[CANCER RESEARCH 63, 4144–4149, July 15, 2003]

Identifying Tumor Origin Using a Gene Expression-based Classification Map[1]

Phillip Buckhaults, Zhen Zhang, Yu-Chi Chen, Tian-Li Wang, Brad St. Croix, Saurabh Saha, Alberto Bardelli, Patrice J. Morin, Kornelia Polyak, Ralph H. Hruban, Victor E. Velculescu, and Ie-Ming Shih[2]

The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins [P. B., T-L. W., S. S., A. B., V. E. V.] and The Department of Pathology [Z. Z., Y-C. C., R. H. H., I-M. S.], Johns Hopkins Medical Institutions, Baltimore, Maryland 21231; The National Cancer Institute, Frederick, Maryland [B. St. C.]; National Institute on Aging, Baltimore, Maryland [P. J. M.]; and Dana-Farber Cancer Institute, Boston, Massachusetts [K. P.]

ABSTRACT

Identifying the primary site in cases of metastatic carcinoma of unknown origin has profound clinical importance in managing cancer patients. Although transcriptional profiling promises molecular solutions to this clinical challenge, simpler and more reliable methods for this purpose are needed. A training set of 11 serial analysis of gene expression (SAGE) libraries was analyzed using a combination of supervised and unsupervised computational methods to select a small group of candidate genes with maximal power to discriminate carcinomas of different tissue origins. Quantitative real-time PCR was used to measure their expression levels in an independent validation set of 62 samples of ovarian, breast, colon, and pancreatic adenocarcinomas and normal ovarian surface epithelial controls. The diagnostic power of this set of genes was evaluated using unsupervised cluster analysis methods. From the training set of 21,321 unique SAGE transcript tags derived from 11 libraries, five genes were identified with expression patterns that distinguished four types of adenocarcinomas. Quantitative real-time PCR expression data obtained from the validation set clustered tumor samples in an unsupervised manner, generating a self-organized map with distinctive tumor site-specific domains. Eighty-one percent (50 of 62) of the carcinomas were correctly allocated in their corresponding diagnostic regions. Metastases clustered tightly with their corresponding primary tumors. A classification map diagnostic of tumor types was generated based on expression patterns of five genes selected from the SAGE database. This expression map analysis may provide a reliable and practical approach to determine tumor type in cases of metastatic carcinoma of clinically unknown origin.

INTRODUCTION

In the United States, ~51,000 patients (or 4% of all new cancer cases) present annually with metastases arising from occult primary carcinomas of unknown origin (1). Adenocarcinomas represent the most common metastatic tumors of unknown primary site. Although these patients often present at a late stage, outcome can be positively affected by accurate diagnoses followed by appropriate therapeutic regimens specific to different types of adenocarcinoma (2). The lack of unique microscopic appearance of the different types of adenocarcinomas challenges morphological diagnosis of adenocarcinomas of unknown origin. The application of tumor-specific serum markers in identifying cancer type is theoretically feasible, but such markers are generally not available at present (3, 4). Transcription profiling of thousands of genes using DNA microarrays has successfully classified tumors according to the site of origin (5, 6), but the need for relatively large amounts of RNA and lack of a unified standard for both data collection and analysis make it difficult to deploy the current platform in a clinical setting. SAGE,[3] on the other hand, measures absolute expression levels through a tag counting approach, allowing data to be obtained and compared from different samples analyzed in isolation. However, the throughput of this approach is low, making it inappropriate for routine clinical applications. Quantitative real-time PCR is a reliable method for assessing gene expression levels from relatively small amounts of tissue (7). Although this approach has recently been successfully applied to the molecular classification of breast tumors into prognostic subgroups based on the analysis of 2,400 genes (8), the measurement of such a large number of randomly selected genes by PCR is clinically impractical.

To overcome the above limitations, we developed a rational strategy that uses a combination of supervised and unsupervised computational methods to analyze SAGE-derived gene expression data and identify a minimum number of genes with maximum ability to separate adenocarcinomas according to their tissue origins. We then used quantitative real-time PCR to query the expression levels of these genes and used novel unsupervised computational analyses to prospectively evaluate their ability to correctly categorize adenocarcinomas of the ovary, breast, colon, and pancreas.

MATERIALS AND METHODS

Tissue Samples. A total of 62 surgically removed samples were retrieved from the tumor bank, Department of Pathology, Johns Hopkins Hospital. The acquisition of human tissue material was approved by the local Institutional Review Board. The panel of carcinomas included 20 ovarian serous carcinomas (11 primaries and 9 metastases), 11 breast carcinomas (10 primaries and 1 metastasis), 6 pancreatic carcinomas (all primaries), and 20 colorectal carcinomas (9 primaries and 11 metastases). Total RNA was purified, and cDNA was synthesized using standard protocols. Briefly, frozen tissues were minced and placed in the TRIzol reagent (Invitrogen, Carlsbad, CA). Total RNA was isolated, and contaminating genomic DNA was removed using the DNA-free kit (Ambion, Austin, TX). cDNA was prepared using oligo dT primers and diluted for PCR. H&E-stained sections were prepared from a portion of the frozen tumors and reviewed by a surgical pathologist (I-M. S.) to confirm the diagnosis and assess the percentage of viable tumor cells in the tissue sections. Tissue samples with >60% stromal contamination or <50% viable tumor cells were not included in analysis. Five primary cultures of ovarian surface epithelium were obtained from normal ovaries that were surgically removed, along with uteri, because of benign uterine disorders.

Analysis Software. UMSA is a modified Support Vector Machine learning algorithm that allows the incorporation of estimated data distribution into the derivation of an optimal soft-margin classifier (9). ProPeak (3Z Informatics, Charleston, SC) is a Java-based software package that implements the UMSA algorithm. Unsupervised clustering with SOM was performed using SOM toolbox 2.0 for Matlab (10).[4] Additional Matlab code was developed by the authors to display the estimated two-dimensional probability distribution patterns of each of the tumor sites and define the diagnostic regions within a derived SOM. PCR primers for the selected genes were designed using Primer3.[5]

Received 12/26/02; accepted 5/7/03.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by Public Health Service Grants CA97527 (to I-M. S.) and P50-CA62924 (to R. H. H.) from the National Cancer Institute, NIH, Department of Health and Human Services, and the United States Department of Defense Research Grant OC010017 (to I-M. S.).

[2] To whom requests for reprints should be addressed, at The Department of Pathology, The Johns Hopkins University School of Medicine, 418 North Bond Street, B-315, Baltimore, MD 21231. E-mail: ishih@jhmi.edu.

[3] The abbreviations used are: SAGE, serial analysis of gene expression; UMSA, unified maximum separability analysis; SOM, self-organizing map; Ct, threshold cycle number; APP, amyloid precursor protein.

[4] Internet address: http://www.cis.hut.fi.

[5] Internet address: http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Tag Selection from the SAGE Database. SAGE libraries were sequenced, and the resulting gene expression data were collated and warehoused as part of the Cancer Genome Anatomy Project public database for gene expression (11). The hematopoietic cell library was reported previously (12). Tags and their expression levels were retrieved from the Internet (12–17).[6] Molecular information regarding the libraries analyzed, as well as the primary expression data, are available under the following ProbeSet ID numbers: Colon Cancer 1 (CoCa1) GSM756; Colon Cancer 2 (CoCa2) GSM755; Ovarian Cancer 1 (OvCa1) GSM735; Ovarian Cancer 2 (OvCa2) GSM736; Ovarian Cancer 3 (OvCa3) GSM737; Pancreatic Cancer 1 (PaCa1) GSM743; Pancreatic Cancer 2 (PaCa2) GSM744; Breast Cancer 1 (BrCa1) GSM670; Breast Cancer 1 Metastasis (BrCa1Met) GSM671; Breast Cancer 2 (BrCa2) GSM672; and Breast Cancer 2 Metastasis (BrCa2Met) GSM673. In the supervised selection process, the tags were first filtered to retain only those that were abundant (greater than eight tags/100,000 total tags) in tumors and absent in hematopoietic cells. Tag counts were subsequently analyzed by the UMSA algorithm in ProPeak to select several top-ranked tags for each of the four tumor sites according to their contribution in separating the specific carcinoma from the rest of the data. The resultant most informative tags were then subjected to unsupervised two-way hierarchical cluster analysis (centered correlation similarity metric with complete linkage clustering, no transformation or normalization; Ref. 18)[7] to evaluate the site specificity of the tumor-associated groups of tags. The tags were matched to the Unigene database, and well-characterized genes were selected for PCR primer design and further analysis.

Gene Expression. Quantitative real-time PCR was performed to determine gene expression levels in a panel of 20 ovarian carcinomas (11 primaries and 9 metastases), 11 breast carcinomas (10 primaries and 1 metastasis), 6 pancreatic carcinomas (all primaries), 20 colorectal carcinomas (9 primaries and 11 metastases), and five normal ovarian surface epithelial cultures using the protocol described previously (19). Primers were designed for representative well-characterized candidate genes to test the performance in quantitative real-time PCR, and those containing robust and specific PCR products without detectable primer dimers were selected for analysis. Approximately 16–100 ng of cDNA were included in the real-time PCR, which was performed using an iCycler, and Cts were obtained using the iCycler Optical system interface software (Bio-Rad Lab, Hercules, CA). Averages in the Ct of duplicate measurements were obtained. The results were expressed as the difference between the Ct of the diagnostic gene and Ct of a control gene (APP), for which expression is relatively constant among the SAGE libraries analyzed. In cases where no gene expression was observed, a cutoff Ct value of 45 cycles was used.

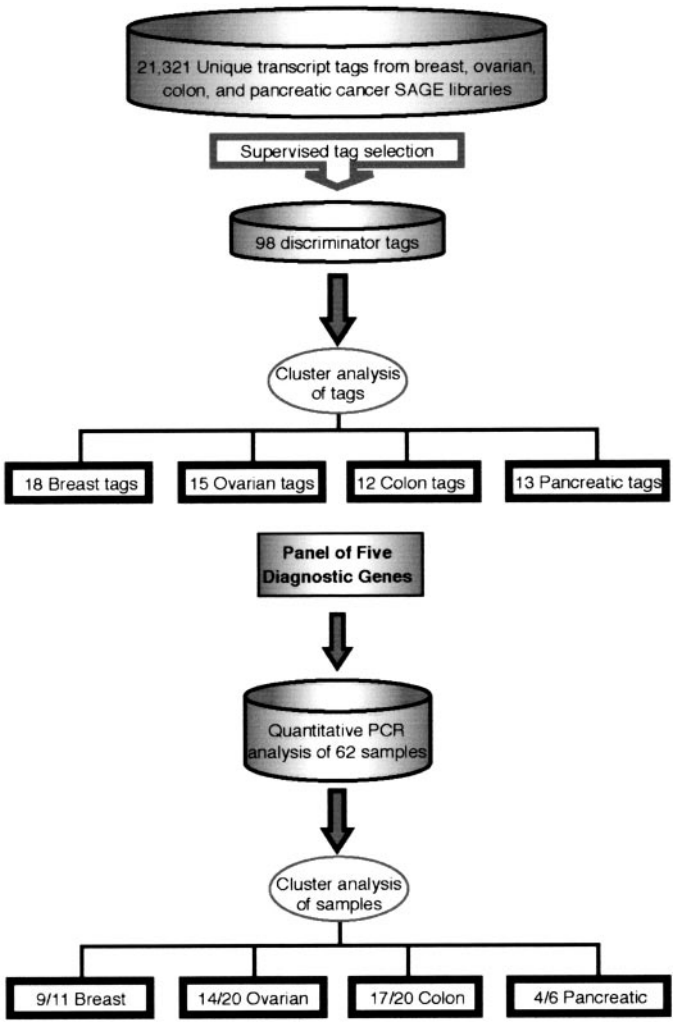

Fig. 1. Outline of data analysis. Tags were subjected to a combination of supervised and unsupervised selection algorithms. Expression levels of genes corresponding to informative transcript tags were measured by quantitative real-time PCR, and the expression data were used to identify tissue type in a test set of 62 clinical samples.

RESULTS

Identification of Diagnostic Genes. Analysis of SAGE libraries constructed from surgical samples of ovarian, breast, pancreatic, and colon adenocarcinomas identified a total of 21,321 unique SAGE tags that were observed two or more times (Fig. 1). To identify genes that were potential candidates for differential diagnosis of these tumor types, we selected tags that were highly expressed (eight or more occurrences/100,000 total tags) in one or more of the tumor samples. This resulted in the identification of 6,396 unique and highly expressed cancer-associated tags. To remove transcripts derived from hematopoietic infiltrates present in variable amounts in these tumors, we eliminated tags that were also present in a SAGE library constructed from tumor-derived, purified hematopoietic cells (12). As a result, a total of 3,747 unique, cancer-specific tags was obtained and subjected to further analysis.

UMSA (9) was used to identify a subset of genes that would provide the most discriminating differential gene expression information. The UMSA approach evaluates and ranks the contribution of each transcript tag toward the optimal separation of SAGE libraries derived from different tissues. Although we only analyzed four tumor sites, this UMSA approach can become valuable in analyzing SAGE libraries obtained from many tissue sites in the future extended study. This analysis resulted in a total of 98 unique transcript tags including 15 tags for ovarian, 12 tags for colon, 13 tags for pancreatic, and 18 tags for breast carcinomas that were calculated to be most informative for tumor site-specific classification. This was corroborated by a two-way hierarchical cluster analysis of both tumors and tags based on the expression levels of this reduced number of tags. This relatively small subset of transcript tags (<0.5% of all tags) was sufficient to correctly cluster the SAGE libraries into groups of similar tissue origins (Fig. 2, horizontal tree). Additionally, tags clustered into nodes of similar expression patterns and had clearly visible associations with the above carcinomas arising from different tissue types (Fig. 2, vertical tree). The tags were matched to the Unigene database,[8] and three well-characterized genes from each node were randomly selected to evaluate their performance in quantitative real-time PCR. Five of these genes showed robust and specific PCR products without detectable primer dimer formation and were further analyzed for their diagnostic potential (Table 1). The selection of these five genes from the candidate pool was arbitrary and driven by technical limitations (*i.e.,* capacity and success of real-time PCR). Most tags present in the pancreatic cluster did not match known genes and were not pursued

[6] Internet address: http://www.ncbi.nlm.nih.gov/SAGE/.
[7] Internet address: http://rana.lbl.gov/EisenSoftware.htm.
[8] Internet address: http://www.ncbi.nlm.nih.gov/UniGene/.

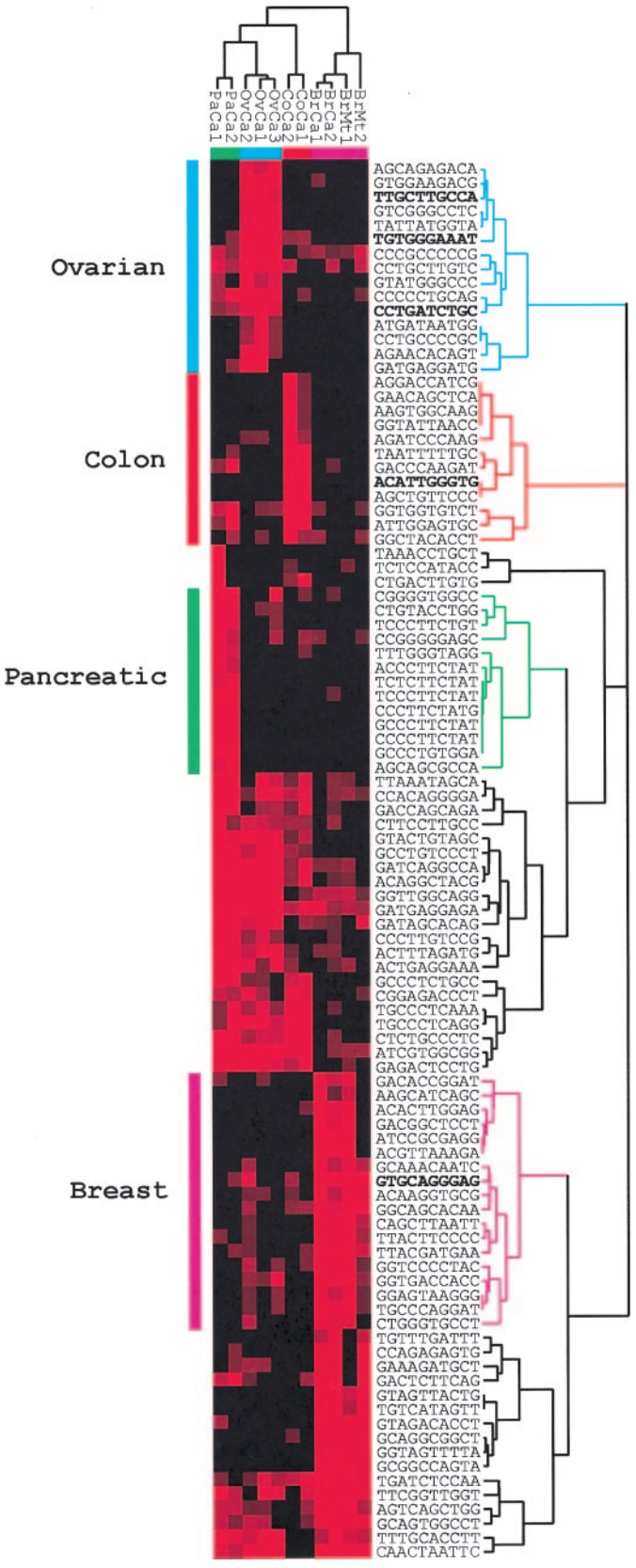

Fig. 2. Cluster analysis and identification of diagnostic tags. The most informative 98 tags (vertical tree) were subjected to two-way hierarchical cluster analysis with different SAGE libraries (horizontal tree) as the training set. Seventy-one cancer type-specific tags are identified and include 15 tags for ovarian, 12 tags for colon, 16 tags for pancreatic, and 28 tags for breast carcinomas. A total of five genes with their tag sequences bolded were selected for real-time PCR. *PaCa,* primary pancreatic carcinoma; *OvCa,* primary ovarian carcinoma; *BrCa,* primary breast carcinoma; *BrMt,* breast carcinoma metastases.

further. Among all of the genes analyzed in the SAGE libraries, APP (Unigene ID 177486) demonstrated a relatively constant expression level and was selected as the control for subsequent expression analyses.

Validation of Diagnostic Genes. To prospectively evaluate their diagnostic power, the expression levels of the five genes selected from the above training set were evaluated by quantitative real-time PCR in a separate set of 62 samples of breast, colon, ovarian, and pancreatic primary carcinomas and metastases and normal ovarian surface epithelium. Ct differences between each diagnostic gene and the control gene (APP) were calculated, and the expression data obtained from this validation set were subjected to unsupervised hierarchical cluster analysis. This analysis was able to correctly classify 44 of 62 samples according to their sites of origin (Fig. 3). A pancreatic cluster was not evident by this analysis. With a sensitivity of 71% (44 of 62), real-time PCR results clearly validated the diagnostic genes' ability to classify different types of adenocarcinomas

In an attempt to improve the resolving power of the expression data, and display results in terms of diagnostic probability, an unsupervised two-dimensional SOM analysis of the quantitative real-time PCR data was performed (10). As shown in Fig. 4, each of the five genes demonstrated a gradient in expression levels with the highest expression present in certain types of cancers. For practical application of this analysis, we generated a two-dimensional gene expression-based classification map according to different cancer types (Fig. 5). Most adenocarcinomas clustered within their well-defined map districts bounded by iso-probability curves. This type of analysis allowed for a rapid visual assessment of the probabilities of each cancer diagnosis. Based on our computational analysis, most of the probability maps demonstrated a single diagnostic domain with the exception of the breast carcinoma probability map, which exhibited two diagnostic domains. By integrating the spatial distribution patterns of these five individual estimated tissue type cancer probability maps, a combined classification map was established (Fig. 5, right bottom corner) with borders between the domains determined by maximizing the overall sensitivity for all diagnostic domains. This allowed categorization of the adenocarcinomas and normal samples into their specific domains with a sensitivity of 80.6% (50 of 62). Metastases clustered tightly with their corresponding primary tumors. In the ovarian cancer probability map, all ovarian cancer specimens were of serous types, the most common type of ovarian cancer, with the exception of two specimens, which were located outside the border of the ovarian cancer diagnostic domain. Histologically, both cases were retrospectively found to be unusual nonserous subtypes of ovarian cancer (20).

DISCUSSION

The results described above demonstrate that expression analysis of a small set of genes can reliably discriminate adenocarcinomas of different tissue origins. These findings suggest that gene expression-based probability maps composed of spatial domains computed by the expression patterns of a few informative genes can provide a molecular platform to predict adenocarcinoma of clinically unknown origin. It is noteworthy that the expression patterns of several genes selected by this computational process have been observed previously by other methods to be relatively tissue specific for certain types of carcinomas, *e.g.,* MUC16 has been recently identified as the gene encoding the CA125 antigen (21, 22), which has long been known as an ovarian cancer-associated marker and routinely used to monitor the response of therapy and detect tumor recurrence in ovarian cancer patients (20). Additionally, ceruloplasmin and SLPI have been reported among the most highly up-regulated markers in ovarian carcinoma (13). Similar to a previous report (23), our finding supports the utility of the SAGE

Table 1 *Panel of diagnostic genes*

Category	SAGE tag	Unigene no.	Gene	Transcript tag counts from cancer SAGE libraries[a]										
				CoCa1	CoCa2	OvCa1	OvCa2	OvCa3	BrCa1	BrCa1MET	BrCa2	BrCa2MET	PaCa1	PaCa2[b]
Colon	ACATTGGGTG	351719	FABP1	56	221	0	0	0	0	0	0	0	0	0
Ovarian	TTGCTTGCCA	296634	CP	0	0	23	41	12	0	0	0	0	0	0
Ovarian	CCTGATCTGC	98502	MUC16	0	0	28	14	12	0	0	0	0	3	5
Ovarian	TGTGGGAAAT	251754	SLPI	3	5	34	62	143	0	0	0	0	0	3
Breast	GTGCAGGGAG	79414	PDEF	5	0	0	12	0	61	63	58	44	0	3
Control	GTGAAACCCC	177486	APP	681	834	508	381	387	269	213	546	415	681	562

[a] Normalized to 100,000 total tags.
[b] CoCa, colon cancer; OvCa, ovarian cancer; BrCa, breast cancer; PaCa, pancreatic cancer; BrCaMET, breast cancer metastasis.

database and robustness of our computational methods in identifying candidate genes for validation. Among those tumor samples that failed to cluster into their diagnostic domains, two ovarian carcinomas were retrospectively found to be nonserous subtypes. This result suggests that our method could be used to define histological subtypes within a tumor type.

As compared with the global transcriptional profiling that analyzes thousands of genes, our reductionist approach, deploying a two-dimensional, expression-based classification map based on a few highly informative genes, has several advantages in determining the origin of a metastatic carcinoma of clinically unknown primary: (*a*) only a relatively small amount of tissue material is required. This feature is particularly attractive because minimally invasive techniques such as small biopsies using fine needle, endoscope, and laparoscope approaches are frequently used; (*b*) by reducing thousands of genes to a limited set of informative genes, the background noise introduced by the vast majority of irrelevant genes can be minimized, and the information gained from each sample is maximized. A similar approach has been shown effective in a specialized cDNA array with only hundreds of genes (24). Accordingly, sensitivity can be improved using this approach; (*c*) quantitative real-time PCR has become a reliable assay without the need for sophisticated instrumentation (7). In contrast to DNA array hybridization and SAGE, the assays can be performed in a shorter period of time (~5 h) at low cost; and (*d*) although bioinformatics tools have recently been applied to microarray data and shown utility in predicting both cancer

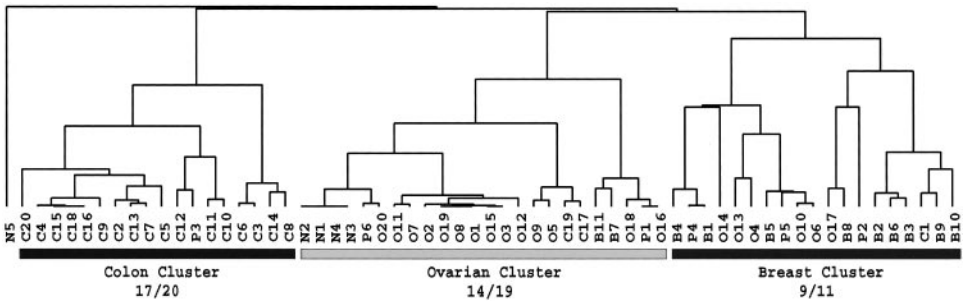

Fig. 3. Hierarchical cluster analysis of gene expression in a validation set of 62 carcinoma specimens. Quantitative real-time PCR data were evaluated using unsupervised hierarchical cluster analysis. Forty of 50 colon, ovarian, and breast carcinomas cluster to their site of origin. Four of five normal ovarian surface epithelial controls also clustered. No pancreatic cluster was observed. *P,* pancreatic carcinoma; *O,* ovarian carcinoma; *C,* colon carcinoma; *B,* breast carcinoma; *N,* normal ovarian surface epithelial controls.

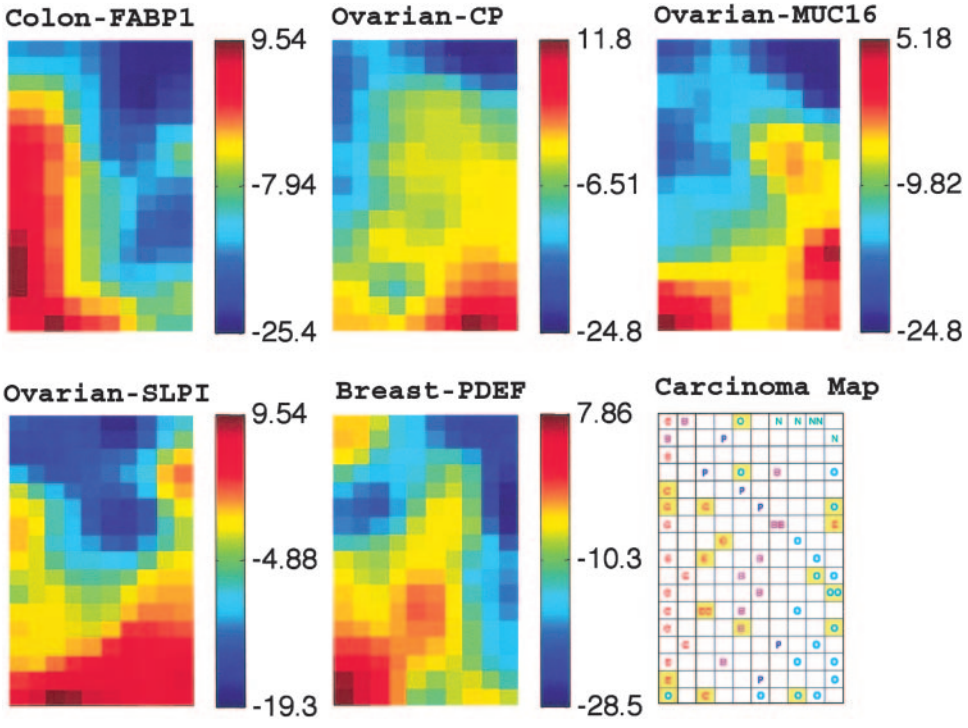

Fig. 4. Distribution of samples and gene expression levels using SOM analysis. Expression data from quantitative real-time PCR analyses are represented as pseudo-color gradients corresponding to the Ct differences between experimental and control genes (*bars* to the *right* of each *panel*). The placements of corresponding clinical samples by SOM are indicated in the *right bottom panel*. *O,* ovarian carcinoma; *B,* breast carcinoma; *P,* pancreatic carcinoma; *C,* colon carcinoma; *N,* normal ovarian surface epithelial controls. Metastases are noted by the *yellow boxes*.

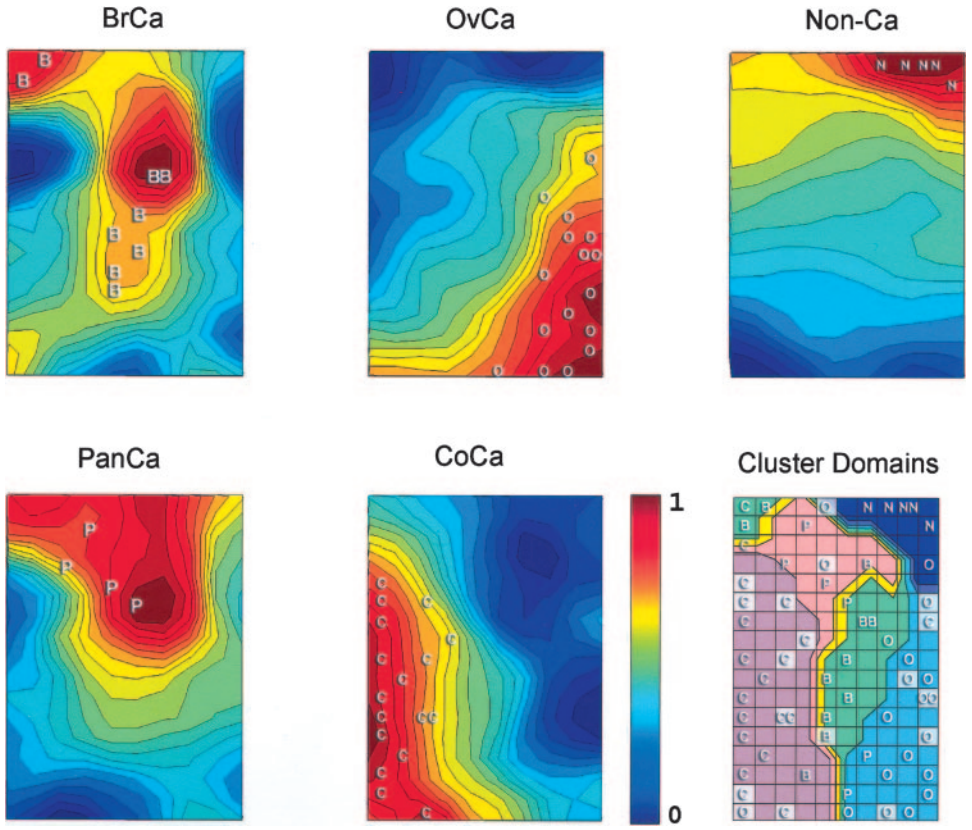

Fig. 5. Expression-based classification map for primary site determination. A probability surface shown as a pseudo-color map was estimated for each cancer type, and individual clinical samples were projected onto the map according to their positions by SOM. The highest probability (*1*) is shown in *red* and lowest (*0*) in *blue*. Iso-probability curves for all carcinoma types are outlined. The landscape of the integrated classification map from the five individual tissue-type probability maps is shown in the *right bottom panel*. The borders were generated along the contiguous saddle points, which represent equal probability between neighboring domains, and maximize the diagnostic sensitivity for all five tissue types. *P,* pancreatic carcinoma; *O,* ovarian carcinoma; *C,* colon carcinoma; *B,* breast carcinoma; *N,* normal ovarian surface epithelial controls. Metastases are noted in the *lighter shaded boxes*.

diagnosis (25) and outcome (26), our computational modeling of expression data through a surface probability map provides a quantitative appreciation of the probability of tissue origin in the tested sample. The results of this cancer classification map in combination with clinical profiles, including patients' demography, the prevalence-specific cancer types, and other clinical parameters, could be clinically useful in determining the cancer origin from unknown primary.

In conclusion, this report provides evidence that a gene expression-based probability map could prove to be a powerful tool aiding the diagnosis of metastatic carcinomas of unknown origin. However, before this approach becomes a practical cancer diagnostic tool, several issues need to be addressed. In this study, the sensitivity in differentiating various types of samples was 80.6%. Sensitivity might be improved in the following ways. Additional gene markers from the candidate list of SAGE tags (Fig. 2) can be tested to determine whether either can be increased over that obtained from the current selected genes. Our approach can also be used in conjunction with emerging protein or peptide markers that are relatively tissue specific, yet are of suboptimal sensitivity or specificity. Although this report only included four major cancer types to demonstrate the feasibility of this method, it would be important to extend our approach to a larger set of samples from a wide variety of common cancer types and generate a more comprehensive cancer diagnostic map. Finally, a more precise landscape of the probabilities and classification map will evolve as more samples are analyzed.

ACKNOWLEDGMENTS

We thank Drs. K. Kinzler and B. Vogelstein and the members of The Molecular Genetics laboratory, Johns Hopkins Oncology Center, for their critical comments.

REFERENCES

1. ACS Cancer Facts & Figures 2001: American Cancer Society, 2001.
2. Hillen, H. F. Unknown primary tumours. Postgrad. Med. J., *76:* 690–693, 2000.
3. Milovic, M., Popov, I., and Jelic, S. Tumor markers in metastatic disease from cancer of unknown primary origin. Med. Sci. Monit., *8:* MT25–MT30, 2002.
4. Diamandis, E. P., Fritsche, H. A., Lilja, H., Chan, D. W., and Schwartz, D. R. (eds.). Tumor Markers: Physiology, Pathobiology, Technology and Clinical Applications. Washington, DC: AACC Press, 2002.
5. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, *98:* 15149–15154, 2001.
6. Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. Molecular classification of multiple tumor types. Bioinformatics, *17:* S316–S322, 2001.
7. Freeman, W. M., Walker, S. J., and Vrana, K. E. Quantitative RT-PCR: pitfalls and potential. Biotechniques, *26:* 112–122, 115–124, 1999.
8. Iwao, K., Matoba, R., Ueno, N., Ando, A., Miyoshi, Y., Matsubara, K., Noguchi, S., and Kato, K. Molecular classification of primary breast tumors possessing distinct prognostic properties. Hum. Mol. Genet., *11:* 199–206, 2002.
9. Zhang, Z., Page, G., and Zhang, H. (eds.). Applying Classification Separability Analysis to Microarray Data. New York: Kluwer Academic Publishers, 2001.
10. Kohonen, T. (ed.). Self-organizing Maps, Ed. 1. New York: Springer-Verlag, 1995.
11. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. A public database for gene expression in human cancers. Cancer Res., *59:* 5403–5407, 1999.
12. St. Croix, B., Rago, C., Velculescu, V., Traverso, G., Romans, K. E., Montgomery, E., Lal, A., Riggins, G. J., Lengauer, C., Vogelstein, B., and Kinzler, K. W. Genes expressed in human tumor endothelium. Science, *289:* 1197–1202, 2000.
13. Hough, C. D., Sherman-Baust, C. A., Pizer, E. S., Montz, F. J., Im, D. D., Rosenshein, N. B., Cho, K. R., Riggins, G. J., and Morin, P. J. Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. Cancer Res., *60:* 6281–6287, 2000.
14. Hough, C. D., Cho, K. R., Zonderman, A. B., Schwartz, D. R., and Morin, P. J. Coordinately up-regulated genes in ovarian cancer. Cancer Res., *61:* 3869–3876, 2001.
15. Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res., *61:* 5697–5702, 2001.
16. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. Science, *276:* 1268–1272, 1997.

17. Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. SAGEmap: a public gene expression resource. Genome Res., *10:* 1051–1060, 2000.
18. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.
19. Morrison, T. B., Weis, J. J., and Wittwer, C. T. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques, *24:* 954–958, 960–962, 1998.
20. Shih, I-M., Sokoll, L. J., and Chan, D. W. Tumor markers in ovarian cancer. *In:* E. P. Diamandis, H. A. Fritsche, H. Lilja, D. W. Chan, and M. K. Schwartz (ed.), Tumor Markers Physiology, Pathobiology, Technology and Clinical Applications, pp. 239–252. Philadelphia: AACC Press, 2002.
21. Yin, B. W., and Lloyd, K. O. Molecular cloning of the CA125 ovarian cancer antigen: identification as a new mucin, MUC16. J. Biol. Chem., *276:* 27371–27375, 2001.
22. Yin, B. W., Dnistrian, A., and Lloyd, K. O. Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. Int. J. Cancer, *98:* 737–740, 2002.
23. Dennis, J. L., Vass, J. K., Wit, E. C., Keith, W. N., and Oien, K. A. Identification from public data of molecular markers of adenocarcinoma characteristic of the site of origin. Cancer Res., *62:* 5999–6005, 2002.
24. Sawiris, G. P., Sherman-Baust, C. A., Becker, K. G., Cheadle, C., Teichberg, D., and Morin, P. J. Development of a highly specialized cDNA array for the study and diagnosis of epithelial ovarian cancer. Cancer Res., *62:* 2923–2928, 2002.
25. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, *286:* 531–537, 1999.
26. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, *415:* 436–442, 2002.