

Population-Based Molecular Prognosis of Breast Cancer by Transcriptional Profiling

Yan Ma,^{1,2} Yong Qian,⁶ Liang Wei,^{1,2} Jame Abraham,^{1,3} Xianglin Shi,⁶ Vincent Castranova,⁶
E. James Harner,² Daniel C. Flynn,^{1,4} and Lan Guo^{1,5}

Abstract Purpose: The purpose of this study is to predict breast cancer recurrence and metastases and to identify gene signatures indicative of clinicopathologic characteristics using gene expression patterns derived from cDNA microarray.

Experimental Design: Expression profiles of 7,650 genes were investigated on an unselected group of 99 node-negative and node-positive breast cancer patients to identify prognostic gene signature of recurrence and metastases. The identified gene signature was validated on independent 78 patients with primary invasive carcinoma (T₁/T₂ and N₀) and on 58 patients with locally advanced breast cancer (T₃/T₄ and/or N₂). The gene predictors were identified using a combination of random forests and linear discriminant analysis function.

Results: This study identified a new 28-gene signature that achieved highly accurate disease-free survival and overall survival (both at $P < 0.001$, time-dependent receiver operating characteristic analysis) in individual breast cancer patients. Patients categorized into high-risk, intermediate-risk, and low-risk groups had distinct disease-free survival ($P < 0.005$, Kaplan-Meier analysis, log-rank test) in three patient cohorts. A strong association ($P < 0.05$) was identified between risk groups and tumor size, tumor grade, estrogen receptor and progesterone receptor status, and HER2/*neu* overexpression in the studied cohorts. We also identified 14-gene predictors of nodal status and 9-gene predictors of tumor grade.

Conclusions: This study has established a population-based approach to predicting breast cancer outcomes at the individual level exclusively based on gene expression patterns. The 28-gene recurrence signature has been validated as quantifying the probability of recurrence and metastases in patients with heterogeneous histology and disease stage.

Breast cancer patients with the same stage of disease can have markedly different clinical outcomes. Traditional diagnostic and prognostic factors may stratify patients with molecularly distinct diseases into the same group based on morphologic

assessments. It remains a critical issue to reliably identify specific high-risk breast cancer patients for recurrent and metastatic diseases. Molecular prediction is the future direction of personalized cancer care. Microarray technologies have fostered tremendous advances in molecular diagnosis and prognosis of breast cancer (1–11). A recent report has described a clinically applied multigene assay to predict recurrence of tamoxifen-treated, node-negative, and estrogen receptor (ER)-positive breast cancer (12). A population-based approach to molecular prognosis of breast cancer and rational individualization of therapy is needed.

In this work, we present a population-based study to predict recurrence and metastases of breast cancer by using gene expression patterns in tumors. Sotiriou et al. (8) originally undertook the population-based study from a regional cancer center where 350 new patients a year were referred in from a population of 1.5 million. Within the years 1993 to 1995, 700 new breast cancer cases were seen, of which 99 cases were representative of the population with comparable overall survival adjusted for tumor size and nodal status. Almost all of these 99 patients received adjuvant therapy after surgery, according to accepted standards of practice at that time. Based on the gene profiling on these 99 tumors, we identified a new 28-gene signature to predict recurrence of breast cancer. Furthermore, the accuracy of the 28-gene signature in disease-free survival prediction was validated ($P < 0.001$) on two

Authors' Affiliations: ¹Mary Babb Randolph Cancer Center, Departments of ²Statistics, ³Medicine and Division of Hematology/Oncology, ⁴Microbiology, Immunology, and Cell Biology, and ⁵Community Medicine, West Virginia University; and ⁶The Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, West Virginia

Received 9/5/06; revised 12/5/06; accepted 1/8/07.

Grant support: NIH/National Center for Research Resources grant P20 RR16440-03 and West Virginia University grant RDG NT10017W (L. Guo) and NIH/National Cancer Institute grant 1R01CA119028-01 (X. Shi).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (<http://clincancerres.aacrjournals.org/>).

Disclaimer: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

Requests for reprints: Lan Guo, Mary Babb Randolph Cancer Center/ Department of Community Medicine, West Virginia University, 1814 HSS, 1 Medical Center Drive, Morgantown, WV 26506-9300. Phone: 304-293-6455; Fax: 304-293-4667; E-mail: lguo@hsc.wvu.edu.

©2007 American Association for Cancer Research.

doi:10.1158/1078-0432.CCR-06-2222

independent patient cohorts (1, 7). The cohort of van't Veer et al. (1) contained 78 patients with sporadic cancer all under the age of 55 years with no lymph node involvement, who were not treated with adjuvant chemotherapy. The cohort of Sorlie et al. (7) included 78 cases, 51 of which were with locally advanced T₃/T₄ and/or N₂ breast cancer treated with primary chemotherapy. In this study, a recurrence score function was used to define a patient's risk for relapse and metastases for individualized clinical decision-making. Patients categorized into high risk, low risk, and intermediate risk had distinct disease-free survival ($P < 0.005$, Kaplan-Meier analysis, log-rank test) in all three cohorts. A strong association ($P < 0.05$) was identified between risk groups and tumor size, tumor grade (13), ER and progesterone receptor (PR) status, and HER2/*neu* overexpression in the studied patient cohorts (1, 7, 8). The recurrence score was also predictive of overall

survival ($P < 0.001$) in the cohorts from Sotiriou et al. (8) and Sorlie et al. (7). To further reveal the molecular pathogenesis of breast cancer, we also identified 14-gene predictors of nodal status and 9-gene predictors of tumor grade using the data from Sotiriou et al. (8) in this population-based study.

Materials and Methods

Patients. Three patient cohorts from the previous publications (1, 7, 8) were analyzed in this study. The cohort from Sotiriou et al. (8) contained the tumor samples from 99 patients with primary local breast carcinoma obtained from the John Radcliffe Hospital from January 1993 to December 1994. All of the 99 tumors were invasive ductal carcinomas: 46 patients were node negative and 53 were node positive.

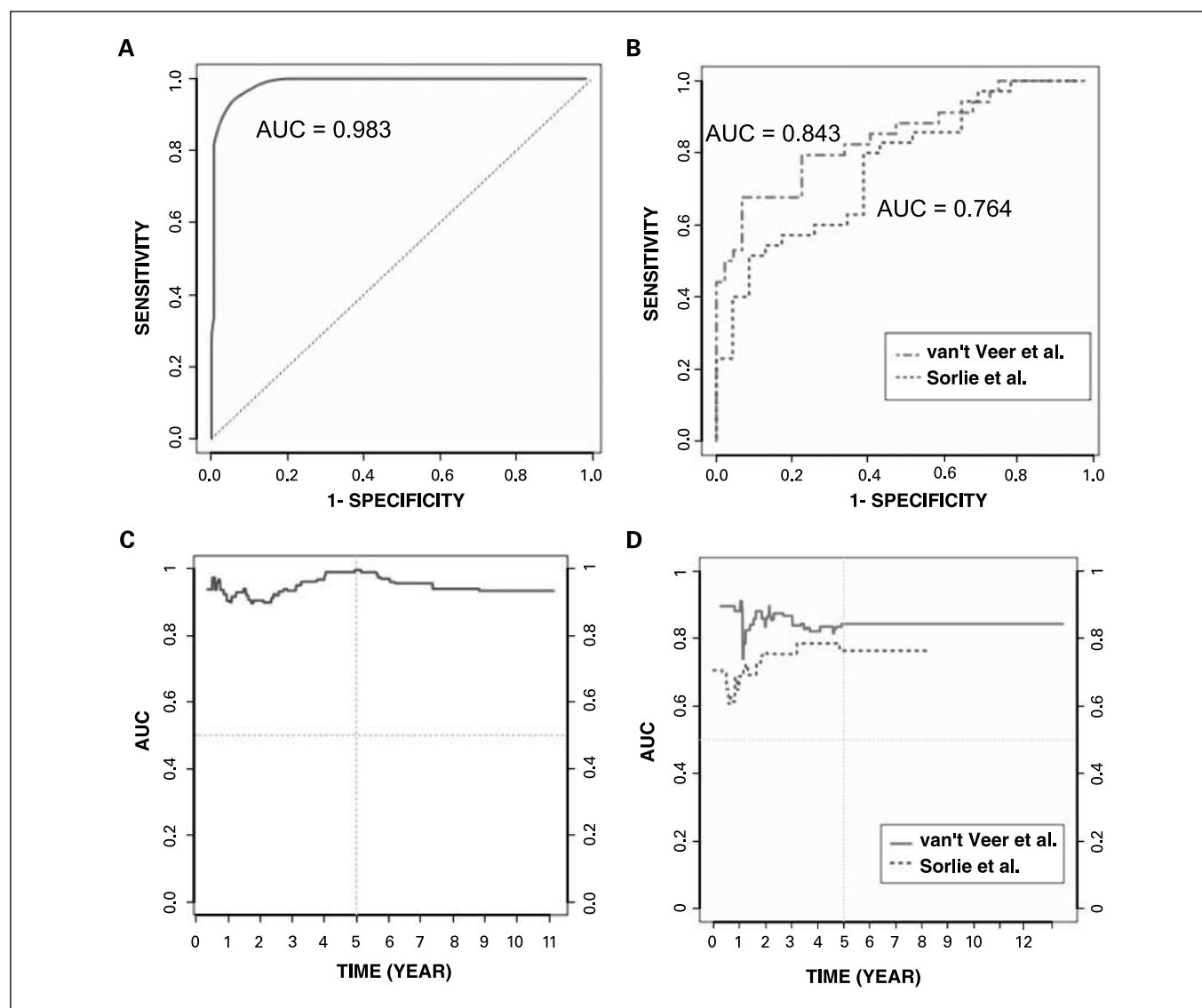


Fig. 1. Time-dependent ROC analyses of the 28-gene signature in disease-free survival prediction in three patient cohorts. *A*, time-dependent ROC ($t = 5$ y) curve of the 28-gene signature on the training set from Sotiriou et al. (8) AUC of 0.983. *B*, time-dependent ROC ($t = 5$ y) curves of the 28-gene signature on two validation sets. AUC of 0.843 with 25 overlapping genes on data from van't Veer et al. (1) and AUC of 0.764 with eight overlapping genes on data from Sorlie et al. (7). *C*, AUC in years 1 to 11 during follow-up after surgery in the patient cohort from Sotiriou et al. (8). *D*, AUC in years 1 to 13 during follow-up after surgery on two independent patient cohorts (1, 7).

Table 1. The 28-gene signature for predicting breast cancer recurrent and metastatic potential

Gene	UniGene cluster ID	Gene name
Unknown	Hs.463079	
TOMM70A	Hs.227253	Translocase of outer mitochondrial membrane 70 homologue a
MCF2	Hs.387262	Mcf.2 cell line-derived transforming sequence
RAD52 Pseudogene	Hs.552577	
MCM2	Hs.477481	Mcm2 minichromosome maintenance deficient 2, mitotin
C18B11	Hs.173311	RNA pseudouridylylate synthase domain containing 2
SEC13L	Hs.301048	Seh1-like
SLC25A5	Hs.522767	Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5
PLSCR1	Hs.130759	Phospholipid scramblase 1
RAD50	Hs.242635	
Unknown		
INPPL1	Hs.523875	Inositol polyphosphate phosphatase-like 1
Unknown	Hs.439445	
TXNRD1	Hs.434367	Thioredoxin reductase 1
PBX2	Hs.509545	Pre-B-cell leukemia transcription factor 2
SSBP1	Hs.490394	Single-stranded DNA binding protein 1
Unknown	Hs.448229	
PDGFRA	Hs.74615	Platelet-derived growth factor receptor
Unknown	Hs.49433	
DDOST	Hs.523145	Dolichyl-diphosphooligosaccharide-protein glycosyltransferase
Unknown	Hs.497723	
S100P	Hs.2962	S100 calcium binding protein
FAT	Hs.481371	Fat tumor suppressor homologue 1
FGF2	Hs.284244	Fibroblast growth factor 2
INSM1	Hs.89584	Insulinoma-associated 1
IRF5	Hs.521181	IFN regulatory factor 5
SMARCD2	Hs.250581	Swi/snf related, matrix associated, actin-dependent regulator of chromatin, subfamily d, member 2
MAP2K2	Hs.465627	Mitogen-activated protein kinase kinase 2

The data were publicly available at the *Proceedings of the National Academy of Sciences* Web site.⁷

The cohort from Sorlie et al. (7) contained 78 breast carcinomas (71 ductal, 5 lobular, and 2 ductal carcinomas *in situ* obtained from 77 different individuals; two independent tumors from one patient diagnosed at different times). Fifty-one of the patients were part of a prospective study on locally advanced breast cancer (T₃/T₄ and/or N₂ tumors) treated with preoperative doxorubicin monotherapy. A total of 58 patients with complete follow-up information were included in this study. The detailed clinical information of all samples is available at the *Proceedings of the National Academy of Sciences* journal Web site and the author Web site.⁸

The clinical data of 78 patients from van't Veer et al. (1) were analyzed in this study, including 34 patients developed metastasis within 5 years and 44 patients continued to be disease-free after 5 years. Twenty patients with hereditary breast cancer (BRCA1 or BRCA2 carriers) were excluded from this study. All of these patients were diagnosed between 1983 to 1996 with primary invasive carcinoma (T₁ or T₂), no axillary metastases (N₀), and all under the age of 55 years. The complete patient data are publicly available at the *Nature* journal Web site and the author Web site.⁹

Acquisition of gene expression profiles. There were 7,650 genes assayed by cDNA microarrays on 99 patient samples as described previously by Sotiriou et al. (8). RNA was isolated by using the Trizol method (Invitrogen, Carlsbad, CA). Total RNA from the Universal Human Reference (Stratagene, La Jolla, CA) was amplified and used as a reference for cDNA microarray analysis (8). The detailed protocols for RNA amplification and cDNA probe labeling and hybridization are available online.¹⁰

The data set from Sorlie et al. (7) includes 9,216 genes screened on 78 patient samples. Total RNA was isolated by phenol-chloroform extraction (Trizol, Life Technologies), and mRNA was purified by either magnetic separation using Dynabeads (Dyna) or the Invitrogen FastTrack 2.0 kit (7). All experiments and microarray assays were described previously (7) and the detailed protocols are available at Stanford University Web site.^{11,12}

The data provided by van't Veer et al. (1) contain 24,500 genes screened on 98 patient samples. The detailed protocols for RNA isolation, cRNA labeling, and gene expression profiling using microarray were described in the original publication (1).

Missing value replacement and gene matching. The data from Sotiriou et al. (8) were used as training set. The genes that had missing values in more than five samples were excluded from the study. For the remaining genes, missing values were replaced by using the *EMV* package in software R.¹³ The missing values were estimated based on a *k*-nearest neighbor algorithm (*k* = 20). This algorithm first selects *k* nearest genes that do not contain any missing values to the one with at least one missing value, based on the Euclidean distance. Then, the missing values are replaced by the average of the neighbors. Genes screened on different microarray platforms were matched by their UniGene Cluster IDs with the interactive *MatchMiner* (14) Web site interface.¹⁴

Feature selection for biomarker identification. Feature selection is important to identify relevant and important genes and to remove irrelevant genes and noise from large scale microarray data sets. A combination of random forests (15) and linear discriminant analysis (LDA) was used to identify gene signatures for predicting breast cancer

⁷ <http://www.pnas.org/cgi/content/full/100/18/10393>

⁸ http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=95

⁹ <http://www.rii.com/publications/2002/vantveer.html>

¹⁰ <http://nciarray.nci.nih.gov/reference/index.html>

¹¹ <http://cmgm.Stanford.edu/pbrown/>

¹² <http://genome-www.stanford.edu/molecularportraits/>

¹³ <http://www.r-project.org>

¹⁴ <http://discover.nci.nih.gov/matchminer/MatchMinerInteractiveLookup.jsp>

Table 1. The 28-gene signature for predicting breast cancer recurrent and metastatic potential (Cont'd)

Function	Cancer involvement	Breast cancer	Gene expression in high-risk group vs. low-risk group	Involvement
Mitochondrial structural protein			Decreased	<0.05
Guanine nucleotide exchange factor	(+)	(+)	Increased	<0.01
			Decreased	<0.05
DNA replication	(+)	(+) Biomarker (22)	Decreased	<0.01
			Increased	<0.01
mRNA export, nuclear pore distribution, and cell division	(+)		Increased	<0.05
Mitochondrial carrier			Increased	<0.01
Lipid transfer signaling	(+)		Increased	<0.01
DNA repair	(+)	(+) Biomarker (23)	Decreased	<0.05
			Increased	<0.01
Lipid metabolism	(+)		Decreased	<0.05
			Decreased	<0.01
Antioxidant and redox regulator	(+)	(+)	Increased	<0.01
Transcriptional repressor and tumor suppressor	(+)		Increased	<0.05
DNA binding protein	(+)		Increased	<0.01
			Increased	<0.01
Growth factor receptor	(+)	(+) Biomarker (24, 25)	Not significantly different	
			Decreased	<0.05
Structure			Increased	<0.01
			Decreased	<0.05
Cell differentiation	(+)	(+) Biomarker (26, 27)	Increased	<0.01
Cell signaling suppressor	(+)		Not significantly different	
Signaling transduction	(+)	(+) Biomarker (28)	Decreased	<0.05
Transcriptional repressor	(+)		Not significantly different	
Tumor suppressor gene	(+)		Increased	<0.01
Chromatin remodeling	(+)	(+)	Decreased	<0.05
Signaling transduction	(+)	(+)	Decreased	<0.05

recurrence/metastases, tumor grade, and nodal status. Random forests of software *R* was first used to identify a small subset of genes from the original microarray data. LDA of software SAS¹⁵ was used to further refine the gene signature.

Random forests is a generalization of the standard tree algorithms (16). The basic step of random forests is to form diverse tree classifiers from a single training set. Each tree is built on a bootstrap sample from the training set. The variables used for splitting the tree nodes are a random subset of the whole variables set. The classification decision of a new case is obtained by majority voting (unless the cutoff value is user defined) over all trees. In random forests, about one third of the cases in the bootstrap sample are not used in growing the tree. These cases are called "out-of-bag" cases and are used to evaluate the algorithm performance. A very important function of random forests is variable importance evaluation. The importance of a variable is defined in terms of its contribution to classification accuracy. Based on the variable importance measure, backward elimination was used to identify the gene subset with the smallest out-of-bag error rate. Here, the out-of-bag error rate was not used to assess the prediction accuracy of the identified gene subsets. Instead, it served as a stopping rule for feature selection. The varSelRF package of software *R* (17) was used according to the following steps: (1) build a forest with *N* trees and obtain a ranking of variable importance; (2) remove 20% of the least important variables; (3) construct a new forest with *K* trees; (4) repeat steps 2 and 3 until two genes are left; and (5) select the gene subset with the smallest out-of-bag error rate.

In the experiments, we chose *N* = 3,000 and *K* = 1,000 because the large number of trees in the initial forests are likely to produce stable importance measures (17). We did not follow the "1-SE rule" as suggested by Diaz-Urriarte et al. (17). This rule chooses the smallest gene

subset, whose error rate is within one SE of the minimum error rate of all forests. We used the "0-SE rule," which identifies the gene subset with the smallest out-of-bag error rate. The "0-SE rule" usually selects more genes than the "1-SE rule" does. Because further gene filtering would be done by using LDA, we chose the gene subsets with the lowest prediction error using random forests.

Discriminant analysis was used to determine which variables discriminate two or more naturally occurring groups in prognosis. Given several variables as the data representation, each class is modeled as multivariate normal distribution with a covariance matrix and a mean vector. Instances are classified to the label of the nearest mean vector based on Mahalanobis distance. The decision surfaces between classes become linear if the classes have a common covariance matrix. When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function. Such function is determined by a measure of generalized square distance, which is based on the pooled covariance matrix as well as the prior probabilities of group membership. The generalized squared distance D_i^2 from input *x* to class *i* is

$$D_i^2(x) = d_i^2(x) + g(i)$$

where $d_i^2(x) = (x - m_i)'V^{-1}(x - m_i)$ is the squared distance from *x* to group *I*; m_i is the *p*-dimensional mean vector for group *I*; *V* is the pooled covariance matrix and $g(i)$ depends on the prior probability of class *i*. In practice, the prior probability can be assumed as equal for all groups (refer to SAS Users' Manual). In this study, we assumed equal prior probability and thus $g(i) = 0$. *x* is classified into class *I*, if $D_i^2(x)$ is the smallest among all the distance measures. We selected the gene markers using backward selection of stepwise discriminant analysis with software SAS.

¹⁵ <http://www.sas.com>

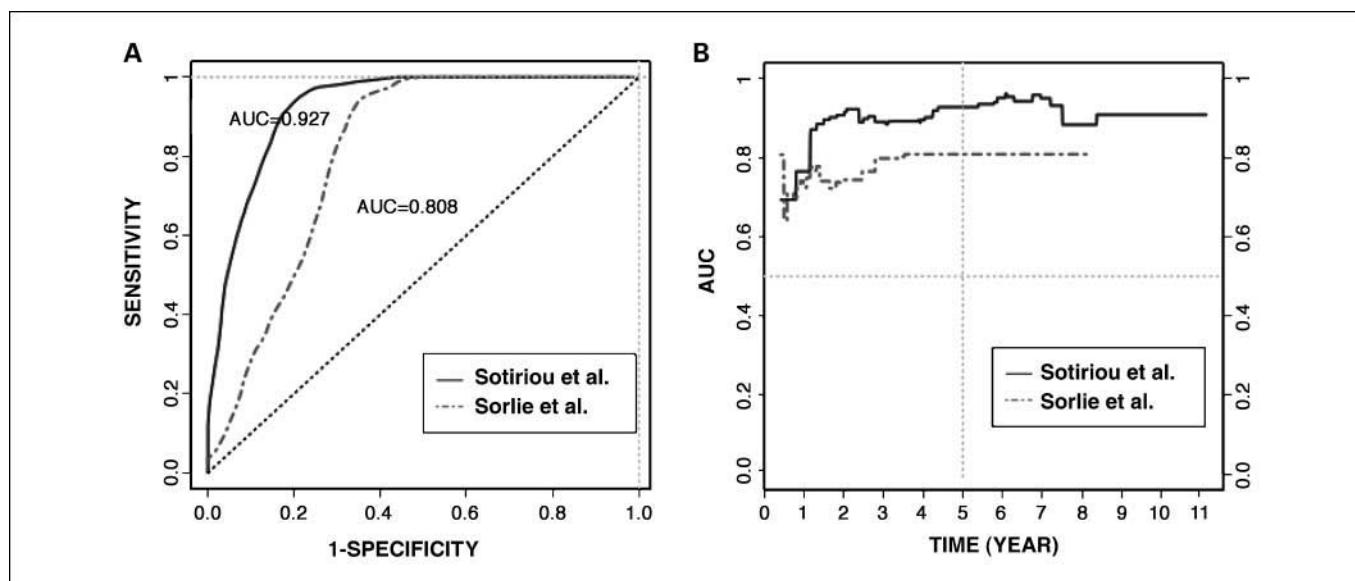


Fig. 2. Time-dependent ROC analyses of the 28-gene signature in overall survival prediction. *A*, time-dependent ROC curves at *t* of 5 y. AUC of 0.927 on data from Sotiriou et al. (8) and AUC of 0.808 on data from Sorlie et al. (7). *B*, the AUC of overall survival prediction during the follow-up after surgery.

Classifier evaluation. LDA of software SAS was used to refine the gene signature obtained from random forests and assess the classification accuracy of models in predicting 5-year relapse-free survival, tumor grade, and nodal status based on the identified gene signatures. Leave-one-out cross-validation was used in the evaluation (18).

To evaluate the accuracy of survival prediction, time-dependent receiver operating characteristic (ROC) analysis for censored data (19, 20) was done with software *R*. Time-dependent ROC analysis extends the concepts of sensitivity, specificity, and ROC curves for time-dependent binary disease variables in censored data. In our study, the binary disease variable $R_i(t) = 1$, if patient *i* has recurrent or metastatic breast cancer before time *t*; otherwise, $R_i(t) = 0$. For a diagnostic marker *M*, both sensitivity and specificity are defined as a function of time *t*:

$$\text{sensitivity}(c, t) = P\{M > c \mid R(t) = 1\}$$

$$\text{specificity}(c, t) = P\{M \leq c \mid R(t) = 0\}$$

A $ROC(t)$ is a function of *t* at different cutoffs *c*. A time-dependent ROC curve is a plot of sensitivity(*c*, *t*) versus 1 – specificity(*c*, *t*). The area under the ROC curve (AUC) can be used as an accuracy measure of the ROC curve. A higher prediction accuracy is evidenced by a larger $AUC(t)$ (19, 20).

Statistical methods. To estimate a patient’s recurrent and metastatic potential, risk scores were generated by fitting the identified gene predictors in a Cox regression model as covariates. The distribution of the risk scores was used to classify the patients into three groups: high risk, low risk, and intermediate risk. Kaplan-Meier analysis was used to assess the disease-free survival probability of three risk groups in the studied patient cohorts (1, 7, 8). To evaluate the association between risk groups and clinicopathologic variables in three patient cohorts (1, 7, 8), the χ^2 test or Fisher’s exact test was used. All statistical testing was done with software *R*.

Results

Gene expression–based prediction of recurrence and metastases. The expression profiles and clinical data from Sotiriou

et al. (8) were used as the training set to predict recurrence in individual breast cancer patients. To identify gene predictors of breast cancer recurrence, 96 tumor samples were used in the classification, and three samples were excluded because their 5-year relapse status could not be determined. In this population-based cohort, 59 patients had relapse within 5 years after surgery, and 37 remained relapse-free for a 5-year interval. To reliably identify good and poor prognostic tumors, a powerful three-step supervised classification method was used. In brief, based on the expression profiles of 7,091 genes after data preprocessing, 66 genes were first identified by using random forests. Then, LDA analysis further refined the recurrence signature to 28 genes (Table 1; Supplementary Fig. S1). Based on the expression profiles of these 28 genes, LDA classified 5-year relapse status for the 96 patients, achieving an overall accuracy of 0.92 (88 of 96), a sensitivity of 0.90 (53 of 59), and a specificity of 0.95 (35 of 37). To evaluate relapse-free survival prediction in the complete cohort between the years 1 and 11 after surgery, a Cox proportional hazards model (21) was built on the 28-gene signature and the risk score was used to construct the time-dependent ROC curve. The AUC (5-year) was 0.983 (Fig. 1A) and remained 0.92 between years 8 and 11 during the follow-up (Fig. 1C).

To evaluate the prognostic power of our identified gene signature, two independent validation sets were used (1, 7). From each validation set, we identified the overlapped genes with our 28-gene signature. Eight genes were found in the data generated by Sorlie et al. (7), including one unknown gene. Twenty-five genes were obtained from the data generated by van’t Veer et al. (1), in which 4 genes were duplicated. Using these overlapped genes, time-dependent ROC analyses were done to evaluate relapse/metastases prediction on two independent patient cohorts (Fig. 1B and D). The AUC (5-year) on the data from van’t Veer et al. (1) was 0.843 with 25 overlapped genes in predicting metastatic potential. The AUC (5-year) was 0.764 on the data from Sorlie et al. (7) with eight overlapped genes in the relapse-free survival prediction (Fig. 2B). The

patients from van't Veer et al. (1) were with primary invasive carcinoma (T_1 or T_2) and no lymph node involvement (N_0); whereas 88% (51 of 58) of the patients from Sorlie et al. (7) were with locally advanced (T_3/T_4 and/or N_2) breast cancer. The results showed that our identified gene predictors from the population-based study achieved accurate disease-free survival prediction in patients with heterogeneous histology and disease stages.

Time-dependent ROC analysis showed that the identified 28-gene signature was also predictive of overall survival ($P < 0.001$; Fig. 2). In the prediction of overall survival, the AUC (5-year) was 0.927 on the data from Sotiriou et al. (8) with all 28 genes and 0.808 on data from Sorlie et al. (7) with 8 overlapped genes.

The identified 28-gene signature of recurrent and metastatic potential is unique. There is no overlap between our gene signature and previously reported survival signatures (1, 8, 12). A literature search found that 17 genes of this 28-gene signature are related to tumorigenesis, and 9 genes are directly linked to breast cancer pathogenesis (Table 1). Furthermore, among the nine breast cancer-related genes, five genes are the established breast cancer biomarkers [i.e., *MCM2* (22), *RAD50* (23), *PDGFRA* (24, 25), *S100P* (26, 27), and *FGF2* (28; Table 1).

Significant association of expression profile-defined risk groups with clinicopathologic characteristics. The clinical variables, such as nodal status, tumor size, tumor grade, ER status, and HER2/*neu* overexpression in breast cancer patients affect the disease outcomes. In three studied cohorts (1, 7, 8), the distribution of the recurrence risk scores in Cox modeling was used to stratify patients into three groups: high risk, low risk, and intermediate risk (the details are provided in Supplementary Materials). Kaplan-Meier analysis showed that disease-free survival was significantly different ($P < 0.005$, log-rank tests) among the risk groups in all patient cohorts (Fig. 3). The clinical characteristics of each risk group in the studied cohorts were reported in Supplementary Tables S5 to S7, including average disease-free survival days, ER and PR status, HER2/*neu* overexpression, nodal status, age, tumor size, grade, and treatment received. The results indicated that the expression profiles of the identified 28-gene signature were strongly associated with the clinicopathologic variables, including tumor size, tumor grade, ER and PR status, and HER2/*neu* overexpression ($P < 0.05$; Table 2). There was insufficient evidence to support the association between ER status and risk groups in Sorlie et al. (7). The reason might be that 80% of the patients in this study exhibited ER positive. No strong evidence was found that age (representing menopausal status) of the patients was associated with the expression profiles of the tumors (Table 2).

Gene predictors of tumor grade and nodal status. To further reveal the molecular pathogenesis of breast cancer, we also identified 14-gene predictors of nodal status and 9-gene predictors of tumor grade (Table 3; Supplementary Fig. S1; Supplementary Table S1) using the combination of random forests and LDA. Based on the expression profiles of the 14-gene signature, LDA correctly classified 80% (79 of 99) of nodal status (node positive versus node negative) in the patients from Sotiriou et al. (8), with a sensitivity of 83% (44 of 53) and a specificity of 76% (35 of 46). Among these 14 nodal status predictor genes, the functions of four genes are unknown,

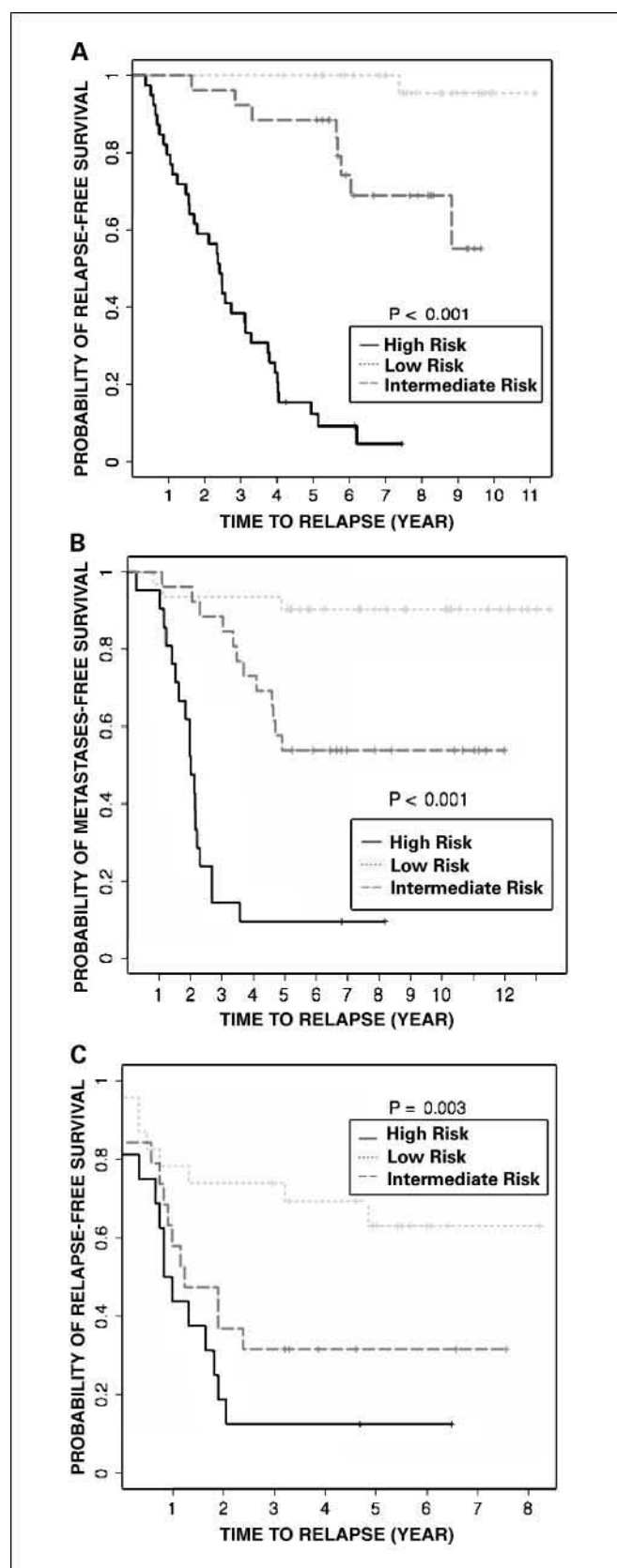


Fig. 3. Kaplan-Meier analysis of disease-free survival of risk groups in three patient cohorts (1, 7, 8). **A.** Kaplan-Meier analysis on data from Sotiriou et al. (8). **B.** Kaplan-Meier analysis on data from van't Veer et al. (1). **C.** Kaplan-Meier analysis on data from Sorlie et al. (7).

whereas the remaining 10 genes are all related to cancer development and progression. Furthermore, six genes are directly related to breast cancer pathogenesis (Table 3). It has also been found that VEGFB is a well-established breast cancer nodal status biomarker (29), whereas STK12 (30, 31) and BIRC3 (32) are the biomarkers for breast cancer risk and survival.

Based on the expression profiles of the nine-gene signature, LDA correctly classified 85% (84 of 99) of tumor grade (grade 1 or 2 versus grade 3; ref. 13) in the patients from Sotiriou et al. (8), with a sensitivity of 87% (39 of 45) and a specificity of 83% (45 of 54). Interestingly, for these nine-gene predictors, all except *ALDH3A2* (aldehyde dehydrogenase 3 family for fat metabolism) are related to cancer development and progression. More importantly, six genes are directly related to breast cancer pathogenesis. Furthermore, *RUNX1* was identified as a molecular target of breast cancer, which is associated with the transforming growth factor- β signal transduction pathway (Table 3; ref. 33).

Discussion

In this population-based study, we were able to quantify the likelihood of local and distant recurrence in breast cancer patients with heterogeneous histologies and stages. The use of the gene expression-based prognostic model provides an accurate estimate of the risk of recurrence and metastases in individual patients. The patients in different recurrence risk categories had distinct disease-free survival ($P < 0.001$) in three studied cohorts. The recurrence gene signature can also accurately predict overall survival. This feature is remarkable because ~50% of patients died in the absence of recurrent breast cancer (1). In addition, the recurrence gene signature predicts the relapse-free interval and metastases-free interval. Therefore, the expression profile-defined prognostic model robustly predicts all the outcomes we examined.

Microarray analyses of breast cancer have identified gene expression profiles associated with patient survival. Sorlie et al. (7) found the expression profiles to distinguish ER-positive from ER-negative tumors with distinct outcomes in a locally advanced cohort treated with primary chemotherapy. van't Veer et al. (1) established a 70-gene signature to predict metastatic potential in an untreated, node-negative cohort. Sotiriou et al. (8) showed the concordance with these previous analyses in node-positive and node-negative patients with the majority

receiving adjuvant treatment. Here, we present a powerful systems biology approach to identify unique gene predictors of breast cancer recurrence and metastases, which achieved highly accurate prognosis at the individual level in these three cohorts (1, 7, 8). Our results are remarkable in their high prognostic accuracy for the data from these previous studies despite the differences in patient populations, treatments used, and technology platforms used.

The variables, such as patient's age, tumor size, measurement of ER/PR (by ligand-binding assay), and *HER2/neu* status (by fluorescence *in situ* hybridization), are routinely used as predictors of recurrence in breast cancer and are incorporated into current treatment guidelines (34, 35). The results of this study show that the expression profiles of our recurrence gene predictors were strongly associated ($P < 0.05$) with tumor size, tumor grade, ER, PR, and *HER2/neu* in the studied patient cohorts. The gene expression profiles in tumors were not significantly associated with the patient's age in all three cases.

We evaluated the recurrence gene signature in the context of the interobserver variability in tumor grading that is typical in oncology practice. Tumor grade correlates with the likelihood of recurrence when analyzed in large populations of patients (12). However, previous studies have also reported that the grading of breast cancer entails subjectivity, leading to considerable variability among pathologists (12, 36–40). Our results indicate that the expression of a recurrence gene signature is strongly associated ($P < 0.05$) with tumor grade in the patients from all three cohorts.

A recent report by a Breast Task Force serving the American Joint Committee on Cancer did not add tumor grade to its staging criteria because of the interobserver variability problem in the current grading system (39). To produce a consistent and standard system for tumor grading, we identified a nine-gene signature to predict tumor grade. The expression profiles of the nine genes (Table 3) accurately classified 85% of the tumor grades in the population-based cohort from Sotiriou et al. (8), with a sensitivity of 87% and a specificity of 83%. The prediction accuracy of our identified nine-gene grade signature was notably high.

The high dimensionality of microarray data has complicated major diagnostic and prognostic breakthroughs in cancer treatment and puts a premium on innovative data mining methods. This article presents a novel computational gene selection system for the identification of molecular

Table 2. The association of expression profile-defined risk groups and clinicopathologic variables in three patient cohorts

Risk groups	P		
	Sotiriou et al. (8)	van't Veer et al. (1)	Sorlie et al. (7)
Age* (<50 or \geq 50 y)	0.243	0.458	0.095
Tumor size (<2 or >2 cm)	0.006	0.047	
Tumor grade (1/2 vs 3)	0.041	0.004	0.001
ER status	0.011	0.004	0.296
PR status		0.001	
<i>HER2/neu</i>	0.020		

*The percentage of patients who were at least 50 y old was 74%, 28%, and 69% in the cohorts from Sotiriou et al. (8), van't Veer et al. (1), and Sorlie et al. (7), respectively.

Table 3. The 14-gene predictors of breast cancer nodal status and the 9-gene predictors of breast cancer tumor grade

Gene	UniGene cluster ID	Gene name	Function	Cancer involvement	Breast cancer involvement
14-Gene predictors of nodal status					
<i>TLR5</i>	Hs.114408	<i>Toll-like receptor 5</i>	Immune system	(+)	
<i>FLJ21128</i>	Hs.96852	<i>Hypothetical protein flj21128 (preferred)</i>			
<i>RBMX</i>	Hs.380118	<i>RNA binding motif protein, x-linked</i>	Apoptosis modulator	(+)	(+)
Unknown	Hs.522309				
<i>HOXD1</i>	Hs.83465	<i>Homeo box d1</i>	Embryonic development	(+)	
Unknown					
Unknown	Hs.390738				
<i>VEGFB</i>	Hs.78781	<i>Vascular endothelial growth factor b</i>	Signaling transduction	(+)	(+) Biomarker (29)
<i>STK12</i>	Hs.442658	<i>Aurora kinase b</i>	Cell cycle	(+)	(+) Biomarker (30, 31)
<i>MAPK12</i>	Hs.432642	<i>Mitogen-activated protein kinase 12</i>	Signaling transduction	(+)	(+)
<i>BIRC3</i>	Hs.127799	<i>Baculoviral IAP repeat-containing 3</i>	Apoptosis suppressor	(+)	(+) Biomarker (32)
<i>ITGA7</i>	Hs.524484	<i>integrin, $\alpha 7$</i>	Signaling transduction	(+)	
<i>CHC1L</i>	Hs.25447	<i>Chromosome condensation 1-like</i>	Tumor suppressor	(+)	
<i>SCYB14</i>	Hs.483444	<i>Chemokine (c-x-c motif) ligand 14</i>	Tumor suppressor	(+)	(+)
9-Gene predictors of tumor grade					
<i>ALDH3A2</i>	Hs.499886	<i>Aldehyde dehydrogenase 3 family</i>	Fat metabolism		
<i>NK4</i>	Hs.943	<i>Interleukin-32</i>	Immune response, apoptosis	(+)	(+)
<i>BUB1</i>	Hs.469649	<i>Bub1 budding uninhibited by benzimidazoles 1 homologue</i>	Cell cycle	(+)	(+)
<i>RUNX1</i>	Hs.149261/ Hs.278446	<i>Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)</i>	Transcription factor	(+)	(+) Biomarker (33)
<i>ZSIG37</i>	Hs.201398	<i>C1q and tumor necrosis factor-related protein 1</i>	Immune response, apoptosis	(+)	
<i>SSI-1</i>	Hs.50640	<i>Suppressor of cytokine signaling 1</i>	Signaling transduction	(+)	(+)
<i>HDAC2</i>	Hs.3352	<i>Histone deacetylase 2</i>	Transcription	(+)	(+)
<i>HMG2</i>	Hs.434953	<i>High-mobility group box 2</i>	Structure protein	(+)	(+)
<i>NFIX</i>	Hs.257970	<i>Nuclear factor κB (ccat-binding transcription factor)</i>	Gene expression and DNA replication	(+)	

signatures. An integration of random forests and LDA was used in this study. Random forests is characterized as an effective machine learning method for processing noisy large-scale data sets. Therefore, we used this algorithm to filter out noninformative genes sequentially from the original microarray data until a small subset of genes was obtained. Then, LDA was used to further filter out more genes. Our

previous studies have shown that an integrative feature selection system based on random forests and other state-of-the-art techniques enables the identification of important biomarkers (41, 42). This study is another successful example of the combinatorial gene selection system, providing a general guideline for developing robotic prognosis toward personalized medicine.

References

- van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
- Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439:353–7.
- Glinisky GV, Higashiyama T, Gliniskii AB. Classification of human breast cancer using gene expression profiling as a component of the survival predictor algorithm. *Clin Cancer Res* 2004;10:2272–83.
- Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. *Lancet* 2003;361:1590–6.
- Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis, and treatment selection. *Nat Rev Cancer* 2005;5:845–56.
- Murphy N, Millar E, Lee CS. Gene expression profiling in breast cancer: towards individualising patient management. *Pathology* 2005;37:271–7.
- Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001;98:10869–74.
- Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003;100:10393–8.
- van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;347:1999–2009.
- West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 2001;98:11462–7.
- Zhao H, Langerod A, Ji Y, et al. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol Biol Cell* 2004;15:2523–36.
- Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.
- Pinder SE, Murray S, Ellis IO, et al. The importance of the histologic grade of invasive breast carcinoma and response to chemotherapy. *Cancer* 1998;83:1529–39.
- Bussey KJ, Kane D, Sunshine M, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol* 2003;4:R27.
- Breiman L. Random forests. *Machine Learning* 2001;45:5–32.
- Gentleman R, Huber W, Carey VJ, Irizarry RA, Dudoit S, editors. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Springer; 2005.
- Diaz-Uriarte R, Alvarez dA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006;7:3.
- Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. San Mateo (CA): Morgan Kaufmann; 1995. p. 1137–43.
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56:337–44.

20. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;61:92–105.
21. Cox D. Regression models and life-tables (with discussion). *J R Stat Soc Ser B Methodol* 1972;34:187–220.
22. Gonzalez MA, Pinder SE, Callagy G, et al. Minichromosome maintenance protein 2 is a strong independent prognostic marker in breast cancer. *J Clin Oncol* 2003;21:4306–13.
23. Tammiska J, Seal S, Renwick A, et al. Evaluation of RAD50 in familial breast cancer predisposition. *Int J Cancer* 2006;118:2911–6.
24. Carvalho I, Milanezi F, Martins A, Reis RM, Schmitt F. Overexpression of platelet-derived growth factor receptor α in breast cancer is associated with tumour progression. *Breast Cancer Res* 2005;7:R788–95.
25. Jechlinger M, Sommer A, Moriggl R, et al. Autocrine PDGFR signaling promotes mammary cancer metastasis. *J Clin Invest* 2006;116:1561–70.
26. Guerreiro DSI, Hu YF, Russo IH, et al. S100P calcium-binding protein overexpression is associated with immortalization of human breast epithelial cells *in vitro* and early stages of breast cancer development *in vivo*. *Int J Oncol* 2000;16:231–40.
27. Schor AP, Carvalho FM, Kemp C, Silva ID, Russo J. S100P calcium-binding protein expression is associated with high-risk proliferative lesions of the breast. *Oncol Rep* 2006;15:3–6.
28. Granato AM, Nanni O, Falcini F, et al. Basic fibroblast growth factor and vascular endothelial growth factor serum levels in breast cancer patients and healthy women: useful as diagnostic tools? *Breast Cancer Res* 2004;6:R38–45.
29. Gunningham SP, Currie MJ, Han C, et al. VEGF-B expression in human primary breast cancers is associated with lymph node metastasis but not angiogenesis. *J Pathol* 2001;193:325–32.
30. Cox DG, Hankinson SE, Hunter DJ. Polymorphisms of the AURKA (STK15/Aurora kinase) gene and breast cancer Risk (United States). *Cancer Causes Control* 2006;17:81–3.
31. Tchatchou S, Wirtenberger M, Hemminki K, et al. Aurora kinases A and B and familial breast cancer risk. *Cancer Lett*. Epub 2006 Jun 6.
32. Span PN, Tjan-Heijnen VC, Heuvel JJ, de Kok JB, Foekens JA, Sweep FC. Do the survivin splice variants modulate or add to the prognostic value of total survivin in breast cancer? *Clin Chem* 2006;52:1693–700. Epub 2006 Jul 27.
33. Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z, Jeffrey SS. A molecular 'signature' of primary breast cancer cultures; patterns resembling tumor tissue. *BMC Genomics* 2004;5:47.
34. Eifel P, Axelson JA, Costa J, et al. National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. *J Natl Cancer Inst* 2001;93:979–89.
35. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Thurlimann B, Senn HJ. Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. *Ann Oncol* 2005;16:1569–83.
36. Davis BW, Gelber RD, Goldhirsch A, et al. Prognostic significance of tumor grade in clinical trials of adjuvant therapy for breast cancer with axillary lymph node metastasis. *Cancer* 1986;58:2662–70.
37. Hopton DS, Thorogood J, Clayden AD, MacKinnon D. Observer variation in histological grading of breast cancer. *Eur J Surg Oncol* 1989;15:21–3.
38. Robbins P, Pinder S, de Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. *Hum Pathol* 1995;26:873–9.
39. Singletary SE, Allred C, Ashley P, et al. Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol* 2002;20:3628–36.
40. Theissig F, Kunze KD, Haroske G, Meyer W. Histological grading of breast cancer. Interobserver, reproducibility, and prognostic significance. *Pathol Res Pract* 1990;186:732–6.
41. Guo L, Ma Y, Ward R, Castranova V, Shi X, Qian Y. Constructing molecular classifier for accurate prognosis of lung adenocarcinoma. *Clin Cancer Res* 2006;12:3344–54.
42. Ma Y, Ding Z, Qian Y, et al. Predicting cancer drug response by proteomic profiling. *Clin Cancer Res* 2006;12:4583–9.