

## Markers of Adenocarcinoma Characteristic of the Site of Origin: Development of a Diagnostic Algorithm

Jayne L. Dennis,<sup>1</sup> Torgeir R. Hvidsten,<sup>7</sup> Ernst C. Wit,<sup>2</sup> Jan Komorowski,<sup>7</sup> Alexandra K. Bell,<sup>3</sup> Ian Downie,<sup>4</sup> Jacqueline Mooney,<sup>4</sup> Caroline Verbeke,<sup>5</sup> Christopher Bellamy,<sup>6</sup> W. Nicol Keith,<sup>1</sup> and Karin A. Oien<sup>1,4</sup>

**Abstract Purpose:** Patients with metastatic adenocarcinoma of unknown origin are a common clinical problem. Knowledge of the primary site is important for their management, but histologically, such tumors appear similar. Better diagnostic markers are needed to enable the assignment of metastases to likely sites of origin on pathologic samples.

**Experimental Design:** Expression profiling of 27 candidate markers was done using tissue microarrays and immunohistochemistry. In the first (training) round, we studied 352 primary adenocarcinomas, from seven main sites (breast, colon, lung, ovary, pancreas, prostate and stomach) and their differential diagnoses. Data were analyzed in Microsoft Access and the Rosetta system, and used to develop a classification scheme. In the second (validation) round, we studied 100 primary adenocarcinomas and 30 paired metastases.

**Results:** In the first round, we generated expression profiles for all 27 candidate markers in each of the seven main primary sites. Data analysis led to a simplified diagnostic panel and decision tree containing 10 markers only: CA125, CDX2, cytokeratins 7 and 20, estrogen receptor, gross cystic disease fluid protein 15, lysozyme, mesothelin, prostate-specific antigen, and thyroid transcription factor 1. Applying the panel and tree to the original data provided correct classification in 88%. The 10 markers and diagnostic algorithm were then tested in a second, independent, set of primary and metastatic tumors and again 88% were correctly classified.

**Conclusions:** This classification scheme should enable better prediction on biopsy material of the primary site in patients with metastatic adenocarcinoma of unknown origin, leading to improved management and therapy.

Most cancers present at their site of origin—that is, it is the primary tumor which causes symptoms in the patient who then attends their doctor. Some 10% to 15% of cancers, however, present as metastases in solid organs, body cavities or lymph nodes (1). Most of these secondary tumors are adenocarcinomas (1–4), for which the seven commonest primary sites are

breast, colon, lung, ovary, pancreas, prostate, and stomach (1). The prognosis and therapy of patients with metastatic adenocarcinoma are linked to the site of origin, so these sites, and others, are investigated by clinical examination, radiology, and serum tumor markers (2–4). If no primary cancer is found, then the metastatic deposit is usually biopsied, to confirm the diagnosis of malignancy and to subtype the tumor. Unfortunately, adenocarcinomas metastatic from different locations have similar microscopic appearances, which confound identification of the primary site (5). Patients with metastatic adenocarcinoma of unknown origin make up around 3% of all cancer patients and this category is among the 10 most common malignancies (4).

Traditionally, cancer classification has been based on clinical criteria and histopathology, which is itself based on tissue and cell morphology. These phenotypic abnormalities, however, are underpinned by changes in gene expression, at the mRNA and ultimately protein levels (6). The latter can be exploited in tissue samples through immunohistochemistry, which is already the standard method for lymphoma categorization (7). Our aim was to establish a similar classification scheme for adenocarcinomas.

For diagnostic pathology, ideally we want to: use as few markers as possible; correctly predict origin for many tumors (high coverage—sensitivity); and misclassify few tumors (high accuracy—specificity; refs. 8, 9). A number of tumor markers are already in use for this purpose (10, 11), including

**Authors' Affiliations:** <sup>1</sup>Cancer Research UK Centre for Oncology and Applied Pharmacology; <sup>2</sup>Department of Statistics and <sup>3</sup>Division of Cancer Sciences and Molecular Pathology, University of Glasgow; <sup>4</sup>Department of Pathology, Glasgow Royal Infirmary, Glasgow; <sup>5</sup>Department of Pathology, St. James' Hospital, Leeds; <sup>6</sup>Department of Pathology, University of Edinburgh, Edinburgh, United Kingdom; and <sup>7</sup>Linnaeus Centre for Bioinformatics, University of Uppsala, Uppsala, Sweden

Received 11/3/04; revised 2/8/05; accepted 3/1/05.

**Grant support:** Cancer Research UK through Departmental core funding, a Clinician Scientist Fellowship (K.A. Oien), and Program Grant (W.N. Keith); the University of Glasgow through the Florence Houston Bequest (J.L. Dennis); the European Union for an Access to Infrastructure Grant (K.A. Oien and J.L. Dennis through J. Komorowski); and Keystone Symposia for a Travel Grant (J.L. Dennis).

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.

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

**Requests for reprints:** Karin A. Oien, Centre for Oncology and Applied Pharmacology, Cancer Research UK, Beatson Laboratories, University of Glasgow, Garscube Estate, Switchback Road, Glasgow G61 1BD, United Kingdom. Phone: 44-141-330-3506; Fax: 44-141-330-4127; E-mail: k.oien@beatson.gla.ac.uk.

©2005 American Association for Cancer Research.

cytokeratins 7 and 20 (CK7 and CK20) and thyroid transcription factor 1 (TTF1; refs. 12, 13). However, we wished first to ensure that all potential candidates were evaluated. In a previous study, we used hierarchical clustering to show that the molecular profiles of the seven most common adenocarcinomas are characteristic of their primary site, and furthermore, that this is maintained during metastasis (14). We then developed a comprehensive bioinformatic approach that identified key differences between primary sites (14).

In the present study, by profiling selected candidate site-specific markers, we aimed to develop a scheme for the immunohistochemical evaluation of adenocarcinomas that improves the prediction of primary sites. The classifier would need to be applicable in routine diagnostic practice and so its utility was validated in a second independent set of primary and metastatic tumors. High-throughout expression profiling at a histologic level was enabled by the use of tissue microarrays (15).

## Materials and Methods

Ethical approval for the use of clinical material in this research was given by the North Glasgow University Hospitals National Health Service Trust.

**Selection of candidate markers.** Through bioinformatic analyses, we identified 61 candidate markers which appeared to be either specific to one or restricted to two or three primary sites (14). Of these, 14 were taken forward to tissue profiling: carcinoembryonic antigen, chromogranin A, clusterin, fatty acid binding protein, immunoglobulin- $\kappa$  light chain, insulin, mammaglobin 1, pancreatic polypeptide, pepsinogen C, prostate-specific antigen, prostatic acid phosphatase, surfactants A and B, and trefoil factor 2. This was integrated with the profiling of 13 immunohistochemical markers which are clinically established or have more recently emerged from the literature: CA125, CA19-9, CDX2, CK7, CK20, estrogen receptor, gross cystic disease fluid protein 15 (GCDFFP-15), hepatocyte, lysozyme, mesothelin, mucin 5AC, trefoil factor 1, and TTF1 (8, 9, 12, 13, 16–27).

**Selection and assembly of tumors for tissue profiling.** Our focus was on adenocarcinomas from the seven main sites. The sample numbers were proportional to their frequency of presentation with metastatic disease, rather than the overall incidence (1). The absolute numbers were chosen to yield statistical power, assuming marker sensitivity and specificity of at least 65% and 75%. The first round comprised 352 primary tumors in four tissue microarrays, with 261 adenocarcinomas from the main sites: 35 breast (28 ductal and 7 lobular), 47 colon, 46 lung, 28 ovary (18 serous and 10 mucinous), 53 pancreas, 18 prostate, and 34 stomach tumors (26 intestinal and 8 diffuse). Other tumors which arise at the same or nearby sites or enter the differential diagnosis were also included, as listed in Fig. 1A. The second round comprised 100 primary and 30 paired metastatic tumors, from the seven main locations only, in two tissue microarrays, with the numbers of primaries and metastases, respectively, given in brackets: 17 breast (10 + 3 ductal and 3 + 1 lobular), 22 colon (17 + 5), 26 (20 + 6) lung, 13 ovary (7 + 2 serous and 3 + 1 mucinous), 27 pancreas (20 + 7), 7 prostate (all primary), 18 stomach (10 + 4 intestinal and 3 + 1 diffuse).

Cases were identified from pathology records. Tissue microarrays were constructed using Beecher Instruments' Manual Tissue Arrayer and 0.6 mm punches (Silver Spring, MD). Test samples were arrayed in triplicate (28), alongside duplicate internal control cores. The tissue microarray blocks were sectioned onto slides using the Paraffin Tape-Transfer System, Instrumedics, Inc. (Hackensack, NJ). Representative sections were checked by H&E staining.

**Immunohistochemistry.** Twenty-six primary antibodies were used according to the supplier's instructions for dilution and antigen retrieval: 10 as detailed in Table 1, plus a further 16: CA19-9, clusterin,

fatty acid binding protein, trefoil factor 2, and surfactant protein B from Novocastra Laboratories, Ltd. (Newcastle upon Tyne, United Kingdom); carcinoembryonic antigen, chromogranin A, hepatocyte, immunoglobulin  $\kappa$ , prostatic acid phosphatase, and surfactant protein A from Dako, Ltd. (Ely, United Kingdom); trefoil factor 1 and mucin 5AC from Lab Vision, Ltd. (Newmarket, United Kingdom); pepsinogen C from Abcam, Ltd. (Cambridge, United Kingdom); pancreatic polypeptide from Harlan Sera-Lab (Loughborough, United Kingdom); and insulin from Diagnostics Scotland (Edinburgh, United Kingdom).

Immunohistochemistry was done by the standard indirect method, using Envision ChemMate from Dako, according to the manufacturer's instructions. The tissue microarray sections required longer dewaxing in xylene (20 minutes). Antigen retrieval was done enzymatically [0.1% trypsin/0.1% calcium chloride in Tris buffer (pH 7.6) at 37°C for 25 minutes] or with heat [0.005% EDTA in Tris buffer (pH 8.0) in a microwavable pressure cooker for 5 minutes]. Immunohistochemistry was done on a Dako Autostainer in the first round and manually by a different scientist in the second round.

**In situ hybridization.** IMAGE clones for mammaglobin 1 (109005) and pepsinogen C (10187) were obtained from the Medical Research Council Geneservice (Babraham, United Kingdom). Plasmid preparation was done and clones confirmed by sequencing. The template was linearized then digoxigenin-UTP-labeled riboprobes were generated by transcription with T7 or Sp6 RNA polymerase, using the DIG RNA Labelling Kit from Roche Diagnostics, Ltd. (Lewes, United Kingdom). *In situ* hybridization was done as previously described (29).

**Microscopic assessment and scoring.** The stained slides were assessed without knowledge of other clinical or pathologic data. For each sample, the three cores were scored individually and the highest figure used (30). The first round slides were scored separately by two observers (J. Dennis and K. Oien, a consultant histopathologist); cores with differing results were reviewed and agreement reached. The second round was assessed by one observer (K. Oien). The intensity of staining was scored on a categorical scale, from 0 to 3, where: 0 was absent; 1 was very weak, dubious staining; 2 was definite, mild, or moderate staining; and 3 was definite, strong staining (16, 26, 31). Only staining in tumor cells, in the appropriate nuclear or cytoplasmic location, was scored. Representative images of positive staining are presented in Supplementary Fig. 1.

**Data analysis.** The data were stored and analyzed using Microsoft Access. To identify subtle associations, we used Rosetta, available online at <http://www.idi.ntnu.no/~aleks/rosetta> (32). Rosetta is a software package implementing a supervised learning method, based on rough set theory and Boolean reasoning. It induces minimal decision rules, where each rule is constructed using the minimal information needed to discern objects. Sensitivity and specificity were calculated using standard formulae (16).

## Results

**Tissue profiling of candidate markers in primary adenocarcinomas.** For the first round of expression profiling, we studied 352 primary tumor samples with 27 candidate markers: 25 by immunohistochemistry alone, 1 by *in situ* hybridization (mammaglobin 1) alone, and 1 by both (pepsinogen C). Staining was scored blindly: for analysis, we regarded only convincing staining (scores of 2 and 3) as positive (16, 26, 31).

Our aim was then to select the best markers, and the best combination in which to apply them: the ideal would be to have one highly sensitive and specific marker for each of the seven sites. First, we analyzed the database manually: some associations were obvious, such as prostate cancer and its close-to-ideal marker prostate-specific antigen. To identify more subtle associations, we used the Rosetta system, which produces minimum rules, which describe and separate groups of objects,

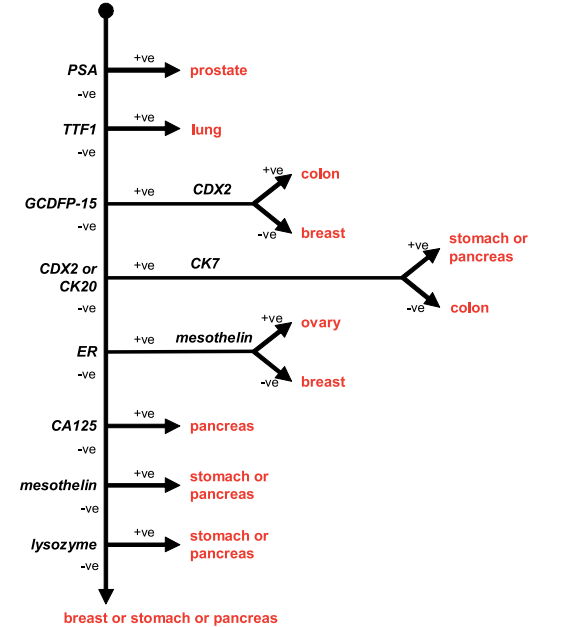
**A** Table of immunohistochemistry results from first round: ten selected markers

Primary site	No.	% Positivity of each marker									
		PSA	TTF1	GCD FP15	CDX2	CK20	CK7	ER	Meso thelin	CA 125	Lyo zyme
Breast	35	0	0	54	0	0	83	77	3	6	14
Ovary serous	18	0	0	6	0	0	89	83	94	89	0
Ovary mucinous	10	10	0	10	20	30	40	40	30	20	20
Endometrial adenoca	10	10	0	10	0	0	60	30	50	90	10
Pancreas	53	0	2	2	0	19	96	0	47	53	51
Ampullary adenoca	6	0	0	17	17	83	83	0	83	67	67
Cholangiocarcinoma	10	0	0	0	10	30	80	0	70	50	90
Stomach	34	3	3	0	18	18	35	0	21	9	85
Esophageal adenoca	21	0	0	0	10	33	48	0	19	10	57
Colon	47	0	0	9	83	68	4	2	2	0	53
Prostate	18	100	11	0	0	0	11	0	0	0	6
Lung	46	0	91	4	2	2	91	9	39	39	43
Lung small cell ca	3	0	100	0	0	0	0	0	0	0	0
Lung squamous ca	7	0	57	0	0	0	43	0	14	14	14
Mesothelioma	6	0	0	0	0	0	67	0	50	17	17
Esophageal squ ca	7	0	0	0	0	0	0	0	0	0	0
Renal cell ca	15	0	0	0	0	0	0	0	0	0	7
Hepatocellular ca	6	0	0	0	0	0	0	0	0	0	0

**B** Sensitivity and specificity of each marker in first round of immunohistochemistry

Marker	Tumors of	% Sensitivity	% Specificity
PSA	Prostate	100	99
TTF1	Lung	91	98
CDX2	Colon	83	96
CDX2	Colon and stomach	56	98
CK20	Colon	68	91
CK20	Colon, stomach and pancreas	36	97
GCDFP-15	Breast	54	96
ER	Breast and ovary	74	95
CA125	Ovary and pancreas	88	88
Mesothelin	Ovary and pancreas	85	85
Lysozyme	Stomach and pancreas	65	69
CK7	(Stomach and pancreas) versus colon	72	96

**C** Decision tree derived from first round data



**D** First round data: site of origin predicted from table

Actual primary site	Predicted primary site									
	Breast	Colon	Lung	Ovary	Panc-reas	Panc OR Stomach	Stomach	Prostat e	Nega-tive	Missing
Breast	31			1					1	2
Colon	2	44				1				
Lung		1	41		1	2				1
Ovary serous				18						
Ovary mucinous		2		5			1	1		1
Pancreas	2		1	5	1	39			1	4
Stomach		2	1			21	5			4
Prostate								18		

**E** First round data: site of origin predicted from decision tree

Actual primary site	Predicted primary site									
	Breast	Colon	Lung	Ovary	Panc-reas	Panc OR Stomach	Pa, St OR Brst	Prostat e	Nega-tive	Missing
Breast	28					1	5			1
Colon	2	43				1	1			
Lung		1	41		2	1				1
Ovary serous	1			14	2	1				
Ovary mucinous	2	3		2		1		1		1
Pancreas	1		1		21	23	4			3
Stomach		7	1		3	21	1	1		
Prostate								18		

**F** Second round data: site of origin predicted from table

Actual primary site	Predicted primary site									
	Breast	Colon	Lung	Ovary	Panc-reas	Panc OR Stomach	Stomach	Prostat e	Nega-tive	Missing
Breast	13	3			1					
Colon		15	5			1				1
Lung			18	5		1	1			1
Ovary serous				6	2					1
Ovary mucinous	1			2	1					
Pancreas		1	1	1	2	1	16	2		2
Stomach		1	1			11	2	1		2
Prostate								7		

**G** Second round data: site of origin predicted from decision tree

Actual primary site	Predicted primary site									
	Breast	Colon	Lung	Ovary	Panc-reas	Panc OR Stomach	Pa, St OR Brst	Prostat e	Nega-tive	Missing
Breast	13	3			1					
Colon		16	5			1	1	1		1
Lung			18	5		1	1	1		1
Ovary serous				6	1	1				1
Ovary mucinous	1			1	1					
Pancreas		1	1	1	2	7	3	9	1	2
Stomach		2	1			1	10	2	1	
Prostate								7		

**H** Table of immunohistochemistry results for primary tumors from both rounds

Primary site	No.	% Positivity of each marker									
		PSA	TTF1	GCD FP15	CDX2	CK20	CK7	ER	Meso thelin	CA 125	Lyo zyme
Breast	52	0	0	49	0	0	87	79	4	13	11
Colon	69	0	0	6	86	78	3	1	7	1	43
Lung	72	0	90	3	1	1	90	6	36	38	35
Ovary serous	27	0	0	4	0	0	93	81	96	93	7
Ovary mucinous	14	7	0	14	14	21	57	50	36	36	29
Pancreas	80	0	4	1	2	15	94	0	49	50	51
Stomach	52	2	2	0	21	21	50	0	19	8	81
Prostate	25	100	8	4	0	0	0	8	0	0	4

**Fig. 1.** A, table of immunohistochemistry results from the first round of tissue profiling for the 10 selected markers. The data for the remaining 17 markers are provided in Supplementary Table 1. The percentage positivity of each marker in each tumor type is given as a number and as background shading. B, the sensitivity and specificity of each marker. C, this diagnostic decision tree was built using the ranked specificities and sensitivities. To apply this, we start at the top. If the tumor is prostate-specific antigen – positive, then it is from prostate; if not, then we move down the tree, and so on. Pancreatic and gastric tumors mostly had to be grouped together. D-G, the results of applying the diagnostic table and decision tree, respectively, are presented in D and E for the first round data and F and G for the second round. In this format, termed a confusion matrix, correct predictions lie along the diagonal, shaded black; the incorrect predictions lie in the white areas. “Missing” tumors for the tree are those lacking any marker along its decision-making pathway, and for the table are those lacking at least five assessable markers. In addition, some tumors lacked any positive markers: these tumors have been called “negative” in the matrix for the diagnostic table and are regarded as errors when calculating predictive success; with the tree, such tumors fall into the “breast or stomach or pancreas” bin at the bottom. The metastatic tumors in the second round are shown in red. H, table of immunohistochemical results combined from both first and second rounds of tissue profiling.

**Table 1.** Details of selected markers and antibodies used

Antibody target	Clone	Source	Dilution	Antigen retrieval
CA125	Ov185:1	Novocastra	1:400	Heat
CDX2	AMT28	Novocastra	1:100	Heat
CK7	OV-TL	Dako	1:500	Heat
CK20	Ks20.8	Dako	1:500	Heat
Estrogen receptor	6F11	Novocastra	1:50	Heat
GCDFP-15	23A3	Novocastra	1:20	Heat
Lysozyme	Polyclonal	Dako	1:5,000	Enzymatic
Mesothelin	5B2	Novocastra	1:20	Heat
Prostate-specific antigen	ER-PR8	Dako	1:100	Heat
TTF1	SPT249	Novocastra	1:50	Heat

and which are ranked by specificity and sensitivity (32). To illustrate this process, for lung, candidate markers included TTF1 and surfactant proteins A and B. All three markers stained around 90% of lung cancers and thus were highly sensitive. Whereas TTF1 and surfactant A were essentially limited to lung tumors, surfactant B was also present in up to 13% of other adenocarcinomas and thus was less specific. TTF1 was selected as the lung marker on the basis of its strong and easily discriminated nuclear staining. In contrast with these lung candidates, the positive markers for breast, estrogen receptor, and GCDFP-15, stained lower proportions of breast cancers but seemed to be independent, such that some tumors were positive only with one or the other, hence both markers were selected (24).

This analysis led to a smaller panel of 10 markers: CA125, CDX2, CK7, CK20, estrogen receptor, GCDFP-15, lysozyme, mesothelin, prostate-specific antigen, and TTF1. The immunohistochemical results for these markers are presented in Fig. 1A. In breast, ovary and stomach, major tumor subtypes exist and were assessed: the expression profiles differed between subtypes only for ovary. The percentage positivity is given as a number and as background shading, for easy visual comparison. The sensitivity and specificity of each marker is given in Fig. 1B. Note that sensitivity is simply the number (or shading intensity) in the appropriate cell in Fig. 1A; and specificity is the (lack of) shading in the column for each marker outside the "correct" cell. Representative images of positive staining for the 10 selected markers are presented in Supplementary Fig. 1. The immunohistochemical results for the remaining 17 candidates are provided in Supplementary Table 1. In addition, the marker data have been summarized in a second graphic format, through the application of hierarchical clustering: the resulting heat-maps are shown in Supplementary Fig. 2.

As already indicated, a few markers, on their own, were diagnostic of a single primary site: prostate-specific antigen for prostate; and TTF1 for lung. Some markers, on their own, were diagnostic of a subset of tumors: estrogen receptor for breast or ovary; and CDX2 for gastrointestinal tumors (ovarian mucinous tumors excluded and discussed below). Other markers, in combination, were diagnostic of a single tumor site: estrogen receptor, mesothelin, and CA125 together indicate ovary. Some tumors had similar profiles that were difficult to separate: pancreas and stomach (between these two, CK7 negativity made pancreas unlikely; and CA125

positivity, especially with mesothelin positivity, made pancreas more likely).

Primary carcinomas which arise at the same sites or might enter the differential diagnosis of metastatic deposits were also studied: their expression profiles resembled the anatomically and developmentally closest "main" adenocarcinoma. Thus, endometrial adenocarcinoma was similar to ovarian cancer, and adenocarcinomas of the duodenal ampulla and bile ducts (cholangiocarcinoma) resembled pancreatic ductal tumors. Hepatocellular carcinomas expressed none of these markers: because the liver is commonly involved in metastatic spread, this facilitates the distinction between primary and secondary tumors.

**Development and application of a diagnostic classification scheme.** Pathologists are accustomed to using information in table format, where all data is considered simultaneously in making a diagnosis. However, we decided to also build a decision tree (Fig. 1C), using the ranked specificities and sensitivities of the markers. To apply this, we start at the top, with the most specific marker. If the tumor is prostate-specific antigen-positive, then it is from the prostate; if not, then we move down the tree, and so on. Pancreatic and gastric tumors mostly had to be grouped together. An alternative decision tree was built using the conventional statistical approach of recursive partitioning; it is shown for comparison in Supplementary Fig. 3, but was not selected for further use.

We then returned to our first round data and attempted to predict the primary site using these classifiers. The decision tree was used as depicted (Fig. 1C). With the diagnostic table, in most cases, a single site could be assigned. In a few cases, however, two sites seemed equally likely, and agreement of either with the actual site was regarded as correct. The four common pairs were: stomach or pancreas, stomach or colon (CDX2 and/or CK20 positive but also mesothelin and/or CA125 positive), ovary or pancreas (CK7, mesothelin, and CA125 positive but negative for estrogen receptor), and breast or ovary (estrogen receptor and CK7 positive but negative for GCDFP-15 and either mesothelin or CA125).

The results for these first round predictions are presented in Fig. 1D and E. The primary site was correctly predicted in 88% with the table and in 87% with the tree. The incorrect predictions could be divided into four main types. First, with both methods, errors were commonly between the tumor pairs

above, where one site alone had been predicted. Second, where diagnosis relies on a single highly specific marker, such as TTF1 in lung, the few negative cases are misclassified. Third, some tumors simply fail to stain with any marker: using the decision tree, such tumors fall into the "bin" at the bottom (breast or stomach or pancreas). Except for the lack of immunohistochemical staining, these tumors did not seem histologically distinct from the majority: for example, they were not particularly poorly differentiated. Fourth, ovarian mucinous tumors are an oddity. They are well-recognized to express gastric, colonic, and/or pancreatic markers and so it is not surprising that their expression profile overlaps with those of gastrointestinal tumors (33).

**Testing of classification scheme in an independent set of tumors.** We then validated the diagnostic table and decision tree in a separate set of primary tumors and extended the analysis to metastatic tumors. Use of independent samples checks for "over-fitting" of the classification scheme to the original data. We constructed two tissue microarrays, containing 100 primary tumors and 30 paired metastases, from the seven main sites. The tissue microarrays were stained with the 10 markers, scored, and the primary site predicted using table and tree separately.

The results for these second round predictions are presented in Fig. 1F and G. The primary and metastatic tumor pairs had similar expression profiles, showing that our data from the primary tumors are applicable not only to other data sets but also to metastatic adenocarcinomas. Correct assignment was obtained again in 87% of primary and metastatic tumors with the table and in 89% with the tree: the figures for the metastases alone were 71% and 83%, respectively. Note that the metastatic set lacked tumors from the prostate, from which the primaries were always correctly classified: their absence alone lowers the success rate of the classifier. Most of the misclassified tumors were in pairs (i.e., a primary and its matched metastasis), with the incorrect predictions as described for the original data set. The immunohistochemical results from both rounds are combined in Fig. 1H.

## Discussion

Adenocarcinoma of unknown origin are a common clinical problem, so we are far from the first to address this diagnostic dilemma from a pathologic standpoint (8–11, 16, 19, 26, 34). However, over the past few years, a number of novel markers have emerged, including CK7, CK20, TTF1, CDX2, and mesothelin (12, 13, 17, 22, 23, 25). We believe that this is the first study to profile large numbers of these newer genes alongside established markers in a comprehensive manner, and to develop a rational classification scheme which has been successfully tested in an independent data set including metastatic tumor samples.

During this research, we performed both internal and external validation. The data were shown to be reproducible through: repetition of staining and scoring, staining by different scientists and methods in the first and second rounds, and by two assessors in the first round. The study design is comparable with the literature: the use of triplicate cores is regarded as standard (28), our numbers of missing or otherwise unassessable cores is low (3%; refs. 28, 31), and our scoring system is well established (16, 26, 31).

In general, our results agree with previous studies on the primary sites and individual markers tested here, as shown in Table 2 (8, 9, 12, 13, 16, 17, 19–27, 35–37), although for some markers (CK20 and CDX2), our sensitivities are at the lower end of the range. For CK20, most previous studies have used full tissue sections rather than tissue microarrays (13, 35, 36), and although there is generally concordance of staining between whole sections and tissue microarrays, expression levels of some genes are lower in tissue microarrays (31). Because the present classification is likely to be used primarily with small biopsy material, tissue microarrays arguably better reflect the situation in practice. For CDX2, Werling et al. (17) found widespread expression in the gut, with strong, uniform nuclear staining in colon, and weaker, more heterogeneous staining in stomach and pancreas, whereas we and others found that CDX2 was relatively specific for colon, with convincing staining absent from most gastric tumors, and in particular from all pancreatic adenocarcinomas (Fig. 1A and Table 2; ref. 25). This apparent discrepancy may simply indicate that our scoring criteria are more stringent, thus excluding heterogeneous, weak staining of non-colonic tumors, and illustrates the eternal difficulty of determining what threshold should be used for positivity (38).

Prostate-specific antigen and TTF1 positivity in a few non-prostatic (mainly gynecologic) and non-lung tumors, respectively, was surprising, and emphasizes that the study was fully "blinded". In fact, Dako's datasheet indicates that prostate-specific antigen is expressed outside the prostate, for example in endometrium. TTF1 positivity could represent a hitherto unrecognized neuroendocrine carcinoma component (12), but a more mundane explanation may apply: in the second round, for example, one tumor pair was submitted as pancreatic cancer with a lung metastasis, but both expressed TTF1, thus it is likely that the primary was in fact the lung cancer.

The use of decision trees is not yet established in pathology, but intuitively developed algorithms for this application have been described (8, 9). DeYoung et al. (8) emphasized that markers must be interpreted in a particular sequence for their predictive values to be maximized, so that the pathologist moves from the most specific to the least specific markers, as with our decision tree (Fig. 1C). In this way, as tumor sites are excluded, even relatively nonspecific markers further down the tree yield useful information (8).

In the development of any classification scheme, a figure of 88% correct in a data set with seven classes would be regarded as excellent. Clearly, this leaves some 12% of tumors which were not correctly classified. Likely reasons have already been described in Results, but note that, even in these cases, the list of likely primary sites has often been narrowed down. Importantly, our predictions were based purely on the numerical immunohistochemical scores, in isolation from both clinical information, such as gender, site of metastasis, and radiology, and the microscopic appearance of the tumor. In practice, both may be helpful. For example, on histology, the presence of "signet ring" cells suggests a gastric or breast origin, calciphosphates suggest ovary and "dirty" necrosis suggests colon. The integration of such data may overcome some shortcomings of the present approach, for example, the difficulty in discriminating certain tumor pairs with immunohistochemistry alone. The question has been addressed by Sheahan et al. (5),

**Table 2.** Comparison of immunohistochemistry results from first round with the literature

	Prostate-specific antigen	TTF1	GCDFP-15	CDX2	CK20	CK7	Estrogen receptor	Mesothelin	CA125	Lysozyme
Breast	<b>0, 0</b>	0, 0, <b>0</b>	33, <b>52, 54</b> , 62, 67, 72, 74, 77	0, 0, <b>0</b>	0, 0, 0, 0, <b>0, 4, 7, 19</b>	70, <b>83, 89</b> , 93, 93, 96	32, 33, 58, 60, 63, <b>73, 77</b>	3, <b>3, 6</b>	<b>6, 13, 13, 13</b> , 23, 24	<b>14</b>
Colon	<b>0, 0</b>	0, 0, <b>0</b>	0, 0, 0, 0, 0, 0, 0, <b>9</b>	<b>83, 90, 99</b>	65, <b>68, 73</b> , 84, 88, 92, 92, 93, 100	<b>4, 5, 6, 7</b> , 9, 16, 23	0, 0, 0, 0, <b>2, 13</b>	<b>2, 4, 22</b>	<b>0, 4, 4, 9</b> , 10, 13	<b>53</b>
Lung	<b>0, 9</b>	66, 75, <b>91</b>	0, 0, 0, 0, <b>4, 4, 6</b>	0, 0, <b>2</b>	0, <b>2, 8, 9</b> , 10, 10	<b>91, 96</b> , 100, 100	0, 0, 3, <b>9, 11</b>	22, 24, <b>39</b>	20, 20, 20, <b>35, 39</b>	<b>43</b>
Ovary	<b>0, 0</b>	<b>0, 0</b>	0, 0, 0, 0, <b>4, 4, 6</b>	0, 0, <b>0</b>	0, 0, 0, 4, 19	83, <b>89, 91</b> , 100, 100	4, 12, 34, <b>50, 83</b>	<b>94, 95, 100</b>	61, 63, 80, <b>89, 91, 96</b>	<b>0</b>
Pancreas	<b>0, 0</b>	0, 0, <b>2</b>	0, 0, 0, 2	0, 0, 32	0, <b>19, 35</b> , 39, 44, 62	87, 92, <b>95, 96</b>	0, 0, <b>0</b>	<b>47, 75, 86</b>	48, 48, <b>53</b>	<b>51</b>
Stomach	<b>0, 3</b>	<b>0, 3</b>	0, 0, 0, 3	<b>18, 20, 70</b>	<b>18, 30, 41, 50</b> , 51, 54, 68	<b>35, 38, 51</b> , 60, 71	0, 0, 2	10, <b>21, 29</b>	7, <b>9, 11</b>	<b>38, 85</b>
Prostate	<b>96, 100</b>	0, 0, <b>11</b>	<b>0, 10, 15</b>	<b>0, 1, 4</b>	0, 0, <b>0, 21</b>	0, <b>0, 12, 29</b>	10, <b>11</b>	0, 0, <b>0</b>	<b>0, 2, 2</b>	<b>6</b>

NOTE: Data from the literature in regular font, from refs. (8, 9, 12, 13, 16, 17, 20–27, 35–37); our data in boldface. Diffuse staining taken as positive in refs. (13, 22, 23). Ovarian tumors are serous (or at least nonmucinous).

who studied the contributions of morphology and clinical data to diagnosis. Two pathologists evaluated 100 metastatic adenocarcinomas from 10 primary sites. On morphology alone, the correct primary site was the first choice in 26% for both pathologists. Addition of gender and site of metastasis enabled the correct first choice in 50% and 55%, respectively (5).

These predictions improve with the use of immunohistochemistry, although surprisingly few comprehensive studies have been published. DeYoung et al. (8) studied over 2,800 epithelial malignancies (a wide range including germ cell tumors, mesotheliomas, and solid carcinomas) and found that a panel of 14 markers yielded the correct primary site in 67%. Brown et al. (26) studied metastatic adenocarcinomas from five sites (colon, breast, ovary, lung, and upper gastrointestinal tract). Using four markers (carcinoembryonic antigen, CA19-9, CA125, and BCA225), the primary site was correctly predicted in 66%; we have profiled three of these, but only CA125 was among the most informative. Legendijk et al. (9) studied primary and metastatic adenocarcinomas from three sites (colon, breast, and ovary) and found that six markers (CK7, CK20, CA125, carcinoembryonic antigen, estrogen receptor, and GCDFP-15) yielded correct classification in 80% to 90%. Interestingly, our difficulty in first identifying and then separating out gastric and pancreaticobiliary cancers is a general experience (5, 26). In clinical practice, this is in fact not a significant problem. Diffuse (signet-ring cell) gastric cancers can be identified as such on histology, whereas the primary tumors of intestinal-type gastric cancers are generally visible at endoscopy; and if not already performed, radiologic assessment may also provide clarification. Thus, upper gastrointestinal and pancreatico-biliary cancers can generally be separated on clinical grounds once this differential diagnosis is raised on pathologic material.

It is important to establish that markers informative for the site of origin in a primary malignancy retain that utility in

metastases. Our site-specific genes are essentially related to epithelial differentiation, rather than to tumor proliferation or invasion. For these markers, our current and previous data certainly suggest that the primary tumor and corresponding metastasis share similar profiles (14), and this has been the experience of others (9, 16). Clearly, further testing in biopsy samples from additional metastatic samples and from metastases of unknown origin would be worthwhile.

The dissemination of any diagnostic classification to other laboratories clearly raises challenges already being faced with testing for therapeutic markers such as estrogen receptor in breast cancer (28). Categorization of a marker as positive or negative is influenced by technical issues such as specimen fixation, tissue processing, and immunohistochemical conditions including antibody dilution and pretreatment, as well as by interpretative aspects including the observer's threshold for positivity and other variables (8, 35). Because the decision tree is based on marker specificity and has generalized well to unseen cases (in the second round), it should be reasonably robust for transfer to other centers, and it is flexible so that new markers may be added (8). Of course, in practice, any immunohistochemical classification should be interpreted in the appropriate histologic and clinical context.

Identification of the primary site in patients with metastatic adenocarcinoma is important for prognosis and rational therapy (2–4). The nihilism of the past has changed because of the emergence of relatively specific and increasingly effective chemotherapeutic regimens for metastatic adenocarcinomas from each site (39). Breast and prostate cancers may respond to hormonal manipulation, platinum- or taxane-based regimens are used with ovarian and lung tumors, 5-fluorouracil-based combinations are the mainstay of colorectal cancer management, and combination therapy with 5-fluorouracil and a platinum agent is the most widely used first-line

regimen for gastric cancer. In contrast, gemcitabine alone is the treatment of choice for palliation of pancreatic cancer but would be regarded as suboptimal therapy for other tumors. Thus, prediction of the likely primary site enables the selection of a regimen optimal for antitumor activity, with its implications for palliation and possibly survival, whereas avoiding unnecessary toxicity.

For the investigation of patients with occult primary disease, it is generally accepted that immunohistochemical assessment of biopsy material is both useful and cost-effective, compared with exhaustive investigations using other modalities, predom-

inantly radiology and endoscopy (2–4, 8, 39). We present here a validated approach to the immunohistochemical evaluation of small tumor samples that should enable the assignment of a likely site of origin for most patients with metastatic adenocarcinoma.

### Acknowledgments

We thank Rod K. Ferrier for technical help with *in situ* hybridization, Jane Hair for searches of pathology records, and Drs. Euan Stronach for performing hierarchical clustering and Jeffrey Evans for oncologic advice.

### References

- Nystrom JS, Weiner JM, Heffelfinger-Juttner J, Irwin LE, Bateman JR, Wolf RM. Metastatic and histologic presentations in unknown primary cancer. *Semin Oncol* 1977;4:53–8.
- Hillen HFP. Unknown primary tumours. *Postgrad Med J* 2000;76:690–3.
- Varadhachary GR, Abbruzzese JL, Lenzi R. Diagnostic strategies for unknown primary cancer. *Cancer* 2004; 100:1776–85.
- Pavlidis N, Briasoulis E, Hainsworth J, Greco FA. Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Cancer* 2003;39: 1990–2005.
- Sheahan K, O'Keane JC, Abramowitz A, et al. Metastatic adenocarcinoma of an unknown primary site. A comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status. *Am J Clin Pathol* 1993;99: 729–35.
- Liotta LA, Petricoin E. Molecular profiling of human cancer. *Nat Rev Genet* 2000;1:48–56.
- Jaffe ES. Hematopathology: integration of morphologic features and biologic markers for diagnosis. *Mod Pathol* 1999;12:109–15.
- DeYoung BR, Wick MR. Immunohistologic evaluation of metastatic carcinomas of unknown origin: an algorithmic approach. *Semin Diagn Pathol* 2000;17: 184–93.
- Lagendijk JH, Mullink H, van Diest PJ, Meijer GA, Meijer CJ. Immunohistochemical differentiation between primary adenocarcinomas of the ovary and ovarian metastases of colonic and breast origin. Comparison between a statistical and an intuitive approach. *J Clin Pathol* 1999;52:283–90.
- Pecciarini L, Giulia Cangi M, Doglioni C. Identifying the primary sites of metastatic carcinoma: the increasing role of immunohistochemistry. *Curr Diagn Pathol* 2001;7:168–75.
- Hammar SP. Metastatic adenocarcinoma of unknown primary origin. *Hum Pathol* 1998;29:1393–402.
- Kaufmann O, Dietel M. Thyroid transcription factor-1 is the superior immunohistochemical marker for pulmonary adenocarcinomas and large cell carcinomas compared to surfactant proteins A and B. *Histopathology* 2000;36:8–16.
- Tot T. Adenocarcinomas metastatic to the liver: the value of cytokeratins 20 and 7 in the search for unknown primary tumors. *Cancer* 1999;85:171–7.
- Dennis JL, Vass JK, Wit EC, Keith WN, Oien KA. Identification from public data of molecular markers of adenocarcinoma characteristic of the site of origin. *Cancer Res* 2002;62:5999–6005.
- Kononen J, Bubendorf L, Kallioniemi A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844–7.
- Perry A, Parisi JE, Kurtin PJ. Metastatic adenocarcinoma to the brain: an immunohistochemical approach. *Hum Pathol* 1997;28:938–43.
- Werling RW, Yaziji H, Bacchi CE, Gown AM. CDX2, a highly sensitive and specific marker of adenocarcinomas of intestinal origin: an immunohistochemical survey of 476 primary and metastatic carcinomas. *Am J Surg Pathol* 2003;27:303–10.
- Machado JC, Nogueira AM, Carneiro F, Reis CA, Sobrinho-Simoes M. Gastric carcinoma exhibits distinct types of cell differentiation: an immunohistochemical study of trefoil peptides (TFF1 and TFF2) and mucins (MUC1, MUC2, MUC5AC, and MUC6). *J Pathol* 2000;190:437–43.
- Lagendijk JH, Mullink H, Van Diest PJ, Meijer GA, Meijer CJ. Tracing the origin of adenocarcinomas with unknown primary using immunohistochemistry: differential diagnosis between colonic and ovarian carcinomas as primary sites. *Hum Pathol* 1998;29:491–7.
- Kaufmann O, Deidesheimer T, Muehlenberg M, Deicke P, Dietel M. Immunohistochemical differentiation of metastatic breast carcinomas from metastatic adenocarcinomas of other common primary sites. *Histopathology* 1996;29:233–40.
- Loy TS, Quesenberry JT, Sharp SC. Distribution of CA125 in adenocarcinomas. An immunohistochemical study of 481 cases. *Am J Clin Pathol* 1992;98:175–9.
- Frierson HF Jr, Moskaluk CA, Powell SM, et al. Large-scale molecular and tissue microarray analysis of mesothelin expression in common human carcinomas. *Hum Pathol* 2003;34:605–9.
- Ordóñez NG. Application of mesothelin immunostaining in tumor diagnosis. *Am J Surg Pathol* 2003; 27:1418–28.
- Wick MR, Lillemo T, Copland GT, Swanson PE, Manivel JC, Kiang DT. Gross cystic disease fluid protein-15 as a marker for breast cancer: immunohistochemical analysis of 690 human neoplasms and comparison with  $\alpha$ -lactalbumin. *Hum Pathol* 1989;20:281–7.
- Moskaluk CA, Zhang H, Powell SM, Cerilli LA, Hampton GM, Frierson HF Jr. Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Mod Pathol* 2003;16:913–9.
- Brown RW, Campagna LB, Dunn JK, Cagle PT. Immunohistochemical identification of tumor markers in metastatic adenocarcinoma. A diagnostic adjunct in the determination of primary site. *Am J Clin Pathol* 1997;107:12–9.
- Zamecnik J, Kodet R. Value of thyroid transcription factor-1 and surfactant apoprotein A in the differential diagnosis of pulmonary carcinomas: a study of 109 cases. *Virchows Arch* 2002;440:353–61.
- Camp RL, Charette LA, Rimm DL. Validation of tissue microarray technology in breast carcinoma. *Lab Invest* 2000;80:1943–9.
- Doughty SE, Ferrier RK, Hillan KJ, Jackson DG. The effects of ZENECA ZD8731, an angiotensin II antagonist, on renin expression by juxtaglomerular cells in the rat: comparison of protein and mRNA expression as detected by immunohistochemistry and *in situ* hybridization. *Toxicol Pathol* 1995;23:256–61.
- Hsu FD, Nielsen TO, Alkushi A, et al. Tissue microarrays are an effective quality assurance tool for diagnostic immunohistochemistry. *Mod Pathol* 2002;15: 1374–80.
- Swierczynski SL, Maitra A, Abraham SC, et al. Analysis of novel tumor markers in pancreatic and biliary carcinomas using tissue microarrays. *Hum Pathol* 2004;35:357–66.
- Hvidsten TR, Laegreid A, Komorowski J. Learning rule-based models of biological process from gene expression time profiles using gene ontology. *Bioinformatics* 2003;19:1116–23.
- Tenti P, Aguzzi A, Riva C, et al. Ovarian mucinous tumors frequently express markers of gastric, intestinal, and pancreatobiliary epithelial cells. *Cancer* 1992;69:2131–42.
- Ellis IO, Hitchcock A. Tumour marker immunoreactivity in adenocarcinoma. *J Clin Pathol* 1988;41:1064–7.
- Tot T. Cytokeratins 20 and 7 as biomarkers: usefulness in discriminating primary from metastatic adenocarcinoma. *Eur J Cancer* 2002;38:758–63.
- Chu P, Wu E, Weiss LM. Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms: a survey of 435 cases. *Mod Pathol* 2000;13:962–72.
- Park SY, Kim HS, Hong EK, Kim WH. Expression of cytokeratins 7 and 20 in primary carcinomas of the stomach and colorectum and their value in the differential diagnosis of metastatic carcinomas to the ovary. *Hum Pathol* 2002;33:1078–85.
- Goldstein NS, Bassi D. Cytokeratins 7, 17, and 20 reactivity in pancreatic and ampulla of Vater adenocarcinomas. Percentage of positivity and distribution is affected by the cut-point threshold. *Am J Clin Pathol* 2001;115:695–702.
- Mintzer DM, Warhol M, Martin AM, Greene G. Cancer of unknown primary: changing approaches. A multidisciplinary case presentation from the Joan Karnell cancer center of Pennsylvania hospital. *Oncologist* 2004;9:330–8.