Patterns of Gene Expression That Characterize Long-term Survival in Advanced Stage Serous Ovarian Cancers

Andrew Berchuck,[1] Edwin S. Iversen,[2,3] Johnathan M. Lancaster,[1] Jennifer Pittman,[3] Jingqin Luo,[3] Paula Lee,[1] Susan Murphy,[1] Holly K. Dressman,[4] Phillip G. Febbo,[5] Mike West,[3] Joseph R. Nevins,[4] and Jeffrey R. Marks[6]

Abstract

Purpose: A better understanding of the underlying biology of invasive serous ovarian cancer is critical for the development of early detection strategies and new therapeutics. The objective of this study was to define gene expression patterns associated with favorable survival.

Experimental Design: RNA from 65 serous ovarian cancers was analyzed using Affymetrix U133A microarrays. This included 54 stage III/IV cases (30 short-term survivors who lived <3 years and 24 long-term survivors who lived >7 years) and 11 stage I/II cases. Genes were screened on the basis of their level of and variability in expression, leaving 7,821 for use in developing a predictive model for survival. A composite predictive model was developed that combines Bayesian classification tree and multivariate discriminant models. Leave-one-out cross-validation was used to select and evaluate models.

Results: Patterns of genes were identified that distinguish short-term and long-term ovarian cancer survivors. The expression model developed for advanced stage disease classified all 11 early-stage ovarian cancers as long-term survivors. The *MAL* gene, which has been shown to confer resistance to cancer therapy, was most highly overexpressed in short-term survivors (3-fold compared with long-term survivors, and 29-fold compared with early-stage cases). These results suggest that gene expression patterns underlie differences in outcome, and an examination of the genes that provide this discrimination reveals that many are implicated in processes that define the malignant phenotype.

Conclusions: Differences in survival of advanced ovarian cancers are reflected by distinct patterns of gene expression. This biological distinction is further emphasized by the finding that early-stage cancers share expression patterns with the advanced stage long-term survivors, suggesting a shared favorable biology.

Epithelial ovarian cancers recapitulate all the histologic patterns observed in the müllerian derived structures of the genital tract, including the serous pattern of the fallopian tube, the endometrioid pattern of the endometrium, and the clear cell and mucinous patterns of the endocervix. Despite the histologic heterogeneity, most deaths are attributable to the serous type, which comprises 60% of cases and has a propensity to present at an advanced stage. Only 5% of women with invasive serous ovarian cancers have early-stage (I/II) disease that is confined to the pelvis and these patients have an 80% to 90% probability of survival (1). Advanced (stage III/IV) cases generally are disseminated throughout the peritoneal cavity at diagnosis. Contemporary surgical management includes removal of the primary ovarian tumor and resection of metastases. The goal is to leave minimal residual disease before administration of chemotherapy with platin and taxane drugs, as optimal surgical debulking is associated with better survival (2). Although the majority of patients with advanced disease achieve complete clinical responses, more than 90% develop recurrences and median survival is about 3 to 4 years (3, 4). The amount of residual cancer present after initial surgery is a strong prognostic factor (2), but some optimally debulked cases exhibit *de novo* resistance to chemotherapy and have poor survival. At the other extreme, a small minority of suboptimally debulked cancers are exquisitely sensitive to chemotherapy and, similar to early-stage cases, never relapse. This range of outcomes is thought to be attributable, at least in part, to variations in underlying biological characteristics of ovarian cancers. Host factors such as nutritional status and immune function also likely play a role in determining survival.

Prior studies that have sought to elucidate the molecular determinants of outcome in serous ovarian cancers generally have focused on single genes. Alterations in several oncogenes and tumor suppressor genes have been described in serous ovarian cancers (5). Mutation of the *TP53* tumor suppressor gene is the most common genetic alteration in ovarian cancers, occurring in at least 70% of advanced stage cases (6). Among patients with advanced stage disease, however, neither *TP53* mutational status (6) nor other single gene alterations have proven strong predictors of outcome.

The development of microarray technology now permits analysis of expression levels of thousands of genes in a cancer. Our group and others have shown that expression microarrays can be used to identify unique characteristics of tumors that can predict clinical phenotypes and survival in breast cancer (7, 8). In the present study, we used Affymetrix U133A microarrays containing over 22,000 probe sets to define expression patterns that distinguish between short-term (<3 years) and long-term (>7 years) survival in advanced serous ovarian cancers. Bayesian linear discriminant and classification tree analyses yielded models that classified cases correctly when subjected to leave-one-out cross-validation. In addition, these models classified all early-stage ovarian cancers as having gene expression patterns similar to that of the cancers from long-term survivors with advanced stage disease suggesting a shared favorable biology.

Materials and Methods

Patients and tissue acquisition

Invasive serous ovarian cancer tissue was snap-frozen at initial surgery before any chemotherapy in 65 women treated at Duke University Medical Center between 1988 and 2001 under the auspices of a protocol approved by the Duke Institutional Review Board. This study included 54 advanced stage (III/IV) cases and 11 early-stage (I/II) cases that were surgically staged. The stage III/IV cases all received platin-based combination chemotherapy. Survival was less than 3 years in 30 cases and greater than 7 years in 24 cases. Median survival of the short-term survivors was 17.5 months. Median survival of the long-term survivors was 107.5 months. Twelve long-term survivors remain alive, seven of whom never developed recurrent disease. None of the patients in this study died of causes other than ovarian cancer. None of the 11 patients with invasive stage I/II disease have died, although one with stage IIC disease is in remission after treatment of recurrent disease. Eight have survived for more than 5 years, and three for more than 3 years since surgery.

Microarray analysis

Frozen tissue samples were embedded in optimum cutting temperature medium and sections were cut and mounted on slides. The slides were stained with H&E to assure that at least 60% of the cellular content was comprised of cancer cells. Approximately 30 mg of tissue were added to a chilled BioPulverizer H tube (Bio101 Systems, Carlsbad, CA). Lysis buffer from the Qiagen Rneasy Mini kit was added and the tissue homogenized for 20 seconds in a Mini-Beadbeater (Biospec Products, Bartlesville, OK). Tubes were spun briefly to pellet the garnet mixture and reduce foam. The lysate was transferred to a new 1.5 mL tube using a syringe and a 21-gauge needle, followed by passage through the needle 10 times to shear genomic DNA. Total RNA was extracted using the Qiagen Rneasy Mini kit. Two extractions were done for each tumor and the total RNA pooled at the end of the Rneasy protocol, followed by a precipitation step to reduce volume. Quality of the RNA was checked by an Agilent 2100 Bioanalyzer. Labeled probes for Affymetrix DNA microarray analysis were prepared according to the instructions of the manufacturer. Biotin-labeled cRNA, produced by *in vitro* transcription, was fragmented and hybridized to the Affymetrix U133A GeneChip arrays (22,283 probe sets, http://www.affymetrix.com/support/technical/byproduct.affx?product=hgu133, Santa Clara, CA) at 45°C for 16 hours and then washed and stained using the GeneChip Fluidics. The arrays were scanned to a target intensity of 500 by a GeneArray Scanner and patterns of hybridization detected as light emitted from the fluorescent reporter groups incorporated into the target and hybridized to oligonucleotide probes. The microarray data (RMA and MAS format) and accompanying clinical data are available at http://data.cgt.duke.edu/clinicalcancerresearch.

Validation of microarray results

RNA from ovarian cancers analyzed by microarrays was used to validate expression of *MAL* (which exhibits the highest fold difference in expression between the long-term and short-term survivors), *TP53*, and *IGF2* by quantitative real-time PCR. One microgram of total RNA was reverse transcribed with random hexamer primers using the AMV First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland). Subsequent real-time PCR reactions (TaqMan Assays-On-Demand, *MAL*, Hs00242748_m1; *TP53*, Hs00153340_m1; *IGF2*, Hs00171254_m1; Applied Biosystems, Foster City, CA) were done according to the recommendations of the manufacturer on an ABI Prism 7900HT Sequence Detection System (Applied Biosystems) with the exceptions that a 25 μL reaction volume was used with 50 total cycles. Relative expression levels of each gene were obtained by normalization to the expression level of *BETA-2-MICROGLOBULIN (B2M)* (TaqMan Assays-on-Demand, Hs99999907_m1) in assays done in tandem with each of the three genes analyzed.

Statistical methodologies

Expression was calculated using the robust multiarray average algorithm (9) implemented in the Bioconductor (http://www.bioconductor.org) extensions to the *R* statistical programming environment (10). Robust multiarray average generates a background-corrected and quantile-normalized measure of expression (11) on the log 2 scale of measurement. It has been shown to be superior to other methods for detecting differential expression in spike-in experiments (9). The 22,283 probe sets were screened to remove 68 control genes, those with a small variance, and those expressed at low levels. Genes of which correlation with outcome was greater than 0.4 in absolute value were used to develop predictive models for survival via binary tree (12, 13) and linear discriminant (14, 15) analyses. Similar approaches have been taken in other studies of various cancer types (16, 17).

Tree analysis. Bayesian classification tree analysis recursively partitions the sample into subgroups according to levels of covariates and associates with each a predictive probability for the outcome variable (18). With categorical or continuous covariates, the partitioning is based on an underlying nonparametric model that is consistent with the retrospective sampling design we employ, and a conservative approach was taken to select predictors that define significant partitions (19). We generated multiple trees using this approach and based predictions on model averages (i.e., model-weighted predictions of those trees). In each leave-one-out cross-validation analysis, a single tree dominated the model average. In summarizing the overall tree analysis, we report the tree which dominated the most cross-validation runs.

Linear discriminant analysis. We employed a method to identify linear discriminant models defined by subsets of the expression variables that are predictive of survival. To identify parsimonious models and to avoid overfitting the data, this approach involves a variable that penalizes model complexity. For each value of that variable that we entertain, out-of-sample predictions were generated using a model average of the top five models identified. The model average associated with the value of the penalty variable achieving the best out-of-sample predictive performance was reported.

Combined tree and linear discriminant (hybrid) modeling. Tree and linear discriminant function predictions were averaged to form

composite predictions. Predictions based on the two forms of analyses were equally weighted. Ninety-percent predictive interval estimates were calculated by averaging upper and lower limits of the 90% predictive intervals calculated under the two models. This procedure is conservative insofar as coverage probabilities exceed the nominal unless the predictions are perfectly positively associated (in which case they agree).

Cross-validation analyses. Leave-one-out cross-validation was used to select and evaluate models. This allows for choice among tree and linear discriminant models of varying complexities on the basis of ability to predict samples not used to fit the models. Here, both variable/"feature" selection (the choice of which subset of the large list of potential expression and clinical measures to include in the model) and estimation (the determination of how to include those chosen) were cross-validated. For each model, out-of-sample predictive probabilities were thresholded at 1/2 to determine predictive classifications and out-of-sample classification accuracy was defined as the fraction of predicted and true classifications in agreement. Tree and linear discriminant models with best out-of-sample classification accuracy were chosen for inclusion in the predictive model. The two-sided Wilcoxon rank sum test for two sample comparisons was used in descriptive analyses. Fold change was calculated as 2^d, where d is mean (log 2 scaled) expression in the comparison group minus that in the baseline group; a fold change of 1 corresponds to no difference, 2 to a doubling, etc. Unless otherwise specified, short-term survivors serve as baseline.

External validation. Gene expression data recently became available from an independent set of ovarian cancers (20) and this was used to perform an external validation of the genes that comprise our linear discriminant model for survival. The data were organized into two classes and a Kaplan-Meier curve generated to test the difference in outcome between the two classes as previously described (21). Genes used in our survival model were mapped from U133 microarrays to U95Av2 using the "best match" provided by Affymetrix (http://www.netaffx.com). Expression data for the matching U95Av2 feature sets for all 68 tumors were loaded into Cluster (22) and processed for hierarchical clustering (no additional filtering, genes and arrays were median centered thrice, genes and arrays were normalized once). Genes and arrays were clustered while calculating weights using the uncentered complete correlation metric and displayed using Treeview (http://rana.lbl.gov/EisenSoftware.htm). The two major classes were identified and assigned class membership (0 or 1) and, together with their observed outcome (recurrence or censorship) and time to outcome (provided by S. Cannistra and D. Spentzos), were used to generate a Kaplan-Meier plot and a log-rank statistical using Prism statistical software (GraphPad, San Diego, CA).

Results

Demographic and clinical features of the 54 women with advanced stage (III/IV) ovarian cancer and of 11 with early-stage (I/II) ovarian cancer are shown in Table 1. Among the advanced stage cases, the short-term survivors had less favorable clinical features including somewhat higher grade and substage and a lower frequency of optimal debulking (largest residual tumors after primary surgery less than 1 cm in diameter). There was no difference in age at diagnosis, but preoperative CA125 levels were somewhat higher among short-term survivors ($P = 0.06$).

Gene expression data were generated from each cancer using Affymetrix GeneChip arrays. The microarray data were initially screened to discard those genes that did not vary or were expressed at low levels, leaving 7,821 for predictive model development. A table of these genes, ordered by association with survival category, their fold changes between long-term and short-term survivors, and full descriptions, appears in Supplementary Material. In addition, clinical features including age at diagnosis, pretreatment CA125 level, and a binary indicator of the maximal size of the largest residual tumor nodule after primary surgery (optimal versus suboptimal debulking) were available for predictive modeling.

Utilization of classification and regression tree models to distinguish short-term and long-term survival. Our previous work has described the use of classification and regression trees together with Bayesian analysis to build predictors of recurrence in breast cancer (7). The most effective of these models makes

---

Table 1. Demographic and clinical characteristics of women with serous ovarian cancer

	Stage III/IV		Stage I/II (*n* = 11)
	Long-term survivors (*n* = 24)	Short-term survivors (*n* = 30)	
	n (%)	*n* (%)	*n* (%)
Race			
White	21 (88%)	21 (70%)	11 (100%)
Black	2 (8%)	7 (23%)	—
Other	1 (4%)	2 (7%)	—
Median age (y)	62	59	52
Median CA 125 level (range)	370 (21-11,710)	1,422 (50-16,510)	105 (13-741)
Stage			
IIIA	1 (4%)	—	
IIIB	2 (8%)	1 (3%)	IA, 5 (46%)
IIIC	21 (88%)	23 (77%)	IC, 2 (18%)
IV	—	6 (20%)	IIC, 4 (36%)
Histologic grade			
Well differentiated	1 (4%)	1 (9%)	
Moderately differentiated	15 (63%)	17 (57%)	5 (45.5%)
Poorly differentiated	8 (33%)	13 (43%)	5 (45.5%)
Optimal debulking	14 (58%)	11 (37%)	

Fig. 1. Genes that are differentially expressed between short-term and long-term ovarian cancer survivors. Expression intensity is normalized to range from 0 to 1. High gene expression is represented in yellow; low gene expression is represented in blue with intermediate levels of expression represented in shades of dark yellow through green to light blue. Genes are plotted in rows; expression values for individual tumors are plotted in columns. Genes are ordered by increasing correlation with survival from bottom to top. Genes above the black horizontal line have positive correlation whereas those below have negative correlation. Strength of association increases moving away from the horizontal line to the top and bottom of the plot. Columns correspond to tumors, with those from short-term survivors depicted to the left of the vertical black line and those from long-term survivors to the right.
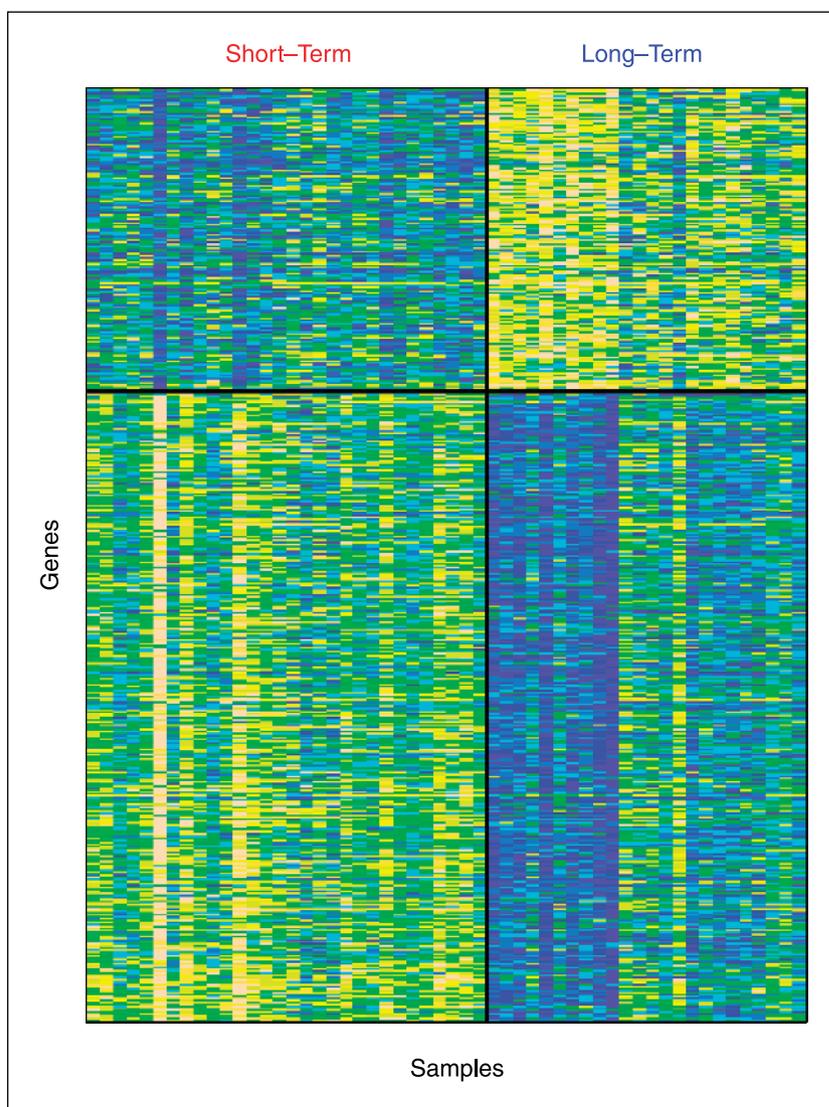
use of a combination of clinical and gene expression data. We have thus used a similar approach to building a classifier that could distinguish long-term and short-term survival in ovarian cancer.

For this purpose, genes with absolute values of correlation to survival greater than 0.4 were used, netting 478 genes from the pool of 7,821 screened genes (Fig. 1). These genes were then used to build a classifier/predictor that could accurately distinguish short-term and long-term survivors, making use of classification and regression tree analysis methods as we have previously described (18). The best tree models involved expression of two or three genes (and no clinical measures) and achieved 90.7% out-of-sample predictive accuracy, correctly classifying 20 of 24 (83.3%) long-term survivors and 29 of 30 (96.7%) short-term survivors. An example tree from these analyses (Fig. 2A) stratifies the patient population into three groups according to expression of cleavage stimulation factor subunit 3 (*CSTF3*) and ATP-binding cassette, subfamily D, member 3 (*ABCD3*). These predictors appeared in the tree with the highest posterior probability in 98% of cross-validation analyses. The gene expression–based tree analysis was compared with the most highly accurate tree models involving only

the clinical variables, including CA125. These trees, one for each left-out sample, achieved 67% cross-validation accuracy (63% among short-term survivors, 71% among long-term survivors) and involved only debulking status and age at diagnosis. An example of a tree that uses only clinical data is depicted in Fig. 2B. Clearly, the ability to distinguish long-term and short-term survival was much improved by the gene expression data.

In examining gene expression in the early-stage (I/II) cancers, the long-term and short-term survivors were used as a training set and the early-stage cases as a test set. This analysis classifies all early-stage cases as long-term survivors with an associated 93.6% average predicted probability. The tree with the highest posterior probability based on the training set (Fig. 2C) divides the training cases into four groups according to expression of pyruvate dehydrogenase complex (*PDX1*), mRNA clone DKFZp762G207 (*DKFZ*), and translocon-associated protein α (*TAPα*).

Expression summaries for the genes used in the clinical and gene-based trees involved in the cross-validation predictions of the long-term and short-term survivors are presented in Table 2. This table includes fold differences in expression of these genes
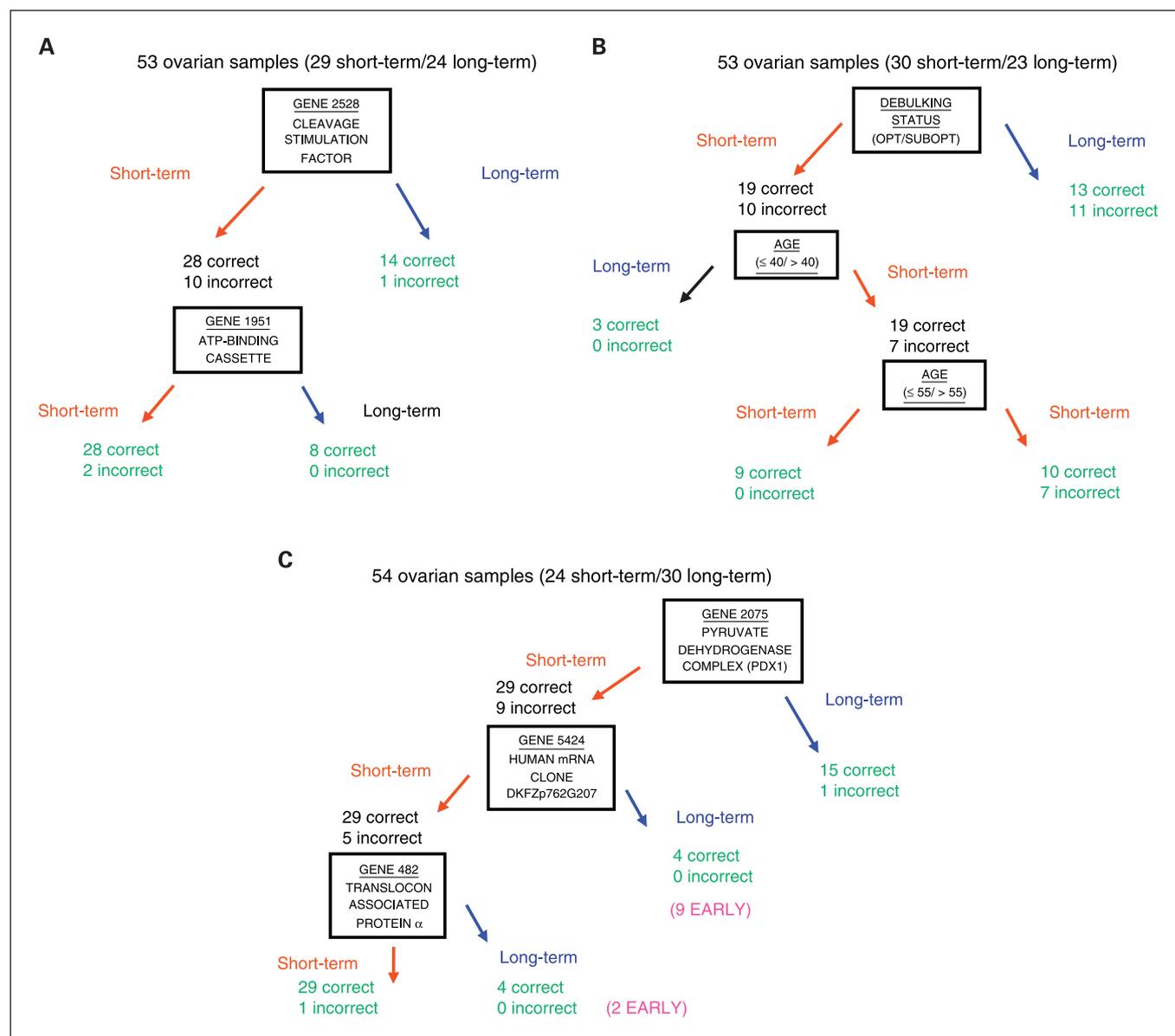
A 53 ovarian samples (29 short-term/24 long-term)

GENE 2528
CLEAVAGE
STIMULATION
FACTOR

Short-term → 28 correct / 10 incorrect

Long-term → 14 correct / 1 incorrect

GENE 1951
ATP-BINDING
CASSETTE

Short-term → 28 correct / 2 incorrect

Long-term → 8 correct / 0 incorrect

B 53 ovarian samples (30 short-term/23 long-term)

DEBULKING
STATUS
(OPT/SUBOPT)

Short-term → 19 correct / 10 incorrect

Long-term → 13 correct / 11 incorrect

AGE
(≤ 40/ > 40)

Long-term → 3 correct / 0 incorrect

Short-term → 19 correct / 7 incorrect

AGE
(≤ 55/ > 55)

Short-term → 9 correct / 0 incorrect

Short-term → 10 correct / 7 incorrect

C 54 ovarian samples (24 short-term/30 long-term)

GENE 2075
PYRUVATE
DEHYDROGENASE
COMPLEX (PDX1)

Short-term → 29 correct / 9 incorrect

Long-term → 15 correct / 1 incorrect

GENE 5424
HUMAN mRNA
CLONE
DKFZp762G207

Short-term → 29 correct / 5 incorrect

Long-term → 4 correct / 0 incorrect
(9 EARLY)

GENE 482
TRANSLOCON
ASSOCIATED
PROTEIN α

Short-term → 29 correct / 1 incorrect

Long-term → 4 correct / 0 incorrect
(2 EARLY)

Fig. 2. Classification tree models that predict survival. *A,* gene expression – based tree analysis for prediction of short-term versus long-term survival. High cleavage stimulation factor expression is associated with 95% predicted probability of long-term survival. Cases with low cleavage stimulation factor expression are further stratified by ATP-binding cassette expression (high expression—88.8% predicted long-term survival; low expression—9.7% predicted long-term survival). *B,* clinical tree for prediction of short-term versus long-term survival. This tree divides the patient population into four groups, each with a distinct predictive long-term survival probability. It first divides with respect to debulking status: those optimally debulked from the first group, where 60.2% are predicted to be long-term survivors. Next, it subdivides the suboptimally debulked group twice by age: the first subdivision isolates "early-onset" cases (age ≤ 40; 94.2% predicted long-term survivors); the second separates "late-onset premenopausal" cases (40 < age ≤ 55; 7% predicted long-term survivors) from "late-onset postmenopausal" cases (age > 55; 45.4% predicted long-term survival). *C,* gene expression – based tree analysis to predict early-stage ovarian cancers. Nine early-stage cases had low levels of pyruvate dehydrogenase complex but high levels of DKFZ, giving a 92.9% predicted probability of long-term survival. The remaining two early cases have low levels of both pyruvate dehydrogenase complex and DKFZ but high levels of translocon-associated protein α, with an associated 96.9% predicted probability of long-term survival.

in long-term versus short-term survivors and early-stage cancers versus short-term survivors.

Linear discriminant models. Although the tree-based models did well in discriminating long-term and short-term survivors, the models were not completely effective. As such, we have examined other methods for analysis. To complement the tree analysis, the strength of which is to identify abruptly nonlinear associations between predictors and outcome, a variant of linear discriminant analysis was employed to identify parsimonious multivariate discriminant functions given the large pool of potential candidate expression measures (23).

The best multivariate linear discriminant functions correctly classified 19 of 24 (79.2%) long-term survivors and 27 of 30 (90%) short-term survivors, achieving 85.2% out-of-sample classification accuracy. These classifiers also predicted all 11 early-stage cases to be long-term survivors. The discriminant functions involved 186 genes across leave-one-out analyses and reflected models allowing for simple conditional dependencies among gene expression measures given survival category (the case $P = 2$). The top five genes of representation were clone T-cell differentiation protein (*MAL*), APMCF1 protein (*APMCF1*), diphosphoinositol polyphosphate phosphohydrolase type 2

(*NUDT4*), plakophilin 4 (*PKP4*), and signal sequence receptor (*SSR1*). Estimates of fold difference in expression of these genes and those appearing in 15 or more leave-one-out models between short-term and long-term survivors with 95% confidence intervals (CI) are presented in Table 2. The *MAL* gene, which was the most highly up-regulated in short-term survivors (3-fold), was included in 54 of 54 linear discriminant models used in leave-one-out analyses. None of the other genes was used in all 54 models. *MAL* was 29-fold higher in short-term survivors compared with early-stage cases.

Combined models. The fact that the tree-based model and the linear discriminant model identified different genes as predictors of long-term and short-term survival suggested the possibility that they addressed different aspects of the underlying biology and, thus, that a combination of the two models would be more effective than either alone. The best tree and discriminant function predictions were combined through a simple predictive model average. This composite model correctly classified 20 of 24 (83.3%) long-term survivors and 29 of 30 (96.7%) short-term survivors and reached an overall

Table 2. Genes used in tree and linear discriminant models of short-term versus long-term survival

Gene name	Affymetrix ID	Gene symbol	Fold difference Long/Short (95% CI)	P	No. times in Top 5 trees	Top 5 LDM	Fold difference Early/Short (95% CI)	P
Cleavage stimulation factor subunit 3	203947_at	CSTF3	1.33 (1.17-1.51)	0.00004	54	0	0.68 (0.57-0.81)	0.00007
ATP-binding cassette, subfamily D (ALD), member 3	202850_at	ABCD3	1.41 (1.17-1.69)	0.0004	53	0	1.33 (1.11-1.60)	0.00253
Histamine *N*-methyltransferase	204112_s_at	HNMT	1.65 (1.30-2.10)	0.0001	2	0	3.79 (2.85-5.05)	<0.000001
Mal, T-cell differentiation protein	204777_s_at	MAL	0.33 (0.19-0.59)	0.00027	0	54	0.03 (0.02-0.07)	<0.000001
APMCF1 protein	218140_x_at	APMCF1	1.39 (1.19-1.63)	0.00009	0	37	0.83 (0.68-1.01)	0.0665
Nudix (nucleoside diphosphate linked moiety X)-type motif 4	212181_s_at	NUDT4	1.62 (1.32-1.99)	0.00002	0	29	0.51 (0.42-0.61)	<0.000001
Plakophilin 4	201928_at	PKP4	1.48 (1.25-1.76)	0.00003	0	29	0.79 (0.55-1.15)	0.215
Signal sequence receptor, α (translocon-associated protein α)	200891_s_at	SSR1	1.63 (1.35-1.96)	0.000004	0	28	5.99 (4.88-7.34)	<0.000001
Protein kinase, cyclic AMP-dependent, regulatory, type I, β	212559_at	Hs.1519	0.74 (0.66-0.83)	0.000004	0	27	1.29 (1.12-1.48)	0.00064
Nudix (nucleoside diphosphate linked moiety X)-type motif 4	206302_s_at	NUDT4	1.72 (1.35-2.19)	0.00004	0	25	1.59 (1.29-1.95)	0.00005
Vesicle-associated membrane protein-associated protein B and C	202550_s_at	VAPB	1.40 (1.22-1.60)	0.00001	0	25	0.37 (0.31-0.45)	<0.000001
Heat shock 27 kDa protein 2	205824_at	HSPB2	0.73 (0.63-0.83)	0.00003	0	23	0.80 (0.64-1.01)	0.0613
Steroid 5 α-reductase	204675_at	SRD5A1	1.69 (1.33-2.14)	0.00005	0	21	0.91 (0.65-1.27)	0.553
Wolfram syndrome 1 (wolframin)	202908_at	WFS1	1.33 (1.14-1.55)	0.00052	0	20	0.61 (0.50-0.75)	0.00002
Connector enhancer of KSR-like (*Drosophila* kinase suppressor of ras)	204740_at	CNK1	0.73 (0.63-0.86)	0.00017	0	20	0.54 (0.41-0.71)	0.00005
Procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase)	202620_s_at	PLOD2	2.32 (1.52-3.55)	0.00022	0	19	0.21 (0.12-0.34)	<0.000001
Plakophilin 4	201929_s_at	PKP4	1.59 (1.32-1.90)	0.000005	0	19	1.99 (1.53-2.59)	<0.000001
Putative RNase III	218269_at	RNASE3L	1.44 (1.22-1.70)	0.00005	0	18	0.91 (0.69-1.19)	0.468
Stress-associated endoplasmic reticulum protein 1	200969_at	—	1.64 (1.28-2.10)	0.0002	0	17	1.26 (0.89-1.80)	0.189
Acyl-protein thioesterase, lysophospholipase II	215566_x_at	LYPLA1	0.71 (0.62-0.81)	0.000002	0	17	0.35 (0.28-0.43)	<0.000001
Tetraspan NET-6 protein	217979_at	NET-6	1.69 (1.27-2.24)	0.00048	0	16	1.03 (0.72-1.45)	0.884
General transcription factor IIH, polypeptide 1	202451_at	GTF2H1	1.53 (1.28-1.82)	0.00001	0	16	3.60 (2.87-4.50)	<0.000001
Embryonic ectoderm development protein	209572_s_at	EED	1.67 (1.37-2.03)	0.000003	0	16	1.02 (0.79-1.32)	0.876
KIAA0624 protein	214734_at	KIAA0624	1.39 (1.17-1.64)	0.00025	0	15	0.09 (0.07-0.10)	<0.000001
ATP-binding protein associated with cell differentiation	203008_x_at	ATP	1.46 (1.23-1.74)	0.00005	0	15	0.84 (0.70-1.02)	0.0772
YY1 transcription factor	200047_s_at	YY1	1.47 (1.27-1.72)	0.000003	0	15	5.82 (4.59-7.38)	<0.000001

Abbreviation: LDM, linear discriminant model.

out-of-sample predictive accuracy of 90.7%. Further, it classified all 11 early-stage cases as long-term survivors.

Figure 3A to C are plots of leave-one-out out-of-sample validations generated by the best tree, discriminant, and composite (hybrid) models. Classifications are obtained by thresholding predictive probabilities at 0.5. The hybrid model combines two structurally different models, each with 90.7% leave-one-out out-of-sample classification accuracy, to form a composite model with 92.6% classification accuracy, suggesting that the tree and discriminant models provide complementary information for accurate classification.

In addition, the models trained on the advanced stage data were used to classify the survival outcomes of the 11 early-stage cases. Both the linear discriminant model and tree models individually classified all of these samples to be long-term survivors as did the hybrid model (Fig. 3). These results suggest

that aspects of the tumor biology of advanced stage cancers associated with long-term survival are shared by early-stage ovarian cancers. Figure 4 presents a plot of out-of-sample point predictions using the combined model. This presents the estimated probability of a sample representing a long-term survivor together with 90% CI for the predictions.

External validation of linear discriminant model. The analysis of ovarian cancers from our institution identified a set of genes that was predictive of short-term versus long-term survival using internal cross-validation techniques. To determine if the same set of genes was associated with prognosis in an independent set of ovarian cancers, we used publicly available data from a recent report (20). Due to a difference in microarray platform, our model could not be directly tested using this data. However, using a previously described methodology (21) we clustered the 68 tumors from the
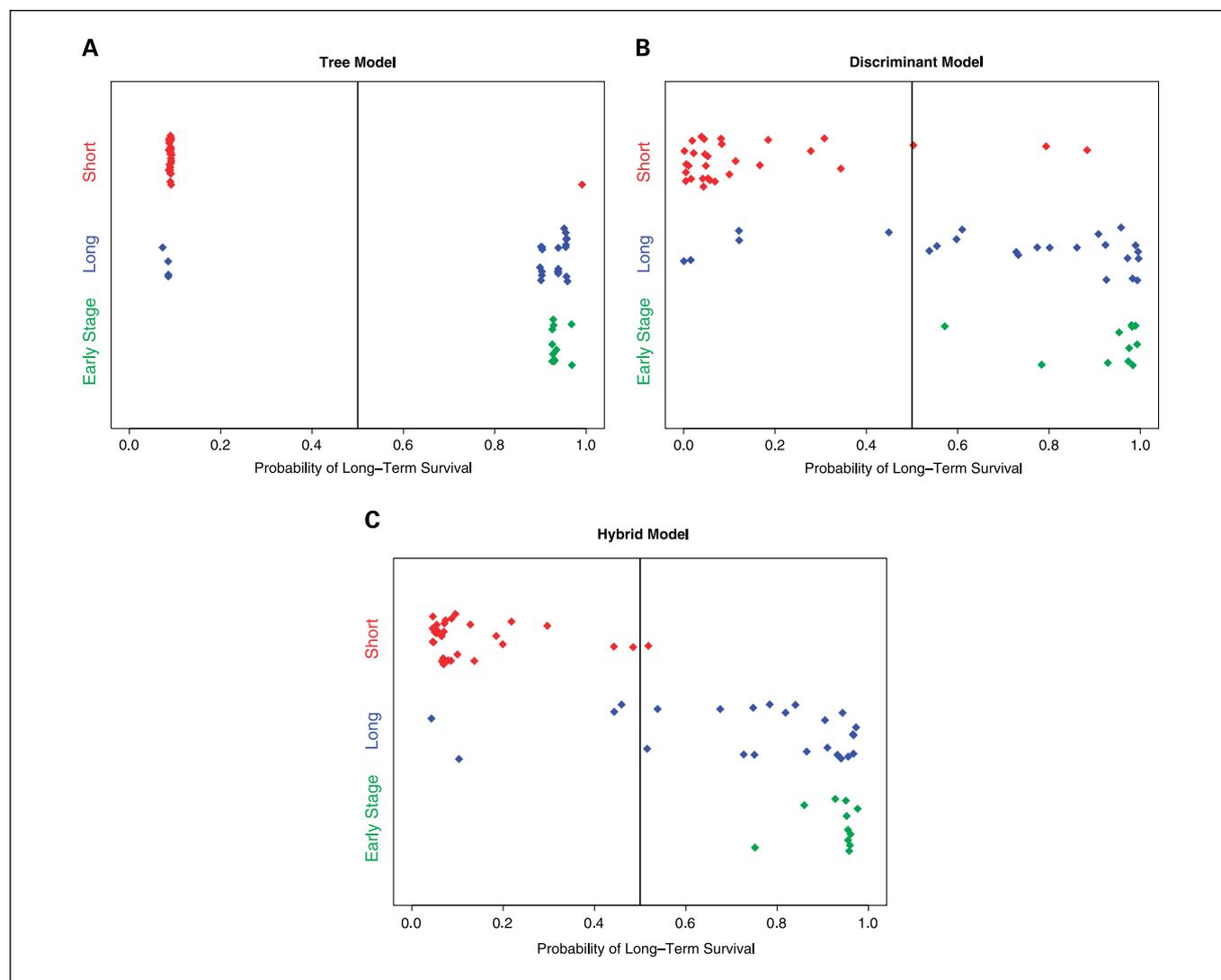

Fig. 3. Cross-validation predictions of survival. Plot of individual leave-one-out out-of-sample predictions (on the *x*-axis) for the 54 advanced stage cancers by actual survival group for the tree (*A*), linear discriminant (*B*), and hybrid models (*C*). The short-term survivors are plotted in red and the long-term survivors are plotted in blue. In addition, out-of-sample predictions for the 11 early-stage tumors are plotted in green (note that these samples were not used in "model training"). Points are jittered in the direction of the *y*-axis to achieve their separation. A vertical line is plotted at 1/2, indicating the classification threshold used for determining out-of-sample predictive accuracy (cases with probability of long-term survival greater than 1/2 were declared long-term survivors; those less than 1/2 as short-term survivors). The predictive accuracy of the hybrid model using a threshold of 1/2 is 90.7%.

Fig. 4. Out-of-sample point predictions with 90% error bars for the hybrid model. Cases are ordered along the x-axis according to the rank of their point prediction, ordered smallest to largest. Short-term survivors are depicted in red, long-term survivors in blue, and early-stage cases in green.
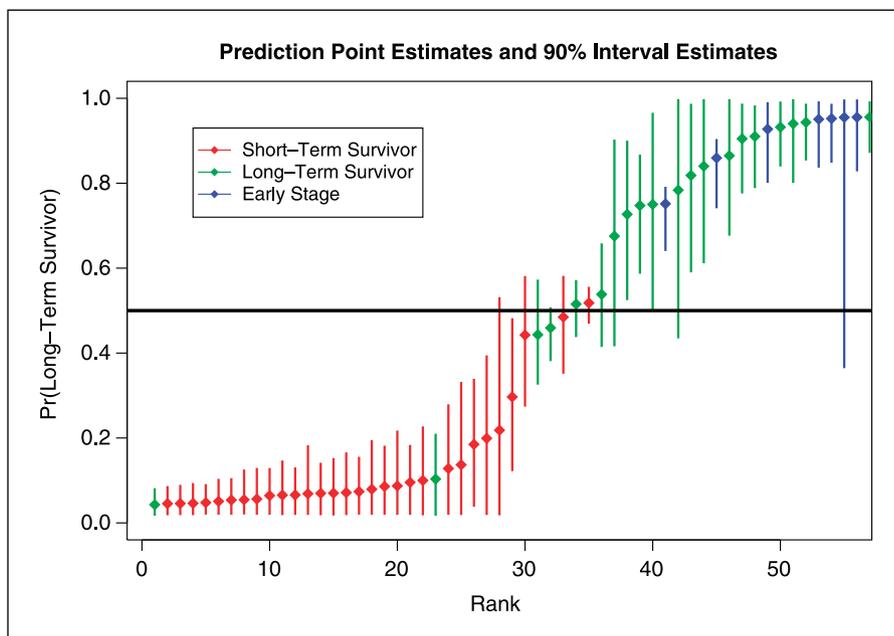
independent set based on the expression of the features representing the same genes identified in our linear discriminant analysis. A significant difference in outcome was found between the two major clusters identified ($P = 0.0074$; Fig. 5). A subset of features including *EED*, *WFS1*, *PKP4*, *SRD5A1*, *KIAA0624*, and *YY1* were again identified as having increased expression in cancers from long-term survivors. Further details of this analysis are available on our website http://data.cgt.duke.edu/clinicalcancerresearch.

Quantitative real-time PCR validation of gene expression. Quantitative real-time PCR was used to validate the differences in expression patterns that were established by microarray analysis. Quantitative real-time PCR was done for *MAL* using 21 short-term and 18 long-term survivors (Fig. 6A). The correlation coefficient was −0.81 (95% CI, −0.89 to −0.66). The linear regression model was highly significant with a P value of <0.0001 and an r^2 value of 0.65, demonstrating agreement between the two methods in quantifying *MAL*

Fig. 5. External validation using the independent data set of Spentzos et al. (20). *A*, dendrogram of 68 tumors based on the expression of features on the U95Av2 microarrays representing genes identified in the linear discriminant model. *B*, Kaplan-Meier plot of survival for the two major classes identified during tumor clustering (survival is in months; $P = 0.0074$).
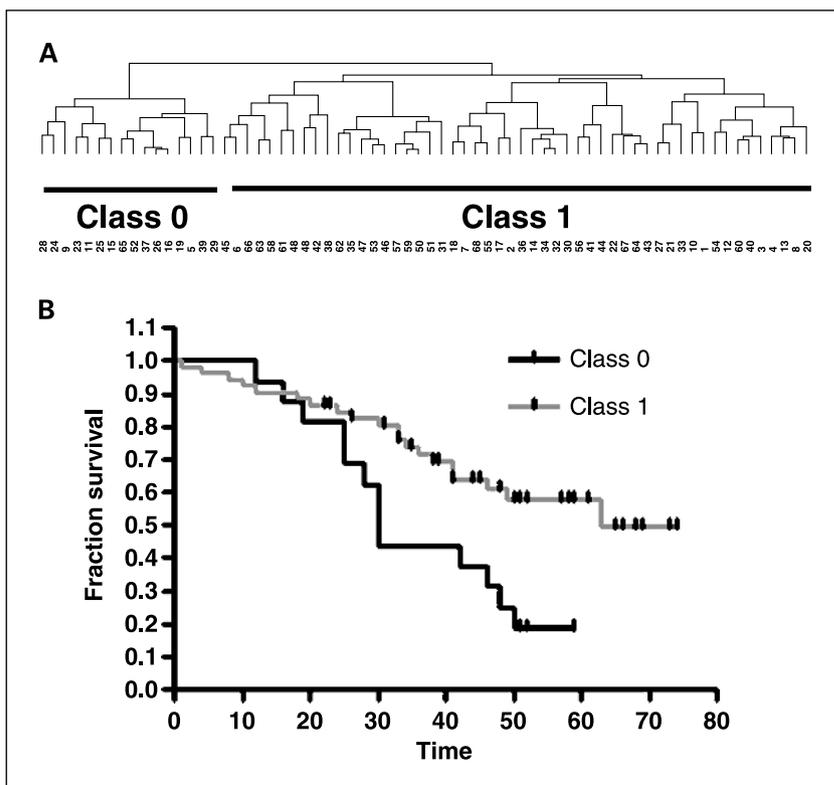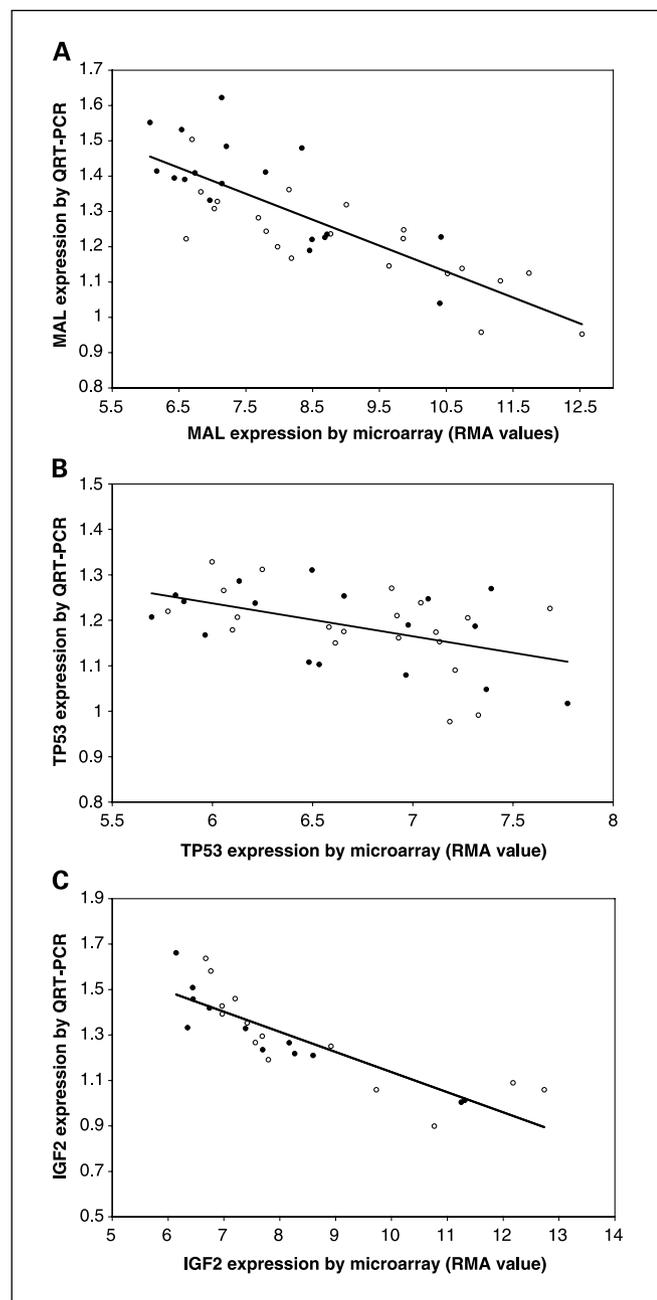
Fig. 6. Validation of differential gene expression between short-term and long-term survivors using quantitative real-time PCR. Linear regression model of serous ovarian tumors showing the relationshp between the expression of *MAL* (*A*), *TP53* (*B*), and *IGF2* (*C*) mRNA by microarray analysis (*x*-axis) versus quantitative real-time PCR (*y*-axis). *Y*-axis values were detrmined by normalizing the cycle of threshold values for each sample to those obtained for β2-microglobulin. ○, short-term survivors; ●, long-term survivors.

and C). Combined, these results indicate a robust statistically significant relationship between microarray expression data values and expression values measured by quantitative real-time PCR.

Discussion

Approximately 10% of epithelial ovarian cancers are attributable to inherited mutations in high-penetrance susceptibility genes (*BRCA1* and *BRCA2*), whereas the majority of cases are sporadic. All ovarian cancers likely arise due to alterations that disrupt molecular pathways involved in regulation of proliferation, apoptosis, and DNA repair, and a number of genes have been identified that are involved in the development of some ovarian cancers (e.g., *TP53*, HER-2/*neu*, and K-*ras*). These studies emphasize that substantial molecular heterogeneity exists between cancers (5). However, analysis of individual candidate genes based on their involvement in pathways related to carcinogenesis is inefficient and has yielded relatively few relevant genes. In addition, none of the genetic alterations described thus far has augmented the ability of conventional clinical staging systems based on the extent of disease to predict survival of women with ovarian cancer.

Microarrays that assess global patterns of gene expression have proven useful in defining and predicting clinical phenotypes in a variety of cancer types (24). This includes studies of breast cancer in which several groups have described gene expression patterns that serve as prognostic tools to define risk of recurrence (8, 25) as well as lymph node involvement (7). The work of our own group has shown the value in utilizing methods of analysis that sample multiple forms of data, both clinical and multiple gene expression patterns, so as to achieve a more precise discrimination and prediction of outcome for individual patients (7, 18). This same logic of utilizing multiple forms of data as well as methods of analysis has been applied in the present study to more accurately achieve a classification and prediction of ovarian cancer survival.

Several groups have applied expression array technology to the analysis of ovarian cancers. Many of these studies have compared gene expression between normal ovarian epithelial cells and ovarian cancers. Numerous genes have been identified that seem to be up- or down-regulated in the process of malignant transformation (26–29). In addition, microarrays have shown patterns of gene expression that distinguish between histologic types (30) and stages (31).

In the present study, we used expression arrays to identify gene expression patterns that reflect patient survival. The groups of patients used represent the extremes with respect to outcome (i.e., those who survived either less than 3 years or greater than 7 years). The availability of frozen tumor samples from a significant number of long-term survivors was a strength of this study as was the use of novel and complementary statistical approaches. The observation that no gene was more than 3-fold differentially expressed emphasizes the power of the gene expression patterns versus the analysis of any single gene. Spentzos et al. (20) recently reported the only other study in which microarrays have been used to predict outcome in ovarian cancer. This study involved a group of 68 patients who were not selected based on survival duration. The investigators used Affymetrix U95 microarrays to develop a 115 gene model that classified cases into unfavorable (median survival, 30 months)

expression. For *TP53*, 37 samples were analyzed by quantitative real-time PCR (20 short-term survivors and 17 long-term survivors) whereas for *IGF2*, 26 samples were analyzed (14 short-term survivors and 12 long-term survivors). The correlation coefficients were −0.47 (95% CI, −0.69 to −0.18) for *TP53* and −0.85 (95% CI, −0.93 to −0.70) for *IGF2*. Linear regression models were again highly significant with *P* values for these associations of 0.0028 for *TP53* and <0.0001 for *IGF2* with r^2 values of 0.2278 and 0.7301, respectively (Fig. 6B

and favorable (median survival not yet reached) groups that exhibited significantly different survival. In this study there was no attempt to predict survival of individual patients, but the results are consistent with ours and together suggest that clinical differences in outcome are reflected in global patterns of gene expression that can be appreciated using microarrays.

In view of the potential for false-discovery using microarrays, external validation using independent sets of samples is a critical step. In this regard, although we could not directly "validate" our linear discriminant model using the Spentzos et al. data set due to differences in microarray platform (20), we did confirm that expression of the genes that comprise our linear discriminant model held prognostic value in this independent set of tumors processed at a different institution using different microarrays. Additional independent validation studies are needed before patterns of gene expression can be used for clinical predictions.

Another clinically relevant finding in our study was that patterns of gene expression in advanced stage cancers from long-term survivors were shared by the early-stage cancers, none of which have been fatal. This provides compelling evidence that the favorable clinical outcome of both long-term survivors with advanced stage disease and early-stage cases is attributable to a shared underlying pattern of molecular alterations. These findings may have implications for screening as an approach to decrease ovarian cancer mortality, as they suggest that cancers diagnosed at an early stage are less rapidly progressive and likely to have a more favorable outcome even if detected at an advanced stage. Conversely, the observation that none of the early-stage ovarian cancers exhibited patterns of gene expression similar to the more virulent cancers from short-term survivors suggests that the most highly lethal ovarian cancers may not be easily amenable to early detection.

Many of the genes that were critical components of the patterns that discriminated between long-term and short-term survivors are known to affect the virulence of the malignant phenotype. The *MAL* gene (T-lymphocyte maturation-associated protein) was the most differentially expressed (3-fold higher in cancers from short-term compared with long-term survivors) and also was up-regulated in short-term survivors relative to early-stage disease (29-fold). *MAL* was included in 54 of 54 linear discriminant models used in leave-one-out analyses, whereas none of the other genes was used in all 54 models. Expression of MAL has been shown previously in ovarian cancers, most notably clear cell and serous cancers (30). The *MAL* protein has several hydrophobic domains and has been shown to be a component of the protein machinery for apical transport in epithelial polarized cells (32, 33). *MAL* also is a component of membrane rafts, which are micro-domains that play a central role in signal transduction acting as a scaffold in which molecules of signal transduction pathways can interact (33). Further, *MAL* has been identified as a gene involved in resistance to cancer therapy. *MAL* was the most differentially expressed gene in microarray experiments that compared IFN-α–sensitive and –resistant cutaneous

T-cell lymphoma cell lines (34). The emergence of IFN-α resistance in this *in vitro* model was accompanied by a 5.5-fold up-regulation of *MAL*. In addition, high *MAL* expression was found to correlate with poor response to therapy in a cohort of patients with cutaneous T-cell lymphoma (34).

Heat shock protein 27 (*HSP27*) is another gene implicated in resistance to therapy that was more highly expressed in cancers from short-term survivors relative to long-term survivors (1.37-fold) and early-stage cases (1.24-fold). *HSP27* is a member of the heat shock protein family of proteins that is normally up-regulated in response to thermal injury and other stresses. Expression of *HSP27* has been shown to augment the ability of cancer cells treated with cytotoxic chemotherapy to evade apoptosis, and this enhances cell survival (35, 36). This apoptotic resistance is thought to be due to inhibition of activity of the apoptosome and caspase activation complex as well as regulation of protea-some-mediated degradation of apoptosis-regulatory proteins. Several groups have found that high *HSP27* expression increases chemoresistance of ovarian cancer cell lines *in vitro* (37, 38) and this can be reversed by treatment with antisense *HSP27* (38). In addition, high expression of *HSP27* correlates with poor survival in several types of cancers including glioma (39), breast cancer (40), and ovarian cancer (38).

Lysophospholipase II (*LYPLA2*), elevated in cancers from patients with short-term survival, may have a biochemical link to the disease process. This is one of a family of enzymes that can convert lysophopholipids to lysophophatidic acid, a molecule commonly elevated in ovarian cancer ascitic fluid (41). Lysophophatidic acid itself has been shown to have a number of biological effects including increasing proliferation, survival, invasiveness, and resistance to cisplatin (42). Ovarian cancer cells have been shown to produce lysophophatidic acid and the increased levels of the *LYPLA2* enzyme may be a significant component of this biosynthesis (41).

Finally, it has been suggested that androgens play a role in both the pathogenesis and growth of ovarian cancers (43). The *SRD5A1* gene, which was more highly expressed in cancers of long-term survivors, is involved in converting testosterone to its active form dihydrotestosterone. Higher levels of *SRD5A1* in ovarian cancers of long-term survivors may reflect maintenance of hormonal responsiveness in ovarian cancers with an inherently less aggressive phenotype, as is the case in breast and prostate cancers.

The value of gene expression–based predictors of prognosis in advanced ovarian cancer will not be fully realized until additional therapies are available for those destined to have poor survival following conventional chemotherapy. In this regard, the expression profiles may not only predict the likelihood of long-term survival following platin chemo-therapy but may also yield clues to individual genes involved in tumor development, progression, and response to therapy. It is likely that some of the most differentially expressed genes, such as those discussed above, will represent appealing therapeutic targets.

References

1. Young RC. Three cycles versus six cycles of adjuvant paclitaxel (Taxol)/carboplatin in early stage ovarian cancer. Semin Oncol 2000;3:8–10S.

2. Markman M, Bundy BN, Alberts DS, et al. Phase III trial of standard-dose intravenous cisplatin plus pacli-taxel versus moderately high-dose carboplatin fol-lowed by intravenous paclitaxel and intraperitoneal cisplatin in small-volume stage III ovarian carcinoma: an intergroup study of the Gynecologic Oncology

Group, Southwestern Oncology Group, and Eastern Cooperative Oncology Group. J Clin Oncol 2001;19: 1001 – 7.

3. McGuire WP, Hoskins WJ, Brady MF, et al. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. N Engl J Med 1996;334:1 – 6.

4. Hoskins WJ, McGuire WP, Brady MF, et al. The effect of diameter of largest residual disease on survival after primary cytoreductive surgery in patients with suboptimal residual epithelial ovarian carcinoma. Am J Obstet Gynecol 1994;170:974 – 9.

5. Boyd JA, Berchuck A. Oncogenes and tumor suppressor genes. In: Hoskins WJ, Perez CA, Young RC, Barakat R, Markman M, Randall M, editors. Principles and Practice of Gynecologic Oncology. Lippincott: Williams and Wilkins; 2005. p. 93 – 122.

6. Havrilesky L, Hamdan H, Darcy K, Leon J, Bell J, Berchuck A. Relationship between p53 mutation, p53 overexpression and survival in advanced ovarian cancers treated on Gynecologic Oncology Group studies #114 and #132. J Clin Oncol 2003;21: 3814 – 25.

7. West M, Blanchette M, 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.

8. Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003; 361:1590 – 6.

9. 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.

10. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4: 249 – 64.

11. Ihaka R, Gentleman R. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299 – 314.

12. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185 – 93.

13. Breiman L. Statistical modeling: the two cultures. Stat Sci 2001;16:199 – 225.

14. Chipman H, George E, McCulloch RE. Bayesian CART model search. J Am Stat Assoc 1998;93: 935 – 48.

15. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002; 97:77 – 87.

16. Silva DPA. Efficient variable screening for multivariate analysis. J Multivariate Anal 2001;76:35 – 62.

17. Simon R, Desper R, Papadimitriou C, et al. Chromosome abnormalities in ovarian adenocarcinoma: III. using breakpoint data to infer and test mathematical models for oncogenesis. Genes Chromosomes Cancer 2000;28:106 – 20.

18. Boulestriex A-L, Tutz G, Strimmer K. A CART-based approach to discover emerging patterns in microarray data. Bioinformatics 2003;19:2465 – 72.

19. Pittman J, Huang E, Nevins JR, Wang Q, West M. Bayesian analysis of binary prediction tree models. Biostatistics 2004;5:587 – 601.

20. Spentzos D, Levine DA, Ramoni MF, et al. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. J Clin Oncol 2004;22:4648 – 58.

21. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33:49 – 54.

22. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression. Proc Natl Acad Sci U S A 1998;95: 14863 – 8.

23. Wermuth N. Linear recursive equations, covariance selection, and path analysis. J Am Stat Assoc 1980; 75:963 – 72.

24. Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol 2002;20:1932 – 41.

25. 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.

26. Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A 2001;98:1176 – 81.

27. Ono K, Tanaka T, Tsunoda T, et al. Identification by cDNA microarray of genes involved in ovarian carcinogenesis. Cancer Res 2000;60:5007 – 11.

28. Lancaster JM, Dressman HK, Whitaker RS, et al. Gene expression patterns that characterize advanced stage serous ovarian cancers. J Soc Gynecol Investig 2004;11:51 – 9.

29. Schummer M, Ng WV, Bumgarner RE, et al. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 1999;238:375 – 85.

30. Schwartz DR, Kardia SL, Shedden KA, et al. Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. Cancer Res 2002;62:4722 – 9.

31. Shridhar V, Lee J, Pandita A, et al. Genetic analysis of early- versus late-stage ovarian tumors. Cancer Res 2001;61:5895 – 904.

32. Rancano C, Rubio T, Correas I, Alonso MA. Genomic structure and subcellular localization of MAL, a human T-cell-specific proteolipid protein. J Biol Chem 1994;269:8159 – 64.

33. Alonso MA, Milan J. The role of lipid rafts in signalling and membrane trafficking in T lymphocytes. J Cell Sci 2001;114:3957 – 65.

34. Tracey L, Villuendas R, Ortiz P, et al. Identification of genes involved in resistance to interferon-α in cutaneous T-cell lymphoma. Am J Pathol 2002;161: 1825 – 37.

35. Concannon CG, Gorman AM, Samali A. On the role of Hsp27 in regulating apoptosis. Apoptosis 2003; 8:61 – 70.

36. Garrido C, Schmitt E, Cande C, Vahsen N, Parcellier A, Kroemer G. HSP27 and HSP70: potentially oncogenic apoptosis inhibitors. Cell Cycle 2003;2: 579 – 84.

37. Yamamoto K, Okamoto A, Isonishi S, Ochiai K, Ohtake Y. Heat shock protein 27 was up-regulated in cisplatin resistant human ovarian tumor cell line and associated with the cisplatin resistance. Cancer Lett 2001;168:173 – 81.

38. Langdon SP, Rabiasz GJ, Hirst GL, et al. Expression of the heat shock protein HSP27 in human ovarian cancer. Clin Cancer Res 1995;1:1603 – 9.

39. Iwadate Y, Sakaida T, Hiwasa T, et al. Molecular classification and survival prediction in human gliomas based on proteome analysis. Cancer Res 2004;64: 2496 – 501.

40. Vargas-Roig LM, Gago FE, Tello O, Aznar JC, Ciocca DR. Heat shock protein expression and drug resistance in breast cancer patients treated with induction chemotherapy. Int J Cancer 1998;79:468 – 75.

41. Fang X, Schummer M, Mao M, et al. Lysophosphatidic acid is a bioactive mediator in ovarian cancer. Biochim Biophys Acta 2002;1582:257 – 64.

42. Hu YL, Albanese C, Pestell RG, Jaffe RB. Dual mechanisms for lysophosphatidic acid stimulation of human ovarian carcinoma cells. J Natl Cancer Inst 2003;95:733 – 40.

43. Risch HA. Hormonal etiology of epithelial ovarian cancer, with a hypothesis concerning the role of androgens and progesterone. J Natl Cancer Inst 1998;90:1774 – 86.