[CANCER RESEARCH 62, 5999–6005, November 1, 2002]

Advances in Brief

Identification from Public Data of Molecular Markers of Adenocarcinoma Characteristic of the Site of Origin[1,2]

Jayne L. Dennis, J. Keith Vass, Ernst C. Wit, W. Nicol Keith, and Karin A. Oien[3]

Cancer Research UK Department of Medical Oncology, University of Glasgow, Cancer Research UK Beatson Laboratories, Glasgow G61 1BD, Scotland, United Kingdom [J. L. D., W. N. K., K. A. O.]; Beatson Institute for Cancer Research, Cancer Research UK Beatson Laboratories, Glasgow G61 1BD, Scotland, United Kingdom [J. K. V.]; Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland, United Kingdom [E. C. W.]; and University Department of Pathology, Glasgow Royal Infirmary, Glasgow G4 0SF, Scotland, United Kingdom [K. A. O.]

Abstract

Patients presenting with metastatic adenocarcinoma of unknown origin are a common clinical problem. Their optimal management and therapy are facilitated by identification of the primary site, yet histologically these tumors are almost identical. Better tumor markers are needed to enable the assignment of metastases to likely sites of origin. In this study, hierarchical clustering of public serial analysis of gene expression data showed that adenocarcinomas and their metastases cluster according to their site of origin. A novel bioinformatic approach was developed to exploit the differences between these clusters, using diverse sources: public expression data from serial analysis of gene expression and digital differential display; and the published literature, including microarray studies. Sixty-one candidate tumor markers with expression predicted to be characteristic of the site of origin were identified. Eleven genes were tested by reverse transcription-PCR in primary adenocarcinomas from a range of sites, and seven (64%) were site-restricted. Analysis of public gene expression data sets is a powerful method for the identification of clinically relevant tumor markers.

Introduction

Oncologists are commonly presented with patients with metastatic adenocarcinoma for which the primary tumor site is unknown (1). Optimal clinical management of these patients is enabled by diagnosis of the site of origin. By far the most common adenocarcinomas arise in lung, colon, breast, prostate, stomach, ovary, and pancreas, but their disease prognosis and therapy vary. These organs are therefore investigated as potential primary sites (2–4). Unfortunately, adenocarcinomas metastatic from these different locations have almost identical microscopic appearances, which do not assist diagnosis of the site of origin.

Traditionally, cancer classification has been based on histopathological and clinical data. New technologies, however, have enabled genome-wide analysis of gene expression patterns, which have suggested that individual cancer phenotypes exhibit characteristic expression profiles [examples include leukemia (5) and breast cancer (6)]. We wanted to compare gene expression profiles from adenocarcinomas from a range of sites commonly known to give rise to metastatic disease to identify tumor markers characteristic of the site of origin.

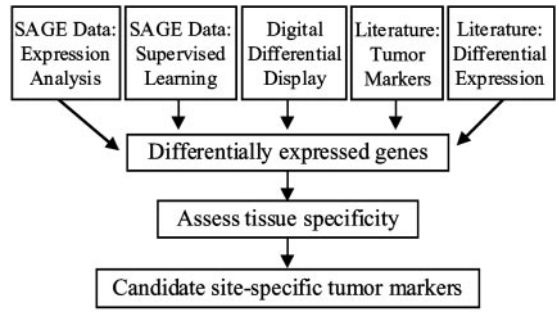

Fig. 1. Candidate tumor markers were identified through a two-stage process: identification of differentially expressed genes; and assessment of the tissue specificity of these genes. Differentially expressed genes were identified through analysis of five data sets; assessment of tissue specificity used publicly available SAGE data.

Our approach centered on data generated by SAGE,[4] which is based on the isolation of short sequence tags (10 bp) from individual mRNA species (7). Sequencing of linked tags allows efficient characterization of transcripts and the digital representation of their expression levels. The production of absolute transcript numbers in a digital format allows comparison between data sets generated by different laboratories. SAGE data are currently available publicly via the Cancer Genome Anatomy Project.[5]

In this study, hierarchical clustering of publicly available SAGE data from a range of adenocarcinomas confirmed the presence of differences in gene expression between tumor sites. These differences were then exploited using a bioinformatic approach that incorporated data from a large panel of tumors investigated using the various technologies currently available for global gene expression profiling. The approach, outlined in Fig. 1, had two phases: identification of differentially expressed genes; followed by assessment of their tissue specificity. Differentially expressed genes were identified on-line and from published literature from experiments using SAGE, microarrays, DDD, differential display, and subtractive hybridization. These 515 genes were then investigated in a wider set of SAGE data for selection on the basis of gene expression *in silico* specific to adenocarcinoma from one site of origin or restricted to two primary sites. Sixty-one candidate markers emerged. The expression pattern of 11 of these markers was then validated by RT-PCR in a range of primary adenocarcinomas.

Received 5/3/02; accepted 9/11/02.

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 Cancer Research UK and the University of Glasgow.

[2] Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org).

[3] To whom requests for reprints should be addressed, at Cancer Research UK Department of Medical Oncology, Cancer Research UK Beatson Laboratories, University of Glasgow, Garscube Estate, Switchback Road, Glasgow G61 1BD, Scotland, United Kingdom. Phone: 44-0-141-330-3506; Fax: 44-0-141-330-4127; E-mail: k.oien@beatson.gla.ac.uk.

[4] The abbreviations used are: SAGE, serial analysis of gene expression; CEA, carcinoembryonic antigen; CK, cytokeratin; DDD, digital differential display; EST, expressed sequence tag; IHC, immunohistochemistry; NCBI, National Center for Biotechnology Information; PSA, prostate-specific antigen; RT-PCR, reverse transcription-PCR; TFF2, trefoil factor 2; PSCA, prostate stem cell antigen.

[5] www.ncbi.nlm.nih.gov/CGAP/.

Table 1 *SAGE libraries used to identify potential tumor markers*

Libraries used to assess the tissue specificity of differentially expressed genes are listed. A subset of these libraries, indicated in bold type, was also used for hierarchical clustering and supervised learning. Two additional gastric libraries were used for hierarchical clustering and supervised learning (gastric cancer-G234 and gastric cancer xenograft X101) but are not listed because they were not used to assess tissue specificity of differentially expressed genes. Additional library details are available from the NCBI website. Abbreviations: T, library derived from a primary tumor; M, library derived from a metastatic tumor; CL, library derived from a tumor cell line; DCIS, ductal carcinoma *in situ*.

Breast	Ovary	Prostate	Pancreas	Colon
95-259 (T)	A2780 (CL)	A− (CL)	CAPAN 1 (CL)	Caco 2 (CL)
95-260 (M)	ES2-1 (CL)	A+ (CL)	CAPAN 2 (CL)	HCT116 (CL)
95-347 (T)	ML 10-10 (CL)	**Chen tumor Pr (T)**	HS766T (CL)	RKO (CL)
95-348 (M)	OV1063-3 (CL)	LN-1 (T)	**Panc 91-16113 (T)**	SW837 (CL)
DCIS (T)	OVCA432-2 (CL)	Chen LNCaP (CL)	**Panc 96-6252 (T)**	**Tu102 (T)**
DCIS 2 (T)	OVP-5 (CL)	CPDR LNCaP C (CL)	Panc 1 (CL)	**Tu98 (T)**
Duke 95-349 (T)	**OVT-6 (T)**	LNCaP no DHT (CL)		
LacZ (CL)	**OVT-7 (T)**	CPDR LNCaP T (CL)		
MDA453 (CL)	**OVT-8 (T)**	**PR317 prostate tumor (T)**		
PTEN (CL)		PrCA-1 (T)		
MCF7 control 0 (CL)		TSU (CL)		
MCF7 control 3 (CL)				
MCF7 estrogen 10 (CL)				
MCF7 estrogen 3 (CL)				
SKBR3 (CL)				

Materials and Methods

Hierarchical Clustering. Fifteen SAGE libraries, derived from adenocarcinomas originating from different sites, were downloaded from the NCBI website.[6] These tumor libraries, from breast, colon, ovary, pancreas, prostate, and stomach, are listed in Table 1. At the time of analysis, no SAGE tumor libraries from lung were publicly available, so lung adenocarcinomas were not included in the hierarchical clustering. Tags were extracted from sequence files using the SAGE 3.04 program (7). Similarity/dissimilarity between libraries was studied using the statistical analysis program "R". First, the Canberra distance metric was applied, which is robust to missing values, sensitive to small values, and ignores comparisons when both observations are zero (8). Then, the resulting dissimilarity matrix was analyzed by "hclust" complete hierarchical clustering, which is based on a method of Murtagh (9). To group libraries of similar origin together, the distance metric was found to be more important than the clustering method, and of the metrics available in "R," Canberra gave the best clustering of libraries into obvious classes.

Marker Identification. Fig. 1 outlines the approach used to identify potential markers of adenocarcinoma characteristic of the site of origin. There were two phases: (*a*) initial identification of differentially expressed genes using multiple data sets; and (*b*) determination of tissue specificity or tissue restriction of differentially expressed genes. Initial identification of differentially expressed genes used five data sets: (*a*) expression analysis of SAGE data; (*b*) supervised learning using SAGE data; (*c*) DDD data; (*d*) literature reporting putative tumor markers; and (*e*) literature reporting differentially expressed genes (all as detailed below). Differentially expressed genes identified from these five sources were evaluated for tissue specificity or tissue restriction via analysis of additional SAGE data.

Expression Analysis of SAGE Data. Five SAGE libraries from adenocarcinomas of breast, colon, ovary, pancreas, and prostate (96-349, Tu102, OVT7, Panc-96-6252, and PR317 prostate tumor) were downloaded from the NCBI website. A local SAGE library of primary gastric adenocarcinoma was also used. At the time of analysis, no SAGE tumor libraries from lung were publicly available, so lung adenocarcinomas were not included in the expression analysis. Using the SAGE 3.04 program, tags were extracted, and each library was compared with the five other libraries in a series of pair-wise comparisons. Statistical significance for these comparisons was based on Monte Carlo simulations within the program with a cutoff of $P < 0.001$. Tag fold change was defined in Microsoft Access as (fx/Σ library 1 tags)/(fx/Σ library 2 tags) where fx represents the frequency of a given tag and Σ library 1 tags represents the total tags in the library of interest. Differential expression was defined as a >10-fold increase in expression over all other libraries in combination with $P < 0.001$.

SAGE Data: Supervised Learning. Supervised learning was performed on the same data set of 15 SAGE libraries used for hierarchical clustering. Genes with low expression levels (total level of <20 over all 15 tumors) were eliminated, and square root transformations of the remaining data were taken. Iterative logistic regression was used to determine the usefulness of each gene in discriminating each tumor from the remaining tumors. The predictive power of the genes was then assessed by calculating the predicted class label for a randomly omitted tumor and comparing that with the true class label. A calibration score was calculated as $P_{estimated} - P_{true}$, where P_{true} is either 1 (correct) or 0 (false); a low score represents a good prediction. This cross-validation process was repeated, and an average prediction score was determined. Calculations were performed using the program Splus 2000. For each tumor, the 20 tags with the best predicting scores were selected.

DDD. DDD, a bioinformatic tool available via the NCBI website,[7] compares the frequencies of ESTs (or uncharacterized cDNAs) between tumor expression libraries in the Cancer Genome Anatomy Project database. Sixteen pooled EST libraries were compared, representing adenocarcinomas from breast (Br15, Br17, and Br18), colon (Co11, Co18, and Co22), lung (Lu26 and Lu27), ovary (Ov1, Ov2, and Ov8), pancreas (Pan1), prostate (Pr3, Pr8, and Pr23), and stomach (Gas4). DDD identifies ESTs differentially expressed with a statistical significance of $P < 0.05$ (Fisher's exact test). The 25 most differentially expressed ESTs were selected.

Literature: Tumor Markers. The Medline database from 1997 to 2000 was searched for papers describing putative tumor markers for sites relevant to adenocarcinoma of unknown origin (breast, colon, lung, ovary, pancreas, prostate, and stomach). One hundred and eighteen potential markers were identified.

Literature: Differential Expression. A literature search was performed of publications between 1998 and 2001 on large-scale expression analysis of tumors relevant to adenocarcinoma of unknown origin. Twenty-seven papers were chosen (10–36), which used diverse technologies including microarrays, differential display, and subtractive hybridization, as well as SAGE. These reported a total of 530 differentially expressed genes, from which 87 were selected on the basis of experimental evidence of higher expression in the tumors of interest by Northern blotting, IHC, or PCR-based methods.

Tissue Specificity or Restriction. The differentially expressed genes and putative tumor markers thus identified were then tested for their tissue specificity (Fig. 1) against a wider panel of 47 SAGE libraries relevant to the problem of adenocarcinoma of unknown primary site, as detailed in Table 1. Expression information from these SAGE libraries is accessible via the "gene to tag mapper" tool available at the NCBI website.[8] For a given gene, this tool displays the potential SAGE tags in the transcript, the frequencies of those tags in each SAGE library in the database, and the normalized level of expression (tags/million) therein. Libraries were separated according to tissue of origin and expression level. Expression levels were grouped into abundance classes, and each class was assigned a weighting. Scores were then calculated based on weight and number of libraries for each expression class. This process resulted in a numerical representation of the expression level of each gene in five tissues. These numbers were then converted to a color scale that was used to prioritize genes for subsequent validation. Candidate tumor markers were

[6] ftp.ncbi.nih.gov/pub/sage/seq.

[7] www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs.

[8] www.ncbi.nlm.nih.gov/SAGE/SAGEcid.cgi.

defined as genes expressed highly in one (specific) or two (restricted) tissues and at low levels in all other tissues. This definition was fulfilled by 61 of the 515 differentially expressed genes already identified. Further information was then sought on each of the 61 candidate genes individually, through literature searches, NCBI's UniGene database, and the Weizmann Institute's GeneCards database.[9] Eleven genes with the strongest evidence for expression restricted to the desired tissues were selected for validation by RT-PCR.

Tissue and cDNA Samples. Primary tissues used included adenocarcinomas from breast, ovary, stomach, pancreas, and lung (two samples of the first four types and one sample of the last type were obtained). Ethical approval for the use of clinical material was given by the North Glasgow University Hospitals National Health Service Trust. The LNCaP prostate adenocarcinoma cell line was also used, along with cDNA samples of prostatic and lung adenocarcinomas [Invitrogen (Paisley, United Kingdom) and Origene (Rockville, MD), respectively].

RNA Isolation and RT-PCR. Tissues were homogenized using a Ribolyser (Hybaid, Ashford, United Kingdom). Total RNA was isolated using Trizol reagent (Invitrogen). RNA was reversed transcribed to cDNA using SuperScript First Strand Synthesis System for RT-PCR (Invitrogen) with oligo(dT) primers. PCR was then performed; primer sequences and conditions are available from the authors on request.

Results

Clustering. Hierarchical clustering identifies differences between data sets and thus separates them into classes and subclasses, eventually forming a hierarchy that is displayed as a dendrogram. This method was applied to 15 SAGE libraries of adenocarcinomas from breast, colon, ovary, pancreas, prostate, and stomach. In the resulting dendrogram, the libraries clustered according to their site of origin (Fig. 2). Two pairs of libraries were derived from primary and metastatic breast tumors from the same patients (95–347 and 95–348; 95–259 and 95–260); these libraries clustered together.

Identification of Differentially Expressed Genes. Five data sets were used: (*a*) analysis of SAGE expression data; (*b*) supervised learning using SAGE data; (*c*) information from DDD; (*d*) literature describing putative tumor markers; and (*e*) literature reporting differentially expressed genes. A total of 515 differentially expressed genes emerged, with 128, 157, 25, 118, and 87 derived from each source, respectively.

Identification of Tumor Markers Characteristic of the Site of Origin. The 515 differentially expressed genes and putative tumor markers were then tested for tissue specificity against a wider panel of 47 SAGE libraries relevant to adenocarcinoma of unknown primary site. Candidate tumor markers were defined as genes expressed highly in one (specific) or two (restricted) tissues and at low levels in all other tissues. Sixty-one transcripts emerged, including both known tumor markers and genes not previously reported as markers (Fig. 3). Among the established markers were CEA and PSA, which are both valuable and widely used clinically. CEA expression *in silico* was restricted to colonic and pancreatic tumors, where it was present at high levels (*orange color*). PSA was abundant in prostatic carcinoma (*red*) but low elsewhere (*blue*). Genes found to be tissue specific or tissue restricted but not previously regarded as tumor markers included lipophilin B and glutathione peroxidase 2. The complete set of results is available online as supplementary information.[2]

Validation of Candidates by RT-PCR. Eleven candidate genes identified by these bioinformatic approaches as showing tissue-specific or tissue-restricted expression were then validated by RT-PCR on tumor tissues and the LNCaP cell line (Fig. 4). The expression patterns of 3 of these 11 genes exactly corroborated the *in silico* predictions. The results for an additional four genes were broadly similar, but with some variance. These are shown in Fig. 4 by *double*

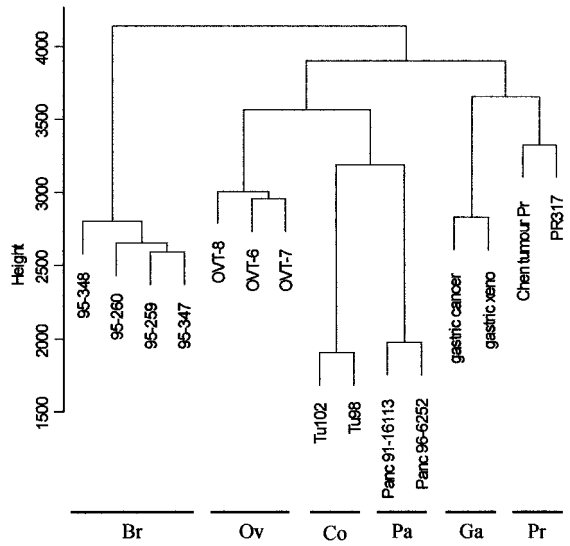

Fig. 2. Hierarchical clustering of SAGE data was used to cluster 15 adenocarcinomas spanning six sites. Two pairs of libraries were derived from corresponding primary and metastatic breast tumors (95–347 and 95–348; 95–259 and 95–260). *Br*, breast tumors; *Ov*, ovarian tumors; *Co*, colon tumors; *Pa*, pancreatic tumors; *Ga*, gastric tumors; *Pr*, prostate tumors.

and *single asterisks*, respectively. The candidates included genes already recognized as tumor markers, namely, PSA, mammaglobin 1, TFF2 (human spasmolytic polypeptide), pepsinogen C, and surfactant A (Fig. 4*A*), as well as the previously uncharacterized markers lipophilin B and glutathione peroxidase 2 (Fig. 4*B*).

Discussion

Adenocarcinomas of unknown origin are commonly encountered in clinical practice. These tumors arise in and spread from primary sites including lung, colon, pancreas, stomach, breast, ovary, and prostate (2–4), which vary in clinical outcome. Knowledge of the primary site facilitates appropriate clinical management and therapy of patients (37). The common glandular epithelial origin of these tumors results in histological similarity in the presenting metastases that is not suggestive of the site of origin. The only exception is metastatic colonic adenocarcinoma, which may exhibit a characteristic appearance comprising large glandular structures with columnar epithelium and extensive necrosis. Identification of better molecular markers to assist diagnosis of the primary site is therefore required. To date, few studies have examined expression patterns across a range of adenocarcinomas (38, 39).

Publicly available SAGE data representing 15 tumors from sites relevant to the problem of adenocarcinoma of unknown origin were downloaded from the NCBI website. Hierarchical clustering of this data showed that tumors clustered according to site of origin, suggesting that there are similarities between tumors of common origin and differences between primary sites. These similarities and differences may be exploited to assist in prediction of the tissue of origin of metastatic adenocarcinoma and underlie our research approach.

Recently, Ramaswamy *et al.* (38) generated microarray expression data for a very large series of tumor types, including adenocarcinomas from the seven sites studied here, lymphoma, leukemia, malignant melanoma, and central nervous system tumors. Hierarchical clustering successfully divided the carcinomas from the other tumor types but did not, unlike our study, separate the adenocarcinomas according to their site of origin. This may be due to their use of measured (microarray) data rather than the counted (SAGE) data here. Alternatively, their chosen algorithm could have masked subtle differences between tumors of epithelial origin.

[9] www.ncbi.nlm.nih.gov/UniGene and bioinformatics.weizmann.ac.il/cards.

A

Gene	Tag Sequence	Data Set	For	Br	Ov	Pr	Pa	Co
Amnionless protein	CGGGAGTCGG	SL	Co					
Carboxypeptidase E	TTTACAAAGA	SAGE	Pr					
Cartilage oligomeric matrix protein	CGGGGTGGCC	SAGE	Pa					
CEA; Carcinoembryonic antigen	AAGGATAAAA	Lit/SAGE	Co, Pa					
Chloride channel, calcium activated, family member 1	GAACAGCTCA	SAGE	Co					
Chromogranin A		Lit	Lu					
Clusterin (Apolipoprotein J)		Lit	Ov					
Collagen type IV alpha 5		Lit	Ga					
Copine III	TGACCAAAAC	SL	Br					
Cyclin dependent kinase inhibitor 1B		Lit	Ga					
Differentially expressed in hematopoietic lineages	GCAAATAAAT	SL	Ga					
DKFZP546B167	ACTTAAGGAA	SAGE	Pr					
DNA-binding protein amp. expr. of surfactant protein B		Lit	Lu					
Enhancer of rudimentary (Drosophila) homolog	GCTAAACTGC	SAGE	Br					
EST	TGGGCAGCTG	SL	Co					
EST wk sim to MUC2_Human mucin 2 precursor	ATGAAACTTC	SL	Pr					
ESTs	TCTTTATTAG	SAGE	Pr					
ESTs	AGTGCTCTTT	SAGE	Ga					
ESTs	ATGTAAAAAC	SAGE	Ga					
ESTs wk sim to A56154ABI substrate ena(D.melanogaster)	GCCCAGGTCA	SAGE	Co					
Fatty acid binding protein	TAGCAGACCC	SAGE	Co					
Glutathione peroxidase 2	**GGTGGTGTCT**	**SAGE**	**Co**					
H. sapiens cDNA: FLJ21409 fis clone COL03924	AGTATGACCT	SL	Co					
H. sapiens cDNA: FLJ22066 fis clone HEP10611	CTTGGGTTTT	SL	Co					
H. sapiens HSPC323 mRNA partial cds	CACCCACTGC	SL	Ga					
H. sapiens mRNA cDNA DKFZp586N0121	TTGAAACCCC	SL	Co					
Homo sapiens cDNA FLJ12389 fis, clone MAMMA1002671	CTGTGAAAAA	SAGE	Pr					
Homo sapiens cDNA: FLJ21968 fis, clone HEP05670	AAAACTTTTG	SAGE	Br					
Human DNA sequence from clone RPS-1187J4	CCAAGGTGGC	SAGE	Ga					
Hypothetical protein FLJ23384		Lit	Lu					
Immunoglobulin kappa constant	AGGGTCCCCG	SAGE	Ov					
Immunoglobulin kappa variable	AGGGTCCCCG	SAGE	Ov					
Insulin	GCCCTGTGGA	SAGE	Pa					
Insulin like growth factor 2 (somatomedin A)	TACAAAATCG	SL	Co					
Interleukin 2		Lit	Ga					
Kallikrein 2	CTGTGGTTAA	SAGE	Pr					
Kallikrein 3 (Prostate Specific Antigen, PSA)	**GGATGGGGAT**	**DDD/Lit/SAGE**	**Pr**					
KIAA0741		DDD	Lu					
KIAA0876	**AACGCTGCGA**	**SAGE**	**Br**					
KIAA0965 protein	ATGCAAATTA	SL	Ga					
KIAA0966	ATGTAAAAAG	SAGE	Ga					
Lipophilin B	**TAAAAACTTT**	**SAGE**	**Br**					
Lung cancer candidate		Lit	Lu					
Mammaglobin 1		**Lit**	**Br**					
Matrix metalloproteinase 7 (matrilysin, uterine)	AACTTGGCCA	SL	Pa					
Matrix metalloproteinase 9		Lit	Lu					
Metallothionein 1L		**Lit**	**Br**					
Microseminoprotein beta	CCTATCAGTA	DDD/SAGE	Pr					
NK homeobox (Drosophila) family 3, A	TCTTTATTAG	SAGE	Pr					
Pancreatic polypeptide	AGCAGCGCCA	Lit/SAGE	Pa					
Pepsinogen C	**AGTGCTCTTT**	**SAGE**	**Ga**					
PR domain containing 10	**AGATGCTAAA**	**SAGE**	**Co**					
Prostate stem cell antigen		**Lit**	**Pa/Pr**					
Prostatic acid phosphatase (PAP)	TCTAGAGAAC	DDD, SAGE	Pr					
Serine/threonine kinase		Lit	Lu					
Sjogren syndrome Ag B (autoantigen La)	ATGAAAATGG	SAGE	Ga/Pa					
Spermine synthase		Lit	Pr					
Surfactant B		Lit	Lu					
Surfactant, pulmonary associated protein A (Surfactant A)		**Lit**	**Lu**					
Transforming growth factor alpha		Lit	Ga					
Trefoil factor 2	**AAATCCTGGG**	**Lit/SL**	**Pa**					

B

Fig. 3. *A*, expression profiles of 61 candidate genes determined to have tissue-specific expression after analysis of SAGE data. Expression patterns of genes shown in *bold* were validated by RT-PCR. Tag sequences are given for genes identified through analysis of SAGE data. The "For" column indicates the site for which the gene was predicted to act as a marker after analysis of differential expression. *B*, color scale used to represent expression level: low expression is represented by *blue*, and high expression is represented by *red*. *Br*, breast; *Ov*, ovary; *Pr*, prostate; *Pa*, pancreas; *Co*, colon; *Ga*, gastric; *SL*, supervised learning; *Lit*, literature.

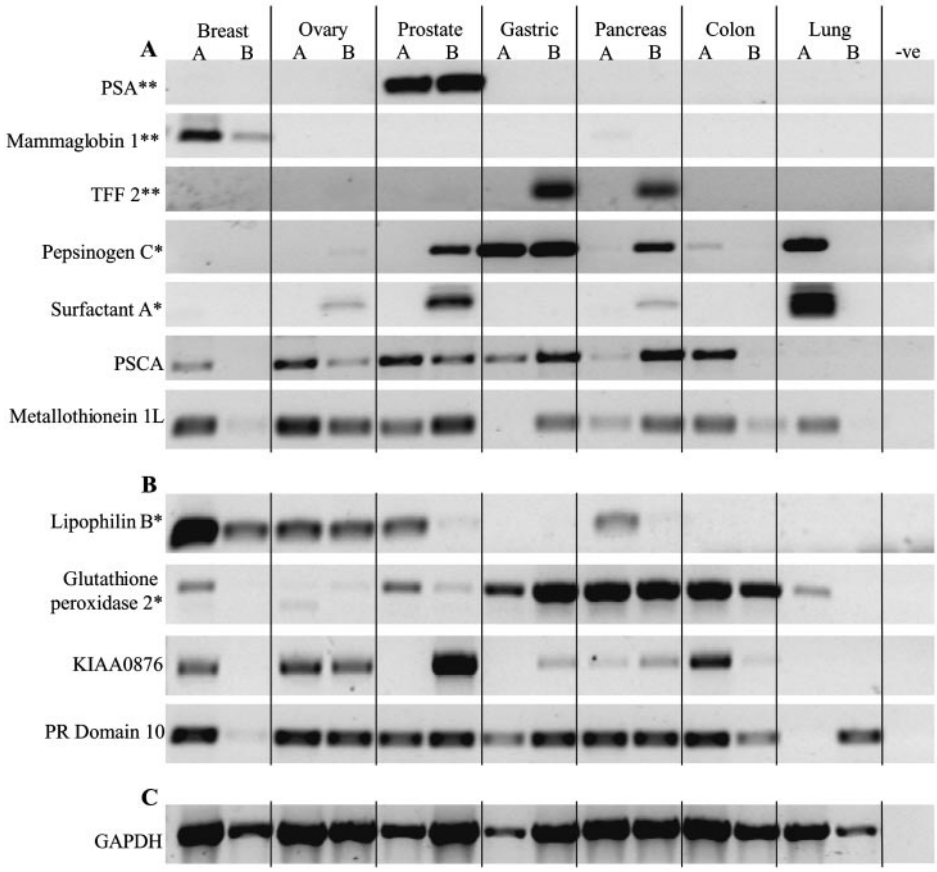

Fig. 4. *A*, RT-PCR for established tumor markers. *B*, RT-PCR for novel tumor markers. *C*, RT-PCR for the internal control, GAPDH. Expression patterns by RT-PCR that exactly corroborate the *in silico* predictions are indicated by *double asterisks*; those that are broadly similar but have some variance are indicated by *single asterisks*. Prostate sample B was the LNCaP cell line. *PR Domain 10*, PR domain containing 10; *GAPDH*, glyceraldehyde-3-phosphate dehydrogenase.

The clustered data included two pairs of breast carcinoma samples derived from primary tumors and corresponding lymph node metastases. These four samples clustered together. This suggests that the expression patterns of the metastatic tumors more closely resembled primary tumors of the same origin than primary adenocarcinomas from alternative sites. Primary and metastasis pairs also showed similar expression profiles in a cDNA microarray study of lung adenocarcinomas (40). Few other large-scale studies of primary and metastatic tumors exist, but data from the many immunohistochemical studies that have looked at small numbers of genes suggest phenotypic similarity (41).

Candidate tumor markers were identified through a bioinformatic approach centered on the analysis of SAGE data. This approach comprised two phases: identification of differentially expressed genes using expression data generated through a range of technologies; and then evaluation of the tissue specificity or tissue restriction of these genes using SAGE data. This approach resulted in the selection of the 61 candidates shown in Fig. 3. Among these genes were the known tumor markers PSA (which differs from PSCA) and CEA. Their clinical value in the diagnosis and management of patients with prostatic and colorectal tumors, respectively, is well-established (42, 43). The presence of these genes self-validates the approach and suggests that the other genes thus identified may prove similarly useful.

Eleven of these candidate genes were chosen for validation by RT-PCR in clinical material. RT-PCR is an established method for the rapid assessment of candidate genes in a range of samples [for example, see Scheurle *et al.* (26)]. RT-PCR results for seven (64%) of these genes were consistent with the expression patterns predicted by bioinformatics: three agreed exactly; and four were broadly similar but with some variance. This compares well with previous studies [for example, Scheurle *et al.* (26) achieved concordance in 3 of 12 known genes]. Seven genes had previously been reported as tumor markers, either in a tissue-specific manner or by being up-regulated in tumors when compared with normal tissues. These genes were PSA [prostate (42)], mammaglobin 1 (breast), TFF2 (also called human spasmolytic polypeptide, specific for pancreas), pepsinogen C (stomach), surfactant A (lung), PSCA (pancreas), and metallothionein 1L (pancreas).

For PSA, mammaglobin 1, and TFF2, the *in silico* and RT-PCR analyses agreed exactly. Pepsinogen C was abundant in the gastric cancers by RT-PCR but was also present in one of the two lung and pancreatic adenocarcinomas and in the prostate cell line. Expression of pepsinogen C in tissues other than the stomach had not been predicted by our bioinformatics approach but is described in Unigene and in the literature; nevertheless, these sources report the highest levels by far of pepsinogen C to be in gastric tissues. Surfactant A was abundant in one of the lung adenocarcinomas by RT-PCR, as predicted, but was also present in the prostate cell line. In fact, expression of surfactant A in the prostate, at lower levels than in the lung, has recently been reported (44). *In silico* analysis had predicted that PSCA would be abundant in pancreatic tumors and would be present, but at lower levels, in breast, colon, ovary, prostate, and stomach tumors. Not surprisingly, PSCA was found by RT-PCR in all of these samples.

The expression patterns of four genes not previously reported as tumor markers were also validated by RT-PCR. These genes, all identified through analysis of SAGE data, were lipophilin B, glutathione peroxidase 2, KIAA0876, and PR domain containing 10. By both bioinformatics and RT-PCR, lipophilin B was restricted to breast, ovary, and prostate tumors, and glutathione peroxidase 2 was expressed at higher levels in colon and pancreatic carcinomas than in breast, ovary, and prostate tumors. Conversely, although KIAA0876, PR domain containing 10, and the known tumor marker metallothio-

nein 1L were predicted be differentially expressed in breast, colon, and pancreatic tumors, respectively, none of the three displayed tissue specificity by RT-PCR. This is likely to be due to variability within and between the tumors from which the SAGE data were generated and those used for our validation. The expression patterns of these genes remain to be determined in a larger set of clinical samples.

Analysis of the expression data and RT-PCR results generated from clinical samples shows that few potential markers are expressed in a truly tissue-specific manner. We anticipate that information from a number of genes used together in a marker panel would be required to assist in predicting the primary site of adenocarcinomas of unknown origin. Such marker panels are already well established for the diagnosis and classification of malignant lymphoma in pathological specimens. Moreover, Su *et al.* (39) recently reported multiclass tumor classification using a set of 11 genes that correctly assigned 83% of test tumors into classes. This suggests that it should also be possible to use the similar small number of genes identified here in a marker panel to assign adenocarcinomas to different classes, according to site.

Application of such a marker panel should be achievable through IHC, which localizes the protein product of a gene in microscopic sections. The advantage of using IHC over other technologies (for example, cDNA microarrays) is that IHC is already in routine use in histopathology laboratories, where the diagnosis of metastatic adenocarcinoma is usually confirmed on a tissue biopsy or cell sample from the patient. Few immunohistochemical markers have yet been used for this purpose. The most valuable are CK20 and CK7, with CK20 positivity suggesting an origin in the gut (41), and thyroid transcription factor-1, which is associated with lung tumors (45). Consequently, no new technology would be needed, and any new panel of tumor markers for adenocarcinomas could be taken forward swiftly into clinical use. Ultimately, this depends on demonstrating that the gene expression profiles described here translate to protein expression patterns that can correctly predict tumor class, and our results provide a resource for further pathological testing.

In conclusion, novel and known tissue-specific and tissue-restricted tumor markers for adenocarcinoma have been identified through the analysis of publicly available expression data and validated in clinical material. The ability of these markers to categorize adenocarcinomas according to their tissue of origin can now be tested in a broader panel of archival tumor samples. Improved prediction of likely primary site in patients with adenocarcinomas of unknown origin should lead to better assessment of their prognosis and optimal, tailored therapy.

Acknowledgments

We thank Dr. Kenneth W. Kinzler for SAGE software.

References

1. Hillen, H. F. Unknown primary tumours. Postgrad. Med. J., *76:* 690–693, 2000.
2. Ayoub, J-P., Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N., and Abbruzzese, J. L. Unknown primary tumours metastatic to liver. J. Clin. Oncol., *16:* 2105–2112, 1998.
3. Katagiri, H., Takahashi, M., Inagaki, J., Sugiura, H., Ito, S., and Iwata, H. Determining the site of the primary cancer in patients with skeletal metastasis of unknown origin. Cancer (Phila.), *86:* 533–537, 1999.
4. Tattersall, M. H. N. Unknown primary cancers. *In:* R. R. Love (ed.), Manual of Clinical Oncology, 6th ed., pp. 532–541. Geneva: Springer-Verlag, 1994.
5. Golub, T. R., Slomin, 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 (Wash. DC), *286:* 531–537, 1999.
6. Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L. H., Borg, A., Ferno, M., Peterson, C., and Meltzer, P. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res., *61:* 5979–5984, 2001.
7. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. Science (Wash. DC), *270:* 484–487, 1995.
8. Mardia, K. V., Kent, J. T., and Bibby, J. M. Multivariate Analysis. London: Academic Press, 1979.
9. Murtagh, F. Multidimensional Clustering Algorithms. COMPSTAT Lectures 4. Wuerzburg, Germany: Physica-Verlag, 1985.
10. Argani, P., Rosty, C., Reiter, R. E., Wilentz, R. E., Murugesan, S. R., Leach, S. D., Ryu, B., Skinner, H. G., Goggins, M., Jaffee, E. M., Yeo, C. J., Cameron, J. L., Kern, S. E., and Hruban, R. H. Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma. Cancer Res., *61:* 4320–4324, 2001.
11. Backert, S., Gelos, M., Kobalz, U., Hanski, M. L., Bohm, C., Mann, B., Lovin, N., Gratchev, A., Mansmann, U., Moyer, M. P., Riecken, E-O., and Hanski, C. Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array. Int. J. Cancer, *82:* 868–874, 1999.
12. Hibi, K., Liu, Q., Beaudry, G. A., Madden, S. L., Westra, W. H., Wehage, S. L., Yang, S. C., Heitmiller, R. F., Bertelsen, A. H., Sidransky, D., and Jen, J. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res., *58:* 5690–5694, 1998.
13. Hippo, Y., Yashiro, M., Ishii, M., Taniguchi, H., Tsutsumi, S., Hirakawa, K., Kodama, T., and Aburatani, H. Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes. Cancer Res., *61:* 889–895, 2001.
14. Hough, C. D., Sherman-Baust, C. A., Pizer, E. S., Montz, F. J., Im, D. D., Rosenhein, 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.
15. Ismail, R. S., Baldwin, R. L., Fang, J., Browning, D., Karlan, B. Y., Gasson, J. C., and Chang, D. D. Differential gene expression between normal and tumor-derived ovarian epithelial cells. Cancer Res., *60:* 6744–6749, 2000.
16. Kitahara, O., Furukawa, Y., Tanaka, T., Kihara, C., Ono, K., Yanagawa, R., Nita, M. E., Takagi, T., Nakamura, Y., and Tsunoda, T. Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Res., *61:* 3544–3549, 2001.
17. Lin, B., Ferguson, C., White, J. T., Wang, S., Vessella, R., True, L. D., Hood, L., and Nelson, P. S. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2. Cancer Res., *59:* 4180–4184, 1999.
18. Lin, B., White, J. T., Ferguson, C., Bumgarner, R., Friedman, C., Trask, B. J., Ellis, W., Lange, P., Hood, L., and Nelson, P. S. PART-1: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12. Cancer Res., *60:* 858–863, 2000.
19. Manda, R., Kohno, T., Matsuno, Y., Takenoshita, A., Kuwano, H., and Yokota, J. Identification of genes (*SPON2* and *C20orf2*) differentially expressed between cancerous and non-cancerous lung cells by mRNA differential display. Genomics, *61:* 5–14, 1999.
20. Martin, K. J., Graner, E., Li, Y., Price, L. M., Kritzman, B. M., Fournier, M. V., Rhei, E., and Pardee, A. B. High-sensitivity array analysis of gene expression for the early detection of disseminated breast tumor cells in peripheral blood. Proc. Natl. Acad. Sci. USA, *98:* 2646–2651, 2001.
21. Nacht, M., Ferguson, A. T., Zhang, W., Petroziello, J. M., Cook, B. P., Gao, Y. H., Maguire, S., Riley, D., Coppola, G., Landess, G. M., Madden, S. L., and Sukumar, S. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. Cancer Res., *59:* 5164–5470, 1999.
22. Notterman, D. A., Alon, U., Sierk, A. J., and Levine, A. J. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma and normal tissue examined by oligonucleotide arrays. Cancer Res., *61:* 3124–3130, 2001.
23. Parle-McDermott, A., McWilliam, P., Tighe, O., Dunican, D., and Croke, D. T. Serial analysis of gene expression identifies putative metastasis-associated transcripts in colon tumour cell lines. Br. J. Cancer, *83:* 725–728, 2000.
24. Peterson, S., Heckert, C., Rudolf, J., Schluns, K., Tchernitsa, O. I., Schafer, R., Dietel, M., and Petersen, I. Gene expression profiling of advanced lung cancer. Int. J. Cancer, *86:* 512–517, 2000.
25. Ryu, B., Jones, J., Hollingsworth, M. A., Hruban, R. H., and Kern, S. E. Invasion-specific genes in malignancy: serial analysis of gene expression comparisons of primary and passaged cancers. Cancer Res., *61:* 1833–1838, 2001.
26. Scheurle, D., DeYoung, M. P., Binninger, D. M., Page, H., Jahanzeb, M., and Narayanan, R. Cancer gene discovery using digital differential display. Cancer Res., *60:* 4037–4043, 2000.
27. Schummer, M., Ng, W. V., Bumgarner, R., Nelson, P. S., Schummer, B., Bednarski, D. W., Hassell, L., Baldwin, R. L., Karlan, B. Y., and Hood, L. Comparative hybridisation of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene (Amst.), *238:* 375–385, 1999.
28. Sgori, D. C., Teng, S., Robinson, G., LeVangie, R., Hudson, J. R., and Elkahloun, A. G. *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res., *59:* 5656–5661, 1999.
29. Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B., and Pastan, I. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. Proc. Natl. Acad. Sci. USA, *95:* 300–304, 1998.
30. Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., Cook, B. P., Dufault, M. R., Ferguson, A. T., Gao, Y., He, T-C., Hermeking, H., Hiraldo, S. K., Hwang, P. M., Lopez, M. A., Luderer, H. F., Mathews, B., Petroziello, J. M., Polyak, K., Zawel, L., Zhang, W., Zhang, X., Zhou, W., Haluska, F. G., Jen, J., Sukumar, S., Landes, G. M., Riggins, G. J., Vogelstein, B., and Kinzler, K. W. Analysis of human transcriptomes. Nat. Genet., *23:* 387–388, 1999.
31. Waghray, A., Schober, M., Feroze, F., Yao, F., Virgin, J., and Chen, Y. Q. Identification of differentially expressed genes by serial analysis of gene expression in human prostate cancer. Cancer Res., *61:* 4283–4286, 2001.
32. Wang, K., Gan, L., Jeffery, E., Gayle, M., Gown, A. M., Skelly, M., Nelson, P. S., Ng, W. V., Schummer, M., Hood, L., and Mulligan, J. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. Gene (Amst.), *229:* 101–108, 1999.

33. Wang, T., Hopkins, D., Schmidt, C., Silva, S., Houghton, R., Takita, H., Repasky, E., and Reed, S. G. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. Oncogene, *19:* 1519–1528, 2000.
34. Xu, J., Stolk, J. A., Zhang, X., Silva, S. J., Houghton, R. L., Matsumura, M., Vedvick, T. S., Leslie, K. B., Badaro, R., and Reed, S. G. Identification of differentially expressed genes in human prostate cancer using a combination of subtraction and microarray analysis. Cancer Res., *60:* 1677–1682, 2000.
35. Xu, L. L., Stackhouse, B. G., Florence, K., Zhang, W., Shanmugam, N., Sesterhenn, I. A., Zou, Z., Srikantan, V., Augustus, M., Roschke, V., Carter, K., McLeod, D. G., Moul, J. W., Soppet, D., and Srivastava, S. PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor is over-expressed in prostate cancer. Cancer Res., *60:* 6568–6572, 2000.
36. Zhou, W., Sokoll, L. J., Bruzek, D. D. J., Zhang, L., Velculescu, V. E., Goldin, S. B., Hruban, R. H., Kern, S. E., Hamilton, S. R., Chan, D. W., Vogelstein, B., and Kinzler, K. W. Identifying markers for pancreatic cancer by gene expression analysis. Cancer Epidemiol. Biomark. Prev., *7:* 109–112, 1998.
37. Culine, S., Fabbro, M., Ychou, M., Romieu, G., Cupissol, D., and Pujol, H. Chemotherapy in carcinomas of unknown primary site: a high density intensity policy. Ann. Oncol., *10:* 569–575, 1999.
38. Ramaswamy, S., Tamayo, P., Rifkin, R. M., Mukherjee, S., Yeang, C-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. Multi-class cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, *98:* 15149–15154, 2001.
39. Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. E., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., and Hampton, G. M. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res., *61:* 7388–7393, 2001.
40. Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. USA, *98:* 13784–13789, 2001.
41. Tot, T. Adenocarcinoma metastatic to the liver. Cancer (Phila.), *85:* 171–177, 1999.
42. Semjonow, A., Albrecht, W., Bialk, P., Gerl, A., Lamerz, R., Schmid, H. P., and van Poppel, H. Tumour markers in prostate cancer: EGTM recommendations. Anticancer Res., *19:* 2799–2801, 1999.
43. Klapdor, R., Aronsson, A-C., Duffy, M. J., Hansson, L. O., Khalifa, R., Lamerz, R., Nilsson, O., and Van, D. Tumour markers in gastrointestinal cancers: EGTM recommendations. Anticancer Res., *19:* 2811–2815, 1999.
44. Khubchandani, K. R., and Snyder, J. M. Surfactant protein A (SP-A): the alveolus and beyond. FASEB J., *15:* 59–69, 2001.
45. Zamecnik, J., and Kodet, R. Value of thyroid transcription factor-1 and surfactant apoprotein A in the differential diagnosis of pulmonary carcinomas: a study of 109 cases. Virchows Arch., *440:* 353–361, 2002.

Identification from Public Data of Molecular Markers of Adenocarcinoma Characteristic of the Site of Origin

Jayne L. Dennis, J. Keith Vass, Ernst C. Wit, et al.

Cancer Res 2002;62:5999-6005.