

A Cell Proliferation Signature Is a Marker of Extremely Poor Outcome in a Subpopulation of Breast Cancer Patients

Hongyue Dai,¹ Laura van't Veer,² John Lamb,¹ Yudong D. He,¹ Mao Mao,¹ Bernard M. Fine,³ Rene Bernards,² Marc van de Vijver,² Paul Deutsch,³ Alan Sachs,¹ Roland Stoughton,^{1,4} and Stephen Friend^{1,3}

¹Rosetta Inpharmatics LLC, (A wholly owned subsidiary of Merck & Co. Inc.) Seattle, Washington; ²Division of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, the Netherlands Cancer Institute, Amsterdam, the Netherlands; ³Merck Research Laboratories, Merck and Co., Inc., West Point, Pennsylvania; and ⁴GHC Technologies, Inc., La Jolla, California

Abstract

Breast cancer comprises a group of distinct subtypes that despite having similar histologic appearances, have very different metastatic potentials. Being able to identify the biological driving force, even for a subset of patients, is crucially important given the large population of women diagnosed with breast cancer. Here, we show that within a subset of patients characterized by relatively high estrogen receptor expression for their age, the occurrence of metastases is strongly predicted by a homogeneous gene expression pattern almost entirely consisting of cell cycle genes (5-year odds ratio of metastasis, 24.0; 95% confidence interval, 6.0-95.5). Overexpression of this set of genes is clearly associated with an extremely poor outcome, with the 10-year metastasis-free probability being only 24% for the poor group, compared with 85% for the good group. In contrast, this gene expression pattern is much less correlated with the outcome in other patient subpopulations. The methods described here also illustrate the value of combining clinical variables, biological insight, and machine-learning to dissect biological complexity. Our work presented here may contribute a crucial step towards rational design of personalized treatment. (Cancer Res 2005; 65(10): 4059-66)

Introduction

Being able to predict outcome and to understand the biological mechanisms leading to poor outcome are two key objectives in cancer research. Recently, important new diagnostic and prognostic information for various cancers has been provided by gene expression profiling studies (1-14). However, a major challenge raised by these studies is to develop appropriate strategies for integrating gene expression data with clinical and histopathologic variables to further improve predictive power and provide deeper insights into disease mechanisms that will have clinical impact.

We previously established a 70-gene-based prognostic classifier (3) for breast cancers diagnosed before age 55. This classifier outperformed clinical predictors and showed good potential in selecting good outcome patients and thereby minimizing overtreatment (15). However, the group of patients that were predicted to have a poor outcome did not have uniform outcomes, with many (52%) patients not developing metastases (mean follow-up of ~8 years). Moreover, the 70 prognostic genes are involved in a

variety of biological processes and thus provided limited insight into biological mechanisms that affect clinical outcome. The uniform gene expression pattern for good outcome patients and heterogeneous patterns for the poor outcome patients in refs. (3, 15) suggest that the biological processes associated with good outcome are more homogeneous than those associated with poor outcome. These observations led to two topics that are the focus of the current study: (a) identifying a subset of patients with high risk to poor outcome and (b) identifying a coherent set of genes that provide biological insight into the mechanisms responsible for poor outcome.

Gene expression alone is likely to identify a subset of patients that are dominated by poor outcome only if the relevant patient groups have a distinctive gene expression pattern. When this is not the case, it may be possible to use clinical measures and existing understanding (even if incomplete) of the disease process to impose specific patient stratification to guide the machine-learning phase of gene expression analysis to develop a prognostic classifier. Such an integrated approach to find optimal prognostic classifiers is the subject of this study.

Specifically, we used the estrogen receptor (ER) level and its variation with age at diagnosis to subdivide the patients. ER status has a marked influence on the gene expression in breast cancer, affecting the expression of >10% of the genes in breast tumors (2, 3, 5, 16, 17), and is generally thought to have an important impact on survival (15, 18-20). Age is also prognostic, with breast cancer in younger patients having a poorer outcome (21). These two variables have been previously used as independent prognostic factors, and interestingly, it has recently been reported that the percentage of ER+ breast carcinomas increases with patient age (22).

The current study shows that using this combination of clinical variables, a subgroup of patients is identified in which expression of proliferation-associated genes is a very strong predictor of outcome. In contrast, proliferation index and tumor grade (a histologic assessment about the aggressiveness of cell growth) have only limited predictive power when used without preselection of patients (see, e.g., refs. 15, 23-27).

Materials and Methods

Tumor samples

Three hundred and eleven breast carcinoma samples that satisfied the selection criteria defined in ref. (15), and described in recent publications (3, 15), were included in the analysis. Specially, these samples include 295 samples from our cohort study (15) and 16 nonredundant samples from our initial study (3). All clinical data are shown in Table S1 of the Supplementary Information.

Data analysis

Estrogen receptor level. ER level was measured by a 60-mer oligonucleotide on our human microarray. Because every individual sample

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

Requests for reprints: Stephen Friend, Merck Research Laboratories, Merck and Co. Inc., PO Box 4, WP14-2500, 770 Sumneytown Pike, West Point, PA 19486. Phone: 215-652-7313; Fax: 215-993-4114; E-mail: stephen_friend@merck.com.

©2005 American Association for Cancer Research.

was compared with a pool of all samples, the ratio to pool was used to measure the relative expression level. We used the same threshold of -0.65 on $\log_{10}(\text{ratio})$ to separate the ER+ group from the ER- group as previously established in ref. (3).

Classification method. The basic algorithms for classification used here are the same as previously used in ref. (3), except for changes listed below.

Feature selection and performance evaluation. For the prognosis in each group, we started by filtering noninformative genes as described in ref. (3). The second step involved a double loop of leave-one-out cross-validation procedure, with the first loop to select the “training samples” (see section below), and the second loop to evaluate the performance. Prognostic features were selected based on the training samples by their correlation to outcome and were reselected during each step of leave-one-out cross-validation. See Supplementary Information for more details.

Identifying homogeneous patterns and dominant mechanism by iterative training sample selection. We developed a method called “iterative training sample selection”, or “homogeneous pattern” in order to reveal the dominant mechanisms. In the first step, only the samples of those patients who had metastases within 5 years or who were metastases-free with more than 5 years of follow-up time were used as the training set. Based on these training samples, a complete leave-one-out cross-validation (including reselecting features) process was done. During this step, the number of features was fixed at 50 genes (the number is chosen to provide a stable classifier by our algorithm). The training samples that were not correctly classified (poor samples correlating more to the average good, or vice versa) by this leave-one-out cross-validation process were further removed from the training set in the second round of leave-one-out cross-validation (see Fig. S6 in Supplementary Information for training samples used for current study). This is the opposite of the “boost” algorithm (28). The boost algorithm increases the weight of the misclassified samples in the training for improving the accuracy of the classifier. The current algorithm focuses on the most common prediction rule (mechanism) within the data set by excluding the “unpredictable” from the training set for robust feature selection. With this method, we selected a very homogeneous group of genes which happened to all be associated with the cell cycle. Due to the homogeneous expression pattern, the classifier accuracy is relatively insensitive to the number of features included in the classifier. Even though improved classifier accuracy is not the objective of this algorithm, it resulted in an improved accuracy in this study, probably due to the identification of a robust feature set.

Error rate and odds ratio, threshold in the final leave-one-out cross-validation. Unless otherwise stated, the error rate is the average error rate from two populations: poor outcome samples misclassified as good divided by total poor samples, and good outcome samples misclassified as poor over total good samples. We report two odds ratios for a given threshold: the overall odds ratio and 5-year odds ratio (5-year odds ratio was calculated from those samples with more than 5-year metastases-free or metastasized within 5 years). The threshold was applied to $\text{cor1} - \text{cor2}$, where “cor1” is the correlation to the “average good profile” in the training set, and “cor2” is the correlation to the “average poor profile” in the training set. The threshold in the final round of leave-one-out cross-validation was determined by a method described in Supplementary Information to avoid overestimating the performance.

Correlation calculation. The correlation between each gene’s expression $\log(\text{ratio})$ and the end point data (final outcome) was calculated using the Pearson correlation coefficient. The correlation between each patient’s profile and the average good profile and average poor profile is the cosine product (without mean subtraction).

Kaplan-Meier plot. Only the patients belonging to the original 295 cohort samples were used for the Kaplan-Meier plot. Overall survival was defined by death from any cause. In the analysis of distant metastasis-free probabilities, patients whose first event was distant metastases were counted as failures; all other patients were censored at the date of their last follow-up, non-breast cancer death, local-regional recurrence or second

primary malignancy, including contralateral breast cancer. Time was measured from the date of surgery. Metastasis-free curves were drawn using the method of Kaplan and Meier and compared using the log-rank test.

Results

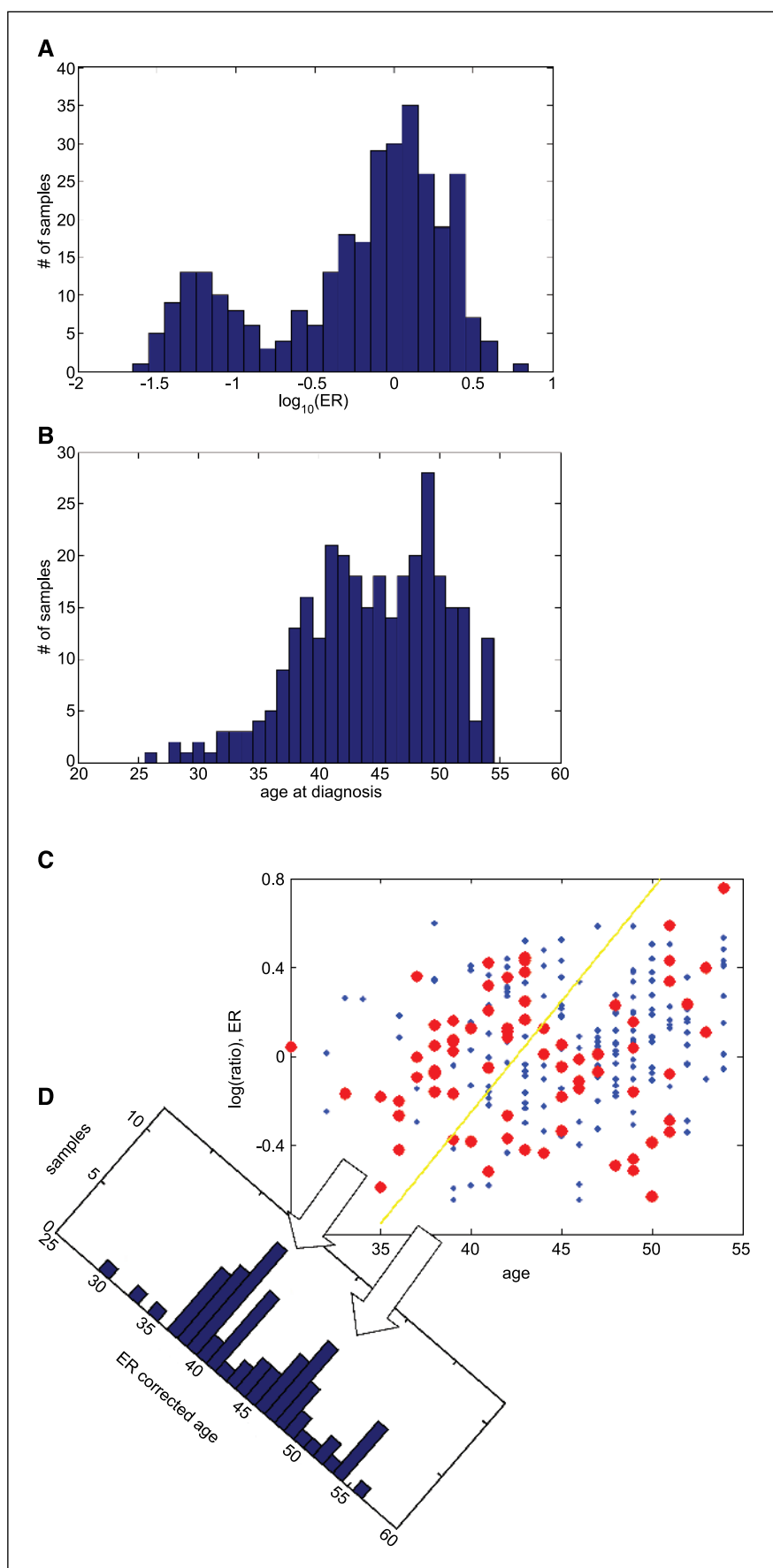
Estrogen receptor level dependence on age and sample stratification based on this dependence. Because the ER level is a dominant factor in breast cancers at the molecular level (3, 16, 29), and age is a prognostic factor (25, 30), we grouped patients based on these two characteristics. When the ER level obtained from the microarray measurements was plotted versus age (Fig. 1C) for the ER-positive patients, we noticed that the patients were not evenly distributed. There seems to be a paucity of samples in the top left and bottom right corners, suggesting a general trend of ER expression increasing with age. This trend is further supported by including patients with age at diagnosis greater than 55 years from a separate data set [see Fig. S1 of Supplementary Information, where the lack of samples in the bottom right is more obvious ($P = 1.1 \times 10^{-9}$), and the ER expression level of ER+ patients relative to the ER- population is even higher]. In addition to this general trend, we also observed indications of two subpopulations as separated by the yellow line in Fig. 1C and D. A bimodality test based on Monte-Carlo simulation shows that the probability of observing such a separation by chance is $<0.01\%$ (see Supplementary Information).

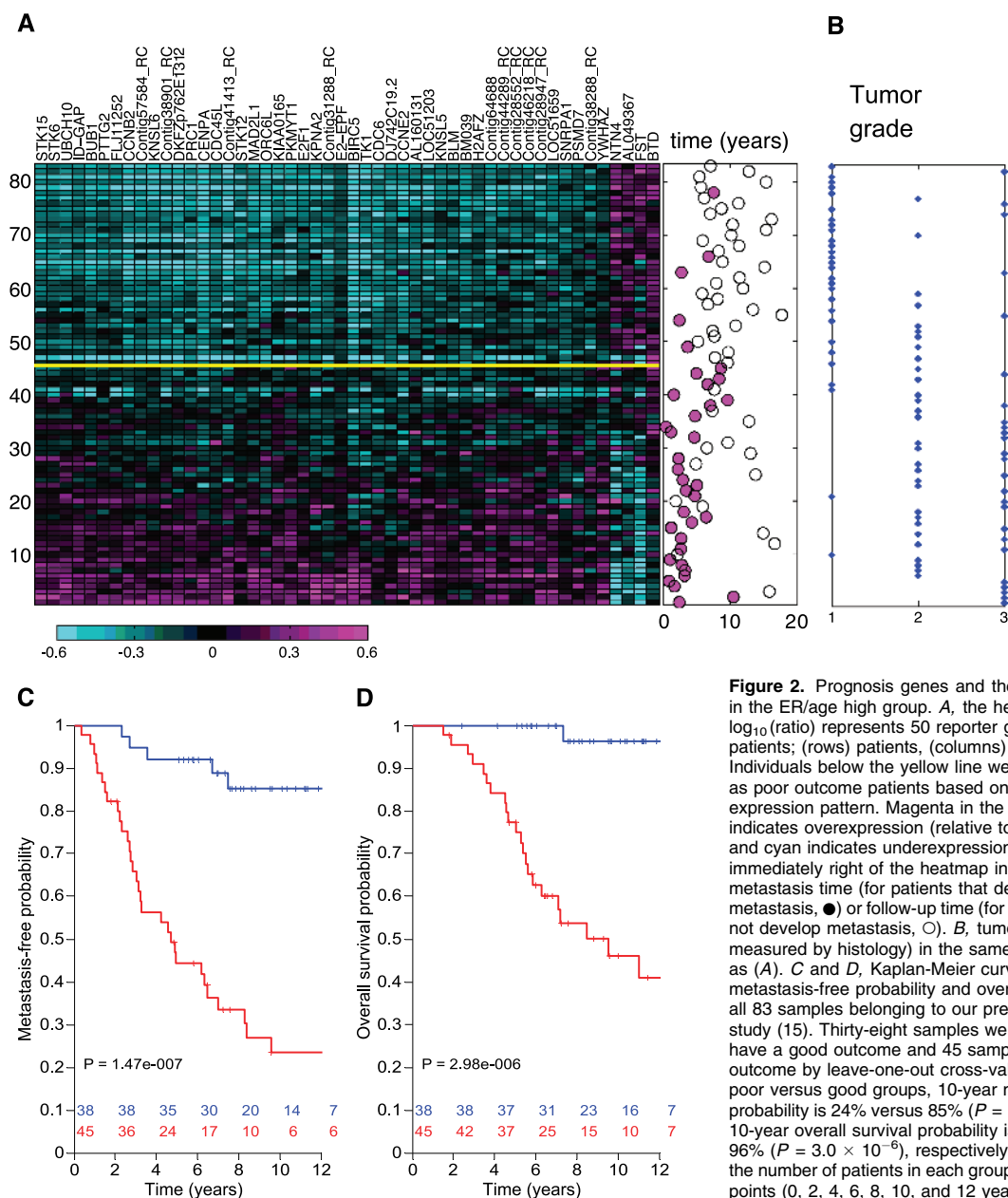
This division of patients based on the combination of ER expression and age triggered further investigations. As shown in Fig. 1, we stratified the breast cancer samples based on their ER expression level and age. The samples above the yellow line are termed “ER/age high” group (which means high ER expression for their age; 83 patients, all from the 295-patient cohort in ref. 15), and the samples below the yellow line as “ER/age low” group (156 patients, 143 from the 295 patient cohort in ref. 15).

Overall outcome is poor in the ER/age high group. The fraction of patients who developed metastases was 43% in the ER/age high group, and 24% in the ER/age low group. The probability of observing such an asymmetry in metastases rate by chance is 3×10^{-11} . This drastic difference in metastasis rate provides additional evidence of two subpopulations within the ER+ patients.

Cell cycle genes are strongly prognostic in ER/age high group, but less or nonprognostic in other groups. Within the ER/age high group, we identified a group of 50 prognostic reporter genes that were highly correlated with the outcome (see Materials and Methods and Table S3 in Supplementary Information). Moreover, the expression of these prognostic genes is relatively homogeneous as indicated by high similarity in expression patterns among those genes as shown in Fig. 2A. Leave-one-out cross-validation, including reporter selection, yielded an odds ratio for metastasis of 14.6 [95% confidence interval (CI) 4.7-45.4] and 5-year odds ratio (see Materials and Methods) of 24.0 (95% CI, 6.0-95.5; see Table 1 for summary information). In the group of patients predicted to have a poor outcome, 31 out of 45 (69%) developed metastases (mean follow-up time, 7.1 years). The 10-year metastasis-free probability is only 24% (for Kaplan-Meier plots, the leave-one-out cross-validation was used to predict samples into “good” and “poor” prognosis groups, Fig. 2C). In contrast, in the group predicted to have a good

Figure 1. ER level versus age for ER-positive breast cancer patients at primary diagnosis. *A*, ER level distribution for the breast tumors in current study. The ER level was measured relative to the average level of all samples using microarrays. *B*, age (years) distribution of all samples in current study. *C*, scatter plot of ER level versus age. The yellow line is used to stratify the ER+ samples [characterized by ER $\log(\text{ratio}) > -0.65$] into ER/age (years) high (above the line) and ER/age low (below the line) groups. The ER- patients characterized by ER $\log(\text{ratio}) < -0.65$ were not shown in this plot. (Red dots) patients who developed metastases; (blue dots) good outcome patients who were metastases-free within follow-up time interval. *D*, histogram of metastasis patients when projected along the yellow line with age + $10 \times \log(\text{ER})$ ["ER-corrected age", or simply, the ages where the line crosses the ER $\log(\text{ratio}) = 0$].





outcome, only 5 out of 38 patients (13%) developed metastases, and the 10-year metastasis-free probability is 85%. It is noteworthy that the overall survival rate at 10 years is only 46% for the poor prognosis group, in comparison with 96% for the good prognosis group (Fig. 2D).

Examination of molecular functions and biological processes of the 50 prognostic genes reveals that many of the highly expressed genes in tumors of poor outcome are cell cycle-associated genes (for example, *STK6*, *STK12*, *CCNB2*, *CCNE2*, *BUB1*, *CDC6*, *CDC25A*, *CDC45L*, *MAD2L1*, *RBL2*, *E2F1*, *KNSL5*, *UBE2C*, *UBCH10*, *PKMYT1*, and *BIRC5*). Further examination of these genes in synchronized HCT116 colon cancer cell line reveals that 16 genes are G₁ phase specific (overexpressed in G₁ phase) and 25 genes are G₂-M specific (see Supplementary Information). Thirty of the

50 prognostic genes are also supported by a previous study (31) as cell cycle-related genes.

Overexpression of cell cycle genes is indicative of cell proliferation, which in turn is known to be associated with poor outcome. Patients whose tumors have a high proliferation rate have an increased risk (10-20%) of metastasis or death (see, for example, refs. 32, 33). This relatively small difference in outcome may be due primarily to the fact that cell proliferation has less of an impact on outcome in the ER/age low patients (Fig. 3A) and essentially no impact on outcome in the ER- patients (Fig. 3D). When the same classifier was applied to the ER/age low group (the ER+ patients not included in the ER/age high group), the overall odds ratio for metastasis is 1.59 (95% CI, 0.74-3.41) and 5-year odds ratio is 3.51 (95% CI, 1.24-10.0). To construct a

classifier, a threshold is used to separate poor outcome from good outcome predictions. Even with a threshold reoptimized for the ER/age low group, the overall odds ratio is only 2.79 (95% CI, 1.31-5.95) and the 5-year odds ratio is 5.29 (2.04-13.7), far less than those for the ER/age high group. This limited power is shown in the Kaplan-Meier plots in Fig. 3B and C. With the reoptimized threshold, the separation between the predicted good and poor group measured by the metastasis-free probability and overall survival probability is only approximately 20% at 10 years. In the ER— group (Fig. 3D), almost all of the patients have evidence of high proliferation, yet only 43% of patients develop metastases. The error rate for predicting metastasis is approximately 50% (no predictive value), no matter what threshold is chosen for the classifier. Figure 3E and F show that almost all samples were predicted to have a poor outcome due to the high expression of proliferation genes.

Tumor grade correlates with the microarray findings. If we select relatively young patients with breast cancer with relatively high ER levels in their tumors (i.e., “ER+ER/age high” group), we find that histologic tumor grade alone has strong prognostic power (Fig. 2B). Low grade (grade = 1) is associated with low risk of metastasis, and high grades (grades = 2 or 3) are associated with poor outcome. The overall odds ratio for metastasis is 5.9 (95% CI, 2.0-18.0), and 5-year odds ratio is as high as 12.5 (95% CI, 2.6-59.3). These predictions are not as strong as those based on the gene expression classifier, but much better than the predictions by clinical variables for the entire patient populations, where the hazard ratio is typically around 2 (refs. 25, 32, 34). Figure 4 compares the metastasis-free probability and overall survival rate for low (grade = 1) and high grades (grade = 2 or 3) in the ER/age high group and ER/age low, respectively. As shown in Fig. 4A and B, high grades in the ER/age high group accurately predict poor outcome (10-year metastasis-free probability is 38%, overall survival rate is 53%), and the separation between the predicted poor and good group is significant ($P < 0.001$). However, in the ER/age low group (Fig. 4C and D), high grades are no longer associated with very poor (10-year metastasis-free probability is 70%, overall survival rate is 78%), and the separation between the predicted poor and good group is not significant ($P > 0.1$). The poor performance of tumor grade in the ER/age low group explains why grade is only a limited prognostic variable for the entire patient population. The good performance of grade in the ER/age high group, on the other hand, provides independent support for the rules learned from the gene expression data. Thus, the performance of a well-known prognostic indicator such as tumor grade can be refined and improved by combination with the ER/age variable. Additionally, the gene expression analysis provided mechanistic insight into the well known prognostic indicator of tumor grade, suggesting that in the ER/age high group, proliferation is associated with elevated grade.

Results are robust in the training and validation fashion. We used 3-fold cross-validation to show the robustness of our process. We randomly selected two-thirds of the samples (54 out of 83 in the ER/age high group) as the training set, and the remaining one-third (29 samples) were used as the validation set. The classifier (including reselecting features) was built based on the training samples only. To avoid the arbitrariness in dividing samples into “training” and “validation”, we repeated the procedure >800 times with a different division each time. The overall odds ratio for the ER/age high group in the validation samples is 9.1 (95% CI, 3.0-27.7) and the 5-year odds ratio is 21.9 (95% CI, 4.6-104.7). These numbers agree well with the performance estimated by the leave-one-out cross-validation process, demonstrating the robustness of our current approach.

Results are robust to the choice of exact division of patient groups. To gain more confidence that these are truly two distinct subgroups, it is important to examine whether the loss of prognostic power from the ER/age high group to the ER/age low group occurs at a relatively discrete boundary, or is continuous. Thus, we developed a classifier for all ER+ samples. Due to the homogeneous pattern method we used (see Materials and Methods), the prognostic genes are again almost entirely cell cycle-related (see Fig. S7 and “Prediction accuracy versus dividing line position in the ER/age plot” in the Supplementary Information). We then determined the prediction accuracy for the population of patients above the yellow line of Fig. 1C as we moved the yellow line position (in parallel to the line in the figure). As shown in Fig. 5, the error rate increased continuously as the yellow line shifted from left to right, but interestingly, became constant after it passed the position indicated in Fig. 1C (see also Fig. S8 of Supplementary Information for more details). This result suggests that (a) the strong prognostic power of the cell cycle genes in the ER/age high group is robust to the choice of exact division of patient groups, and (b) the ER/age low group is not simply a continuum of the ER/age high group because the error rate did not continue to increase as one moved through the ER/age low group.

Discussion

We have observed that for patients with breast cancer with relatively high ER expression level for their age (i.e., the ER/age high group), cell proliferation is a strong predictor for metastasis. This predictive power is greatly reduced for other groups of patients with breast cancer.

Although the present study is based on patients less than 55 years of age, our conclusions are unlikely to change when older patients are included. As shown in Fig. 1C, the lack of patients with ages greater than 50 years in the ER/age high group indicates that an additional number of older patients will not affect the

Table 1. Performance of cell cycle genes in predicting outcome

	Overall odds ratio (95% CI)	5-year odds ratio (95% CI)	10-year metastasis-free probability	10-year overall survival probability	<i>P</i> value of log-rank test (metastasis-free)
ER/age high	14.6 (4.7-45.4)	24.0 (6.0-95.5)	85% versus 24%	96% versus 46%	1.5×10^{-7}
ER/age low	2.79 (1.31-5.95)	5.29 (2.04-13.7)	82% versus 62%	92% versus 70%	0.006
ER—	2.37 (0.23-23.9)	2.61 (0.26-26.6)	75% versus 47%	75% versus 45%	0.46



www.aacrjournals.org

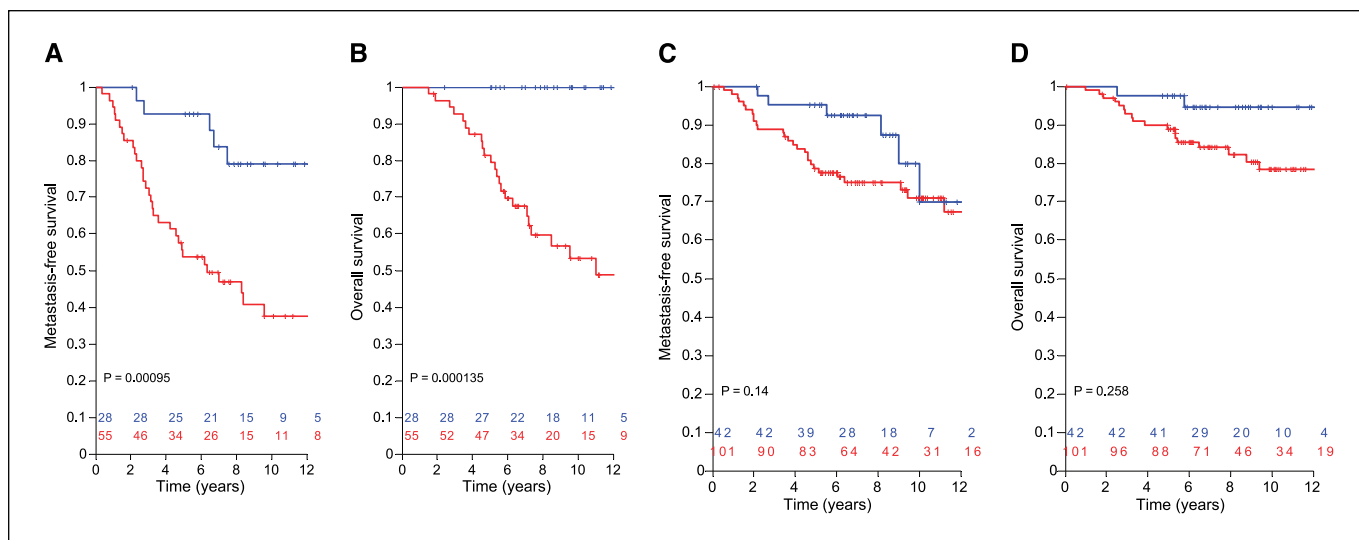


Figure 4. Tumor grade is a strong prognostic variable in the ER/age high group, but not in the ER/age low group. *A* and *B*, Kaplan-Meier curves for metastasis-free probability and overall survival by using tumor grade as a prognostic predictor in the ER/age high group. Twenty-eight samples were predicted to have a good outcome (grade = 1) and 55 samples with poor outcome (grade = 2 or 3). For poor versus good groups, 10-year metastasis-free probability is 38% versus 79% ($P = 0.001$) and 10-year overall survival is 53% versus 100% ($P = 0.0001$), respectively. Both curves show that tumor grade has strong prediction power in this group. *C* and *D*, Kaplan-Meier curves for metastasis-free probability and overall survival by using tumor grade as a prognostic predictor in the ER/age low group. Forty-two samples were predicted to have a good outcome (grade = 1) and 101 samples with poor outcome (grade = 2 or 3). For poor versus good groups, 10-year metastasis-free probability is 71% versus 70% ($P = 0.14$) and 10-year overall survival is 78% versus 95% ($P = 0.26$), respectively. Both curves show that tumor grade has next to no prediction power in this group.

performance of proliferation genes in this group. The inclusion of older age patients is unlikely to change the reduced prognostic power in other groups either, because the prognostic value of tumor grade for the entire patient population (the majority are older aged) is not as strong as that for the ER/age high subgroup we observed.

The different degrees of association between cell proliferation and poor outcome in different groups of patients confirms the concept that breast cancer pathogenesis and tumor maintenance is heterogeneous, with different subtypes likely having independent pathways of tumor progression. Previous prognostic factors for metastases are limited by their applications to all patients, but can be improved when applied to the right subgroup of patients as shown in the current paper.

It is worth noting that even though the patients in the ER/age high group are clinically heterogeneous, the incidence of distant metastases is strongly predicted by a biologically uniform set of genes, indicating that proliferation is the prime driving force for disease progression. In contrast, in other breast cancer subgroups, factors in addition to tumor cell proliferation may also be important in determining outcome.

The results revealed by the expression data in the ER/age high group has important clinical implications. In particular, the prognosis of patients in this group may be predicted solely using the combination of particular clinical and histopathologic variables. For example, one can use an immunohistochemical measurement of ER level if it has enough accuracy (the immunohistochemical measure of ER correlates with mRNA level of ER, see, for example, ref. 3). Otherwise, PCR measure of mRNA abundance of ER and patients' age at diagnosis can be used to select the ER/age high patients and to test whether tumor grade has a significant prognostic power. If validated, this would have a significant impact on the treatment decisions for these patients.

Biologically, the fact that grouping patients based on ER expression level and age yielded good results might imply that there is an important mechanism governing the relationship between ER expression level and patient age. After seeing good performance with such stratification, we assessed the error rate using various stratifications along the ER axis or age axis independently. None of them did as well as the approach using the ER and age dependence (see Supplementary Information).

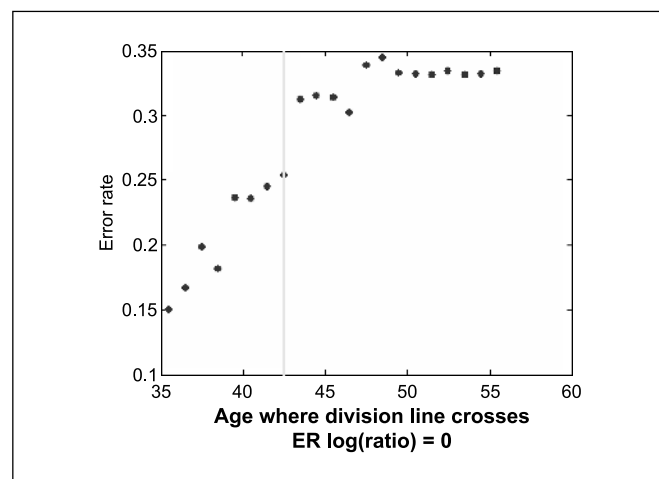


Figure 5. The leave-one-out cross-validation error rate versus patient stratification. A line parallel to the yellow line in Fig. 1C was considered. As one moves this line from left to right, the number of patients above the line increases. The leave-one-out cross-validation error rate for the patients above the line (cumulative error rate) is plotted as a function of the line position. The line position is indicated by the X-position (in age) where the line crosses $\log(\text{ER ratio}) = 0$. The vertical line in this figure is in the same position as the yellow line in Fig. 1C. This figure shows that the error rate increases continuously as the line moves from left to right, and becomes constant after it passes the vertical line.

From a data mining point of view, combining gene expression with other types of information represents a promising new direction. Gene expression data obtained from clinical samples are generally difficult to interpret because they provide only a snapshot of a complicated disease state. Integrating clinical information with gene expression is crucial for the interpretation of this rich and complicated information. From a model prediction point of view, Pittman et al. (35) made good progress in improving prediction accuracy by including gene expression and clinical variables in a decision tree. In this study, instead of equally mixing clinical data with gene expression data in a machine-learning model, we used clinical variables to stratify the patients.

It is not clear why patients with high ER/age seem to be so biologically distinct. It is possible that tumors in young patients with high ER have a unique propensity to depend on the identified proliferation-associated genes. It is noteworthy that a homotogeneous prognostic gene expression pattern was identified in this group, and confirmation in independent populations would support the significance of this unexpected finding.

In conclusion, by combining ER expression level and age, we identified a group of patients with relatively poor outcome. Within this group, a gene expression classifier identifies a subgroup of patients with an almost 70% chance of metastasis. Importantly, this gene expression classifier suggests that cell proliferation is the driving mechanism associated with poor outcome. These results suggest that further refinements of diagnostic predictors may more often be generated by combining different informative clinical and molecular variables. The integrative approach used in this study also shows the value of moving beyond single-variable statistical comparisons when introducing new prognostic markers.

Acknowledgments

Received 11/5/2004; revised 1/18/2005; accepted 2/18/2005.

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.

We thank Drs. Peter Linsley, Doug Bassett, Vladimir Svetnik, Richard Raubertas, and I-Ming Wang for their critical and fruitful discussions. Dr. Jerald Radich also made valuable comments that helped to improve the manuscript. The cRNA samples of HCT116 provided by Dr. Carolyn Buser-Doepner and processed by Dr. Steven Bartz were used for the Supplementary Information.

References

- Zajchowski DA, Bartholdi MF, Gong Y, et al. Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. *Cancer Res* 2001;61:5168-78.
- 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.
- 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.
- Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003;100:8418-23.
- Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;406:747-52.
- Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-13.
- Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531-7.
- DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-60.
- Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403:503-11.
- Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* 2004;113:913-23.
- Ishibashi Y, Hanyu N, Nakada K, et al. Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma. *Cancer Res* 2003;63:5159-64.
- Latil A, Bieche I, Chene L, et al. Gene expression profiling in clinically localized prostate cancer: a four-gene expression model predicts clinical behavior. *Clin Cancer Res* 2003;9:5477-85.
- Pang ST, Dillner K, Wu X, et al. Gene expression profiling of androgen deficiency predicts a pathway of prostate apoptosis that involves genes related to oxidative stress. *Endocrinology* 2002;143:4897-906.
- Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816-24.
- 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.
- Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001;61:5979-84.
- 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.
- Surowiak P, Dziegiel P, Zabel M, Matkowski R, Kornafel J. Prognostic value of immunocytochemical estimation of estrogen receptor (ER) and of pS2 estrogen-dependent protein in cells of mammary ductal carcinoma. Analysis of five-year course of the disease. *Folia Histochem Cytobiol* 2001;39:143-4.
- Pichon MF, Broet P, Magdelenat H, et al. Prognostic value of steroid receptors after long-term follow-up of 2257 operable breast cancers. *Br J Cancer* 1996;73:1545-51.
- Collett K, Hartveit F, Skjaerven R, Maehle BO. Prognostic role of oestrogen and progesterone receptors in patients with breast cancer: relation to age and lymph node status. *J Clin Pathol* 1996;49:920-5.
- Maggard MA, O'Connell JB, Lane KE, et al. Do young breast cancer patients have worse outcomes? *J Surg Res* 2003;113:109-13.
- Foulkes WD, Metcalfe K, Sun P, et al. Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type. *Clin Cancer Res* 2004;10:2029-34.
- van Diest PJ, van der Wall E, Baak JP. Prognostic value of proliferation in invasive breast cancer: a review. *J Clin Pathol* 2004;57:675-81.
- Simpson JF, Gray R, Dressler LG, et al. Prognostic value of histologic grade and proliferative activity in axillary node-positive breast cancer: results from the Eastern Cooperative Oncology Group Companion Study, EST 4189. *J Clin Oncol* 2000;18:2059-69.
- Bijker N, Peterse JL, Duchateau L, et al. Risk factors for recurrence and metastasis after breast-conserving therapy for ductal carcinoma-in-situ: analysis of European Organization for Research and Treatment of Cancer Trial 10853. *J Clin Oncol* 2001;19:2263-71.
- Mirza AN, Mirza NQ, Vlastos G, Singletary SE. Prognostic factors in node-negative breast cancer: a review of studies with sample size more than 200 and follow-up more than 5 years. *Ann Surg* 2002;235:10-26.
- Reed W, Hannisdal E, Boehler PJ, et al. The prognostic value of p53 and c-erb B-2 immunostaining is overrated for patients with lymph node negative breast carcinoma: a multivariate analysis of prognostic factors in 613 patients with a follow-up of 14-30 years. *Cancer* 2000;88:804-13.
- Freund Y. Boosting a weak learning algorithm by majority. *Information Comput* 1995;121:256-85.
- 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.
- Kaufmann M. Review of known prognostic variables. *Recent Results Cancer Res* 1996;140:77-87.
- Whitfield ML, Sherlock G, Saldanha AJ, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 2002;13:1977-2000.
- Liu S, Edgerton SM, Moore DH II, Thor AD. Measures of cell turnover (proliferation and apoptosis) and their association with survival in breast cancer. *Clin Cancer Res* 2001;7:1716-23.
- Medri L, Volpi A, Nanni O, et al. Prognostic relevance of mitotic activity in patients with node-negative breast cancer. *Mod Pathol* 2004;17:1024.
- Michels JJ, Marnay J, Delozier T, Denoux Y, Chasle J. Proliferative activity in primary breast carcinomas is a salient prognostic factor. *Cancer* 2004;100:455-64.
- Pittman J, Huang E, Dressman H, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A* 2004;101:8431-6.