American Journal of Pathology, Vol. 159, No. 4, October 2001
Copyright © American Society for Investigative Pathology

Short Communication

Organ-Specific Molecular Classification of Primary Lung, Colon, and Ovarian Adenocarcinomas Using Gene Expression Profiles

Thomas J. Giordano,* Kerby A. Shedden,† Donald R. Schwartz,‡ Rork Kuick,§ Jeremy M. G. Taylor,¶ Nana Lee,‡ David E. Misek,§ Joel K. Greenson,* Sharon L. R. Kardia,‖ David G. Beer,** Gad Rennert,†† Kathleen R. Cho,* Stephen B. Gruber,‡ Eric R. Fearon,‡ and Samir Hanash§

From the Departments of Pathology, Statistics,† Internal Medicine,‡ Pediatrics,§ Biostatistics,¶ Epidemiology,‖ and Surgery,** The University of Michigan, Ann Arbor, Michigan; and the National Cancer Control Center,†† Clalit Health Services and Technion University, Haifa University, Haifa, Israel*

Molecular classification of tumors based on their gene expression profiles promises to significantly refine diagnosis and management of cancer patients. The establishment of organ-specific gene expression patterns represents a crucial first step in the clinical application of the molecular approach. Here, we report on the gene expression profiles of 154 primary adenocarcinomas of the lung, colon, and ovary. Using high-density oligonucleotide arrays with 7129 gene probe sets, comprehensive gene expression profiles of 57 lung, 51 colon, and 46 ovary adenocarcinomas were generated and subjected to principle component analysis and to a cross-validated prediction analysis using nearest neighbor classification. These statistical analyses resulted in the classification of 152 of 154 of the adenocarcinomas in an organ-specific manner and identified genes expressed in a putative tissue-specific manner for each tumor type. Furthermore, two tumors were identified, one in the colon group and another in the ovarian group, that did not conform to their respective organ-specific cohorts. Investigation of these outlier tumors by immunohistochemical profiling revealed the ovarian tumor was consistent with a metastatic adenocarcinoma of colonic origin and the colonic tumor was a pleomorphic mesenchymal tumor, probably a leiomyosarcoma, rather than an epithelial tumor. Our results demonstrate the ability of gene expression profiles to classify tumors and suggest that determination of organ-specific gene expression profiles will play a significant role in a wide variety of clinical settings, including molecular diagnosis and classification. ***(Am J Pathol 2001, 159:1231–1238)***

Molecular classification of tumors by high-throughput comprehensive technologies for assaying gene expression, such as high-density oligonucleotide and cDNA microarrays, offers the potential to radically alter the practice of surgical pathology and oncology. Using these technologies, it may be possible to identify clinically relevant subsets of tumors that would otherwise be indistinguishable by conventional histopathological assessment. In principle, expression-profiling analyses should identify tumors more likely to invade, recur, and/or metastasize, and the approach should allow improved prediction of response to specific therapeutic regimens and clinical outcome. Data from a recent study of a large cohort of lymphomas supports this view. Specifically, large B-cell lymphomas could be divided based on gene expression profiles into two subtypes associated with different survival rates.[1] A similar study classified cutaneous malignant melanomas based on gene expression profiles.[2]

Another major anticipated benefit of these technologies is the establishment of organ- and tumor-specific profiles that, among other potential benefits, might assist with the diagnostic work-up of patients with metastatic cancer of unknown origin at the time of initial diagnosis. A comprehensive library of unique gene expression profiles of all of the major tumor types would permit a definitive diagnosis in the absence of pertinent clinical history,

Supported by funds from the National Cancer Institute (NIH CA84952 and CA81488) and in part by the Tissue Core of the University of Michigan Comprehensive Cancer Center (NIH CA46952).

Accepted for publication July 6, 2001.

Address reprint requests to Thomas J Giordano, M.D., Ph.D., University of Michigan Health System, Department of Pathology, Room 2G332/Box 0054, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0054. E-mail: giordano@med.umich.edu.

imaging studies, and/or surgical exploration, thus simplifying the diagnostic evaluation. For example, comparing the gene expression profile of a patient's brain lesion to the gene expression library may be sufficient to establish a diagnosis of primary lung adenocarcinoma in the absence of thoracic imaging studies. Furthermore, these profiles might assist in the diagnosis of histologically similar primary tumor types, such as in distinguishing poorly differentiated lung carcinoma from malignant mesothelioma. However, at this point, it has not been established if it will be possible to define gene expression profiles that will discriminate the major tumor types in an organ-specific manner. A crucial step in establishing the diagnostic relevance of gene expression profiles is to compare profiles of histologically similar tumors, such as adenocarcinomas, from different organs. In this study, we compared the gene expression profiles of 154 primary adenocarcinomas of lung, colon, and ovary and demonstrated these profiles could discriminate the tumors in an organ-specific manner. In addition, we identified genes that are potentially useful as diagnostic markers for these tumors.

Materials and Methods

Tumors and Histopathology

The primary tumors analyzed in this study were derived from several sources. The lung adenocarcinomas were procured from the University of Michigan Health System between 1994 and 1999. The colon adenocarcinomas were procured from five Israeli hospitals as part of the Molecular Epidemiology of Colorectal Cancer Study, a collaborative project between the University of Michigan and the National Cancer Control Center, Haifa, Israel (NIH CA81488). The ovarian tumors were procured from several sources, including the University of Michigan Health System, the Cooperative Human Tissue Network, and Cornell New York Hospital. All procedures were approved by the University of Michigan Institutional Review Board (IRB-Medicine).

All tumors were processed in a similar manner. Frozen tumor samples were embedded in OCT freezing media (Miles Scientific, Naperville, IL), cryotome sectioned (5 μm), and evaluated by routine hematoxylin and eosin (H&E) stains by one of three surgical pathologists. Whenever possible, the corresponding H&E sections from paraffin blocks were also evaluated. Areas of relatively pure tumor (at least 70% tumor cells) were selected for RNA isolation. All grades of differentiation were exhibited by the tumors.

RNA Isolation

Single isolates of tumor samples were homogenized in the presence of Trizol reagent (Life Technologies, Gaithersburg, MD) and total cellular RNA was purified according to manufacturer's procedures. RNA samples were further purified using RNeasy spin columns (Qiagen, Valencia, CA) and used to prepare cRNA probes. RNA quality of the lung and ovary tumors was assessed by 1% agarose gel electrophoresis in the presence of ethidium bromide. Samples that did not reveal intact and approximately equal 18S and 28S ribosomal bands were excluded from further study (5% of the lung and 17% of the ovary cases).

cRNA Synthesis and Gene Expression Profiling

This study used commercially available high-density microarrays (Affymetrix, Santa Clara, CA) that produce gene expression levels on 7129 known genes and expressed sequence tags (HuGeneFL Array). Preparation of cRNA, hybridization, and scanning of the arrays were performed according to manufacturer's protocols. Briefly, 5 μg of total RNA was used to generate double-stranded cDNA by reverse transcription using a cDNA synthesis kit (Superscript Choice System; Life Technologies, Inc., Rockville, MD) that uses an oligo(dT)$_{24}$ primer containing a T7 RNA polymerase promoter 3′ to the poly T (Geneset, La Jolla, CA), followed by second-strand synthesis. Labeled cRNA was prepared from the double-stranded cDNA by *in vitro* transcription by T7 RNA polymerase in the presence of biotin-11-CTP and biotin-16-UTP (Enzo, Farmington, NY). The labeled cRNA was purified over RNeasy columns. Fifteen μg of cRNA was fragmented at 94°C for 35 minutes in 40 mmol/L of Tris-acetate, pH 8.1, 100 mmol/L of potassium acetate, and 30 mmol/L of magnesium acetate. The cRNA was then used to prepare 300 μl of hybridization cocktail (100 mmol/L MES, 1 mol/L NaCl, 20 mmol/L ethylenediaminetetraacetic acid, 0.01% Tween 20) containing 0.1 mg/ml of herring sperm DNA (Promega, Madison, WI) and 500 μg/ml of acetylated bovine serum albumin (Life Technologies, Inc.). Before hybridization, the cocktails were heated to 94°C for 5 minutes, equilibrated at 45°C for 5 minutes, and then clarified by centrifugation (16,000 × *g*) at room temperature for 5 minutes. Aliquots of this hybridization cocktail containing 10 μg of fragmented cRNA were hybridized to HuGeneFL arrays at 45°C for 16 hours in a rotisserie oven at 60 rpm. The arrays were washed using nonstringent buffer (6× SSPE) at 25°C, followed by stringent buffer (100 mmol/L MES, pH 6.7, 0.1 mol/L NaCl, 0.01% Tween 20) at 50°C. The arrays were stained with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR), washed with 6× sodium chloride, sodium phosphate, EDTA (SSPE buffer), incubated with biotinylated anti-streptavidin IgG, stained again with streptavidin-phycoerythrin, and washed again with 6× SSPE. The arrays were scanned using the GeneArray scanner (Affymetrix). Image analysis was performed with GeneChip software (Affymetrix).

Statistical Analysis

The HuGeneFL chip consists of 7129 probe sets, each representing a transcript. Each probe set typically consists of 20 perfectly complementary 25 base long probes as well as 20 mismatch probes that are identical except for an altered central base. We subtract the mismatch probe values from the perfect match values and average the middle 50% of these differences as the expression measure for that probe set.

A quantile normalization procedure was used to adjust for differences in the probe intensity distribution across different chips. We applied a monotone linear spline to each chip that mapped quantiles 0.02 up to 0.98 (in increments of 0.02) exactly to the corresponding median quantiles for all 154 samples. Then, the transform log(100 + max(X + 100; 0)) was applied to the data from each chip.

We built a classifier out of our 154 training samples as follows. We selected N markers from each of the three tumor classes, giving 3N markers in all (a range of values for N was considered, as discussed in the Results section). To classify a new sample of unknown tumor type, we compute the correlation coefficient between the 3N markers on the unidentified sample and the same markers on each of the 154 training samples. The class identities of the five training samples having the greatest correlation with the unclassified sample are then considered. If three or more of these samples belong to a common class, then this is the predicted class for the unclassified sample. Otherwise, the prediction is considered to be indeterminate. This strategy is known in the classification literature as "five-nearest neighbors with majority voting."[3]

We used a cross-validation procedure to estimate the error rate of our classifier. In the generic procedure, we set aside a single validation sample, leaving 153 samples to train a classifier, as described above. Note that compared to the classifier that would be used in practice, this new classifier will use a slightly different set of markers as well as having one fewer training sample. Moreover, this classifier does not have access to the expression values or tissue type of the single held-out sample. We then use this classifier to predict the class of the held-out sample, and record whether this prediction is correct. This process is repeated 154 times, with each sample being held out exactly once. The aggregate error rate across the 154 predictions is used as an estimate for the error rate that would be expected to occur in practice.

Immunohistochemistry

Routine immunohistochemistry was performed using formalin-fixed, paraffin-embedded sections using the avidin-biotin complex method.[4] The following antibodies, dilutions, and pretreatment conditions were used: anti-keratin (CAM 5.2, 1:10, trypsin pretreatment; Becton-Dickinson, San Jose, CA), anti-human epithelial keratins (AE1:AE3, 1:800, no pretreatment; Roche Diagnostics, Indianapolis, IN), anti-human melanoma (HMB45, 1:25, no pretreatment; DAKO, Carpinteria, CA), anti-cow S-100 (1:500, no pretreatment; DAKO), anti-vimentin (1:800, no pretreatment; DAKO), anti-human cytokeratin 20 (CK20) (1:25, DAKO Protease 1 pretreatment; DAKO), anti-human CK7 (1:25, DAKO Protease 2 pretreatment; DAKO), anti-CEA (monoclonal D-14, 1:8, DAKO Protease 1 pretreatment: E-Z-EM, Westbury, NY), anti-α smooth muscle actin (1:1600, no pretreatment; Sigma Chemical Co., St. Louis, MO), anti-human muscle actin (HHF35, 1:100, DAKO Protease 2 pretreatment; DAKO) and anti-c-KIT (1:100, citrate buffer pretreatment; DAKO).

Results

Gene Expression Profiles Distinguish Lung, Colon, and Ovary Adenocarcinomas and Identify Differentially Expressed Genes

Comprehensive gene expression profiles of 57 lung, 51 colon, and 46 ovary primary carcinomas were generated using high-density oligonucleotide arrays with 7129 probe sets, which in total interrogated some 6800 genes. To provide a visual assessment of relationships between the tumors based on gene expression, we considered each sample to be represented by a point in a 7129 multidimensional space, with each coordinate given by a gene expression level. Several views of this set of points were generated using principal component analysis (PCA), which locates the two-dimensional views that capture the greatest amount of variability in the data. We note that these views were based solely on aggregate expression variation and no references to the tissue classifications were made. We generated four views in all, by stratifying the genes into four quarters of equal size, and then applying PCA to the measurements in each quarter separately (Figure 1). The strata were formed based on the average expression level of each transcript across the 154 samples; the first stratum contained the 25% of transcripts with the least average abundance, the second stratum contained the next 25% of the genes, and so on.

The views determined by PCA (Figure 1) indicated substantial differences in gene expression between the three tumor types throughout the range of expression measures. Thus, not only did genes with high average expression levels allow the three tumor types to be distinguished from each other (Figure 1, top quarter), but genes that were expressed at low average levels were informative as well (Figure 1, first quarter). Particularly notable was the wide margin that separated lung and colon tumors (Figure 1, top quarter). It was also evident that substantial heterogeneity occurred within each tumor type, with the ovarian tumors showing the greatest heterogeneity, and the colon tumors showing the least. Much of the substructure in the various panels of Figure 1 was attributed to identified differences in tumor histopathology. The ovarian tumors could be divided into two subsets. A minor subset highlighted in Figure 1 (top quarter) clustered separately from the major ovarian set, overlapped with colon tumors in the second and third quarterlies (Figure 1), and consisted exclusively of mucinous ovarian tumors. However, some mucinous tumors also clustered with the major group. The colon tumor that was substantially separated from the remainder was a sarcoma (see below), in contrast to the remaining colon tumors, all of which were adenocarcinomas. The ovarian tumor most completely embedded within the colon cohort was subsequently determined to have a number of fea-

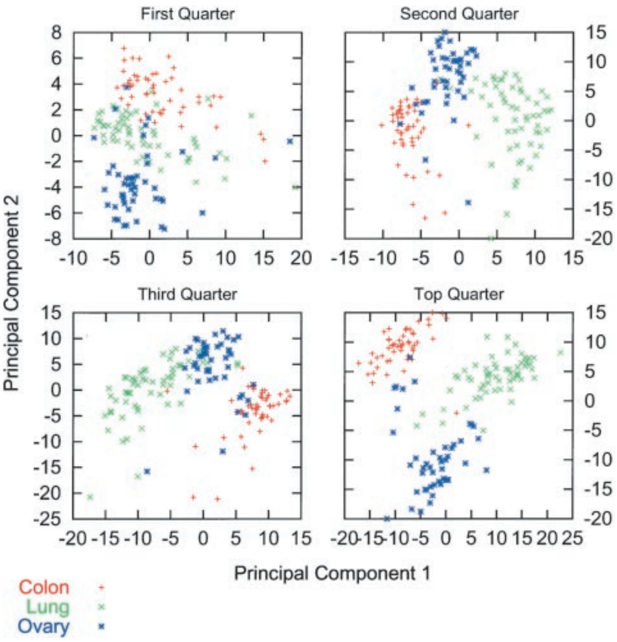

Figure 1. PCA of all genes. Four PCA views of the gene expression profiles, generated by stratifying the all of the genes analyzed into four quarters of equal size (see Results). Colon cohort, red; lung cohort, green; and ovary cohort, blue.

tures (see below) indicating it was a colonic adenocarcinoma metastatic to the ovary.

Next we applied a gene selection procedure to identify a set of markers for each tumor type (lung, colon, ovary). Specifically, we sought to identify the genes that had much greater expression in one tumor type compared to either of the other two types. The difference between the average expression of each gene within a given tumor type and the larger of the two average expression values for the same gene in the other two types was computed. Because the data were log-transformed, this value can be interpreted as the logarithm of the ratio between the geometric mean expression in the more highly expressing of the other two types. The result is that for each tumor class, we obtained a ranking of the genes, with the genes having high rank being the strongest markers for the tumor type. The top 20 markers for each type are shown in Table 1. Three of the 60 genes were identified twice and one three times, as they are represented more than once in the arrays with distinct probe sets.

A substantial fraction of the genes that were assayed exhibited differential expression between the three tumor types. More than 2000 were statistically significant at the 5% level using analysis of variance, although many of these exhibited less than a twofold change between the class means. Thirty genes (29 unique) exhibited greater than fivefold increased levels of expression in the colon tumor type relative to the other two types. The corresponding number of genes with fivefold greater levels of expression for lung and ovary are, 36 (32 unique) and 32 (31 unique) genes, respectively. The markers selected for use in classification had between 2.1-fold and >200-fold greater average expression in the type for which they were a marker compared to the other two types.

We then built a classifier out of our 154 tumor samples as described in Materials and Methods. Values of $N = 3$, 5, 7, 10, 15, and 20 for each tumor type were considered for the number of markers per tumor class that were made available to the classifier. Using fewer than seven markers (six unique genes) led to a degradation in performance, whereas more than 10 markers did not provide any improvement. When 10 markers (nine unique colon and ovary genes and seven unique lung genes) were considered, 152 of 154 samples were correctly classified. This is considered to be the best possible result, because the probable colonic metastasis diagnosed as an ovarian primary represented an apparent erroneous diagnosis and the colonic sarcoma could not have been correctly classified as it represented the only sarcoma in the study.

Additionally, to provide another visual assessment of relationships between the tumors based on the most differentially expressed genes, we used PCA as above using the gene expression data for 60 genes (55 unique), the top 20 from each tumor type identified by the classifier. The PCA view derived from these genes is shown (Figure 2) and was very similar to the view derived from the top quarter of expressed genes (Figure 1), but showed more distinct separation of the tumor types. The two absolute outlier tumors identified by the classifier were clearly seen as outliers and the subset of mucinous ovarian tumors identified previously were again ascertained as colon-like.

Histopathological and Immunohistochemical Investigation of Outlier Cases

The two absolute outlier tumors, one colonic and one ovarian, were further investigated by routine immunohistochemical methods to further define their histopatholog-

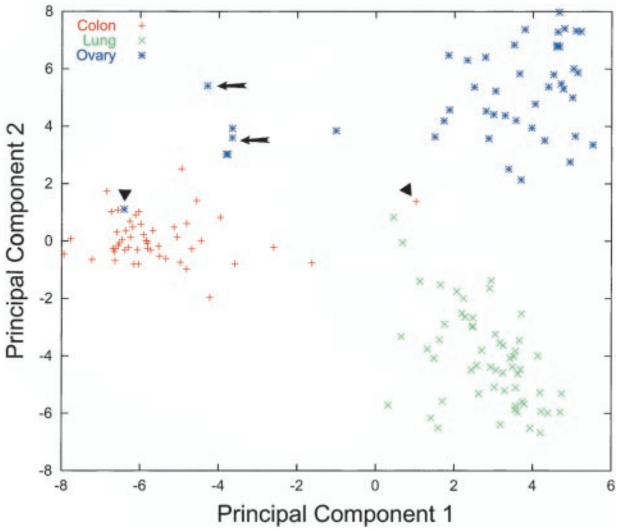

Figure 2. PCA of selected genes. PCA view of the top 60 differentially expressed genes (20 from each tumor type). **Arrowheads** show the two outlier tumors, one within the colonic group and one distinct from the three tumor groups. **Arrows** highlight the group of mucinous ovarian tumors that are seen as colon-like. Colon cohort, red; lung cohort, green; and ovary cohort, blue.

Table 1. Top 20 Differentially Expressed Genes for Each Tumor Type

Probe set	Gene name	Unigene description
Colon genes		
M10050	FABP1	Fatty acid binding protein 1, liver
AB006781	LGALS4	Lectin, galactoside-binding, soluble, 4 (galectin 4)
X83228	CDH17	Cadherin 17, LI cadherin (liver-intestine)
M35252	TM4SF3	Transmembrane 4 superfamily member 3
X68314	GPX2	Glutathione peroxidase 2 (gastrointestinal)
U07969	CDH17	Cadherin 17, LI cadherin (liver-intestine)
U51095	CDX1	Caudal type homeo box transcription factor 1
L08044	TFF3	Trefoil factor 3 (intestinal)
M29540	CEACAM5	Carcinoembryonic antigen-related cell adhesion molecule 5
U79725	GPA33	Glycoprotein A33 (transmembrane)
X52003	TFF1	Trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)
X12901	VIL1	Villin 1
M76180	DDC	Dopa decarboxylase (aromatic L-amino acid decarboxylase)
J05412	REG1A	Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)
J05257	DPEP1	Dipeptidase 1 (renal)
X73501	KRT20	Cytokeratin 20
M22430	PLA2G2A	Phospholipase A2, group IIA (platelets, synovial fluid)
M82962	MEP1A	Meprin A, alpha (PABA peptide hydrolase)
U27333	FUT3	Fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group included)
U51096	CDX2	Caudal type homeo box transcription factor 2
Lung genes		
M68519	SFTPA2	Surfactant, pulmonary-associated protein A2
M24461	SFTPB	Surfactant, pulmonary-associated protein B
M30838	SFTPA2	Surfactant, pulmonary-associated protein A2
M13686	SFTPA1	Surfactant, pulmonary-associated protein A1
J03890	SFTPC	Surfactant, pulmonary-associated protein C
S71043	NULL	Homo sapiens SNC73 protein (SNC73) mRNA, complete cds
U43203	TITF1	Thyroid transcription factor 1
HG3925-HT4195	SFTPA2	Surfactant, pulmonary-associated protein A2
Y09267	FMO2	Flavin containing monooxygenase 2
X82850	TITF1	Thyroid transcription factor 1
X53331	MGP	Matrix Gla protein
HG544-HT544	ECGF1	Endothelial cell growth factor 1 (platelet-derived)
U05861	AKR1C1	Aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase)
M87789	IGHG3	Immunoglobulin heavy constant gamma 3 (G3m marker)
X64072	ITGB2	Integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit)
M63438	IGKC	Immunoglobulin kappa constant
HG3044-HT3742	FN1	Fibronectin 1
M34996	HLA-DQA1	Major histocompatibility complex, class II, DQ alpha 1
L48516	PON3	Paraoxonase 3
X57809	NULL	Human anti-streptococcal/anti-myosin immunoglobulin lambda light chain variable region mRNA, partial cds
Ovary genes		
X03635	ESR1	Estrogen receptor 1
M11433	RBP1	Retinol-binding protein 1, cellular
X07438	RBP1	Retinol-binding protein 1, cellular
U90336	PEG3	Paternally expressed 3
HG1496-HT1496	DLK1	Delta-like homolog (Drosophila)
X51630	WT1	Wilms tumor 1
X92744	DEFB1	Defensin, beta 1
J05428	UGT2B7	UDP glycosyltransferase 2 family, polypeptide B7
J00306	SST	Somatostatin
U66838	CCNA1	Cyclin A1
M59979	PTGS1	Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)
U28368	ID4	Inhibitor of DNA binding 4, dominant negative helix-loop-helix protein
X04470	SLPI	Secretory leukocyte protease inhibitor (antileukoproteinase)
U85707	MEIS1	Meis1 (mouse) homolog
X58079	S100A1	S100 calcium-binding protein A1
U17280	STAR	Steroidogenic acute regulatory protein
U65011	PRAME	Preferentially expressed antigen in melanoma
M68516	SERPINA5	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5
M63379	CLU	Clusterin (complement lysis inhibitor, SP-40,40, sulfated glycoprotein 2, testosterone-repressed prostate message 2, apolipoprotein J)
S37730	IGFBP2	Insulin-like growth factor binding protein 2 (36kD)

ical classification. This approach is typically used by practicing surgical pathologists to determine the likely origin of tumors of uncertain primary site. The colonic tumor, a pleomorphic spindle-cell neoplasm on routine H&E (Figure 3A), was correctly diagnosed at the contributing hospital as a nonepithelial neoplasm, probably sarcoma, but was included in the colon cancer cohort because all of these tumors were derived from a population-based study of incident, invasive colorectal cancers, almost all of which are adenocarcinomas. Immunostains for low- and high-molecular weight cytokeratins, CK7 and CK20, S-100, HMB-45, vimentin, muscle-specific actin, and smooth muscle actin showed the neoplastic cells to be negative for cytokeratins (Figure 3B), S-100 (not shown), and HMB-45 (not shown), strongly positive for vimentin (Figure 3C), and focally positive for both actins (not shown). This immunohistochemical profile, together with the histopathology, is diagnostic of high-grade leiomyosarcoma. Gastrointestinal stromal tumor was not an appropriate diagnosis based on histopathology, a negative KIT immunostain (not shown), and no detectable expression of the *c-kit* gene on the array (not shown).

The outlier ovarian tumor was originally diagnosed as a primary mucinous ovarian adenocarcinoma (Figure 3D). Review of the microarray expression data for CK7 and CK20 and CEA showed low levels of expression for CK7 and high levels of expression for CK20, in sharp contrast to other tumors in the ovarian cohort (high CK7 and low CK20). The CEA data were not informative. Immunohistochemical stains for CK7, CK20, and CEA, performed to validate the expression data and to investigate the possibility that this tumor was metastatic to ovary, showed strong and diffuse tumor immunoreactivity for CK20 (Figure 3E), whereas CK7 showed no immunoreactivity (Figure 3F). CEA showed strong and diffuse immunoreactivity of the tumor cells and associated mucin (Figure 3G). These results offered strong support for the view that this tumor was a colonic adenocarcinoma metastatic to ovary rather than a primary ovarian adenocarcinoma.

Classification Using Known Diagnostic Markers

In practice, it is possible to discriminate most colon, ovary, and lung adenocarcinomas with histopathology and a limited immunohistochemical profile that includes markers CK7, CK20, and thyroid transcription factor (TTF)-1, in which colon tumors are CK7 (−), CK20 (+), and TTF-1 (−), ovary tumors are CK7 (+), CK20 (−), and TTF-1 (−), and lung tumors are CK7 (+), CK20 (−), and TTF-1 (+). In our cohort of tumors, using the array expression data for these three markers and using the same statistical approach, these markers correctly classified 138 of 154 samples, with 152 of 152 being the best possible outcome (see above).

Discussion

We have generated comprehensive gene expression profiles from 154 primary lung, colon, and ovarian adenocarcinomas and used two statistical methods, PCA and cross-validated prediction based on differentially expressed genes, to identify differences that allow discrimination of tumors in an organ-specific manner. Our results demonstrate strong discrimination of these tumors based solely on gene expression profiles.

A previous proof of principle gene expression-profiling study reported molecular discrimination of acute myelogenous leukemia from acute lymphoblastic leukemia.[5] Although these leukemias can be separated on morphological grounds in the majority of cases, there are tumors that require adjuvant methodologies, such as flow cytometry, for accurate diagnosis. Similarly, the current study compares gene expression profiles of histologically similar tumors, ie, adenocarcinomas, from three common sites that usually can be morphologically separated but sometimes pose diagnostic difficulties. For instance, the histopathological separation of primary ovarian carcinomas from metastases from a gastrointestinal source is a common diagnostic dilemma, as is illustrated by the outlier ovarian tumor in this study. This tumor, diagnosed as primary ovarian mucinous adenocarcinoma, was shown to be of probable colonic origin by immunohistochemical evaluation of low-molecular weight cytokeratin proteins 7 and 20 and CEA, markers with demonstrated diagnostic utility.[6] The ability of the expression profiles to identify this tumor as not belonging to the ovarian cohort further validates this approach. Thus, it would be interesting to determine the utility of this approach to the evaluation of mucinous tumors of the abdominal cavity of unknown origin.

The other nonconforming case, a primary invasive colorectal cancer, was similarly investigated and shown to be a colonic sarcoma by histopathology and immunohistochemical profile. This case also provides further validation of the gene expression profiles. Comparison of the expression data from the sarcoma to the colonic adenocarcinoma cohort illustrates some of the differences in specific gene expression. Examples of such differentially expressed genes include vimentin and enteric smooth muscle-γ2 actin, genes previously shown to be expressed in sarcomas and useful diagnostic immunohistochemical markers,[7,8] and other genes not previously shown to be expressed in sarcomas (for example, C-type lectin superfamily member 2; *CLECSF2*).

Many of the potential differentially expressed genes identified by this method have previously been shown to be either diagnostically useful and/or expressed in an organ-specific manner. TTF-1 is a nuclear protein expressed in pulmonary and thyroid epithelium,[9] plays a role in lung and thyroid development,[10] and is an immunohistochemical marker used to distinguish primary from metastatic lung adenocarcinoma.[11,12] TTF-1, as well as several surfactant-related genes, was identified as one of the genes whose expression is primarily restricted to lung adenocarcinomas. Similarly, the CDX1 and CDX2 genes, which encode intestine-specific transcription factors,[13,14] were identified as preferentially expressed in the colon tumors, as was CK20, a known marker of colonic adenocarcinoma.[6] Finally, the estrogen receptor 1 gene, known as a marker of breast and gynecological malignancies,[15]

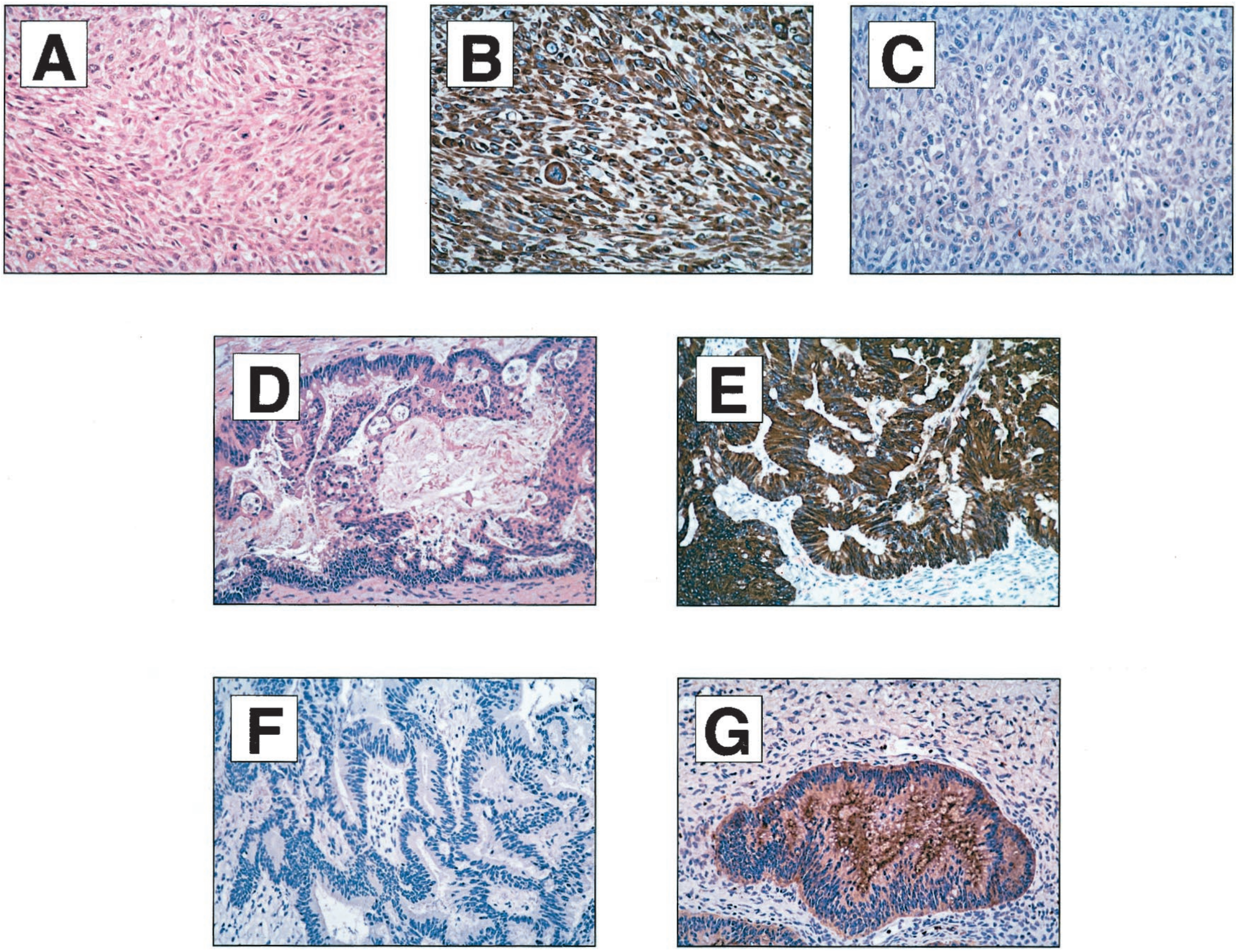

Figure 3. Histopathological and immunohistochemical analysis of colonic (**A–C**) and ovarian outlier tumors (**D–G**). **A:** H&E stain. **B:** Vimentin immunostain. **C:** Cytokeratin cocktail immunostain. **D:** H&E stain. **E:** CK20 immunostain. **F:** CK7 immunostain, and **G:** CEA immunostain. Original magnifications, ×200.

was identified as preferentially expressed in the ovarian tumors. The independent identification by this study of these known organ-specific markers provides strong validation of the utility of gene expression profiling as a highly effective gene discovery tool.

Using expression data for three genes commonly used as immunohistochemical markers in clinical practice, the molecular approach was able to correctly classify the large majority (91%) of tumors. However, the success of this relatively limited diagnostic work-up when compared with the analysis of thousands of genes by microarrays should not diminish the merits of global gene expression profiling. As molecular profiling is extended to a larger number of tumor types, it will be necessary to use additional numbers of genes to define organ-specific profiles. Furthermore, as shown by this study and others,[1,2,16] this approach is quite fruitful as a discovery tool to identify additional diagnostic markers.

The implications of this study are broad, suggesting that the establishment of gene expression profiles can be used to classify neoplasms in an organ-specific manner, one of the charges of the National Cancer Institute's "Director's Challenge" program for the molecular classification of cancer. The success of this study of three common tumor types suggests it will be feasible to extend this approach to a comprehensive cohort of tumors. The availability of such a comprehensive cancer map will profoundly impact clinical cancer care through improved diagnosis, prognosis, and treatment.

Acknowledgments

We thank Donita Sanders, Michelle Lizyness, Tina Fields, Susan Ahrens, Sylvia Zelenka, Joseph Bonner, and Lynn Tomsho for excellent technical and administrative support; Doug Selby, Doug Johnson, John Weeks, Michele LeBlanc, and Nancy McAnsh, technicians of the University of Michigan Comprehensive Cancer Center Tissue Core, for tissue procurement and immunohistochemistry; and Dr. Lora Hedrick Ellenson of Weill Medical College of Cornell University for contribution of a few ovarian tumors.

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tram T, Yu X, Powel JI, Yang L, Marti GE, Moore T, Hudson Jr J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Cahn WC,

Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403:503–511

2. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000, 406:536–540
3. Ripley B: Pattern Recognition and Neural Networks. Cambridge University Press, 1996
4. Sheibani K, Tubbs RR: Enzyme immunohistochemistry. Technical aspects. Semin Diagn Pathol 1984, 1:235–250
5. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286: 531–537
6. Chu P, Wu E, Weiss LM: Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms: a survey of 435 cases. Mod Pathol 2000, 13:962–972
7. Leader M, Collins M, Patel J, Henry K: Vimentin. An evaluation of its role as a tumor marker. Histopathology 1987, 11:63–72
8. Skalli O, Gabbiani G, Babai F, Seemayer TA, Pizzolato G, Schurch W: Intermediate filament proteins and actin isoforms as markers for soft tissue tumor differentiation and origin. Am J Pathol 1988, 130:515–531
9. Li C, Cai J, Pan Q, Minoo P: Two functionally distinct forms of NKX2.1 protein are expressed in the pulmonary epithelium. Biochem Biophys Res Comm 2000, 270:462–468
10. Lazzaro D, Price M, De Felice M, Di Lauro R: The transcription factor TTF-1 is expressed at the onset of thyroid and lung morphogenesis and in restricted regions of the foetal brain. Development 1996, 113:673–678
11. Pelosi G, Fraggetta F, Pasini F, Maisonneuve P, Sonzogni A, Iannucci A, Terzi A, Bresaola E, Valduga F, Lupo C, Viale G: Immunoreactivity for thyroid transcription factor-1 in stage I non-small cell carcinomas of the lung. Am J Surg Pathol 2001, 25:363–372
12. Reis-Filho JS, Carrilho C, Valenti C, Leitao D, Ribeiro CA, Ribeiro SG, Schmidt FC: Is TTF1 a good immunohistochemical marker to distinguish primary from metastatic lung adenocarcinomas? Pathol Res Pract 2000, 196:835–840
13. Ee HC, Erler T, Bhathal PS, Young GP, James RJ: Cdx-2 homeodomain protein expression in human and rat colorectal adenoma and carcinoma. Am J Pathol 1995, 147:586–592
14. Silberg DG, Furth EE, Taylor JK, Schuck T, Chiou T, Traber PG: CDX1 protein expression in normal, metaplastic, and neoplastic human alimentary tract epithelium. Gastroenterology 1997, 113:478–486
15. Brandenberger AW, Tee MK, Jaffe RB: Estrogen receptor alpha and beta mRBAs in normal ovary, ovarian serous cystadenocarcinoma and ovarian cell lines: down-regulation of ER-beta in neoplastic tissues. J Clin Endocrinol Metab 1998, 83:1025–1028
16. Welch JB, Zarrinkar PP, Sapinoso LM, Kern SG, Behling CA, Monk BJ, Lockhart DJ, Burger RA, Hampton GM: Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identified candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci USA 2001, 98:1176–1181