

# Molecular Classification of Human Cancers Using a 92-Gene Real-Time Quantitative Polymerase Chain Reaction Assay

Xiao-Jun Ma, PhD; Rajesh Patel, PhD; Xianqun Wang, PhD; Ranelle Salunga, BSc; Jaji Murage, BSc; Rupal Desai, BSc; J. Todd Tuggle, BSc; Wei Wang, PhD; Shirley Chu, BSc; Kimberly Stecker, BSc; Rajiv Raja, PhD; Howard Robin, MD; Mat Moore, PhD; David Baunoch, PhD; Dennis Sgroi, MD; Mark Erlander, PhD

• **Context.**—Correct diagnosis of the tissue origin of a metastatic cancer is the first step in disease management, but it is frequently difficult using standard pathologic methods. Microarray-based gene expression profiling has shown great promise as a new tool to address this challenge.

**Objective.**—Adoption of microarray technologies in the clinic remains limited. We aimed to bridge this technological gap by developing a real-time quantitative polymerase chain reaction (RT-PCR) assay.

**Design.**—We constructed a microarray database of 466 frozen and 112 formalin-fixed, paraffin-embedded (FFPE) samples of both primary and metastatic tumors, measuring expression of 22 000 genes. From the microarray database, we used a genetic algorithm to search for gene combinations optimal for multitumor classification. A 92-gene RT-PCR assay was then designed and

used to generate a database for 481 frozen and 119 FFPE tumor samples.

**Results.**—The microarray-based K-nearest neighbor classifier demonstrated 84% accuracy in classifying 39 tumor types via cross-validation and 82% accuracy in predicting 112 independent FFPE samples. We successfully translated the microarray database to the RT-PCR platform, which allowed an overall success rate of 87% in classifying 32 different tumor classes in the validation set of 119 FFPE tumor samples.

**Conclusions.**—The RT-PCR-based expression assay involving 92 genes represents a powerful tool for accurately and objectively identifying the site of origin for metastatic tumors, especially in the cases of cancer of unknown primary. The assay uses RT-PCR and routine FFPE samples, making it suitable for rapid clinical adoption.

(*Arch Pathol Lab Med.* 2006;130:465–473)

Correct diagnosis of the tissue origin of a cancer contributes critically into treatment decisions because current therapies are based largely on anatomical site. However, accurate determination of the origin of a metastatic cancer has been challenging.<sup>1,2</sup> Identifying the primary site of certain metastatic cancers is particularly difficult, resulting in approximately 3% to 5% of all tumors being classified into the category of cancer of unknown primary (CUP).<sup>3,4</sup> Currently, histopathology, based on tissue architecture and cell morphology is the cornerstone of cancer diagnosis, but it can be subjective. In a study as-

sessing the accuracy of pathologists, less than 50% of metastatic tumors of known primary sites were correctly diagnosed.<sup>5</sup> Therefore, developing new methods of objectively and accurately identifying tissue origin and classifying histologic subtypes of all cancers remains an important clinical need.

The current standard of care employs immunohistochemistry as an adjunct to conventional clinical and pathological investigations, especially in tracing the site of origin for CUP cases.<sup>6,7</sup> However, even using a growing panel of antibodies, the success rate in identifying the origin of metastatic tumors is limited to ~67%.<sup>7</sup> For patients presenting with CUP, the success rate drops to 25%, even after exhaustive workup.<sup>8,9</sup> Recently, gene expression profiling using microarrays has shown great promise providing highly accurate diagnosis of cancer types at the molecular level.<sup>10–17</sup> In a seminal study, a support vector machine algorithm classifier was demonstrated to provide 78% classification accuracy in predicting 14 common tumor types.<sup>14</sup> Bloom et al<sup>10</sup> combined multiple tumor datasets to obtain a large collection of tumors spanning 21 tumor types, and built a neural network-based classifier with 85% accuracy. More recently, Tothill et al<sup>17</sup> built a 14-class support vector machine classifier with 89% overall accuracy. The success of these studies supports the premise that, despite varying degrees of dedifferentiation and

---

Accepted for publication December 13, 2005.

From Research and Development, Arcturus Bioscience, Inc, Carlsbad and Mountain View, Calif (Drs Ma, Patel, X. Wang, Mr Salunga, Mr Murage, Ms Desai, Mr Tuggle, Dr W. Wang, Ms Chu, Ms Stecker, and Drs Raja and Erlander); Research & Development, US Labs, Irvine, Calif (Drs Moore and Baunoch); Department of Pathology, Sharp Memorial Hospital, San Diego, Calif (Dr Robin); and Department of Pathology, Harvard Medical School, Molecular Pathology Research Unit, Massachusetts General Hospital, Boston (Dr Sgroi).

Drs Ma, Patel, X. Wang, W. Wang, Raja, and Erlander, Messrs Salunga, Murage, and Tuggle, and Mss Desai, Chu, and Stecker are employed by Arcturus Bioscience, Inc, and have a financial interest in the 92-gene assay. All other authors have no relevant financial interest in the products or companies described in this article.

Reprints: Mark Erlander, PhD, Arcturus BioScience, Inc, 2715 Loker Ave, West, Carlsbad, CA 92008 (e-mail: merlander@arcturusdx.com).

colonization in different tissue milieu, a tumor retains sufficient signatures of their cellular origin. Therefore, it is theoretically possible to build a comprehensive gene expression database spanning a majority of tumor types and use it as a clinical diagnostic tool. However, one major technical hurdle exists for such a strategy in the clinical setting: microarray technology remains complex and time-consuming, and it is currently limited as a research tool. In the clinic, real-time quantitative polymerase chain reaction (RT-PCR) is the "gold standard" for gene expression analysis, but it has a much lower capacity in measuring, at most, a few hundred genes.

In this study we used a comprehensive microarray-based tumor database to develop a 92-gene panel RT-PCR assay for classifying 32 tumor classes, which is, to our knowledge, the broadest coverage of tumor types reported to date. The conversion of a microarray database to the RT-PCR platform makes it suitable for rapid clinical adoption.

## MATERIALS AND METHODS

### Tumor Samples

Tumor samples were obtained from several sources (CytoMx, Boston, Mass; ProteoGenex, Culver City, Calif; Cureline, South San Francisco, Calif; Massachusetts General Hospital, Boston). Patient consent and institutional review board approval were obtained for all samples in accordance with the National Institutes of Health human research study guidelines. Diagnosis of histologic types was made by full pathologic workup including hematoxylin-eosin (H&E) staining and immunohistochemistry when necessary. All cases were reviewed by at least 2 independent pathologists. Frozen tumor samples were embedded in optimal cutting temperature (OCT) compound and 5- $\mu$ m sections were prepared; one section from each sample was evaluated by H&E staining. Formalin-fixed, paraffin-embedded (FFPE) tumor samples were sectioned into 7- $\mu$ m sections and evaluated by H&E staining. The average tumor content of all samples was approximately 65%.

### RNA Extraction and Amplification

For frozen tumor samples, a whole-tissue section, 5  $\mu$ m in thickness, was used for RNA isolation. For FFPE samples, RNA was isolated from four 7- $\mu$ m-thick tissue sections using the Paradise Reagent System (Arcturus Bioscience, Inc, Mountain View, Calif). Isolated RNA was subjected to deoxyribonuclease treatment followed by T7 in vitro transcription using the RiboAmp kit (frozen samples) or the Paradise system (FFPE samples) according to manufacturer's instructions (Arcturus).

For microarray analysis, both frozen and FFPE samples were amplified for 2 rounds. Labeled cRNA was generated during the second round in the presence of 5-[3-aminoallyl]uridine 5'-triphosphate (Sigma-Aldrich, St Louis, Mo). Universal Human Reference RNA (Stratagene, San Diego, Calif) was amplified in the same manner. The purified amplified RNA was then conjugated to Cy5 (experimental samples) or Cy3 (reference sample) dye (Amersham Biosciences, Piscataway, NJ).

For RT-PCR analysis (TaqMan, Applied Biosystems, Foster City, Calif), 2 rounds of T7-based amplification were performed for frozen samples and one round for the FFPE samples.

### Microarray Analysis

A custom-designed 22000-gene oligonucleotide (60mer) microarray was fabricated using ink-jet in situ synthesis technology (Agilent Technologies, Palo Alto, Calif). Cy5-labeled sample RNA and Cy3-labeled reference RNA were cohybridized at 65°C, 1 $\times$  hybridization buffer (Agilent Technologies). Slides were washed at 37°C with 0.1 $\times$  SSC/0.005% Triton X-102. Image analysis was performed using image analysis software from Agilent.

Raw Cy5/Cy3 ratios per array were normalized using nonlin-

ear local regression<sup>18</sup>; no background adjustment was made because of generally low local background values. Between-array normalization was accomplished by subtracting the median log<sub>2</sub> ratios genewise across all arrays.

### Multiclass K-Nearest Neighbor Classification

We implemented a weighted K-nearest neighbor (KNN) algorithm as follows. To classify a sample of unknown class, we first calculated its distance,  $d$ , to each instance in the training set either as the vector angle theta (for microarray data)<sup>19</sup> or 1-Pearson correlation coefficient (for RT-PCR data). The top KNNs were examined for their class labels, and the unknown case was assigned to the class with the largest summed weight ( $1/d$ ).

To assign a confidence level for each KNN prediction, we used the hypergeometric distribution to compute the probability of observing  $l$  instances of tumor class  $L$  among  $k$  nearest neighbors given the total number  $n$  of class  $L$  instances and the total number  $N$  of all instances in the training set:

$$p(l; N, n, k) = \frac{\binom{k}{l} \binom{N-k}{n-l}}{\binom{N}{n}}$$

$P$  values were then converted to confidence categories as follows: high,  $P < .0001$ ; medium,  $P \geq .0001$  to  $P < .001$ ; low,  $P \geq .001$  to  $P < .01$ ; unclassifiable,  $P \geq .01$ .

Starting with microarray data with 22000 genes, it was important to include only informative genes for distance calculation. We used leave-one-out cross-validation to select the optimal number  $n$  of genes per tumor class. During each round of leave-one-out cross-validation, the top  $n$  genes were selected by using the differential gene expression analysis software package called linear models for microarray data (limma).<sup>20</sup>

### Gene Selection via Genetic Algorithms

To reduce the search space for the genetic algorithm (GA), we first identified candidate genes differentially expressed between different tumor classes at different granularity levels. We traversed the tumor taxonomy tree starting from the root node, extracting the top 30 genes differentially expressed between the children nodes. To select genes at each step, we used limma to rank all genes on the array according to evidence of differential expression and then performed receiver operating characteristic analysis on the top 1000 genes to further select the top 30 with the largest partial area under the curve (at 0.2) values.<sup>21</sup>

We adapted the GA-maximum likelihood algorithm described by Ooi and Tan<sup>22</sup> to use KNN. The GA parameters were population = 100, generations = 200, crossover rate = 0.8, mutation rate = 0.005, and the universal stochastic sampling method for selection. The fitness score was the 3-fold cross-validation accuracy on the training set (KNN with  $k = 5$ ).

We conducted preliminary GA/KNN runs to determine the approximate size range to search for and found that gene sets with sizes from 60 to 80 performed similarly and better than those in the lower-size ranges. We then conducted 100 independent runs using the size range 61 to 80 as input parameter. After each run, the best gene set was obtained by sorting, in ascending order, the sum of cross-validation error in the training set ( $X_c$ ) and the independent test set error ( $X_i$ ), then by  $X_i$ , and finally by  $X_c$ , as previously described.<sup>22</sup>

### TaqMan PCR Assay

TaqMan assays were designed using Primer Express (Applied Biosystems) and using minor groove binder probes to aim for short amplicon sizes (<80 bp).

An aliquot (200–500 ng) of the amplified RNA from each tumor sample was converted into cDNA via reverse transcription using the Paradise Reagent System. TaqMan assays using 1/30th of the reverse transcribed material were performed in a 10- $\mu$ L volume in a 384-well plate using the ABI 7900HT instrument (Applied

**Table 1. Description of Tumor Classifications**

| Tumor Type                   | n  | Histologic Types   |
|------------------------------|----|--|
| Adrenal                      | 9  | Pheochromocytoma   |
| Brain                        | 32 | Astrocytoma, glioblastoma, glioblastoma multiforme oligoastrocytoma, oligodendroglioma                               |
| Breast                       | 49 | Adenocarcinoma of breast ductal, medullary, lobular, mucinous, papillary   |
| Carcinoid-intestine          | 10 | Carcinoid tumor of duodenum, ileum   |
| Cervix-adeno                 | 8  | Adenocarcinoma of cervix, endometrioid   |
| Cervix-squamous              | 14 | Carcinoma of cervix adenosquamous, large and squamous cell nonkeratinizing, squamous cell nonkeratinizing            |
| Endometrium                  | 23 | Adenocarcinoma of endometrium endometrioid, endometrioid with squamous features, papillary serous                    |
| Gallbladder                  | 5  | Adenocarcinoma of gallbladder  |
| Germ-cell-ovary              | 9  | Teratoma immature, mature, carcionid, ovary yolk sac   |
| GIST                         | 10 | Gastrointestinal stromal tumor of stomach  |
| Kidney                       | 15 | Carcinoma of kidney renal cell chromophil, papillary, chromophobe with sarcomatoid features, clear and granular cell |
| Leiomyosarcoma               | 14 | Leiomyosarcoma, leiomyosarcoma metastatic  |
| Liver                        | 14 | Carcinoma of liver hepatocellular, carcinoma of liver hepatocellular metastatic                                      |
| Lung-adeno-large cell        | 26 | Adenocarcinoma of lung, bronchioloalveolar, mucinous, large cell, large cell neuroendocrine                          |
| Lung-small                   | 10 | Carcinoma of lung small cell   |
| Lung-squamous                | 12 | Carcinoma of lung squamous cell  |
| Lymphoma-B cell              | 11 | Lymphoma extranodal marginal zone B cell, large B-cell diffuse   |
| Lymphoma-Hodgkin             | 9  | Lymphoma hodgkin, nodular sclerosing   |
| Lymphoma-T cell              | 5  | Lymphoma peripheral T cell   |
| Meningioma                   | 8  | Meningioma, fibroblastic, meningothelial, secretory  |
| Mesothelioma                 | 12 | Mesothelioma of pleura, epithelial, mixed  |
| Osteosarcoma                 | 10 | Osteosarcoma   |
| Ovary-clear                  | 16 | Adenocarcinoma of ovary clear cell, clear cell papillary serous  |
| Ovary-serous                 | 19 | Adenocarcinoma of ovary papillary serous   |
| Pancreas                     | 26 | Adenocarcinoma of pancreas, mucinous, acinar cell  |
| Prostate                     | 13 | Adenocarcinoma of prostate   |
| Skin-basal cell              | 8  | Carcinoma of skin basal cell, nodular  |
| Skin-melanoma                | 13 | Malignant melanoma   |
| Skin-squamous                | 10 | Carcinoma of skin squamous cell  |
| Small and large bowel        | 41 | Adenocarcinoma of colon, duodenum, rectum, small intestine   |
| Soft-tissue-liposarcoma      | 5  | Liposarcoma  |
| Soft-tissue-MFH              | 13 | Histiocytoma malignant fibrous, myxofibrosarcoma   |
| Soft-tissue-sarcoma-synovial | 9  | Sarcoma synovial, biphasic, monophasic   |
| Stomach-adeno                | 12 | Adenocarcinoma of stomach, mucinous, signet ring cell  |
| Testis-other                 | 15 | Carcinoma of testis embryonal, germ cell tumor, teratoma, mixed germ cell  |
| Testis-seminoma              | 13 | Seminoma of testis   |
| Thyroid-follicular-papillary | 14 | Carcinoma of thyroid follicular, papillary   |
| Thyroid-medullary            | 7  | Carcinoma of thyroid medullary   |
| Urinary bladder              | 29 | Carcinoma of bladder, transitional cell, papillary, urothelial   |

\* GIST indicates gastrointestinal stromal tumor; MFH, malignant fibrous histiocytoma.

Biosystems). The samples were heated to 50°C for 2 minutes, 95°C for 10 minutes, and followed by 45 cycles of 95°C for 15 seconds and 60°C for 1 minute.

**PCR Data Analysis**

For each sample, the threshold cycle (CT) values for the 5 reference genes were averaged to obtain CT<sub>ref</sub>. The relative expression level of a target gene was expressed as ΔCT = CT<sub>ref</sub> - CT<sub>target</sub>. The ΔCT values for each gene were then z-transformed by subtracting the mean and dividing by the standard deviation across all samples. Frozen and FFPE samples were transformed separately.

**Statistical Analysis**

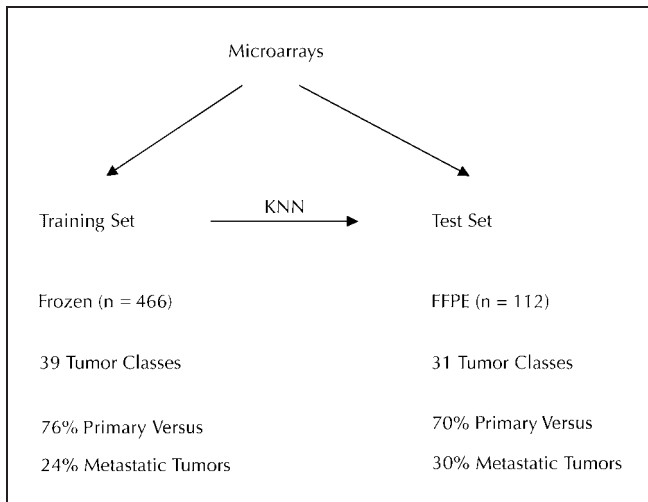
All statistical procedures were carried out in R, an open source statistical programming environment (<http://www.r-project.org>). The Bioconductor R packages were used extensively for microarray data analysis (<http://bioconductor.org>). Gene ontology analysis was performed using Gostat.<sup>23</sup>

**RESULTS**

**Construction of Microarray Tumor Database**

To develop a comprehensive reference database to aid in tumor classification, we selected 578 tumors represent-

ing a wide spectrum of tissue origins and histologic subtypes. We annotated the tumor samples into 39 tumor classes based on tissue origin and histopathology (Table 1). In selecting the tumor samples, the following issues were considered. First, our database should span the majority of tumor sites, including both epithelial and non-epithelial origins. Second, within each tissue/organ site, we should cover all major histologic subtypes, representing tumors of different cell types within the same tissue/organ. Third, because the primary goal of this study is to aid in the diagnosis of metastatic cancer, metastatic tumors were selected whenever possible. However, as is well known, biopsies from metastatic cancer are less readily available than those from primary tumors. Therefore, our tumor collection consists of 75% primary and 25% metastatic tumors. Note that these tumors do not necessarily represent the frequency of metastatic cancers in the population but rather were selected to provide adequate sample sizes for expression profiling. A final consideration is the clinical setting of the envisioned diagnostic test. Currently, biopsy samples in the clinic have been stored main-



**Figure 1.** Schematic of microarray database construction and multiclass prediction. KNN indicates K-nearest neighbor; FFPE, formalin-fixed, paraffin-embedded.

ly as FFPE tumor blocks; therefore, we wanted to assess the performance of our diagnostic test on FFPE samples. Of the 578 tumor samples, 466 were from fresh-frozen biopsies and 112 were from FFPE blocks.

We performed gene expression profiling on these tumor samples using an oligonucleotide microarray of 22,000 genes. The RNA from a mixture of a diverse set of cell lines was used as reference in 2-color hybridization. Expression values of each gene thus were represented as log ratios (base 2) relative to this reference.

### Multiclass Classification by Microarray Database

Before we attempted to translate the microarray database to the RT-PCR platform, we needed to establish that the gene expression data we collected would allow accurate classification for the many tumor types included herein. Previous attempts using various algorithms have demonstrated high accuracies in predicting, at most, 21 tumor types.<sup>10</sup> We used the frozen samples as the training set for building our classifiers and reserved the FFPE samples solely for independent validation (Figure 1).

We chose the classic KNN algorithm to build our multiclass classifier. We first performed leave-one-out cross-validation on the training set; that is, holding out each sample in turn as a test to calculate classification error. Importantly, within each leave-out-one cross-validation round, discriminating genes were reselected, excluding the held-out sample using a linear model procedure.<sup>20</sup> Selecting 30 genes per tumor class during leave-out-one cross-validation and using  $k = 5$ , KNN classified the frozen sample set with an overall accuracy of 84% (Table 2). In comparison, when we randomly permuted the tumor class labels, we obtained an overall accuracy of about 4% from a 5-fold cross-validation. Therefore, our gene expression data demonstrated highly significant and accurate classification for as many as 39 tumor classes.

We next examined classification performance according to tumor differentiation. For the subset of tumor samples with available tumor differentiation information ( $n = 191$ ), prediction accuracies were 94%, 84%, and 71% for well-differentiated, moderately differentiated, and poorly or undifferentiated tumors, respectively, indicating a de-

**Table 2.** Performance of Microarray-Based K-Nearest Neighbor Classifier\*

| Tumor Class                  | Frozen |          | FFPE |          |
|------------------------------|--------|----------|------|----------|
|                              | n      | Accuracy | n    | Accuracy |
| Adrenal                      | 7      | 1.00     | 2    | 1.00     |
| Brain                        | 16     | 0.94     | 16   | 1.00     |
| Breast                       | 43     | 0.98     | 6    | 0.83     |
| Carcinoid-intestine          | 8      | 0.88     | 2    | 1.00     |
| Cervix-adeno                 | 8      | 0.38     |      |          |
| Cervix-squamous              | 13     | 0.77     | 1    | 1.00     |
| Endometrium                  | 13     | 0.85     | 10   | 0.80     |
| Gallbladder                  | 5      | 0.00     |      |          |
| Germ-cell-ovary              | 8      | 0.00     | 1    | 0.00     |
| GIST                         | 10     | 0.90     |      |          |
| Kidney                       | 11     | 0.91     | 4    | 1.00     |
| Leiomyosarcoma               | 13     | 0.77     | 1    | 1.00     |
| Liver                        | 14     | 0.93     |      |          |
| Lung-adeno-large cell        | 18     | 0.72     | 8    | 0.38     |
| Lung-small                   | 8      | 0.75     | 2    | 1.00     |
| Lung-squamous                | 10     | 0.70     | 2    | 1.00     |
| Lymphoma-B                   | 7      | 0.71     | 4    | 0.25     |
| Lymphoma-Hodgkin             | 9      | 0.89     |      |          |
| Lymphoma-T cell              | 5      | 0.40     |      |          |
| Meningioma                   | 7      | 1.00     | 1    | 1.00     |
| Mesothelioma                 | 10     | 1.00     | 2    | 1.00     |
| Osteosarcoma                 | 7      | 0.86     | 3    | 1.00     |
| Ovary-clear                  | 14     | 0.79     | 2    | 1.00     |
| Ovary-serous                 | 14     | 0.93     | 5    | 0.80     |
| Pancreas                     | 24     | 0.92     | 2    | 1.00     |
| Prostate                     | 11     | 0.91     | 2    | 1.00     |
| Skin-basal cell              | 5      | 1.00     | 3    | 0.33     |
| Skin-melanoma                | 10     | 1.00     | 3    | 1.00     |
| Skin-squamous                | 6      | 0.83     | 4    | 0.50     |
| Small and large bowel        | 33     | 0.94     | 8    | 1.00     |
| Soft-tissue-liposarcoma      | 5      | 0.60     |      |          |
| Soft-tissue-MFH              | 11     | 0.82     | 2    | 1.00     |
| Soft-tissue-sarcoma-synovial | 7      | 1.00     | 2    | 1.00     |
| Stomach-adeno                | 8      | 0.25     | 4    | 0.50     |
| Testis-other                 | 14     | 0.93     | 1    | 1.00     |
| Testis-seminoma              | 10     | 1.00     | 3    | 1.00     |
| Thyroid-follicular-papillary | 12     | 1.00     | 2    | 1.00     |
| Thyroid-medullary            | 7      | 1.00     |      |          |
| Urinary bladder              | 25     | 0.80     | 4    | 0.75     |
| Overall                      | 466    | 0.84     | 112  | 0.82     |

\* FFPE indicates formalin-fixed, paraffin-embedded; GIST, gastrointestinal stromal tumor; MFH, malignant fibrous histiocytoma.

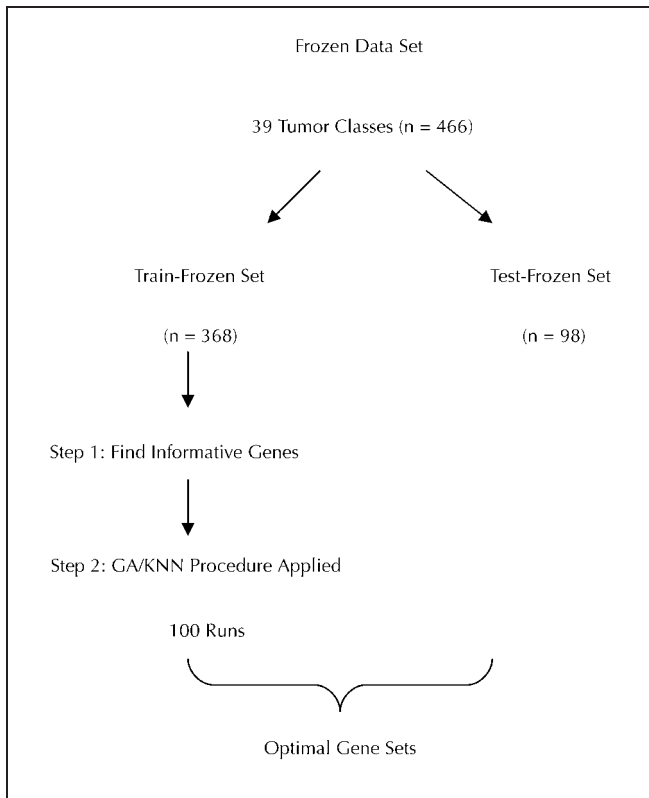
crease in predictability with increasing degree of dedifferentiation ( $P = .02$ ). Note that the 71% accuracy for poorly differentiated or undifferentiated tumors is much higher than the 30% reported in a previous study.<sup>14</sup>

We then tested whether prediction accuracies differ between primary and metastatic tumors. From the leave-one-out cross-validation results, the accuracies for primary and metastatic tumors were the same, at 84%, indicating no loss in performance for metastatic tumors. This was in agreement with previously reported findings.<sup>10,16</sup>

Finally, we selected 30 genes per tumor class using the entire set of frozen samples, resulting in 979 distinct genes. A classifier was then built using these genes and the entire training set to predict the FFPE samples ( $n = 112$ ), yielding an overall accuracy of 82% (Table 2). Therefore, the classifier built using frozen samples demonstrated comparable performance for FFPE samples.

### Gene Selection by Genetic Algorithm

Our 39-class KNN classifier using 979 highly informative genes demonstrated high accuracy in classification of



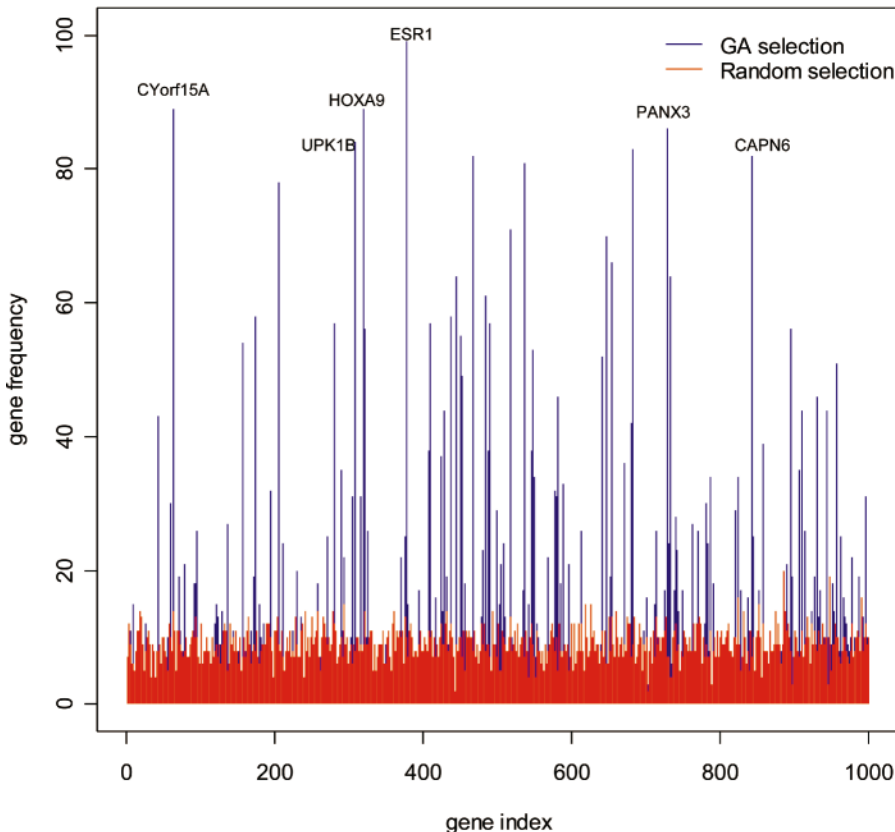
**Figure 2.** Schematic of gene selection procedure. GA/KNN indicates genetic algorithm/K-nearest neighbor.

both frozen and FFPE samples. The current gold standard for quantitative gene expression analysis in the clinic is RT-PCR. However, RT-PCR has limited throughput in measuring hundreds of genes. Therefore, we sought to develop a gene panel of less than 96 genes that can be conveniently manipulated on a single 96-well plate.

For a 39-class prediction task, the information content of each prediction is  $\log_2(39) = 5.29$  bits. Therefore, a theoretic minimum of 6 independent "genes," each carrying 1 bit of information, would be sufficient for such a task. Although not biological, this exercise provided one key but counterintuitive insight into the property of such "informative" genes: a maximally informative gene should not be expressed uniquely in 1 tumor class, but instead should be expressed in half of the tumor classes, and the combination of 6 such different genes would allow unique assignment of a sample's tumor class. Therefore, searching for combinations of genes approximating this property would be expected to result in compact gene sets. Recently, GA-based search methods have been applied to gene expression datasets and have produced such gene sets for multiclass prediction.<sup>24,25</sup>

We adopted the search strategy of Ooi and Tan<sup>22</sup> to select candidate gene sets for developing our PCR gene panel. We conducted our gene searches using the frozen tumor data set in 2 steps (Figure 2). To minimize overfitting, we split the frozen sample set into a training set (n = 368, 80%) and a test set (n = 98, 20%), keeping the same proportions of tumor-type representation in each set. These 2 partitions are referred to as train-frozen and test-frozen, respectively, to distinguish them from the fully blinded

### Frequency Plot: GA Selection vs. Random Selection



**Figure 3.** Gene selection frequency via genetic algorithm/K-nearest neighbor.

**Table 3. List of Selected 87 Genes**

| Accession | Gene Symbol | Description*  |
|-----------|-------------|---|
| AA456140  | PANX3       | Pannexin 3  |
| AA745593  | BATF        | Basic leucine zipper transcription factor, ATF-like                                   |
| AA765597  | SPRED2      | Sprouty-related, EVH1 domain containing 2   |
| AA782845  | SLC35F3     | Solute carrier family 35, member F3   |
| AA865917  |             | Hypothetical LOC389142  |
| AA946776  | FGF9        | Fibroblast growth factor 9 (glia-activating factor)                                   |
| AA993639  | FLJ10748    | Hypothetical protein FLJ10748   |
| AB038160  | TMPRSS3     | Transmembrane protease, serine 3  |
| AF104032  | SLC7A5      | Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5        |
| AF133587  | RTDR1       | Rhabdoid tumor deletion region gene 1   |
| AF301598  | EMX2        | Empty spiracles homolog 2 ( <i>Drosophila</i> )                                       |
| AF332224  | CYorf15A    | Chromosome Y open reading frame 15A   |
| AI041545  | KDELR2      | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2             |
| AI147926  | CSF2RB      | Colony-stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)     |
| AI309080  | KCNJ11      | Potassium inwardly rectifying channel, subfamily J, member 11                         |
| AI341378  | CPEB2       | Cytoplasmic polyadenylation element binding protein 2                                 |
| AI457360  | ERN2        | Endoplasmic reticulum to nucleus signalling 2   |
| AI620495  | MEIS1       | Meis1, myeloid ecotropic viral integration site 1 homolog (mouse)                     |
| AI632869  | UPK1B       | Uroplakin 1B  |
| AI683181  | PRDM6       | PR domain containing 6  |
| AI685931  | KIBRA       | KIBRA protein   |
| AI802118  | SLC6A13     | Solute carrier family 6 (neurotransmitter transporter, GABA), member 13               |
| AI804745  |             |   |
| AI952953  |             |   |
| AI985118  | C14orf105   | Chromosome 14 open reading frame 105  |
| AJ000388  | CAPN6       | Calpain 6   |
| AK025181  | LOC91464    | RAX-like homeobox   |
| AK027147  | TITF1       | Hypothetical protein LOC253970  |
| AK054605  | FLJ11539    | Hypothetical protein FLJ11539   |
| AL023657  | SH2D1A      | SH2 domain protein 1A, Duncan disease (lymphoproliferative syndrome)                  |
| AL039118  | FOXG1B      | Forkhead box G1A  |
| AL110274  |             |   |
| AL157475  | C8orf13     | Chromosome 8 open reading frame 13  |
| AW118445  | CELSR2      | Cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, <i>Drosophila</i> ) |
| AW194680  | HOXD11      | Homeobox D11  |
| AW291189  |             | Hypothetical LOC388416  |
| AW298545  | KIAA1904    | KIAA1904 protein  |
| AW445220  | LY6K        | Lymphocyte antigen 6 complex, locus K   |
| AW473119  | ESR1        | Estrogen receptor 1   |
| AY033998  | ELAVL4      | ELAV (embryonic lethal, abnormal vision, <i>Drosophila</i> )-like 4 (Hu antigen D)    |
| BC000045  | VGLL1       | Vestigial like 1 ( <i>Drosophila</i> )  |
| BC001293  | HOXC10      | Homeobox C10  |
| BC001504  | PYCR1       | Pyroline-5-carboxylate reductase 1  |
| BC001639  | SLC43A1     | Solute carrier family 43, member 1  |
| BC002551  | CDC43       | Cell division cycle associated 3  |
| BC004331  | HSDL2       | Hydroxysteroid dehydrogenase like 2   |
| BC004453  | HTR3A       | 5-hydroxytryptamine (serotonin) receptor 3A   |
| BC005364  | C10orf59    | Chromosome 10 open reading frame 59   |
| BC006537  | HOXA9       | Homeobox A9   |
| BC006881  | PPARG       | Peroxisome proliferative activated receptor, gamma                                    |
| BC006819  | S100P       | S100 calcium binding protein P  |
| BC008764  | KIF2C       | Kinesin family member 2C  |
| BC008765  | SDC1        | Syndecan 1  |
| BC009084  | SELENBP1    | Selenium binding protein 1  |
| BC009237  | TSHR        | Thyroid-stimulating hormone receptor  |
| BC010626  | KIF12       | Kinesin family member 12  |
| BC011949  | CA2         | Carbonic anhydrase II   |
| BC012926  | EPS8L3      | EPS8-like 3   |
| BC013117  | RG517       | Regulator of G-protein signalling 17  |
| BC015754  | CADPS       | Ca <sup>2+</sup> -dependent secretion activator                                       |
| BC017586  | MGC26610    | Calcyphosine-like   |
| BE552004  |             | CDNA FLJ44317 fis, clone TRACH3000586   |
| BE962007  | COX11       | COX11 homolog, cytochrome c oxidase assembly protein (yeast)                          |
| BF224381  |             | Hypothetical LOC400951  |
| BF437393  |             |   |
| BF446419  | PCANAP6     | Prostate cancer-associated protein 6  |
| BF592799  | PRKCQ       | Protein kinase C, theta   |
| BI493248  | IBSP        | Integrin-binding sialoprotein (bone sialoprotein, bone sialoprotein II)               |
| H05388    | ZNF365      | Hypothetical protein LOC283045  |
| H07885    |             | Transcribed locus   |
| H09748    | BCL11B      | B-cell CLL/lymphoma 11B (zinc finger protein)   |
| M95585    | HLF         | Hepatic leukemia factor   |

Table 3. Continued

| Accession | Gene Symbol    | Description*   |
|-----------|----------------|--|
| N64339    | <i>GJB6</i>    | Gap junction protein, beta 6 (connexin 30)   |
| NML000065 | <i>C6</i>      | Complement component 6   |
| NML001337 | <i>CX3CR1</i>  | Chemokine (C-X3-C motif) receptor 1  |
| NML003914 | <i>CCNA1</i>   | Cyclin A1  |
| NML004062 | <i>CDH16</i>   | Cadherin 16, KSP-cadherin  |
| NML004063 | <i>CDH17</i>   | Cadherin 17, LI cadherin (liver-intestine)   |
| NML004496 | <i>FOXA1</i>   | Forkhead box A1  |
| NML006115 | <i>PRAME</i>   | Preferentially expressed antigen in melanoma   |
| NML019894 | <i>TMPRSS4</i> | Transmembrane protease, serine 4   |
| NML033229 | <i>TRIM15</i>  | Tripartite motif-containing 15   |
| R15881    | <i>CHRM3</i>   | Cholinergic receptor, muscarinic 3   |
| R45389    |                | CDNA clone IMAGE:4797120   |
| R61469    |                | Transcribed locus, moderately similar to NP_775622.1 hypothetical protein LOC270028<br>[ <i>Mus musculus</i> ] |
| X69699    | <i>PAX8</i>    | Paired box gene 8  |
| X96757    | <i>MAP2K6</i>  | Mitogen-activated protein kinase kinase 6  |

\* ATF indicates ascites tumor fluid; EGF, epidermal growth factor; and CLL, chronic lymphatic leukemia.

FFPE dataset. Because the most informative genes are not necessarily those that are uniquely expressed in 1 tumor class, we sought for genes with expression patterns of varying degrees of tumor specificity. To do this, we constructed an approximate tumor taxonomy tree in which tumors are organized in a hierarchical manner covering both broad tumor categories (e.g., epithelial vs nonepithelial) and specific tumor subtypes. Using the train-frozen dataset, we selected 30 genes at each tree node using a combination of linear models and receiver operating characteristic analysis (see "Materials and Methods"), resulting in a total of 1001 distinct genes. In the second step, using these genes as input, we applied the GA/KNN procedure. We configured our GA/KNN algorithm to find gene sets between 61 and 80 genes in size and we conducted 100 independent runs. From each GA run, the best gene set was chosen by evaluating both leave-out-one cross-validation accuracy on train-frozen and the predictive accuracy on the test-frozen dataset.

Because the FFPE dataset was never used to select these gene sets, it allowed an unbiased estimate of classification performance. The 100 "optimized" gene sets had an average prediction accuracy of 80% on the FFPE dataset, indicating that there existed many gene sets capable of accurately discriminating 39 tumor classes. As a comparison, we randomly picked 1000 sets of genes between 61 and 80 in size from the 22k array and used them to predict the FFPE dataset. The average accuracy from these random sets was 49%, indicating that the GA-derived gene sets were indeed significantly optimized and generalizable to independent samples. The best overall set, consisting of 74 genes, had leave-out-one cross-validation accuracy of 91% on train-frozen, and prediction accuracies of 87% on test-frozen and 86% on the fully blinded FFPE test set. Of note, the 74-gene classifier performed better than the 979-gene classifier obtained earlier, demonstrating that it is possible to derive highly compact gene sets using GA without compromising performance.

Among the 100 GA-derived gene sets, certain genes appeared with high frequency. For example, *ESR1* was selected 99 times, 9 genes were selected more than 80 times, and 90 genes were selected more than 20 times, much more frequent than the 8 times expected by chance alone (Figure 3). This behavior of high reproducibility of gene selection by GA has been shown before, and it has been

demonstrated that genes most frequently selected by GA/KNN are most informative for classification.<sup>25</sup>

Taken together, between the 74-gene top-performing set and the list of 90 frequently selected genes, 38 genes were shared, resulting in 126 distinct genes as candidates for RT-PCR assay development.

#### Development of a 92-Gene RT-PCR Panel

TaqMan PCR assays were designed for the 126 GA/KNN-derived genes and performed on 400 frozen samples, a subset of samples in the microarray database. Requiring (1) high correlations ( $R^2 > 0.5$ ) between PCR and microarray data and (2) amplicons located within the last 500 bases of the transcript or within 150 bases from the original microarray probe resulted in 87 genes (Table 3). Among the final 87 genes, 60 were present in the 74-gene set.

As expected, all of the 87 genes were expressed in multiple tumor types. Approximately 80% of these genes have functional annotation, including several well-characterized tumor markers (*ESR1*, *TITF1*, *S100P*, *PRAME*, *PCANAP6*, and *UPK1B*). Gene ontology analysis of the 87-gene set indicated 2 significantly overrepresented groups.<sup>23</sup> Of the 63 genes with gene ontology mapping information,<sup>26</sup> 12 are DNA-binding transcription factors ( $P = .009$  for overrepresentation), and 13 are localized to plasma membrane ( $P = .03$ ). These transcription factors include many that are important in the development of cell and tissue types (*HOXA9*, *HOXC10*, *HOXD11*, *EMX2*, *MEIS1*, *FOXA1*, *FOXP1B*, *PAX8*, and *ESR1*). The plasma membrane protein group includes receptors for neurotransmitters (*HTR3A*, *CHRM3*) and cytokines (*CSF2RB* and *CXCR1*), ion channels (*KCNJ11*, *SLC6A13*, *SLC7A5* and *SLC43A1*), and cell-cell junction components (*PANX3*, *GJB6*, *CDH16*, and *CDH17*). Perhaps, the combinatorial expression pattern of these transcription factors and cell-surface proteins offers the most telltale features about the developmental lineage and tissue function of a cancer's origin, which make them excellent markers for tumor classification.

In addition, 5 reference genes with relatively invariant expression across the broad spectrum of tissue and disease types in the microarray database were selected for input normalization. Taken together, we developed a 92-gene RT-PCR assay that was optimized to classify as many as 39 tumor classes.

**Table 4. Classification Performance of 92-Gene Real-Time Polymerase Chain Reaction Assay\***

| Tumor Class                  | Frozen |          | FFPE |          |
|------------------------------|--------|----------|------|----------|
|                              | n      | Accuracy | n    | Accuracy |
| Adrenal                      | 8      | 1.00     | 1    | 1.00     |
| Brain                        | 16     | 1.00     | 3    | 1.00     |
| Breast                       | 41     | 1.00     | 1    | 1.00     |
| Carcinoid-intestine          | 7      | 0.57     | 2    | 1.00     |
| Cervix-adeno                 | 8      | 0.50     | 2    | 0.50     |
| Cervix-squamous              | 13     | 0.69     | 3    | 0.67     |
| Endometrium                  | 12     | 0.75     | 3    | 0.67     |
| Gallbladder                  | 7      | 0.00     | 0    | —        |
| Germ-cell                    | 32     | 0.78     | 9    | 0.78     |
| GIST                         | 10     | 0.90     | 3    | 1.00     |
| Kidney                       | 10     | 0.90     | 4    | 1.00     |
| Leiomyosarcoma               | 13     | 0.92     | 3    | 0.33     |
| Liver                        | 13     | 1.00     | 2    | 1.00     |
| Lung-adeno-large cell        | 16     | 0.63     | 3    | 0.00     |
| Lung-small                   | 9      | 0.89     | 5    | 0.40     |
| Lung-squamous                | 10     | 0.90     | 3    | 1.00     |
| Lymphoma                     | 23     | 1.00     | 10   | 1.00     |
| Meningioma                   | 8      | 1.00     | 3    | 1.00     |
| Mesothelioma                 | 10     | 1.00     | 5    | 0.80     |
| Osteosarcoma                 | 7      | 0.86     | 2    | 1.00     |
| Ovary                        | 27     | 1.00     | 5    | 1.00     |
| Pancreas                     | 24     | 0.96     | 3    | 1.00     |
| Prostate                     | 17     | 0.94     | 7    | 1.00     |
| Skin-basal cell              | 4      | 1.00     | 4    | 0.75     |
| Skin-melanoma                | 11     | 0.82     | 4    | 0.75     |
| Skin-squamous                | 8      | 1.00     | 3    | 1.00     |
| Small and large bowel        | 35     | 0.97     | 6    | 0.83     |
| Soft-tissue                  | 27     | 0.89     | 8    | 0.88     |
| Stomach-adeno                | 10     | 0.30     | 3    | 0.00     |
| Thyroid-follicular-papillary | 12     | 1.00     | 3    | 1.00     |
| Thyroid-medullary            | 8      | 0.88     | 0    | —        |
| Urinary bladder              | 25     | 0.88     | 6    | 1.00     |
| Overall                      | 481    | 0.88     | 119  | 0.82     |

\* FFPE indicates formalin-fixed, paraffin-embedded; GIST, gastrointestinal stromal tumor.

### Performance of the 92-Gene RT-PCR Panel

After establishing the 92-gene RT-PCR assay, we re-created the tumor classification database by performing the 92-gene assay on a total of 481 frozen tumor samples, 452 (94%) of which were part of the original microarray database. Because of low sample sizes in some of the 39 tumor classes, we consolidated the germ cell tumors (ovary and testis), lymphomas (B cell, T cell, and Hodgkin), ovarian cancer (clear cell and serous subtypes), and soft tissue sarcoma (liposarcoma, malignant fibrous histiocytoma, and synovial) categories, resulting in 32 tumor classes.

The classification performance of the PCR-based 92-gene database was first assessed by leave-out-one cross-validation. Of the 32 tumor classes, 24 of the 32 tumor types were predicted with more than 80% sensitivity, and the overall accuracy was 88% (Table 4). This high level of performance was similar to that from the original microarray database (result not shown), indicating that our attempt to translate the microarray database to the PCR-based platform was successful. However, because the 87-gene set was derived from the microarray database containing mostly the same samples, the leave-out-one cross-validation accuracy estimates on the frozen samples may be biased.

To independently test the performance of the 92-gene PCR assay, we used the frozen tumor PCR database to classify 119 independent FFPE tumor samples. Again, 21

tumor classes showed more than 80% sensitivity (Table 4). Because of the limited sample size within each tumor class in the test set, there are large confidence intervals in accuracy per tumor class. However, the PCR-based classifier had an overall accuracy of 82%, with a 95% confidence interval between 74% and 89%, far exceeding that by random chance alone (~4%). Some of the misclassifications were made between related tumor types or subtypes. For example, the 3 stomach adenocarcinomas were misclassified as digestive system cancers (1 as pancreatic and 2 as bowel cancers), and 1 large cell lung carcinoma was misclassified as the lung squamous subtype.

We then assigned a confidence level (high, medium, low, or unclassifiable) to each prediction result, accounting for both the degree of consensus voting and the relative representation of tumor classes in the training database (see "Materials and Methods"). Of the 119 predictions, 86 (72%), 8 (7%), and 17 (14%) were assigned with high, medium, and low confidence, respectively, and 8 (6.7%) were unclassifiable. The overall accuracies of the high, medium, low confidence groups were 94%, 75% and 59%, respectively. Excluding the 8 unclassifiable cases, the adjusted overall accuracy of classification was 87%.

### COMMENT

Multiple studies have demonstrated that microarray-based gene expression profiling enables accurate tumor classification.<sup>10-17</sup> In this study, we also demonstrated that a microarray-based classifier enabled accurate classification for as many as 39 tumor classes. Importantly, our microarray-based classifier was equally accurate in identifying tumor types based on biopsies from either the primary site or the site of metastasis, indicating that gene expression patterns of the tumor cells are sufficiently robust in the presence of heterologous tissue elements. Furthermore, we demonstrated that the expression-based tumor classification was only modestly affected by loss of cell differentiation. This result suggests that an expression-based classification tool would be particularly useful for poorly or undifferentiated tumors, which often confound even the most experienced pathologist.<sup>5,27</sup>

Although microarray-based classifiers are powerful, adopting them for clinical use is problematic for at least 3 reasons. First, the breadths of previously published microarray databases are limited. The most comprehensive database prior to this study spanned 21 tumor types, but it did not include liver and germ cell tumors, for example.<sup>10</sup> Because subsets of germ cell tumors are highly treatable, it is important to correctly identify these tumors.<sup>28</sup> Second, most of these studies demonstrated the need to use hundreds to thousands of discriminating genes in the classifiers.<sup>10,14</sup> The common approach has been to rank genes based on their tumor class-specific differential expression and then select the highest ranking genes to build the classifiers. This forward-search approach generates gene lists with high "redundancy"; that is, many genes provide the same information to the classifier because of high levels of gene-gene correlations. On the other hand, using a support vector machine-based backward-selection scheme, it was shown that all probe elements (>16 000) on the microarrays were required to yield the best classifier.<sup>14</sup>

Clearly, classifiers using these gene lists are beyond the scope of current methods employed in clinical practice. To address this technological issue, recently, Tothill et al<sup>17</sup> selected a 79-gene panel to design a RT-PCR assay. However,

to reduce the gene number to a manageable size, the authors also reduced the number of tumor types for the assay from 14 to 5, which would limit its usefulness in the real world. In this study, we took advantage of proven feature selection strategies to derive a set of 87 genes amenable to routine RT-PCR. Finally, almost all published studies required frozen tumor biopsies for analysis, which would require changing the current practice of formalin fixation, a serious impediment to clinical adoption. The present work demonstrates that translating from microarrays to RT-PCR is possible without reducing the scope of the diagnostic test or compromising performance.

To conclude, we have developed a 92-gene RT-PCR panel that can accurately classify a broad spectrum of tumor types. We envision 2 applications for such a diagnostic test. First, current investigative diagnostic procedures for CUP patients, including blood and urine analysis, radiographs and computed tomographic scans, and immunohistochemistry is both costly and time-consuming.<sup>29</sup> Incorporating an expression-based test such as the 92-gene panel assay would both greatly expedite the diagnostic process and result in significant reduction in the diagnosis of CUP. Second, even the diagnosis of metastatic lesions in general would benefit from such a molecular test as an adjunct to either confirm a diagnosis or suggest additional testing to reach a final diagnosis. Finally, we anticipate that the performance of the 92-gene test will improve as the database grows and the classifier algorithm "learns" from it. Increased precision in tumor classification would be expected to improve the clinical management and outcome of patients with metastatic cancer.

We thank Yien Tran, Diem Tran, Philip McQuarry, Ana Tassin, Paul Amon, Li Ding, Walter Lee, and Eden Estepa-Sabal for excellent technical and operational assistance.

#### References

1. Burton EC, Troxclair DA, Newman WP 3rd. Autopsy diagnoses of malignant neoplasms: how often are clinical diagnoses incorrect? *JAMA*. 1998;280:1245-1248.
2. Raab SS, Grzybicki DM, Zarbo RJ, Meier FA, Geyer SJ, Jensen C. Anatomic pathology databases and patient safety. *Arch Pathol Lab Med*. 2005;129:1246-1251.
3. Pavlidis N, Fizazi K. Cancer of unknown primary (CUP). *Crit Rev Oncol Hematol*. 2005;54:243-250.
4. Varadhachary GR, Abbruzzese JL, Lenzi R. Diagnostic strategies for unknown primary cancer. *Cancer*. 2004;100:1776-1785.
5. Sheahan K, O'Keane JC, Abramowitz A, et al. Metastatic adenocarcinoma of an unknown primary site: a comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status. *Am J Clin Pathol*. 1993;99:729-735.

6. 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.
7. DeYoung BR, Wick MR. Immunohistologic evaluation of metastatic carcinomas of unknown origin: an algorithmic approach. *Semin Diagn Pathol*. 2000;17:184-193.
8. Hainsworth JD, Greco FA. Treatment of patients with cancer of unknown primary site. *Important Adv Oncol*. 1991;173-190.
9. Raber MN, Abbruzzese JL, Frost P. Unknown primary tumors. *Curr Opin Oncol*. 1992;4:3-9.
10. Bloom G, Yang IV, Boulware D, et al. Multi-platform, multi-site, microarray-based human tumor classification. *Am J Pathol*. 2004;164:9-16.
11. Buckhaults P, Zhang Z, Chen YC, et al. Identifying tumor origin using a gene expression-based classification map. *Cancer Res*. 2003;63:4144-4149.
12. 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.
13. Giordano TJ, Shedden KA, Schwartz DR, et al. Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. *Am J Pathol*. 2001;159:1231-1238.
14. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*. 2001;98:15149-15154.
15. Shedden KA, Taylor JM, Giordano TJ, et al. Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *Am J Pathol*. 2003;163:1985-1995.
16. Su AI, Welsh JB, Sapinoso LM, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*. 2001;61:7388-7393.
17. 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*. 2005;65:4031-4040.
18. Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 2002;30:e15.
19. Kuruvilla FG, Park PJ, Schreiber SL. Vector algebra in the analysis of genome-wide expression data. *Genome Biol*. 2002;3:research 0011.1-0011.11.
20. Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods*. 2003;31:265-273.
21. Pepe MS, Longton G, Anderson GL, Schummer M. Selecting differentially expressed genes from microarray experiments. *Biometrics*. 2003;59:133-142.
22. Ooi CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*. 2003;19:37-44.
23. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*. 2004;20:1464-1465.
24. Tan Y, Shi L, Tong W, Wang C. Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Res*. 2005;33:56-65.
25. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*. 2001;17:1131-1142.
26. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25:25-29.
27. Kern WH, Abbott M. The determination of unknown primary sites based upon the histologic appearance of metastases. *Surg Gynecol Obstet*. 1980;151:73-76.
28. Dowell JE. Cancer from an unknown primary site. *Am J Med Sci*. 2003;326:35-46.
29. Abbruzzese JL, Abbruzzese MC, Lenzi R, Hess KR, Raber MN. Analysis of a diagnostic strategy for patients with suspected tumors of unknown origin. *J Clin Oncol*. 1995;13:2094-2103.