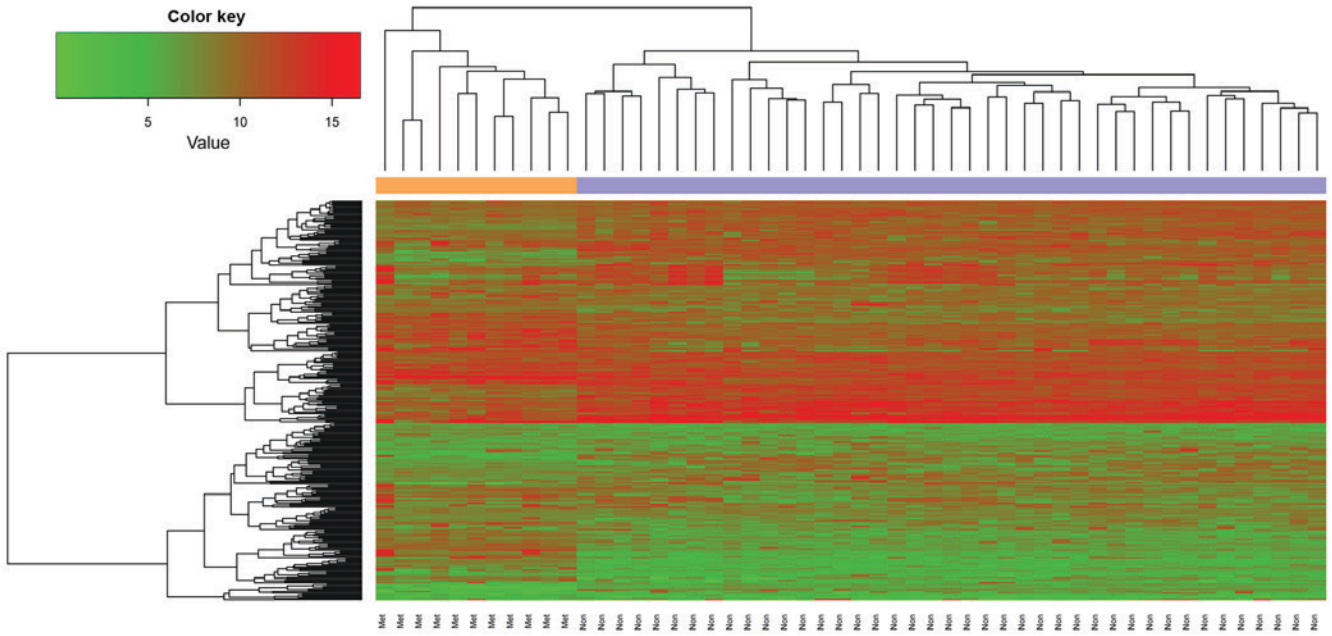

Figure 5. Clustering heatmap of the top 30 genes and samples in the training dataset. The color gradient from red to green represents the changes in expression level from high to low. The bars represent the samples (orange refers to metastatic samples; purple refers to non-metastatic samples). Met, metastatic samples; Non, non-metastatic samples.

an accuracy of 94.38% (168/178) in 35 metastatic samples and 143 non-metastatic samples, with an AUROC of 0.958 (Fig. 7B). Based on the 30 feature genes, the survival time of patients with metastatic breast cancer was significantly shorter

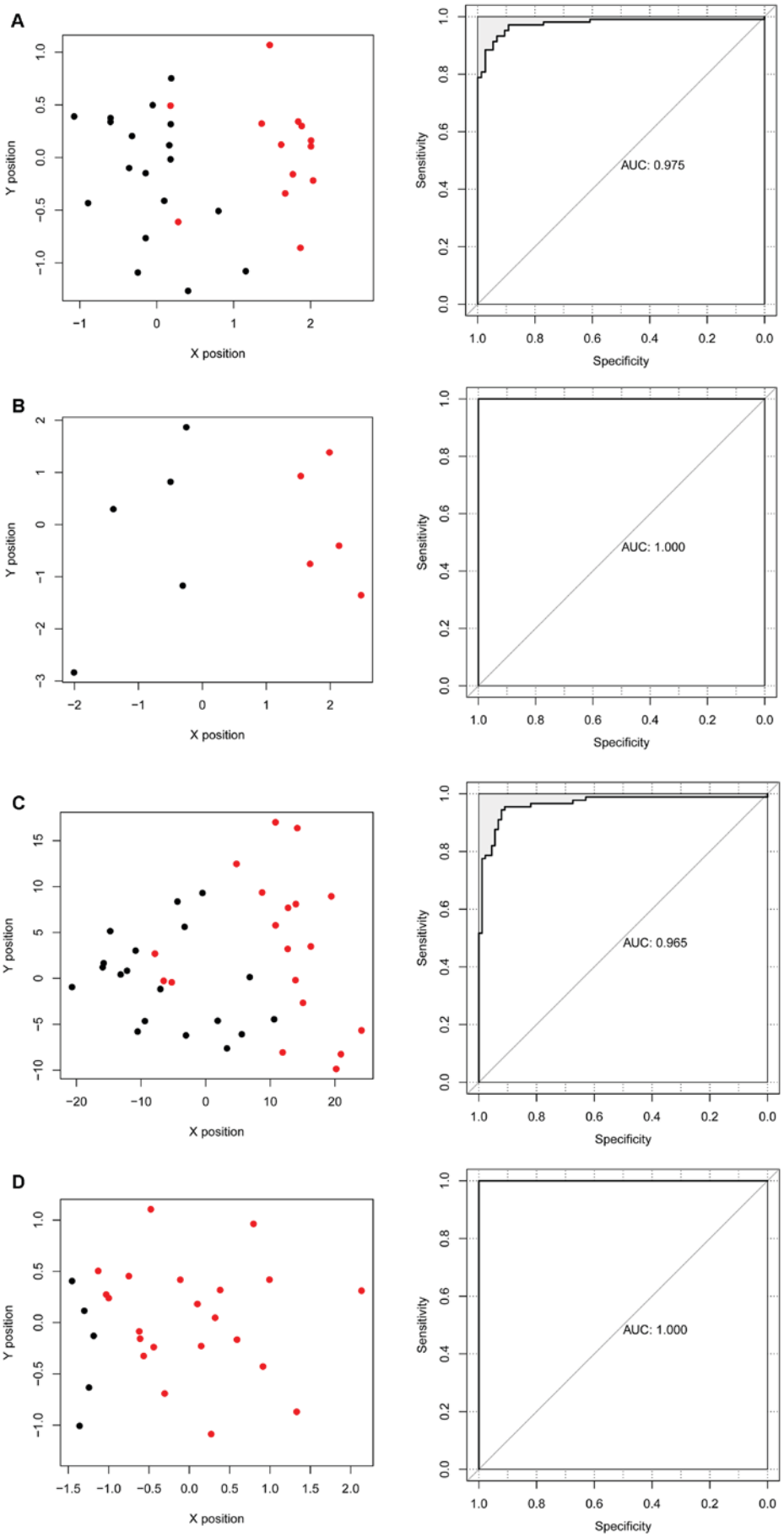

Figure 6. Classification results on other microarray profiles, including (A) GSE29431, (B) GSE39494, (C) GSE43837 and (D) GSE46826. Non-metastatic samples are marked in black and metastatic samples are marked in red. The receiver operating characteristic curves of the classifier are displayed on the right-hand side. AUC, area under the curve.

Table VI. Enriched pathways of the 30 feature genes.

Pathway	P-value	Genes
hsa05200: Pathways in cancer	$1.11x10^{-5}$	*AKT1, CDKN1A, ETS1, RUNX1T1, NOS2, RUNX1, MYC, PTEN, CDK2*
hsa04012: ErbB signaling pathway	$3.85x10^{-3}$	*AKT1, CDKN1A, SHC1, MYC*
hsa04115: p53 signaling pathway	$2.60x10^{-2}$	*CDKN1A, PTEN, CDK2*
hsa04110: Cell cycle	$2.81x10^{-5}$	*CDKN1A, MYC, CDK2*

AKT1, protein kinase B serine/threonine kinase 1; CDKN1A, cyclin-dependent kinase inhibitor 1A; ETS1, ETS proto-oncogene 1 transcription factor; RUNX1, runt related transcription factor 1; RUNX1T1, RUNX1 translocation partner 1; NOS2, nitric oxide synthase 2; MYC, myelocytomatosis proto-oncogene protein; PTEN, phosphatase and tensin homolog; ErbB, Erb-B2 receptor tyrosine kinase 2; SHC1, SHC Adaptor Protein 1.

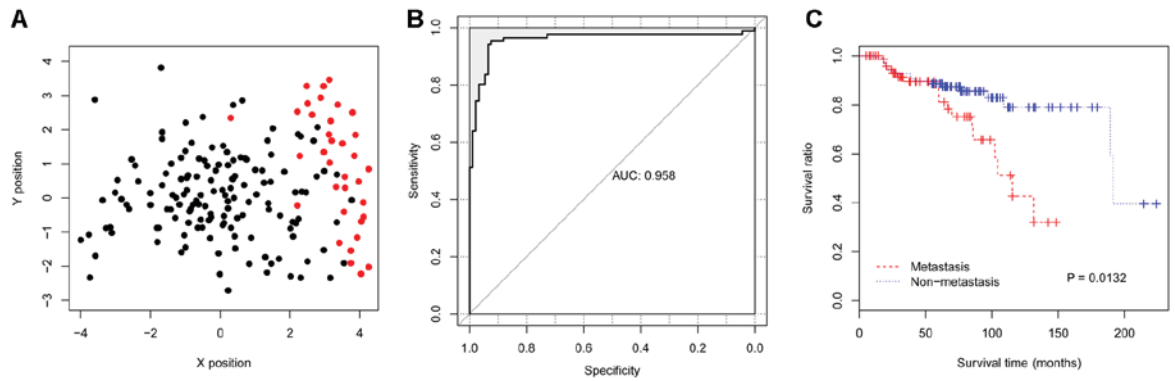

Figure 7. Classification effect of the support vector machine classifier on an independent sample from The Cancer Genome Atlas database. (A) The spot graph of the different samples (non-metastatic samples are marked in black and metastatic samples are marked in red). (B) The receiver operating characteristic curve and (C) the survival curve. AUC, area under the curve.

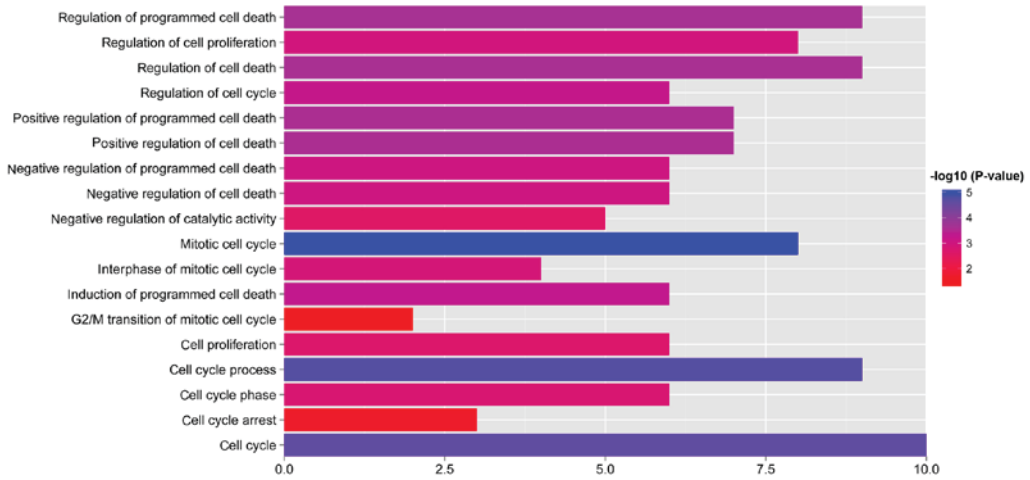

Figure 8. Enriched functions of the 30 feature genes. Gene numbers are displayed on the x-axis. The color represents the -log (P-value) and the changes from red to blue represents high -log (P-value) to low -log (P-value).

than the patients with non-metastatic breast cancer, and the survival status was worse (Fig. 7C).

Function and pathway enrichment. The 30 feature genes in the SVM classifier were utilized for function and pathway enrichment. The results indicated that cell cycle associated functions and pathways were the most significant terms (Fig. 8; Table VI).

Discussion

As breast cancer metastasis accounts for the majority of breast cancer mortalities, there have been a number of reports analyzing DEGs associated with metastasis in breast cancer. Some previous studies have identified the markers associated with metastasis using the protein-network based approach (21-23). Walsh *et al* (24) identified tripartite motif

containing 25 as a key determinant of breast cancer metastasis using an integrated transcriptional interaction network. In the present study, MetaQC package was firstly applied to conduct QC tests for the different profiles as the MetaQC package is the quantitative and objective tool in the determination of the inclusion/exclusion criteria for meta-analysis (8). The DEGs between metastatic and non-metastatic samples in the dataset were identified using the MetaDE package, which contains 12 state of the art genomic meta-analysis methods that detect DEGs (7). In the present study, a total of 541 feature genes were identified between metastatic and non-metastatic samples.

The PPI network of DEGs was constructed and was comprised of 307 feature genes and 586 interactions, among which 220 nodes exhibited higher expression levels in metastatic samples and 87 nodes exhibited lower expression levels in metastatic samples when compared with non-metastatic samples. Feature genes were ranked according to their BC that quantifies the importance of a vertex within a graph (25,26). The top 10 genes with the highest BC values included *NXF1*, *CDK2*, *MYC*, *CUL5*, *SHC1*, *CLTC*, *NCL*, *WDR1*, *PSMD2* and *TERF2*. *CDKN1A*, *E2F1* and *MYC* were the genes that interacted with *CDK2*.

Then, the SVM classifier of screened feature genes was constructed to evaluate the classification performance. The SVM classifier constructed by the top 30 feature genes (which included *AKT1*, *CDKN1A*, *ETS1*, *RUNX1T1*, *NOS2*, *RUNX1*, *MYC*, *PTEN* and *CDK2*, for example) was able to distinguish metastatic samples from the non-metastatic samples; this was proved by the clustering analysis. Overall, the classifier displayed good performance with a correct rate, specificity, PPV and NPV of >0.89, sensitivity >0.84 and an AUROC of >0.96. The verification on an independent dataset exhibited an accuracy of 94.38% and an AUROC of 0.958 for the 35 metastatic samples and 143 non-metastatic samples. The survival time of the metastatic samples was revealed to be shorter than the non-metastatic samples, based on the analysis of these 30 feature genes. Cell cycle associated functions and pathways were the most significant terms of the 30 feature factors.

CDK2 is reported to exert important roles in cell cycle regulation and is associated with tumor aggressiveness and poor prognosis (27,28). Kim *et al* (29) demonstrated that the specific activity of CDK2 could be used as a prognostic indicator for early breast cancer. Roesley *et al* (30) also identified that CDK2 phosphorylates breast cancer metastasis suppressor 1 (BRMS1) on Serine 237 and the mutation can prevent BRMS1 from suppressing cell migration. In addition, sirtuin 2 (SIRT2)-mediated inhibition of the migration of fibroblasts can be antagonized by the CDK2-induced SIRT2 phosphorylation (31). *CDKN1A* (also known as p21), one of the CDK inhibitor genes, contributes to cell cycle progression (32). Variant genotypes of *CDKN1A* were observed to be associated with an increased risk of breast cancer in the Chinese female population (33). When mammalian cells are exposed to DNA damaging agents, CDKN1A will inhibit cyclin/CDK2 complexes and participate in mediating growth arrest (34). The CDK2/CDKN1A ratio is considered to be a predictive factor of major clinical events in patients with oral squamous cell carcinoma (35). *E2F1* is a target of cellular (c)-Myc that promotes cell cycle progression (36). The *E2F1* mRNA levels are a strong determinant of clinical outcome in primary breast cancer (37). The CDK2-E2F1 signaling pathway exerts a pivotal role in regulating the G1 to S phase transition in the cell cycle (38). The interactions between CDK2/CDKN1A and CDK2/E2F1 identified in the present study indicated that they may influence the metastasis of breast cancer via their effect on the cell cycle.

The proto-oncogene *c-MYC* encodes a transcription factor that regulates cell growth, proliferation and apoptosis. *c-MYC* is commonly amplified in breast cancer and promotes the phenotypic transformation of mammary cells by synergistically interacting with transforming growth factor α (39). *MYC* gene amplification is often acquired in lethal distant breast cancer metastases of unamplified primary tumors (40), and the overexpression of *MYC* significantly decreased the metastasis of breast cancer cells to lung (41).

In conclusion, in the present study a SVM classifier was constructed to assess the possibility of breast cancer metastasis, which exhibited high accuracy in several independent datasets. The *CDK2*, *CDKN1A*, *E2F1* and *MYC* genes were highlighted as the potential feature genes for metastatic breast cancer, which may interact synergistically by influencing the cell cycle. The results provided some potential markers for breast cancer metastasis, which may also be prospective precise treatment targets for metastatic breast cancer. In the group's future studies, the expression levels of the potential feature genes will be validated in clinical samples by reverse transcription-quantitative polymerase chain reaction or immunohistochemical staining.

References

1. DeSantis C, Ma J, Bryan L and Jemal A: Breast cancer statistics, 2013. CA Cancer J Clin 64: 52-62, 2014.
2. Jemal A, Siegel R, Xu J and Ward E: Cancer statistics, 2010. CA Cancer J Clin 60: 277-300, 2010.
3. Weigelt B, Peterse JL and van't Veer LJ: Breast cancer metastasis: Markers and models. Nat Rev Cancer 5: 591-602, 2005.
4. Sleeman J and Steeg PS: Cancer metastasis as a therapeutic target. Eur J Cancer 46: 1177-1180, 2010.
5. Khan S, Shukla S, Sinha S, Lakra AD, Bora HK and Meeran SM: Centchroman suppresses breast cancer metastasis by reversing epithelial-mesenchymal transition via downregulation of HER2/ERK1/2/MMP-9 signaling. Int J Biochem Cell Biol 58: 1-16, 2015.
6. Joyce JA and Pollard JW: Microenvironmental regulation of metastasis. Nat Rev Cancer 9: 239-252, 2009.
7. Chen DT, Hernandez JM, Shibata D, McCarthy SM, Humphries LA, Clark W, Elahi A, Gruidl M, Coppola D and Yeatman T: Complementary strand microRNAs mediate acquisition of metastatic potential in colonic adenocarcinoma. J Gastrointest Surg 16: 905-913, 2012.
8. Kang DD, Sibille E, Kaminski N and Tseng GC: MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis. Nucleic Acids Res 40: e15, 2012.
9. Qi C, Hong L, Cheng Z and Yin Q: Identification of metastasis-associated genes in colorectal cancer using metaDE and survival analysis. Oncol Lett 11: 568-574, 2016.
10. Fan ZG, Wang KA and Lu BL: Feature selection for fast image classification with support vector machines. In: International Conference on Neural Information Processing Springer 3316, pp1026-1031, 2004.
11. Guyon I, Weston J, Barnhill S and Vapnik V: Gene selection for cancer classification using support vector machines. Machine learning 46: pp389-422, 2002.
12. Järvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP and Monni O: Are data from different gene expression microarray platforms comparable? Genomics 83: 1164-1168, 2004.
13. Smyth GK: Limma: Linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor Springer, pp397-420, 2005.

14. Wang X, Li J, Tseng GC and Wang MX: Package 'MetaDE'. 2012.
15. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, *et al*: The BioGRID interaction database: 2017 update. Nucleic Acids Res 45: D369-D379, 2017.
16. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, *et al*: Human protein reference database-2009 update. Nucleic Acids Res 37 (Database Issue): D767-D772, 2009.
17. Bader GD, Betel D and Hogue CW: BIND: The biomolecular interaction network database. Nucleic Acids Res 31: 248-250, 2003.
18. Smoot ME, Ono K, Ruscheinski J, Wang PL and Ideker T: Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics 27: 431-432, 2011.
19. Chang KC, Wang Y, Bodine PV, Nagpal S and Komm BS: Gene expression profiling studies of three SERMs and their conjugated estrogen combinations in human breast cancer cells: Insights into the unique antagonistic effects of bazedoxifene on conjugated estrogens. J Steroid Biochem Mol Biol 118: 117-124, 2010.
20. Servant N, Gravier E, Gestraud P, Laurent C, Paccard C, Biton A, Brito I, Mandel J, Asselain B, Barillot E and Hupé P: EMA-A R package for easy microarray data analysis. BMC Res Notes 3: 277, 2010.
21. Chuang HY, Lee E, Liu YT, Lee D and Ideker T: Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140, 2007.
22. Jonsson PF, Cavanna T, Zicha D and Bates PA: Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7: 2, 2006.
23. Sodek KL, Evangelou AI, Ignatchenko A, Agochiya M, Brown TJ, Ringuette MJ, Jurisica I and Kislinger T: Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT). Mol Biosyst 4: 762-773, 2008.
24. Walsh LA, Alvarez MJ, Sabio EY, Reyngold M, Makarov V, Mukherjee S, Lee KW, Desrichard A, Turcan Ş, Dalin MG, *et al*: An integrated systems biology approach identifies TRIM25 as a key determinant of breast cancer metastasis. Cell Rep 20: 1623-1640, 2017.
25. Kourtellis N, Francisci Morales GD and Bonchi F: Scalable online betweenness centrality in evolving graphs. IEEE Transact Know Data Eng 27: 2494-2506, 2015.
26. Brandes U: On variants of shortest-path betweenness centrality and their generic computation. Soc Net 30: 136-145, 2008.
27. Lee MH and Yang HY: Regulators of G1 cyclin-dependent kinases and cancers. Cancer Metastasis Rev 22: 435-449, 2003.
28. Kourea H, Koutras A, Scopa C, Marangos MN, Tzoracoeleftherakis E, Koukouras D and Kalofonos HP: Expression of the cell cycle regulatory proteins p34cdc2, p21waf1, and p53 in node negative invasive ductal breast carcinoma. Mol Pathol 56: 328-335, 2003.
29. Kim S, Nakayama S, Miyoshi Y, Taguchi T, Tamaki Y, Matsushima T, Torikoshi Y, Tanaka S, Yoshida T, Ishihara H and Noguchi S: Determination of the specific activity of CDK1 and CDK2 as a novel prognostic indicator for early breast cancer. Ann Oncol 19: 68-72, 2008.
30. Roesley SNA, Suryadinata R, Morrish E, Tan AR, Issa SM, Oakhill JS, Bernard O, Welch DR and Šarčević B: Cyclin-dependent kinase-mediated phosphorylation of breast cancer metastasis suppressor 1 (BRMS1) affects cell migration. Cell Cycle 15: 137-151, 2016.
31. Pandithage R, Lilischkis R, Harting K, Wolf A, Jedamzik B, Lüscher-Firzlaff J, Vervoorts J, Lasonder E, Kremmer E, Knöll B and Lüscher B: The regulation of SIRT2 function by cyclin-dependent kinases affects cell motility. J Cell Biol 180: 915-929, 2008.
32. Weiss RH, Marshall D, Howard L, Corbacho AM, Cheung AT and Sawai ET: Suppression of breast cancer growth and angiogenesis by an antisense oligodeoxynucleotide to p21(Waf1/Cip1). Cancer Lett 189: 39-48, 2003.
33. Ma H, Jin G, Hu Z, Zhai X, Chen W, Wang S, Wang X, Qin J, Gao J, Liu J, *et al*: Variant genotypes of CDKN1A and CDKN1B are associated with an increased risk of breast cancer in Chinese women. Int J Cancer 119: 2173-2178, 2006.
34. Bianco S, Jangal M, Garneau D and Gévry N: LRH-1 controls proliferation in breast tumor cells by regulating CDKN1A gene expression. Oncogene 34: 4509-4518, 2015.
35. Nagata M, Kurita H, Uematsu K, Ogawa S, Takahashi K, Hoshina H and Takagi R: Diagnostic value of cyclin-dependent kinase/cyclin-dependent kinase inhibitor expression ratios as biomarkers of locoregional and hematogenous dissemination risks in oral squamous cell carcinoma. Mol Clin Oncol 3: 1007-1013, 2015.
36. Matsumura I, Tanaka H and Kanakura Y: E2F1 and c-Myc in cell growth and death. Cell Cycle 2: 333-338, 2003.
37. Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, Benz CC, Flury R, Dieterich H, Spyratos F, Eppenberger U and Eppenberger-Castori S: Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. Breast Cancer Res 9: R33, 2007.
38. Chen XZ, Cao ZY, Chen TS, Zhang YQ, Liu ZZ, Su YT, Liao LM and Du J: Water extract of *Hedyotis Diffusa* Willd suppresses proliferation of human HepG2 cells and potentiates the anticancer efficacy of low-dose 5-fluorouracil by inhibiting the CDK2-E2F1 pathway. Oncol Rep 28: 742-748, 2012.
39. Amundadottir LT, Johnson M, Merlino G, Smith GH and Dickson RB: Synergistic interaction of transforming growth factor alpha and c-myc in mouse mammary and salivary gland tumorigenesis. Cell Growth Differ 6: 737-748, 1995.
40. Singhi AD, Cimino-Mathews A, Jenkins RB, Lan F, Fink SR, Nassar H, Vang R, Fetting JH, Hicks J, Sukumar S, *et al*: MYC gene amplification is often acquired in lethal distant breast cancer metastases of unamplified primary tumors. Mod Pathol 25: 378-387, 2012.
41. Liu H, Radisky DC, Yang D, Xu R, Radisky ES, Bissell MJ and Bishop JM: MYC suppresses cancer metastasis by direct transcriptional silencing of αv and β3 integrin subunits. Nat Cell Biol 14: 567-574, 2012.