Imaging, Diagnosis, Prognosis

Clinical Cancer Research

Expression Signature Developed from a Complex Series of Mouse Models Accurately Predicts Human Breast Cancer Survival

Mei He[1], David P. Mangiameli[1,4], Stefan Kachala[5], Kent Hunter[2], John Gillespie[6], Xiaopeng Bian[3], H.-C. Jennifer Shen[1], and Steven K. Libutti[1,7]

Abstract

Purpose: The capability of microarray platform to interrogate thousands of genes has led to the development of molecular diagnostic tools for cancer patients. Although large-scale comparative studies on clinical samples are often limited by the access of human tissues, expression profiling databases of various human cancer types are publicly available for researchers. Given that mouse models have been instrumental to our current understanding of cancer progression, we aimed to test the hypothesis that novel gene signatures possessing predictability in clinical outcome can be derived by coupling genomic analyses in mouse models of cancer with publicly available human cancer data sets.

Experimental Design: We established a complex series of syngeneic metastatic animal models using a murine breast cancer cell line. Tumor RNA was hybridized on Affymetrix MouseGenome-430A2.0 GeneChips. With the use of Venn logic, gene signatures that represent metastatic competency were derived and tested against publicly available human breast and lung cancer data sets.

Results: Survival analyses showed that the spontaneous metastasis gene signature was significantly associated with metastasis-free and overall survival ($P < 0.0005$). Consequently, the six-gene model was determined and showed statistical predictability in predicting survival in breast cancer patients. In addition, the model was able to stratify poor from good prognosis for lung cancer patients in most data sets analyzed.

Conclusions: Together, our data support that novel gene signature derived from mouse models of cancer can be used for predicting human cancer outcome. Our approaches set precedence that similar strategies may be used to decipher novel gene signatures for clinical utility. *Clin Cancer Res; 16(1); 249–59.* ©*2010 AACR.*

Cancer is responsible for about one third of all mortalities in the United States, whereas metastatic disease is responsible for >90% of all cancer-related deaths (1). Subsequently, metastatic competency represents one of the most heavily investigated topics of modern medicine, science, and industry. Hanahan and Weinberg (2) have organized the plethora of cellular abnormalities into six basic competency traits that must be acquired for a malignancy to thrive: "self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis." These competencies are thought to be the product of alterations attained by the tumor early in the clinical timeline. Coupled with the increasing heterogeneity of the cell population of the tumor, multiple phenotypes may arise with varying levels and tendencies of metastatic competency (3).

Animal models have been important to our current understanding of malignant and metastatic progression (4). The use of different models and techniques, such as *in vivo* passaging for phenotype purification, transgenic animals for specific molecular manipulation, and *in vivo* and *ex vivo* models for screening of cancer therapies, has led to invaluable functional insights. Importantly, these animal model systems have allowed us to develop useful dogmatic philosophies about the causes of malignant transformation and novel strategies to further investigate malignant behavior (5, 6).

Authors' Affiliations: [1]Tumor Angiogenesis Section, Surgery Branch; [2]Metastasis Susceptibility Section, Laboratory of Cancer Biology and Genetics; [3]Center for Biomedical Informatics and Information Technology, National Cancer Institute, NIH, Bethesda, Maryland; [4]Division of Surgical Oncology, Department of Surgery, College of Physicians and Surgeons, Columbia University; [5]Department of Surgery, New York-Presbyterian Hospital/Weill Cornell Medical Center, New York, New York; [6]SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, Maryland; and [7]Department of Surgery, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York

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

Corresponding Author: Steven K. Libutti, Department of Surgery, Montefiore Medical Center, Albert Einstein College of Medicine, 4th Floor, Greene Medical Arts Pavilion, 3400 Bainbridge Avenue, Bronx, NY 10467-2490. Phone: 718-920-4231; Fax: 718-882-1279; E-mail: slibutti@montefiore.org.

doi: 10.1158/1078-0432.CCR-09-1602

Translational Relevance

In this study, we devised an analytic strategy that led to the development of a novel gene signature for predicting cancer patient outcome. Our study is significant in that it minimizes the need to have direct access to human tissue samples while maximizing the utility of publicly available clinical data sets and mouse models of cancer for generating novel genomic assays for clinical purpose. Our innovated approach that combined mouse models of cancer with publicly available clinical data set sets precedent that similar strategy can be applied to other cancer types. Upon further validation analyses and prospective studies, the novel gene signature described here could potentially be coupled with other genomic-based assays to assist physicians in the management of cancer patients.

Another valuable and recent scientific success has been the development and use of high throughput assays such as microarray expression analysis. Molecular profiling with this technology has led to derivation of gene signatures for various cancer types (7–9) and gained utility in the management of selected cancer patients. For instance, two genomic based assays currently serve as surrogate indicators to determine which early stage breast cancer patients are likely to benefit from adjuvant chemotherapy (10).

While various efforts of banking clinical samples are underway, many researchers are still limited by the access and cost of obtaining sufficient amount of human tissues for experimental purpose. However, a wealth amount of expression profiling data sets on different human tumor histology is publicly available. Thus, this study aimed to test the hypothesis that novel gene signatures possessing predictability of clinical outcome can be derived by coupling genomic analyses in mouse models of cancer with publicly available clinical data sets. Specifically, we established a complex series of metastatic mouse models using a murine breast cancer cell line. Using microarray expression analysis, Venn logic, and clinical data sets, the six-gene signature was derived and showed accuracy in predicting breast cancer patient survival. We believe that this six-gene model represents a general metastatic competency gene signature because this six-gene model can stratify prognosis outcomes in lung cancer patient cohorts as well. Together, we showed that novel gene signature for predicting cancer patient outcome can be developed by coupling properly designed mouse model systems with publicly available clinical data sets. In addition, our study is significant in that it minimizes the need to have direct access to human tissue samples while maximizing the utility of publicly available clinical data sets for generating novel genomic assays for clinical purpose.

Materials and Methods

Animal models and tissue procurement

All animal studies were in accord with the NIH Animal Care and Use Committee Guidelines.

Embolic liver and lung metastatic model

Liver metastases splenic vein model (LvMsv). Murine breast adenocarcinoma (4T1) cells (American Type Culture Collection: The Global Bioresource Center) were harvested from cell culture flasks, washed three times in HBSS, and adjusted to a final concentration of 1×10^7 cells/mL. Cell preparations were kept on ice until injection. BALB/c mice were anesthetized with isoflurane and prepared for surgery under sterile conditions. Animals were positioned in right lateral recumbency, shaved, and wiped with 70% ethanol. A left subcostal incision, ~10 mm long, was made, and the peritoneum was opened. The spleen was exposed and gently retracted; the gastrosplenic ligament and short gastric vessels were identified and divided, leading to complete mobility of the spleen on its hilar pedicle. The spleen was then extracorporealized and positioned on sterile saline soaked gauze. Next, cell suspension (200 μL) was slowly injected into the upper splenic pole, using a 30-gauge needle (Becton Dickinson). After injection, slight pressure was applied to spleen to achieve hemostasis and minimize extrasplenic seeding. Five minutes were elapsed to allow portal vein embolization. Splenectomy vis-à-vis application of a medium Ligaclip (Ethicon Endo-Surgery, Inc.) to splenic vessels and sharp excision of the organ followed. The abdominal cavity was then closed en masse with 9-mm wound autoclips (Roboz Surgical). Animals were monitored and sacrificed when they became moribund. Livers were examined with 2× surgical loupes, and hepatic metastases were immediately resected, snap frozen in liquid nitrogen, and ultimately stored at -80°C.

Lung metastases tail vein model. 4T1 cells were prepared as LvMsv model, adjusted to 5×10^6 cells/mL, and kept on ice until injection. Tail veins of female BALB/c mice were cannulated with a 27-gauge needle, and the animals were administered 50 μL of cell suspension. After 14 d, they were sacrificed, and the tracheobronchopulmonary tree was resected and insufflated with PBS. The lung metastases were resected under surgical loupes, snap frozen in liquid nitrogen, and stored at -80°C.

Spontaneous liver and lung metastases model

Tumor cell suspension (100 μL; 1×10^7 cells/mL) was prepared as LvMsv model and then injected into the left cephalad mammary gland of BALB/c mice. After 14 d, the resultant orthotopic tumors were excised under sterile conditions, and the tumor was immediately snap frozen in liquid nitrogen and stored at -80°C. The wound was closed with autoclips. After an additional 14 d, animals were sacrificed, and the spontaneous liver and lung metastases were procured as described earlier.

Microarray and statistical analysis

To minimize individual variation, tumor samples were used from three individual mice, from each metastatic

animal model. Twenty cryostat sections (10 μm) were cut in all samples under RNase free condition and stored at -80°C. Sections were stained with hematoxylin and eosin by pathologist (J. Gillespie), and only tumor area was microdissected. Total RNA was immediately isolated using the PicoPure RNA Isolation Kit (Arcturus). Total RNA (30 ng) from each sample was used in the reverse transcription of two consecutive rounds of linear amplification, first using the MessageAmp II aRNA Amplification Kit (Ambion), followed by biotin labeling using the MessageAmp II-Biotin Enhanced Kit (Ambion). RNA concentrations were measured by NanoDrop ND-1000 (NanoDrop). The quality of RNA preparations was assessed with Bioanalyzer RNA 6000 NanoLabChip Kit (Agilent Technology). All samples included in this study had a 28S/18S rRNA ratio of >1.5, with an average of 2.0. Each biotinylated cRNAs (20 μg) was fragmented and hybridized to an Affymetrix Mouse Genome 430A2.0 Array GeneChip (Affymetrix). Arrays were scanned using standard Affymetrix protocols. Image analysis and probe quantification was done with the Affymetrix GeneChip Operating Software, which produced raw probe intensity data.

Raw intensity profiles were analyzed using Partek Genomics Suite Software (Partek, Inc.). Robust microarray analysis was applied for normalization. Significantly regulated genes were defined as those genes from one experimental group whose expression was statistically and significantly different from another group by virtue of multiway ANOVA. Resulted ratios were transformed into log 2 values and used as expression levels for genes in metastatic gene signatures. Genes included in the lists were further selected with a false discovery rate of <10% using Partek Genomics Suite Software. Each probe set was treated as a separate gene, whereby averaging of the triplicate led to the defined data of the respective gene. Validation was through Cox's proportional hazard regression using estimated hazard ratios and clinicopathologic data. Kaplan-Meier survival analysis was applied to generate predictive values for gene signatures.

Clustering

Hierarchical cluster analysis was carried out with Stanford University Cluster Software (11). The average linkage and uncentered Pearson correlation distance measure were used as the similarity metric for clustering of genes and arrays. The clusters were visualized using Tree View.[8]

Application of gene signatures to public data sets

To compare expression data from the mouse and human data sets, a common correspondence has to be made between probes on the mouse arrays with probes on the human arrays. To map our mouse signature to public data sets of human arrays, we first matched mouse signature gene symbols to human gene symbols by using a mouse-human homology gene list provided by Microarray Data Base (Center for Cancer Research, National Cancer Institute, NIH). We then used the gene symbol identifier to match genes represented in different microarray data sets. For cDNA microarrays, genes with fluorescent hybridization signals at least 1.5-fold greater than the local background fluorescent signal in the reference channel (Cy3) were considered adequately measured and were selected for further analyses. For Affymetrix microarray data, signal intensity values were *z*-transformed into ratios, and genes with technically adequate measurements obtained from at least 90% of the samples in a given data set were selected for analysis. Gene value was generated by the averaging of each probe set within a given experimental group. The patterns of expression in published data sets were subsequently analyzed according to our gene signature. Averaged linkage clustering was done using Cluster Software. After application of each signature, the sample data from each public data set was segregated into two classes based on the first bifurcation of its hierarchical dendrogram. This most proximal bifurcation represents the most fundamental surrogate of fidelity of the samples profile with the tested signature. Survival analysis was done on each class that resulted from the grouping.

Published data sets

Data from Gene Expression Omnibus database.[9]

Breast cancer data sets

van de Vijver data set. This was a validation study on a predictive expression signature, which involved 295 young patients with early stage breast cancer, of which 151 were lymph node negative, 226 were estrogen receptor positive, and 110 had received adjuvant chemotherapy (12, 13).

GSE4922 data set. This was a derivation study for the molecular profiling of the histologic grading of breast cancer; the patients used are referred to as the Uppsala cohort. Of the 316 patients in the cohort, 249 were used to derive the molecular profile, of which 211 of them were estrogen receptor positive, 81 were lymph node positive, and 58 showed p53 mutation. Eighty-six patients that overlapped with the GSE2990 data set were excluded, leaving 163 patients in this analysis. Data were originally published by Bergh et al. (14) and reinvestigated by Ivshina et al. (13).

GSE2034 data set. This was a derivation and validation analysis of a gene signature for the prediction of breast cancer patient outcomes. It consisted of 286 lymph node–negative breast cancer patients who never received adjuvant chemotherapy and of which 209 were estrogen receptor positive. Data were published by Wang et al. (15).

GSE1456 data set. This study was a derivation and validation analysis of a predictive gene signature for the outcomes of women with breast cancer. It involved 159 patients with breast cancer, of which 82% were estrogen receptor positive, 62% were lymph node negative, and 79% were treated with adjuvant chemotherapy. Data were published by Pawitan et al. (16).

[8] Available from: http://rana.lbl.gov/EisenSoftware.htm.

[9] Available from: http://www.ncbi.nlm.nih.gov/geo with accession code.

GSE2990 data set. This study was a derivation and validation analysis of a correlative gene signature aimed at histologic grade. It involved 189 women with breast cancer, of which 160 were lymph node negative. Sixty-four estrogen receptor–positive samples were used to derive a signature that effectively differentiates outcomes and grade. Data were published by Sotiriou et al. (17).

GSE7390 data set. This study was a multicenter validation trial to evaluate the clinical utility of a gene signature for the management of early node negative breast cancer. Their analysis involved 198 patients, of which we excluded 22 because of overlap with the GSE2990 data set. Data were published by Desmedt et al. (18).

Lung cancer data sets

GSE4573 data set. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It consisted of 130 patients with squamous cell carcinomas from all stages. Data were published by Raponi et al. (19).

GSE11117 data set. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It involved 41 chemotherapy-naïve non–small cell lung carcinoma (NSCLC) patients. Data were published by Baty et al.[10]

Data sets published by National Cancer Institute director's challenge consortium for the molecular classification of lung adenocarcinoma and Shedden et al. (8).

Moffitt Cancer Center data set. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It involved 79 patients with NSCLC of all stages.

University of Michigan Cancer Center data set. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It involved 177 patients with NSCLC of all stages.

The Dana-Farber Cancer Institute data set. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It involved 82 patients with NSCLC of all stages.

Memorial Sloan-Kettering Cancer Center. This was a derivation and validation analysis of a gene signature for the prediction of lung cancer patient outcomes. It involved 104 patients with NSCLC of all stages.

Survival analysis

Kaplan-Meier estimates and log-rank testing were used to construct survival curves. Statistical significance was evaluated using Cox regression analysis of hazard ratios. Overall survival was defined as the time interval between the first dates of any form of treatment and the last follow-up date or date of death; patients alive at the date of last follow-up were censored at that date. Metastasis-free survival was defined as the interval from the first treatment day to the day of the diagnosis of distant metastases. All other patients were censored on their date of last follow-up, including alive without disease, alive with locoregional recurrence, alive with a second primary cancer, and death from an alternate cause. The relapse-free survival was defined as the time interval between the date of breast surgery and the date of a diagnosed relapse or last follow-up. Women who developed contralateral breast cancer were censored. The data reported in this study were based on the 10-y survival in the van de Vijver, GSE4922, GSE2990, and Moffitt Cancer Center data sets; 5-y survival in the GSE2034, GSE1456, GSE4573, University of Michigan Cancer Center, and The Dana-Farber Cancer Institute data sets; 12-y survival in the GSE7390 data set; and 4-y survival in the GSE11117 and Memorial Sloan-Kettering Cancer Center data sets. Patients with missing survival data or those that were reported to have zero follow-up time were excluded from survival analyses. All reported *P* values are two sided. Multivariate analysis by Cox proportional hazard regression and all survival statistics were done in Partek Genomics Suite.

Selecting the six-gene model from the spontaneous metastasis gene signature (SpMGS)

To further evaluate the prognostic value of each gene within the signatures, intercohort multivariate Cox proportional hazards analysis of each signature gene was done in three breast cancer data sets. Genes significantly correlated with patient outcomes ($P < 0.05$) were determined for each data sets. Only genes with $P < 0.05$ and present in at least one of three data sets were selected. Among the three data sets, a total of 17 unique genes were derived from the original 79 SpMGS. Twelve of these had hazard ratio of >1, of which 6 genes were predictive in all three data sets. This served as the logic and derivation of the new six-gene model. Survival analysis was done on the three original public breast cancer data sets (van de Vijver, GSE4922, and GSE2034) using the six-gene model. In addition, the six-gene model was tested against three additional independent public breast cancer data sets (GSE1456, GSE2990, and GSE7390).

Results

Metastases gene signatures derived from mouse models

Using a murine breast cancer cell line, a complex series of metastatic mouse models were established as shown in Fig. 1A. Spontaneous metastasis to the liver and lungs developed after resection of the primary breast tumors, whereas embolic metastasis was derived from direct inoculation of tumor cells through the systemic or portal venous system, respectively. Gene expression profiling was done on the six different tumor types collected (Fig. 1A). Based on statistical analyses, we identified genes that were significantly and differentially expressed between the metastatic tumor types (spontaneous and embolic) and primary tumor. As shown in Fig. 1B, 194 unique genes (corresponding

[10] Available from: http://www.ncbi.nlm.nih.gov/geo.

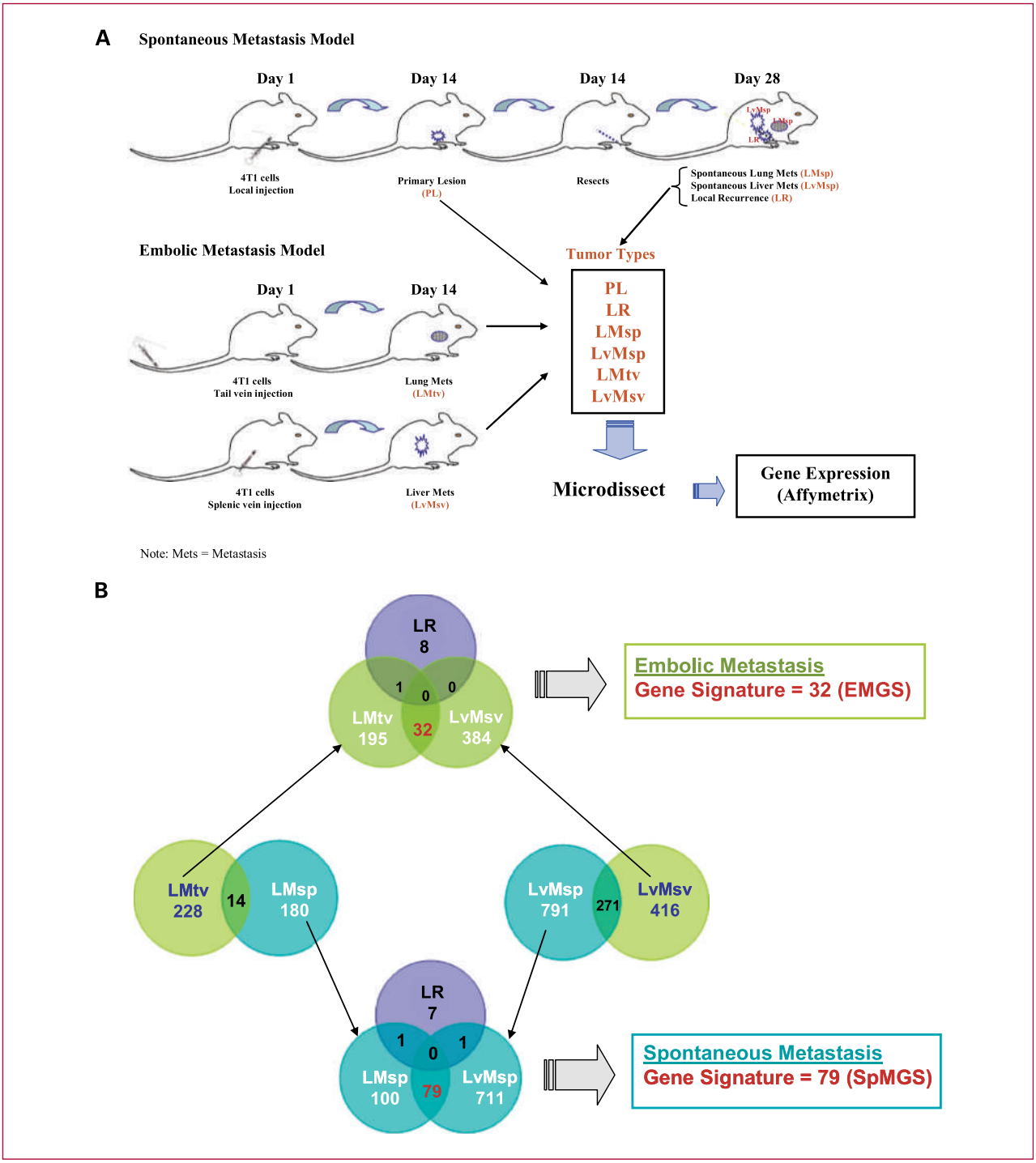

Fig. 1. *A,* establishment of spontaneous (*top*) and embolic (*bottom*) metastatic animal models. Murine breast adenocarcinoma 4T1 cells were used to generate the models, and six different tumor types were procured as indicated. After microdissection, gene expressions were analyzed using Affymetrix microarrays. *B,* generation of EMGS and SpMGS using Venn diagrams. Statistical analyses identified genes that were significantly and differentially expressed between the metastatic tumor types and primary tumors. Further experimental details can be found in the Materials and Methods.

to 226 gene probe sets) associated with spontaneous lung metastasis; 1,062 unique genes (corresponding to 1,203 gene probe sets) associated with spontaneous liver metastasis; 242 unique genes (corresponding to 271 gene probes sets) associated with embolic lung metastasis; 687 unique genes associated (corresponding to 788 gene probe sets) with embolic liver metastasis (LvMsv); and only 9 unique genes associated with local recurrence. The embolic lesions

allowed us to control for the ambient changes in gene expression associated with tumor growth in an alternate parenchyma, which were present despite the earlier steps needed to gain metastatic competency. Using Venn logic, we excluded the ambient changes and targeted the alternate expression patterns as a source for predictive power. Thus, we generated a SpMGS containing 79 genes and an embolic metastasis gene signature (EMGS) containing 32 genes.

Expression of gene signature from mouse model in human breast cancer

To evaluate the prognostic value of the metastatic gene signatures and to determine which of the 79 SpMGS genes were more predictive with metastasis-free survival, we used three publicly available data sets of human breast cancer expression data and correlating clinical outcomes. These included the van de Vijver, GSE4922, and GSE2034 gene sets.

To facilitate visualization and identify subgroups of patients that expressed the SpMGS, we organized the gene expression patterns and samples using hierarchical clustering. We segregated patients into two classes in which patients in class 2 exhibited the metastatic signature, whereas those in class 1 did not (Supplementary Fig. S1). To correlate clinical outcome, we calculated the probability of remaining free of distant metastases and overall survival, given the genetic expression class for signature.

van de Vijver data set. Kaplan-Meier curves showed a significant association between the SpMGS and metastasis-free and overall survival ($P < 0.0005$). This analysis indicates that the risk for metastasis was significantly higher for patients in class 2 than class 1. Class 1 had better metastases-free and overall survival (85% and 94% at 5 years, and 76% and 84% at 10 years, respectively) compared with class 2 (64% and 77% at 5 years, and 51% and 63% at 10 years, respectively; Fig. 2A). The univariate hazard ratio was 0.36 ($P < 0.00003$) for metastasis and 0.33 ($P = 0.00014$) for death. Multivariable proportional hazards analysis confirmed that the SpMGS classification was a significant independent factor in predicting disease outcome ($P = 0.003$). The SpMGS was a sensitive predictor of distant metastases, with hazard ratio of 0.46 (Table 1).

A univariate Cox proportional hazards model was used to evaluate the association of our signature with clinical outcome in each category, stratified for multiple clinical parameters. As summarized in Table 2, the prognostic profile based on SpMGS was accurate in predicting the outcome of disease. Comparing patients in class 1 with those in class 2 revealed a hazard ratio for distant metastases of 0.43 for lymph node–negative patients and 0.28 for lymph node–positive patients ($P < 0.05$ for both). Similarly, the prognostic profile was strongly associated with disease outcome in groups of patients with tumor diameter ≤20 mm (hazard ratio, 0.33; $P = 0.002$) and tumor diameter >20 mm (hazard ratio, 0.45; $P = 0.02$), as well as in patients with age ≤45 years (hazard ratio, 0.30; $P = 0.00007$) and age >45 years (hazard ratio, 0.46; $P = 0.05$). Furthermore, the SpMGS could be used to stratify tumors of well and intermediate differentiation into good and poor prognostic subcategories (hazard ratio, 0.24 and 0.26, respectively; $P < 0.05$) but was less correlative with the stratification of poorly differentiated lesions ($P = 0.67$). The clinical corollary was significant for tumors that were estrogen receptor positive (hazard ratio, 0.36; $P < 0.05$) but not for those that were estrogen receptor negative. This analysis also showed that SpMGS was a strong predictor of improved outcomes in the group of patients who did or did not receive chemotherapy (hazard ratio, 0.25 and 0.43, respectively; $P < 0.05$).

GSE4922 and GSE2034 data sets. A similar analysis was done on GSE4922 and GSE2034 data sets to predict overall survival in GSE4922 data set and relapse-free survival in GSE2034 data set. The survival analysis showed that the risk for death or metastasis in both data sets was significantly higher among patients with an expression profile associated with SpMGS class 2 [hazard ratio, 0.55 ($P = 0.019$) and 0.47 ($P = 0.0013$), respectively; Fig. 2B and C].

It should also be noted that when similar analysis was done using the 32-gene EMGS on the three data sets, the predictive outcomes were either statistically insignificant or not as powerful as the SpMGS (Supplementary Fig. S2).

To determine if SpMGS is unique from previously published work, we cross-referenced our SpMGS to other human breast cancer gene profiles. SpMGS has only one gene (*PTDSS1*) in common with the 70-gene signature by van't Veer et al. (20), one gene (*FOS*) in common with the 264-gene signature by Ivshina et al. (13), and one gene (*TOB2*) in common with the 186-gene signature Liu et al. (21). Together, these results indicated that our mouse-derived SpMGS was an independent new expression profile that had prognostic value when applied to human disease.

Evaluation of gene signature and a six-gene model

To further evaluate the prognostic value of each gene within the signatures, we did multivariate Cox proportional hazards analysis of each signature gene in different data sets based on clinical information. In SpMGS, 17 of 79 genes were present in at least one of the three breast cancer data sets and had significant sensitivity in their ability to assign prognosis ($P < 0.05$). More importantly, 12 of these 17 (70.6%) had a hazard ratio of >1 (Table 3), indicating that upregulation of those genes will lead to poor prognosis. In contrast, 16 of 32 genes from EMGS present in all three data sets had a significant association with prognosis profile ($P < 0.05$), noting that only 4 of these (25%) had a hazard ratio of >1.

The genes with high hazard ratios were considered high yield components of the predictive model. As such, of the 12 genes (hazard ratio, >1) in SpMGS subgroup, six genes that were present in all three data sets were selected. This six-gene model consists of the following genes: *Abcf1*, *Coro1c*, *Dpp3*, *Preb*, *Ptdss1*, and *Ube3a* (Supplementary Table S1). We next tested six-gene model for its predictive power as a stand-alone expression signature. Survival analysis on the original three public data sets indicated that the six-gene model is powerful in predicting patient outcome (Fig. 3, *top*). As expected, similar to the 79 SpMGS,

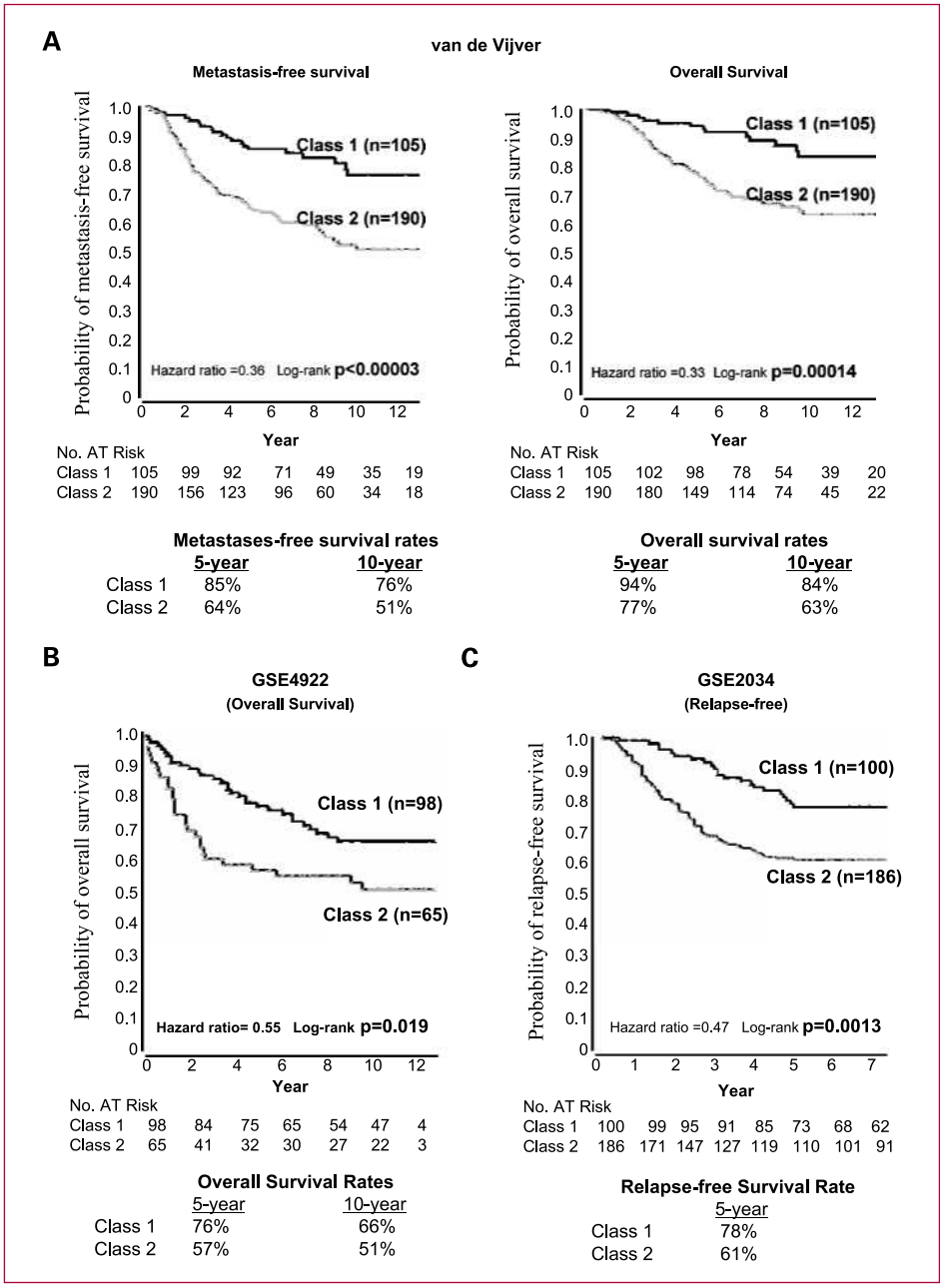

Fig. 2. Kaplan-Meier analysis of the probability that patients would: remain free of metastases and overall survival in van de Vijver data set (*A*), overall survival in GSE4922 data set (*B*), and relapse-free survival in GSE2034 data set (*C*). Patients who exhibited the metastatic signature (SpMGS) were assigned class 2 (*grey*), whereas those who did not were assigned class 1 (*black*). Hazard ratios and *P*s are within each graph. The 5-y and 10-y survival rates are at the bottom for each data set.

the six-gene model also predicted survival independent of known clinical variables based on multivariable proportional hazards analysis using the van de Vijver data set (Supplementary Table S2). Because it is likely that gene expression profiles will affect future clinical decision making, it has been emphasized that predictive models should be validated independent of its training data sets. Therefore, the six-gene model was tested against three additional independent publicly available breast cancer data sets (Fig. 3, *bottom*). Data revealed a significant association between the six-gene model and relapse-free survival in GSE1456 and GSE2990 data sets and overall survival in GSE7390 data set ($P = 0.0009$, $P = 0.03$, and $P = 0.018$ by log-rank test, respectively). Notably, in all data sets tested, patients with poor prognosis correlated largely with upregulation of the six genes based on cluster analysis.

Table 1. Multivariable proportional hazards analysis of the risk for distant metastasis as a first event in van de Vijver data set based on SpMGS

	HR	*P*
SpMGS	0.46	0.003
Primary tumor size (≤2 vs >2 cm)	0.62	0.03
Node (negative vs positive)	0.79	0.45
Age (<45 vs ≥45 y)	2.05	0.0009
Chemo (no vs yes)	1.54	0.17
ER (negative vs positive)	1.1	0.69
Differentiation		
Intermediate vs well	2.15	0.03
Poor vs well	2.8	0.004

Abbreviations: Chemo, chemotherapy exposure; ER, estrogen receptor; HR, hazard ratio.

Expression of six-gene model in human lung cancer

Based on our experimental design, we hypothesized that the six-gene model represents a general metastatic competency signature. Thus, we investigated whether the six-gene model plays a role in predicting prognosis outcome in cancer types other than breast cancer. Subsequently, we applied the six-gene model to six independent publicly available human lung cancer data sets to predict the overall survival. As summarized in Supplementary Table S3, the six-gene model was able to stratify poor from good prognosis with statistical significance in GSE4573 and Moffitt Cancer Center data sets ($P = 0.04$ and $P = 0.03$, respectively). Although the predictions of other data sets [GSE11117, University of Michigan Cancer Center, The Dana-Farber Cancer Institute, and Memorial Sloan-Kettering Cancer Center] were not statistically significant, they trended toward poor prognosis ($P = 0.09$, $P = 0.08$, $P = 0.07$, and $P = 0.09$, respectively) and were well separated by Kaplan-Meier curves (Supplementary Fig. S3).

Table 2. Univariate Cox proportional hazard model: class 1 versus class 2 hazard ratio for metastasis-free survival according to SpMGS

Clinical patients	HR	*P*	Total patients
Node positive	0.28	0.0009	144
Node negative	0.43	0.006	151
Tumor size ≤ 2 cm	0.33	0.002	150
Tumor size > 2 cm	0.45	0.02	140
Age ≤ 45 y	0.3	0.00007	166
Age > 45 y	0.46	0.05	129
Chemo			
Yes	0.25	0.002	110
No	0.43	0.003	185
ER positive	0.36	0.0003	226
ER negative	0.75	0.63	69
Differentiation			
Poor	0.87	0.67	119
Intermediate	0.24	0.0008	101
Well	0.26	0.03	75

NOTE: This analysis included data of the 295 breast cancer patients in van de Vijver data set, with the prognostic role of the metastases signatures tested within each patient category.

Table 3. Cox regression analysis; the genes had significant sensitivity in predicting favorable or poor prognosis ($P < 0.05$) in three data sets (van de Vijver, GSE4922, and GSE2034 data sets)

SpMGS			EMGS		
Symbol	HR (gene)	*P* (gene)	Symbol	HR (gene)	*P* (gene)
ABCF1	**2.60**	<0.001	**GNAI1**	**2.30**	<0.001
PREB	**2.05**	0.007	**HEPH**	**1.85**	0.012
PAPOLA	**2.04**	0.013	**C9orf58**	**1.43**	0.031
PTDSS1	**2.00**	<0.001	**TGFB1I1**	**1.35**	0.009
DOCK7	**1.87**	<0.001	DPEP1	0.83	0.032
HSPA9A	**1.79**	0.023	FOLR2	0.82	0.030
CORO1C	**1.71**	0.002	DSP	0.82	0.049
DPP3*	**1.63**	0.005	TMEM30B	0.81	0.048
ANAPC5	**1.29**	0.009	LUM	0.78	0.042
FBXW11	**1.26**	0.042	KLF15	0.77	0.018
UBE3A†	**1.24**	0.046	TSC22D3	0.75	0.004
ATP6V1C1	**1.23**	0.031	ATP1B1	0.73	0.003
HSPC117	0.80	0.018	ELN	0.69	0.006
XBP1	0.68	<0.001	BHLHB5	0.67	0.015
FOS	0.66	0.013	CXCL12	0.64	<0.001
TOB2	0.47	0.050	SPARCL1‡	0.57	<0.001
HCRT	0.43	0.046			

NOTE: Genes with hazard ratio of >1 are in boldface.
*Indicated that hazard ratio and *P* value were 1.52 and 0.01, respectively, in other data set.
†Indicated that hazard ratio and *P* value were 0.64 and 0.047, respectively, in other data set.
‡Indicated that hazard ratio and *P* value were 0.7 and 0.003, respectively, and 0.7 and 0.006, respectively, in other data sets.

Discussion

The malignant process surmounts several fairly distinct hurdles in its ultimate heterotopic progression (22). This process, so contra-aligned with host survival, led us to focus on why such an endowment is granted to these abnormal cells. We surmised, as have others, that unique

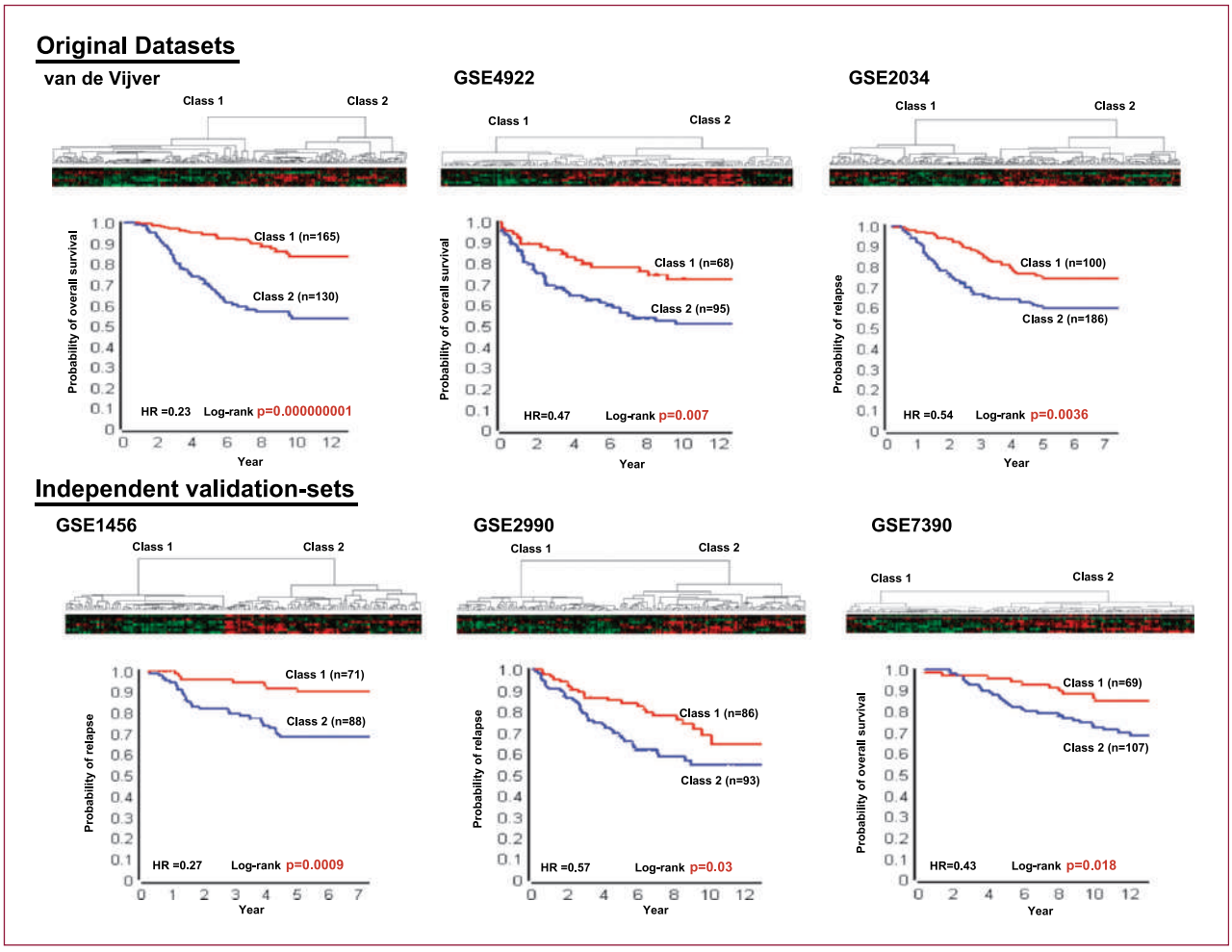

Fig. 3. Survival analysis based on six-gene model in six independent human breast cancer data sets. Three original data sets (*top*) and three additional independent validation sets (*bottom*) were used to validate the predictive power of the six-gene model. Class 2 (*blue*) included patients who exhibited the six-gene signature, whereas class 1 (*red*) included patients who did not. The hazard ratios and *P*s are within each graph.

genetic aberrancies cause, encourage, or certainly allow this to occur (23). Specific tissue tropisms are additionally confounding because it is evident that there is a differential between differing tumor histology and their resultant metastatic profiles (24). If early metastatic competence occurs in the setting of vast cellular heterogeneity, would our signature stay durably accurate within and across patients? In devising a model that accurately identifies the genetic perturbations responsible for metastases, we felt that looking at the differential expression between the primary and metastatic lesions was not enough. Breast cancer growing in lung tissue should have genetic expression alterations despite how it arrived there. This ambient organ-imposed expression alteration confounds a straightforward approach toward detecting metastatic competency genes. We deduced using Venn logic that, by subtracting the ambient gene profile from the primary and spontaneously metastatic tumor gene profiles, we could derive the constitutive metastatic competency genes found in the spontaneously metastasizing cancer. Embolic lung and liver mouse models served to provide the respective ambient gene profiles (EMGS). Incorporating multiple tropisms (lung and liver) allowed us to have internally generated controls for genetic interpretive quality assessment. In addition, it allowed us to categorize gene sets into tropism-specific metastatic competency genes if they were unique to specific organ tropisms or general metastatic competency genes if they were present in both tropisms. The SpMGS represents the theoretical general metastatic competency genes.

The SpMGS is composed of 79 unique genes, given our stringencies. When profiled against publicly available human breast cancer data sets, this signature significantly correlates with nodal status, tumor size, age, response to chemotherapy administration, estrogen receptor positivity, and favorable histologic differentiation. In addition, the signature was significantly predictive of patient survival outcomes across three independent data sets. The hazard

ratio shown in Table 3 further showed that the SpMGS is composed of genes with higher predictive yield than the EMGS. The consistency of predictive power across unrelated patient cohorts underscores and validates the accuracy of the SpMGS, as well as the approach toward its derivation.

With the intention of further amplifying the clinical yield, we queried the SpMGS in association with the public data sets and culled six genes that contributed the most to the predictive ability of the signature. This six-gene model was not only more portable than the SpMGS but was also highly predictive of survival outcomes when tested against three additional independent data sets of human breast cancer. We further evaluated the applicability of this six-gene model to lung cancer patients and showed predictive value in the survival analysis. This was not as powerful or durable as for breast cancer but still showed a significant clear disparity in those patients who did well and those who had poorer outcomes. However, the applicability of this six-gene model to other tumor histology remained to be tested.

Genomic assays have proven extremely important to the clinical management of early breast cancer patients. Two commercially available assays have allowed physicians to identify patients who are at low risk for recurrence and subsequently may forego morbid adjuvant chemotherapy (12, 20). Our six-gene model offers a similar utility, although it is more portable and perhaps more applicable to a wider cancer patient population. Because of its portability, it could conceivably be transformed into a hospital-based assay, which would presumably lower the cost of currently available extramural expensive assays. Despite the promising performance of the six-gene model shown in this study, comparative analysis with existing genomic assays and prospective clinical validation will be crucial to show the clinical potential of the six-gene model.

In addition to the potential clinical benefit of the six-gene model, a significant finding of this study is that gene signatures derived from mouse models can be used to predict human cancer patient outcomes. With the increasing numbers of clinical data sets accessible for analysis, constraints (e.g., ethical, fiscal, and logistic) associated with using human samples can be mitigated with properly designed mouse model systems. Our analytic approach (Supplementary Fig. S4, study flow chart) that involved the use of multiple animal models to derive a novel gene signature applicable to humans sets precedence that similar strategies may be used, albeit cautiously, for cross-species analysis.

Furthermore, specific knockup or knockdown abrogative studies can be done in our animal model systems to decipher the functional importance of the six genes, which are correlated with metastatic competency and subsequent human survival in breast cancer. These animal model systems may also serve to provide surrogate markers for screening of therapeutic targets. It should also be noted that the six genes in our signature are novel and independent of the genes used in the existing genomic assays for breast cancer patients. Because most of the genes in this signature have not previously been linked to metastasis, detailed molecular studies will help to determine functional roles of these genes in metastatic competency and their utilities as potential therapeutic targets.

In summary, a complex use of animal models, microarray, and statistics has led us to a gene signature that accurately and consistently predicts human breast cancer patient survival. This six-gene signature may also possess clinical utility for predicting survival outcomes in other cancer types such as lung cancer. There are certain constrains that restrict our ability to use human subjects for the most optimal methods warranted by a clinical or scientific question. The design of this study underscores the utility of animal models in the development of clinical assays when properly coupled with existing clinical data sets. It is imperative that we remain cognizant to all possible revenues derived from well designed and carefully analyzed animal models.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Qingrong Chen and Yonghong Wang for their valuable input on statistical analysis of this study. Their willingness to share their knowledge greatly helped the author get up to speed in an area previously with little familiarity. On a personal level, they have been wonderful colleagues to work with.

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.

Received 6/22/09; revised 8/21/09; accepted 9/28/09; published OnlineFirst 12/22/09.

References

1. Sporn MB. The war on cancer. Lancet 1996;347:1377–81.
2. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
3. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nat Rev Cancer 2003;3:453–8.
4. Langley RR, Fidler IJ. Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis. Endocr Rev 2007;28:297–321.
5. Kang Y, Siegel PM, Shu W, et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell 2003;3:537–49.
6. Mangiameli DP, Blansfield JA, Kachala S, et al. Combination therapy targeting the tumor microenvironment is effective in a model of human ocular melanoma. J Transl Med 2007;5:38.
7. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nat Genet 2005;37 Suppl:S38–45.

8. Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008;14:822–7.
9. Tomida S, Takeuchi T, Shimada Y, et al. Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. J Clin Oncol 2009;27:2793–9.
10. Driouch K, Landemaine T, Sin S, Wang S, Lidereau R. Gene arrays for diagnosis, prognosis and treatment of breast cancer metastasis. Clin Exp Metastasis 2007;24:575–85.
11. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
12. 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.
13. Ivshina AV, George J, Senko O, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 2006;66:10292–301.
14. Bergh J, Norberg T, Sjogren S, Lindgren A, Holmberg L. Complete sequencing of the *p53* gene provides prognostic information in breast cancer patients, particularly in relation to adjuvant systemic therapy and radiotherapy. Nat Med 1995;1:1029–34.
15. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
16. Pawitan Y, Bjohle J, Amler L, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 2005;7:R953–64.
17. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262–72.
18. Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res 2007;13:3207–14.
19. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66:7466–72.
20. 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.
21. Liu R, Wang X, Chen GY, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 2007;356:217–26.
22. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
23. Fidler IJ. The organ microenvironment and cancer metastasis. Differentiation 2002;70:498–505.
24. Kang Y. New tricks against an old foe: molecular dissection of metastasis tissue tropism in breast cancer. Breast Dis 2006;26:129–38.