Developing a radiosensitivity gene signature for Caucasian patients with breast cancer

YE JI[1-3], QINGHUA JIANG[4], GUOHAO JIAN[1-3], HAITONG SUN[1-3], YAMIN WANG[4], HUALONG QIN[5], SHUYU ZHANG[3], JIANPING CAO[3] and ZAIXIANG TANG[1,2,6]

[1]Department of Biostatistics, School of Public Health, and [2]Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, Medical College of Soochow University, Suzhou, Jiangsu 215123; [3]School of Radiation Medicine and Protection and Collaborative Innovation Center of Radiation Medicine of Jiangsu Higher Education Institutions, Soochow University, Suzhou, Jiangsu 215006; [4]Department of Basic Science, Changzhou Vocational Institute of Engineering, Changzhou, Jiangsu 213164; [5]Department of Thoracic Surgery, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215006; [6]Center for Genetic Epidemiology and Genomics, Medical College of Soochow University, Suzhou, Jiangsu 215123, P.R. China

Received January 11, 2018; Accepted June 28, 2018

DOI: 10.3892/or.2018.6567

Abstract. Adjuvant radiotherapy is an important clinical treatment option for patients with breast cancer. However, for Caucasian patients, the clinical benefit of adjuvant radiotherapy can differ from African-American patients with respect to the overall survival. The goal of the current study was to develop a gene signature and to pre-identify patients likely to benefit from radiotherapy. Using publicly available breast cancer data from The Cancer Genome Atlas, a new cross-validation procedure was proposed for developing a gene signature and predicting radiosensitive patients. The results demonstrated that the predicted radiosensitive patients who received radiotherapy exhibited a significantly better survival, while the effect of radiotherapy was not significant for predicted non-radiosensitive patients. Further hierarchical cluster analysis revealed that the predicted sensitivity for each patient corresponded closely to the results of the cluster analysis. Collectively, the findings of the current study demonstrated that a radiosensitive molecular signature can be used to identify radiosensitive Caucasian patients with breast cancer.

Introduction

Breast cancer is a malignant disease that is estimated to affect >245,000 women each year in the USA alone, according to a recent report by the American Cancer Society (1). It was estimated that the mortality rate of breast cancer in 2017 will reach >40,500 women in the U.S. (2). The disease, which can also affect men, is caused by uncontrollable cell growth in the breast. Numerous treatment options currently exist for patients with breast cancer. Among the local treatments available, radiotherapy is an essential component of the therapeutic regimen for these patients. However, in clinical practice, not all patients benefit from radiotherapy. The need for radiation depends on several clinical factors, including the type of surgery, whether the cancer has spread to the lymph nodes or elsewhere in the body, and, in certain cases, the patient age.

While radiotherapy has a pronounced effect on cancer cells, it also affects healthy tissue in the area being treated. Whether a patient should be offered radiotherapy will depend on the individual situation, and the effects of treatment may also vary from one patient to another. The effect of radiotherapy on patients who had breast-conserving surgery is strongly positive, reducing the 10-year risk of any first recurrence and the 15-year risk of breast cancer-associated mortality according to a meta-analysis involving >10,000 patients (3). However, another meta-analysis suggested that the cumulative incidence of breast-cancer-specific mortality and overall survival were not significantly improved for the patients who had breast-conserving surgery plus radiotherapy in the Swedish ductal carcinoma *in situ* trial (4). A recent meta-analysis also indicated that radiotherapy could slightly reduce the risk of local relapse in older patients if breast cancer was diagnosed early. However, reduced local relapse cannot translate into significant survival benefits (5). These previous studies have not provided uniform, positive results regarding radiotherapy for breast cancer patients.

Considerable controversy remains with respect to the use of radiotherapy for breast cancer (6). Firstly, it can be difficult to predict the patient's sensitivity to radiotherapy. In addition, late and chronic toxicity, as well as other side-effects of radiotherapy, such as breast pain and dermatitis, are often a concern (7,8). In the era of precision medicine, biology-driven

Correspondence to: Dr Zaixiang Tang, Department of Biostatistics, School of Public Health, Medical College of Soochow University, 199 Renai Road, SIP, Suzhou, Jiangsu 215123, P.R. China
E-mail: tangzx@suda.edu.cn

Key words: gene signature, radiosensitivity, radiotherapy, sensitivity prediction, breast cancer

personalized radiotherapy in breast cancer using biomarkers to guide exclusive radiotherapy and combination therapy have started to emerge (9,10). Several biomarkers based on single gene or molecular expression with specificity in predicting the survival benefit have been developed and validated (11-15). Gene signatures (which refers to a group of genes), have been used to identify radiosensitive patients in numerous types of cancer, including glioblastoma, cervical, colorectal, and head and neck cancer (16-20). Only a few effective radiosensitive gene signatures have been developed for breast cancer (11,21,22). A new adaptive procedure for simultaneously developing and validating the gene signature was proposed in the current study. It is argued that if a powerful radiosensitivity signature is developed, then it may possible to effectively identify the right patients with breast cancer to receive radiotherapy.

In the present study, the RNAseq data for Caucasian patients with breast cancer obtained from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) were utilized to develop a gene signature for predicting radiosensitive patients. Since only one dataset was obtained from TCGA, it was difficult to identify another independent dataset with survival outcomes and the same RNAseq data to perform independent-sample validation. Furthermore, due to the difficulty of collecting clinical samples and the large number of potential genes available for analysis, the development of a reliable diagnostic classifier using early nonrandomized phase II data is often not feasible. To overcome these difficulties, an internal cross-validation was performed via the cross-validated adaptive signature design that combined the gene signature development and the validation test in a single trial, as originally introduced by Freidlin *et al* (23), Freidlin and Simon (24), and Tang *et al* (25). The previous procedure of selecting informative genes was improved in the present study, and this novel approach was extended to the proportional hazard model. Thus, a radiosensitive gene signature was developed for predicting radiosensitive Caucasian patients with breast cancer.

Materials and methods

Study samples. The clinical data and normalized RNAseq expression data were downloaded from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/; updated August 2017) through the R package TCGA assembler (26). The raw clinical data for 1,097 patients were available. Survival and radiotherapy information from several clinical files were combined, including the patients' clinical records and follow-up information. After removing patients with no clear radiotherapy or survival information, 1,007 patients were selected for further analysis. The expression data originally included 1,219 patients with expression data on 20,530 genes. Duplicated patients and genes without clear names were first removed from the raw data, and then 1,097 patients with 20,502 gene expression values were obtained. Subsequently, only clinical data for Caucasian breast cancer patients (n=702) were extracted and merged with the expression data. Thus, a dataset with 700 Caucasian breast patients was obtained. Genes with a maximum expression value of 20 were excluded as they exhibited almost no expression. The gene expression values for some patients were zero, called zero expression. For a gene, if >30% patients had zero expression, this gene would be removed. Next, we calculated the variance of gene expression values for each gene. We then kept the top 80% genes according to the variances for these genes. These data cleaning steps resulted in a total of 700 patients with 13,516 gene expression profiles for final analysis. The procedure for data processing is summarized in Fig. 1. The missing values in clinical data were filled in by multiple imputation using the R package *mice (https://cran.r-project.org/web/packages/mice/index.html)*. The cleaned clinical data are summarized in Table I.

Definition of radiosensitivity and gene signature. The definition of radiosensitivity differs from that in laboratory research based on cell and tissues. In the present study, radiosensitive patients were defined as those who experienced better survival after receiving radiotherapy. To develop a radiosensitive gene signature for identifying radiosensitive patients, a model was established using the following assumptions: i) there is a subset of s sensitive genes that significantly interact with radiotherapy; ii) the effect of the interaction between gene and radiotherapy usually strongly contributes to the overall survival. The survival benefit of radiotherapy is associated with these predictive genes through the Cox proportional hazards model, as follows:

$$h(t|X)=h_0(t)\exp(r\lambda+x_1b_1+x_2b_2+\cdots+x_sb_s+rx_1i_1+rx_2i_2+\cdots+rx_si_s)$$

In this equation, $h_0(t)$ is the baseline hazard function, λ is the effect of radiotherapy, r is an indicator for radiotherapy (with 1 indicating radiotherapy and 0 otherwise), s is the number of sensitive genes, x_1 to x_s are the gene expression values, b_1 to b_S are the main effects of these s sensitive genes, and i_1 to i_s are the radiotherapy-expression interaction effects that reflect the degree to which the effect of radiotherapy on survival is influenced by the expression levels of sensitive genes.

From a single sensitive gene, the hazard ratio (HR) can be estimated by $\exp(r\lambda + x_jb_j + rx_ji_j)$. If the radiotherapy-expression interaction effect value is negative, patients who overexpress the sensitive gene will have a higher survival probability with radiotherapy than without radiotherapy, since a small HR (<1) is obtained. It was assumed that a fraction of the patient population overexpresses several of these sensitive genes. Then, these patients would be expected to have a relatively high probability of survival. These patients were referred to as radiosensitive patients. These sensitive genes were termed the radiosensitive gene signature.

Gene signature development and cross-validation procedure. Freidlin and Simon (24) in 2005 and Freidlin *et al* (23) in 2010 developed a novel cross-validated adaptive signature design to identify sensitive patients in clinical trial for binary outcome. Following the framework of these studies, this approach was modified and extended to the proportional hazards model, and used to develop radiosensitive gene signature for sarcoma data in a previous study (25). In the present study, the procedure was further improved. An updated three-step K-fold cross-validated procedure for gene signature development was used, including the training, prediction and validation steps, and this procedure is described herein.

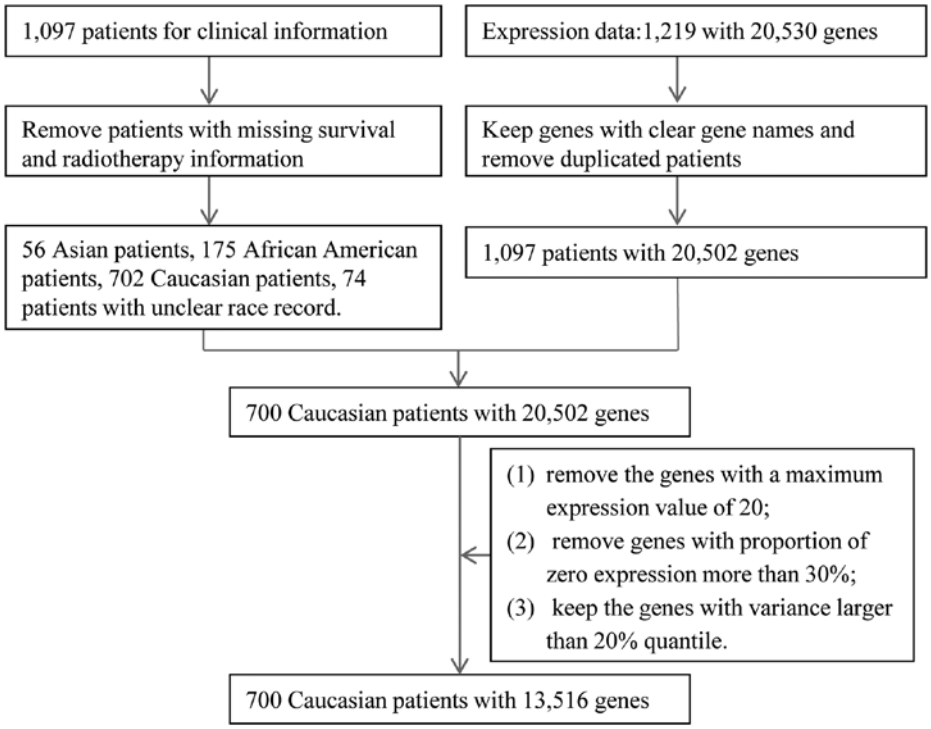

Figure 1. Workflow of data cleaning steps.

Training step (step 1). The data were randomly split into *K* parts with the same sample size (usually *K*=10). Next, -*K*-1 part patients were used as training data to fit models and predict the radiosensitive patients in the left-out part (validation data). In the training data, a Cox proportional hazards model was fit for each gene (*j*):

$$h(t \mid X) = h_0(t)\exp(r\lambda + x_j b_j + r x_j i_j)$$

The genes with negative interaction effects (i_j) were selected for further prediction. Subsequently, P-values for all interaction effects were used to rank the genes.

Prediction step (step 2). The top significant *g* sensitive genes were used to build a gene signature and calculate an index, termed the nominal HR (nHR), as follows:

$$nHR = \exp(r\lambda + \sum_{i}^{g}(x_j b_j + r x_j i_j)),$$

All these *g* genes for patients in the validation data (*k*-th part) were combined to calculate nHR. In this equation, λ referred to the main effect of radiotherapy. Patients in the validation set who had an nHR value lower than a specified threshold *R* Each patient only appeared once in each of the validation datasets. Subsequent to these two steps, each patient was classified as either radiosensitive or non-radiosensitive.

Validation step (step 3). For predicted radiosensitive patients, a log-rank test was then performed to assess the difference in survival between the radiotherapy and non-radiotherapy groups at a specified significant level of P<0.05. A significantly improved survival indicated that radiotherapy would be beneficial to radiosensitive patients. In these cases, the prediction of radiosensitive patients was also considered accurate, and the gene signature including *g* genes was considered effective.

The procedure outlined earlier in this study has two key tuning parameters, *g* and *R*, in the prediction step. The optimal values of the tuning parameters *g* and *R* are not usually known in advance. For this reason, all possible combinations of *g* and *R* can be tried and tested. A nested inner loop of the *K*-fold cross-validation approach can be used on the training data to select the best tuning parameter values without affecting the statistical validity of the procedure. Similar procedures for the two key tuning parameters (25) and for more tuning parameters have also been reported in previous studies (23,24).

In the Training step of the present study, including (*K*-1) part patients, the tuning parameters *g* and *R* were selected empirically by selecting the values that provided the highest power to predict sensitive patients on a set of possible combinations of *g* and R. In practice, the following approach based on 10-fold cross-validation is recommended for the selection of the best combination from a set of M possible combinations using the Training step (*K*-1) patients only.

Part 1 of the approach involves randomly splitting the data (*K*-1 part patients only) to *T* parts with the same sample size, and the use of *T*=10 is recommended in this step. Next, the T^{th} part patients are removed, and step 1 of the three-step procedure described earlier is performed on the remaining patients. Using step 2 of the three-step procedure, whether the *t*-th part patients are classified as sensitive is determined according to different possible tuning parameter combinations. For the two tuning parameters, top *g* genes from 1 to 200 significant genes are empirically tested. These genes were usually significantly interacted with gene expression, with significant *i* effects.

Table I. Clinical characteristics of patients, and results of univariate and multivariate Cox regression analysis (n=700).

		Univariate analysis		Multivariable analysis	
Characteristics	No. of patients	HR (95% CI)	P-values	HR (95% CI)	P-values
Age (range, 26-90), years					
<60	364	1			
≥60	336	2.40 (1.50-3.84)	$2.51x10^{-4}$	3.20 (1.77-5.78)	<0.001
History of malignancy					
No	663	1			
Yes	36	2.34 (1.01-5.44)	0.048	2.51 (0.92-6.89)	0.073
NA	1				
Margin status					
Negative	613	1			
Positive	43	0.42 (0.13-1.36)	0.148	1.31 (0.56-3.08)	0.533
Close	19	0.78 (0.21-2.91)	0.717	3.18 (0.88-11.56)	0.078
NA	25				
Histological type					
Lobular	165	1			
Ductal	461	0.85 (0.49-1.48)	0.568	1.26 (0.63-2.54)	0.519
Other	61	1.12 (0.51-2.46)	0.778	1.92 (0.79-4.66)	0.148
NA	13				
T stage					
T1	194	1			
T2	392	1.57 (0.85-2.89)	0.151	1.43 (0.70-2.93)	0.327
T3	96	2.32 (1.13-4.78)	0.022	1.25 (0.49-3.23)	0.640
T4	17	4.02 (1.51-10.69)	0.005	2.61 (0.76-8.91)	0.126
NA	1				
N stage					
N0	340	1			
N1	223	1.61 (0.61-2.85)	0.105	1.62 (0.80-3.28)	0.177
N2	79	3.33 (1.67-6.63)	0.001	5.66 (2.23-14.37)	<0.001
N3	48	4.52 (2.06-9.89)	$1.64x10^{-4}$	4.70 (1.63-13.53)	0.004
NA	10				
M stage					
M0	597	1			
M1	11	5.16 (2.49-10.68)	$9.76x10^{-6}$	1.59 (0.55-4.64)	0.394
NA	92				
Surgery type					
Simple mastectomy	139	1			
Lumpectomy	165	1.05 (0.47-2.34)	0.913	0.62 (0.24-1.59)	0.317
Modified radical	232	1.62 (0.80-3.28)	0.183	0.84 (0.35-2.00)	0.689
Other	120	0.86 (0.38-1.94)	0.722	0.66 (0.26-1.65)	0.371
NA	44				
ER					
ER^-	125	1			
ER^+	537	1.27 (0.68-2.38)	0.460	1.24 (0.53-2.93)	0.623
NA	38				
PR					
PR^-	192	1			
PR^+	467	1.00 (0.60-1.67)	0.986	0.55 (0.27-1.12)	0.099
NA	41				
HER					
HER^-	336	1			

Table I. Continued.

Characteristics	No. of patients	Univariate analysis		Multivariable analysis	
		HR (95% CI)	P-values	HR (95% CI)	P-values
HER^{+-}	135	1.00 (0.51-1.97)	0.998	0.84 (0.31-2.27)	0.724
HER^{+}	91	1.17 (0.51-2.70)	0.705	0.97 (0.43-2.16)	0.937
NA	108				
Chemotherapy					
No	102	1			
Yes	597	0.41 (0.22-0.77)	0.006	0.30 (0.14-0.65)	0.002
NA	1				
Radiotherapy					
No	295	1			
Yes	405	0.69 (0.44-1.09)	0.113	0.66 (0.37-1.19)	0.169

HR, hazard ratio; 95% CI, 95% confidence interval; NA, data not available; ER, estrogen receptor; PR, progesterone receptor; HER, human epidermal growth factor receptor.

Then, *R* ranging between 0.005 and 0.5 is also assessed in increments of 0.005. Subsequently, a total M=20,000 possible tuning parameter combinations are examined. The procedure is repeated for values of *t*=1 to 10, allowing for each study patient to be predicted exactly one time under the tuning parameter combinations. All M=20,000 possible tuning parameter combinations are attempted, and M subsets of sensitive patients are then formed, each corresponding to a set of tuning parameters.

In part 2 of the approach, the survival among the predicted sensitive patients who received the radiotherapy and the predicted sensitive patients who did not received the radiotherapy is compared for each of the M subsets. Subsequently, the tuning parameter combination that provides the smallest P-value in log-rank test is selected. This tuning parameter combination would then be used to filter the sensitive patients on validation patients at step 2 (Prediction step).

The approach described in the present study preserves the validity of predicting radiosensitive patients in the K^{th} subset, as only the data from the (*K*-1) parts are used to determine the tuning parameters. This procedure is a nested inner loop of *K*-fold cross-validation applied only in the Training step (*K*-1) patients. In this procedure, *T*=10 is recommended, since 10-fold cross-validation usually has a small and stable bias and error (27). Leave-one-out cross-validation (LOOCV) could also be applied to obtain a stable result (27). However, LOOCV can be very time-consuming to implement.

In addition, for different (*K*-1) patients in the Training step, the tuning parameters (*g*, *R*) may differ. Theoretically, the reselection of the tuning parameters (*g*, *R*) or significant genes for different loops of the cross-validation is essential to the validity of the approach (28). However, this does not suggest that the classifications and selection are unstable or that the classifier will provide an accurate prediction for independent data. Good genomic signatures are generally not unique (22,23). As described by Freidlin *et al* (23), in order to save computational time, the first cross-validation subset could be used to select the turning parameters (*g*, *R*).

Statistical analysis. Log-rank test was used to compare the survival curve between two different groups. Then, P-value was obtained by the log-rank test. Univariate and multivariate Cox regression were performed to evaluate various clinical factors associated with overall survival, using the R package 'Survival' (https://cran.r-project.org/web/views/Survival.html). The Wald test was used to get P value. R package 'rms' (https://cran.r-project.org/web/packages/rms/) was used to plot the survival curve. In addition, hierarchical clustering analysis was also used in our analysis. Agreement analysis was also performed to calculate the kappa coefficient. P<0.05 was considered to indicate a statistically significant difference.

Results

Survival analysis of clinical information. Asian patients were excluded in our later analysis, due to small sample size. Only 56 patients were obtained. African-American patients were also excluded due to the significantly improved overall survival observed under radiotherapy treatment for these patients (Fig. 2A). The clinical information of the 700 Caucasian breast cancer patients included in the present study is summarized in Table I. Univariate and multivariable analyses were subsequently performed to investigate the associations among the overall survival and clinical factors. The results indicated that radiotherapy was not a significant clinical factor associated with the overall survival of Caucasian patients (Table I and Fig. 2B). Among the tested parameters, only the age, clinical N stage and chemotherapy were found to be significantly associated with overall survival.

Development of a radiosensitive gene signature. In order to develop the radiosensitive gene signature, the proposed procedure described in the Materials and methods was implemented. Usually, when the three-step 10-fold cross-validation procedure is performed, then 10 different radiosensitive gene signatures may be developed, due to the reselection of the

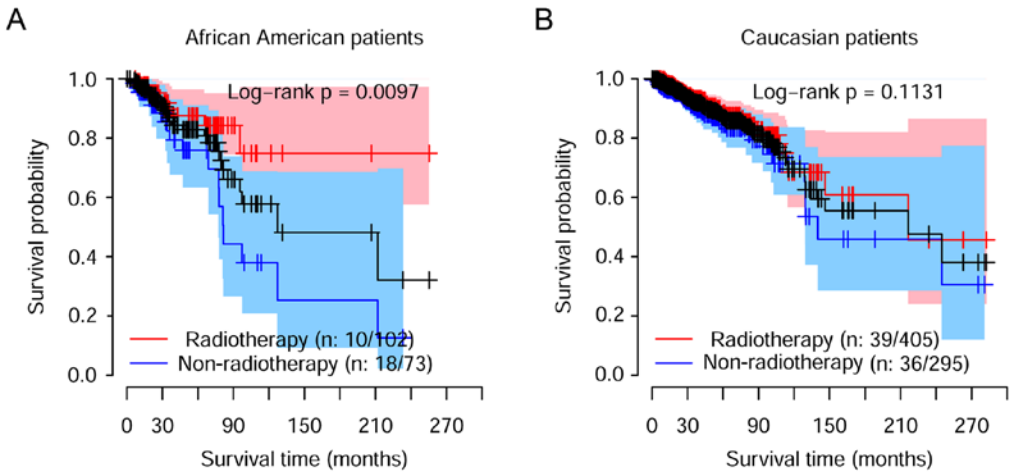

Figure 2. Survival curves of (A) African-American and (B) Caucasian breast cancer patients receiving radiotherapy and non-radiotherapy treatments. Values in the parentheses indicate the number of mortality cases over the sample size for each group.

Table II. Genes (n=30) included in the radiosensitive gene signature and their interaction effect with radiotherapy.

ID	Gene name	Gene effect	P-value	Radiotherapy effect	P-value	Interaction effect	P-value
1	ERMP1	0.7625 (0.1643)	<0.0001	-0.3684 (0.2355)	0.1178	-0.9536 (0.2514)	0.0001
2	PAFAH1B2	0.5700 (0.1793)	0.0015	-0.3345 (0.2407)	0.1647	-0.8705 (0.2484)	0.0005
3	SMARCA2	0.2629 (0.1264)	0.0376	-0.4324 (0.2437)	0.0760	-0.8241 (0.2376)	0.0005
4	ABCB7	0.4112 (0.1447)	0.0045	-0.3764 (0.2482)	0.1293	-0.7819 (0.2289)	0.0006
5	UNC119B	0.4640 (0.1156)	0.0001	-0.3097 (0.2500)	0.2154	-0.8101 (0.239)	0.0007
6	SFMBT1	0.2491 (0.1482)	0.0927	-0.5282 (0.2575)	0.0402	-0.909 (0.2709)	0.0008
7	APPL1	0.3662 (0.1740)	0.0353	-0.4189 (0.239)	0.0796	-0.8143 (0.2475)	0.0010
8	LUC7L2	0.1615 (0.1576)	0.3054	-0.446 (0.2427)	0.0661	-0.8455 (0.2581)	0.0011
9	LYPD6B	0.3162 (0.1047)	0.0025	-0.4338 (0.2589)	0.0938	-0.9623 (0.2975)	0.0012
10	ZNF445	0.2786 (0.1447)	0.0541	-0.4609 (0.2413)	0.0561	-0.8018 (0.2483)	0.0012
11	PRKCE	0.3720 (0.1386)	0.0073	-0.3507 (0.2354)	0.1362	-0.7738 (0.2399)	0.0013
12	DAG1	0.3849 (0.1520)	0.0114	-0.3986 (0.2428)	0.1006	-0.7417 (0.231)	0.0013
13	ERLIN1	0.2730 (0.1347)	0.0427	-0.3915 (0.2430)	0.1072	-0.744 (0.2318)	0.0013
14	ZNF780A	0.3305 (0.1277)	0.0096	-0.4579 (0.2419)	0.0583	-0.96 (0.3011)	0.0014
15	PMPCB	0.0422 (0.1502)	0.7788	-0.5548 (0.2560)	0.0302	-0.838 (0.2629)	0.0014
16	PCNP	-0.0037 (0.1347)	0.9782	-0.5086 (0.2499)	0.0419	-0.7305 (0.2317)	0.0016
17	SACM1L	0.2809 (0.1658)	0.0903	-0.4364 (0.2469)	0.0771	-0.7705 (0.2444)	0.0016
18	EPO	0.5185 (0.1232)	<0.0001	-0.3322 (0.2344)	0.1564	-0.5009 (0.161)	0.0019
19	CTNNB1	0.3909 (0.1806)	0.0305	-0.4398 (0.2415)	0.0686	-0.9346 (0.3013)	0.0019
20	NOC3L	-0.0199 (0.1742)	0.9089	-0.5609 (0.2578)	0.0296	-0.9785 (0.3164)	0.0020
21	ATAD1	0.1935 (0.1537)	0.2082	-0.4187 (0.2389)	0.0797	-0.7345 (0.2377)	0.0020
22	MAGI1	0.3359 (0.1554)	0.0306	-0.3714 (0.2389)	0.1199	-0.7802 (0.2526)	0.0020
23	ZZZ3	0.0486 (0.1466)	0.7399	-0.5104 (0.2511)	0.0421	-1.0561 (0.343)	0.0021
24	ZNF621	0.5117 (0.1846)	0.0056	-0.4001 (0.2395)	0.0949	-0.8123 (0.2641)	0.0021
25	NEK4	0.2223 (0.1600)	0.1648	-0.4859 (0.2509)	0.0528	-0.7855 (0.2567)	0.0022
26	LYPD6	0.3941 (0.1401)	0.0049	-0.473 (0.2565)	0.0651	-1.129 (0.3711)	0.0023
27	WDR48	0.2213 (0.0944)	0.0191	-0.4068 (0.2383)	0.0878	-0.6069 (0.2021)	0.0027
28	EFCAB14	0.4517 (0.1670)	0.0068	-0.3418 (0.2409)	0.1559	-0.6998 (0.2349)	0.0029
29	MLH1	0.6132 (0.1682)	0.0003	-0.3237 (0.2400)	0.1776	-0.7012 (0.2356)	0.0029
30	NMD3	0.3923 (0.1507)	0.0093	-0.3597 (0.2381)	0.1309	-0.7238 (0.2439)	0.0030

significant genes for each of the 10 training datasets. For the cross-validation procedure, the parameter reselection is essential (28). This does not indicate that the gene signature is unstable or that the classifier will not accurately predict

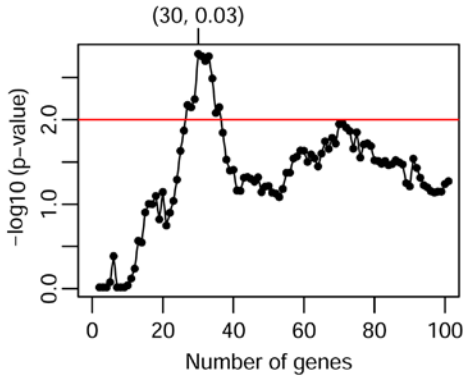

Figure 3. Profile of -log10 (P-values) obtained by log-rank tests between the radiotherapy and non-radiotherapy groups of predicted radiosensitive patients. The gene signature, including top 30 significant genes with a threshold of R=0.03, is found to provide an accurate prediction with the smallest P-values ($P=1.65x10^{-3}$).

independent data. Genomic signatures are generally not unique (22,23). As suggested by Freidlin *et al* (23), in order to save computational time, the first cross-validation training dataset can be used to select the tuning parameter. This suggestion was followed in the present study, and the first training dataset was used to select the tuning parameters, *g* and *R*. As a result, the top *g*=30 significant genes and a threshold R of 0.03 were included in the gene signature for predicting the radiosensitive patients. Under this tuning parameter, the minimum, log-rank tests were performed to compare the survival rate between the patients who received radiotherapy and the patients who did not receive radiotherapy (Fig. 4). The survival curves for predicted radiosensitive patients who received radiotherapy and non-radiotherapy treatment are shown in Fig. 4A, while the comparison between radiotherapy and non-radiotherapy for predicted non-radiosensitive patients are displayed in Fig. 4B. Furthermore, the survival among radiosensitive and non-radiosensitive patients who received radiotherapy treatment was compared, as shown in Fig. 4C. These results suggested that the predicted radiosensitive patients strongly benefited from radiotherapy. For the patients predicted to be non-radiosensitive, there was no significant difference in survival between radiotherapy and non-radiotherapy treatment. In addition, there was no significant difference in the survival of predicted radiosensitive and non-radiosensitive patients when they all did not receive radiotherapy treatment, as shown in Fig. 4D. Taken together, as expected, the radiosensitive gene signature was able to identify radiosensitive patients accurately.

The aforementioned analysis provided results by log-rank test to demonstrate the differences between two groups in the univariate analysis. Subsequently, multivariable analysis was further performed to assess the effect of radiotherapy on overall survival for the predicted radiosensitive and non-radiosensitive patients. The adjusted factors included the

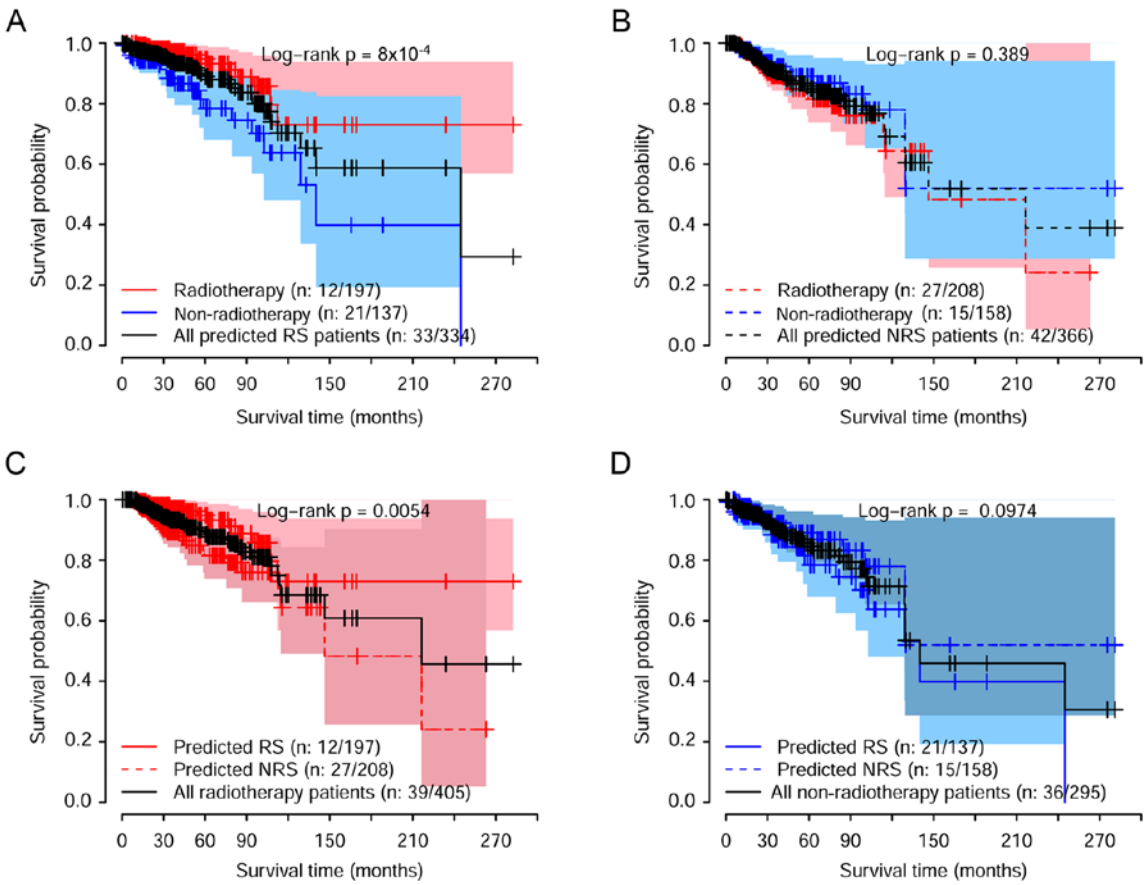

Figure 4. Survival curves of the predicted (A) RS and (B) NRS patients receiving radiotherapy and non-radiotherapy. Comparison of survival curves of RS and NRS patients who (C) received radiotherapy or (D) did not receive radiotherapy. The colored areas denote the 95% confidence intervals of the survival rate. Values in the parentheses indicate the number of mortality cases over the sample size for each group. RS, radiosensitive; NRS, non-radiosensitive.

Table III. HR estimation for patients receiving RT vs. NRT treatment, and for patients predicted to be RS vs. NRS.

A. RT vs. NRT

Patients	Data	HR (95% CI)	P-value
Predicted RS	Raw	0.32 (0.15-0.64)	0.0015
	Adjusted	0.15 (0.05-0.45)	0.0007
Predicted NRS	Raw	1.32 (0.70-2.48)	0.3911
	Adjusted	1.80 (0.72-4.52)	0.2123

B. Predicted RS vs. NRS

Patients	Data	HR (95%CI)	P-value
Received RT	Raw	0.39 (0.20-0.78)	0.0073
	Adjusted	0.36 (0.16-0.81)	0.0140
Received NRT	Raw	1.74 (0.90-3.38)	0.1000
	Adjusted	1.89 (0.80-4.47)	0.1500

P-values were estimated by Wald test. The adjusted factors are age, chemotherapy, history of other malignancy, histologic type, margin status, T, N, M stage. HR, hazard ratio; 95% CI, 95% confidence interval; RT, radiotherapy; NRT, non-radiotherapy; RS, radiosensitive; NRS, non-radiosensitive.

age (at initial pathologic diagnosis), chemotherapy, history of other malignancy, histologic type, margin status, and clinical T, N, and M stages. Table III lists the univariate and multivariable analysis results. For predicted radiosensitive patients, radiotherapy strongly improved the overall survival, with raw and adjusted HR values of 0.32 [95% confidence interval (CI), 0.15-0.64] and 0.15 (95% CI, 0.05-0.45), respectively. By contrast, for predicted non-radiosensitive patients, radiotherapy may not be an effective clinical treatment, with a nonsignificant adjusted HR of 1.80 (95% CI, 0.72-4.52) detected. Among the patients who received radiotherapy, the radiosensitive patients had a significantly higher probability of survival as compared with the non-radiosensitive patients, with a significant adjusted HR value of 0.36 (95% CI, 0.16-0.81). In addition, there was no significant difference in survival between the predicted radiosensitive and non-radiosensitive patients when none of them received radiotherapy. The univariate and multivariable analysis results suggested that the prediction of radiosensitive patients was accurate and effective, and that the gene signature was effective and accurate for predicting the radiosensitive patients.

Gene signature and cluster analysis. The gene signature included 30 genes, all of which significantly interacted with radiotherapy. According to the procedure of developing a gene signature, higher gene expression was associated with stronger sensitivity, which indicated better overall survival. Therefore, the pattern of expression of the selected 30 genes may be correlated with the prediction of radiosensitivity. The expression pattern of the selected 30 genes was extracted to perform hierarchical clustering analysis using the R package *pheatmap (https://cran.r-project.org/web/packages/pheatmap/index.html)*. As shown in Fig. 5, the 700 patients were classified into two groups. The predicted radiosensitive and non-radiosensitive patients were denoted by the blue and yellow bars, respectively. The results indicated that the predicted radiosensitive and non-radiosensitive patients closely matched the left and right branches of hierarchical cluster analysis, respectively. Sensitivity prediction and cluster analysis demonstrated exact matches for ~83% of patients. Agreement analysis was also performed between the results of the methods. The kappa coefficient value of 0.66 with P<0.001 suggested that the gene signature facilitated sensitivity prediction.

Discussion

Radiotherapy is a common clinical method used in the treatment of breast cancer. In the present study, the TCGA breast cancer dataset was downloaded, and survival analysis was performed using data of Caucasian breast cancer patients. The results suggested that radiotherapy did not significantly improve the overall survival. We then developed a radiosensitive gene signature for Caucasian patients with breast cancer. However, only one dataset had sufficient clinical information. For this reason, in accordance with the procedure published by Freidlin and Simon (24) in 2005 and Freidlin *et al* (23) in 2010, the adaptive gene signature development method for predicting which Caucasian patients are radiosensitive was proposed and updated in the present study. The method combined the development of a gene signature and validation into a single adaptive inner-loop procedure. The results suggested that this gene signature was powerful. From the result of Fig. 4A, we can see that for the predicted radiosensitive patients, the overall survival was significant different between the radiotherapy group and non-radiotherapy group. The overall survival was significantly better for predicted sensitive patients compared with predicted non-sensitive patients (Fig. 4C). For predicted non-radiosensitive patients, there was no significant difference in the overall survival between the radiotherapy group and non-radiotherapy group (Fig. 4B).

In the present study, new tumor event information obtained from TCGA, which included local recurrence, new primary tumor and distant metastasis tumor, was summarized. For these patients who received radiotherapy treatment, there was a lower rate of new tumor events at 22.3% (43/150) in the predicted radiosensitive group, compared with 26.6% (54/149) in the predicted non-radiosensitive group. This result partially supported the prediction of radiosensitive patients.

The current study not only developed a radiosensitive gene signature, but also identified genes that may be associated with the molecular foundation of breast cancer. For instance, a recent study suggested that *ERMP1* is broadly expressed in a high percentage of breast, colorectal, lung and ovary cancer cases, regardless of their stage and grade. This gene may function as a molecular starter to trigger the survival response induced by extracellular stresses (29). Using the metabolic mapping platforms, researchers also identified that *PAFAH1B3* may act as a key metabolic driver of breast cancer pathogenicity that is upregulated in primary human breast tumors and correlated with poor prognosis (30). The majority of other genes identified in the

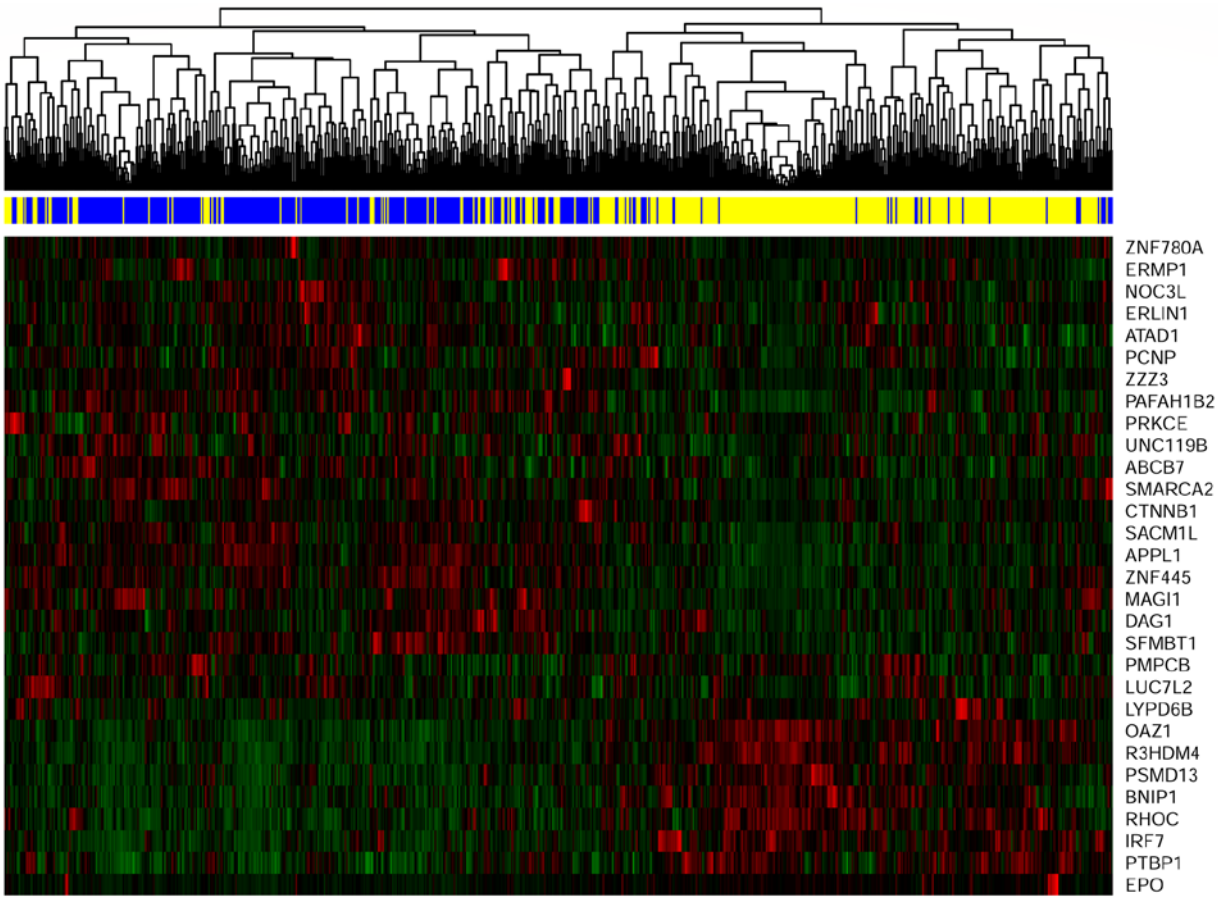

Figure 5. Hierarchical clustering analyses. Hierarchical clustering was used to determine the expression pattern of the selected 30 genes. The top blue and yellow bands denote the predicted radiosensitive and non-radiosensitive patients, respectively.

current study have also been reported to be involved in various types of cancer and diseases. These include *ABCB7* in myelodysplastic syndromes (31), *SFMBT1* in cervical cancer (32), and *APPL1* in gastric cancer (33,34). These results may provide helpful evidence for further research into breast cancer. Furthermore, it is worth noting that these genes exhibited significant interaction with radiotherapy, which indicates that these genes may also be associated with radiosensitivity. However, the associations between the expression of these genes and radiotherapy treatment have not received attention by researchers. The current study results may also provide evidence for further basic research on radiosensitivity.

To apply the gene signature identified in the present study for predicting novel radiosensitive cases, the HR for each gene must be calculated using the standard expression value of RNAseq according to the equation $\exp(r\lambda + x_j b_j + r x_j i_j)$. In this equation, r equals 1, denoting radiotherapy, and x_j is the expression level. Other coefficients of the formula for each gene are listed in Table II. The product of these HR (nHR) values was then compared with a threshold of 0.03. In the case where a patient exhibits an nHR value lower than the threshold, then this patient can be identified as radiosensitive and would be expected to gain an overall survival benefit.

It would be ideal to use clinical trials to develop and validate the gene signature for predicting patients who are most likely to respond to radiotherapy. However, there are often several barriers; for example, the gene signature itself may not be available by the beginning of the trial. In the current study, a new adaptive gene signature development procedure was improved and proposed. The adaptive design described in the present study may be useful in such situations, only one dataset obtained, offering the development and validation of a gene signature in one study (23,24). In the inner-loop procedure, 10-fold cross-validation is recommended, since it permits the maximization of the portion of study patients contributing to the development of the diagnostic signature and the minimization of prediction error (27). In addition to 10-fold cross-validation, a split-sample method and LOOCV are often mentioned in internal validation; however, the implementation of LOOCV can be time-consuming (27). In the present study, 10-fold cross-validation was implemented to develop a gene signature and for further validation.

Although the adaptive method was proposed and used in previous studies (25,35,36), the procedure in the Training step was further updated in the present study. Only significant genes with negative interaction effects were selected for further prediction. This selection strategy increased the power of the gene signature, since the genes included in the signature were all positively associated with improved overall survival.

In conclusion, we proposed a new adaptive gene signature development procedure. Using the proposed method, we developed a radiosensitive gene signature for breast cancer. The result showed that, compared with predicted

non-radiosensitive patients, the predicted radiosensitive patients had a better survival when they received radiotherapy. This result suggested the proposed method was effective, and the gene signature for radiosensitive prediction was accurate.

Acknowledgements

The authors acknowledge the contributions of the TCGA Research Network.

Funding

This study was supported by the National Natural Science Foundation of China (grant nos. 81773541 and 81573253, awarded to ZT), Key Investigation and Development Program of China (2016YFC0904700 and 2016YFC0904702, awarded to JC), and a project funded by Changzhou Vocational Institute of Engineering (grant no. CDGZ2015030, awarded to QJ).

Availability of data and materials

Data were downloaded from https://cancergenome.nih.gov/ through the R package TCGA assembler.

Authors' contributions

ZT, QJ, JC and YJ participated in the study conception and design. YJ, GJ, HS and YW performed real data analysis. YJ, QJ, ZT, SZ and HQ wrote the manuscript. SZ and HQ were also involved in the conception of the study. YJ, QJ and ZT revised and edited the manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. American Cancer Society: Cancer facts and figures 2016. Journal 2016.
2. Siegel RL, Miller KD and Jemal A: Cancer statistics, 2017. CA Cancer J Clin 67: 7-30, 2017.
3. Early Breast Cancer Trialists' Collaborative Group (EBCTCG); Darby S, McGale P, Correa C, Taylor C, Arriagada R, Clarke M, Cutter D, Davies C, Ewertz M, Godwin J, *et al*: Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: Meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet 378: 1707-1716, 2011.
4. Wärnberg F, Garmo H, Emdin S, Hedberg V, Adwall L, Sandelin K, Ringberg A, Karlsson P, Arnesson LG, Anderson H, *et al*: Effect of radiotherapy after breast-conserving surgery for ductal carcinoma in situ: 20 years follow-up in the randomized SweDCIS Trial. J Clin Oncol 32: 3613-3618, 2014.
5. Huang XZ, Chen Y, Chen WJ, Zhang X, Wu CC, Zhang CY, Sun SS and Wu J: Effect of radiotherapy after breast-conserving surgery in older patients with early breast cancer and breast ductal carcinoma in situ: A meta-analysis. Oncotarget 8: 28215-28225, 2017.
6. Krug D, Baumann R, Budach W, Dunst J, Feyer P, Fietkau R, Haase W, Harms W, Piroth MD, Sautter-Bihl ML, *et al*: Current controversies in radiotherapy for breast cancer. Radiat Oncol 12: 25, 2017.
7. Ahmadloo N, Kadkhodaei B, Omidvari S, Mosalaei A, Ansari M, Nasrollahi H, Hamedi SH and Mohammadianpanah M: Lack of prophylactic effects of aloe vera gel on radiation induced dermatitis in breast cancer patients. Asian Pac J Cancer Prev 18: 1139-1143, 2017.
8. Karlsson P: Postoperative radiotherapy after DCIS: Useful for whom? Breast 34 (Suppl 1): S43-S46, 2017.
9. Hirst DG and Robson T: Molecular biology: The key to personalised treatment in radiation oncology? Br J Radiol 83: 723-728, 2010.
10. Tsoutsou PG, Durham AD and Vozenin MC: A need for biology-driven personalized radiotherapy in breast cancer. Breast Cancer Res Treat 167: 603-604, 2018.
11. Jang BS and Kim IA: A radiosensitivity gene signature and PD-L1 status predict clinical outcome of patients with invasive breast carcinoma in The Cancer Genome Atlas (TCGA) dataset. Radiother Oncol 124: 403-410, 2017.
12. Lai Y, Chen Y, Lin Y and Ye L: Down-regulation of LncRNA CCAT1 enhances radiosensitivity via regulating miR-148b in breast cancer. Cell Biol Int 42: 227-236, 2018.
13. Liu G, Wang H, Zhang F, Tian Y, Tian Z, Cai Z, Lim D and Feng Z: The effect of VPA on increasing radiosensitivity in osteosarcoma cells and primary-culture cells from chemical carcinogen-induced breast cancer in rats. Int J Mol Sci 18: E1027, 2017.
14. Zhang X, Li Y, Wang D and Wei X: miR-22 suppresses tumorigenesis and improves radiosensitivity of breast cancer cells by targeting Sirt1. Biol Res 50: 27, 2017.
15. Zhou ZR, Yang ZZ, Wang SJ, Zhang L, Luo JR, Feng Y, Yu XL, Chen XX and Guo XM: The Chk1 inhibitor MK-8776 increases the radiosensitivity of human triple-negative breast cancer by inhibiting autophagy. Acta Pharmacol Sin 38: 513-523, 2017.
16. Salendo J, Spitzner M, Kramer F, Zhang X, Jo P, Wolff HA, Kitz J, Kaulfuß S, Beißbarth T, Dobbelstein M, *et al*: Identification of a microRNA expression signature for chemoradiosensitivity of colorectal cancer cells, involving miRNAs-320a, -224, -132 and let7g. Radiother Oncol 108: 451-457, 2013.
17. Spitzner M, Emons G, Kramer F, Gaedcke J, Rave-Fränk M, Scharf JG, Burfeind P, Becker H, Beissbarth T, Ghadimi BM, *et al*: A gene expression signature for chemoradiosensitivity of colorectal cancer cells. Int J Radiat Oncol Biol Phys 78: 1184-1192, 2010.
18. Hall JS, Iype R, Senra J, Taylor J, Armenoult L, Oguejiofor K, Li Y, Stratford I, Stern PL, O'Connor MJ, *et al*: Investigation of radiosensitivity gene signatures in cancer cell lines. PLoS One 9: e86329, 2014.
19. Pramana J, Van den Brekel MW, van Velthuysen ML, Wessels LF, Nuyten DS, Hofland I, Atsma D, Pimentel N, Hoebers FJ, Rasch CR, *et al*: Gene expression profiling to predict outcome after chemoradiation in head and neck cancer. Int J Radiat Oncol Biol Phys 69: 1544-1552, 2007.
20. Imadome K, Iwakawa M, Nakawatari M, Fujita H, Kato S, Ohno T, Nakamura E, Ohkubo Y, Tamaki T, Kiyohara H, *et al*: Subtypes of cervical adenosquamous carcinomas classified by EpCAM expression related to radiosensitivity. Cancer Biol Ther 10: 1019-1026, 2010.
21. Eschrich SA, Fulp WJ, Pawitan Y, Foekens JA, Smid M, Martens JW, Echevarria M, Kamath V, Lee JH, Harris EE, *et al*: Validation of a radiosensitivity molecular signature in breast cancer. Clin Cancer Res 18: 5134-5143, 2012.
22. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ and Perou CM: Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355: 560-569, 2006.
23. Freidlin B, Jiang W and Simon R: The cross-validated adaptive signature design. Clin Cancer Res 16: 691-698, 2010.
24. Freidlin B and Simon R: Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 11: 7872-7878, 2005.
25. Tang Z, Zeng Q, Li Y, Zhang X, Ma J, Suto MJ, Xu B and Yi N: Development of a radiosensitivity gene signature for patients with soft tissue sarcoma. Oncotarget 8: 27428-27439, 2017.

26. Zhu Y, Qiu P and Ji Y: TCGA-assembler: Open-source software for retrieving and processing TCGA data. Nat Methods 11: 599-600, 2014.
27. Molinaro AM, Simon R and Pfeiffer RM: Prediction error estimation: A comparison of resampling methods. Bioinformatics 21: 3301-3307, 2005.
28. Simon R, Radmacher MD, Dobbin K and McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95: 14-18, 2003.
29. Grandi A, Santi A, Campagnoli S, Parri M, De Camilli E, Song C, Jin B, Lacombe A, Castori-Eppenberger S, Sarmientos P, *et al*: ERMP1, a novel potential oncogene involved in UPR and oxidative stress defense, is highly expressed in human cancer. Oncotarget 7: 63596-63610, 2016.
30. Mulvihill MM, Benjamin DI, Ji X, Le Scolan E, Louie SM, Shieh A, Green M, Narasimhalu T, Morris PJ, Luo K, *et al*: Metabolic profiling reveals PAFAH1B3 as a critical driver of breast cancer pathogenicity. Chem Biol 21: 831-840, 2014.
31. Dolatshad H, Pellagatti A, Liberante FG, Llorian M, Repapi E, Steeples V, Roy S, Scifo L, Armstrong RN, Shaw J, *et al*: Cryptic splicing events in the iron transporter *ABCB7* and other key target genes in *SF3B1*-mutant myelodysplastic syndromes. Leukemia 30: 2322-2331, 2016.
32. Jiang Z, Song Q, Zeng R, Li J, Li J, Lin X, Chen X, Zhang J and Zheng Y: MicroRNA-218 inhibits EMT, migration and invasion by targeting SFMBT1 and DCUN1D1 in cervical cancer. Oncotarget 7: 45622-45636, 2016.
33. Zhai JS, Song JG, Zhu CH, Wu K, Yao Y and Li N: Expression of APPL1 is correlated with clinicopathologic characteristics and poor prognosis in patients with gastric cancer. Curr Oncol 23: e95-e101, 2016.
34. Liu Y, Zhang C, Zhao L, Du N, Hou N, Song T and Huang C: APPL1 promotes the migration of gastric cancer cells by regulating Akt2 phosphorylation. Int J Oncol 51: 1343-1351, 2017.
35. Zhou J, Wu X, Li G, Gao X, Zhai M, Chen W, Hu H and Tang Z: Prediction of radiosensitive patients with gastric cancer by developing gene signature. Int J Oncol 51: 1067-1076, 2017.
36. Tang Z, Zeng Q, Li Y, Zhang X, Suto MJ, Xu B and Yi N: Predicting radiotherapy response for patients with soft tissue sarcoma by developing a molecular signature. Oncol Rep 38: 2814-2824, 2017.