Osteopontin Identified as Lead Marker of Colon Cancer Progression, Using Pooled Sample Expression Profiling

Deepak Agrawal, Tingan Chen, Rosalyn Irby, John Quackenbush, Ann F. Chambers, Marianna Szabo, Alan Cantor, Domenico Coppola, Timothy J. Yeatman

Background: **New tumor markers and markers of tumor progression are needed for improved staging and for better assessment of treatment of many cancers. Gene expression profiling techniques offer the opportunity to discover such markers. We investigated the feasibility of sample pooling strategy in combination with a novel analysis algorithm to identify markers.** ***Methods:*** **Total RNA from human colon tumors (n = 60) of multiple stages (adenomas; cancers with modified Astler Collier stages B, C, and D; and liver metastases) were pooled within stages and compared with pooled normal mucosal specimens (n = 10) by using oligonucleotide expression arrays. Genes that showed consistent increases or decreases in their expression through tumor progression were identified. Northern blot analysis was used to validate the findings. All statistical tests were two-sided.** ***Results:*** **More than 300 candidate tumor markers and more than 100 markers of tumor progression were identified. Northern analysis of 11 candidate tumor markers confirmed the gene expression changes. The gene for the secreted integrin-binding protein osteopontin was most consistently differentially expressed in conjunction with tumor progression. Its potential as a progression marker was validated (Spearman's ρ = 0.903; *P*<.001) with northern blot analysis using RNA from an independent set of 10 normal and 43 tumor samples representing all stages. Moreover, a statistically significant correlation between osteopontin protein expression and advancing tumor stage was identified with the use of 303 additional specimens (human cancer = 185, adenomas = 67, and normal mucosal specimens = 51) (Spearman's ρ = 0.667;** *P*<.001). ***Conclusions:*** **Sample pooling can be a powerful, cost-effective, and rapid means of identifying the most common changes in a gene expression profile. We identified osteopontin as a clinically useful marker of tumor progression by use of gene expression profiling on pooled samples.** **[J Natl Cancer Inst 2002;94:513–21]**

The detection of clinically useful tumor markers whose expression predicts tumor stage or outcome is an important priority in cancer research. The identification of these markers, however, is not a simple biologic problem. Unlike clonal cell cultures, the molecular analysis of human tissue samples necessarily involves heterogeneous cell populations whose messenger RNA (mRNA) composition is proportionally complex. Similarly, the variability in gene expression from one individual tissue sample to another is substantial *(1)* and may obscure common patterns of gene expression that are predictive of clinical outcome. These and

© Oxford University Press

other problems linked to the identification of tumor markers with prognostic significance from biologically complex data sets necessitate the development of novel approaches to identify such markers.

Genome-wide expression profiling is potentially well suited to addressing the multifaceted problems associated with the discovery of clinically useful tumor markers *(2)*. Microarray-based expression profiling has the capacity to evaluate thousands of genes from many different tumor tissues simultaneously in a single experiment. It has become increasingly clear, however, that the value of gene expression data is enhanced when tumors are first microdissected before analysis *(3)*. This process ensures that the majority of the tumor sample submitted for analysis is composed of viable, non-necrotic tumor with limited incorporation of normal adjacent tissues and stroma that could artifactually influence the resultant gene expression profile. Unfortunately, this process leads to a significant reduction in the volume of tumor available for analysis.

Realizing the practical limitations imposed by restricted access to large quantities of individual tumors and the complexity associated with the analysis of large datasets from numerous tumors, we hypothesized that a sample pooling strategy before gene expression profiling might be effective. For a marker to be clinically relevant, it must be notably overexpressed or underexpressed in the majority of the tumor samples of a given histology. For the marker to have prognostic significance, it should also show expression alterations concordant with tumor stage or clinical outcome. We proposed that signals for useful markers (common to the majority of samples) would be positively reinforced if multiple samples were pooled. Contributions from genes that are altered in only a minority of tumor specimens and are, thus, less useful as markers would be minimized. This pooling design differs significantly from the standard approach of examining gene expression profiles of tumor samples individually and then mathematically analyzing the pooled data to derive gene clusters *(4)*. This latter design carries the advantage of identifying the variation of the expression of each gene from one tumor to another; this is a feature that is lost with sample pooling.

The goal of this study was to determine the feasibility of detecting tumor markers and markers of progression in human colon cancer by using gene expression profiling in combination with a novel analysis algorithm, to assess pooled tumor samples. At present, few tumor markers have been identified that have demonstrated clinical utility. Using 70 human normal and tumor tissue samples of various grouped stages, we examined the capacity of Affymetrix oligonucleotide arrays containing both 6800 and 12 000 elements to uncover tumor markers with potential clinical relevance.

SUBJECTS AND METHODS

Normal and Tumor Tissues

Bulk tumors obtained from the tissue procurement facility of the H. Lee Moffitt Cancer Center were histologically confirmed, grossly dissected, and snap frozen in liquid nitrogen within 20 minutes of surgical resection. Tumors for this project were grouped according to stage but selected at random from a large list of banked, frozen tumors.

Tumors that were used to detect candidate tumor markers included 10 modified Astler Collier (AC) stage C microdis-

sected tumors and their paired normal mucosa. An additional 12 AC stage C tumors and their normal mucosal pairs were used to demonstrate that genes identified by sample pooling analyses were also differentially expressed in an independent validation set of tumors.

Sixty tumors of different clinical stage, in addition to 10 normal mucosal specimens, were used for all microarray experiments designed to elucidate tumor progression markers (adenomas without evidence of cancer [n = 10], modified AC stage B1 cancers [partial thickness invasion, lymph node negative; n = 10], AC stage C2 cancers [full thickness invasion, lymph node positive; n = 10], AC stage D cancers [primary tumor metastatic to distant organs; n = 10], and resected liver-metastatic foci [n = 20]). RNA derived from 43 additional tumors was used to perform northern blot analyses using the osteopontin probe. Microdissected (when noted) samples represented bulk frozen samples that were examined using a frozen section technique to identify and eliminate any regions containing normal adjacent tissue, intervening stroma, or necrotic regions. Microdissection was performed with the use of a scalpel tip by a single pathologist (D. Coppola). Although it is not possible to eliminate all intervening stroma and cellular infiltrates without performing laser capture microdissection, these tissues were judged to contain more than 90% tumor cells. For each sample, total RNA was prepared by TRIZOL (Invitrogen, Carlsbad, CA) extraction, quantified, and validated for integrity by gel electrophoresis. Samples were either pooled in equimolar amounts (5–10 samples/pool) or used independently for GeneChip™ (Affymetrix, Santa Clara, CA) hybridization. All tumors were obtained without personal identifiers under an approved Institutional Review Board protocol (No. 5937).

Oligonucleotide Arrays

All of the experiments described in this study were performed with either the first edition Human HuFl 6800 (6800 elements) or the second edition HuU95A (12 000 elements) GeneChip™. The HuFL6800 chip contains probes corresponding to 5000 named genes (based on National Center for Biotechnology Information [NCBI] UniGene build 139, as provided by Affymetrix), whereas the HuU95A contains more than 12 000 probe sets corresponding to 8900 named genes (UniGene Build 139). The reduced number of named genes allows for multiply redundant probe sets for individual genes such as osteopontin. Each GeneChip™ was hybridized using targets synthesized from 10 μg of starting material (total RNA); pooled samples were made by combining either 1 μg (for 10-sample pools) or 2 μg (for five-sample pools) of total RNA from each component tissue sample. Target synthesis, hybridization, and posthybridization staining were performed using standard protocols as recommended by the manufacturer (Affymetrix).

Standard Array Analysis

Stained chips were scanned on a GeneArray™ Scanner (Affymetrix), and data files were processed as summarized below by GeneChip™ software (Affymetrix). Each gene on the chip is assayed by measuring fluorescence intensity resulting from hybridization to 16–20 oligonucleotide probe pairs; each pair consists of a perfect match (PM) complement and a one base mismatched (MM) variant to a gene-specific sequence. GeneChip™ software generates a mean intensity for each gene by first calculating the difference between each of the PM and MM probe

pairs and then averaging these differences across the gene-specific probe set, yielding an average intensity value for each gene. One problem with this algorithm is that it can generate negative expression measures if there is substantial hybridization to one or more of the MM probes when the MM complement is greater than that of the PM.

A New Algorithm for Analysis

As detailed in the Results section below, the presence of negative expression values prompted us to construct a novel algorithm for assessing expression. Under the assumption that greater hybridization to a MM probe than to its PM partner indicates that the particular probe pair is poorly selected, we eliminated from further consideration any probe pair with a negative difference. In addition, to ensure that the estimate of expression was not based on the biased representation of a few probe pairs, we only included genes that had 16 or more positive probe pairs. Although fewer probe pairs can be used, we chose 16 as a conservative limit. This approach assures positive intensity measures for all genes for which we can accurately assess expression, and it eliminates those genes for which the chip does not provide reliable data. Intensity values were averaged across all good probe sets to provide a single measure for each gene.

Identification of Candidate Tumor Markers and Progression Genes

Gene expression as a function of tumor stage was measured using average intensities generated by the standard and modified methods described above. Genes with a negative average intensity value were excluded from the GeneChip™ results obtained by the standard method. Comparisons of gene expression in tumors relative to normal mucosa were calculated as ratios (fold change) of the mean intensity values for each gene. A ratio of 1 is used to represent no difference in gene expression between a tumor and normal mucosa.

Candidate tumor markers were selected by querying the gene expression database, derived from AC stage C2 tumors, for genes whose expression in tumor tissues versus normal mucosa was increased or decreased more than twofold. Only genes that met this criterion for all tumor stages after applying the modified algorithm described above were selected. These thresholds are determined by the limits of the current technology and represent reproducible and reliable cutoff points for selecting genes whose behavior can be validated by other techniques (e.g., northern blot analysis).

Candidate tumor progression markers were selected by the identification of genes that exhibited an overall pattern of progressively increasing expression concordant with advancing tumor stage (such as normal mucosa < adenoma < AC stage B1 < AC stage C2 < AC stage D < liver metastases), recognizing a twofold difference as a minimally acceptable biologically significant change in expression. A similar approach identified those genes that decreased progressively with advancing tumor stage. The expression of all genes on the HuU95A chip is available on a public Web site (http://cancer.tigr.org/data/pooling.shtml).

Validation by Northern Blot Analyses

Potential tumor markers and progression genes were validated by northern blot analysis using tumors distinct from those used in microarray analysis. For northern blot analysis, 10 µg of total RNA was extracted, submitted to gel electrophoresis, blotted, and then hybridized with radiolabeled, gene-specific probes as described. Ethidium bromide was used to stain the gels and to control for the equivalent loading of lanes; alternatively, northern blots were reassessed with radiolabeled GAPDH probes. Expression levels were quantified by densitometry.

Immunohistochemical Analysis

Using stage-oriented human colon cancer paraffin-embedded tissue microarrays (catalog Nos. CR200 and CR50; Clinomics Laboratories, Inc., Frederick, MD) and tissues from the H. Lee Moffitt Cancer Center, we stained 303 tissue samples (185 human cancers, 67 adenomas, and 51 normal mucosal specimens) with hematoxylin and eosin (H & E) (Richard-Allan Scientific, Kalamazoo, MI) using standard histologic techniques. Tissue sections were also subjected to immunostaining for osteopontin with the murine anti-human osteopontin monoclonal antibody mAb53 *(5)*, using the avidin-biotin peroxidase complex technique (Vectastain Elite ABC Kit; Vector Laboratories, Burlingame, CA), following the manufacturer's instructions. We used antibody at a 1 : 750 dilution, after microwave antigen retrieval (four cycles of 5 minutes each on high in 0.1M citrate buffer). The microwave used is an 1100 W Emerson Model AT 736 (Emerson & Cuming Microwave Products, Randolph, MA). The stain was semiquantitatively examined by two independent pathologists using a scale from 0 to III (none, weak, moderate, strong). Statistical correlations were assessed using Spearman's ρ.

RESULTS

Performance of Pooling: Standard Affymetrix Algorithm

To test whether analysis of pooled samples could accurately reflect gene expression in individual tumor samples and to verify that pooled samples provided reproducible data, we surveyed expression in RNA derived from five individual, microdissected human colon tumor samples of the same stage (AC stage C2) with the use of the HuFL6800 GeneChip™. In addition, we measured gene expression for two sets of pooled RNA samples (Pools 1 and 2). One set was derived from the five individual tumors (Pool 1), and a second set was constructed from five additional, independent tumors (Pool 2).

Measured RNA expression levels for each of the tumors were compared gene by gene with those measured for Pool 1 and Pool 2, as well as with a "calculated pool" (Pool C) constructed by averaging gene expression levels across each of the five tumors. In addition, both individual tumor samples and pools were compared with measured expression levels from normal mucosa. The results of these comparisons, shown as scatter plots, are summarized in Fig. 1. In general, the measured expression of genes in the five individual tumor samples correlated extremely well with both those for the calculated pool (Pool C) ($R^2 \geq 0.9371$) and for the corresponding physical pool (Pool 1) ($R^2 \geq 0.8867$). The correlation between the individual samples and the independently derived pool (Pool 2) was nearly as good, and both physical pools correlated extremely well ($R^2 = 0.9309$). Furthermore, all tumor samples correlated much better with each other than with measured expression in normal mucosa. This suggested that pooling samples can, in a single assay, provide results that summarize the expression of individual tu-
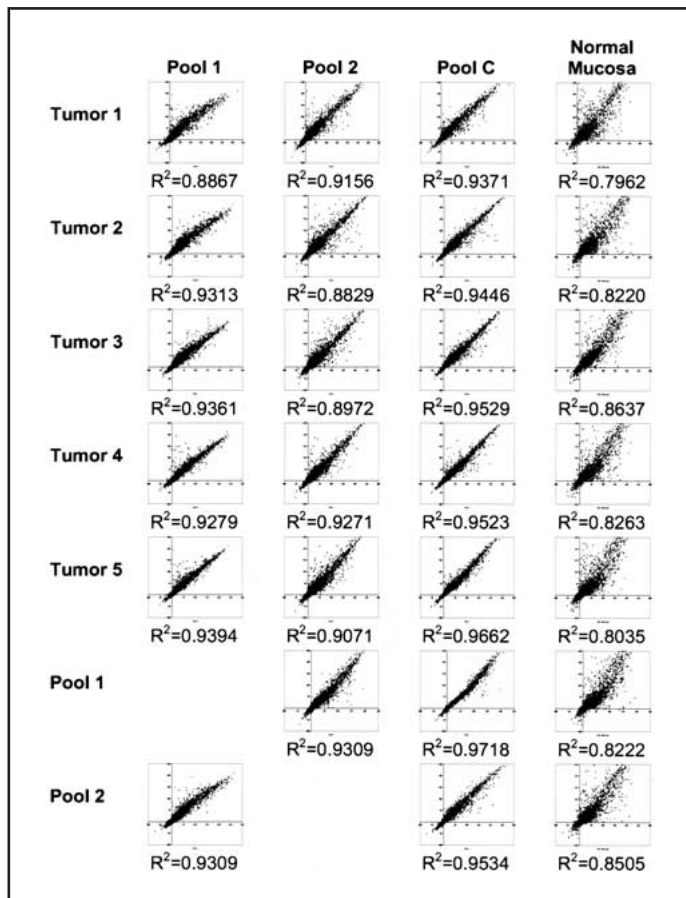
Fig. 1. Scatter plots and squared correlation coefficients were calculated by the use of measured gene expression levels from the Affymetrix Human HuFl 6800 GeneChip™. Gene expression in five individual tumors, a pool containing equimolar quantities of RNA from those five samples (Pool 1), and a second pool constructed from five independent tumors (Pool 2) were compared with each of the pools; a "calculated pool" (Pool C) was computed as an average across the five individual tumors, and expression was measured in normal mucosa. Each scatter plot shows expression in the sample listed in the **left column** (y-axis) plotted against the measured expression for the sample listed directly above in the **top row** (x-axis). Note that the highest correlation is between Pool 1 and Pool C, as expected. Furthermore, the correlation between the individual tumor samples and the pools is much higher than between any of the tumor samples (either individual or pooled) and normal mucosa. These results suggest that the pools maintain patterns of gene expression representative of that in tumors and distinct from normal tissue.

mor samples, and that independently derived pools can provide nearly identical measures of gene expression.

When, however, we compared the genes identified as being the most differentially expressed (defined as genes with expression relative to normal mucosa increased or decreased more than twofold), we discovered that only 38% (400/1042) of genes differentially expressed in the pool were also expressed in the majority (at least three of five) of the individual tumors used to construct the physical pool (Table 1). This might suggest that, although there is good correlation between the pools and individual tumors for the entire gene set, the most differentially expressed genes are not well represented by the pools. Further inspection of the raw gene expression intensity data, however, showed that many of the genes for which there was lack of agreement in expression between the individual tumors and the pools occurred when one or more of the samples exhibited negative values for expression intensity.

516 ARTICLES

Table 1. Effect of a new algorithm for identification of tumor markers using pooled samples

	No. of candidate tumor markers identified	
	Standard analysis	Modified analysis
Differentially expressed genes based on pool*	1042	339
Differentially expressed genes based on individual samples†	2058	77
Agreement based on majority‡	400 (38%)	228 (67%)
Agreement based on t test§	366 (18%)	77 (100%)
κ statistic§	−0.056	0.366

*Number of genes whose expression intensity levels in tumors versus normal mucosa were changed more than twofold for all of the assessed genes.

†A one-sample t test was performed on the five individual tumors to identify important genes using the individual sample data whose mean log (base 2) expression was statistically significantly (P<0.05) greater than 1 or less than −1.

‡Number of genes whose differential expression in the physical pool (Pool 1) was mirrored by a similar change in gene expression in the majority (three or more) of the five individual tumors tested.

§Differentially expressed genes identified by this method were compared with those derived from a more than twofold threshold to derive a κ statistic.

Performance of Pooling: Improved Algorithm

We postulated that we might be able to improve the predictive value of the pool for the most differentially expressed genes by modifying the standard analysis algorithm provided by Affymetrix, which uses all of the probe sets on the array without any assessment of their relative quality. An evaluation of the data across the seven chips used in our pooling analysis revealed that a large number of probe pairs yielded negative values; i.e., the MM probes exhibited greater hybridization intensities than their PM counterparts, suggesting that these probe pairs do not accurately reflect gene expression.

To address this problem, we developed a new algorithm to identify and omit probe pairs that had negative (PM – MM) intensities in the normal mucosal samples. To improve the accuracy of this approach, we eliminated genes from our consideration that had fewer than 16 valid oligonucleotide probe pairs. Although fewer probe pairs can be used, we chose 16 as a conservative limit.

Validation of New Algorithm

Application of the new algorithm to the data processed by the standard algorithm generated a reduced set of informative genes (339), for which the pooled sample now predicted the gene expression of the majority of individual tumors (in at least three of five tumors) for 67% (228) of the genes (Table 1). This represents a substantial improvement in performance of the pool in predicting the gene expression of individual samples by the new algorithm. As might be expected, 77 (100%) of the 77 genes that were differentially expressed in all five of the individual tumors were correctly predicted by the pooled sample. Moreover, 75 (91%) of 82 genes that were predicted to be differentially expressed in four of five individual tumors and 78 (63%) of 123 genes predicted in three of five tumors were also correctly predicted by the pool. Of the 52 genes that were differentially expressed in the majority of individual samples and not correctly predicted by the pool, the mean fold change (increase or decrease) across the individual samples was less than twofold in 51 of 52 cases. This suggests that pooling is less effective for genes

Journal of the National Cancer Institute, Vol. 94, No. 7, April 3, 2002

whose expression is only marginally changed in a subset of the five tumors.

Using the new algorithm, standard statistical analysis using a one-sample, two-sided t test to derive differentially expressed genes from the five individual samples confirms that 77 (100%) of 77 of these genes were also identified in the pool (κ coefficient = 0.366) (Table 1). This represents a considerable improvement over the standard analysis algorithm, in which 366 (18%) of 2058 genes identified as differentially expressed in the individual samples agreed with the pool (κ coefficient = –0.056).

Validation of Sample Pooling

To rigorously validate the results derived from the analysis of pooled samples (training set), we performed northern blot analysis to assess the expression of individual genes in a set of up to 12 unrelated AC stage C2 tumors (validation set) not used to derive the pools. We randomly selected 11 of the top 20 genes from this revised data set of 339 candidate tumor markers for further evaluation in this test. These genes appeared to be overexpressed or underexpressed by more than twofold in the tumor pool when compared with the normal mucosal pool and were also present in at least three of the five tumor samples. Of these selected gene probes, nine of 11 identified gene expression changes concordant with those predicted by the pool in the majority of individual samples tested (Fig. 2). Importantly, our data reveal that sample pooling does not exclude the ability to detect genes that show decreased expression. For example, in addition to the experimentally validated genes tested above, we noted that multiple genes, identified in our analysis of a small number of pooled samples with the HuFl 6800 GeneChip™, were also reported to show decreased expression in an analysis of 18 individual colon adenocarcinomas *(6)* (e.g., guanylin [M97496], DRA [L02785], and tetranectin [X64559]). Collectively, these data demonstrate that sample pooling accurately identifies genes with altered expression in the majority of tumors, allowing identification of potential markers with clinical utility in defining molecular fingerprints of tumors.

Osteopontin: Leading Candidate Marker of Tumor Progression

Our primary goal in this study was the identification of tumor progression markers that might ultimately be used individually or collectively to predict clinical outcome. To that end, we performed two sets of experiments using the pooling strategy with the modified analysis algorithm outlined above. First, five sets (n = 10 tissues/set) of bulk (not microdissected) human colon normal and neoplastic tissues were selected at random from the H. Lee Moffitt Cancer Center Tumor Bank; these tissues represented key stages in tumor progression: normal mucosa, adenomas, AC stage C2 cancers (C2), and two groups of liver metastases (LM1 and LM2, respectively). Total RNA from each set of tissues was extracted and pooled (10 tumors/pool) to derive five pooled RNA samples that were evaluated using the HuFL6800 GeneChip™. Gene expression data for each pooled tissue set relative to normal mucosal samples were calculated, based on the modified algorithm outlined above, to identify genes whose expression increased or decreased sequentially with tumor stage. This analysis identified osteopontin as a leading candidate marker of tumor progression with (LM1 and LM2) > AC stage C2 tumors > adenomas > normal mucosa (Fig. 3, A).
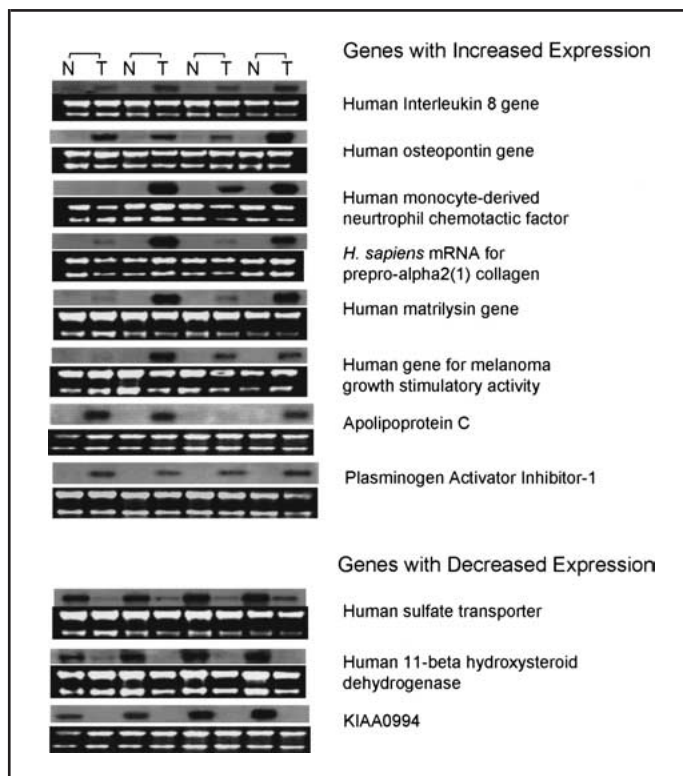
Fig. 2. Northern blot analysis validation of gene expression of a subset of 11 genes randomly selected from the top 20 of more than 330 candidate tumor markers identified using pooled specimen RNA. With the use of radiolabeled probes specific for messages predicted to be overexpressed or underexpressed by microarray analysis, northern blot analyses of numerous Astler Collier (AC) stage C2 colon cancers (T) (distinct from those used in the pool, n = 12) and their paired adjacent normal colonic mucosa (N), were performed (**upper rows**). Gel loading was controlled by monitoring ethidium bromide staining of 28s and 18s ribosomal bands (**lower rows**). Note that for nearly all 11 probes tested, the prediction of the pool matched with the majority (more than two of four samples) of individual tumor samples tested.

In an effort to increase our ability to identify novel tumor markers, we also used the second edition HuU95a GeneChip™, which contains 12 000 elements, to assess extended sets of tumor samples. Six pools of RNA were assembled representing normal mucosa, adenomas, AC stage B1 cancers (B1), AC stage C2 cancers (C2), AC stage D cancers (D), and resected liver metastases (LM), with each pool containing 10 individual microdissected samples from the appropriate tissue. Expression for each pool was assayed as described previously, and fold changes in expression were calculated relative to normal mucosa. Candidate progression markers were selected by identifying genes that increased or decreased in expression concordant with tumor progression. Among these genes, osteopontin again emerged as the leading candidate, consistently showing progression and the highest level of differential expression at each tumor stage (Fig. 3, B). The HuU95a GeneChip™ contained two different probe sets that recognize osteopontin, although probe set 2 produced stronger hybridization. Both probe sets produced similar results that demonstrated increasing osteopontin expression relative to normal mucosa with advancing tumor stage, confirming the results of the HuFL6800 GeneChip™ performed on bulk tissue samples.

Using this approach, we were able to identify 107 genes that displayed both marked tumor progression and a significant dif-
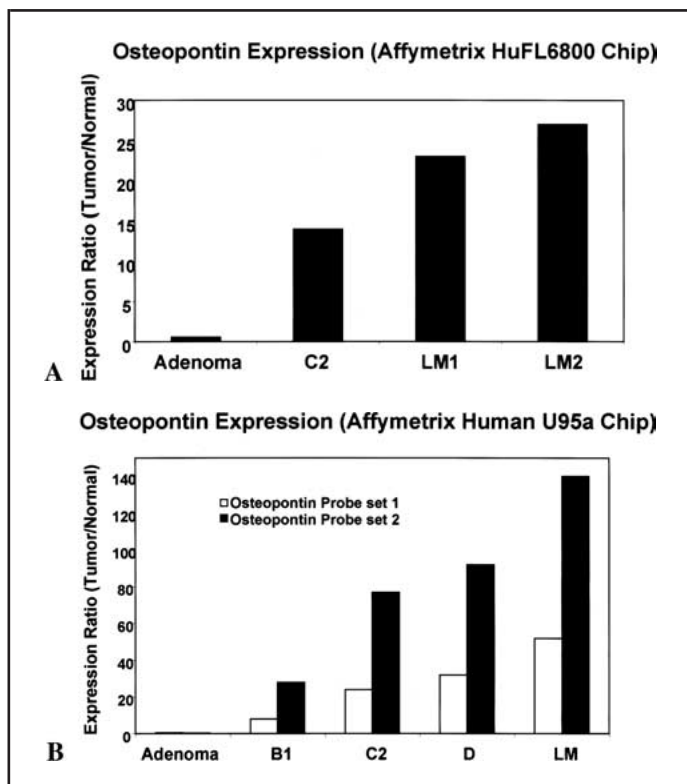
Fig. 3. A) Osteopontin-specific oligonucleotide-based gene expression analysis of pooled, bulk human colon tumor specimens of progressive stage (n = 10/stage). Osteopontin gene expression analysis for adenomas, Astler Collier (AC) stage C2 cancers, and liver metastases (from two separate groups), performed using an Affymetrix HuFl6800 element gene array, shows progressive increases in expression intensity with advancing stage. **B**) Osteopontin gene expression analysis for microdissected tumor specimens derived from adenomas, AC stage B1, AC stage C2, and AC stage D cancers, and liver metastases (n = 10 tumors/stage), performed using a U95a 12 000-element gene array, confirmed and refined observations resulting from the use of the HuFl6800 element array. Expression analysis of two different sets of osteopontin oligonucleotides (probes 1 and 2) confirmed similar osteopontin expression patterns. Probe 2 is more efficient than probe 1 and results in greater hybridization intensities stage for stage.

ferential expression relative to normal mucosa—characteristics favoring a clinically useful tumor marker. Hierarchical clustering analysis (average linkage using a Euclidean distance metric) *(7)* of these 107 genes shows that multiple copies of the same gene cluster together (two probe sets recognizing osteopontin top the list), as do tumors of related stage, and that invasive tumors cluster away from adenomas (Fig. 4).

To validate the tumor-stage-dependent increase in osteopontin expression, we performed northern blot analyses on numerous, randomly selected human colon cancer primary and metastatic specimens (n = 43) not used in the microarray analyses. The analysis demonstrates a clear correlation (Spearman's ρ = 0.903; *P*<.001) between increasing osteopontin fold expression relative to normal mucosa (mean ± SD) and tumor stage (Fig. 5). Representative northern blot analyses of individual tumors show the same result and further demonstrate that, even within the same patient, there is a progressive increase in osteopontin expression in a cancer that evolved from an adjacent adenoma (Fig. 6, A–D).

To demonstrate that mRNA expression was predictive of osteopontin protein expression, 303 paraffin-embedded, archival

normal and tumor samples representing the full range of clinical stages for colon cancer were stained with anti-osteopontin monoclonal antibody, using immunohistochemical techniques (Fig. 6, E). A highly significant correlation (*P*<.001) was demonstrated for osteopontin protein expression when cancers and adenomas were compared with normal mucosal specimens. In addition, a significant correlation (Spearman's ρ = 0.667; *P*<.001) between the degree of osteopontin protein expression and advancing AC stage was identified (Table 2).

DISCUSSION

At present, there are few tumor markers that have clinical utility in the management of colon cancer. Even the application of the most widely used marker, carcinoembryonic antigen, has been recently called into question *(8)*. For these reasons, the identification of candidate tumor markers that can be used to predict outcome or to derive biologic insight regarding the mechanisms underlying tumor progression would be valuable.

Microarray analyses are well suited for identifying tumor markers. There is promise that these analyses may be capable of identifying molecular fingerprints that predict outcome, independent of current staging systems. However, because our primary goal was to identify individual tumor markers with potential clinical utility, we chose to discover markers that were correlated with AC stage, the current gold standard for determining colorectal cancer prognosis. Detection of markers reproducing current staging systems, although valuable individually, might also prove integral to molecular signatures derived independently of stage.

Although a detailed analysis of numerous individual tumor specimens is generally ideal because it permits the assessment of gene expression variability from sample to sample, our results strongly suggest that sample pooling is an effective alternative strategy. This is particularly true when the goal is to rapidly identify tumor markers that are expressed by the majority of tumors in a population. The inherent benefits of sample pooling are multiple. By pooling tissue specimens, the RNA requirement per tumor is proportionately reduced, making sample pooling attractive when tissue banks can provide only limited quantities of microdissected sample per patient. The risk of a single specimen contributing bias to the pool is also proportionately reduced with increasing sample size. And finally, because the number of microarray chips necessary for the study is reduced, the computational requirements for analysis are reduced as well.

Although standard Affymetrix algorithms demonstrated that pooled data correlated highly with data derived from individual tumors, these correlations were artificially elevated by the inclusion of many genes whose expression was low (approximately onefold differences) or not substantially different from the normal mucosal specimens. Unfortunately, these genes lacking differential expression are not prime candidates for clinically useful tumor markers. In contrast, genes that are differentially expressed are of great interest. Standard Affymetrix algorithms were relatively ineffective in identifying genes in the pool that predicted similar expression in individual tumors. Because standard Affymetrix algorithms appeared to incorporate all of the microarray chip data, including probe sets that demonstrate greater hybridization for the MM probes than for the PM probes, we postulated that a modified algorithm eliminating negative values for (PM – MM) might improve our capacity to identify tumor markers in pooled samples. Because the probes on these

Adenoma | Dukes B1 | Dukes C2 | Dukes D | Mets

Probe_id	GenBank	Name
2092_s_at	J04765	HUMOSTRO Human osteopontin mRNA, complete cds
34342_s_at	AF052124	Homo sapiens clone 23810 osteopontin mRNA, complete cds
35315_at	X02544	Human mRNA for alpha1-acid glycoprotein (orosomucoid)
1693_s_at	D11139	HUMTIMP Human gene for tissue inhibitor of metalloproteinases, partial sequence
39959_at	AL031983	dJ271M21.6 (Diubiquitin)
311_s_at		Fibronectin, Alt. Splice 1
31720_s_at	M10905	Human cellular fibronectin mRNA
32305_at	J03464	Human collagen alpha-2 type I mRNA, complete cds, clone pHCOL2A1
32306_g_at	J03464	Human collagen alpha-2 type I mRNA, complete cds, clone pHCOL2A1
37200_at	J04162	Human leukocyte IgG receptor (Fc-gamma-R) mRNA, complete cds
32186_at	M80244	Human E16 mRNA, complete cds
38126_at	J04599	Human hPGI mRNA encoding bone small proteoglycan I (biglycan), complete cds
38111_at	X15998	H.sapiens mRNA for the chondroitin sulphate proteoglycan versican, V1 splice-variant; precursor peptide
38112_g_at	X15998	H.sapiens mRNA for the chondroitin sulphate proteoglycan versican, V1 splice-variant; precursor peptide
671_at	J03040	HUMSPARC Human SPARC/osteonectin mRNA, complete cds
32307_at	V00503	Human mRNA encoding Pro-alpha-2 chain of type I procollagen. (major part)
34703_f_at	AA151971	zo30b03.r1 Homo sapiens cDNA, 5 end /clone=IMAGE-588365 /clone_end=5
1385_at	M77349	HUMTGFBIG Human transforming growth factor-beta induced gene product (BIGH3) mRNA, complete cds
863_g_at	U04313	HSU04313 Human maspin mRNA, complete cds
408_at	X54489	HSMGSAG Human gene for melanoma growth stimulatory activity (MGSA)
37187_at	M36820	Human cytokine (GRO-beta) mRNA, complete cds
32821_at	AI762213	wi54d04.x1 Homo sapiens cDNA, 3 end /clone=IMAGE-2394055 /clone_end=3
2036_s_at	M59040	HUMCD44B Human cell adhesion molecule (CD44) mRNA, complete cds
1126_s_at	L05424	HUMSCG19 Human cell surface glycoprotein CD44 (CD44) gene, 3 end of long tailed isoform
31472_s_at	AF098641	Homo sapiens CD44 isoform RC (CD44) mRNA, complete cds
434_at	X03473	HSHIS10G Human gene for histone H1(0)
37209_g_at	AJ001612	Homo sapiens mRNA for L-3-phosphoserine-phosphatase homologue
36781_at	X01683	Human mRNA for alpha 1-antitrypsin
1664_at		Insulin-Like Growth Factor 2
38138_at	D38583	Human mRNA for calgizzarin, complete cds
37724_at	V00568	Human mRNA encoding the c-myc oncogene
40790_at	AB004066	Homo sapiens mRNA for DEC1, complete cds
35255_at	AF098799	Homo sapiens RanBP7/importin 7 mRNA, complete cds
1519_at	J04102	HUMETS2A Human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA, complete cds
38805_at	X89750	H.sapiens mRNA for TGIF protein
33436_at	Z46629	Homo sapiens SOX9 mRNA
38763_at	L29254	Human (clone P1-5) L-iditol-2 dehydrogenase gene
34873_at	Y16241	Homo sapiens mRNA for nebulette
40385_at	U64197	Homo sapiens chemokine exodus-1 mRNA, complete cds
34319_at	AA131149	zo16d05.r1 Homo sapiens cDNA, 5 end /clone=IMAGE-587049 /clone_end=5
39695_at	M3156	Human decay-accelerating factor mRNA, complete cds
35981_at	D17291	Human gene for regenerating protein I beta, complete cds
439_at	X07876	HSIRP Human mRNA for irp protein (int-1 related protein)
38178_at	L40802	Homo sapiens 17 beta-hydroxysteroid dehydrogenase (17-HSD) gene
40739_at	M83670	Human carbonic anhydrase IV mRNA, complete cds
33512_at	Z70295	H.sapiens GCAP-II gene
38549_at	AF026941	Homo sapiens cig5 mRNA, partial sequence
39726_at	J04040	Human glucagon mRNA, complete cds
38848_at	AI732905	of72a12.x5 Homo sapiens cDNA, 3 end /clone=IMAGE-1435870 /clone_end=3
928_at	L02785	HUMDRA Homo sapiens colon mucosa-associated (DRA) mRNA, complete cds
33611_g_at	AL049977	Homo sapiens mRNA; cDNA DKFZp564C122 (from clone DKFZp564C122)
33651_at	AB013456	Homo sapiens hAQP8 mRNA for aquaporin 8, complete cds
334_s_at		Vasoactive Intestinal Peptide
34637_f_at	M12963	Human class I alcohol dehydrogenase (ADH1) alpha subunit mRNA, complete cds
33069_f_at	U06641	Human UDP glucuronosyltransferase mRNA, partial cds
39385_at	M22324	Human aminopeptidase N/CD13 mRNA encoding aminopeptidase N, complete cds
34572_at	M76424	H.sapiens carbonic anhydrase VII (CA VII) gene
36067_at	AB000887	Homo sapiens mRNA for EBI1-ligand chemokine, complete cds
40095_at	J03037	Human carbonic anhydrase II mRNA, complete cds
35265_at	AF007216	Homo sapiens sodium bicarbonate cotransporter (HNBC1) mRNA, complete cds
33007_at	AC002302	Homo sapiens Chromosome 16 BAC clone CIT987-SKA-345G4 ~complete genomic sequence
32110_at	AB011095	Homo sapiens mRNA for KIAA0523 protein, partial cds
39637_at	U14528	Human sulfate transporter (DTD) mRNA, complete cds
33068_f_at	U08854	Human UDP glucuronosyltransferase precursor (UGT2B15) mRNA, complete cds
37482_at	U37100	Homo sapiens aldose reductase-like peptide mRNA, complete cds
31623_f_at	K01383	Human metallothionein-I-A gene, complete coding sequence
39594_f_at	R93527	yq35f10.r1 Homo sapiens cDNA, 5 end /clone=IMAGE-197803 /clone_end=5
926_at	J03910	HUMMT2A Human (clone 14VS) metallothionein-IG (MT1G) gene, complete cds
33491_at	X63597	H.sapiens si mRNA for sucrase-isomaltase
35717_at	AB020629	Homo sapiens mRNA for KIAA0822 protein, complete cds
38519_at	U68233	Human farnesol receptor HRR-1 (HRR-1) mRNA, complete cds
34002_at	M77144	Human type II 3-beta hydroxysteroid dehydrogenase/5-delta -4-delta isomerase gene, complete cds
33062_at	U95301	Human calcium-dependent group X phospholipase A2 mRNA, complete cds
34974_at	Y13323	Homo sapiens mRNA for disintegrin-protease
35224_at	AF070569	Homo sapiens clone 24659 mRNA sequence
33462_at	D13626	Human mRNA for KIAA0001 gene, complete cds
38407_r_at	AI207842	ao89h09.x1 Homo sapiens cDNA, 3 end /clone=IMAGE-1953089 /clone_end=3
39681_at	AF060568	Homo sapiens promyelocytic leukemia zinc finger protein (PLZF) gene, complete cds
35449_at	U11276	Human hNKR-P1a protein (NKR-P1A) mRNA, complete cds
38587_at	M18079	Human, intestinal fatty acid binding protein gene, complete cds, and an Alu repetitive element
31744_at	D89479	Homo sapiens mRNA for ST1B2, complete cds
37821_at	AF041260	Homo sapiens AIBC1 (AIBC1) mRNA, complete cds
38151_at	AF002672	Homo sapiens breast cancer suppressor candidate 1 (bcsc-1) mRNA, complete cds
37123_at	D88154	Homo sapiens mRNA for villin-like protein, complete cds
35974_at	U10485	Human lymphoid-restricted membrane protein (Jaw1) mRNA, complete cds
34038_at	U76343	Human liver GABA transport protein mRNA, 3 end
36454_at	AF037335	Homo sapiens carbonic anhydrase precursor (CA 12) mRNA, complete cds
35275_at	AL050025	Homo sapiens mRNA; cDNA DKFZp564D066 (from clone DKFZp564D066)
36066_at	AB020635	Homo sapiens mRNA for KIAA0828 protein, partial cds
33382_at	M92449	Human LTR mRNA, 3 end of coding region and 3 flank
34775_at	AF065388	Homo sapiens tetraspan NET-1 mRNA, complete cds
40775_at	AL021786	Human DNA sequence from PAC 696H22 on chromosome Xq21.1-21.2.
39697_at	U26726	Human 11-beta-hydroxysteroid dehydrogenase type 2 mRNA, complete cds
33613_at	AA806239	oc21e02.s1 Homo sapiens cDNA /clone=IMAGE-1341530
39087_at	U28249	Human 11kd protein mRNA, complete cds
40666_at	AF039918	Homo sapiens CD39L4 (CD39L4) mRNA, complete cds
35214_at	AF061016	Homo sapiens UDP-glucose dehydrogenase (UGDH) mRNA, complete cds
39510_r_at	AL049932	Homo sapiens mRNA; cDNA DKFZp564H2416 (from clone DKFZp564H2416)
40330_at	AL050031	Homo sapiens mRNA; cDNA DKFZp566E0224 (from clone DKFZp566E0224)
33709_at	AF067224	Homo sapiens cGMP phosphodiesterase A2 (PDE9A) mRNA, complete cds
38717_at	AL050159	Homo sapiens mRNA; cDNA DKFZp586A0522 (from clone DKFZp586A0522)
35674_at	AB023211	Homo sapiens mRNA for KIAA0994 protein, partial cds
33672_at	U59209	Homo sapiens C19steroid specific UDP-glucuronosyltransferase mRNA, complete cds
33672_r_at	U59209	Homo sapiens C19steroid specific UDP-glucuronosyltransferase mRNA, complete cds
34612_at	AL022315	dJ1177I5.3 (Lectin, Galactose-binding, soluble, 2 (Galectin 2, S-Lac Lectin 2, HL14))
36247_f_at	M12272	Homo sapiens alcohol dehydrogenase class I gamma subunit (ADH3) mRNA, complete cds
32392_s_at	M57951	Human bilirubin UDP-glucuronosyltransferase isozyme 2 mRNA, complete cds

Fig. 4. Hierarchical clustering analysis of 107 genes selected from the 12 000 gene set, which show progressive increases or decreases in expression with stage, permits the rapid identification of both tumor-specific markers and markers predictive of stage progression. Those genes showing progressive fold increases or decreases in gene expression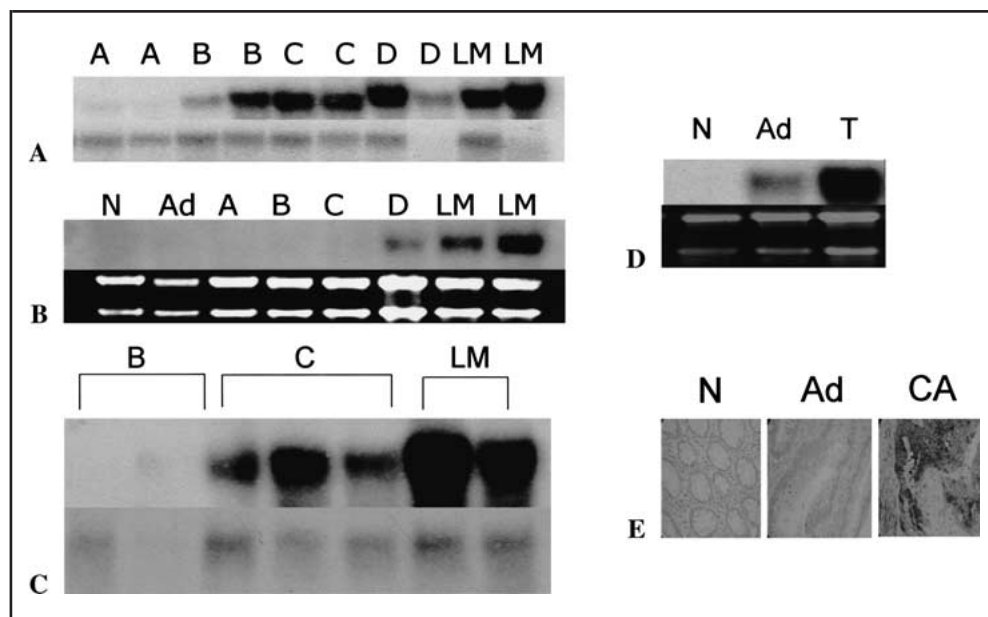 relative to normal mucosa are shown proportionally in **red** and **green,** respectively. This analysis also demonstrates a relationship between multiple copies of the same gene, which cluster together, as do tumors of related stage; conversely, genes expressed in invasive tumors cluster away from noncancerous adenomas. Osteopontin, represented by two probe sets, is the leading candidate progression marker on this list.
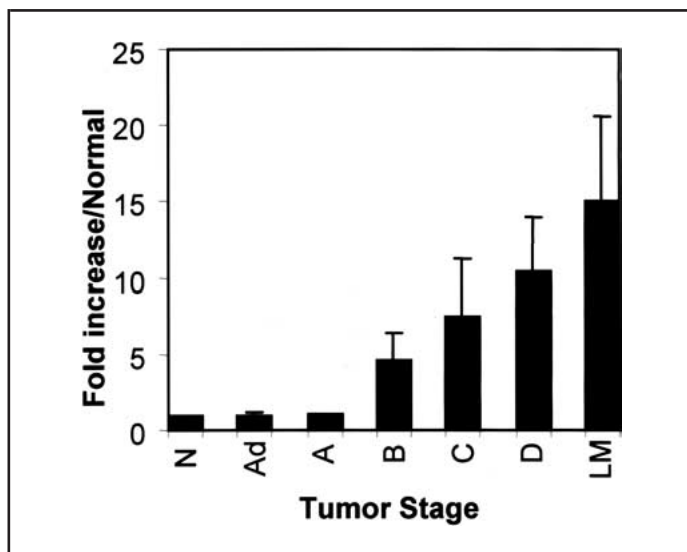
Fig. 5. Composite northern blot analysis of osteopontin messenger RNA expression levels (fold increases relative to normal mucosa, mean ± SD) in 43 human colon tumors derived from numerous modified Astler Collier (AC) stages. Human tumors used in these analyses were different from those used to perform the microarray analyses. Sample sizes were as follows: normal mucosa (N) (n = 8), adenomas (Ad) (n = 3), AC stage A cancers (A) (n = 3), AC stage B cancers (B) (n = 7), AC stage C cancers (C) (n = 13), AC stage D cancers (D) (n = 7), and liver metastases (LM) (n = 10). Osteopontin gene expression increases concordant with advancing tumor stage (Spearman value = 0.903, $P<.001$).

chips consist of relatively short oligonucleotides, a significant level of hybridization might occur with closely related sequences that could contribute substantial intensity and, thereby, interfere with measurements for any gene-specific probe. Although other explanations cannot be excluded, we determined that elimination of these probe sets from calculations of gene expression provided an improved dataset with a greater capacity to predict the expression of individual tumors from a pool of the same tumors.

By applying a pooling strategy, we were able to identify a large set (339) of candidate tumor markers. These genes displayed overexpression or underexpression by twofold or more relative to normal colonic mucosa in the majority of tumor samples tested. From the top 20 genes on the list, 11 different genes were selected at random and validated by northern blot analyses of independent tumors (not used to construct the sample pools). This list of genes includes genes associated with tumor invasion (matrilysin), adhesion (prepro-α2[1] collagen), and possible tumor growth (human gene for melanoma growth-stimulatory activity).

Of further interest was the derived set of genes linked to tumor progression. These genes were derived from five new pooled sets of tumors from the same pathologic stage. We were able to identify 107 candidate tumor progression markers. These were analyzed using the University of California San Diego HAPI software *(9)* (http://array.ucsd.edu/hapi/), which links the gene identities to literature citations. Of these identities, 25 (23%) had previously been linked to the digestive system, with 49 (46%) being linked to neoplasms and 14 (13%) showing direct association with digestive disease neoplasia.

Among the progression genes, osteopontin was identified as the leading candidate, validating the results from our initial analysis that identified tumor markers. Osteopontin is a secreted, integrin-binding protein that has already been reported as a marker of tumor progression in breast *(5,10),* lung *(11),* and prostate cancer *(12).* The results presented here provide the first data to suggest that osteopontin is a strong marker of colon cancer progression. Our results indicated that as colon tumors progress from normal mucosa and the adenoma stage, where osteopontin expression is not easily detectable, to invasive cancers (some AC stage B, most AC stage C), to metastatic primary cancers (AC stage D), and to resected liver metastases, they acquire sequentially increased osteopontin expression. In particular, osteopontin induction was most notable in liver metastases where fold increases were as high as 10- to 20-fold over adenomas and normal mucosal samples. These mRNA results

Fig. 6. Osteopontin RNA and protein expression increases with advancing tumor stage. **A–C)** Representative northern blots showing that osteopontin expression increases concordant with advancing tumor stage in the majority of evaluable tumors: normal mucosa (N) < adenoma (Ad) < Astler Collier (AC) stage A cancer (A) < AC stage B cancer (B) < AC stage C cancer (C) < AC stage D cancer (D) < liver metastases (LM). Tumors were derived from 24 different individuals. **Upper panels** represent osteopontin expression; **lower panels** represent glyceraldehyde-3-phosphate dehydrogenase (GAPDH) expression as loading control. **D)** Osteopontin expression in an AC stage C cancer is substantially higher than the adjacent adenoma from which it evolved in the same individual. Ethidium bromide stain of ribosomal bands (**lower panel**) was the loading control. **E)** Representative immunohistochemical analysis demonstrates that osteopontin protein expression is cytoplasmic and that staining in normal mucosa and adenoma is substantially less than that of invasive cancer.

Table 2. Correlation between osteopontin protein expression determined by immunohistochemical staining and tumor progression

	Stain (+)	Stain (−)
Cancer*	176 (95.1%)	9 (4.9%)
Adenoma	27 (40.3%)	40 (59.7%)
Normal intestinal mucosa	4 (7.8%)	47 (92.2%)

	No. of tissues stained			
Stain intensity (0–III)	0	I	II	III
Normal intestinal mucosa†	47	3	0	1
Adenoma	28	6	4	3
AC‡ stage A	5	24	18	16
AC stage B	3	16	29	16
AC stage C	0	1	16	8
AC stage D	1	5	13	14

*Spearman's ρ = 0.752, $P<.001$.

†Spearman's ρ = 0.667, $P<.001$.

‡AC = Astler Collier.

were further validated with immunohistochemical studies of 303 tumor specimens. Osteopontin has been shown to bind to cells via integrins (notably $\alpha v \beta 3$, $\alpha v \beta 1$, and $\alpha v \beta 3$) as well as CD44 [for review, *see (13)*]. Although the biologic functions of osteopontin are not fully understood, it has been implicated in malignancy, immune function, and vascular remodeling, as well as in bone remodeling [for reviews, *see (13–18)*]. Osteopontin was identified as a tumor-associated protein in transformed cells in culture *(15)* and has been shown to be present in some human tumor samples *(19)*. In breast cancer, osteopontin has been shown to contribute functionally to the malignant behavior of the cells *(20)*. Osteopontin has been noted to be present in some colon cancers *(19)*, and a recent report has identified osteopontin overexpression using Affymetrix technology on individually analyzed bulk colon cancers (n = 18) when compared with normal tissues *(6)*. Our study, however, is the first to demonstrate an association of increasing osteopontin expression with tumor progression in colon cancer, making it a candidate tumor marker with potential clinical utility.

Although this study addressed the feasibility of sample pooling, further studies are needed to determine the optimal number of samples needed to construct an informative pool. We have demonstrated that when five to 10 samples are used to construct a pool, informative data can be mined with regard to the identification of both tumor markers and progression markers. Although the identification of individual progression markers is one potential application of sample pooling, it can also be used to rapidly identify larger patterns of tumor markers that may be predictive of clinical outcome. Furthermore, these observations have implications for the design of many microarray-based experiments on both cell lines and tissues.

REFERENCES

(1) Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet 2000;1:48–56.

(2) DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–60.

(3) Gillespie JW, Ahram M, Best CJ, Swalwell JI, Krizman DB, Petricoin EF, et al. The role of tissue microdissection in cancer research. Cancer J 2001; 7:32–9.

(4) Brenton JD, Aparicio SA, Caldas C. Molecular profiling of breast cancer: portraits but not physiognomy. Breast Cancer Res 2001;3:77–80.

(5) Tuck AB, O'Malley FP, Singhal H, Harris JF, Tonkin KS, Kerkvliet N, et al. Osteopontin expression in a group of lymph node negative breast cancer patients. Int J Cancer 1998;79:502–8.

(6) Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 2001;61:3124–30.

(7) Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.

(8) Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen C. An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected colon cancer. JAMA 1993;270:943–7.

(9) Masys DR, Welsh JB, Lynn Fink J, Gribskov M, Klacansky I, Corbeil J. Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics 2001;17:319–26.

(10) Tuck AB, O'Malley FP, Singhal H, Tonkin KS, Harris JF, Bautista D, et al. Osteopontin and p53 expression are associated with tumor progression in a case of synchronous, bilateral, invasive mammary carcinomas. Arch Pathol Lab Med 1997;121:578–84.

(11) Chambers AF, Wilson SM, Kerkvliet N, O'Malley FP, Harris JF, Casson AG. Osteopontin expression in lung cancer. Lung Cancer 1996;15:311–23.

(12) Thalmann GN, Sikes RA, Devoll RE, Kiefer JA, Markwalder R, Klima I, et al. Osteopontin: possible role in prostate cancer progression. Clin Cancer Res 1999;5:2271–7.

(13) Sodek J, Ganss B, McKee MD. Osteopontin. Crit Rev Oral Biol Med 2000;11:279–303.

(14) Denhardt DT, Giachelli CM, Rittling SR. Role of osteopontin in cellular signaling and toxicant injury. Annu Rev Pharmacol Toxicol 2001;41: 723–49.

(15) Senger DR, Perruzzi CA, Papadopoulos A. Elevated expression of secreted phosphoprotein I (osteopontin, 2ar) as a consequence of neoplastic transformation. Anticancer Res 1989;9:1291–9.

(16) Oates AJ, Barraclough R, Rudland PS. The role of osteopontin in tumorigenesis and metastasis. Invasion Metastasis 1997;17:1–15.

(17) Patarca R, Saavedra RA, Cantor H. Molecular and cellular basis of genetic resistance to bacterial infection: the role of the early T-lymphocyte activation-1/osteopontin gene. Crit Rev Immunol 1993;13:225–46.

(18) Furger KA, Menon RK, Tuck AB, Bramwell VHC, Chambers AF. The functional and clinical roles of osteopontin in cancer and metastasis. Curr Mol Med 2001;1:621–32.

(19) Brown LF, Papadopoulos-Sergiou A, Berse B, Manseau EJ, Tognazzi K, Perruzzi CA, et al. Osteopontin expression and distribution in human carcinomas. Am J Pathol 1994;145:610–23.

(20) Tuck AB, Arsenault DM, O'Malley FP, Hota C, Ling MC, Wilson SM, et al. Osteopontin induces increased invasiveness and plasminogen activator expression of human mammary epithelial cells. Oncogene 1999;18: 4237–46.

NOTES