Transcriptomic signature predicts the distant relapse in patients with ER+ breast cancer treated with tamoxifen for five years

HAO ZHOU*, QINGFU LV* and ZHAOJI GUO

Department of General Surgery, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215006, P.R. China

Received March 18, 2017; Accepted September 6, 2017

DOI: 10.3892/mmr.2017.8234

Abstract. Tamoxifen is the most commonly used drug to treat estrogen receptor positive (ER+) breast cancer. However, many patients with ER+ breast cancer have experienced resistance and other adverse side effects following treatment with tamoxifen. Furthermore, clinical and pathological parameters have thus far failed to predict the efficiency of tamoxifen administration. Therefore, gene signature based models for the prediction of survival time of such patients are urgently needed. In the current study, gene expression levels and follow-up information of samples from GSE17705 and GSE22219 databases were used to construct a risk score model based on Cox multivariate regression. The expression levels of 10 genes were included in the model: CCNB2, CCNA2, FOXD1, WSB2, RBPMS, CTDSP1, BIN3, SLBP, EPRS, FTO. The samples in the high-risk group had a relative early distant relapse time period (median survival time of 3.75 years) compared with the patients in the low risk group (median survival time of 6.5 years, P<0.01). For further validation, a further two independent datasets (GSE26971, GSE58644) were assessed. The overall survival time period of patients with high-risk scores in these datasets was significantly longer than those with low-risk scores (P<0.01). Furthermore, the associations between clinical parameters and risk score were investigated, and it was revealed that the risk score was significantly correlated with tumor age, tumor stage and grade. In addition, a 5-year survival nomogram was plotted in order to facilitate the utilization of risk score along with other clinical data. In summary, using the transcriptomic profile, a multi-gene expression based risk score was developed and was revealed as being able to successfully predict the outcome of patients with ER+ breast cancer treated with tamoxifen for 5 years.

Introduction

Breast cancer is the most prevalent malignancy in women worldwide (1). A recent report conducted in China revealed that 272,400 new diagnoses, as well as 70,700 mortalities, occur annually as a result of breast cancer (2). Molecular subtyping of breast cancer is relatively well established, and tamoxifen represents the most common drug prescribed to patients with breast cancer. However, relapse occurs in a large proportion of patients with estrogen-receptor positive (ER+) breast cancer treated with tamoxifen (3), and current clinical practice is insufficient for accurate prognosis. Previous research has identified survival-associated genomic signatures of breast cancer. For example, high expression of the GATA binding protein 3 gene has been reported to be associated with prolonged progression-free survival in patients with ER+ breast cancer (4). Furthermore, patients with a reduced level of Beclin 1 expression demonstrated a higher sensitivity to tamoxifen and a prolonged survival time (5). In addition, a high protein expression level of enhancer of zeste 2 polycomb repressive complex 2 subunit (EZH2) has been reported to be associated with the development of distant metastases in breast cancer (6).

However, the clinical prognostic effect of single molecular biomarkers varies across datasets; whereas a multiple gene expression-based staging method is robust across datasets (7-10). In the present study, a transcriptome-based risk score for the prediction of survival in patients with ER+ breast cancer treated with tamoxifen was developed using the Cox multivariate regression model. Risk scores were developed using cyclin B2 (CCNB2), glutamyl-prolyl-tRNA synthetase (EPRS), α-ketoglutarate dependent dioxygenase, stem-loop binding protein (SLBP), CTD small phosphatase 1 (CTDSP1), cyclin A2 (CCNA2), bridging integrator 3 (BIN3), RNA binding protein with multiple splicing (RBPMS), forkhead box D1 (FOXD1), gene encoding WD repeat of SOCS box containing 2 (WSB2); and the resultant model based upon said genes’ expression levels was revealed to successfully predict survival time in the training and validation datasets (GSE22219, GSE26971 and GSE58644). Median survival time of the high-risk and the low-risk group was 3.75 and 6.5 years, respectively. Furthermore, the associations...
between risk score and clinical parameters were investigated and it was demonstrated that age, grade and stage were significantly associated with risk score. A 5 year survival nomogram was plotted in order to facilitate the utilization of the risk score, which was demonstrated to be an important clinical indicator for prognosis. In conclusion, this study has developed a robust risk score staging system for the prediction of survival in patients with ER+ breast cancer treated with tamoxifen.

Materials and methods

Sample enrollment and data pre-analysis. The following key words were searched for in the Gene Expression Omnibus (GEO) dataset: ‘Breast cancer’, ‘tamoxifen’, ‘expression’ and ‘microarray’; and datasets with <100 ER+ tamoxifen-treated samples, or datasets without survival information, were then manually filtered out. Following this, four datasets, GSE17705, GSE22219, GSE26971 and GSE56884, were then retained for further analysis. Furthermore, samples that were not primary tumor tissue were also excluded during this process. Raw data were then downloaded in the CEL format from the GEO datasets. Following this, background correction and normalization with Robust Multiarray Averaging were carried out using the R package ‘affy’ function ‘rma’ (v1.54.0). Probe and gene names were matched according to the manufacturer-provided annotation file. Genes with more than one complementary probe were merged and the average values were retained as the expression levels for the corresponding genes.

Risk score model development. Cox univariate regression was implemented in both GSE26971 and GSE17005 datasets via correlation of each individual gene’s expression with the survival genes significantly correlated with distant metastasis-free survival time in both GSE26971 and GSE17005 datasets were retained for further analyses as candidate genes. Random forest variable hunting was applied for the selection of a reasonable combination of candidate genes using R package ‘RandomForestSRC’ (v1.9.0). The parameter used was: 100 repeats and 100 iterations. Following this, multivariate Cox regression analysis was carried out in order to develop the linear risk score model using the selected candidate genes, and coefficients were solved with the developed risk score was an effective predictive indicator being 96% in the low-risk group. These results indicated that age, grade and stage were significantly associated with risk score. A 5 year survival nomogram was plotted with the clinical data in the training dataset using R package ‘rms’.

Results

Gene selection and model development. The detailed workflow of gene selection and model development is presented in Fig. 1A. The levels of association between gene expression levels and treatment outcomes (survival data) were assessed using Cox univariate regression. Genes associated with overall survival in both the GSE17705 and GSE26971 datasets were identified, and a total of 48 genes were then selected as candidates. Following this, the random forest variable hunting was performed in order to select for the optimal candidate genes. Following identification of 10 candidate genes (Fig. 1B), risk scores using Cox multivariate regression and expression of 10 genes were then calculated. The coefficients are presented in Fig. 1C, and parameters of Cox regression are shown in Table I. The risk scores were calculated as follows (where gene names represent their respective expression levels): Risk score = (0.299988203)cyclin B2 (CCNB2) + (0.640775607) glutamyl-prolyl-tRNA synthetase (EPRS) + (-0.756716676) α-ketoglutarate dependent dioxygenase (FTO) + (0.132792707) bridging integrator 3 (BIN3) + (-0.618268545) RNA binding protein with multiple splicing (RBPM3) + (0.580014194) forkhead box D1 (FOXD1) + (-0.288974361) gene encoding WD repeat of SOCS box containing 2 (WSB2).

Prognostic values of the risk score in the training dataset. Patients were divided into two groups, a high-risk group or a low-risk group, according to their median risk score. Following this, the difference in survival between the high-risk and the low-risk groups was calculated, and the results revealed that the high-risk group had a reduced relapse-free time compared with the low-risk group, with median survival times of 3.75 vs. 6.5 years, respectively (P<0.001; Fig. 2A). The high-risk group tended to represent early metastasis, and genes with high expression levels tended to have positive coefficients and genes with low expression tended to have negative coefficients (Fig. 2B). The 5-year distant relapse-free survival rate of the high-risk group was 75%; whereas this value was revealed as being 96% in the low-risk group. These results indicated that the developed risk score was an effective predictive indicator for the distant relapse survival time period of patients with ER+ breast cancer treated with tamoxifen.

Risk score performance validation. Considering that the risk score staging system was developed based upon gene expression data in the GSE17705 dataset, there was a potential risk that the model would over-fit to the dataset. In order to assess the robustness of the developed risk score model, three independent datasets (GSE22219, GSE26971 and GSE568644) were used for further validation. Following the locking of the coefficients for the 10 genes, a risk score for each patient was calculated. In addition to patients belonging to the training dataset, the patients belonging to each of the three independent datasets were artificially divided into high-risk and low-risk groups using median risk score values as cutoff values. The patients with high-risk scores tended to have early relapse, as was similarly demonstrated in patients belonging to the training dataset using R package ‘rms’.

Statistical analysis. All statistical analyses were performed using R software (v3.0.1; https://www.r-project.org) and R packages. Normalizations of affymetrix raw data were performed with R package ‘affy’ using the function ‘rma’. The survival analysis and cox probability hazard model construction were performed with R package ‘survival’. Random forest variable hunting was implemented with R package ‘RandomForestSRC’, and receiver operating characteristic (ROC) curves were generated with R package ‘pROC’ (11). The nomogram was plotted with the clinical data in the training dataset using R package ‘rms’.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclin B2 (CCNB2)</td>
<td>0.299988203</td>
</tr>
<tr>
<td>Glutamyl-prolyl-tRNA synthetase (EPRS)</td>
<td>0.640775607</td>
</tr>
<tr>
<td>α-ketoglutarate dependent dioxygenase (FTO)</td>
<td>-0.756716676</td>
</tr>
<tr>
<td>Bridging integrator 3 (BIN3)</td>
<td>0.132792707</td>
</tr>
<tr>
<td>RNA binding protein with multiple splicing (RBPM3)</td>
<td>-0.618268545</td>
</tr>
<tr>
<td>Forkhead box D1 (FOXD1)</td>
<td>0.580014194</td>
</tr>
<tr>
<td>Gene encoding WD repeat of SOCS box containing 2 (WSB2)</td>
<td>-0.288974361</td>
</tr>
</tbody>
</table>
Furthermore, the gene expression profiles for the 10 genes in the both the low-risk and the high-risk groups resemble those demonstrated by the training dataset (Fig. 3B). These results demonstrate that the risk score model is robust.
across datasets for the prediction of distant relapse in patients with ER+ breast cancer treated with tamoxifen.

**Risk score and clinical information.** Subsequently, the associations between clinical parameters (stage, age, grade, lymph
node invasion and primary tumor size) with the risk score were evaluated. As revealed in Fig. 4A, age, tumor stage and grade were significantly associated with the risk score; whereas the other clinical parameters were not (P>0.05). To facilitate the utilization of the risk score, a 5-year distant relapse nomogram was plotted (Fig. 4B). According to this nomogram, risk score was one of the most important metastatic indicators.

**Discussion**

Tamoxifen is the most frequently used drug for the treatment of patients with ER+ breast cancer. However, tamoxifen drug resistance has previously been observed (2). The underlying mechanism of how tamoxifen drug resistance develops remains unclear. In order to predict the survival time of patients treated with tamoxifen, this study has developed a predictive risk score staging system based upon gene expression levels. According to the developed model, the risk score successfully predicted the survival time of patients across both training and test datasets. In addition, associations between risk score and pathological parameters were assessed. The proposed nomogram demonstrated that the risk score was one of the most important indicators for prognosis.

Among the included genes, FOXD1 has previously been reported to promote migration and to be associated with drug resistance in glioma (12). CCNA2 was revealed to correlate closely with distant metastasis-free, recurrence-free and overall survival in breast cancer; in addition, it also contributes to tamoxifen resistance in patients with ER+ breast cancer (13). CCNB2 has previously been demonstrated to serve as an
independent biomarker for invasive breast cancer, and elevated CCNB2 has previously been revealed to be associated with poor patient survival (14). Although little is known about FTO expression and breast cancer, gene polymorphism of FTO has been revealed to be associated with carcinogenesis and survival of patients with breast cancer (15,16). Another gene, CTDSPI, inhibits cancer cell migration and invasion (17). According to recent findings, EPRS is a regulator of cell proliferation in ER+ breast cancer, and reduced EPRS expression has been demonstrated to be associated with decreased distant relapse-free survival in patients treated with tamoxifen for 5 years (18). Enhanced RBPMS expression has been revealed to significantly repress activator protein 1 signaling activity, and thus regulate the proliferation and migration of breast cancer cells (19). The aforementioned candidate genes were either associated with survival of breast cancer patients or tamoxifen resistance/sensitivity, thus explaining why a risk score based upon the expression levels of said genes has proved to be effective for the survival prediction time period of patients with ER+ breast cancer. However, it was revealed that none of the candidate genes were significantly associated with survival across all of the included datasets (data not shown), thus indicating that the expression level of a single gene as a predictive measure for the survival time period of patients with ER+ breast cancer is not as robust as a cumulative risk score.

In conclusion, the current model developed in this study is robust across datasets in the prediction of the survival time of patients with ER+ breast cancer treated with tamoxifen.

References