Supplementary Data for

Population-based Molecular Prognosis of Breast Cancer by Transcriptional Profiling

Yan Ma, Yong Qian, Liang Wei, Jame Abraham, Xianglin Shi, Vincent Castranova, E. James Harner, Daniel C. Flynn, and Lan Guo

Table of Contents

Supplementary Figures

Supplementary Tables

1 Experimental Database

1.1 Training Data Source

The gene expression and clinical data from Sotiriou *et al.* (1) were used as training set. There were 7,650 genes assayed by cDNA microarrays on 99 patient samples. All of the tumors were invasive ductal carcinomas: 46 patients were node negative and 53 were node positive (please refer to the original paper for the details). The data were publicly available as the *supporting information* on the PNAS website: http://www.pnas.org/cgi/content/full/100/18/10393.

1.2 Training Data Pre-processing

The data pre-processing of the training set consists of two steps: (1) remove genes with more than five missing values across all the samples. In this step, 559 genes were eliminated, and (2) replace missing values by using the *EMV* package in software *R* (http://www.r-project.org). 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 containing at least one missing value, based on the Euclidean distance. Then, the missing values are replaced by the average of the neighbors. After the data pre-processing was performed, 7,091 genes remained in the data set.

1.3 Validation Data Sources

Two validation data sets were used in our analysis. One was from a publication by Sorlie *et al.* (2), which is available from the website: http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=95. The data set in Sorlie *et al.* (2) includes 9,216 genes screened on 78 patient samples. We used the expression profiles of 58 patients in our model validation. The remaining cases were not included because patient disease-free survival information was not available.

The other data set was from van't Veer *et al.* (3), which is available at: http://www.rii.com/publications/2002/vantveer.html. This data set contains 24,500 genes screened on 98 patient samples. In the studied cohort, 34 patients developed metastasis within 5 years and 44 patients continued to be disease-free after five years. In addition, there were 18 patients with BRCA1 germline mutations and two BRCA2 carriers. In our validation, we used the cohort of 78 patients, excluding 20 patients with BRCA mutations.

2. Study Design

We built a prediction model for each of three prognostic factors: relapse/metastases potential (relapse/metastases within five years vs. disease-free in five years), nodal status (node negative vs. node positive), and tumor grade (grade 1/2 vs. 3). Using the data set from Sotiriou *et al.* (1), the classification models for nodal status and tumor grade were constructed based on the expression levels of 7,091 genes (after pre-processing) on 99 clinical specimens. The prediction model for 5-year disease-free survival was built on 96 clinical samples, omitting 3 patients whose 5-year relapse-free survival could not be determined.

Marker genes from these prognostic models were identified by using a combination of Random Forests of software *R (http://www.r-project.org)* and Linear Discriminant Analysis of Software *SAS (http://www.sas.com/)*. Random forests were first used to select a small subset of genes, and then, Linear Discriminant Analysis was used to further refine the gene signatures. Feature selection is important to identify relevant and important genes and to remove irrelevant genes and noise from large scale microarray data sets. The random forests algorithm (4) utilizes an ensemble of classification trees. Random forests are characterized as an effective machine learning method for processing noisy large-scale data sets. Therefore, we employed this algorithm to filter out non-informative genes sequentially until a small subset of genes was obtained. Then, SAS PROC STEPDISC procedure for Linear Discriminant Analysis was used to further filter out more genes. Backward elimination in PROC STEPDISC selected a much smaller subset of genes that generated favorable prediction accuracy. The details of both random forests and discriminant analysis are described in the following sections.

3. Random Forests

Random forests are a generalization of the standard tree algorithms (5). The random forests (4) algorithm is an ensemble of un-pruned classification trees. The basic step of random forests is to form *diverse* base tree classifiers from a single training set. Two sources of randomness are introduced: (1) each tree is built upon a bootstrap sample (a random sample taken with replacement) from the training set. Bootstrapping generates diverse versions of training data (6), and (2) only a subset of variables is explored to split each node in the tree. Therefore, the optimal split of a single node is based on a random subset of the variables instead of the whole variables set. Each tree generates its own classification rules. The classification decision for a given input case is made by majority voting over all trees.

About one-third of the cases in the bootstrap sample are not used to construct a classification tree. These samples are called *out-of-bag (OOB)* cases. For each tree, the OOB cases are used to get a classification result. For each sample in the training set, the final classification of the forest is the class having the most votes from the bootstrapped OOB cases. Comparison between this classification and the real class label in the data generates an unbiased estimator of the error rate. Therefore, random forests do not need a separate test set or additional cross-validation to evaluate its results (4).

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 performed to identify the gene subset with the smallest OOB error rate. Here, the OOB 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 *R* (7) 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
(5). Select the gene subset with the smallest OOB error rate

In the experiments, we chose *N* = 3,000 and *K* =1,000 because a large number of trees in the initial forest is likely to produce stable importance measures (7). We did not follow the "1-Standard Error (1-SE) rule" as suggested by Diaz-Uriarte *et al*. (7). This rule chooses the smallest gene subset, whose error rate is within one standard error of the minimum error rate of all forests. We used the "0-Standard Error (0-SE) rule", which identifies the gene subset with the smallest OOB error rate. The "0-SE rule" usually selects more genes than the "1-SE rule". Since further gene filtering would be pursued by using Linear Discriminant Analysis, we chose the gene subsets with the lowest prediction error for modeling disease-free survival, nodal status, and tumor grade (*Table 1*). *Figure 1* shows the feature selection process in each model by using random forests.

Table S1. Summary of feature selection processes for each prediction model. The discriminant function was used to compute the prediction accuracies.

Model	**# of Genes Obtained Using Random Forests**	**# of Genes Obtained Using SAS PROC STEPDISC**	**Accuracy**	**Sensitivity**	**Specificity**
5-Year-Relapse (relapse vs. relapse-free)	66	29	92% (88/96)	90% (53/59)	95% (35/37)
Nodal Status (positive vs. negative)	42	14	80% (79/99)	83% (44/53)	76% (35/46)
Tumor Grade (grade 1/2 vs. 3)	18	9	85% (84/99)	87% (39/45)	83% (45/54)

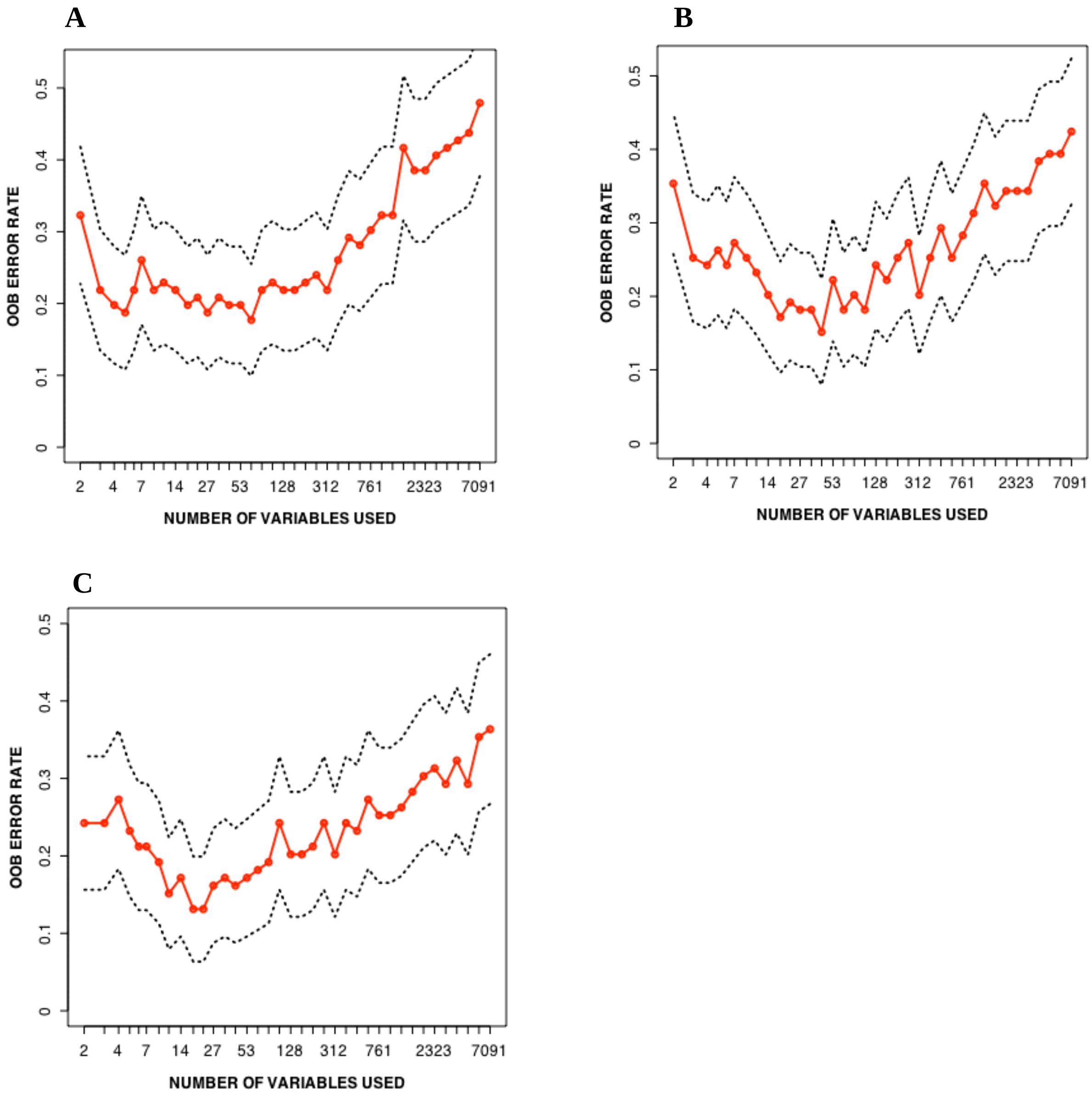

Figure S1. Feature selection using random forests. In each panel, the red line indicates the OOB error rate (vertical axis) at different number of genes (horizontal axis). The two dashed lines are two standard errors above/below the error rates. A. Disease-free survival; B. Nodal Status; C. Tumor Grade.

4. Linear Discriminant Analysis

Discriminant analysis was used to determine which variables discriminate two or more naturally occurring groups in prognosis. Given a number of 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 (SAS User's Guide, version 9.1). 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 $D_i^2(x)$ 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. 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.

A common application of discriminant function analysis is feature selection, i.e., determining the attributes that discriminate between classes. In this study, we selected features using a stepwise backward search with SAS PROC STEPDISC. Initially, all variables (i.e., genes) in the subset identified using random forests are included in the model and the variable that contributes the least to the prediction of class membership then is eliminated. By doing so, one would only keep the variables that contribute the most to the discrimination between classes in the model. The final gene subsets obtained from SAS PROC STEPDISC for each prognostic model are shown in *Tables 2-4*. We used leave-one-out cross-validation to assess the prediction accuracy. The classification accuracies based on the final gene set in each model are shown in *Table 1*.

Table S2. A 28-gene relapse signature. This gene signature achieves 92% accuracy in predicting the relapse status (relapse vs. relapse-free in a 5-year period). Note: '-' means that the gene name is not available.

Gene	Spot ID	Clone_IMAGE	UniGene Cluster ID
-	3912	198917	Hs.463079
TOMM70A	4919	198312	Hs.227253
MCF2	2370	268412	Hs.387262
RAD52 Pseudogene	418	1377154	Hs.552577
MCM2	1881	239799	Hs.477481

C18B11	5984	131988	Hs.173311
SEC13L	6497	757210	Hs.301048
SLC25A5	5182	291660	Hs.522767
PLSCR1	6959	268736	Hs.130759
TXNRD1	7296	789376	Hs.434367
RAD50	2925	261828	Hs.242635
-	6498	46196	
INPPL1	1987	703964	Hs.523875
-	583	501651	Hs.439445
TXNRD1	6736	789376	Hs.434367
PBX2	536	80549	Hs.509545
SSBP1	3434	125183	Hs.490394
-	2403	34396	Hs.448229
PDGFRA	6674	376499	Hs.74615
-	6555	488202	Hs.49433
DDOST	2416	50666	Hs.523145
-	2276	182930	Hs.497723
S100P	5593	135221	Hs.2962
FAT	7009	591266	Hs.481371
FGF2	3514	324383	Hs.284244
INSM1	3061	22895	Hs.89584
IRF5	5962	260035	Hs.521181
SMARCD2	2923	741067	Hs.250581
MAP2K2	1652	769579	Hs.465627

Table S3. A 14-gene signature achieves 80% accuracy in predicting nodal status (positive vs. negative).

Gene	SPOT ID	Well ID	Clone IMAGE	UniGene Cluster ID
TLR5	1635	208694	277229	Hs.114408
FLJ21128	2062	207691	279077	Hs.96852
RBMX	2159	202137	841352	Hs.380118
-	3303	27894	955999	Hs.522309
HOXD1	3607	202214	342593	Hs.83465
-	3735	26914		
-	4151	209569	50635	Hs.390738
VEGFB	4777	28189	167296	Hs.78781
STK12	4825	28957	241029	Hs.442658
MAPK12	5195	27254	309482	Hs.432642
BIRC3	6757	150040	428231	Hs.127799
ITGA7	6932	208400	377671	Hs.524484

CHC1L	7058	200264	768316	Hs.25447
SCYB14	7385	207798	345034	Hs.483444

Table S4. A 9-gene signature achieves 85% accuracy in predicting tumor grade (grade 1/2 vs. 3).

Gene	SPOT ID	Well ID	Clone IMAGE	UniGene Cluster ID
ALDH3A2	879	27777	767804	Hs.499886
NK4	1509	201566	810859	Hs.943
BUB1	3087	208790	781047	Hs.469649
RUNX1	3928	201479	773215	Hs.149261, Hs.278446
ZSIG37	4160	209599	78844	Hs.201398
SSI-1	5865	201673	712668	Hs.50640
HDAC2	6713	28381	712866	Hs.3352
HMG2	2353	28253	341782	Hs.434953
NFIX	3177	200222	754600	Hs.257970

5. Time-dependent ROC Curves

Both sensitivity and specificity are the most widely used statistics to describe a diagnostic test. Sensitivity measures the probability of a positive test among patients with disease, while specificity quantifies the chance of a negative test among patients without disease. Receiver operating characteristics (ROC) curve displays *1- specificity* vs. *sensitivity* of a diagnostic marker for a binary disease variable. ROC analysis interprets the predictive power of a diagnostic test. A good diagnostic test is supported by a marker which is powerful in distinguishing between the two classes of the disease variable. Since many disease outcomes vary over time, 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 prior to time *t*; otherwise, $R_i(t)$ = 0. For a diagnostic marker *M*, both sensitivity and specificity are defined as a function of time *t*:

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

$$specificity(c,t) = P\left\{M \le c \mid R(t) = 0\ \right\}$$

A *ROC(t)* is a function of *t* at different cutoffs *c*. A time-dependent ROC curve is a plot of *1 – specificity(c, t)* vs. *sensitivity(c, t).* The area under the ROC curve (AUC) is used as an accuracy measure of the ROC curve. A higher prediction accuracy is evidenced by a larger *AUC(t) (8;9).*

We identified a 28-gene relapse signature from the training set (1). Five-year disease-free survival prediction accuracy of the 28-gene signature was 0.983 on the training set (1), 0.843 on one validation set from van't Veer *et al.* (3), and 0.764 on another validation set from Sorlie *et al.* (2). A Cox proportional hazards model (10) was built upon the signature and the risk score was used for constructing the time-dependent ROC curve on the training data (*Figure 2* A). *Figure 2* C shows the evolution of the AUC in the time cause on the training data. The horizontal dashed line indicates the AUC of a weak classifier (AUC = 0.50). The vertical dashed line indicates the 5-year cutoff. To validate the discriminatory power of our identified gene signature, two validation sets were used. From each validation set, we identified the genes that are common to our 28-gene signature. Eight genes were found in the data generated by Sorlie *et al.* (2), including one unknown gene. We used Unigene Cluster ID to search for the common genes in this data set, such that unknown genes without any gene names could be identified. Twenty-five genes were obtained from the data generated by van't Veer *et al.* (3), in which four genes were duplicated. Since no Unigene Cluster ID was available in this data set, we used gene names to identify overlapped genes, and found that four genes appeared twice. The time-dependent ROC curves based on these two validation data sets are demonstrated in *Figure 2B*; the AUC vs. time is shown in *Figure 2D*.

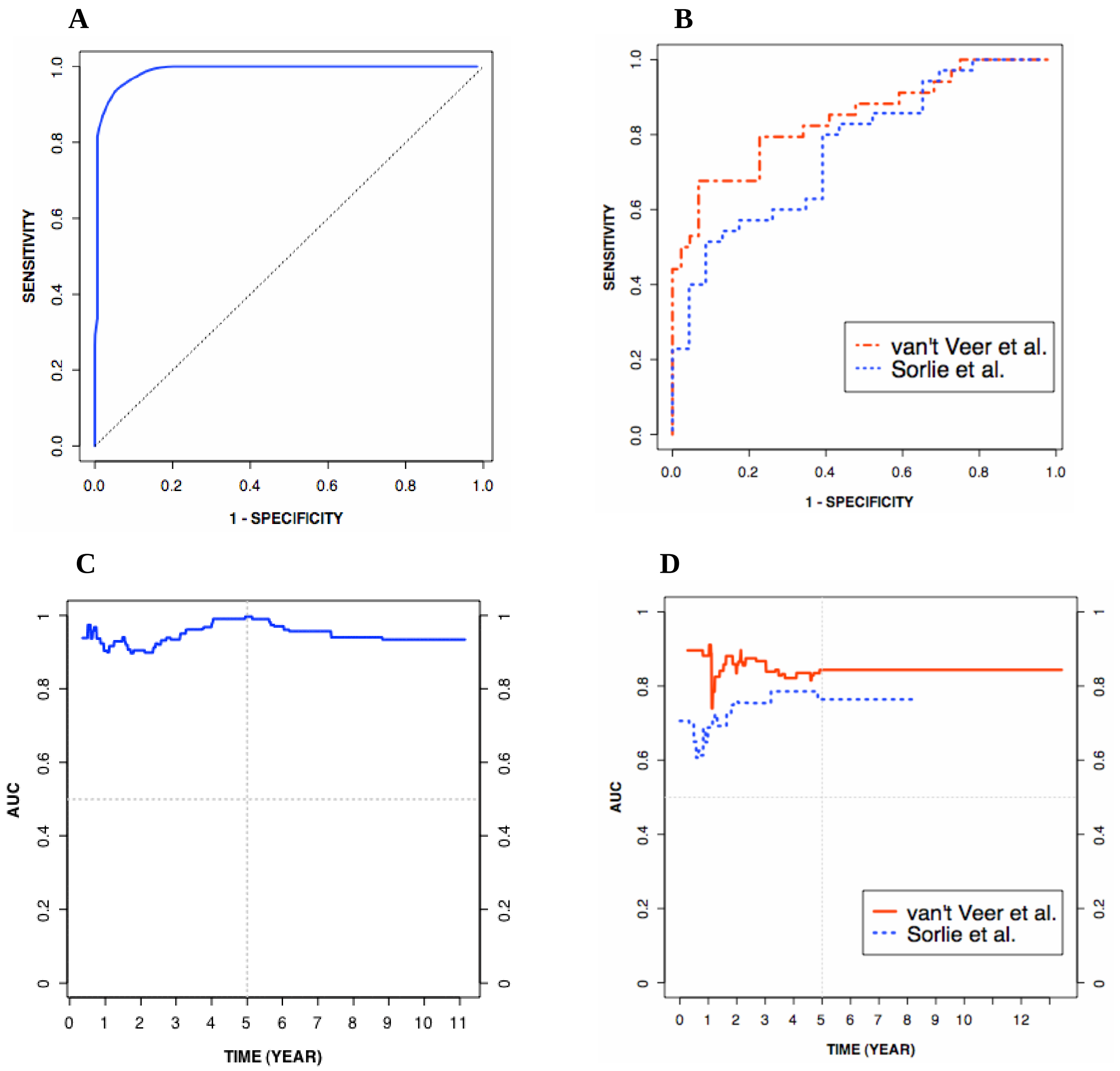

Figure S2. Time dependent ROC analyses of the 28-gene signature in disease-free survival prediction on three patient cohorts.

A. Time dependent ROC (*t* = 5 years) curve of the 28-gene signature on the training set generated by Sotiriou *et al.*(1) AUC = 0.983.

B. Time-dependent ROC (*t* = 5 years) curves of the 28-gene signature on two validation sets. AUC = 0.843 with the 25 overlapping genes on the set from van’t Veer *et al.*(3); AUC = 0.764 with 8 overlapping genes on the set from Sorlie *et al.* (2).

C. Area under the ROC curve in year 1 to year 11 on the training set (1).

D. Area under the ROC curve in year 1 to year 13 on the two validation data sets (2;3).

This 28-gene signature is also predictive of overall survival. Using time-dependent ROC analyses for overall survival time, the prediction accuracy of the 28-gene signature was 0.927 on the training set (1) (*Figure 3* A) and 0.808 on the validation set generated by Sorlie *et al.* (2) (*Figure 3* B).

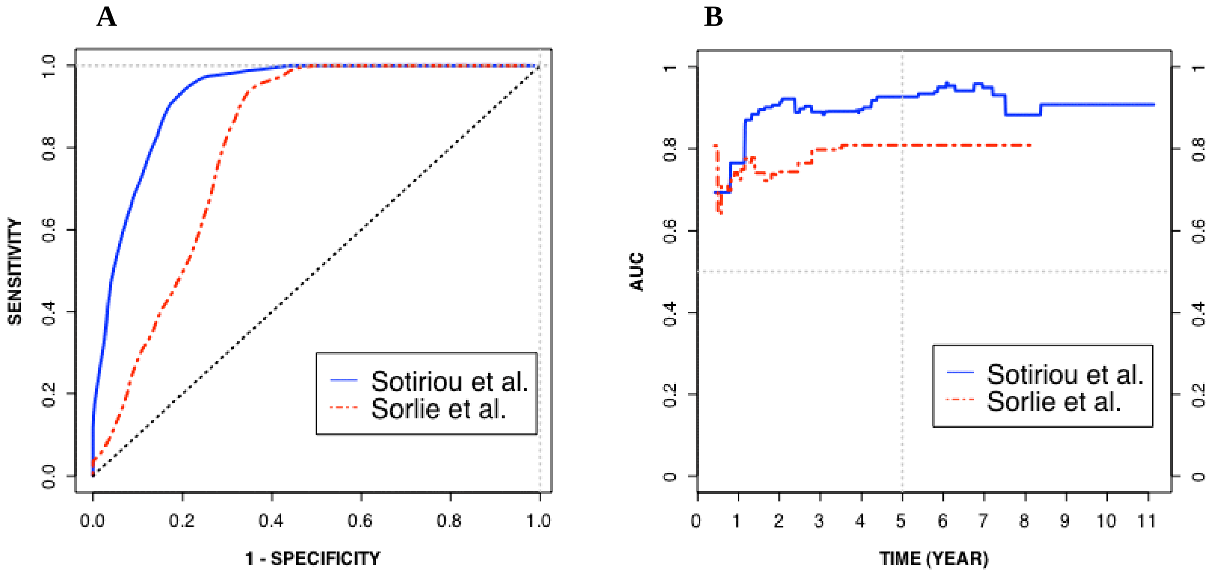

Figure S3. Time dependent ROC analyses of the 28-gene signature in the prediction of overall survival.
A. Time-dependent ROC curves at time = 5 years (Sotiriou et al: AUC = 0.927; Sorlie et al: AUC = 0.808).
B. the area under the ROC curve (AUC) at different time points.

6. Determine Risk Groups

To assess a breast cancer patient’s relapse and metastatic potential, risk scores were generated by using a Cox model of the 28-gene signature, independent of clinical-pathological parameters. A large value of the risk scores indicates a high risk of relapse/metastases, while a small value indicates a lower risk of breast cancer relapse. Our 28-gene signature obtained from the training set (1) was fitted into a Cox regression model as covariates. To avoid overfitting, we randomly split the data set into two subsets – one was used to define risk groups by fitting the model and obtaining the risk score cutoffs; the other subset was used to validate the cutoffs for defining the risk groups. The distribution of the risk scores from the training subset was used to divide the patients into three groups: high-risk, low-risk, and intermediate-risk. The cutoffs defined in the training subset were used to separate the patients in the test subset into high, low and intermediate risk groups.

The percentage of patients categorized into high, low, or intermediate risk group was 39%, 26%, and 35%, respectively. Table 5 displays the clinical characteristics of each risk group, including average relapse-free days, ER status, Her2/neu overexpression, nodal status, age, tumor size, and treatment received on the data from Sotiriou *et al.*(1). Same analysis was applied to the two validation sets. Table 6 summarizes the clinical characteristics of each risk group, including average metastases-free days, ER and PR status, age, tumor size, and tumor grade on the data from van't Veer *et al*. (3). Table 7 summarizes the clinical characteristics of each risk group, including average relapse-free days, ER status, age, and tumor grade on the data from Sorlie *et al.*(2). Kaplan-Meier analyses showed that disease-free survival was significantly different for each risk group in all three data sets ($p < 0.005$, log-rank test; Figure 4).

Table S5. Clinical characteristics of each risk group (Sotiriou *et al.*(1))

Risk Group	**Average RFS (days)**	**% of Age ≥ 50 yrs**	**# of Her-2\neu positive cases**	**% of Tumor Size > 2cm**	**% of Positive Nodal Status**	**% of Chemo**	**% of Hormone**	**% of ER +**
High	969	82%	6	82%	67%	38%	79%	54%
Intermediate	2407	73%	1	58%	50%	35%	85%	58%
Low	2781	65%	0	47%	41%	24%	74%	85%

Table S6. Clinical characteristics of each risk group (van't Veer *et al.* (3))

Risk Group	**% of Patients**	**Average RFS (days)**	**% of Age ≥ 50**	**% of tumor size > 2 cm**	**% of ER +**	**% of PR +**	**% of Tumor Grade 3**
High	27%	884	33%	67%	57%	38%	81%
Intermediate	33%	2284	19%	42%	96%	88%	77%
Low	40%	2988	32%	32%	77%	71%	42%

Table S7. Clinical characteristics of each risk group (Sorlie *et al.* (2))

Risk Group	**% of Patients**	**Average RFS (days)**	**% of Age ≥ 50**	**% of ER +**	**% of Tumor Grade 3**	**% of T3/T4 Tumors**
High	28%	553	50%	69%	81%	94%
Intermediate	32%	801	84%	89%	26%	89%
Low	40%	1376	70%	73%	32%	77%

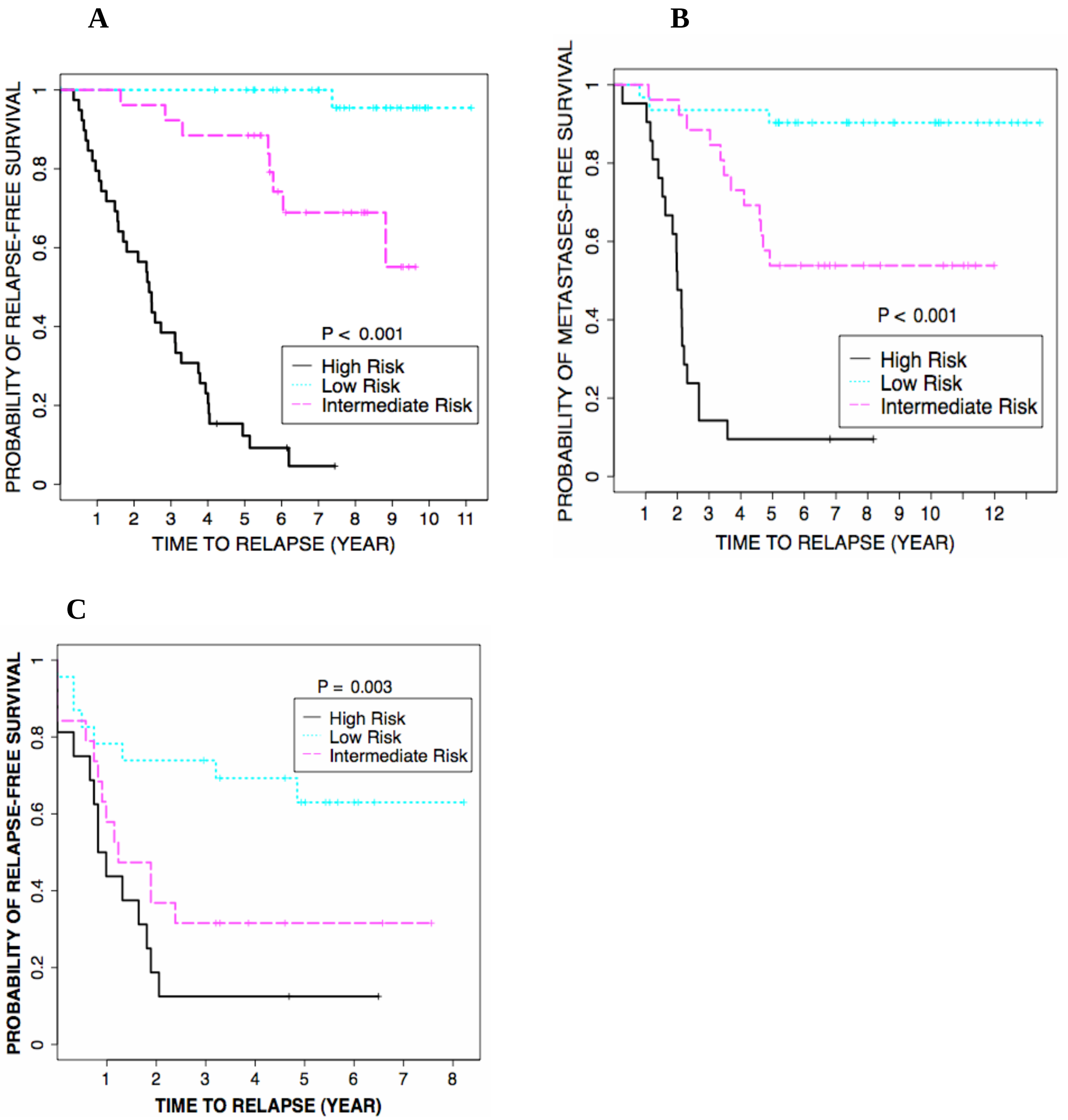

Figure S4. Kaplan-Meier analysis of disease-free survival of three risk groups in three patient cohorts. A. Kaplan-Meier analysis on data from Sotiriou *et al.* (1). B. Kaplan-Meier analysis on data from van't Veer *et al.*(3). C. Kaplan-Meier analysis on data from Sorlie *et al.*(2).

In this study, we also evaluated the association between the risk groups and the clinical-pathological parameters on three data sets (1-3) by using either Chi-square test or Fisher's exact test. Chi square test was used, if its assumptions were satisfied. Otherwise, Fisher's test was used. Table 8 reports the *P* values resulted from the tests. The results indicated that our identified 28-gene relapse/metastases signature was indicative of the clinical parameters including tumor size, grade, ER/PR status, and Her2/neu overexpression.

Table S8. The association between risk groups and clinical-pathological parameters in three patient cohorts.

	***P* Values**		
Risk Groups vs.	Sotiriou *et al.(1)*	van't Veer *et al.(3)*	Sorlie *et al.(2)*
Age [1] (<50 yrs or ≥ 50yrs)	0.243	0.458	0.095
Tumor size (<2 cm or >2cm)	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*	
Her2/neu	0.020*		

[1] The percentage of patients who were at least 50 years old was 74%, 28%, and 69% in the cohorts from Sotiriou *et al. (1),* van't Veer *et al. (3),* and Sorlie *et al. (2),* respectively**.**

To assess the therapeutic benefits for each risk group, average relapse-free survival days were compared for patients receiving adjuvant therapy in each group using the data from Sotiriou *et al.* (1). Specifically, therapeutic effects for patients receiving chemotherapy alone, hormonal therapy alone, or both chemo and hormone therapy were compared for each risk group. The observation in Table 9 is consistent with current clinical practice.

Table S9. Breast cancer therapeutic benefits assessment.

	Average RFS (days)		
Risk Group	Chemo Alone	Hormonal Alone	Chemo + Hormonal
High	613 (5 patients)	1005 (21 patients)	1048 (10 patients)
Intermediate	1478 (1 patient)	2496 (15 patients)	2262 (7 patients)
Low	3632 (4 patients)	2734 (20 patients)	2545 (5 patients)

References

1. 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.

2. 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.

3. '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.

4. Breiman L. Random Forests. *Machine Learning* 2001;45:5-32 .

5. Bioinformatics and Computational Biology Solutions, 1st ed. New York: Springer, 2005.

6. Breiman L. Bagging Predictors. *Machine Learning* 1996;24:123-40.

7. Diaz-Uriarte R, Alvarez dA. Gene selection and classification of microarray data using random forest. *BMC.Bioinformatics.* 2006;7:3.

8. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56:337-44.

9. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;61:92-105.

10. Cox D. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* 1972;34:187-220.