Vol 27 2009

Reprinted from

JOURNAL OF CLINICAL ONCOLOGY

Multicenter Validation of a 1,550-Gene Expression Profile for Identification of Tumor Tissue of Origin

Federico A. Monzon, Maureen Lyons-Weiler, Ljubomir J. Buturovic, C. Ted Rigl, W. David Henner, Christin Sciulli, Catherine I. Dumur, Fabiola Medeiros, and Glenda G. Anderson

www.jco.org

Official Journal of the American Society of Clinical Oncology

ASCO®

Published Ahead of Print on March 30, 2009 as 10.1200/JCO.2008.17.9762
The latest version is at http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2008.17.9762

JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

Multicenter Validation of a 1,550-Gene Expression Profile for Identification of Tumor Tissue of Origin

Federico A. Monzon, Maureen Lyons-Weiler, Ljubomir J. Buturovic, C. Ted Rigl, W. David Henner, Christin Sciulli, Catherine I. Dumur, Fabiola Medeiros, and Glenda G. Anderson

From The Methodist Hospital and The Methodist Hospital Research Institute, Houston, TX; the Clinical Genomics Facility and Department of Pathology, University of Pittsburgh, Pittsburgh, PA; Pathwork Diagnostics, Sunnyvale, CA; Department of Pathology, VA Commonwealth University, Richmond, VA; and the Department of Laboratory Medicine and Pathology, Division of Laboratory Genetics, Mayo Clinic, Rochester, MN.

Submitted May 5, 2008; accepted December 2, 2008; published online ahead of print at www.jco.org on March 30, 2009.

Supported by a sponsored research agreement from Pathwork Diagnostics, Sunnyvale, CA (F.A.M., with a subcontract to C.I.D.).

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Presented in part in poster format at the Annual Meeting of the Association for Molecular Pathology, Los Angeles, CA, November 7-10, 2007; and the Annual Meeting of the United States and Canadian Academy of Pathology, San Diego, CA, March 24-30, 2007.

Corresponding author: Federico A. Monzon, MD, Medical Director of Molecular Diagnostics, The Methodist Hospital, 6565 Fannin St, MS 205, Houston, TX 77030; e-mail: famonzon@tmhs.org.

0732-183X/09/2799-1/$20.00

DOI: 10.1200/JCO.2008.17.9762

ABSTRACT

Purpose
Malignancies found in unexpected locations or with poorly differentiated morphologies can pose a significant challenge for tissue of origin determination. Current histologic and imaging techniques fail to yield definitive identification of the tissue of origin in a significant number of cases. The aim of this study was to validate a predefined 1,550-gene expression profile for this purpose.

Methods
Four institutions processed 547 frozen specimens representing 15 tissues of origin using oligonucleotide microarrays. Half of the specimens were metastatic tumors, with the remainder being poorly differentiated and undifferentiated primary cancers chosen to resemble those that present as a clinical challenge.

Results
In this blinded multicenter validation study the 1,550-gene expression profile was highly informative in tissue determination. The study found overall sensitivity (positive percent agreement with reference diagnosis) of 87.8% (95% CI, 84.7% to 90.4%) and overall specificity (negative percent agreement with reference diagnosis) of 99.4% (95% CI, 98.3% to 99.9%). Performance within the subgroup of metastatic tumors (n = 258) was found to be slightly lower than that of the poorly differentiated and undifferentiated primary tumor subgroup, 84.5% and 90.7%, respectively (P = .04). Differences between individual laboratories were not statistically significant.

Conclusion
This study represents the first adequately sized, multicenter validation of a gene-expression profile for tissue of origin determination restricted to poorly differentiated and undifferentiated primary cancers and metastatic tumors. These results indicate that this profile should be a valuable addition or alternative to currently available diagnostic methods for the evaluation of uncertain primary cancers.

J Clin Oncol 27. © 2009 by American Society of Clinical Oncology

INTRODUCTION

Evidence-based management indicates that a thorough investigation of uncertain primary cancers should be performed to assist in therapeutic decisions.[1,2] This is typically carried out with immunohistochemistry (IHC) panels on the tumor specimen, and advanced whole body or site-directed imaging tests.[1,3-5] This work-up is associated with considerable resources, time, and expense[1,6,7]; however, the primary site remains unidentified in up to 30% of patients who present with an uncertain primary cancer.[1,8,9] Thus new approaches are needed to reduce diagnostic uncertainty in these patients. The use of gene expression–based signatures for classifying tumor tissue of origin (TOO) has been reported,[10-14] and these studies indicated that metastatic and poorly-differentiated specimens pose a significant challenge to gene expression–based classifiers.

To our knowledge, we present the first blinded, multicenter validation study conducted on a gene expression–based test to identify the tissue of origin, the Pathwork Tissue of Origin Test (Pathwork Diagnostics, Sunnyvale, CA). An interlaboratory reproducibility study of the 1,550-gene expression profile has been described previously.[15] Two important aspects of this study are: it is the first clinical validation of significant size (> 500 specimens) to be performed on a test for TOO; and it is the only reported study conducted entirely with metastatic

tumors and poorly differentiated or undifferentiated primary tumors chosen to resemble the expected population of difficult to diagnose cancers.

METHODS

Patients and Tumor Specimens

Tumor specimens or tumor-derived microarray gene expression files from 622 patients were screened for inclusion. Three hundred fifty-one frozen tissue specimens were obtained from the Health Sciences Tissue Bank at the University of Pittsburgh (UPitt), the Mayo Clinic tissue bank, and commercial providers: Cytomyx (Lexington, MA), Proteogenix (Culver City, CA), and Asterand (Detroit, MI). In addition, electronic files of microarray data on 271 tumors were obtained from the International Genomics Consortium (IGC; Phoenix, AZ). Criteria for inclusion for frozen specimens were: $\geq$ 0.1 g of frozen tissue, histologic verification of minimal necrosis ($\leq$ 20% of tumor tissue), and sufficient tumor representation ($\geq$ 60% of tissue examined). Histologic verification was performed by a pathologist at the institution providing the tissue sample, who visually estimated the percent tumor cells. Inclusion criteria for all specimens (tissues and microarray files) were: characterization as a poorly differentiated or undifferentiated primary tumor (American Joint Committee on Cancer grade 3 or 4, or "high grade" in pathology report), or a metastatic tumor; and classification by the original pathology report as one of the 15 tissue types on the Pathwork TOO test panel (Data Supplement Table 1, online only). Sixteen specimens were excluded due to off-panel morphology: 45 due to less than 60% tumor content, 23 due to more than 20% necrosis, and six due to microarray quality control failures. A total of 547 specimens met all inclusion criteria for the validation analyses, with no fewer than 25 specimens for each of the 15 tissues on the panel. Characteristics of patients and tumor specimens are presented in Table 1. All specimens were collected and de-identified under institutional review board approved protocols.

Specimen Processing and Gene Expression Assays

Each specimen processing laboratory was trained to perform the test, and proficiency in performing the assay at each laboratory was verified using known total RNA samples and tissue from known specimens (n = 8 to 10). All laboratories obtained the correct tissue identification for these performance training samples (data not shown).

For the 547 specimens in the validation cohort, 276 frozen tumor tissues were processed at the Clinical Genomics Facility of UPitt, Cogenics (Morrisville, NC), and the Mayo Clinic as outlined in Figure 1. Tissue processing methods have been previously described and additional details are presented in Data Supplement Table 2 (online only).[15] Samples were hybridized to one of three microarrays: Pathwork Diagnostics Pathchip, Affymetrix GeneChip HG-U133A or HG-U133 Plus 2. The arrays were scanned using the Affymetrix GCS3000 scanner and intensity levels calculated using Affymetrix GCOS 1.1.3 or 1.4. The resulting raw intensity data files (.CEL), including the 271-gene expression data files from IGC, were processed at Pathwork Diagnostics for automated analysis and report generation. Probe-level intensity data were transformed into gene expression values and standardized using the 121-gene standardization method whose performance has been previously described,[15,16] before applying the 1,550-gene profile. Data from the 276 frozen tumor specimens have been deposited to the NCBI Gene Expression Omnibus (GEO)[17] under series accession number GSE12630. GEO accession numbers for the 271-gene expression data files from IGC are listed in Data Supplement Table 3 (online only).

1,550-Gene Profile for Tumor Tissue of Origin Identification

The 1,550-gene profile was trained using gene expression data files from a panel of 2,039 tumors comprising 15 tissue types and 60 different morphologies, as illustrated in Figure 2 and detailed in Data Supplement Table 1. The training set included both primary and metastatic tumors and well-differentiated to undifferentiated tumors. None of the validation specimens were used for algorithm training.

The 1,550-gene profile is a proprietary algorithm that uses the expression level of 1,550 transcripts to perform pair-wise comparisons between the test sample and each of the 15 tissues on the test panel. The results are presented as 15 similarity scores, one for each tissue included in the test panel.

Before analysis of the clinical validation study data, the 1,550-gene profile was locked based on its performance with the training data. Similarity score thresholds for determining absence and presence of tissue in the sample were also locked. The similarity scores were probability based, with a reported range from 0 to 100, and all 15 scores sum to 100. A similarity score of 30 or above indicates the presence of a given tissue in the specimen; a similarity score of 5 or less indicates the absence of a given tissue. Similarity scores between 5 and 30 are considered indeterminate. These criteria were used to make a tissue determination for each specimen.

The Pathwork System Software and 1,550-gene profile produced an automated report (Fig 1) for each specimen. An assessment of the biologic

Table 1. Patients and Tumors Characteristics Included in This Study

Characteristic	No.	%
Tumor		
Metastatic	258	47
Primary		
Grade 3	185	34
Grade 4	68	12
Not graded*	36	7
Patient		
Age, years†		
< 50	142	26
50-59	133	24
60-69	139	25
≥ 70	132	24
Sex‡		
Male	254	46
Female	290	53

*Melanoma, thyroid, and lymphoma tumors are not normally graded.
†Age data were available for 546 of 547 patients.
‡Sex data were available for 544 of 547 patients.

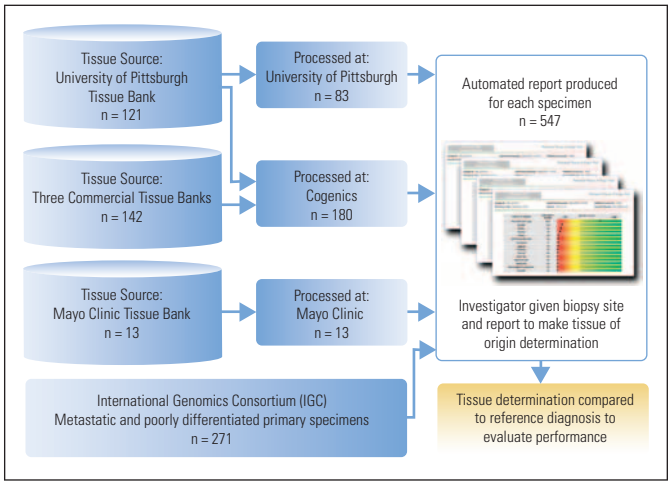

Fig 1. Validation study design. Gene expression data from 547 tumor samples generated by multiple laboratories were processed by the Pathwork Tissue of Origin test (Pathwork Diagnostics, Sunnyvale, CA) software. The test software transformed data into gene expression values, performed data verification and standardization, and generated reports that were evaluated in a blinded fashion by the investigators.

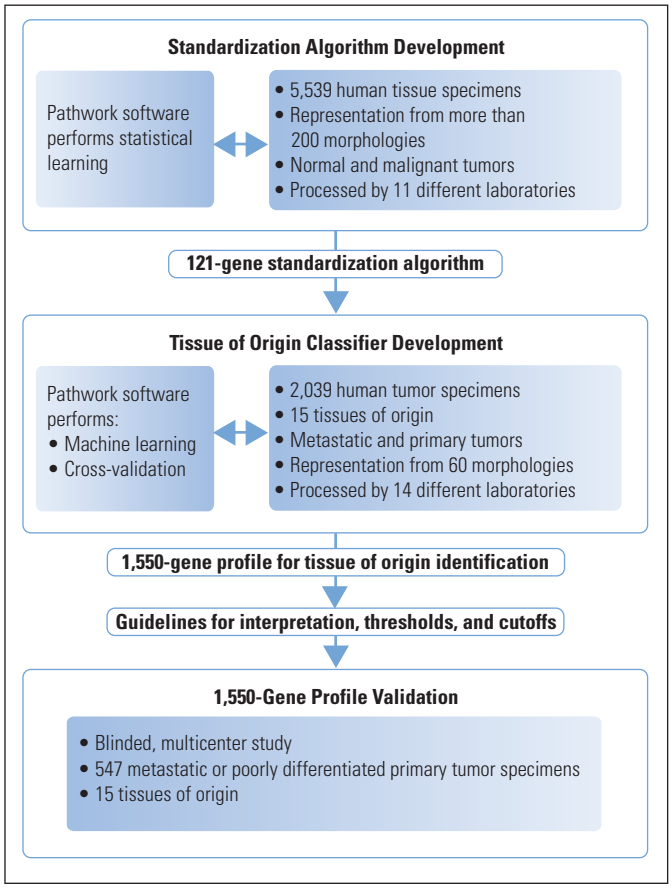

Fig 2. Development of the 1,550-gene profile to identify tissue of origin. A 121-gene standardization algorithm was used. The 1,550-gene profile for tissue of origin identification was trained using 2,039 primary and metastatic tumors. The algorithm was locked and thresholds for positive, negative, and indeterminate calls were predetermined before the multicenter validation.

plausibility for gene-tissue associations for the 60 genes with the strongest correlations with individual tissues is available in the online only Appendix.

Validation Study Design

The objective of this study was to determine the performance characteristics of the Pathwork Tissue of Origin Test in the identification of TOO for a series of metastatic and poorly differentiated or undifferentiated primary tumor specimens of known origin, which was considered the reference diagnosis. These specimens are representative of those that would likely be designated as uncertain primary cancer after initial histologic evaluation. The study evaluated agreement between the tissue determination made using the 1,550-gene profile and the reference diagnosis for each specimen. We also evaluated the nonagreement and indeterminate fractions.

Technical personnel performing the gene expression assays and investigators who interpreted the Pathwork Tissue of Origin Test results for making a tissue determination were blinded to patient sex, histology, or morphology information, and reference diagnosis. When making the tissue determination, investigators were provided only biopsy site and the 15 similarity scores for each specimen. Matching of reference diagnosis and the predicted site of origin was performed by an investigator not involved with any aspect of sample processing or tissue determination who was blinded to all the above information. Results were stratified by type of tissue (primary *v* metastatic), by processing site, and by site of origin, all of which are potential sources of variability.

Statistical Methods

Power calculations were based on the estimated 88% sensitivity found in cross-validation analyses of the training data set. Sample size was determined by calculating the minimum number of samples needed to detect a 5% reduction in performance (ie, a decrease from 88% to 83% sensitivity), determined to be clinically significant. One-tailed calculations indicated that 540 specimens would provide 95% power to detect this difference at a significance level of .05. We targeted no fewer than 25 samples per tissue type with a distribution reflecting the incidence of individual cancers, subject to specimen availability.

For each specimen, a tissue determination was made using the reported similarity scores and criteria described earlier, and compared to the reference diagnosis. A true-positive result was indicated when the tissue determination matched the reference diagnosis. When the tissue determination and the reference diagnosis did not match, the specimen was considered a false positive. For each tissue on the panel, sensitivity (or positive percent agreement) was defined as the ratio of true positive results to the total positive samples analyzed. Specificity (or negative percent agreement) was defined as the ratio: (1 − false positive)/(total tested including indeterminate − total positive). Diagnostic odds ratio was calculated as (sensitivity/(1 − specificity)/((1 − specificity)/sensitivity).[18]

RESULTS

Agreement With Reference Diagnosis

The 1,550-gene profile results showed 87.8% overall agreement with the reference diagnosis (480 of 547; 95% CI, 84.7% to 90.4%) for the 547 specimens. The overall sensitivity (positive percent agreement) and specificity (negative percent agreement) were 87.8% (95% CI, 84.7% to 90.4%) and 99.4% (95% CI, 98.3% to 99.9%), respectively (Table 2). Diagnostic odds ratios for all tissues are significantly greater than one, indicating that each of the individual tests is highly informative. Similarity scores reported for each of the 15 tissues on the panel for all samples are provided in Data Supplement Table 3. Overall rate of nonagreement for these specimens was 7.1% (39 of 547; 95% CI, 5.1% to 9.6%), and the rate of indeterminate calls was 5.1% (28 of 547; 95% CI, 3.4% to 7.3%; Table 3 and Data Supplement Table 4).

Analysis by Relevant Subgroups

The rates of agreement between the test result and the reference diagnosis ranged from 94.1% for breast cancer specimens (n = 68) to 72.0% for gastric and pancreatic cancer specimens (n = 25 each; Table 3). Performance differences between tissue sites were statistically significant ($\chi^2 = 42.02$; $P = .04$; $df = 28$; n = 547).

Performance of the test was found to be somewhat better with primary tumors (90.7% agreement; n = 289) than with metastatic specimens (84.5% agreement; n = 258) (Fisher's exact method two-sided $P = .04$). Rates of agreement between the test result and the reference diagnosis were 88.0%, 84.4%, 92.3%, and 89.7% at study sites 1 (Clinical Genomics Facility), 2 (Cogenics), 3 (Mayo Clinic), and 4 (IGC), respectively, and these differences were not statistically significant ($\chi^2 = 4.4$, $P = .62$; $df = 6$; n = 547).

Nonagreements and Indeterminates

Of the 39 tissue determinations that were in nonagreement with the reference diagnosis, 11 matched the biopsy site for that sample. Of the 28 specimens with indeterminate results, 25 reported no similarity score above 30, and three reported two similarity scores greater than 30, neither of which could be excluded as the biopsy site. In 11 of these 28 indeterminate samples, the highest similarity score was that of the reference diagnosis tissue, and in only one result was the reference diagnosis ruled out due to a similarity score less than 5. When the 28

Table 2. Sensitivity and Specificity of the 1,550-Gene Profile for Tissue of Origin Identification

Reference Diagnosis	Sample		Sensitivity			Specificity		
	Algorithm Development	Multicenter Validation	Positive % Agreement	Ratio	95% CI	Negative % Agreement	Ratio	95% CI
Bladder	62	28	78.6	22/28	59.0 to 91.7	100.0	519/519	99.3 to 100.0
Breast	444	68	94.1	64/68	85.6 to 98.4	98.3	471/479	96.7 to 99.3
Colorectal	253	56	92.9	52/56	82.7 to 98.0	99.2	487/491	97.9 to 99.9
Gastric	52	25	72.0	18/25	50.6 to 87.9	99.4	519/522	98.3 to 99.9
Germ cell	121	30	73.3	22/30	54.1 to 87.7	100.0	517/517	99.3 to 100.0
Hepatocellular	151	25	92.0	23/25	74.0 to 99.0	99.8	521/522	98.8 to 100.0
Kidney	41	39	94.9	37/39	82.7 to 99.4	99.8	507/508	98.9 to 100.0
Melanoma	221	26	80.8	21/26	60.6 to 93.4	99.8	520/521	98.9 to 100.0
Non-Hodgkin's lymphoma	97	33	93.9	31/33	79.8 to 99.3	99.4	511/514	98.3 to 99.9
Non–small cell lung	69	31	87.1	27/31	70.2 to 96.4	98.6	509/516	97.2 to 99.5
Ovarian	189	69	92.8	64/69	83.9 to 97.6	99.0	473/478	97.6 to 99.7
Pancreas	43	25	72.0	18/25	50.6 to 87.9	99.8	521/522	98.9 to 100.0
Prostate	105	26	88.5	23/26	69.8 to 97.6	100.0	521/521	99.3 to 100.0
Soft to tissue sarcoma	122	31	83.9	26/31	66.3 to 94.5	99.4	513/516	98.3 to 99.9
Thyroid	69	35	91.4	32/35	76.9 to 98.2	99.6	510/512	98.6 to 100.0
Overall	2,039	547	87.8	480/547	84.7 to 90.4	99.4	NA	98.3 to 99.9

indeterminate results were excluded, the overall accuracy was 92.5% (480 of 519).

DISCUSSION

Gene expression–based classifiers for clinical applications should demonstrate strong reproducibility in sample processing, analytic performance, and clinical reported result. In this study, we show that the Pathwork Tissue of Origin Test can reliably identify the TOO in 87.8% of the 547specimens tested, and in 84.5% of the metastatic specimens. This compares favorably with current clinical practice standards, such as IHC, which has shown 66% to 88% agreement in blinded tests.[19-22] The performance of this test also compares favorably with other gene expression–based TOO classifiers with reported accuracies in the range of 76% to 89%.[10-14,23,24] Moreover, the results of this clinical validation study are consistent with the 86.8% agreement reported in our previous study.[15]

Published gene expression–based studies that show possible clinical application are criticized for one or more common flaws: reuse of the training samples in reported results, post hoc modification of the algorithm or thresholds, inadequate blinding, inadequate study size, and inappropriate handling of indeterminate results in reported performance.[25,26] Many groups have published multigene algorithms and results that demonstrate the promise of gene expression–based classifiers in TOO identification.[10-14,23,24] These studies have been restricted to smaller numbers of specimens (< 120), often dominated by well-differentiated primary cancers, and have often allowed post hoc modifications or enhancements to the algorithm design or thresholds. For example, in the study by Ma and coauthors where a panel of 92 genes was developed to identify 32 different tumor types, the same training set was repeatedly used to test different iterations of the classifier, and the final performance was evaluated in 119 tumors where representation from each tumor type ranged from 1 to 10 specimens.[23] Thus, in this test, correct identification of one single specimen was interpreted as 100% accuracy for that tissue type. Likewise, in a recent study by Rosenfeld et al, performance of a microRNA–based classifier was evaluated in 83 specimens, and representation of each of 22 tissue types ranged from 2 to 8 samples.[27] Clearly, these studies were inadequately sized to establish true diagnostic performance. In contrast, this validation study used 547 independent specimens with minimum tissue representation of 25 samples. Furthermore, Rosenfeld et al allowed post hoc enhancement of the test's performance by introducing a combination union classifier where sensitivity was calculated based on correct identification of TOO by either one of two algorithms (decision tree or k-nearest neighbor). Overall accuracy for the decision tree alone was 72% (60 of 83) for all samples and 59% (13 of 22) when only metastatic tumor samples were considered.

This is, to our knowledge, the largest clinical validation study of a gene expression assay for TOO determination to date. The study was designed and executed to avoid the common flaws mentioned earlier: all of the specimens used in the validation of the test were newly acquired; the algorithm was locked and thresholds predetermined based on the training set before the analysis of the validation specimens; indeterminate results are appropriately included in the reported performance; specimen identity was masked until the final analysis; and this study is the first to be adequately sized to provide performance data sufficient to support clinical use of a microarray-based test for TOO determination. Other strengths of this study are the wide range of tissues of origin evaluated, the characteristics of the challenging specimens, and the use of multiple laboratories for tissue processing and microarray analysis.

In a clinical scenario, the uncertainty of a tumor's origin usually arises in the context of metastatic and/or poorly differentiated to undifferentiated malignancies, and some of the previously published gene expression–based classifiers have shown decreased performance with less differentiated tumors.[12] Our results show that this test can identify the tissue of origin in poorly differentiated and undifferentiated tumor specimens, which is the clinically relevant population, since well-differentiated tumors rarely present a diagnostic challenge.

Table 3. Effect of Possible Sources of Variability in Tumor Tissue of Origin Test Performance

Performance by	No. of Specimens	Agreement		Nonagreement		Indeterminate	
		No.	%	No.	%	No.	%
Reference diagnosis*							
Bladder	28	22*	78.6	4	14.3	2	7.1
Breast	68	64	94.1	4	5.9	0	< 0.1
Colorectal	56	52	92.9	4	7.1	0	< 0.1
Gastric	25	18	72.0	4	16.0	3	12.0
Germ cell	30	22	73.3	3	10.0	5	16.7
Hepatocellular	25	23	92.0	0	< 0.1	2	8.0
Kidney	39	37	94.9	1	2.6	1	2.6
Melanoma	26	21	80.8	2	7.7	3	11.5
Non-Hodgkin's lymphoma	33	31	93.9	1	3.0	1	3.0
Non–small-cell lung	31	27	87.1	2	6.5	2	6.5
Ovarian	69	64	92.8	3	4.3	2	2.9
Pancreas	25	18	72.0	5	20.0	2	8.0
Prostate	26	23	88.5	1	3.8	2	7.7
Soft tissue sarcoma	31	26	83.9	3	9.7	2	6.5
Thyroid	35	32	91.4	2	5.7	1	2.9
Overall	547	480	87.8	39	7.1	28	5.1
Overall 95% CI		84.7 to 90.4		5.1 to 9.6		3.4 to 7.3	
Metastatic *v* primary tumor samples†							
Metastatic	258	218†	84.5	23	8.9	17	6.6
Poorly and undifferentiated primary	289	262	90.7	16	5.5	11	3.8
At each processing laboratory‡							
IGC	271	243†	89.7	18	6.6	10	3.7
Cogenics	180	152	84.4	15	8.3	13	7.2
CGF-UPitt	83	73	88.0	5	6.0	5	6.0
Mayo clinic	13	12	92.3	1	7.7	0	< 0.1

Abbreviations: IGC, International Genomics Consortium; CGF, Clinical Genomics Facility; UPitt, University of Pittsburgh.
*$\chi^2 = 42.02$; $P = .04$; $df = 28$; N = 547.
†Fisher's exact method two-sided $P = .04$.
‡$\chi^2 = 4.4$; $P = .62$; $df = 6$; N = 547.

Interestingly, we found a small but statistically significant reduction in the accuracy of the test when primary cancers and metastatic tumors were compared (90.7% and 84.5%, respectively). However, the performance in the metastatic samples still compares favorably with IHC, which is the current standard of care for tissue of origin identification. Importantly, similarly sized validation studies of IHC panels in clinical use today have not been performed, and in one of the largest blinded studies of IHC performance, Dennis and coauthors[20] reported 67% accuracy (20 of 30) in metastatic samples using a predetermined panel of ten antibodies.

One of the limitations of our study was the inability to independently verify the reference diagnosis used to assess the accuracy of the test. The diagnosis was extracted from the surgical pathology report that accompanied the specimen at the time it was banked. It is possible that some of these diagnoses are incorrect and this could result in an over- or underestimation of the test's accuracy. Another limitation is the requirement for frozen tissue. In many instances, the need to perform a tissue of origin determination is not known until after the specimen has been fixed. Although for this study we specified the need for ≥ 0.1 g of tumor tissue, the assay requires 1 μg of total RNA; this quantity is obtainable from a needle core specimen if adequate tumor representation is present. However, validation of needle core biopsy material and/or formalin-fixed paraffin embedded tissues should be performed in separate studies.

This test is designed to be interpreted by a pathologist in conjunction with pathologic examination of the tissue and in consultation with the surgeon/oncologist. This is especially important in patients where the differential between a primary and a metastatic tumor is being considered, since the metastatic tumor specimens are expected to contain surrounding noncancerous tissue from the biopsy site. Due to the blinded nature of the study, the pathologists interpreting the TOO test results did not know the morphologic features of the specimen and/or the clinical features of the patient. It is expected that the clinical performance of the test will be favorably influenced by the availability of this information. In addition, it is important to note that although the test was trained and validated on a preselected panel of 15 tumor types which represent approximately 89% of the incident solid tumors[28] that are known to produce distant metastases, the possibility that an uncertain primary cancer might originate from a tissue site not covered by the panel must be considered. It is also important to acknowledge that in certain clinical situations, the need to test a sample that does not meet the quality control criteria for the test ($\geq$ 60% tumor and $\leq$ 20% necrosis) could arise. As described previously, the best assay performance is achieved when these two criteria are met,[15] but there are insufficient data to adequately determine the impact of testing suboptimal specimens. Furthermore, the assay has been approved by the US Food and Drug Administration based on the stated sample quality thresholds.

In conclusion, this study represents the first adequately sized, multicenter validation of a prespecified diagnostic test for tissue of origin determination restricted to poorly differentiated and undifferentiated primary cancers and metastatic tumors. Our results confirm the diagnostic value of the 1,550-gene profile used in the Pathwork Tissue of Origin Test. This test should be a valuable addition or alternative to currently available diagnostic methods for the evaluation of uncertain primary cancers.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: Ljubomir J. Buturovic, Pathwork Diagnostics (C); C. Ted Rigl, Pathwork Diagnostics (C); W. David Henner, Pathwork Diagnostics (C); Glenda G. Anderson, Pathwork Diagnostics (C) **Consultant or Advisory Role:** Federico A. Monzon, Pathwork Diagnostics (C) **Stock Ownership:** Ljubomir J. Buturovic, Pathwork Diagnostics; C. Ted Rigl, Pathwork Diagnostics; W. David Henner, Pathwork Diagnostics; Glenda G. Anderson, Pathwork Diagnostics **Honoraria:** Federico A. Monzon, Pathwork Diagnostics **Research Funding:** Federico A. Monzon, Pathwork Diagnostics **Expert Testimony:** None **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Federico A. Monzon, Ljubomir J. Buturovic, C. Ted Rigl, Glenda G. Anderson
Financial support: Glenda G. Anderson
Administrative support: Maureen Lyons-Weiler, C. Ted Rigl, Christin Sciulli
Provision of study materials or patients: Federico A. Monzon, Maureen Lyons-Weiler, C. Ted Rigl, Fabiola Medeiros
Collection and assembly of data: Federico A. Monzon, Maureen Lyons-Weiler, C. Ted Rigl, Christin Sciulli
Data analysis and interpretation: Federico A. Monzon, Ljubomir J. Buturovic, C. Ted Rigl, W. David Henner, Catherine I. Dumur, Fabiola Medeiros, Glenda G. Anderson
Manuscript writing: Federico A. Monzon, Ljubomir J. Buturovic, W. David Henner, Catherine I. Dumur, Fabiola Medeiros, Glenda G. Anderson
Final approval of manuscript: Federico A. Monzon, Ljubomir J. Buturovic, W. David Henner, Catherine I. Dumur, Glenda G. Anderson

REFERENCES

1. National Comprehensive Cancer Network: NCCN Clinical Practice Guidelines in Oncology, Occult Primary, v. 1.2007. http://www.nccn.org/professionals/physician_gls/PDF/occult.pdf
2. Briasoulis E, Tolis C, Bergh J, et al: ESMO Guidelines Task Force: ESMO minimum clinical recommendations for diagnosis, treatment and follow-up of cancers of unknown primary site (CUP). Ann Oncol 16:i75-i76, 2005 (suppl 1)
3. Pavlidis N, Briasoulis E, Hainsworth J, et al: Diagnostic and therapeutic management of cancer of an unknown primary. Eur J Cancer 39:1990-2005, 2003
4. Bugat R, Bataillard A, Lesimple T, et al: FNCLCC: Summary of the standards, options and recommendations for the management of patients with carcinoma of unknown primary site (2002). Br J Cancer 89:S59-S66, 2003 (suppl 1)
5. Varadhachary GR, Abbruzzese JL, Lenzi R: Diagnostic strategies for unknown primary cancer. Cancer 100:1776-1785, 2004
6. Chu PG, Weiss LM: Keratin expression in human tissues and neoplasms. Histopathology 40: 403-439, 2002
7. Schapira DV, Jarrett AR: The need to consider survival, outcome, and expense when evaluating and treating patients with unknown primary carcinoma. Arch Intern Med 155:2050-2054, 1995
8. Pavlidis N, Merrouche Y: The importance of identifying CUP subsets, in Fizazi K (ed): Carcinoma of Unknown Primary Site. New York, NY, Taylor & Francis Group, 2006, pp 37-48
9. Hillen HF: Unknown primary tumours. Postgrad Med J 76:690-693, 2000
10. Bloom G, Yang IV, Boulware D, et al: Multi-platform, multi-site, microarray-based human tumor classification. Am J Pathol 164:9-16, 2004
11. Buckhaults P, Zhang Z, Chen YC, et al: Identifying tumor origin using a gene expression-based classification map. Cancer Res 63:4144-4149, 2003
12. Ramaswamy S, Tamayo P, Rifkin R, et al: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98:15149-15154, 2001
13. Su AI, Welsh JB, Sapinoso LM, et al: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 61:7388-7393, 2001
14. Tothill RW, Kowalczyk A, Rischin D, et al: An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 65:4031-4040, 2005
15. Dumur CI, Lyons-Weiler M, Sciulli C, et al: Interlaboratory performance of a microarray-based gene expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers. J Mol Diagn 10:67-77, 2008
16. Moraleda J, Grove N, Tran Q, et al: Gene expression data analytics with interlaboratory validation for identifying anatomical sites of origin of metastatic carcinomas. J Clin Oncol 22:862s, 2004 (suppl; abstr 9625)
17. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207-210, 2002
18. Glas AS, Lijmer JG, Prins MH, et al: The diagnostic odds ratio: A single indicator of test performance. J Clin Epidemiol 56:1129-1135, 2003
19. Brown RW, Campagna LB, Dunn JK, et al: Immunohistochemical identification of tumor markers in metastatic adenocarcinoma: A diagnostic adjunct in the determination of primary site. Am J Clin Pathol 107:12-19, 1997
20. Dennis JL, Hvidsten TR, Wit EC, et al: Markers of adenocarcinoma characteristic of the site of origin: Development of a diagnostic algorithm. Clin Cancer Res 11:3766-3772, 2005
21. DeYoung BR, Wick MR: Immunohistologic evaluation of metastatic carcinomas of unknown origin: An algorithmic approach. Semin Diagn Pathol 17:184-193, 2000
22. Park S-Y, Kim B-H, Kim J-H, et al: Panels of immunohistochemical markers help determine primary sites of metastatic adenocarcinoma. Arch Pathol Lab Med 131:1561-1567, 2007
23. Ma XJ, Patel R, Wang X, et al: Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med 130:465-473, 2006
24. Talantov D, Baden J, Jatkoe T, et al: A quantitative reverse transcriptase-polymerase chain reaction assay to identify metastatic carcinoma tissue of origin. J Mol Diagn 8:320-329, 2006
25. Simon R, Radmacher MD, Dobbin K, et al: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95:14-18, 2003
26. Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332-7341, 2005
27. Rosenfeld N, Aharonov R, Meiri E, et al: MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol 26:462-469, 2008
28. Estimated new cancer cases and deaths by sex for all sites, US, 2007. http://www.cancer.org/docroot/MED/content/downloads/MED_1_1x_CFF2007_Estimated_New_Cases_Deaths_by_Sex_US.asp

Acknowledgment

We thank Amelia Hensler and Aprel Delo for tissue procurement and preparation, Jill Hagenkord and Melissa Price for retrieval of pathology information, Rebecca Deeter for statistical analyses, Raji Pillai, PhD, for the biologic plausibility analysis (supplemental methods), Shawn Becker for helpful advice and Jane Seck for editorial support.

Appendix

Biologic Basis for Tissue Classification by the Pathwork Tissue of Origin Test

Introduction

The Pathwork Tissue of Origin (TOO) test (Pathwork Diagnostics, Sunnyvale, CA) uses 1,550 probe sets specific for 1,550 markers or genes to identify the tissue of origin of the specimen being tested. The principle of this genomic test is that the combination of probe sets provides highly specific information on tissue identity which cannot be obtained from any individual probe set.

These 1,550 probe sets were selected using machine learning methods. It is a well-known concern that machine learning methods, when used inappropriately to develop a multiplex classifier, can select markers that are spurious or artifacts (Simon R: J Nat Cancer Inst 97:866-867, 2005; Simon R: Cancer Biomark 2:89-96, 2006; Dupuy A, Simon R: Nat Cancer Inst 98:147-157, 2007). One method for confirmation of the validity of markers selected by such a classifier is biologic plausibility. Biologic plausibility of the TOO test is based on the use of published sources to confirm that the markers selected have been shown by others, using traditional laboratory techniques, to have relevance to the classification, thereby showing independent biologic relevance to the classification.

This report is an analysis of the biologic plausibility of the markers selected for use in the TOO test. The assessment of biologic plausibility for the 1,550 markers used in the TOO test involves the elucidation of the function of these markers and their behavior in different tissue types. Even though understanding the biology of all transcripts in the genome remains one of the important goals of biology, the current incompleteness of this knowledge does not prevent the generation of highly accurate diagnostic tests. Diagnostic tests such as the TOO test make observations on gene expression levels, but do not make any intervention on the genes or affect the underlying biology. Nor is the performance of the TOO test contingent on the knowledge of underlying biology. The authors of this report believe that continuing research on biologic function will provide information on the function of all probe sets in the future. As stated by Simon (Simon R: Cancer Biomark 2:89-96, 2006) "It is, of course, desirable to understand the mechanistic relationship of the components of an expression signature, but the classifier can be validated without such an understanding."

The information that is available in the literature on some of the probe sets in the TOO test signature reveals plausibility for some but not all markers. For several, expression is not restricted to one or a few tissues, and probe sets include those that are essential for cellular functions across all tissues.

It is important to note that the markers or genes do not necessarily have to be involved with cancer-related pathways, because the TOO test is specific for 15 tissues and was not designed to discriminate normal from cancerous tissue. Moreover, the markers or genes do not necessarily have to be present in a given tissue, because absence can provide the basis for the discrimination of that tissue from the other 14 tissues. Also, absence is often under-cited in the literature. This underscores the point that the TOO test is entirely based on empirically and statistically defined gene expression profiles that can form the basis for the test, provided that microarray and algorithm design, along with training, testing, and clinical validation, are performed in a reliable manner.

Objective

Demonstrate that markers or genes identified by the Pathwork Tissue of Origin algorithm have biologic plausibility.

Methods and Materials and Probe Set Ranking

Published sources were used to confirm that the markers selected have been shown, using traditional laboratory techniques, to have independent biologic relevance to the classification.

The clinical validation set was used to assess the correlation between a given tissue of interest and all other tissues for each of the 1,550 markers. This set is independent of the development (training) set.

The probe set ranking process follows, using bladder tissue probe set #1 of 1,550 as an example. (1) The standardized expression (SE) values for each of the 28 bladder specimens and for the remaining specimens were recorded. (2) Each specimen was then assigned a class label of either 1 (positive; ie, bladder) or 0 (negative; ie, not bladder). The Pearson correlation between the SE values, a continuous variable, and the class label, a binary variable, was computed for probe set #1. In this manner, Pearson correlations and *P* values for each of the 1,550 probe sets for bladder were determined. (3) The absolute values of the correlations were sorted in decreasing order, and the top four markers (n = 60 of 1,550, or 3.9%) were examined further as described in this report. This probe set ranking process was repeated for each of the remaining 14 tissues of interest.

Selection Criteria for Literature Searches

First, annotations were obtained for all probe sets on the Pathchip microarray using NetAffx search and annotations engine (www.affymetrix.com). These annotations are updated on a quarterly basis and include gene symbol and gene title, as well as biologic, cellular, and molecular process information from the Gene Ontology project. Second, the gene symbol was used to perform PubMed searches for published articles on the marker of interest. Variations of the search were performed using combinations of the gene symbol, gene title, tissue or malignancy

of interest, to ensure that a reasonable effort was made to find any reports on the biologic basis of the correlation. The articles used for the assessment are listed as parenthetical references.

Biologic Plausibility Criteria

Positively correlated markers. If the published literature showed clear evidence of marker involvement in the tissue malignancy of interest, that is indicated by yes in the biologic basis column in the respective tables below (Appendix Tables A1 to A15); if no link was found, it is indicated by unknown.

Unless there is a publication that has explored the role of a marker in a specific tissue or malignancy and has categorically ruled out any role, the authors propose that there may be a link which this work may be the first to uncover, even though there are no published reports at the present time.

Negatively correlated markers. For a negatively correlated marker with a specific tissue of interest, it is expected that few or no reports of association with this tissue will be found. Negative association is not always reported and thus, the lack of reports is tentatively considered supportive unless evidence of high expression in that tissue is found. This interpretation is indicated by yes in the biologic basis column.

Some negatively correlated markers are broadly expressed in certain tissue types (eg, epithelial cells) and thus the lack of expression in a tissue where this expression is not expected is also considered supportive evidence.

Results

An assessment of biologic plausibility for the top four markers for each tissue is presented in the following Appendix Tables. In some tissues most, if not all, of the top four markers represent genes involved in functions specific to the respective tissues. These include lung, prostate, and thyroid. In others, the top four markers may not represent genes with tissue-specific functions, but their involvement in the specific malignancy has been described in the literature. In other cases, such as pancreas and melanoma, there is no clear indication that the markers represent genes with functions specific to the tissue in question. In such cases, the authors surmise that the underlying biology of the genes is not yet fully understood, and that this may be the first observation of a strong correlation between the tissue of interest and expression of these markers in malignancies. For 10 of 15 tissues, two different probe sets from the same gene are among the top four markers. Tables with the probe set IDs, gene symbols, and gene titles for the top four markers are provided below for each tissue, and a brief discussion of the marker follows each Table (Appendix Tables A1-A15).

The top two markers for bladder (Appendix Table A1) are both probe sets for DHSR2, a member of the short-chain dehydrogenase/reductase SDR family, whose function is not known and significance in metastatic cancer is not evident from the literature (Shafqat N, Shafqat J, Eissner G, et al: Cell Mol Life Sci 63:1205-1213, 2006; Gabrielli F, Donadel G, Bensi G, et al: Eur J Biochem 232(2):473-477, 1995), LYPD3, which is also reported by the gene name C4.4A, has been detected mainly in metastasizing carcinoma cells and has been shown by in situ hybridization to be upregulated during progression of urothelial cancers (Paret C, Bourouba M, Beer A, et al: Int J Cancer 115:724-733, 2005; Hansen LV, Gardsvoll H, Nielsen BS, et al: Biochem J 380:845-857, 2004; Smith BA, Kennedy WJ, Harnden P, et al: Cancer Res 61:1678-1685, 2001; Fletcher GC, Patel S, Tyson K, et al: Br J Cancer 88:579-585, 2003). GALNT1 is involved in O-linked oligosaccharide biosynthesis, and its differential expression is linked to the aberrant carbohydrate antigen expression in gastric, colorectal and epithelial cancers (Cheng SL, Huang Liu R, Sheu JN, et al: Biol Pharm Bull 29:655-669, 2006; Mandel U, Hassan H, Therkildsen MH, et al: Glycobiology 9:43-52, 1999; Radvanyi L, Singh-Sandhu D, Gallichan S, et al: Proc Natl Acad Sci U S A 102:11005-11010, 2005; Chang GT, Jhamai M, van Weerden WM, et al: Endocr Relat Cancer 11:815-822, 2004).

TRPS1 is reported to be overexpressed in breast cancers, based on microarray studies and corroborating methodologies, such as in situ hybridization and immunohistochemistry (Appendix Table A2). It is located on chromosome 8q23-q24 which also bears the *MYC* gene, known to be amplified in breast cancers (Radvanyi L, Singh-Sandhu D, Gallichan S, et al: Proc Natl Acad Sci U S A 102:11005-11010, 2005; Chang GT, Jhamai M, van Weerden WM, et al: Endocr Relat Cancer 11:815-822, 2004; Savinainen KJ, Linja MJ, Saramaki OR, et al: Br J Cancer 90:1041-1046, 2004). IRX5, a homeobox protein, has been studied in cardiac tissue, but no reports are available of its role in normal or malignant breast tissue (Rosati B, Grau F, McKinnon D: J Mol Cell Cardiol. 40(2):295-302, 2006). Similarly, no reports as yet link EFDH1, an EF-hand domain family member, to breast cancer (Lucas B, Grigo K, Erdmann S, et al: Oncogene 24:6418-6431, 2005). SCGB2A2, mammaglobin 1, is overexpressed in breast cancers, and has been proposed for use in breast cancer diagnostics and treatment (Lacroix M: Endocr Relat Cancer 13:1033-1067, 2006; L'Esperance S, Popa I, Bachvarova M, et al: Int J Oncol 29:5-24, 2006; Zafrakas M, Petschke B, Donner A, et al: BMC Cancer 6:88, 2006).

CDX1 and CDX2, homeobox transcription factors, direct the development and maintenance of intestinal epithelium. CDX2 is overexpressed in colorectal cancers (Witek ME, Nielsen K, Walters R, et al: Clin Cancer Res 11:8549-8556, 2005; Erickson LA, Papouchado B, Dimashkieh H, et al: Endocr Pathol 15:247-252, 2004; Xu XL, Yu J, Zhang HY, et al: World J Gastroenterol 10:3441-3454, 2004; Pilozzi E, Onelli MR, Ziparo V, et al: J Pathol 204:289-295, 2004; Guo RJ, Huang E, Ezaki T, et al: J Biol Chem 279:36865-36875, 2004; Appendix Table A3). NOX1 (two of four markers) NADPH oxidase 1, is dominantly expressed in the colon and is implicated in the pathogenesis of colon cancer (Rokutan K, Kawahara T, Kuwano Y, et al: Antioxid Redox Signal 8:1573-1582, 2006; Brewer AC, Sparks EC, Shah AM: Free Radic Biol Med 40:260-274, 2006; Szanto I, Rubbia-Brandt L, Kiss P, et al: J Pathol 207:164-176, 2005; Fukuyama M, Rokutan K, Sano T, et al: Cancer Lett 221:97-104, 2005).

SPINK1, a serine peptidase inhibitor, shows elevated serum levels in patients with gastric cancer compared to those with benign gastrointestinal malignancies (Wiksten JP, Lundin J, Nordling S, et al: Histopathology 46(4):380-388, 2005; Solakidi S, Dessypris A, Stathopoulos GP, et al: Clin Biochem 37:56-60, 2004; Appendix Table A4). LGALS4, a galactose-binding lectin, is differentially expressed in carcinoid tumors in different regions of the gastrointestinal tract, and is higher in tumor than in normal tissues (Rumilla KM, Erickson LA, Erickson AK, et al: Endocr

Pathol 17:243-249, 2006; Lotan R, Ito H, Yasui W, et al: Int J Cancer 56:474-480, 1994). TRIM31 has not yet been reported as a marker for gastric cancers (Dokmanovic M, Chang BD, Fang J, et al: Cancer Biol Ther 1:24-27, 2002). FUT2, a fucosyltransferase, is expressed in normal gastric epithelium, and is implicated in the gastric carcinogenesis process (Lopez-Ferrer A, de Bolos C: Glycoconj J 19:13-21, 2002; Lopez-Ferrer A, de Bolos C, Barranco C, et al: Gut 47:349-356, 2000; Koda Y, Soejima M, Wang B, et al: Eur J Biochem 246:750-755, 1997).

TEAD4, a transcription factor, is amplified and overexpressed in testicular germ cell tumors (Skotheim RI, Autio R, Lind GE, et al: Cell Oncol 28:315-326, 2006; Appendix Table A5). RAB15, a member of the RAS oncogene family, regulates endocytic trafficking. A specific role in germ cell tissue or cancers has not been reported, as is the case for MIER2, mesoderm induction early response 1, a DNA binding protein (Zuk PA, Elferink LA: J Biol Chem 275:26754-26764, 2000; Olkkonen VM, Peterson JR, Dupree P, et al: Gene 138:207-211, 1994; Elferink LA, Anzai K, Scheller RH: J Biol Chem 267:22693, 1992; Howell M, Itoh F, Pierreux CE, Valgeirsdottir S, et al: Dev Biol 214:354-369, 1999). Ubiquitination is required for all developmental stages of spermatogenesis. The authors infer that ubiquitin associated protein 2 (UBAP2) is involved in this process (Kwon J: Exp Anim 56:71-77, 2007).

KCNJ16, a potassium channel family member is expressed in kidneys and may play a role in the regulation of fluid and pH balance (Liu Y, McKenna E, Figueroa DJ, et al: Cytogenet Cell Genet 90:60-63, 2000; Appendix Table A6). The scaffolding protein PDZK1 is also involved in ion exchange (Thomson RB, Wang T, Thomson BR, et al: Proc Natl Acad Sci U S A 102:13331-13336, 2005). Carbonic anhydrase XII (CA12) has been identified as a marker of metastatic renal cell carcinoma (Kim HL, Seligson D, Liu X, et al: J Urol 173:1496-1501, 2005).

Hemopexin (HPX) is a plasma glycoprotein expressed only in the liver. Studies in other organisms suggest that its expression is augmented in hepatocellular carcinogenesis (Darabi A, Gross S, Watabe M, et al: Cancer Lett 95:153-159, 1995; Alam J, Smith A: J Biol Chem 264:17637-17640, 1989; Appendix Table A7). F12, coagulation factor XII is a procoagulant protein involved in activating the intrinsic clotting pathway. Some studies suggest its role in tumor cell progression, angiogenesis, invasion, and metastasis. The structure of F12 includes two EGF homologous domains, suggesting that it might mimic EGF biologic characteristics and act as a growth factor. Members of the coagulation cascade have been targeted in some trials in the treatment of cancer (Wang X, Wang E, Kavanagh JJ, et al: J Transl Med 3:25, 2005). GALT, an evolutionarily conserved enzyme central to D-galactose metabolism, is highly expressed in the liver (Heidenreich RA, Mallee J, Rogers S, et al: Pediatr Res 34:416-419, 1993; Elsas LJ, Lai K, Saunders CJ, et al: Mol Genet Metab 72:297-305, 2001).

In the lung, surfactant, pulmonary-associated protein B (SFTPB) has been found to be a member of a 10-gene classifier that can distinguish between head and neck squamous cell carcinoma (HNSCC) and lung squamous cell carcinoma (Appendix Table A8). It has been postulated that this classifier could determine the origin of squamous cell carcinomas in the lungs of patients with previous head and neck malignancies (Vachani A, Nebozhyn M, Singhal S, et al: Clin Cancer Res 13:2905-2915, 2007). In a separate study, SFTPB was one of three markers that could detect lymph node metastasis in non–small-cell lung cancer (NSCLC) patients and were able to distinguish between benign and positive lymph nodes (Xi L, Coello MC, Litle VR, et al: Clin Cancer Res 12:2484-2491, 2006). Surfactant, pulmonary-associated protein C (SFTPC) has also been identified as a useful diagnostic marker for lung cancer (Li R, Todd NW, Qiu Q, et al: Clin Cancer Res 13:482-487, 2007; Chen Y, Pacyna-Gengelbach M, Deutschmann N, et al: Biochem Biophys Res Commun 353:559-564, 2007; Okubo T, Knoepfler PS, Eisenman RN, et al: Development 132:1363-1374, 2005).

The top four markers for lymphoma are negatively correlated (Appendix Table A9). They appear to serve important cellular functions, in epithelial tumors. SH3BP4 encodes an SH3-domain binding protein. SH3 domains are found in a variety of proteins, including tyrosine kinases, such as Abl and Src, and are involved in cell signaling, and functions related to the cytoskeleton (Kairouz R, Daly RJ: Breast Cancer Res 2:197-202, 2000; Ren R, Mayer BJ, Cicchetti P, et al: Science 259:1157-1161, 1993). EFNA1, or ephrin A1, is expressed in normal epithelial cells, and its overexpression has been described in prostate, gastric, esophageal, colon, lung, liver, mammary, and ovarian cancers so its absence in lymphoid tissues is expected. Its involvement in cell deadhesion and movement may explain the correlation of its overexpression with observed poor prognosis (Xu F, Zhong W, Li J, et al: Anticancer Res 25:2943-2950, 2005; Herath NI, Spanevello MD, Sabesan S, et al: BMC Cancer 6:144, 2006). RHBDF1 is a rhomboid family protease. Interestingly, some ephrin family members are cleaved by such proteases (Pascall JC, Brown KD: Biochem Biophys Res Commun 317:244-252, 2004). Discoidin domain receptor family member 1 (DDR1) belongs to a family of surface receptors that bind to several types of collagen and facilitate cell adhesion that is known to be associated with several cancers, including breast and ovarian cancers, pituitary adenomas, NSCLC, and acute lymphoblastic leukemia (Turashvili G, Bouchal J, Baumforth K, et al: BMC Cancer 7:55, 2007; Ford CE, Lau SK, Zhu CQ, et al: Br J Cancer 96:808-814, 2007; Yoshida D, Teramoto A: J Neurooncol 82:29-40, 2006; Chiaretti S, Li X, Gentleman R, et al: Clin Cancer Res 11:7209-7219, 2005).

There was no evidence in the literature of involvement of the top four melanoma markers with this tissue (Appendix Table A10). Since three of these markers are negatively correlated with melanoma it is not unexpected that no association has been reported. ZFP106 is a zinc finger protein, and such proteins participate in protein-protein and protein-DNA interactions, and can function in the regulation of transcription as well as DNA repair (Witkiewicz-Kucharczyk A, Bal W: Toxicol Lett 162:29-42, 2006). ELMO3 plays a role in cell engulfment and motility (Katoh H, Fujimoto S, Ishida C, et al: Brain Res 1073-1074:103-108, 2006). CD24 is a small, heavily glycosylated cell surface protein which is expressed in hematologic malignancies and in a large variety of solid tumors. Its overexpression enhances the metastatic potential of cancer cells and has been correlated with poor prognosis (Su MC, Hsu C, Kao HL, et al: Cancer Lett 235:34-39, 2006; Lindley S, Dayan CM, Bishop A, et al: Diabetes 54:92-99, 2005).

Mucin 16 (MUC16) is a well-validated cell surface marker for ovarian cancer and is thought to facilitate the peritoneal metastasis of ovarian tumors (Appendix Table A11). It carries the peptide epitope CA125, which is used to monitor the progression and recurrence of ovarian cancer (Chen Y, Clark S, Wong T, et al: Cancer Res 67:4924-4932, 2007; Gubbels JA, Belisle J, Onda M, et al: Mol Cancer 5:50, 2006). Differential expression of MEIS1, a cofactor of homeobox genes, has been shown to be present in myeloid leukemogenesis, but is not evident in ovarian cancer. MEIS1 is a homeodomain transcription factor coexpressed with HOXA9 in most human acute myeloid leukemias (AMLs), and is also

highly expressed in neuroblastomas (Wang GG, Pasillas MP, et al: Blood 106:254-264, 2005; Spieker N, van Sluis P, Beitsma M, et al: Genomics 71:214-221, 2001; Serrano E, Lasa A, Perea G, et al: Acta Haematol 116:77-89, 2006; Camos M, Esteve J, Jares P, et al: Cancer Res 66:6947-6954, 2006). *PAX8* is a regulatory developmental box gene which has been shown to be highly expressed in epithelial ovarian cancer, but absent in the precursor ovarian surface epithelia of healthy individuals (Bowen NJ, Logani S, Dickerson EB, et al: Gynecol Oncol 104:331-337, 2007).

One of the top negatively correlated markers for pancreas is lactate hydrogenase B; aberrant expression has been found in prostate, lung, and colorectal cancers, and also in testicular germ cell tumors and ependymomas (Appendix Table A12). It is thought to act through a mechanism which involves promoter hypermethylation (Chen Y, Zhang H, Xu A, et al: Lung Cancer 54:95-102, 2006; Maekawa M, Inomata M, Sasaki MS, et al: Clin Chem 48:1938-1945, 2002). Another top marker for pancreas is transcobalamin I, a vitamin B12-binding protein that appears to be involved in innate defense against infections, as suggested in a study of acute cholera. It was identified as a marker in NSCLC. Altered vitamin B12 binding was observed in hepatocellular carcinoma (Remmelink M, Mijatovic T, Gustin A, et al: Int J Oncol 26:247-258, 2005; Kanai T, Takabayashi T, Kawano Y, et al: Jpn J Clin Oncol 34:346-351, 2004; Flach CF, Qadri F, Bhuiyan TR, et al: Infect Immun 75:2343-2350, 2007). Annexin A2 is associated with the progression of lung adenocarcinoma, and is also differentially expressed in hepatocellular carcinoma, melanoma, glioma, leukemia, thyroid cancer, and osteosarcoma (Olbryt M, Jarzab M, Jazowiecka-Rakus J, et al: Gene Expr 13:191-203, 2006; Ishiyama T, Kano J, Anami Y, et al: Cancer Sci 98:50-57, 2007; Yoon SY, Kim JM, Oh JH, et al: Int J Oncol 29:315-327, 2006; Wang AG, Yoon SY, Oh JH, et al: Biochem Biophys Res Commun 345:1022-1032, 2006; Tatenhorst L, Rescher U, Gerke V, et al: Neuropathol Appl Neurobiol 32:271-277, 2006; Madoiwa S, Someya T, Hironaka M, et al: Thromb Res 119:229-240, 2007; Musholt TJ, Hanack J, Brehm C, et al: World J Surg 29:472-482, 2005; Olwill SA, McGlynn H, Gilmore WS, et al: Thromb Res 115:109-114, 2005; Guzman-Aranguez A, Olmo N, Turnay J, et al: J Cell Biochem 94:178-193, 2005; Gillette JM, Chan DC, Nielsen-Preiss SM: J Cell Biochem 92:820-832, 2004).

Kallikreins (KLK) are drawing increasing attention for their role as biomarkers for screening, diagnosis, prognosis, and monitoring in various cancers. Two of the top four markers in prostate tissue represent KLK3 (kallikrein-related peptidase 3, or prostate-specific antigen), an extensively studied marker whose involvement in malignancies has been clearly established, and whose level of expression in malignant tissue is the basis of diagnostic tests currently in use for prostate cancer (Paliouras M, Borgono C, Diamandis EP: Cancer Lett 249:61-79, 2007; Appendix Table A13). The other two markers represent KLK2 (kallikrein-related peptidase 2), which has been reported as an emerging complementary prostate cancer biomarker (Pakkala M, Hekim C, Soininen P, et al: J Pept Sci 13:348-353, 2007).

Cadherins, including CDH1 (E-cadherin), are a family of transmembrane glycoproteins that mediate cell-cell adhesion in epithelial cells and play a crucial role in cell differentiation. CDH1-mediated adhesion is lost during most epithelial cancer development (Charrasse S, Comunale F, Gilbert E, et al: Oncogene 23:2420-2430, 2004; Ottaiano A, De Chiara A, Fazioli F, et al: Anticancer Res. 25:4519-4526, 2005; Appendix Table A14). ANKRD25, or ankyrin repeat domain 25, is a growth regulatory factor (Harada JN, Bower KE, Orth AP, et al: Genome Res 15:1136-1144, 2005). Desmosomes are intercellular junctions that tightly link adjacent epithelial cells. Desmoplakin is an essential component of desmosomes and serves to anchor intermediate filaments to the plaque (Wang J, Bu DF, Li T, et al: Br J Dermatol 153:558-564, 2005; Miettinen M: Am J Pathol 138:505-513, 1991). The presence of desmosomes is a feature frequently used to identify epithelial origin by electron microscopy in poorly differentiated malignancies. TPD52 or tumor protein 52 is a gene implicated in cell proliferation, apoptosis and vesicle trafficking, and is frequently over-expressed in cancer. It has been identified as a chromosome 8q21 amplification target in breast and prostate carcinoma (Sims AH, Finnon P, Miller CJ, et al: Int J Radiat Biol 83:409-420, 2007; Boutros R, Fanayan S, Shehata M, et al: Biochem Biophys Res Commun 325:1115-1121, 2004; Tiacci E, Orvietani PL, Bigerna B, et al: Blood 105:2812-2820, 2005; Byrne JA, Balleine RL, Schoenberg Fejzo M, et al: Int J Cancer 117:1049-1054, 2005).

The top four markers for thyroid tissue are probe sets for thyroglobulin, PCSK2, TITF1, and TSHR, all of which perform thyroid-specific functions (Harish K: Endocr Regul 40:53-67, 2006; Puskas LG, Juhasz F, Zarva A, et al: Cell Mol Biol (Noisy-le-grand) 51:177-186, 2005; Weber F, Shen L, Aldred MA, et al: J Clin Endocrinol Metab 90:2512-2521, 2005; Zighelboim I, Goodfellow PJ, Schmidt AP, et al: Clin Cancer Res 13:2882-2889, 2007; Trueba SS, Auge J, Mattei G, et al: J Clin Endocrinol Metab 90:455-462, 2005; Milas M, Mazzaglia P, Chia SY, et al: Surgery 141:137-146, 2007; Appendix Table A15). These have individually been shown by different groups to be involved in malignancies. Thyroglobulin levels in tissue and serum are important in the management of differentiated thyroid cancers (Harish K: Endocr Regul 40:53-67, 2006). PCSK2 and thyroid stimulating hormone receptor have been reported to show differential expression in thyroid cancers and are identified as potential diagnostic markers for thyroid cancer (Puskas LG, Juhasz F, Zarva A, et al: Cell Mol Biol (Noisy-le-grand) 51:177-186, 2005; Weber F, Shen L, Aldred MA, et al: J Clin Endocrinol Metab 90:2512-2521, 2005; Milas M, Mazzaglia P, Chia SY, et al: Surgery 141:137-146, 2007).

Conclusion

The authors conclude that biologic plausibility is evident for the majority of top classifiers (43 of 60; 71%). However, there are some probe sets and genes in this set, for which the underlying biology has not yet been fully understood. This work may represent the first observation of a strong positive or negative correlation between the tissue of interest and expression, or lack thereof, of those markers in specific malignancies.

The essence of the TOO test, demonstrated in the clinical validation study reported in the accompanying text, is that the levels of expression of 1,550 markers or genes used in combination across 15 tissues follow a pattern that allows the test to distinguish tissues of origin with high agreement with clinical truth, defined in this study as the reference diagnosis.

Biologic plausibility has been demonstrated for a subset of the markers used in identifying the 15 tissues of origin included in this test. For example, the top four markers for thyroid show clear biologic plausibility. In other tissues such as bladder, the biologic plausibility of the markers cannot be established from the literature at this time, but the strong statistical relationship of these markers suggests that biologic relevance, although not known at this time, will be revealed as research continues on biologic function of all transcripts in the human genome.

Table A1. Bladder

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
206463_s_at	*DHRS2*	Dehydrogenase/reductase (SDR family) member 2	Unknown
214079_at	*DHRS2*	Dehydrogenase/reductase (SDR family) member 2	Unknown
204952_at	*LYPD3*	LY6/PLAUR domain containing 3	Yes
201724_s_at	*GALNT1*	UDP-*N*-acetyl-alpha-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase 1 (GalNAc-T1)	Unknown

NOTE. Pathwork Tissue of Origin Test from Pathwork Diagnostics, Sunnyvale, CA.

Table A2. Breast

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
209343_at	*EFHD1*	EF-hand domain family, member D1	Unknown
218502_s_at	*TRPS1*	Trichorhinophalangeal syndrome I	Yes
210239_at	*IRX5*	Iroquois homeobox protein 5	Unknown
206378_at	*SCGB2A2*	Secretoglobin, family 2A, member 2	Yes

Table A3. Colon

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
206387_at	*CDX2*	Caudal type homeobox transcription factor 2	Yes
206418_at	*NOX1*	NADPH oxidase 1	Yes
207217_s_at	*NOX1*	NADPH oxidase 1	Yes
206430_at	*CDX1*	Caudal type homeobox transcription factor 1	Yes

NOTE. Pathwork Tissue of Origin Test from Pathwork Diagnostics, Sunnyvale, CA.

Table A4. Gastric Tissue

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
206239_s_at	*SPINK1*	Serine peptidase inhibitor, Kazal type 1	Yes
204272_at	*LGALS4*	Lectin, galactoside-binding, soluble, 4 (galectin 4)	Yes
208170_s_at	*TRIM31*	Tripartite motif-containing 31	Unknown
210608_s_at	*FUT2*	Fucosyltransferase 2 (secretor status included)	Yes

NOTE. Pathwork Tissue of Origin Test from Pathwork Diagnostics, Sunnyvale, CA.

Table A5. Germ Cell

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
41037_at	*TEAD4*	TEA domain family member 4	Yes
59697_at	*RAB15*	RAB15, member RAS onocogene family	Unknown
44822_s_at	*MIER2*	Mesoderm induction early response 1, family member 2	Unknown
219192_at	*UBAP2*	Ubiquitin associated protein 2	Unknown

Table A6. Kidney

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
219564_at	*KCNJ16*	Potassium inwardly-rectifying channel, subfamily J, member 16	Unknown
205380_at	*PDZK1*	PDZ domain containing 1	Unknown
215867_x_at	*CA12*	Carbonic anhydrase XII	Yes
214164_x_at	*CA12*	Carbonic anhydrase XII	Yes

Table A7. Hepatocellular (liver)

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
39763_at	*HPX*	Hemopexin	Yes
210013_at	*HPX*	Hemopexin	Yes
203179_at	*GALT*	Galactose-1-phosphate uridylyltransferase	Unknown
205774_at	*F12*	Coagulation factor XII (Hageman factor)	Unknown

Table A8. Non–Small-Cell Lung

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
213936_x_at	*SFTPB*	Surfactant, pulmonary-associated protein B	Yes
37004_at	*SFTPB*	Surfactant, pulmonary-associated protein B	Yes
211735_x_at	*SFTPC*	Surfactant, pulmonary-associated protein C	Yes
205982_x_at	*SFTPC*	Surfactant, pulmonary-associated protein C	Yes

Table A9. Non-Hodgkin's Lymphoma

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
222258_s_at	*SH3BP4*	SH3-domain binding protein 4	Yes
202023_at	*EFNA1*	Ephrin-A1	Yes
218686_s_at	*RHBDF1*	Rhomboid 5 homolog 1 (*Drosophila*)	Yes
1007_s_at	*DDR1*	Discoidin domain receptor family, member 1	Yes

Table A10. Melanoma

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
217781_s_at	*ZFP106*	Zinc finger protein 106 homolog (mouse)	Unknown
219411_at	*ELMO3*	Engulfment and cell motility 3	Yes
208651_x_at	*CD24*	CD24 molecule	Yes
209771_x_at	*CD24*	CD24 molecule	Yes

Table A11. Ovary

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
220196_at	*MUC16*	Mucin 16, cell surface associated	Yes
204069_at	*MEIS1*	Meis1, myeloid ecotropic viral integration site 1 homolog (mouse)	Unknown
213917_at	*PAX8*	Paired box gene 8	Yes
121_at	*PAX8*	Paired box gene 8	Yes

Table A12. Pancreas

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
213564_x_at	*LDHB*	Lactate dehydrogenase B	Yes
201030_x_at	*LDHB*	Lactate dehydrogenase B	Yes
205513_at	*TCN1*	Transcobalamin I (vitamin B12 binding protein, R binder family)	Unknown
201590_x_at	*ANXA2*	Annexin A2	Yes

Table A13. Prostate

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
209855_s_at	*KLK2*	Kallikrein-related peptidase 2	Yes
209854_s_at	*KLK2*	Kallikrein-related peptidase 2	Yes
204583_x_at	*KLK3*	Kallikrein-related peptidase 3	Yes
204582_s_at	*KLK3*	Kallikrein-related peptidase 3	Yes

Table A14. Soft-Tissue Sarcoma

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
201131_s_at	*CDH1*	Cadherin 1, type 1, E-cadherin (epithelial)	Yes
218418_s_at	*ANKRD25*	Ankyrin repeat domain 25	Unknown
200606_at	*DSP*	Desmoplakin	Yes
201690_s_at	*TPD52*	Tumor protein D52	Yes

Table A15. Thyroid

Probe Set ID	Gene Symbol	Gene Title	Biological Basis
203673_at	*TG*	Thyroglobulin	Yes
204870_s_at	*PCSK2*	Proprotein convertase subtilisin/kexin type 2	Yes
211024_s_at	*TITF1*	Thyroid transcription factor 1/thyroid transcription factor 1	Yes
210055_at	*TSHR*	Thyroid stimulating hormone receptor	Yes

2795WTG03312009

M364