A Quantitative Reverse Transcriptase-Polymerase Chain Reaction Assay to Identify Metastatic Carcinoma Tissue of Origin

Dimitri Talantov,* Jonathan Baden,* Tim Jatkoe,*
Kristina Hahn,* Jack Yu,* Yashoda Rajpurohit,*
Yiqiu Jiang,* Chang Choi,* Jeffrey S. Ross,[†]
David Atkins,* Yixin Wang,* and
Abhijit Mazumder*

From Veridex LLC,* Warren, New Jersey; and the Department of
Pathology and Laboratory Medicine,[†] Albany Medical College,
Albany, New York

Identifying the primary site in patients with metastatic carcinoma of unknown primary origin can enable more specific therapeutic regimens and may prolong survival. Twenty-three putative tissue-specific markers for lung, colon, pancreatic, breast, prostate, and ovarian carcinomas were nominated by querying a gene expression profile database and by performing a literature search. Ten of these marker candidates were then selected based on validation by reverse transcriptase-polymerase chain reaction (RT-PCR) on 205 formalin-fixed, paraffin-embedded metastatic carcinoma specimens originating from these six and from other cancer types. Next, we optimized the RNA isolation and quantitative RT-PCR methods for these 10 markers and applied the quantitative RT-PCR assay to a set of 260 metastatic tumors. We then built a gene-based algorithm that predicted the tissue of origin of metastatic carcinomas with an overall leave-one-out cross-validation accuracy of 78%. Lastly, our assay demonstrated an accuracy of 76% when tested on an independent set of 48 metastatic samples, 37 of which were either a known primary or initially presented as carcinoma of unknown primary but were subsequently resolved. (*J Mol Diagn* 2006, 8:320–329; DOI: 10.2353/jmoldx.2006.050136)

Carcinoma of unknown primary (CUP), wherein metastatic disease is present without an identifiable primary tumor site, represents approximately 3 to 5% of all cancers.^{1,2} The prognosis and therapeutic regimen of cancer patients are dependent on the origin of the primary tumor, underscoring the need to identify the site of the primary tumor.³

A variety of methods are currently used to resolve this problem. Immunohistochemical (IHC) markers, in panels of 4 to 14 markers to improve sensitivity and specificity,

have demonstrated identification accuracies of 66 to 88%.^{4–6} More expensive diagnostic workups include imaging methods such as chest X-ray, computed tomographic scans, and positron emission tomographic scans. Despite these sophisticated technologies, the ability to resolve CUP cases is only 20 to 30% ante mortem.^{7,8}

A promising new approach lies in the ability of genome-wide gene expression profiling to identify the origin of tumors.^{9–12} These studies demonstrate the feasibility of tissue of origin identification based on a gene expression profile. Nevertheless, gene marker candidates identified in such studies must be validated on metastatic tissues to confirm that their tissue-specific expression is preserved in metastasis. Furthermore, the gene expression profiling technology must be able to use formalin-fixed, paraffin-embedded tissue (FFPE), because fixed tissue samples are the standard material in current practice. Formalin fixation results in degradation of the RNA^{13,14} so existing microarray protocols do not perform as reliably. Lastly, the profiling technology must be robust, reproducible, and easily accessible. Quantitative RT-PCR (qRT-PCR) has been shown to generate reliable results from FFPE tissue.^{15–18} Therefore, a more practical approach would be to develop a diagnostic assay based on a smaller set of tissue-specific gene markers and a more robust technology.¹⁹ In several previously published studies, investigators used a high-throughput gene discovery approach (microarray or serial analysis of gene expression [SAGE]) followed by validation of fewer tissue-specific gene markers using RT-PCR.^{20–22}

In this manuscript, we describe a qRT-PCR optimized set of 10 markers and demonstrate a high accuracy of prediction of metastatic carcinoma tissue of origin when used with FFPE metastatic carcinoma tissues.

Materials and Methods

Marker Candidate Selection

An internal database containing normal and malignant gene expression profiles from many in-house cancer projects related to breast, colon, lung, ovary, pancreas,

Accepted for publication January 23, 2006.

Address reprint requests to Abhijit Mazumder, Veridex LLC, 33 Technology Dr., Warren, NJ 07059. E-mail: amazumde@vrus.jnj.com.

and prostate samples (682 in total) was used to select an initial panel of markers. For lung, colon, breast, ovary, and prostate, marker candidates were selected by constructing five dichotomous *t*-tests for each tissue type versus all others. To generate pancreas marker candidates, two queries were performed. In the first query, we selected genes that were overexpressed in pancreatic cancer compared with normal tissue, as identified by *t*-test, and the maximum expression of which was at least twofold higher than the maximum expression in other cancers. In a second query, genes that had at least two "present" calls in pancreas tissues and at most two "present" calls in other tissue types were selected followed by selection of genes overexpressed in pancreatic cancer compared with normal. Results of both pancreas queries were combined.

In addition to gene expression profile analysis, a few markers were selected from the literature. Results of all queries were combined to generate a short list of ToO marker candidates for each tissue type. The sensitivity and specificity of each marker was estimated. Markers that demonstrated the best ability to differentiate tissues by their origin were nominated for RT-PCR testing based on redundancy and complementarity.

FFPE Metastatic Carcinoma of Known Origin and CUP Tissues

A total of 386 FFPE metastatic carcinomas (stages III to IV) of known origin and 24 FFPE prostate primary adenocarcinomas were acquired from a variety of commercial vendors, including Proteogenex (Los Angeles, CA), Genomics Collaborative, Inc. (Cambridge, MA), Asterand (Detroit, MI), Ardais (Lexington, MA), and Oncomatrix (San Marcos, CA). An independent set of 48 metastatic carcinomas of known primary and CUP tissues was obtained from Albany Medical College. For each specimen, patient demographic, clinical, and pathology information was collected as well. The histopathological features of each sample were reviewed to confirm diagnosis and to estimate sample preservation and tumor content. For metastatic samples, diagnoses of metastatic carcinoma and tissue of origin were unequivocally established based on the patient's clinical history and histological evaluation of metastatic carcinoma in comparison with the corresponding primaries.

RNA Isolation from FFPE Samples

RNA isolation from paraffin tissue sections was based on the methods and reagents described in the High Pure RNA Paraffin kit manual (Roche Diagnostics, Indianapolis, IN) with the following modifications. Paraffin-embedded tissue samples were sectioned according to size of the embedded metastasis (2 to 5 mm = $9 \times 10 \mu\text{m}$, 6 to 8 mm = $6 \times 10 \mu\text{m}$, and 8 to ≥ 10 mm = $3 \times 10 \mu\text{m}$). Sections were deparaffinized as described by the manual. The tissue pellet was dried in a 55°C oven for 5 to 10 minutes and resuspended in 100 μl of tissue lysis buffer, 16 μl of 10% sodium dodecyl sulfate, and 80 μl of pro-

teinase K. Samples were incubated in a thermomixer set at 400 rpm for 2 hours at 55°C. Subsequent sample processing was performed according to the High Pure RNA Paraffin kit manual. Samples were quantified by OD 260/280 readings obtained by a spectrophotometer and were stored at -80°C until use.

Quantitative Real-Time Polymerase Chain Reaction for Marker Candidate Prescreening

One microgram of total RNA from each sample was reverse-transcribed with random hexamers using Superscript II reverse transcriptase according to the manufacturer's instructions (Invitrogen, Carlsbad, CA). Primers and MGB probes for the tested gene marker candidates and the control gene ACTB were designed using Primer Express software (Applied Biosystems, Foster City, CA). RT-PCR amplification was performed in a 20- μl reaction mix containing 200 ng of template cDNA, 2 $\times$ TaqMan universal PCR master mix (10 ml) (Applied Biosystems), 500 nmol/L forward and reverse primers, and 250 nmol/L probe. Reactions were run on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). In each assay, a "no-template" control and template cDNA were included in duplicate for both the gene of interest and the control gene.

Optimized One-Step Quantitative Real-Time Polymerase Chain Reaction

Gene-specific primers and hydrolysis probes for the optimized one-step qRT-PCR assay are listed in Table 1. Genomic DNA amplification was excluded by designing our assays around exon-intron splicing sites. Hydrolysis probes were labeled at the 5' nucleotide with FAM as the reporter dye and at the 3' nucleotide with BHQ1-TT as the internal quenching dye.

Quantitation of gene-specific RNA was performed in a 384-well plate on the ABI Prism 7900HT sequence detection system (Applied Biosystems). After the PCR reaction was complete, baseline and threshold values were set in the ABI 7900HT Prism software, and calculated cycle threshold (Ct) values were exported to Microsoft Excel.

For comparison of two-step with one-step RT-PCR reactions, first-strand synthesis of two-step reaction was performed using either 100 ng of random hexamers or gene-specific primers per reaction. Both the one-step and two-step reactions were performed on 100 ng of template (RNA/cDNA). After the PCR reaction was completed, baseline and threshold values were set in the ABI 7900HT Prism software, and calculated Ct values were exported to Microsoft Excel.

Algorithm Development

Linear discriminators were constructed using the MASS (Venables and Ripley) library function "lda" in the R language (version 2.1.1). The model used is dependent on

Table 1. Primer and Probe Sequences, Accession Numbers, and Amplicon Lengths

Target	Accession no.	Forward primer	Reverse primer	Probe sequence (5'FAM-3'BHQ1-TT)	Product length
HUMPSPB	NM_00052	CACAGCCCCGACCTTTGATGA	GGTCCCAGAGCCCGTCTCA	AGCTGTCCAGCTGCAAAGGAAAAGCC	75
TTF1	NM_003317	CCAACCCAGACCCGCGC	CGCCCATGCCGCTCATGTTCA	CCCGCCATCTCCCGCTTCATG	78
DSG3	NM_001944	GCAGAGAAGGAGAAGATAAC- TCAA	ACTCCAGAGATTCCGGTAGGTGA	ATTGCCAAGATTACTTCAGATTACCA	97
CDH17	NM_004063	TCCCTCGGCAGTGGAAGCTTA	TCCTCAAACCTCTGTGTGCCTG- GTA	CCAAAATCAATGGTACTCATGCCCCG- ACTG	94
MG	NM_002411	AGTTGCTGATGGTCCTCATGC	CACCTGTGGATTGATTGTCTT- GGA	CCCTCTCCCAGCACTGCTACGCA	107
PDEF	NM_012391	CGCCACCTGGACATCTGGA	CACTGGTCGAGGCACAGTAGT- GA	GTCAGCGGCCCTGGATGAAAGAGCGG	86
WT1	NM_024426	GCGGAGCCCAATACAGAATA- CAC	CGGGGCTACTCCAGGCACA	TCAGAGGCATTTCAGGATGTGCGACG	80
PSCA	NM_005672	CTGTTGATGGCAGGCTTGGC	TTGCTCACTGGGCTTTGCA	GCAGCCAGGCACTGCCCTGCT	74
F5	NM_000130	TGAAGAAATATCCTGGGATT- ATTCA	TATGTGTATCTTCTGGAATA- TCATCA	ACAAAGGGAAACAGATATTGAAGACTC	87
KLK3	NM_001648	CCCCCAGTGGGTCCTCACA	AGGATGAAACAAGCTGTGCCGA	CAGGAACAAAAGCGTGATCTTGCTGG	82
β -Actin	NM_001101	GCCCTGAGGCACTCTTCCA	CGGATGTCCACGTCACACTTCA	CTTCCTTCTGGGCATGGAGTCTCTG	100
PBGD	NM_000190	CCACACACAGCCTACTTTCC- AA	TACCCACGCGAATCACTCTCA	AACGGCAATGCGGCTGCAACGCGGAA	103

KLK3, kallikrein 3; MG, mammaglobin; PDEF, prostate-derived Ets transcription factor.

the tissue from which the metastasis was extracted from and the gender of the patient. When a lung, colon, or ovarian site of metastasis is encountered, the class prior is set to zero for the class that is equivalent to the site of metastasis. Furthermore, the prior odds are set to zero for the breast and ovary class in male patients, whereas in female patients, the prostate class prior is set to zero. All other prior odds used in the models are assumed to be equal. Furthermore, classification for each sample is based on the highest posterior probability determined by the model for each class. To estimate the performance of the model, we performed leave-one-out cross-validation. In addition to this, the data sets were randomly split in halves, while preserving the proportional relationship between each class, into training and testing sets. This random splitting was repeated three times.

Results

Our experimental work consisted of two major parts. The first part included the reduction from 23 tissue-specific marker candidates to 10, based on the RT-PCR interrogation on FFPE metastatic carcinoma tissues (Figure 1A). The second part included qRT-PCR assay optimization followed by assay implementation on a training set of FFPE metastatic carcinomas, building of a prediction algorithm, its cross-validation, and validation on an independent sample set (Figure 1B).

Sample Characteristics

A total of 386 metastatic carcinomas of known origin (stages III to IV) and 24 prostate primary adenocarcinoma samples were used in the study. The metastatic carcinomas originated from lung, pancreas, colorectal, ovary, prostate, and other cancers. The "other" sample category consisted of metastasis derived from tissues

other than lung, pancreas, colon, breast, ovary, and prostate. Patient characteristics are summarized in Table 2. Samples were separated into two sets: the initial set (205 specimens) used to confirm tissue-specific differential expression of putative marker candidates, and the training set (260 specimens) that incorporated the optimized one-step qRT-PCR procedure and a cross-validated error estimation of a prediction algorithm. The first set of

Figure 1. Experimental workflow: marker candidate selection (A) and assay optimization and prediction algorithm building and testing (B).

Table 2. Summary of Patient Information

	Metastatic carcinoma of known origin	Albany Medical College sample set
Total number	401	48
Average age	57.8 ± 11	62.13 ± 11.7
Gender		
Female	241	20
Male	160	28
Tissue of origin		
Lung	65	9
Pancreas	63	2
Colorectal	61	4
Breast	63	5
Ovarian	82	2
Prostate	27	2
Kidney	8	8
Stomach	7	0
Other*	25	5
Carcinoma of Unknown Primary		11
Histopathological diagnosis		
Adenocarcinoma, moderately/well differentiated	306	27
Adenocarcinoma, poorly differentiated	49	4
Squamous cell carcinoma	16	5
Poorly differentiated carcinoma	16	10
Small cell carcinoma	3	
Melanoma	5	
Lymphoma	3	
Hepatocellular carcinoma	2	
Mesothelioma	1	
Other†	14	2
Metastatic site		
Lymph nodes	73	1
Brain	17	14
Lung	20	7
Liver	75	11
Pelvic region (ovary, bladder, fallopian tubes)	53	2
Abdomen (omentum, mesentery, colon, peritoneum)	91	5
Other (skin, thyroid, chest wall, umbilicus)	44	8
Unknown	2	
Primary (prostate)	26	

*Esophagus, bladder, pleura, liver, gallbladder, bile ducts, larynx, pharynx, non-Hodgkin's lymphoma.

†Small cell mesothelioma, hepatocellular, melanoma, lymphoma.

205 samples included 25 lung, 41 pancreatic, 31 colorectal, 33 breast, 33 ovarian, one prostate, 23 other cancer metastases, and 18 prostate primary cancers. The second set consisted of 260 samples including 56 lung, 43 pancreatic, 30 colorectal, 30 breast, 49 ovarian, 32 other cancer metastases, and 20 primary prostate cancers. Sixty-four specimens, including 16 lung, 21 pancreatic, 15 other metastatic, and 12 prostate primary, carcinomas were from the same tumors in both sets.

The independent sample set obtained from Albany Medical College was composed of 33 CUP specimens with a primary diagnosed in 22 of them and 15 metastatic carcinomas of known origin. For resolved CUPs, a diagnosis was rendered based on results of testing with a panel of IHC markers or clinical diagnostic tests. Patient demographic, clinical, and pathology characteristics are presented in Table 2.

Marker Candidate Selection

Analysis of gene expression profiles of five primary tissues types (lung, colon, breast, ovary, and prostate) resulted in nomination of 13 tissue-specific marker can-

didates for qRT-PCR testing. Not surprisingly, all top candidates have been identified in previous studies and well described in the literature.²³⁻³¹

A special approach was used to identify pancreas-specific markers. Pancreatic ductal adenocarcinoma develops from ductal epithelial cells that compose only a small percentage of all pancreatic cells (with acinar and islet cells composing the majority) in the normal pancreas. Furthermore, pancreatic adenocarcinoma tissues contain a significant amount of adjacent normal tissue.^{32,33} Because of this, our candidate pancreas markers were enriched for genes elevated in pancreas adenocarcinoma relative to normal pancreas cells. As a result of this approach, we identified a novel pancreas cancer-specific marker, coagulation factor V (F5).

In addition, seven marker candidates that have been discussed in the literature were selected, including pancreatic cancer markers (prostate stem cell antigen [PSCA], serine proteinase inhibitor, clade A member 1, cytokeratin 7, matrix metalloprotease 11, and mucin 4^{8,23,32,34-36}), lung squamous cell carcinoma marker

Table 3. Gene Marker Candidates Tested and Testing Summary

Tissue type	Name	Identification method		Marker selection filters			
		Gene expression database query	Literature	Low expression in corresponding metastatic tissue	Marker redundancy	Tissue cross-reactivity	Marker adequate?
Lung	SP-B	x	x				x
	TTF1	x	x				x
	DSG3		x				x
Pancreas	PSCA		x				x
	F5	x					x
	ITGB6	x			x		
	TGM2	x		x			
	MUC4		x	x		x	
	SERPINA 1		x	x			
	KRT7		x	x		x	
	MMP11		x			x	
	HPT1	x	x				x
Colon	FABP1	x	x				
	CDX1	x	x		x		
	GUCY2C	x	x		x		
	PSA	x	x				x
Prostate	hKLK2	x	x		x		
	MGB1	x	x				x
Breast	PIP	x	x			x	
	PDEF		x				x
	HE4	x	x			x	
Ovarian	PAX8	x	x			x	
	WT1	x	x				x

KRT7, cytokeratin 7; MMP11, matrix metalloprotease 11; MUC4, mucin 4; PDEF, prostate-derived Ets transcription factor; SERPINA 1, serine proteinase inhibitor, clade A member 1.

desmoglein 3 (DSG3),³⁷ and the breast marker prostate-derived Ets transcription factor.²⁴

A total of 23 tissue-specific marker candidates were selected for further RT-PCR interrogation on 205 metastatic carcinoma FFPE tissues by qRT-PCR. Of 23 tested markers, 13 were rejected based on their cross-reactivity, low expression level in the corresponding metastatic tissues, or redundancy. Table 3 provides the gene symbols of the tissue-specific markers selected for RT-PCR validation and also summarizes the selection criteria applied to markers. Ten markers were selected for the final version of the assay. The lung markers were human surfactant pulmonary-associated protein B (HUMSPB), thyroid transcription factor 1, and DSG3. The pancreas markers were PSCA and coagulation factor V (F5), and the prostate marker was kallikrein 3. The colorectal marker was cadherin 17. The breast markers were mammaglobin and prostate-derived Ets transcription factor. The ovarian marker was Wilms tumor 1. Mean normalized relative expression values of selected markers in different metastatic tissues are presented in Figure 2.

Optimization of Sample Preparation and qRT-PCR Using FFPE Tissues

We next wanted to optimize the RNA isolation and qRT-PCR methods using fixed tissues before we examined the performance of our marker panel. We first analyzed the effect of reducing the proteinase K incubation time from 16 to 3 hours. We found no effect on yield (data not shown). We did, however, notice that some samples showed longer fragments of RNA when the shorter pro-

teinase K step was used (Figure 3, A and B). For example, when RNA was isolated from a 1-year-old block (C22), we observed no difference in the electropherograms. However, when RNA was isolated from a 5-year-old block (C23), we observed a larger fraction of higher molecular weight RNAs, as assessed by the hump in the shoulder, when the shorter proteinase K digest was used. We found this trend to generally hold when other samples were processed (data not shown), regardless of the organ of origin for the formalin-fixed, paraffin-embedded metastasis. We conclude that shortening the proteinase K digestion time does not sacrifice RNA yields and may aid in isolating longer, less degraded RNA.

We next compared three different methods of reverse transcription: reverse transcription with random hexamers followed by qPCR (two step), reverse transcription with a gene-specific primer followed by qPCR (two step), and a one-step qRT-PCR using gene-specific primers. We isolated RNA from 11 metastases and compared Ct values across the three methods for β -actin, HUMSPB (Figure 3, C and D), and thyroid transcription factor (data not shown). We found statistically significant differences ($P < 0.001$) for all comparisons. For both genes, the reverse transcription with random hexamers followed by qPCR (two-step reaction) gave the highest Ct values, whereas the reverse transcription with a gene-specific primer followed by qPCR (two-step reaction) gave slightly (but statistically significant) lower Ct values than the corresponding one-step reaction. We conclude that optimization of the RT-PCR reaction conditions can generate lower Ct values, which may help in analyzing older paraffin blocks,¹⁸ and that a one-step RT-PCR reaction with

Figure 2. Expression of 10 selected tissue-specific gene marker candidates in FFPE metastatic carcinomas and prostate primary adenocarcinomas. For each plot, the y axis represents the normalized marker expression value.

gene-specific primers can generate Ct values comparable with those generated in the corresponding two-step reaction.

Diagnostic Performance of an Optimized qRT-PCR Assay

We next performed 12 qRT-PCR reactions (10 markers and two housekeeping genes) on a set of 260 formalin-fixed, paraffin-embedded metastases. Twenty-one samples gave high Ct values for the housekeeping genes, so only 239 were used in a heat map analysis. Analysis of the normalized Ct values in a heat map revealed the high specificity of the breast and prostate markers; moderate specificity of the colon, lung, and ovary; and somewhat lower specificity of the pancreas markers (Figure 4).

Using expression values, normalized to the average expression of two housekeeping genes, we developed an algorithm to predict the tissue of origin for the metastasis. We combined the normalized qRT-PCR data with the algorithm and determined the accuracy of the qRT-PCR assay by performing a leave-one-out cross-validation test (LOOCV). For the six tissue types included in the assay, we separately estimated both the number of false-positive calls, when a sample was wrongly predicted as another tumor type included in the assay (pancreas as

colon, for example), and the number of times a sample was not predicted as those included in the assay tissue types (other). Results of the LOOCV are presented in Table 4. The tissue of origin was predicted correctly for 204 of 260 tested samples with an overall accuracy of 78%. A significant proportion of the false-positive calls were due to the cross-reactivity of the marker in histologically similar tissues. For example, three squamous cell metastatic carcinomas originating from pharynx, larynx, and esophagus were wrongly predicted as lung due to DSG3 expression in these tissues. Positive expression of CDH17 in other than colon gastrointestinal carcinomas, including stomach and pancreas, caused false classification of four of six stomach and 3 of 43 pancreatic cancer metastases.

In addition to a LOOCV test, the data were randomly split into three separate pairs of training and testing sets. Each split contained approximately 50% of the samples from each class. In three separate pairs of training and testing sets, assay overall classification accuracies were 77, 71, and 75%, confirming assay performance stability.

Lastly, we tested an independent set of 48 FFPE metastatic carcinomas that included metastatic carcinoma of known primary, CUP specimens with a tissue of origin diagnosis rendered by pathological evaluation including IHC and clinical tests, and CUP specimens that remained

Figure 3. Assay optimization. **A** and **B:** Electropherograms obtained from an Agilent Bioanalyzer. RNA was isolated from formalin-fixed, paraffin-embedded tissue using a 3-hour (**A**) or 16-hour (**B**) proteinase K digestion. Sample C22 (red) was a 1-year-old block, whereas sample C23 (blue) was a 5-year-old block. A size ladder is shown in green. **C** and **D:** Comparison of Ct values obtained from three different qRT-PCR methods: random hexamer priming in the reverse transcription followed by qPCR with the resulting cDNA (RH 2 step), gene-specific (reverse primer) priming in the reverse transcription followed by qPCR with the resulting cDNA (GSP 2 step), or gene-specific priming and qRT-PCR in a one-step reaction (GSP 1 step). RNA from 11 samples was divided into the three methods, and RNA levels for two genes were measured: β -actin (**C**), HUMSPB (**D**). The median Ct value obtained with each method is indicated by the **solid line**.

CUP after diagnostic workup. The tissue of origin prediction accuracy was estimated separately for each category of samples. Table 5 summarizes the assay results. The tissue of origin prediction was, with only a few exceptions, consistent with the known primary or tissue of origin diagnosis assessed by clinical/pathological evaluation including IHC.

The assay also made putative tissue of origin diagnoses for 8 of 11 samples that remained CUP after standard diagnostic tests. One of the CUP cases was especially interesting. A male patient with a history of prostate cancer was diagnosed with metastatic carcinoma in lung and pleura. Serum prostate-specific-antigen (PSA) tests and IHC with PSA antibodies on metastatic tissue were negative, so the pathologist's diagnosis was CUP with an inclination toward gastrointestinal tumors. Our assay strongly (posterior probability, 0.99) predicted the tissue of origin as colon.

Discussion

In this research study, we have used microarray-based expression profiling on primary tumors to identify candidate markers for use with metastases, a paradigm that is

consistent with several recent findings. For example, Weigelt et al²⁴ have shown that gene expression profiles of primary breast tumors are maintained in distant metastases.³⁸ Backus et al²⁴ identified putative markers for detecting breast cancer metastasis using a genome-wide gene expression analysis of primary tissues.²⁴

During the development of the assay, we limited our selection to six cancer types, including lung, pancreas, and colon, which are among the most prevalent in CUP,^{1,7} and breast, ovarian, and prostate, for which treatment could be potentially most beneficial for patients.¹ We note, however, that additional tissue types and markers can be added to our panel as long as the overall accuracy of the assay is not compromised and the logistics of the RT-PCR reactions are not encumbered. It is important to mention that development of the algorithm required a training set for which the diagnosis was known and accurate. That is, only metastases having a known organ of origin could be used. Had we used metastases of unknown origin for which no primary could be found (a more clinically relevant set), we would not have been able to employ our computational approach. This study, therefore, represents the first step toward application of the assay on more clinically relevant samples.

Figure 4. Heat map showing the relative expression levels of the 10 marker panel across 239 samples. Red indicates higher expression. For each sample, ΔCt was calculated by taking the mean Ct of each CUP marker and subtracting the mean Ct of an average of our housekeeping markers ($\Delta Ct = Ct[CUP \text{ marker}] - Ct[\text{ave. HK marker}]$). The minimal ΔCt for each tissue of origin marker set (lung, breast, prostate, colon, ovarian, and pancreas) was determined for each sample. The tissue of origin with the overall minimal ΔCt was scored one, and all other tissues of origin were scored zero. Data were sorted according to pathological diagnosis. Partek Pro was populated with the modified feasibility data, and an heat map was generated.

Our microarray-based studies with primary tissue confirmed the specificity and sensitivity of known markers. As a result, the majority of tissue-specific markers we have used have been documented in the literature as having high specificity for the tissues studied here. We note that a recent study using IHC found that PSCA is overexpressed in prostate cancer metastases.³⁹ Dennis et al²⁰ demonstrated that PSCA could be used as a tumor of origin marker for pancreas and prostate. We also observed strong RNA expression of PSCA in some prostate tissues, but because we have included PSA in our assay, we could segregate prostate and pancreatic cancers. A novel finding of this study was the use of F5 as a complementary (to PSCA) marker for pancreatic tissue of origin. In both the microarray data set with primary tissue

Table 5. Assay Performance on an Independent Set of 48 Samples, Including Metastatic Carcinoma of Known Origin and CUPs

	Tested	Correct	Accuracy
Known mets*	15	11	73.3
Resolved CUP	22	17	77.3
Unresolved CUP	11		

* mets, metastatic carcinomas.

and the qRT-PCR data set with formalin-fixed, paraffin-embedded metastases, F5 was found to complement PSCA.

Our study is the first to combine microarray-based expression profiling with a small panel of qRT-PCR assays. Our microarray studies with primary tissue identified some, but not all, of the same tissue of origin markers as those identified previously by SAGE studies. This finding is not surprising given studies that have demonstrated that a modest agreement between SAGE- and DNA microarray-based profiling data exists and that the correlation improves for genes with higher expression levels.^{40,41} It is noteworthy that Dennis et al⁶ found an accuracy of 83% when a panel of IHC markers, two of which were identical to our markers, was used. However, there are two major distinctions between that study and our study, beyond the fact that different technologies (IHC vs. qRT-PCR) were used. First, our method employs quantitative cutoffs and normalized Ct values rather than a positive or negative output as used by Dennis et al, a fact that may ultimately help in specificity of the assay. Second, we have tested our assay on three sets of metastatic samples comprising 205, 260, and 48 samples, of which only 38 samples in total were primary tissue. However, Dennis et al⁶ tested only 30 metastatic samples. A potential benefit of qRT-PCR over IHC is that the former removes any subjective bias (eg, regarding the intensity of the staining) and does not depend on having a very experienced pathologist capable of interpreting the staining. A side-by-side comparison using the same sample set and using both IHC and our RT-PCR markers will be required to determine which method is superior.

We further improved the qRT-PCR protocol through the use of gene-specific primers in a one-step reaction. To our knowledge, this is the first demonstration of the use of gene-specific primers in a one-step qRT-PCR reaction

Table 4. Assay Cross-Validation by LOOCV Test on 260 FFPE Metastatic Carcinoma and Primary Tissue Samples

Prediction	Tissue of origin							Total
	Breast	Colon	Lung	Ovary	Pancreas	Prostate	Other	
Breast	22	0	2	1	1	0	0	
Colon	1	27	3	2	4	0	4	
Lung	1	2	45	2	3	0	5	
Other	1	1	3	1	4	0	16	
Ovary	5	0	0	43	0	0	1	
Pancreas	0	0	3	0	31	0	6	
Prostate	0	0	0	0	0	20	0	
Total	30	30	56	49	43	20	32	260
No. correct	22	27	45	43	31	20	16	204
Accuracy	73.3	90.0	80.4	87.8	72.1	100.0	50.0	78.5

with formalin-fixed, paraffin-embedded tissue. Other investigators have either performed a two-step qRT-PCR (cDNA synthesis in one reaction followed by qPCR) or used random hexamers or truncated gene-specific primers.^{15–18,42}

In summary, the 78% overall accuracy of our assay compares favorably with other studies.^{4–6,10–12} Although overall our markers demonstrated high specificity, we observed a number of false-positive calls when metastatic carcinoma tissue of origin was wrongly predicted because of marker cross-reactivity in other than target tissues. The challenge was to identify tissue of origin of squamous cell metastatic carcinoma, because the selected marker (DSG3) is squamous cell differentiation specific rather than tissue of origin specific. Its expression in squamous cell carcinomas of tissues other than lung tissue of origin was reported previously.⁴³ Squamous cell carcinoma tissue of origin prediction was challenging in other tissue of origin classification studies even with larger numbers of tissue-specific markers.²² Similar difficulties were observed with the colon marker CDH17, which, although overall specific for colon metastases, was also positive in metastases originating from gastric tumors and occasionally in pancreatic cancer metastases. Expression of CDH17 in gastric cancer and in pancreatic ductal adenocarcinomas was reported previously as well.⁴⁴ Nevertheless, in our opinion, prediction of metastatic carcinoma as gastrointestinal in origin will help to narrow the primary search.

The fact that the assay performance with the 22 resolved CUP samples was similar to that seen with the metastases having a known primary is significant because it suggests that the assay will work on samples that are more likely to be poorly differentiated and more representative of the true clinical dilemma. Further validation of the assay with larger numbers of true and resolved CUP samples will be needed to assess not only the true clinical value of such molecular techniques but also the ability of new information to impact survival and quality of life.

Acknowledgments

We thank Dr. Emma Du (Scripps Clinic, La Jolla, CA) and Dr. Scott Granter (Harvard Medical School, Boston, MA) for help with pathology evaluation of the samples used in the study; Haiying Wang (Veridex) for providing data from the proteinase K incubation study; Tom Briggs (Veridex) for providing preliminary data on reverse transcription methods; Lance Boling (Kelly Scientific Services) for technical assistance with RNA isolation; and Dr. Yi Zhang (Veridex) for support in development of predictive algorithm.

References

1. Ghosh L, Dahut W, Kakar S, Posadas EM, Torres CG, Cancel-Santiago R, Ghosh BC: Management of patients with metastatic cancer of unknown primary. *Curr Probl Surg* 2005, 42:12–66
2. Mintzer DM, Warhol M, Martin A-M, Greene G: Cancer of unknown

- primary: changing approaches, a multidisciplinary case presentation from the Joan Karnell Cancer Center of Pennsylvania Hospital. *Oncologist* 2004, 9:330–338
3. Greco FA, Rodriguez GI, Shaffer DW, Hermann R, Litchy S, Yardley DA, Burris HA III, Morrissey LH, Erland JB, Hainsworth JD: Carcinoma of unknown primary site: sequential treatment with paclitaxel/carboplatin/etoposide and gemcitabine/irinotecan. A Minnie Pearl cancer research network phase II trial. *Oncologist* 2004, 9:644–652
4. 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, 107:12–19
5. DeYoung BR, Wick MR: Immunohistologic evaluation of metastatic carcinomas of unknown origin: an algorithmic approach. *Semin Diagn Pathol* 2000, 17:184–193
6. Dennis JL, Hvidsten TR, Wit EC, Komorowski J, Bell AK, Downie I, Mooney J, Verbeke C, Bellamy C, Keith WN, Oien KA: Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. *Clin Cancer Res* 2005, 11:3766–3772
7. Pavlidis N, Briasoulis E, Hainsworth J, Greco FA: Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Can* 2003, 39:1990–2005
8. Varadhachary GR, Abbruzzese JL, Lenzi R: Diagnostic strategies for unknown primary cancer. *Cancer* 2004, 100:1776–1785
9. Dennis JL, Oien KA: Hunting the primary: novel strategies for defining the origin of tumors. *J Pathol* 2005, 205:236–247
10. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimtrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF Jr, Hampton GM: Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 2001, 61:7388–7393
11. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* 2001, 98:15149–15154
12. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J, Yeatman TJ: Multi-platform, multi-site, microarray-based human tumor classification. *Am J Pathol* 2004, 164:9–16
13. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P: Unlocking the archive-gene expression in paraffin-embedded tissue. *J Pathol* 2001, 195:66–71
14. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K: Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Res* 1999, 27:4436–4443
15. Abrahamsen HN, Steiniche T, Nexø E, Hamilton-Dutoit SJ, Sorensen BS: Towards quantitative mRNA analysis in paraffin-embedded tissues using real-time reverse transcriptase-polymerase chain reaction. *J Mol Diagn* 2003, 5:34–41
16. Specht K, Richter T, Muller U, Walch A, Werner M, Hofler H: Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. *Am J Pathol* 2001, 158:419–429
17. Godfrey TE, Kim S-H, Chavira M, Ruff DW, Warren RS, Gray JW, Jensen RH: Quantitative mRNA expression analysis from formalin-fixed, paraffin-embedded tissues using 5' nuclease quantitative reverse transcription-polymerase chain reaction. *J Mol Diagn* 2000, 2:84–91
18. Cronin M, Pho M, Dutta D, Stephens JC, Shak S, Kiefer MC, Esteban JM, Baker JB: Measurement of gene expression in archival paraffin-embedded tissue. *Am J Pathol* 2004, 164:35–42
19. Ramaswamy S: Translating cancer genomics into clinical oncology. *N Engl J Med* 2004, 350:1814–1816
20. 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
21. Buckhaults P, Zhang Z, Chen Y-C, Wang T-L, St. Croix B, Saha S, Bardelli A, Morin PJ, Polyak K, Hruban RH, Velculescu VE, Shih I-M: Identifying tumor origin using a gene expression-based classification map. *Cancer Res* 2003, 63:4144–4149
22. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, Biankin AV, Sutherland RL, Henshall SM, Fong K, Pollack JR, Bowtell DDL, Holloway AJ: An expression-based site of origin diagnostic method designed for clin-

- ical application to cancer of unknown origin. *Cancer Res* 2005, 65:4031–4040
23. Argani P, Rosty C, Reiter RE, Wilentz RE, Murugesan SR, Leach SD, Ryu B, Skinner HG, Goggins M, Jaffee EM, Yeo CJ, Cameron JL, Kern SE, Hruban RH: Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma. *Cancer Res* 2001, 61:4320–4324
24. Backus J, Laughlin T, Wang Y, Belly R, White R, Baden J, Min CJ, Mannie A, Tafta L, Atkins D, Verbanac KM: Identification and characterization of optimal gene expression markers for detection of breast cancer metastasis. *J Mol Diagn* 2005, 7:327–336
25. Cunha AC, Weigle B, Kiessling A, Bachmann M, Rieber EP: Tissue-specificity of prostate specific antigens: comparative analysis of transcript levels in prostate and non-prostatic tissues. *Cancer Lett* 2006, 236:229–238
26. Borgono CA, Michael IP, Diamandis EP: Human tissue kallikreins: physiologic roles and applications in cancer. *Mol Cancer Res* 2004, 2:257–280
27. McCarthy DM, Maitra A, Argani P, Rader AE, Faigel DO, Van Heek NT, Hruban RH, Wilentz RE: Novel markers of pancreatic adenocarcinoma in fine-needle aspiration: mesothelin and prostate stem cell antigen labeling increases accuracy in cytologically borderline cases. *Appl Immunohistochem Mol Morphol* 2003, 11:238–243
28. Hwang H, Quenneville L, Yaziji H, Gown AM: Wilms tumor gene product: sensitive and contextually specific marker of serous carcinomas of ovarian surface epithelial origin. *Appl Immunohistochem Mol Morphol* 2004, 12:122–126
29. Fleming TP, Watson MA: Mammaglobin, a breast-specific gene, and its utility as a marker for breast cancer. *Ann NY Acad Sci* 2000, 923:78–89
30. Nakamura N, Miyagi E, Murata S, Kawaoi A, Katoh R: Expression of thyroid transcription factor-1 in normal and neoplastic lung tissues. *Mod Pathol* 2002, 15:1058–1067
31. Khoor A, Whitsett JA, Stahlman MT, Halter SA: Expression of surfactant protein B precursor and surfactant protein B mRNA in adenocarcinoma of the lung. *Mod Pathol* 1997, 10:62–67
32. Prasad NB, Biankin AV, Fukushima N, Maitra A, Dhara S, Elakhloun AG, Hruban RH, Goggins M, Leach SD: Gene expression profiles in pancreatic intraepithelial neoplasia reflect the effects of Hedgehog signaling on pancreatic ductal epithelial cells. *Cancer Res* 2005, 65:1619–1626
33. Ishikawa M, Yoshida K, Yamashita Y, Ota J, Takada S, Kisanuki H, Koinuma K, Choi YL, Kaneda R, Iwao T, Tamada K, Sugano K, Mano H: Experimental trial for diagnosis of pancreatic ductal carcinoma based on gene expression profiles of pancreatic ductal cells. *Cancer Sci* 2005, 96:387–393
34. Fukushima N, Sato N, Prasad N, Leach SD, Hruban RH, Goggins M: Characterization of gene expression in mucinous cystic neoplasms of the pancreas using oligonucleotide microarrays. *Oncogene* 2004, 23:9042–9051
35. Jones LE, Humphreys MJ, Campbell F, Neoptolemos JP, Boyd MT: Comprehensive analysis of matrix metalloproteinase and tissue inhibitor expression in pancreatic cancer: increased expression of matrix metalloproteinase-7 predicts poor survival. *Clin Cancer Res* 2004, 10:2832–2845
36. Moniaux N, Andrianifahanana M, Brand RE, Batra SK: Multiple roles of mucins in pancreatic cancer, a lethal and challenging malignancy. *Br J Cancer* 2004, 91:1633–1638
37. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: Classification of human lung adenocarcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001, 98:13790–13795
38. Weigelt B, Glas AM, Wessels LFA, Witteveen AT, Peterse JL, van't Veer LJ: Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc Natl Acad Sci USA* 2003, 100:15901–15905
39. Lam JS, Yamashiro J, Shintaku IP, Vessella RL, Jenkins RB, Horvath S, Said JW, Reiter RE: Prostate stem cell antigen is overexpressed in prostate cancer metastases. *Clin Cancer Res* 2005, 11:2591–2596
40. van Ruisen F, Ruijter JM, Schaaf GJ, Asgharnegad L, Zwijnenburg DA, Kool M, Baas F: Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix Gene Chips. *BMC Genomics* 2005, 6:91
41. Kim HL: Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Exp Mol Med* 2003, 35:460–466
42. Mikhitarian K, Reott S, Hoover L, Allen A, Cole DJ, Gillanders WE, Mitas M: Enhanced detection of RNA from paraffin-embedded tissue using a panel of truncated gene-specific primers for reverse transcription. *BioTechniques* 2004, 36:1–4
43. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, Butterfoss D, Xiang D, Zanation A, Yin X, Shockley WW, Weissner MC, Dressler LG, Shores CG, Yarbrough WG, Perou CM: Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* 2004, 5:489–500
44. Takamura M, Sakamoto M, Ino Y, Shimamura T, Ichida T, Asakura H, Hirohashi S: Expression of liver-intestine cadherin and its possible interaction with galectin-3 in ductal adenocarcinoma of the pancreas. *Cancer Sci* 2003, 94:425–430