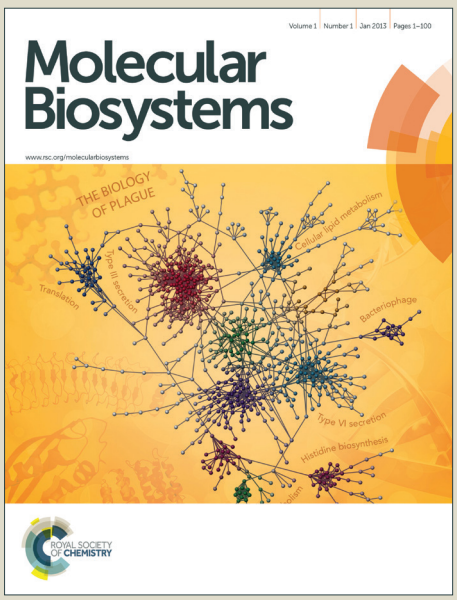
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

A 20-gene signature in predicting chemoresistance of taxane-based chemotherapy of breast cancer

Dong-Xu He[1,#,*], Yu-Dong Xia[#,2], Xiao-Ting Gu[3], Jian Jin[3], Xin Ma[3,*]

[1]National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi 214122, China; [2]E-GENE Technologies, Co., Ltd., Shenzhen 518083, China; [3]Department of Cellular and Molecular Pharmacology, School of Medicine and Pharmaceutics, Jiangnan University, Wuxi 214122, China.

[#]these authors contributed equally to the manuscript.
*Corresponding authors:
Dong-Xu He, PhD
National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University,
Wuxi, China
Tel./ Fax: 86-510-85918229
Email: hedongxu@jiangnan.edu.cn

Xin Ma, PhD
Department of Cellular and Molecular Pharmacology, School of Medicine and Pharmaceutics, Jiangnan University,
Wuxi, China
Tel./ Fax: 86-510-85918219
Email: maxin@jiangnan.edu.cn

Abstract

To date, there is no effective marker to predict chemoresistance in cancers. In this study, we aimed to find a signature that can detect chemoresistance to taxane-based therapies in breast cancer. By studying the gene-expression profiling in discovery cohorts with 92 taxane-resistant and 68 sensitive patients, a 20-gene taxane-based chemotherapy signature (TAXSig) and a TAXSig equation were developed. The TAXSig and its equation were later validated in five further independent datasets with a total of 659 patients. In general, the TAXSig equation easily and effectively discriminated chemoresistant from sensitive individuals. The TAXSig-discriminated groups showed significant differences in clinical outcomes both in estrogen-receptor-positive and -negative (ER^-) breast cancer patients, while TAXSig was especially powerful in discriminating ER^- patients who had a good prognosis and were chemosensitive. In conclusion, TAXSig is a reliable, effective, and reproducible means of classifying chemoresistance to taxane-based therapies in breast cancer.

Key words: chemoresistance; taxane-based chemotherapy signature; clinical outcomes

1. Introduction

Chemoresistance is one of the primary causes of failure in the chemotherapeutic treatment of most human tumors, and diagnosing the sensitivity of a tumor before chemotherapy may greatly improve the efficiency of therapies and the quality of life

of patients.

Although chemoresistance can be divided into acquired or intrinsic based on the initial response to the first therapy, common mechanisms have been identified that giving tumor cells resistance to a variety of structurally and functionally distinct agents [1, 2]. Therefore, recognizing the common mechanisms in different chemoresistant tumor cells may help to predict the response of a patient to a specific chemotherapy. For example, P-glycoprotein plays an important role in the chemoresistance of tumor cells by pumping various chemotherapeutic drugs out of the cell before they exert their cytotoxic effects [3]. Also, glutathione-S-transferase is another factor that mediates chemoresistance by increasing the anti-oxidant and anti-apoptotic capacity of tumor cells [4]. To date, although increasing numbers of key factors are being identified in chemoresistant tumor cells, many of these studies have focused only on the mechanism of a single factor in certain tumor cell lines or in small numbers of clinical samples. Therefore, the results of these studies may not be supported in the clinic due to the difference between *in vitro* studies of tumor cell lines and a real tumor mass, as well as the heterogeneity of human tumors. Therefore, in order to more effectively diagnosis chemoresistance, it may be a good strategy to identify one common mechanism that involves many factors and is shared by large number of clinical tumors.

Gene-expression profiling of clinical tumor samples provide a new means to generate ‘signatures’ for detecting certain features of cancer, such as metastasis and chemoresistance. These signatures contain genes whose expression commonly

changes in the indicated types of cancers [5-8]. The most famous example is the development of MammaPrint [8], which is a 70-gene breast cancer gene signature that has been cleared by the U.S. Food and Drug Administration to assess the probability of metastasis in breast cancer.

In this study, by using previously-published gene-expression profiles from breast cancer patients who were resistant to taxane-based chemotherapy, we identified 20 genes that are differentially expressed in chemoresistant and chemosensitive breast cancers. A taxane-based chemotherapy signature (TAXSig) was then generated. TAXSig predicts the chemoresponse to taxane-based chemotherapy and the outcomes of breast cancer patients.

2. Materials and Methods

2.1 Preparation of expression data

We collected publicly-available datasets of breast cancers in the Gene Expression Omnibus (GEO) with enough information about either the response to taxane-based chemotherapy or the outcome of each patient. The datasets were produced by whole-genome microarrays and with a medium to large sample size (Table 1). The raw data with hybridization probes were consolidated with the Entrez GeneID and the gene names were Gene Ontology (GO)-annotated with the Perl programming language. The gene expression level was normalized and $\log_2$-transformed, and the Wilcoxon rank sum test was used to calculate the difference in the gene expression value between chemoresistant and chemosensitive patients.

Table 1. Publicly-available gene expression data analyzed in this study

GEO Data source	Chemo -therapy	No. of arrays	Institution	Reference	Platform	No. of Gene IDs
GSE349&350	Docetaxel	24	Baylor College of Medicine (USA)	Chang *et al.* [9]	GPL8300	12085
GSE25055	Taxane	310	Nuvera Biosciences (USA)	Hatzis *et al.*[10]	GPL96	20967
GSE25065	anthracycline	198				
GSE22220	N/A	216	University of Oxford (UK)	Buffa *et al.*[11]	GPL6098	13344
GSE22049	N/A	16 out of 60	University of North Carolina at Chapel Hill (USA)	Bockhorn *et al.*[12]	GPL1390, 887, 885	22575

2.2 Hierarchical clustering

Hierarchical clustering was performed with Cluster 3.0 software. The raw expression data were $\log_2$-transformed and mean-centered by subtracting the means of each gene between different patients so that the mean or median value of each row was 0. Complete hierarchical clustering was then performed and the results were visualized in TreeView software.

2.3 Discriminative model

Twenty genes were selected as a signature to discriminate chemoresistant from chemosensitive patients (taxane-based chemotherapy signature, TAXSig). The Bayesian discriminative method using leave-one-out cross-validation in SPSS was used to assess the validity and robustness of TAXSig in distinguishing the two phenotypic states. With this method, the mRNA levels of these genes were used to classify patients and we generated a discriminative equation to give each patient a TAXSig score. The mean value of the scores was used as a threshold for chemosensitive or resistant individuals [13-15]. The performance of the TAXSig was evaluated by the area under the receiver operating characteristic curve (AUC, Matlab).

2.4 Survival analysis

Distant relapse-free survival (DRFS) and relapse-free survival (RFS) were considered as events for all survival analysis. Survival curves were analyzed by the Kaplan-Meier method and compared with the Log-rank method (SPSS). Hazard ratios (HR) between the chemoresistant and chemosensitive groups were calculated

using COX regression (backward stepwise selection procedure (Wald), SPSS). Odds ratios (ORs) throughout all of the datasets were calculated using Review Manager software to yield forest plots.

3 Results

3.1 Generating the 20-gene TAXSig and discriminative equation in the discovery cohort

GSE349 and GSE350 (including chemoresistant and chemosensitive patients respectively, termed GSE349&350 in the following), as well as GSE25055 were used together as discovery cohorts to generate the TAXSig for predicting chemoresistance to taxane-based chemotherapy. GSE349&350 contains 24 patients who were either resistant or sensitive to docetaxel treatment. GSE25055 contains 310 patients with different degrees of response to taxane-anthracycline chemotherapy. In GSE25055, we defined patients with a pathologically complete response (pCR) after chemotherapy as chemosensitive, while patients with an extensive residual cancer burden (RCB-III) were considered to be chemoresistant. As a result, 79 of 310 patients in GSE25055 were chemoresistant, while 57 were chemosensitive.

Then the p-value (Wilcoxon rank sum test, $p < 0.05$) for each gene between resistant and sensitive patients was calculated separately in the two data sets. The significantly-changed genes were then compared between the two data sets to find overlaps, and the process generated 124 genes (supplemental table 1).

The overlapped genes were ranked according to their p-values. Because it would

be ideal to find a signature with few genes but good predictive power, we chose the top 50 genes with the smallest p-values in either GSE349&350 or GSE25055 as the basis of the signature in the first-round selection.

These 50 genes were applied to GSE349&350 and GSE25055 using a preliminary Bayesian discriminative method. The genes showed only ~65% success in discriminating the chemoresponse in these datasets (data not shown). So we then selected or eliminated the genes one by one using the following method: each gene was left out of the 50-gene set one at a time, the discriminative model was refitted using the remaining genes, and the chemoresistance was predicted and the success rate was calculated. If one gene caused a greater successful rate when omitted from the 50-gene set, it was eliminated; and *vice-versa*. Finally, a 20-gene TAXsig was generated, showing the best success rate.

The genes were hierarchically-clustered and GO-annotated (figure 1 and supplemental table 2). The implications of these genes for cancer progression and drug resistance were also analyzed by searching for previously-published studies (supplemental table 2), and the results showed that these genes regulate different biological processes.

We performed leave-one-out validation with the Bayesian discriminative method to predict the chemoresistant status of the patients [8, 11] in GSE349&350 and 25055 using TAXSig. With this method, two discriminative equations were generated based on the mRNA levels of the 20 genes in TAXSig. Then a score was calculated from the equation for each patient. Finally, a threshold value was generated as the mean of the

scores of all patients. TAXSig discriminated the GSE349&350 patients with a success rate of 100%, while correctly discriminating 89.9% of the chemoresistant patients and 86.0% of the chemosensitive patients in GSE25055 (Figure 2 A and B). The ability of TAXSig to predict the chemoresponse in GSE25055 was further tested by the AUC, which was 0.876 ± 0.06 (Figure 2 C), indicating that TAXSig performs well in predicting the chemoresponse in GSE25055.

Because of the differences of microarray platforms and groups of patients between GSE349&350 and GSE25055, TAXSig generated completely different discriminative equations for the two datasets. As a result, the chemoresistant patients were given positive scores with the GSE349&350 equation, but the scores were negative with the GSE25055 equation.

However, although the equation from GSE349&350 showed a greater success rate, GSE25055 involved more patients, so the discriminative equation from GSE25055 was theoretically more reliable and precise than that from GSE349&350. Indeed, when we used the equation from GSE25055 to calculate the scores for patients in GSE349&350, it also showed a good ability to discriminate (Figure 2 D), but the equation from GSE349&350 performed worse for GSE25055 (data not shown). Therefore, we defined the equation form GSE25055 as the TAXSig equation as follows:

$$Y_k = DBI_k + ATG9A_k + TNFRSF10C_k + \ldots\ldots + LSM6_k - 33.449$$

Where the TAXSig score Y_k is calculated for the k^{th} patient with its mRNA level for the 20 genes in the TAXSig. The coefficients for each gene are omitted for clarity

in this equation but are shown in table 2. For each patient, if their TAXSig score is less than the threshold value, they are defined as chemoresistant, and *vice versa*. The scores calculated by TAXSig for the patients in GSE349&350 and 25055 are shown in supplemental table 3.

To exclude the possibility that the TAXSig is dependent on the specific algorithm derived from discovery cohorts in predicting the chemoresistance, we applied a distinct classification method, i.e. logistic regression [7, 16], to determine the ability of the TAXSig in discrimination. As the result, the TAXSig still discriminated the chemoresistant patients from the chemosensitive ones in GSE349&350 and 25055, and both success rate (Figure 2 D) and incorrectly discriminated individuals (supplemental table 2) were very similar with results from TAXSig equation.

Table2. The coefficients of the TAXSig discriminative equation

	gene	coefficient		gene	coefficient
1	DBI	0.22	**11**	GOLGA2	0.087
2	ATG9A	0.751	**12**	GNAI3	-0.115
3	TNFRSF10C	-0.123	**13**	DTNA	-0.128
4	FGFR1	-0.107	**14**	TUBGCP3	0.207
5	PRKCI	0.856	**15**	PDXK	0.068
6	ATF3	0.253	**16**	BTN3A3	0.202
7	TNPO2	0.401	**17**	CDKN2C	0.436
8	UBE3B	0.132	**18**	DCTN1	-0.43
9	TOR1A	0.969	**19**	NDUFA6	-0.296
10	SATB2	-0.207	**20**	LSM6	0.276

3.2 Predicting clinical outcomes with TAXSig

We later applied TAXsig to the 174 patients in GSE25055 other than the 79 chemoresistant and 57 chemosensitive patients in this dataset and found that the chemoresponse of these patients was indeterminate. The chemoresponse of these patients were predicted according to their TAXSig scores (supplemental table 3).

Later, we divided all 310 patients in GSE25055 into chemoresistant and chemosensitive groups, and Kaplan-Meier curves were then used to calculatd the differences in DRFS between the two groups. This showed that the DRFS rate in chemoresistant patients was significantly lower than that of chemosensitive patients (Figure 3 A).

Furthermore, using the multivariate Cox proportional hazards model, TAXSig was tested for its association with DFSR together with other clinical indicators (Table 3). We found that TAXSig, estrogen receptor (ER) status, lymph-node metastatic status, and tumor T-grade were covariates with independent prognostic value for distant relapse/death. The HR of TAXSig for DRFS was 7.303, indicating that the signature is strongly associated with distant relapse/death. Similarly, the HRs of T-stage and nodes were 1.39 and 1.46, indicating that they are significantly associated with distant relapse/ death. On the other hand, ER status showed an HR of 0.212, indicating that ER-positive status is negatively associated with distant relapse/death.

Based on the COX analysis that ER status interferes with the DRFS, we then grouped TAXSig-separated patients according to their ER status, and the DRFS difference was calculated again by Kaplan-Meier curves. As figure 3 B and C showed,

This showed that chemoresistant patients had a lower DRFS rate in both the ER-positive (ER^+) and ER-negative (ER^-) groups than chemosensitive individuals. In addition, the DRFS decreased more dramatically in ER^- chemoresistant patients than those who were ER^+.

Table 3. Multivariate Cox regression analysis of the TAXSig for predicting distance relapse in breast cancer patients (n=310)

Variable	Categories	Sig.	HR	95.0% CI for HR Lower	95.0% CI for HR Upper
TAXSig	Resistant vs Sensitive	0.000	7.303	3.165	14.765
ER	Positive vs negative	0.000	0.212	0.122	0.366
PR	Positive vs negative	0.747	0.882	0.412	1.890
Her2	Positive vs negative	0.324	2.060	0.489	8.671
age	26~75	0.950	1.001	0.976	1.026
T Stage	T1,T2, T3,T4	0.026	1.390	1.039	1.858
Nodes	N0,N1, N2,N3	0.03	1.460	1.136	1.877
Grade	1,2, 3,4	0.573	1.139	0.725	1.788

Method = Backward Stepwise (Wald); HR: Hazard ratio. ER= Estrogen receptor; PR= Progesterone receptor; Her2= human epidermal growth factor receptor 2; nodes= lymph nodes metastasis.

3.3 Analysis of TAXSig in validation cohort

TAXsig was then validated in GSE25065, which included 198 patients with information about their chemoresponse to taxane-anthracycline chemotherapy and clinical outcomes. Twenty-three patients were defined as chemosensitive when they were pCR and 31 as resistant when they were RCB-III after chemotherapy. The TAXSig equation was then applied, and 73.9% of the chemosensitive and 83.9% of the chemoresistant patients were correctly discriminated (Figure 4 A).

We then defined the chemoresponse of all 198 patients, and the DRFS was compared between those who were chemoresistant and those who were sensitive. Consistent with the discovery cohort, the DRFS in chemoresistant patients decreased significantly (Figure 4 B), which was also found in both ER^+ and ER^- patients. Also, the chemoresistant ER^- patients showed the worst outcomes.

Furthermore, TAXSig was validated in GSE41998 [17], which included 127 patients receiving sequential neoadjuvant therapy starting with AC treatment (doxorubicin and cyclophosphamide) for 3 weeks, followed by paclitaxel for 12 weeks. Thirty-four patients were defined as chemosensitive when they were pCR and 93 as resistant when they were not after paclitaxel treatment. After TAXSig calculation of, 79.4% of the chemosensitive and 80.6% of the chemoresistant patients were correctly discriminated (Figure 4 E).

Finally, TAXSig was also tested in chemotherapeutic regimes without taxel agents (GSE4779 [5]) but receiving 5-fluorouracil, epirubicin, and cyclophosphamide treatment. Patients were defined as chemosensitive when they were pCR and resistant when they were not after treatment. As a result, 61.3% patients of were successfully

classified as chemosensitive and 63.4% as resistant (Figure 4 F).

In order to test the effect of TAXSig equation to discriminate chemoresponse, logistic regression was applied in GSE25065 and 41998 with TAXSig. It was found the result from logistic regression was still similar with that from TAXSig equation, as they did in discovery cohorts (Figure 4 G and H).

3.4 Meta-analysis of clinical outcomes by TAXSig

To further test the ability of TAXSig to predict clinical outcomes, meta-analysis was performed to combine the results from TAXSig analysis in different datasets. To obtain the meta-analysis, two more data sets with information about RFS were analyzed by TAXSig (GSE22220 [11] and GSE22049 [12]). Because these datasets do not contain chemoresponse information, patients were ranked according to their TAXSig scores, and individuals with the top 50% scores were considered to be chemosensitive. As a result, the patients grouped as chemosensitive showed greater RFS rates (Figure 5 A and C), and the difference between chemoresistant and sensitive patients were more significant in the ER^- groups (Figure 5 B).

Finally, the meta-analysis of relapse/death events were analyzed by combining the results from GSE25055, 25065, 22220 and 22059 with a total of 739 breast cancer cases. After TAXSig divided the patients into chemoresistant and chemosensitive groups, the relapse/death numbers in each group were summarized and the ORs were calculated. All of the ORs for relapse/death were <1 in the sensitive groups (Figure 5 D), which means that TAXSig-predicted chemoresistant patients had an overall negative correlation with the bad outcomes of relapse/death.

4 Discussion

Taxanes are a group of chemotherapeutic agents widely used in the treatment of metastatic and early breast cancer. However, currently there are no valid biomarkers to predict resistance to these agents [18]. Sensitivity and resistance to taxane are highly complex due to the clinical heterogeneity of breast cancers, and using single gene biomarkers to assess taxane response seldom produces conclusive results.

Therefore, we did not design this study to discover specific genes for resistance to taxane-based chemotherapies, but set out to identify patterns of several genes that could be used as a predictive signature in breast cancer patients with taxane resistance. We found a taxane-resistance signature by analyzing the gene expression profiling of 92 taxane-resistant and 68 taxane-sensitive patients in a discovery cohort. We constructed a 20-gene TAXSig, which contained genes that were commonly changed in taxane-resistant breast cancer samples. According to the GO category of the included genes, they regulate various biological processes of cancer cells that favor the development of chemoresistance. We then searched previously-published studies to find the signaling pathways of these genes implicated in chemoresistance, and found that, among the significantly-changed genes in chemoresistant breast cancers, (i) decreased apoptosis and increased cell proliferation were most frequently involved to directly decrease the sensitivity of tumor cells to cytotoxic chemotherapeutic drugs; these genes included ATG9A [19] [20], TNFRSF10C [21], ATF3 [22], GNAI3 [23], and CDKN2C [24]; (ii) cell motility may be increased to help tumor cells to escape from the chemotherapeutic drug and enable metastasis *via* regulation of SATB2 [25],

GOLGA2 [26], DTNA [27], and TUBGCP3 [28]; (iii) FGFR1 [29, 30] and PDXK [31] have been shown to modulate the malignancy transition of tumor cells, which is essential for the development of chemoresistance; and (iv) the epithelial mesenchymal pathway was also included in TAXSig as PRKCI [32, 33] and DCTN1 [34], to generate drug resistance [35]. Therefore, TAXSig contains several pathways essential to the development of chemoresistance, which guarantees its sensitivity in predicting chemoresistance. Furthermore, genes such as DBI, TNPO2, UBE3B, TOR1A, BTN3A3, NDUFA6, and LSM6 were found not to be directly related to cancer progression and chemoresistance, so TAXSig may also open new areas to study the mechanism of chemoresistance.

TAXSig showed a good ability to predict the chemoresponse in four separate cohorts with taxane-based chemotherapy for breast cancer regardless of the subtype and grade of the cancer, suggesting that TAXSig might be useful to predict the taxane response in most breast cancer patients. Furthermore, one of the cohorts used a single drug throughout chemotherapy and another three used taxane-based multi-drug chemotherapy, and TAXSig was able to predict the chemoresistant to both of regimes, suggesting that it is applicable to routine regimes of taxane chemotherapy in breast cancer. We did not assess TAXSig in regimes with more than three drugs because they are seldom used clinically and the gene-expression profiling is not available.

In order to enhance the utility of TAXSig, we then developed a discriminative equation to give each patient a TAXSig score. By comparing the scores of patients within each cohort, the patients could be easily distinguished as chemoresistant or

sensitive. The equation worked well in both of the discovery cohorts, GSE349&350 and GSE25055. In the validation cohorts with taxel-related regimes (GSE25065 and GSE41998), but not in the cohort without taxel as a chemotherapeutic agent (GSE4779), TAXSig still predicted chemoresistant patients well, but the success rate for predicting chemosensitive patients was relatively low. Therefore, it cannot be denied that the results may lead to overtreatment of a small fraction of chemosensitive patients. However, we noted in the GSE25065 cohort, 22 of the 23 chemosensitive patients had T3–T4 stage or AJCC IIA-IIIB tumors, which are thought to be high-risk and should be treated with more complicated regimes. Therefore, the results of TAXSig prediction can still reduce the absolute number of unnecessarily exposures to chemotherapy compared to treatment selection based on the clinicopathological criteria.

Previously, Potti et al. did work similar to ours to assess the sensitivity to taxel-based chemotherapies [16]. They generated several separate signatures predicting the chemoresponse to agents such as docetaxel, paclitaxel, and adriamycin. All of these signatures were generated separately from docetaxel-, paclitaxel-, or adriamycin-resistant/sensitive NCI-60 cell lines. With a method similar to ours, i.e. leave-one-out cross-validation, the signature was validated in lung and ovarian cancer cell lines, as wells as clinical samples of ovarian and breast cancer. However, Potti et al. did not provide a discriminative equation. In the validation of the test for taxel-based therapies, Potti et al first used the same data set as we did (GSE349&350) to validate the single-agent regime. Both their (91.6%) and our studies (overall

accuracy = 23/24 = 95.8%) reached a high accuracy of prediction, suggesting that expression profiling from both cell lines and clinical samples may be a good choice for developing an effective gene-expression signature. Other studies have used both methods to develop cancer signature; Shats et al. and Hsu et al. [6, 7] used cell line data, and Chang et al. and Famer et al. [5, 9] used clinical data, but the later method is more frequently used. However, in validating the response to multiple-agent regimes that use paclitaxel as the taxel-based drug (paclitaxel, 5-FU, adriamycin, and cyclophosphamide are included in the regime), Potti et al. used a different signature developed from paclitaxel-resistant cell lines for this step of prediction, different from our use of the same TAXSig. Clearly, using one TAXSig with a clear discriminative equation is more convenient for clinical application.

In addition, we compared the discriminative ability of TAXSig equation with logistic regression of TAXSig, and two methods generated similar results, suggesting the TAXSig equation is reliable. Previously, the logistic regression was commonly used to discriminate certain phenotypes based on a group of covariants [7, 16], but it is unpractical clinically. By contrast, it is possible that once the TAXSig equation is generated based on a group of patients with known chemoresistance, it may be used to calculate and predict the chemoresponse of patients with unknown chemoresistance and known gene-expression data of the 20 genes.

The survival status after therapy is tightly associated with the chemoresponse. As one might expect, TAXSig also showed a good ability to predict DRFS. Patients predicted to be chemoresistant had an overall lower rate of DRFS. The results were

consolidated in 739 breast cancer patients by meta-analysis, suggesting the chemoresistant patients, as discriminated by TAXSig, overall have worse outcomes. Furthermore, the gene-expression profiling of these patients were from 6 different microarray chips but TAXSig worked well in all of them, suggesting that TAXsig is practical for clinical use.

Furthermore, the TAXSig was powerful in discriminating good-outcome from bad-outcome ER^- patients. Although ER^- patients generally have a worse prognosis, TAXSig was able to identify chemosensitive ER^- patients with a good prognosis, thus avoiding overtreatment of this group. In addition, TAXSig was able to identify ER^+ patients resistant to taxane-anthracycline-endocrine-based chemotherapy, and this group of patients also showed bad clinical outcomes. ER^+ patients generally have a better prognosis than those who are ER^-, and routinely given endocrine therapy. However, a small number of ER^+ patients still showed chemoresistance and a bad prognosis, so it is important to identify ER^+ patients with a bad prognosis or chemoresistance and treat them carefully.

Taken together, in this study we found 20 genes that reliably, effectively, and reproducibly classified patients who were chemoresistant to taxane-based therapies, and excluded those that were chemosensitive. Over the last decade, microarray-based technology has emerged as a new and personalized approach to tumor diagnosis. Therefore, the flexibility of TAXSig between different types of microarray platforms enables its practicability as both a prognostic and a predictive biomarker.

5 Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (81100185, 31200126, and 31371317); The Natural Science Foundation of Jiangsu Province [BK20141109]; the Natural Science Fundation for Distinguished Young Scholars of Jiangsu Province [BK20140004]. to Drs Dongxu He and Xin Ma. We thank Dr. IC Bruce for reading the manuscript.

Reference List

[1] L. Gatti and F. Zunino, Overview of tumor cell chemoresistance mechanisms. Methods Mol. Med. 111 (2005) 127-148.

[2] J.H. Gerlach, N. Kartner, D.R. Bell, and V. Ling, Multidrug resistance. Cancer Surv. 5 (1986) 25-46.

[3] J.H. Gerlach, D.R. Bell, C. Karakousis, H.K. Slocum, N. Kartner, Y.M. Rustum, V. Ling, and R.M. Baker, P-glycoprotein in human sarcoma: evidence for multidrug resistance. J. Clin. Oncol. 5 (1987) 1452-1460.

[4] N. Traverso, R. Ricciarelli, M. Nitti, B. Marengo, A.L. Furfaro, M.A. Pronzato, U.M. Marinari, and C. Domenicotti, Role of glutathione in cancer progression and chemoresistance. Oxid. Med. Cell Longev. 2013 (2013) 972913.

[5] P. Farmer, H. Bonnefoi, P. Anderle, D. Cameron, P. Wirapati, V. Becette, S. Andre, M. Piccart, M. Campone, E. Brain, G. Macgrogan, T. Petit, J. Jassem, F. Bibeau, E. Blot, J. Bogaerts, M. Aguet, J. Bergh, R. Iggo, and M. Delorenzi, A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. Nat. Med. 15 (2009) 68-74.

[6] Y.C. Hsu, H.Y. Chen, S. Yuan, S.L. Yu, C.H. Lin, G. Wu, P.C. Yang, and K.C. Li, Genome-wide analysis of three-way interplay among gene expression, cancer cell invasion and anti-cancer compound sensitivity. BMC. Med. 11 (2013) 106.

[7] I. Shats, M.L. Gatza, J.T. Chang, S. Mori, J. Wang, J. Rich, and J.R. Nevins, Using a stem cell-based signature to guide therapeutic selection in cancer. Cancer Res. 71 (2011) 1772-1780.

[8] M.J. van de Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, d. van, V, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, and R. Bernards, A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. 347 (2002) 1999-2009.

[9] J.C. Chang, E.C. Wooten, A. Tsimelzon, S.G. Hilsenbeck, M.C. Gutierrez, R. Elledge, S. Mohsin, C.K. Osborne, G.C. Chamness, D.C. Allred, and P. O'Connell, Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 362 (2003) 362-369.

[10] C. Hatzis, L. Pusztai, V. Valero, D.J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, M. Martin, J. Cotrina, H. Gomez, R. Hubbard, J.I. Chacon, J. Ferrer-Lozano, R. Dyer, M. Buxton, Y. Gong, Y. Wu, N. Ibrahim, E. Andreopoulou, N.T. Ueno, K. Hunt, W. Yang, A. Nazario, A. DeMichele, J. O'Shaughnessy, G.N. Hortobagyi, and W.F. Symmans, A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305 (2011) 1873-1881.

[11] F.M. Buffa, C. Camps, L. Winchester, C.E. Snell, H.E. Gee, H. Sheldon, M. Taylor, A.L. Harris, and J. Ragoussis, microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res. 71 (2011) 5635-5645.

[12] J. Bockhorn, R. Dalton, C. Nwachukwu, S. Huang, A. Prat, K. Yee, Y.F. Chang, D. Huo, Y. Wen, K.E. Swanson, T. Qiu, J. Lu, S.Y. Park, M.E. Dolan, C.M. Perou, O.I. Olopade, M.F. Clarke, G.L. Greene, and H. Liu, MicroRNA-30c inhibits human breast tumour chemotherapy resistance by regulating TWF1 and IL-11. Nat. Commun. 4 (2013) 1393.

[13] T. Hastie, L. Sleeper, and R. Tibshirani, Flexible covariate effects in the proportional hazards model. Breast Cancer Res. Treat. 22 (1992) 241-250.

[14] P.A. Lachenbruch, Discriminant diagnostics. Biometrics 53 (1997) 1284-1292.

[15] G. J McLachlan, Discriminant analysis and statistical pattern recognition. 1992,Wiley

[16] A. Potti, H.K. Dressman, A. Bild, R.F. Riedel, G. Chan, R. Sayer, J. Cragun, H. Cottrill, M.J. Kelley, R. Petersen, D. Harpole, J. Marks, A. Berchuck, G.S. Ginsburg, P. Febbo, J. Lancaster, and J.R. Nevins, Genomic signatures to guide the use of chemotherapeutics. Nat. Med. 12 (2006) 1294-1300.

[17] C.E. Horak, L. Pusztai, G. Xing, O.C. Trifan, C. Saura, L.M. Tseng, S. Chan, R. Welcher, and D. Liu, Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. Clin. Cancer Res. 19 (2013) 1587-1595.

[18] S. Murray, E. Briasoulis, H. Linardou, D. Bafaloukos, and C. Papadimitriou, Taxane

resistance in breast cancer: mechanisms, predictive biomarkers and circumvention strategies. Cancer Treat. Rev. 38 (2012) 890-903.

[19] C. He and D.J. Klionsky, Atg9 trafficking in autophagy-related pathways. Autophagy. 3 (2007) 271-274.

[20] M.C. Maiuri, E. Zalckvar, A. Kimchi, and G. Kroemer, Self-eating and self-killing: crosstalk between autophagy and apoptosis. Nat. Rev. Mol. Cell Biol. 8 (2007) 741-752.

[21] D. Bernard, B. Quatannens, B. Vandenbunder, and C. Abbadie, Rel/NF-kappaB transcription factors protect against tumor necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL)-induced apoptosis by up-regulating the TRAIL decoy receptor DcR1. J. Biol. Chem. 276 (2001) 27322-27328.

[22] X. Huang, X. Li, and B. Guo, KLF6 induces apoptosis in prostate cancer cells through up-regulation of ATF3. J. Biol. Chem. 283 (2008) 29795-29801.

[23] T. Wu, Y. Li, D. Huang, F. Han, Y.Y. Zhang, D.W. Zhang, and J. Han, Regulator of G-protein signaling 19 (RGS19) and its partner Galpha-inhibiting activity polypeptide 3 (GNAI3) are required for zVAD-induced autophagy and cell death in L929 cells. PLoS. One. 9 (2014) e94634.

[24] K.L. Guan, C.W. Jenkins, Y. Li, M.A. Nichols, X. Wu, C.L. O'Keefe, A.G. Matera, and Y. Xiong, Growth suppression by p18, a p16INK4/. Genes Dev. 8 (1994) 2939-2952.

[25] O. Aprelikova, X. Yu, J. Palla, B.R. Wei, S. John, M. Yi, R. Stephens, R.M. Simpson, J.I. Risinger, A. Jazaeri, and J. Niederhuber, The role of miR-31 and its target gene SATB2 in cancer-associated fibroblasts. Cell Cycle 9 (2010) 4387-4398.

[26] H. Mellor, Cell motility: Golgi signalling shapes up to ship out. Curr. Biol. 14 (2004) R434-R435.

[27] D.S. Bredt, Knocking signalling out of the dystrophin complex. Nat. Cell Biol. 1 (1999) E89-E91.

[28] E.A. Nigg, Centrosome aberrations: cause or consequence of cancer progression? Nat. Rev. Cancer 2 (2002) 815-825.

[29] R. Sharpe, A. Pearson, M.T. Herrera-Abreu, D. Johnson, A. Mackay, J.C. Welti, R. Natrajan, A.R. Reynolds, J.S. Reis-Filho, A. Ashworth, and N.C. Turner, FGFR signaling promotes the growth of triple-negative and basal-like breast cancer cell lines both in vitro and in vivo. Clin. Cancer Res. 17 (2011) 5275-5286.

[30] N. Turner, A. Pearson, R. Sharpe, M. Lambros, F. Geyer, M.A. Lopez-Garcia, R. Natrajan, C. Marchio, E. Iorns, A. Mackay, C. Gillett, A. Grigoriadis, A. Tutt, J.S. Reis-Filho, and A. Ashworth, FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. Cancer Res. 70 (2010) 2085-2094.

[31] S.M. Zhang, W.C. Willett, J. Selhub, D.J. Hunter, E.L. Giovannucci, M.D. Holmes, G.A. Colditz, and S.E. Hankinson, Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer. J. Natl. Cancer Inst. 95 (2003) 373-380.

[32] D. Coradini, P. Boracchi, F. Ambrogi, E. Biganzoli, and S. Oriana, Cell polarity, epithelial-mesenchymal transition, and cell-fate decision gene expression in ductal carcinoma in situ. Int. J. Surg. Oncol. 2012 (2012) 984346.

[33] G. Moreno-Bueno, F. Portillo, and A. Cano, Transcriptional regulation of cell polarity in EMT and cancer. Oncogene 27 (2008) 6958-6969.

[34] S.E. Williams, S. Beronja, H.A. Pasolli, and E. Fuchs, Asymmetric cell divisions promote Notch-dependent epidermal differentiation. Nature 470 (2011) 353-358.

[35] J.M. Lee, S. Dedhar, R. Kalluri, and E.W. Thompson, The epithelial-mesenchymal transition: new insights in signaling, development, and disease. J. Cell Biol. 172 (2006) 973-981.

Figure legends

Figure 1. Hierarchical clustering of patients in GSE349&350 and GSE25055 based on the gene expression of TAXSig. The heat map depicts the two-way hierarchical clustering of 24 (GSE349&350) and 136 (GSE25055) breast tumor samples with 20 genes. Thirteen of the 24 patients in GSE349&350, and 79 of the 136 in GSE25055 were chemoresistant, while the others were chemosensitive. Low (green) and high (red) activity of the genes and predicted clustering of the patients generally divided them into two large clusters of sensitive (blue bar) and resistant (purple bar) individuals.

Figure 2. Performance of TAXSig in the discovery cohort. (A a and b) The signature score and threshold were calculated separately in GSE349&350 and GSE25055 cohorts using the Bayesian discriminative method. The violet dots indicate incorrectly-classified individuals. The accuracy of the signature was calculated as the number of patients correctly classified/total number of resistant or sensitive patients.

(B) The area under the receiver operating characteristic curve (AUC) was calculated to test how well the signature predicted chemoresistance. (C) The TAXSig equation was generated from GSE25055 using the Bayesian discriminative method, and the TAXSig scores were calculated in GSE349&350. (D a and b) The probability of chemoresistance was predicted by the binary logistic regression in GSE349&350 and GSE25055 cohorts. 0.5 was set as cutoff point.

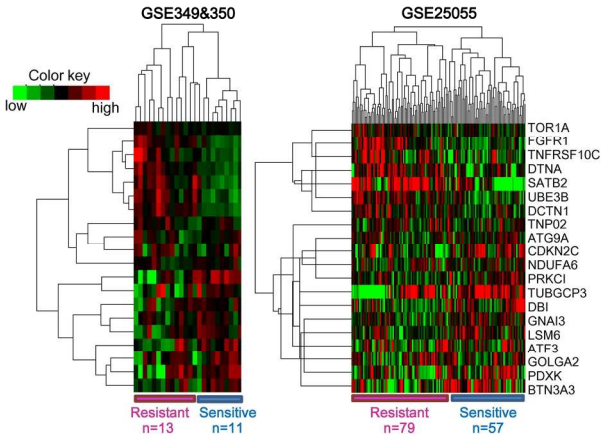
Figure 3. Survival analysis of the discovery cohort using TAXSig. (A) Kaplan-Meier analysis of DRFS (censored at 8 years) in 310 patients in the GSE25055 data set. The patients were grouped into chemoresistant or sensitive by the TAXSig calculation, and their difference in DRFS was calculated. (B and C) Kaplan-Meier analysis of DRFS in ER^+ (B) and ER^- (C) patients in GSE25055.

Figure 4. Validation of TAXSig. (A) The signature score and threshold were calculated using the TAXSig equation in GSE25065. The violet dots represent incorrectly-classified individuals. (B) Kaplan-Meier analysis of DRFS in 198 TAXSig-grouped patients in GSE25065; the difference of DRFS was also calculated for the ER^+ (C) and ER^- (D) patients. (E and F) The signature score and threshold were calculated using the TAXSig equation in GSE41998. The violet dots represent incorrectly-classified individuals. (G and H) The probability of chemoresistance was predicted by the binary logistic regression in GSE25065 and GSE41998 cohorts. 0.5 was set as cutoff point.

Figure 5. Meta-analysis of relapse/death using TAXSig. TAXSig scores were calculated for patients in GSE22220 and GSE22040, and those with the top 50%

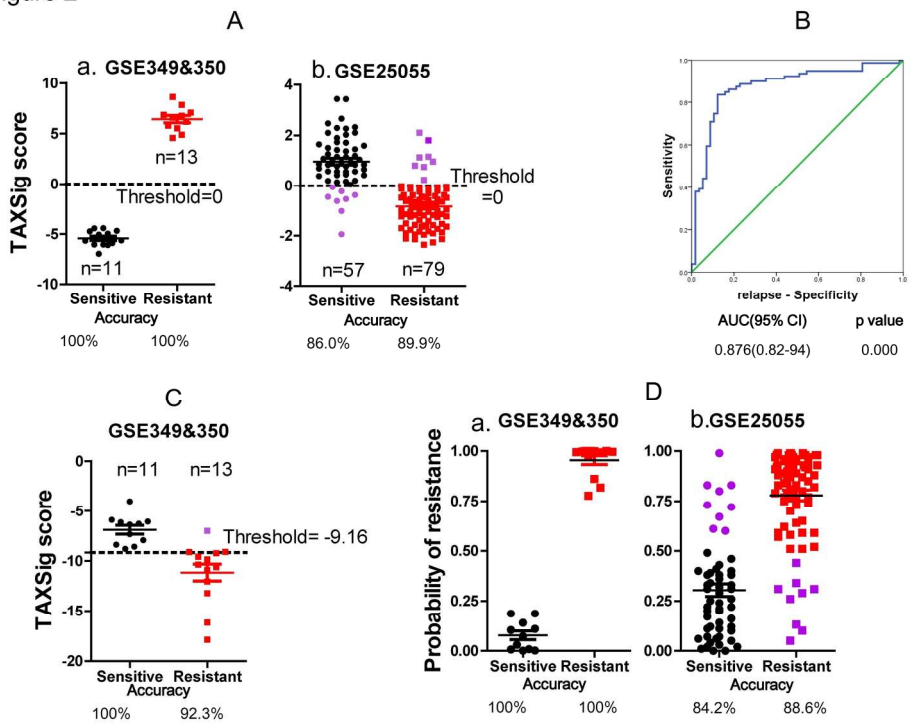
scores were defined as chemosensitive, the rest were chemoresistant. RFSs were then compared between chemoresistant and sensitive patients in the GSE22220 (A) and GSE22049 (C) datasets by Kaplan-Meier analysis. The RFSs of chemoresistant and sensitive ER^- patients in GSE22220 were also calculated separately (B). Finally, the ability of TAXSig to predict clinical outcomes (measured as relapse/death events) in GSE25055, GSE25065, GSE22220, and GSE22049 were analyzed by meta-analysis. The results were visualized by forest plot. Odds ratios (OR) for each dataset are plotted as horizontal bars, the length of the bar represents the 95% confidence interval, and the bars can be compared vertically between datasets. The diamond represents the total OR of the signature in a total of 739 breast cancer cases. The weight means the relative size of each dataset.

Figure 1



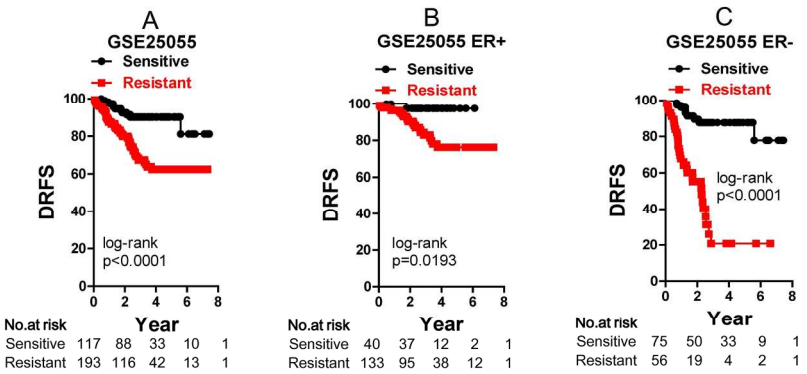
190x142mm (300 x 300 DPI)

Figure 2



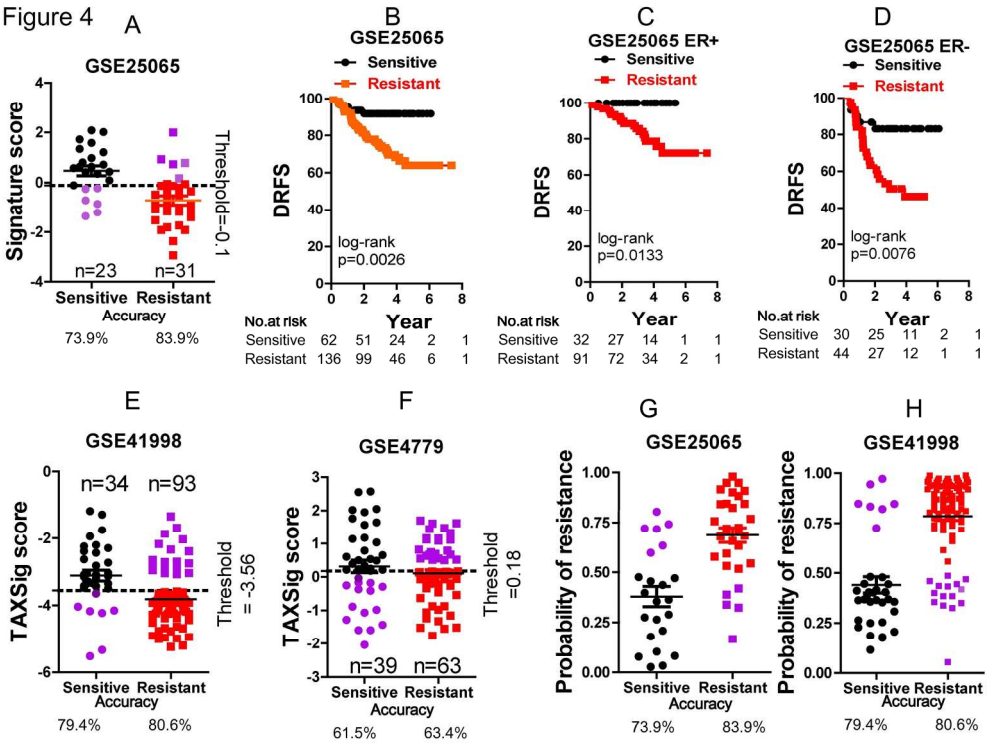
190x142mm (300 x 300 DPI)

Figure 3



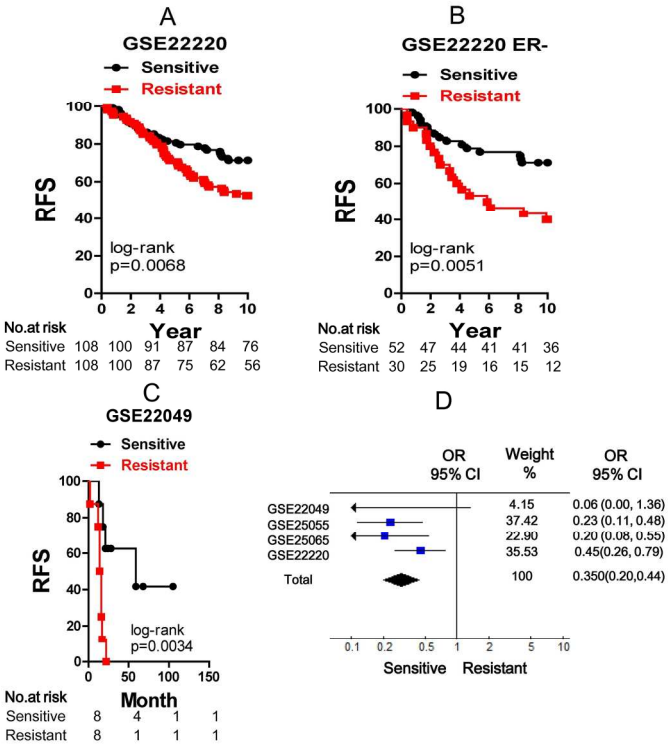
190x142mm (300 x 300 DPI)

Figure 4



190x142mm (300 x 300 DPI)

Figure 5



190x142mm (300 x 300 DPI)