Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures

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

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 gastrointestinal tract, 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 of 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 immunohistochemistry (IHC), 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, gastrointestinal tract, 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 gastrointestinal 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 adenocarcinomas, 1 hepatocellular carcinoma, and 2 gastrointestinal adenocarcinomas. 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.

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 <200, 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.

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 Ps 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...
MOLECULAR CLASSIFICATION OF HUMAN CARCINOMAS

Materials and Methods

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, 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, 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, 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, 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, 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, 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, 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, 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, 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.
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

<table>
<thead>
<tr>
<th>True identity of unknown sample</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PR</td>
</tr>
<tr>
<td>PR</td>
<td>26 (0.564)</td>
</tr>
<tr>
<td>BL</td>
<td>8 (0.343)</td>
</tr>
<tr>
<td>BR</td>
<td>26 (0.267)</td>
</tr>
<tr>
<td>CO</td>
<td>23 (0.279)</td>
</tr>
<tr>
<td>GA</td>
<td>1 (0.115)</td>
</tr>
<tr>
<td>KI</td>
<td>1 (0.115)</td>
</tr>
<tr>
<td>LI</td>
<td>1 (0.045)</td>
</tr>
<tr>
<td>OV</td>
<td>1 (0.045)</td>
</tr>
<tr>
<td>PA</td>
<td>1 (0.045)</td>
</tr>
<tr>
<td>LA</td>
<td>1 (0.045)</td>
</tr>
<tr>
<td>LS</td>
<td>1 (0.045)</td>
</tr>
</tbody>
</table>

* 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).

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). On the basis of gene annotation alone, we recognized many well-described genes whose expression is elevated in tumors. These included MUC-2 and A53 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.

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 ≥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. 2A). 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 Fox5.1, 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. 2C). 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 specificities of the SVM methodology are different.

We found that classification of tumors arising in certain ana-

### Table 2 Prediction of tumor origin of metastatic carcinomas

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

* pap., papillary; ca., carcinoma; Ad, adenocarcinoma; SCC, squamous cell carcinoma.
MOLECULAR CLASSIFICATION OF HUMAN CARCINOMAS

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.

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 highlighted 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.

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

References


Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures


Updated version  Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/61/20/7388

Cited articles  This article cites 20 articles, 8 of which you can access for free at:
http://cancerres.aacrjournals.org/content/61/20/7388.full.html#ref-list-1

Citing articles  This article has been cited by 67 HighWire-hosted articles. Access the articles at:
/content/61/20/7388.full.html#related-urls

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.