[CANCER RESEARCH 61, 7388–7393, October 15, 2001]

Advances in Brief

Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures[1]

Andrew I. Su, John B. Welsh, Lisa M. Sapinoso, Suzanne G. Kern, Petre Dimitrov, Hilmar Lapp, Peter G. Schultz, Steven M. Powell, Christopher A. Moskaluk, Henry F. Frierson, Jr., and Garret M. Hampton[2]

Department of Chemistry, The Scripps Research Institute, La Jolla, California 92037 [A. I. S., P. G. S.]; Genomics Institute of the Novartis Research Foundation, San Diego, California 92121 [J. B. W., L. M. S., S. G. K., P. D., H. L., P. G. S., G. M. H.]; and Departments of Medicine [S. M. P.] and Pathology [C. A. M., H. F. F.], University of Virginia Health System, Charlottesville, Virginia 22908

Abstract

Classification of human tumors according to their primary anatomical site of origin is fundamental for the optimal treatment of patients with cancer. Here we describe the use of large-scale RNA profiling and supervised machine learning algorithms to construct a first-generation molecular classification scheme for carcinomas of the prostate, breast, lung, ovary, colorectum, kidney, liver, pancreas, bladder/ureter, and gastroesophagus, which collectively account for ~70% of all cancer-related deaths in the United States. The classification scheme was based on identifying gene subsets whose expression typifies each cancer class, and we quantified the extent to which these genes are characteristic of a specific tumor type by accurately and confidently predicting the anatomical site of tumor origin for 90% of 175 carcinomas, including 9 of 12 metastatic lesions. The predictor gene subsets include those whose expression is typical of specific types of normal epithelial differentiation, as well as other genes whose expression is elevated in cancer. This study demonstrates the feasibility of predicting the tissue origin of a carcinoma in the context of multiple cancer classes.

Introduction

Effective treatment of cancer patients fundamentally depends on knowledge of the primary anatomical site of tumor origin. Thus, classification of human cancers into distinct groups based on their tissue of origin and histopathological appearance is important for optimal patient management. The use of biological reagents, particularly antibodies for detecting specific tumor antigens by IHC,[3] has contributed significantly toward improving cancer diagnosis and treatment. It is estimated that ~4% of all patients diagnosed with cancer present with metastatic tumors for which the origin of the primary tumor has not been determined (1). On occasion, the primary site for a metastatic tumor is not clearly apparent even after pathological analysis. Thus, predicting the primary tumor site of origin for some of these cancers represents an important clinical objective. We have constructed a first-generation molecular classification scheme for carcinomas of the prostate, breast, colorectum, lung (adenocarcinoma and squamous cell carcinoma), liver, gastroesophagus, pancreas, ovary, kidney, and bladder/ureter, which collectively account for ~70% (~400,000 cases) of all cancer-related deaths in the United States (2). The gene expression signatures discovered by our classification approach include novel tumor-related genes whose encoded proteins may lead to new clinical reagents for successful tumor diagnosis.

Materials and Methods

Tumor Samples. An initial set of 100 primary carcinomas was used for the development of our classification scheme ("training set"). This set of tumors comprised 10 prostate adenocarcinomas, 9 bladder/ureter carcinomas (8 transitional cell carcinomas and 1 squamous cell carcinoma), 10 infiltrating ductal breast adenocarcinomas, 10 colorectal adenocarcinomas, 11 gastroesophageal adenocarcinomas, 11 clear cell carcinomas of the kidney, 6 hepatocellular carcinomas, 10 serous papillary ovarian adenocarcinomas, 6 pancreatic adenocarcinomas, and 17 lung carcinomas (9 adenocarcinomas and 8 squamous cell carcinomas). The set of 75 blinded tumor samples ("test set") included 63 primary tumors and 12 metastatic lesions. The primary tumor samples were 9 lung cancers (4 adenocarcinomas and 5 squamous cell carcinomas), 9 colorectal adenocarcinomas, 13 infiltrating ductal breast adenocarcinomas, 14 prostate adenocarcinomas, 15 papillary serous ovarian carcinomas, 1 hepatocellular carcinoma, and 2 gastroesophageal carcinomas. Metastatic tumors included those arising in the colorectum, ovary, breast, lung, prostate, and kidney. More detailed descriptions of our ovarian and prostate cancer collections have been reported (3, 4). A detailed description of the tumors used in this study is available from our website.[4] The University of Virginia Human Investigation Committee approved the use of the human tissue samples obtained from the University of Virginia. Each specimen was assessed by H&E frozen section examination, and areas rich in tumor were cut from the frozen blocks prior to RNA extraction. Care was taken to avoid as much as possible nonneoplastic epithelium within the tumor samples. Hence, the samples used in this study consisted predominantly of neoplastic cells.

Microarray Hybridization. RNA extraction and hybridization on oligonucleotide microarrays (U95a GeneChip; Affymetrix Incorporated, Santa Clara, CA) was performed as described (4), with the exception that the arrays were hybridized at 50°C for 16–20 h. GeneChip hybridization data were processed and scaled as described (5, 6). We included only those probe sets (9198) whose maximum hybridization intensity (AD) in at least one sample was >200; the other probe sets were excluded (the quantification of gene transcripts with AD values uniformly <200 are typically unreliable). All AD values <20, including negative AD values, were raised to a value of 20, and the data were log transformed. The primary hybridization data are available from our website.[4]

Cancer Classification and Cancer Class Prediction. For each of the 9198 genes that passed the minimal expression threshold, a Wilcoxon rank score (7) was calculated for the group with the highest mean expression *versus* samples from all other groups (implemented in Matlab version 6.0). The 100 genes with the lowest *P*s in each class (total, 1100 genes) were ranked based on their predictive accuracy for discriminating one tumor class *versus* all others using a SVM classifier (8). Specifically, genes were ranked based on their LOOCV accuracy (9). In LOOCV for a given gene, we blinded ourselves to one sample, trained an SVM using the remaining samples, and used the SVM to predict the class identity of the blinded sample (either cancer class *X*, or not cancer class *X*). This process was repeated for all samples in the training set, and an overall prediction accuracy was calculated for each gene. The SVM procedure used

Received 7/23/01; accepted 8/31/01.

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.

[1] Andrew I. Su acknowledges the ARCS Foundation of San Diego and the La Jolla Interfaces in Science Program for predoctoral support.

[2] To whom requests for reprints should addressed, at Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121. Phone: (858) 812-1522; Fax: (858) 812-1746; E-mail: garret_hampton@yahoo.com.

[3] The abbreviations used are: IHC, immunohistochemistry; AD, average difference; SVM, support vector machine; LOOCV, leave-out-one cross-validation; WT, Wilms' tumor.

[4] www.gnf.org/cancer/epican.

here[5] was implemented in the software package R v1.2.2.4. The voting scheme used the 10 genes with the highest SVM/LOOCV accuracy from each class (110 total genes across 11 tumor classes). For each class, a minimum SVM/LOOCV accuracy threshold was set such that at least 10 genes passed; because in each class multiple genes have equivalent accuracy, 216 genes were selected from the 11 classes and were iteratively bootstrapped to obtain an equal number (*i.e.,* 10) of voting genes per class (10). For classifying an unknown sample, prediction scores were calculated using one set of 110 genes (calculated as described below), and final predictions were based on averaged scores over 50 iterations. Hybridization values for our 110-gene predictor set were compared to each sample in our training set. An L1 distance (sum of absolute differences) from the unknown sample to each training sample was calculated. The "class distance" was defined as the mean distance from the unknown sample to the members of that class in the training set. The class to which an unknown sample has the lowest class distance is the predicted identity. A Dixon test for outliers was used to assign a confidence score to each prediction. The Dixon metric is calculated by sorting the vector of mean distances, where $X_i < X_{i+1}$, and computing the value $D = (X_2 - X_1)/(X_n - X_1)$ (11). A Dixon threshold of $D = 0.1$ was empirically set as a conservative boundary for high confidence predictions.

Tissue Microarrays and IHC. Tissue microarrays containing 0.6-mm cores from 265 different zinc formalin-fixed, paraffin-embedded specimens were constructed using a Tissue Microarrayer (Beecher Instruments, Silver Spring, MD). Samples consisted of 36 normal adult epithelial tissues and 229 carcinomas, which included most of the tumors whose transcripts were profiled in the study. Ovarian cancers were profiled as described previously (3), and 16 other independent serous papillary carcinomas of the ovary were included in the tissue microarrays. For IHC on the tissue microarrays and on a whole-tissue section of a normal ovary, the avidin-biotin immunoperoxidase method was performed. After slides had been placed in a citrate buffer and treated with microwave heat for 20 min, the polyclonal anti-WT antibody (C-19; 1:100 dilution; Santa Cruz Biotechnology, Santa Cruz, CA) was applied for 1 h at room temperature. Nuclear immunoreactivity was considered to represent true positivity.

Results and Discussion

We reasoned that a combination of molecular features characteristic of a neoplasm's epithelium of origin, as well as consistent molecular alterations that underlie specific neoplastic phenotypes, might be sufficient to predict the class of an unknown carcinoma; thus, we sought to develop a multiclass molecular classification scheme based on genes whose expression was specific to tumor tissues of each anatomical site. To obtain sufficient data necessary to develop the classification method, we hybridized total RNA from a series of 100 carefully prepared primary tumors from 10 diverse tissue origins (referred to as the training set) on Affymetrix oligonucleotide microarrays containing probe sets for 12,533 genes. We chose primary carcinomas from each anatomical site that represented the most commonly diagnosed histopathologies (*e.g.,* for kidney, we selected predominantly clear cell carcinomas, and for ovary we selected papillary carcinomas of the ovary; see "Materials and Methods" and supplementary information on our website).[4]

Initial analysis of the data by methods that group similarly expressed genes, as well as tumors with similar gene expression (*i.e.,* unsupervised hierarchical clustering; Ref. 12), showed that we could readily group cancers of some anatomical sites, such as those of the prostate and kidney, based solely on the patterns of the most variably expressed genes. In contrast, we found a high degree of similarity between cancers of the colorectum, stomach, bladder/ureter, and lung, making their histological separation difficult on the basis of unsupervised clustering (data not shown; available as supplementary Fig. 1 on our website).[4] We therefore divided the process of multiclass prediction into three components: (*a*) filtering the large data set of gene expression (12,533 genes in 100 tumors; $>1.25 \times 10^6$ data points) to exclude those genes that do not contribute to tumor distinction; (*b*) ranking potentially predictive genes to identify the most accurate tumor-specific classifiers; and (*c*) determining an optimal method by which these genes could be used to "vote" for the likely class of a blinded tumor sample in the context of multiple tumor classes.

We first sought to "prefilter" the data set to identify genes with uniformly high expression among carcinomas of a specific anatomical site and uniformly low expression among carcinomas of all other anatomical sites or histopathologies (Fig. 1*A*). This was achieved using the Wilcoxon rank-sum test (7), which tests the null hypothesis that gene expression in one tumor class is not different from gene expression in any other tumor class. The genes in each class that had significant *P*s represented those that disputed the null hypothesis and defined those that were most different among tumor classes. For carcinomas of all of the anatomical sites that were examined, we were able to identify many such genes (Fig. 1*B*). One hundred of the Wilcoxon-selected genes from each tumor class were then subjected to a "prediction accuracy test." Each of the genes was tested for its ability to discriminate one tumor class from all other tumor classes, using a SVM-learning algorithm (Ref. 8; see "Materials and Methods"), which has been shown to yield good results in gene expression-based classification problems (Ref. 13; Fig. 1). LOOCV was used to blind ourselves sequentially to each of the 100 tumor samples; the SVM was trained on the remaining samples and then used to predict the class of the blinded specimens (9, 10). This test identified >10 genes per tumor class that could predict the class of a blinded tumor in at least 91% of cases (such as lung adenocarcinomas and bladder/ureter carcinomas; Fig. 1*C*). Typically, the accuracy of the classifier genes was higher, ranging up to 100% (*e.g.,* prostate carcinoma; Fig. 1*C*). A voting scheme was developed based on calculating a class distance, by which we could evaluate how molecularly related an unknown sample was to tumors of different classes (see "Materials and Methods"). We also used a confidence score to estimate the strength of each prediction and experimentally determined a confidence threshold that minimized tumor misclassification. Empirically, we determined that a small group of 110 genes, representing 10 genes per tumor class, most accurately predicted the origin of a blinded tumor sample (see "Materials and Methods"). The complete list of genes comprising the multiclass predictor is available as supplementary Table 1 from our website.[4]

Using these optimized parameters, the performance of the classification method was first assessed by predicting the anatomical site of tumor origin for each of the 100 carcinomas in the training set by cross-validation (*i.e.,* LOOCV). Confident predictions were made for 94/100 (94%) of the samples, of which 92 (98%) were correct. The two misclassified cases were a hepatocellular carcinoma and a squamous cell carcinoma of the lung. The 6 cases that did not pass the confidence threshold included 3 gastroesophageal carcinomas, 1 bladder/ureter carcinoma, 1 lung adenocarcinoma and 1 lung squamous cell carcinoma. Of these cases, 5 were actually correctly predicted, but with low confidence. One of the gastroesophageal carcinomas was incorrectly predicted as a lung adenocarcinoma (Table 1; also see supplementary Table 2 on our website).[4] Therefore, in the absence of the confidence threshold, we correctly predicted an anatomical origin of 97 of 100 (97%) tumors. Anecdotally, one of the tumor samples that was originally part of our training set and was labeled as a hepatocellular carcinoma was strongly predicted as a colon cancer by cross-validation (data not shown). Histological reevaluation of this poorly differentiated neoplasm revealed some minor differences between the frozen tissue section representative of the material that we had profiled and a paraffin-embedded tissue section from the patient's tumor. DNA was isolated from both frozen and paraffin-embedded

[5] Developed by E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel.

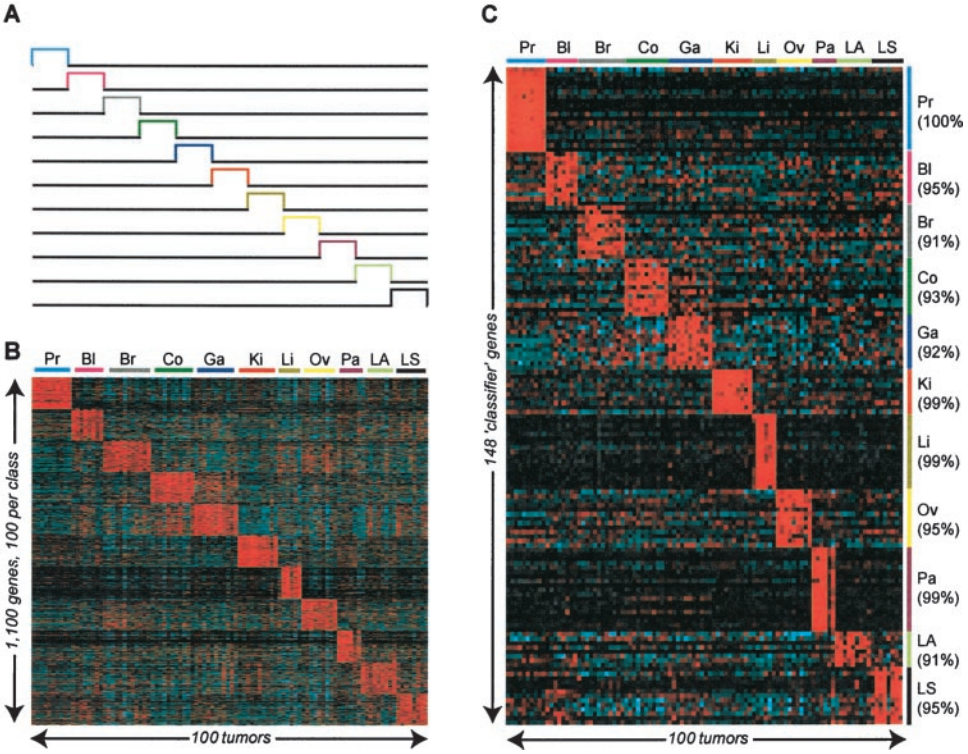

Fig. 1. Selection of tumor-specific genes for cancer class prediction. *A*, schematic diagram depicting the idealized expression profile of tumor-specific genes that the method selects as classifiers. The shape of each profile represents genes that are highly expressed in each cancer type relative to all other tumors in the training set. *B*, 100 genes per tumor class (total, 1100) with the most significant scores in a Wilcoxon rank-sum test for equality were selected as likely candidates for tumor classifiers. *Pr*, prostate; *Bl*, bladder/ureter; *Br*, breast; *Co*, colorectal; *Ga*, gastroesophagus; *Ki*, kidney; *Li*, liver; *Ov*, ovary; *Pa*, pancreas; *LA*, lung adenocarcinomas; *LS*, lung squamous cell carcinoma. *C*, the final refined set of gene classifiers was generated after the genes in *B* were ranked by SVM/LOOCV accuracy. Annotations of the genes from which 110 "predictor" genes were bootstrapped are provided on our website.[4] For clarity, only 8 of 76 predictor genes for lung adenocarcinomas are depicted here. Levels of gene expression (depicted in each *row*) across all samples (*columns*) were median-centered and normalized by "Cluster" and output in "Treeview" (12). *Red*, increased gene expression; *blue*, decreased expression; *black*, median level of gene expression. The color intensity is proportional to the hybridization intensity of a gene from its median level across all samples.

tissue sections and genotyped with a series of polymorphic microsatellite markers. The genotypes from the different sources were substantially different, suggesting that the frozen tissue sample had been mislabeled. These results underscore the use of an objective molecular classification scheme because it depends on objective molecular signatures rather than relying on morphological features of the tumor tissues.

We next applied the classifier to an independent series of 75 carcinoma samples, which were blinded during training of the classifier and queried only after development of the algorithm. This group included some of the tumor classes represented in our training set (specifically, carcinomas of the ovary, prostate, colorectum, lung, breast, and gastroesophagus) as well as 12 metastatic lesions of diverse primary origin (*e.g.*, prostate, breast, ovary, and colon). We made confident and accurate predictions for 64 of 75 (85%) carcinomas within the training set above the empirically set confidence threshold, including 9 of 12 (75%) metastatic carcinomas. The 11 cases that were predicted with low confidence, and therefore not classified, included 4 breast carcinomas, 2 gastroesophageal carcinomas, 1 hepatocellular carcinoma, 1 clear cell

Table 1 *Distribution of class predictions*

Value in each box is the number of samples predicted (from a total of 175) with a given identity by cross-validation in the training set and class prediction in the test set of tumors. The average Dixon confidence score is shown in parentheses.

True identity of unknown sample	Predicted class										
	PR[a]	BL	BR	CO	GA	KI	LI	OV	PA	LA	LS
PR	26 (0.564)										
BL		8 (0.343)									
BR			26 (0.267)								
CO				23 (0.279)							
GA					11 (0.187)					2 (0.044)	
KI						11 (0.502)					
LI			1 (0.115)				5 (0.523)			1 (0.041)	
OV		1 (0.045)						26 (0.317)			
PA									6 (0.529)		
LA		1 (0.020)								13 (0.275)	
LS			1 (0.138)								13 (0.307)

[a] PR, prostate; BL, bladder/ureter; BR, breast; CO, colorectal; GA, gastroesophagus; KI, kidney; LI, liver; OV, ovary; PA, pancreas; LA, lung adenocarcinoma; LS, lung squamous cell carcinoma.

kidney carcinoma, 1 ovarian carcinoma, 1 colorectal carcinoma, and 1 lung adenocarcinoma. Apart from the fact that three of the tested and unclassified cases were metastatic (Table 2), there were no obvious distinguishing features among these carcinomas *versus* other carcinomas in the training or test sets. Of the 11 unclassified cases, 7 were correctly classified, but with low confidence. Thus, a correct anatomical site of tumor origin was predicted for 71 of 75 (95%) cases in the test set, including 11 of the 12 (92%) metastatic lesions (Tables 1 and 2, and supplementary Table 2 on our website).[4] It is important to note that this initial test set does not fully test our class prediction model, specifically because of the lack of sufficient numbers of pancreatic, bladder, and kidney carcinomas. However, our cross-validation results strongly suggest that we would be able to correctly predict an anatomical origin for the majority of carcinomas from these tissue sites.

Most of the genes included in the classifier are expressed in a tissue-specific manner in the epithelium from which the tumors arose and are expressed at similar or elevated levels in the resultant carcinomas (supplementary Table 1).[4] On the basis of gene annotation alone, we recognized many well-described genes whose expression is elevated in tumors. These included *MUC-2* and *A33* in colon cancers, the latter of which has been used as an immunotherapeutic target in advanced colorectal carcinomas (14); mammaglobin-1 (*MGB-1*), which has been found to be a highly sensitive diagnostic marker for micrometastatic breast carcinoma (15); and thyroid transcription factor 1 (*TTF-1*), which has been proposed as a highly accurate marker for the differential diagnosis of lung adenocarcinomas (16). We also identified genes such as uroplakin II (*UPII*), whose expression in bladder carcinoma cells is likely maintained at levels similar to that of normal urothelium. Detection of *UPII* transcripts in circulating bladder cancer cells, however, has been proposed as a sensitive marker of micrometastasis (17).

We also identified genes whose annotations suggested their expression in the stromal cells that surround epithelial tumors or in inflammatory cells. In some cases we subsequently found evidence that suggests their overexpression in malignant epithelia [*e.g.,* the fibroblast activation protein (*FAP-*α) in breast cancers (18)]. In adenocarcinomas of the lung, we identified genes whose annotations indicated the presence of B cells, T cells, macrophages, and neutrophils. We suspect that many of these genes may have been selected because of the relative paucity of "lung-specific" classifiers, and not because these samples necessarily contained higher proportions of infiltrating inflammatory cells relative to the other tumor samples. Conservatively, we suggest that the most reliable classifiers of lung adenocarcinomas probably include those genes with predicted accuracies $>$95%, *i.e., TTF-1*. In pancreas cancers we identified genes whose expression is indicative of acinar cell differentiation. Although we specifically attempted to avoid normal epithelium in all of the tumor samples that we profiled, the highly diffuse nature of pancreatic cancer growth precluded an absolutely complete separation of normal and neoplastic cells. Highly expressed genes within small amounts of normal epithelia may conceivably give rise to some of the signals detected on the arrays. However, it remains a possibility that expression of some of these "acinar" genes is maintained in pancreatic tumor cells.

Table 2 *Prediction of tumor origin of metastatic carcinomas*

Samples of metastatic carcinomas predicted by the classification method. Metastatic sites are in parentheses. Nine of 12 carcinomas (bold) were correctly predicted with high confidence. There were no incorrect predictions with high confidence score. High confidence is defined as above a Dixon score threshold of $>$0.1.

Sample	Prediction	Dixon score	Sample identity
U7	**Ovary**	**0.29**	**Metastatic serous pap.[a] ca. of the ovary (omentum)**
U8	**Ovary**	**0.34**	**Metastatic serous pap. ca. of the ovary (omentum)**
U11	**Ovary**	**0.20**	**Metastatic serous pap. ca. of the ovary (omentum)**
U12	**Colon**	**0.33**	**Metastatic colon ca. (ovary)**
U16	Breast	0.03	Metastatic breast ca. (liver)
U17	Bladder	0.02	Metastatic lung Ad (brain)
U19	**Lung SCC**	**0.36**	**Metastatic lung SCC (liver)**
U40	**Prostate**	**0.54**	**Metastatic prostate ca. (lymph node)**
U41	**Prostate**	**0.47**	**Metastatic prostate ca. (lymph node)**
U42	**Colon**	**0.31**	**Metastatic colon ca. (liver)**
U43	**Colon**	**0.25**	**Metastatic colon ca. (liver)**
UX14	Kidney	0.07	Metastatic kidney ca. (colon)

[a] pap., papillary; ca., carcinoma; Ad, adenocarcinoma; SCC, squamous cell carcinoma.

Because of the inherent difficulty in using gene annotation alone to judge tissue-specific *versus* tumor-elevated gene expression, we next sought to objectively "dissect" some of the predictor gene subsets into tissue-specific genes and tissue-specific/tumor-elevated genes. As an example, we chose 28 of the genes that were $\geq$92% predictive of serous papillary carcinomas of the ovary and compared the expression levels of these genes in an expanded set of 24 ovarian tumor samples against 5 samples of normal ovary, 2 of which were highly enriched for surface ovarian epithelial cells (3). Differential expression was determined for genes whose expression was significantly different in normal and tumor tissues ($P < 0.01$, unpaired t test) and where the mean level of expression in tumor tissues was $>$3 times that in normal tissues. By these criteria, 18 of 28 genes were significantly overexpressed in the tumors (Fig. 2*A*). Among this group of genes were protease M/neurosin/kallikrein 6 (*hK6*), which has been proposed as a candidate serum marker for ovarian cancer (19), and mesothelin (*CAK1*), which is overexpressed in ovarian cancers and is used as a specific target for a novel therapeutic immunotoxin (20). The 10 tissue-specific genes, which included the WT gene (*WT-1*), *smad6*, and *Hox5.1*, most likely represent features of normal ovarian physiology.

We have begun to evaluate the "predictability" of some of the classifier genes in ovarian cancers at the level of the expressed protein. For example, we used a polyclonal antibody specific to the WT protein, whose transcript was highly expressed in ovarian cancers relative to tumors of the other 10 classes, on tissue microarrays containing 229 carcinomas representing tumors from the 10 anatomical sites analyzed in the study. Immunostaining for WT protein was present in nuclei from 18 of 20 (90%) serous papillary carcinomas, whereas nuclear immunoreactivity was absent in the other 209 carcinomas (Fig. 2, *B–E*). As expected from the analysis of classifier gene transcription in ovarian cancers, the normal serous lining epithelium of the ovary was also positive for WT protein (Fig. 2*C*). It should be noted that expression of WT has been reported in other tissues, but in the context of the tissues examined in our classification scheme, WT expression was specific for the ovary.

Transcript profiles of human tumors have previously been used to predict the membership of an unknown sample into one of two, three, or at most four distinct tumor classes (21–23). However, the use of tumor-specific genes to extend these or other discriminant methods to prediction of tumor origin in the context of multiple ($>$10) cancer classes has not been demonstrated and is particularly challenging. We assessed many methods for multiclass prediction during this study, based on either weighted correlation methods (21) or on other supervised learning methods (*e.g.,* Fisher's linear discriminant analysis). Although all of the methods that we used have performed reasonably well, we found that methods such as SVM, which do not make assumptions about the distribution of the data (8), performed significantly better and selected for greater uniformity and specificity among the class-specific predictors. These findings have recently been corroborated (24), although the specifics of the SVM methodology are different.

We found that classification of tumors arising in certain ana-

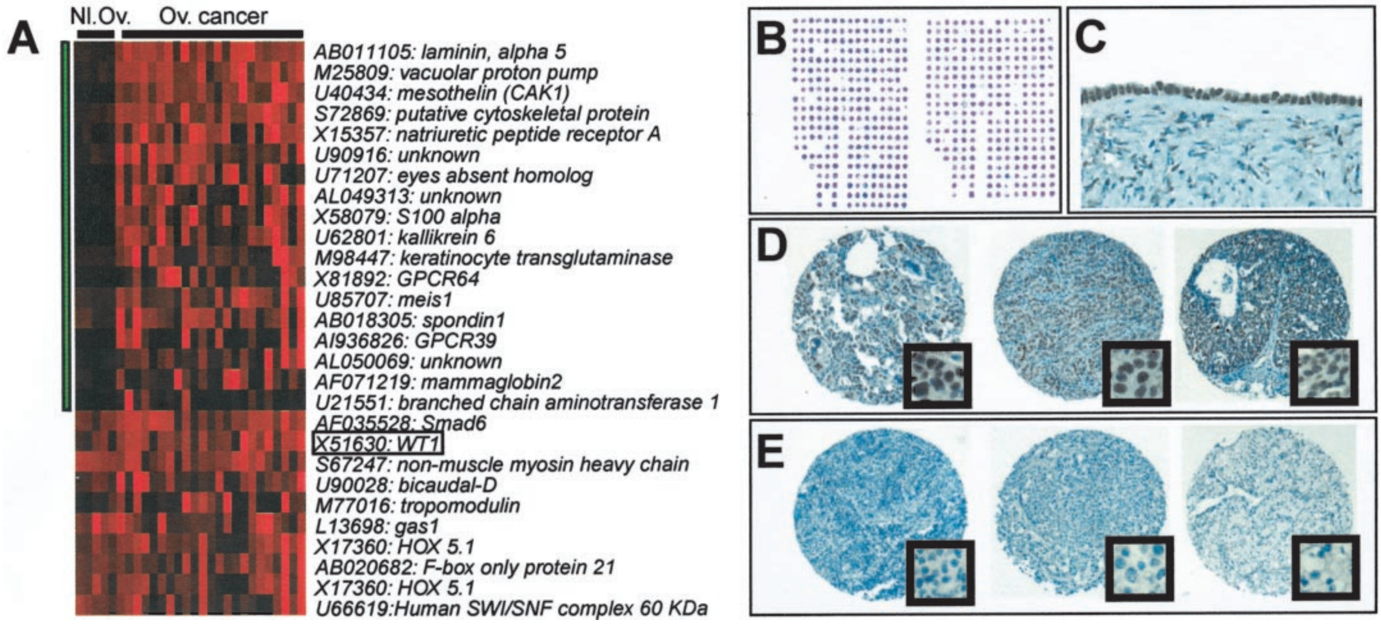
A
Nl.Ov.
Ov. cancer
AB011105: laminin, alpha 5
M25809: vacuolar proton pump
U40434: mesothelin (CAK1)
S72869: putative cytoskeletal protein
X15357: natriuretic peptide receptor A
U90916: unknown
U71207: eyes absent homolog
AL049313: unknown
X58079: S100 alpha
U62801: kallikrein 6
M98447: keratinocyte transglutaminase
X81892: GPCR64
U85707: meis1
AB018305: spondin1
AI936826: GPCR39
AL050069: unknown
AF071219: mammaglobin2
U21551: branched chain aminotransferase 1
AF035528: Smad6
X51630: WT1
S67247: non-muscle myosin heavy chain
U90028: bicaudal-D
M77016: tropomodulin
L13698: gas1
X17360: HOX 5.1
AB020682: F-box only protein 21
X17360: HOX 5.1
U66619:Human SWI/SNF complex 60 KDa
B
C
D
E

Fig. 2. Genes and proteins predictive for serous papillary adenocarcinomas of the ovary. *A*, expression levels of highly predictive classifier genes in normal and malignant samples of the ovary. *Green bars*, differentially expressed genes where the mean level of expression in tumor samples (*Ov. cancer*) is >3 times the mean level of expression in normal tissues (*Nl. Ov.*) and where $P < 0.01$ by an unpaired *t* test. Expression of the *WT-1* gene is highlight by the *box*. Gene expression was normalized and output in Treeview as described in the legend to Fig. 1. *B*, visualization of a tissue microarray containing 36 normal epithelial tissues and 229 carcinomas representative of the 10 anatomical sites of the tumors profiled in the study stained with H&E staining. *C-E*, tissue microarrays stained with an antibody specific to the WT protein. *C*, normal serous lining of the ovary positive for WT. *D*, three serous papillary carcinomas of the ovary positive for WT. *E*, breast, lung, and kidney carcinomas negative for WT (immunoperoxidase technique). *Insets* show magnified view of nuclei.

tomical sites was relatively straightforward because of the large number of unequivocal predictor genes (*e.g.*, 19 genes with 100% predictive accuracy for prostate cancer). In contrast, prediction of other tumors, such as those of the lung, bladder/ureter, or gastroesophagus, was more difficult because of the relative paucity of highly predictive classifier genes. The difficulty in selecting genes whose expression is specific to these cancers reflects a high degree of molecular relatedness, which we had observed in initial analyses of tumor gene expression.[4] For example, blinded gastroesophageal cancers that could not be predicted by our method were assigned as lung tumors (albeit with confidence scores close to zero). Analysis of the entire human transcriptome may uncover tumor-specific genes for those neoplasms that we have shown to have a high similarity in expression profiles.

A striking conclusion from the data presented here is that we could identify subsets of genes with highly restricted, tumor-specific expression for as many as 11 distinct tumor classes, despite well-described tumor heterogeneity and obvious molecular similarities among many divergent tumor classes. The fact that we could successfully use these gene subsets to predict the origin of a given tumor in a majority of cases underscores how strongly characteristic these genes must be for specific histopathological subtypes of cancer. In that regard, it is worth noting that, using as few as 11 genes (*i.e.*, 1 gene per tumor class), we could predict the anatomical origin of up to 91 and 83% of the training and blinded tumor samples, respectively (in the absence of a strict confidence threshold). These results suggest that we can construct custom DNA microarrays for a molecular classification of solid tumors, a resource that will augment traditional site-specific and histopathological classification schemes. The extension of these and other discriminant methods to identify molecular correlates with tumor grade, stage, response to therapy, and outcome will further contribute to the optimal management of patients with cancer.

Acknowledgments

We are grateful to Drs. Elizabeth Winzeler, Yingyao Zhou, Steve Kay, Quinn Deveraux, and Nathanael Gray for discussions and critical reading of the manuscript.

References

1. Hillen, H. F. Unknown primary tumors. Postgrad. Med. J., *76:* 690–693, 2000.
2. Greenlee, R. T., Hill-Harmon, M. B., Murray, T., and Thun, M. Cancer Statistics, 2001. CA Cancer J. Clin., *51:* 15–36, 2001.
3. Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A., and Hampton, G. H. Analysis of gene expression in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA, *98:* 1176–1181, 2001.
4. Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. Analysis of gene expression identifies candidate makers and pharmacologic targets in prostate cancer. Cancer Res., *61:* 5974–5978, 2001.
5. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, K. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol., *14:* 1675–1680, 1996.
6. Wodicka, L., Dong, H., Mittmann, M., Ho, M-H., and Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol., *15:* 1359–1367, 1997.
7. Hollander, M., and Wolfe, D. A. Nonparametric Statistical Inference. New York: John Wiley & Sons, 1973.
8. Vapnik, V. N. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag, 1995.
9. Stone, M. Cross-validation choice and assessment of statistical predictions. J. R. Stat. Soc., *B-36*: 111–114, 1974.
10. Efron, B., and Tibshirani, R. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993.
11. Greller, L. D., and Tobin, F. L. Detecting selective expression of genes and proteins. Genome Res., *9:* 282–296, 1999.
12. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.
13. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, *16:* 906–914, 2000.
14. Tschmelitsch, J., Barendswaard, E., Williams, C. J., Yao, T. J., Cohen, A. M., Old, L. J., and Welt, S. Enhanced antitumor activity of combination radioimmunotherapy

(^{131}I-labeled monoclonal antibody A33) with chemotherapy (fluorouracil). Cancer Res., *57:* 2181–2186, 1997.

15. Ghossein, R. A., Carusone, L., and Bhattacharya, S. Molecular detection of micrometastases and circulating tumor cells in melanoma prostatic and breast carcinomas. In Vivo, *14:* 237–250, 2000.
16. Reis-Filho, J. S., Carrilho, C., Valenti, C., Leitao, D., Ribeiro, C. A., Ribeiro, S. G., and Schmitt, F. C. Is TTF1 a good immunohistochemical marker to distinguish primary from metastatic lung adenocarcinomas? Pathol. Res. Pract., *196:* 835–840, 2000.
17. Li, S. M., Zhang, Z. T., Chan, S., McLenan, O., Dixon, C., Taneja, S., Lepor, H., Sun, T. T., and Wu, X. R. Detection of circulating uroplakin-positive cells in patients with transitional cell carcinoma of the bladder. J. Urol., *162:* 931–935, 1999.
18. Kelly, T., Kechelava, S., Rozypal, T. L., West, K. W., and Korourian, S. Seprase, a membrane-bound protease, is overexpressed by invasive ductal carcinoma cells of human breast cancers. Mod. Pathol., *11:* 855–861, 1998.
19. Diamandis, E. P., Yousef, G. M., Soosaipillai, A. R., and Bunting, P. Human kallikrein 6 (zyme/protease M/neurosin): a new serum biomarker of ovarian carcinoma. Clin. Biochem., *33:* 579–583, 2000.
20. Hassan, R., Viner, J. L., Wang, Q. C., Margulies, I., Kreitman, R. J., and Pastan, I. Anti-tumor activity of K1-LysPE38QQR, an immunotoxin targeting mesothelin, a cell-surface antigen overexpressed in ovarian cancer and malignant mesothelioma. J. Immunother., *20:* 2902–2906, 2000.
21. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash. DC), *286:* 531–537, 1999.
22. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.
23. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., *7:* 673–679, 2001.
24. Yeang, C-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. Molecular classification of multiple tumor types. Bioinformatics, *17* (Suppl. 1)*:* s316–s322, 2001.