[CANCER RESEARCH 61, 3124–3130, April 1, 2001]

Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays

Daniel A. Notterman,[1] Uri Alon,[2] Alexander J. Sierk, and Arnold J. Levine[3]

Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

ABSTRACT

Using an oligonucleotide array containing sequences complementary to ~3200 full-length human cDNAs and 3400 expressed sequence tags (GeneChip, Affymetrix), mRNA expression patterns were probed in 18 colon adenocarcinomas and 4 adenomas. Paired normal tissue was available and analyzed for each of the tumors. Relatively few changes in transcript expression are associated with colon cancer. Nineteen transcripts (0.48% of those detected) had at least 4–10.5-fold higher mRNA expression in carcinoma compared with paired normal samples, whereas 47 transcripts (1.3% of those detected) had at least 4–38-fold or lower expression in the tumor tissue compared with the normal samples. Some of these differences were confirmed by reverse transcription-PCR. Many of these transcripts were already known to be abnormally expressed in neoplastic tissue in general, or colon cancer in particular, and several of these differences were also observed in premalignant adenoma samples. A two-way hierarchical clustering algorithm successfully distinguished adenoma from adenocarcinoma and normal tissue, generating a phylogenetic tree that appropriately represented the clinical relationship between the three tissue types included in the analysis. This supports the concept that genome-wide expression profiling may permit a molecular classification of solid tumors.

INTRODUCTION

The majority of mutations found in tumor cells such as colon carcinoma occur in signal transduction pathways that ultimately regulate transcription factors and therefore a large number of genes and their transcription patterns. It is thus not surprising that abnormalities in gene expression are characteristic of neoplastic tissue (1). Traditional methods of analysis have imposed a practical limit on the number of candidate genes, the expression of which can be conveniently and simultaneously studied. Highly parallel technologies exploiting sample hybridization to oligonucleotide or cDNA arrays permit the expression levels of tens of thousands of genes to be monitored simultaneously and rapidly (2).

Adenocarcinoma of the colon is a well-characterized model of human cancer in which histological progression of tumors has been correlated with specific sequential genetic mutations (3). The tumor is common, and the surgical technique of resection results in both normal and neoplastic tissue that is available for study. In this report, the use of an oligonucleotide array (GeneChip, Affymetrix) to monitor the expression profiles of colon adenocarcinomas, adenomas, and normal colon tissue is described. Of the transcripts represented by this array, 6003 represent unique GenBank accession numbers, about evenly divided between full-length human cDNAs and ESTs[4] that are similar to other known eukaryotic genes. We have described previously the use of clustering analysis to map global differences in gene expression in normal and carcinoma samples in a subset of genes and ESTs (2000 of approximately 6600 genes and ESTs, including some duplicates) contained in a colon cancer database (4). Here, we have combined expression profiles from an additional eight samples (four adenoma samples paired with four normal samples) with a subset of the samples examined previously (only 18 paired adenocarcinoma samples, characterized with respect to tumor cell percentage) and included an analysis of 4000 genes and ESTs with an average expression level of 10 or greater. This permits us to test, in paired normal/neoplastic tissue, whether genome-wide expression monitoring and clustering techniques can differentiate benign from malignant colorectal tumors and to learn how the progression from normal tissue through adenoma and adenocarcinoma is mirrored in changes in gene expression. Our findings indicate that expression of ~1.8% of the genes detected by the array differ between adenocarcinoma and normal colon tissue, and that many of the expression changes noted in the cancers were presaged by those in the adenomas. Furthermore, a clustering algorithm successfully distinguished normal tissue, adenoma, and adenocarcinoma.

MATERIALS AND METHODS

Tissue Samples. Samples of colon adenocarcinoma, adenoma, and paired normal tissue (generally full-thickness colon) from the same patient were obtained from the Cooperative Human Tissue Network. The tissue was snap-frozen in liquid nitrogen within 20–30 min of harvesting and stored thereafter at −80°C. mRNA was extracted from the bulk tissue samples and hybridized to the array as described previously (4). The clinical stage was estimated from accompanying surgical pathology and clinical reports using the Modified Astler-Collier (MAC) system: 8 samples were MAC B, 5 samples MAC C, and 5 samples MAC D. Twelve samples were judged moderately well differentiated, 2 were well differentiated, and 3 were poorly differentiated. The grade was not provided for one sample. For the adenocarcinomas, samples were specifically re-reviewed by a pathologist at the institution from which the sample was contributed, using paraffin-embedded tissue that was adjacent or in close proximity to the frozen sample from which the RNA was extracted. The histological characteristics of the tumor samples, the estimated percentage of contamination with nontumor cells, the presence of mutations in the *p53* gene, and the clinical disease stage are included in the supplemental data tables.[5]

Oligonucleotide Array. The experiments with adenocarcinoma and paired normal tissue were performed with the Human 6500 GeneChip Set (Affymetrix), and the experiments with the adenomas and their paired normal tissue were performed with the Human 6800 GeneChip Set (Affymetrix). Gene intensity information was converted to a mean intensity for each gene by proprietary software (Affymetrix), which includes routines for filtering and centering the data (in these experiments, to 50 intensity units). Expression of genes related to smooth muscle and connective tissue was consistently greater in the normal than the tumor samples, probably because of the greater heterogeneity of tissue type in the normal samples (4). These transcripts were identified by inspecting the gene description in GenBank ($n = 41$) and are not included in Table 2, although they are included in the complete list of transcripts and their associated expression intensities detected in this experiment.[5]

Statistical and Database Methods. Intensity information was exported to Excel or Access (Microsoft, Redmond, CA) as needed. Statistical analysis was performed with Statistica (Stat-Soft, Tulsa, OK). To avoid division by 0 or a

Received 4/11/00; accepted 1/26/01.

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] To whom requests for reprints should be addressed, at Department of Molecular Biology, Princeton University, Princeton University, NJ 08544. Fax: (609) 258-4575; E-mail: Dnotterman@Princeton.edu.

[2] Present address: Weizmann Institute of Science, Rehovot 76100, Israel.

[3] Present address: Rockefeller University, New York, New York 10021.

[4] The abbreviations used are: EST, expressed sequence tag; MGSA, melanoma growth stimulatory activity; SAGE, serial analysis of gene expression.

[5] Internet address: http://microarray.princeton.edu/oncology/.

Table 1 *Transcripts more highly expressed in adenocarcinoma than in paired normal tissue*

Intensity values <10 are adjusted to 10. Only transcripts with a 4-fold difference or greater ($P < 0.001$) in expression intensity between tumor and normal are included. Transcripts shown in bold capital letters were confirmed by RT-PCR. Gene descriptions have been edited.

Accession no.	Description	Intensity in tumor	Intensity in normal	Tumor/ Normal
X54489	**Human gene for MGSA**	105.1	10.0	10.5
U22055	**Human 100 kDA coactivator mRNA, complete cds**[a]	72.9	10.0	7.3
D14657	Human mRNA for *KIAA0101* gene, complete cds	64.8	10.0	6.5
M61832	Human *S*-adenosylhomocysteine hydrolase (AHCY) mRNA, complete cds	123.1	20.7	6.0
M77836	Human pyrroline 5-carboxylate reductase mRNA, complete cds	95.5	17.9	5.3
D21262	Human mRNA for *KIAA0035* gene, partial cds [? nucleolar phosphoprotein]	55.6	10.8	5.2
M36821	Human cytokine (GRO-γ) mRNA, complete cds	141.2	27.6	5.1
L23808	**Human metalloproteinase (HME) mRNA, complete cds**	71.1	14.0	5.1
R08183	Similar to bovin hs 10-kD protein 1(chaperonin 10)(HSPE1)(NM_002157)	439.4	91.1	4.8
L29254	Human (clone D21-1) L-iditol-2 dehydrogenase gene, exon 9, and complete cds	47.2	10.0	4.7
H50438	**M-phase inducer phosphatase 2 (*Homo sapiens*)**	46.8	10.0	4.7
U33286	Human chromosome segregation gene homolog CAS mRNA, complete cds	98.8	21.2	4.7
X54942	**H. sapiens CKSHS2 mRNA for CKS1 protein homologue**	131.9	30.1	4.4
R32511	*H. sapiens* cDNA clone 135395 3′ [RNA POL II subunit]	43.5	10.0	4.3
T87871	*H. sapiens* cDNA clone 115765 3′ [myoblast cell surface antigen 24.1 DS]	42.1	10.0	4.2
X05231	Human mRNA for collagenase (identical to metalloproteinase 1)	41.8	10.0	4.2
R36977	Similar to *H. sapiens* general transcription factor IIIA (GTF3A) mRNA	177.5	43.6	4.1
U17899	Human chloride channel regulatory protein mRNA, complete cds	66.3	16.5	4.0
X54942	**H. sapiens CKSHS2 mRNA for CKS1 protein homologue**	171.9	43.1	4.0

[a]Cds, coding sequence.

negative number, values of 10 or less were set to 10 in deriving Tables 1 and 2. This has the effect of attenuating the reported change in expression when comparison involves a transcript that is not expressed or is expressed at a very low level. Thus, one should use caution in making strict quantitative comparisons between Tables 1 and 2 and data from other highly parallel methods, such as SAGE, or from traditional methods.

Statistical tests used data sets in which values <10 were not adjusted. The paired or unpaired Student's *t* test or the Mann-Whitney *U* test was used as appropriate, with significance set at $P < 0.001$. Statistical comparisons were performed on the 4000 transcripts with the highest mean intensity.

A two-way pairwise average-linkage cluster analysis was applied to organize the expression matrix such that genes and tumors with similar expression patterns are adjacent to one another (5,6). This analysis was performed with Cluster 2.02, and the resulting expression map was visualized with Treeview 1.45.[6] This process also results in a phylogenetic tree, the branch lengths of which reflect the degree of similarity between the tissues. Adenoma and adenocarcinoma samples (and their respective matched normal samples) were hybridized to different versions of the GeneChip. To permit comparison between these types of neoplasm, it was necessary to create a composite database that included only accession numbers (approximately 1800) represented on both GeneChip versions. In developing this composite database, an attempt was made to match all accession numbers on both arrays (rather than the 4000 used for the other analysis). Values of 10 or less were not adjusted. However, prior to application of the cluster algorithm to the composite database, expression levels less than or equal to zero were deleted, the remaining values were log-transformed, and both vectors of the data matrix were centered about the mean and normalized, in that order. Only genes for which 85% or more of the expression values were greater than zero were used for the clustering operation ($n = 1096$).

Once the data matrix is reorganized by the clustering algorithm so that genes are organized by similarity of expression along the vertical axis and tissue samples are organized by similarity of expression along the horizontal axis, it can be visualized as a color map, as described previously (4–6), and as shown in Fig. 3.

RESULTS

Colon adenoma ($n = 4$), adenocarcinoma ($n = 18$), and paired normal colon samples ($n = 22$) were studied (total of 44 samples). Of the ~6600 cDNAs and ESTs represented on the array, ~4000 (62%) were detected at an expression level of 10 or higher.

Relative Expression in Neoplastic and Normal Tissue. For each of the 18 adenocarcinomas, the relative expression of each gene was compared in normal and neoplastic tissue. Fig. 1 plots the average mRNA expression intensity of 4000 genes in normal and tumor samples. Points lying above or below the upper and lower boundaries represent samples in which expression in tumors was either 4-fold higher or 4-fold lower, on average, than the corresponding normal sample. Genes for which expression differences achieved statistical significance are associated with a filled circle ($P < 0.001$).

Tables 1 and 2 list the transcripts displaying a 4-fold or more increase or decrease in expression level that was also significant at the $P < 0.001$ level. Nineteen transcripts (0.48% of those detected at an intensity of 10 or greater) displayed 4–10.5-fold or greater expression in the tumors than the paired normal tissue ($P < 0.001$), whereas 88 transcripts (2.2%) displayed 4–38-fold lower expression in the cancer than the paired normal tissue ($P < 0.001$). After removing the 41 transcripts associated with smooth muscle and connective tissue, 47 remained and are included in Table 2. The choice of a 4-fold boundary for inclusion in Tables 1 or 2 is an arbitrary one, taken to constrain the size of Tables 1 and 2. The complete data set, together with the associated *P*s, is contained in the supplementary data on file.[5]

Consistent with the literature, metalloproteinases were significantly more highly expressed in colonic neoplasia than in normal tissue (*e.g.*, L23808, human metalloproteinase, X05231, and collagenase). One can also discern a group of transcripts not associated previously with colon cancer but either linked to other forms of neoplasia or to regulation of the cell cycle: MGSA, the related cytokine GRO-γ, M_r 100,000 coactivator, ckshs2, CDC25B (an M-phase tyrosine phosphatase), and transcription factor IIIA (GTF3A).

Yet another group of transcripts could be related to altered levels of metabolism (rather than cancer growth *per se*). These include *S*-adenosylhomocysteine hydrolase, pyrroline 5-carboxylate reductase, and L-iditol-2 dehydrogenase.

Several other gene products associated with colon cancer are not included in Table 1, although each had more than a 4-fold greater expression in tumors than in normal tissue, because the *P* associated with the expression difference was marginally greater than 0.001 (ranging from 0.001 to 0.003). These included matrilysin (a matrix metalloproteinase), matrix metalloproteinase 12, osteopontin, and transforming growth factor-β-induced gene product (BIGH3). Other

[6] Both shareware programs available at Internet address: http://rana.stanford.edu/software.

Table 2 *Transcripts were more highly expressed in paired normal tissue than in adenocarcinoma*

Intensity values <10 are adjusted to 10. Only transcripts with a 4-fold difference or greater ($P < 0.001$) in expression intensity between tumor and normal are included. Forty-one transcripts representing smooth muscle or collagen are not shown (see "Materials and Methods"). Transcripts shown in bold capital letters were confirmed by RT-PCR. Gene descriptions have been edited.

Accession no.	Description	Intensity in tumor	Intensity in normal	Normal/ Tumor
M83670	Human carbonic anhydrase IV mRNA, complete cds	10.0	378.	37.9
M97496	***H. sapiens* guanylin mRNA, complete cds**[a]	53.6	1082.9	20.2
X64559	***H. sapiens* mRNA for tetranectin**	10.0	137.8	13.8
T54547	*H. sapiens* cDNA similar to M84526 complement factor D precursor	10.0	119.9	12.0
M95936	Human protein-serine/threonine (AKT2) mRNA, complete cds	10.0	113.5	11.4
T55200	*H. sapiens* cDNA similar to gb:M10942_cds1 human metallothionein-le gene	10.0	84.0	8.4
T46924	*H. sapiens* cDNA similar to gb:U11863 amiloride-sens amine oxidase	15.1	124.0	8.2
L11708	Human 17 β-hydroxysteroid dehydrogenase type 2 mRNA, complete cds	16.6	134.6	8.1
T46933	*H. sapiens* cDNA clone 70843 3′ [11-β dehydrogenase]	11.2	84.9	7.6
H54425	*H. sapiens* cDNA similar to gb:M10942_cds1 human metallothionein-le gene	18.2	135.5	7.4
M26393	Human short chain acyl-CoA dehydrogenase mRNA, complete cds	10.0	71.8	7.2
M82962	Human *N*-benzoyl-L-tyrosyl-*p*-amino-benzoic acid hydrolase α subunit mRNA	10.0	71.3	7.1
J03037	Human carbonic anhydrase II mRNA, complete cds	10.0	65.1	6.5
T72257	*H. sapiens* cDNA similar to gb:L07765 liver carboxylesterase	10.0	63.1	6.3
M84526	Human adipsin/complement factor D mRNA, complete cds	43.9	260.4	5.9
T76971	*H. sapiens* cDNA similar to gb:X64177 *H. sapiens* mRNA for metallothionein	38.3	217.7	5.7
H77597	*H. sapiens* cDNA similar to gb:X64177 *H. sapiens* mRNA for metallothionein	57.0	320.5	5.6
T67986	*H. sapiens* cDNA clone 82030 3′ similar to gb:X14723 clusterin precursor	35.0	195.7	5.6
R99208	*H. sapiens* cDNA clone 200586 3′ similar to gb:X76717 *H. sapiens* MT-1l mRNA	10.0	55.7	5.6
U03749	Human chromogranin A (*CHGA*) gene, exon 8, and complete cds	10.0	55.0	5.5
R93176	Soares 1NFLS *H. sapiens* cDNA similar to gb:M33987 carb. anhydrase I[a]	10.0	53.1	5.3
L02785	***H. sapiens* colon mucosa-associated (DRA), complete cds**[a]	161.0	848.1	5.3
R94967	*H. sapiens* cDNA similar to gb:L11924 hepatocyte growth factor	10.0	52.4	5.2
J03037	Human carbonic anhydrase II mRNA, complete cds	10.0	51.8	5.2
M74509	Human endogenous retrovirus type C oncovirus sequence	15.1	77.9	5.2
L11708	Human 17 β-hydroxysteroid dehydrogenase type 2 mRNA, complete cds	18.7	96.2	5.2
X77777	*H. sapiens* intestinal VIP receptor related protein mRNA	13.7	70.5	5.1
R69552	*H. sapiens* cDNA clone 155302 3′ [glutamate]	10.0	50.6	5.1
R50730	*H. sapiens* cDNA similar to gb:Z19585 thrombospondin 4 precursor	10.0	50.3	5.0
H43887	*H. sapiens* cDNA similar to gb:M84526 complement factor D prec.	84.8	400.4	4.7
U17077	Human BENE mRNA, partial cds	32.5	147.6	4.5
U25138	Human MaxiK potassium channel β subunit mRNA, complete cds	14.9	67.4	4.5
X86693	***H. sapiens* mRNA for hevin like protein**	47.0	212.6	4.5
H57136	*H. sapiens* cDNA similar to SP:A40533 A40533 cAMP-DEP protein kinase	10.0	44.5	4.5
X73502	*H. sapiens* mRNA for cytokeratin 20[a]	55.2	245.6	4.5
J03037	Human carbonic anhydrase II mRNA, complete cds	10.0	44.0	4.4
R70806	*H. sapiens* cDNA similar to gb:X62535 diacylglycerol kinase	10.0	43.9	4.4
T51913	*H. sapiens* cDNA similar to gb:S45630 α crystallin B chain	10.0	43.5	4.3
T50678	*H. sapiens* cDNA contains TAR1 repetitive element [α tryptase]	12.5	53.7	4.3
Z50753	*H. sapiens* mRNA for GCAP-II/uroguanylin precursor[a]	42.9	183.7	4.3
M58286	*H. sapiens* tumor necrosis factor receptor mRNA, complete cds	30.5	130.5	4.3
U08854	Human UDP glucuronosyltransferase precursor (UGT2B15) mRNA, complete cds	30.8	131.2	4.3
X52679	Human ASM-2 mRNA for sphingomyelin phosphodiesterase (EC 3.1.4.12)	10.0	42.2	4.2
T71025	*H. sapiens* cDNA similar to gb:J03910_rna1 human	217.8	893.3	4.1
M12272	*H. sapiens* alcohol dehydrogenase class I γ subunit (ADH3) mRNA	42.9	174.8	4.1
M26683	Human IFN-γ treatment inducible mRNA	37.8	152.0	4.0
D90313	**Human mRNA for biliary glycoprotein, BGPI**[a]	12.7	51.0	4.0

[a] Also identified in the SAGE database, see text.

transcripts known to be more highly expressed in malignant tissue are not contained in Table 1, because the magnitude of their increase in the tumors was less than the cutoff of 4-fold imposed here. Examples include nm23-h2 (3.0-fold), c-myc (2.6-fold), and transcripts coding several ribosomal proteins (s6, s8, p0, and s3; 1.5–1.6-fold; $P < 0.01$). These smaller changes are statistically significant and are probably meaningful, contributing to the altered growth of cells.

A substantial number of transcripts were more highly expressed in normal tissue than in the paired cancer specimens (Table 2). Many of these transcripts simply represent smooth muscle or connective tissues layers more generously included with the normal than the neoplastic tissue samples. Transcripts that are readily identified as such are not included in the Table (see "Materials and Methods"). Forty-seven (1.3% of the total measured) remain and are included in Table 2. Notably, the comparison produced several transcripts shown previously to be down-regulated in cancer. Of particular interest are the following gene products: guanylin, a product of colonic epithelial cells (7); colon mucosa-associated mRNA, down-regulated in adenocarcinoma; tetranectin; hevin; and biliary glycoprotein (BGP1). This indicates that it is possible to glean meaningful information from this analysis although the tissue composition of the normal and neoplastic samples may differ.

To further validate the expression changes detected by the oligonucleotide array, semiquantitative reverse transcription-PCR reactions were used on two tumor/normal tissue pairs in which epithelial tissue was grossly dissected free of underlying stromal and muscular elements prior to RNA extraction. mRNA from these samples was reverse transcribed from oligo-dT primers, and the resulting cDNA was amplified by PCR using gene-specific primers after establishing a dilution at which amplification remained within the linear range. Gene-specific primers for glyceraldehyde-3-phosphate dehydrogenase were included in each amplification tube. The level of expression was determined by densitometric analysis using NIH image software, normalized against glyceraldehyde-3-phosphate dehydrogenase. Al-

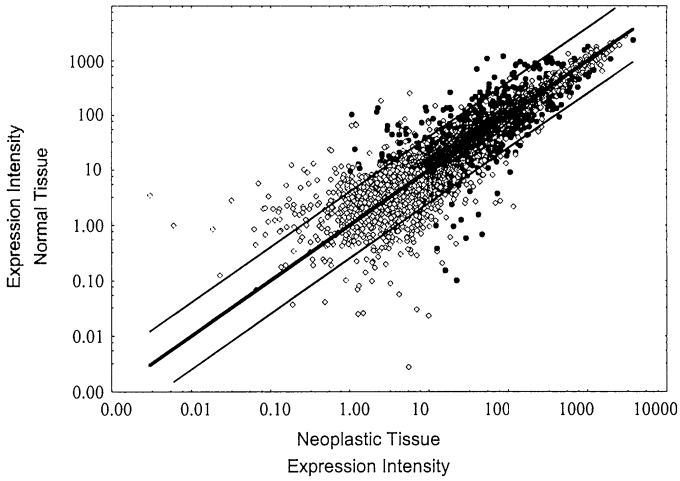

Fig. 1. Expression intensity in normal compared with tumor samples. Colon adenocarcinoma samples and paired normal samples were hybridized to GeneChips (Affymetrix), and the expression levels were analyzed with GeneChip 3.0 analysis software (Affymetrix). The upper and lower boundaries represent a 4-fold difference in the average of the expression of each gene between carcinoma and normal tissue. ●, genes for which average expression in carcinoma was significantly higher or lower than it was in the matched normal sample ($P < 0.001$). Approximately 2% of the 4000 genes analyzed for this figure displayed a statistically significant, 4-fold difference in expression intensity between tumor and matched normal samples. Data from the adenoma samples are not included in this figure. The axes scales are logarithmic.

though the precise magnitude of change in expression was not always recapitulated, the direction and order of magnitude of change predicted by the array was confirmed by reverse transcription-PCR in each case (data not shown). This indicates that for these gene products, at least, the observed changes in expression could not simply be attributed to gross contamination with nonepithelial cells. Those transcripts so validated are printed in bold letters in Tables 1 and 2.

Adenomas are considered the pathological and genetic precursor to adenocarcinoma of the colon (3). To learn whether the abnormalities in mRNA expression observed in carcinomas are already present in the precursor lesion, four adenomas (each paired with normal colon tissue from the same patient) were analyzed for mRNA expression. Interestingly, several of the abnormalities present in the carcinomas were already present in the adenomas. For example, adenomas overexpressed M_r 100,000 coactivator (by 11-fold), BIGH3 (8.9-fold), ckshs2 (2.7-fold), MGSA (2.1-fold), and matrilysin (3.0-fold). This suggests that the genes represented by these transcripts may play a role at a relatively early stage of carcinogenesis. The literature indicates that matrilysin is detected in >80% of adenomas and is not expressed in most normal epithelial cells. This metalloproteinase mRNA may be regulated by β-catenin (also overexpressed in adenoma and colorectal carcinomas) and is an emerging target of cancer therapy (8). Adenomas also displayed attenuated expression of several genes identified as down-regulated in the carcinomas, including the colonic epithelial cell product, guanylin (111-fold), down-regulated in adenocarcinoma (40-fold), and hevin (7.2-fold). Expression of genes in adenomas and carcinomas was correlated ($r = 0.5$, $P < 0.001$, Pearson's).[7] That similar patterns of dysregulation of known cancer pathway genes could be discerned in the adenocarcinoma and adenoma samples is telling, because these patterns could be discerned although bulk adenoma samples are likely to contain substantial quantities of normal colonic tissue. Thus, analysis of bulk tissue may be informative, even if it is not optimal.

A single adenoma or cancer specimen did not generally bias the differences between expression in normal and neoplastic tissue. Some comparisons are shown in Fig. 2, indicating that for many of the transcripts, differences between normal and tumor affected a substantial majority of the samples; this was so for both the adenomas and the carcinomas. There was some variability in gene expression across different tumor and normal samples, perhaps representing a combination of assay heterogeneity, true biological difference, and an imbalance in tissue composition.

Cluster Analysis. To further probe differences between normal tissue, adenomas, and carcinomas, but on a global basis, cluster analysis was performed on all 22 samples. The phylogenetic tree resulting from the hierarchical clustering algorithm is shown in Fig. 3. A striking feature of the phylogenetic tree is that the three tissue types are clearly separated in a manner that respects the conventional histopathological classification of this tumor. Although the carcinomas and their benign precursors, the adenomas, are placed on a different trunk than are the paired normal tissues, these neoplasms are also separated from one another, occupying adjacent branches of the same trunk. Thus, this hierarchical cluster algorithm, operating on a relatively small set of expression data (1096 genes and ESTs), was completely successful in grouping the three types of colon samples on the basis of subtle, distributed differences in gene expression.

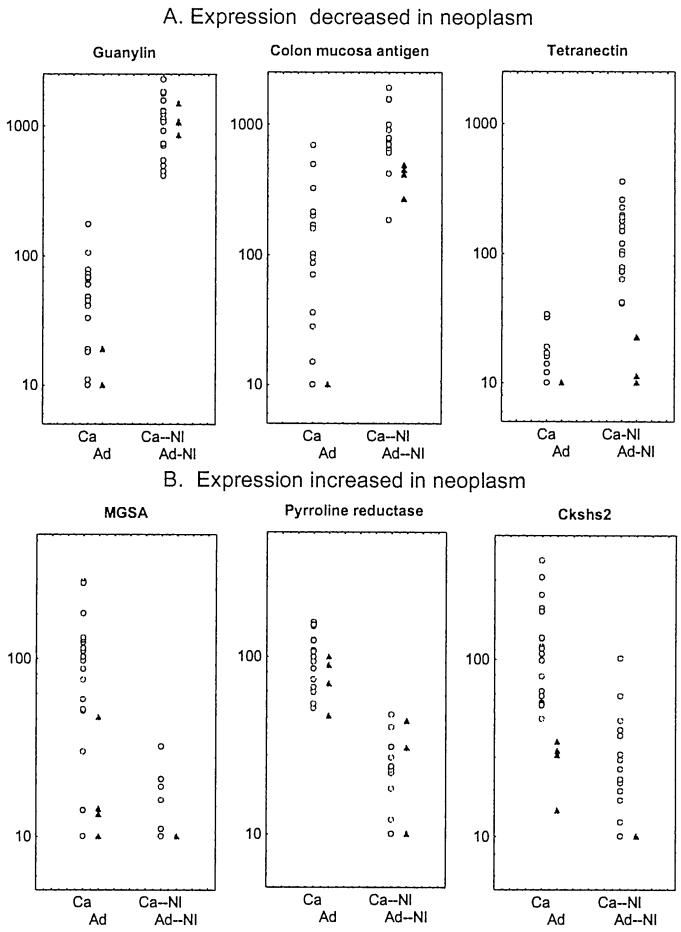

Fig. 2. Comparison of transcript expression levels in individual neoplasms (adenoma and adenocarcinoma) with paired normal tissue samples. *A*, transcripts with decreased expression in tumors. *B*, transcripts with increased expression in tumors. For each transcript, the mean ratio of expression in neoplasm to normal tissue and normal was 4-fold or greater ($P < 0.001$). The ordinate scale is logarithmic. Because of the adjustment of transcripts with an intensity ≤10 to 10, data points may be superimposed and visualized as a single data point. However, for each transcript, the number of samples was the same. *Ca*, carcinoma ($n = 18$); *Ad*, adenoma ($n = 4$); *Ca-Nl*, normal paired with carcinoma ($n = 18$); *Ad-Nl*, normal paired with adenoma ($n = 4$).

[7] A figure displaying this relationship has been placed at Internet address: http://microarray.princeton.edu/oncology/.

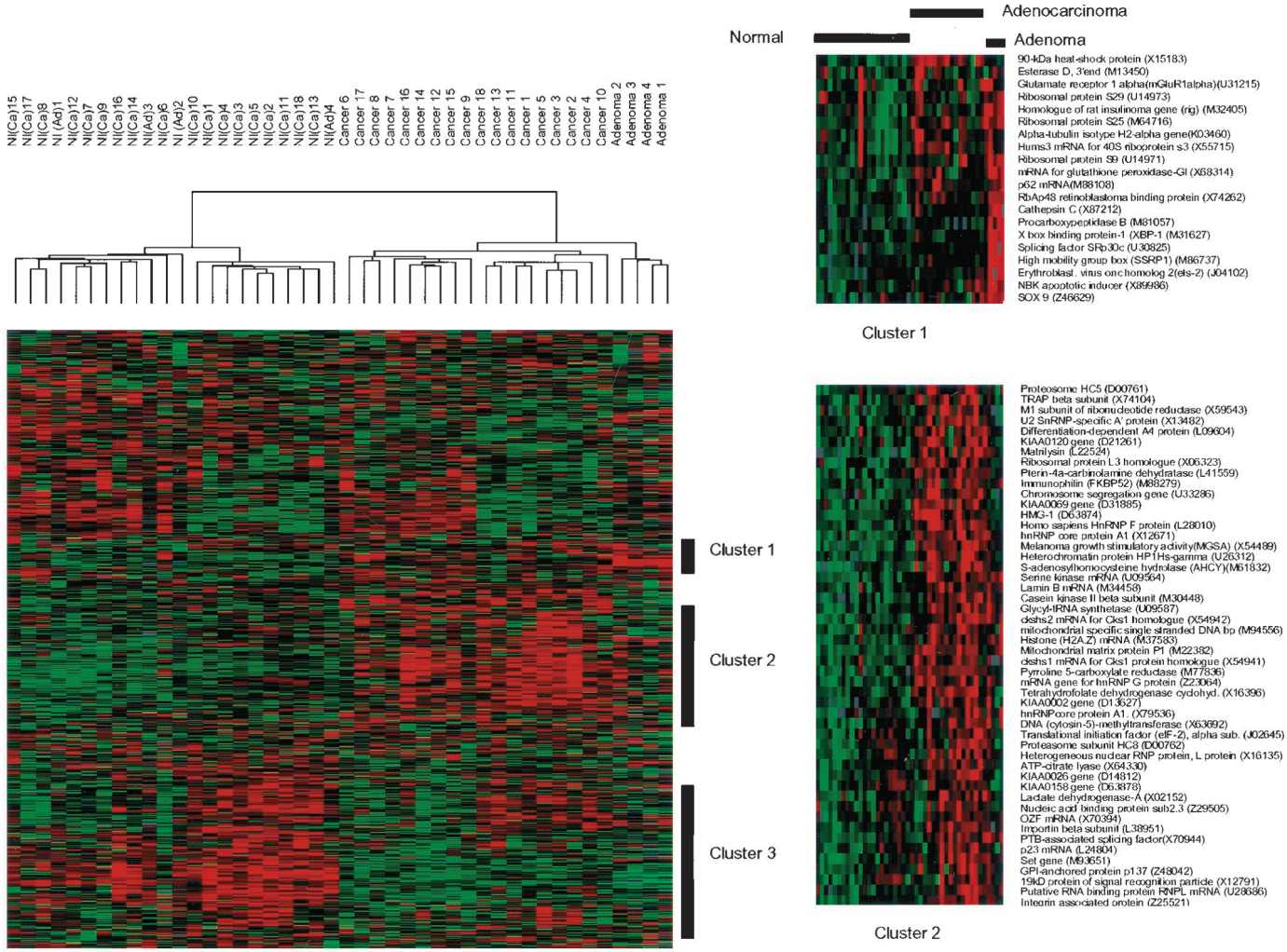

Fig. 3. *Left,* cluster map and phylogenetic tree resulting from a two-way, pairwise, average-linkage cluster analysis. Approximately 1800 genes in common to two versions of GeneChip (Affymetrix) were combined into a single matrix that was then clustered as described in the text. Each color patch in the resulting visual map represents the expression level of the associated gene in that tissue sample, with a continuum of expression levels from *dark green* (lowest) to *bright red* (highest). Missing values are coded as *silver*. The carcinomas and their benign precursors, the adenomas, are placed on an entirely different trunk than are the paired normal tissues. Strikingly, the adenomas and the carcinomas are also separated from one another, occupying adjacent branches of the same trunk. The color map indicates that this hierarchy is associated with several distinct groupings of increased gene expression intensity (clusters 1–3). *Right upper panel, cluster 1:* a cluster of genes that are more intensely expressed in adenoma than in normal tissue or carcinoma. This grouping contained approximately 8–10 genes, the mRNA expression of which appeared to be much higher in the adenoma samples than in associated normal tissue or the carcinomas. *Right lower panel, cluster 2:* a cluster of genes that are more highly expressed in carcinoma than in adenoma or normal tissue. This is a broad and diffuse cluster, which contains many products that were also identified simply by examining expression ratios (Table 1).

Several clusters of genes appear to differentiate adenomas, carcinomas, and their matched normal samples. The three most obvious are:

(*a*) cluster 1. Cluster 1 (Fig. 3, *right upper panel*) represented a group of genes that were more intensely expressed in adenoma than in either adenocarcinoma or normal tissue. This group contained several transcription factors, of which some have been implicated as oncogenes (*XBP-1, SSRP1, ETS-2,* and *SOX9*), ribosomal proteins (S29 and S9), an inducer of apoptosis (NBK), and a splicing factor (SRp30c). It is possible, but remains to be shown, that these gene products play an early role in the transition from adenoma to carcinoma.

(*b*) cluster 2. Cluster 2 (Fig. 3, *right lower panel*) comprised a larger and more diffuse group of genes that were more highly expressed in adenocarcinoma than in adenoma or in normal tissue. This cluster contains many of the gene products already identified (Tables 1 and 2) as being more highly expressed in colorectal neoplasia than in normal tissue (*e.g.,* Ckshs2, MGSA, matrilysin, and diverse products related to proliferation and metabolic rate). In some instances the cluster map reveals that expression in adenomas was intermediate between the normal samples and the carcinomas (*e.g.,* pterin-4a-carbinolamine dehydratase, casein kinase II, matrilysin, MGSA, MRL3, a human ribosomal protein L3 homologue, and several components of heterogeneous nuclear riboproteins). Intermediate expression of these genes, some of which are related to cell growth or proliferation, is consistent with the existing histological classification in which adenoma occupies a biological position between normal colon epithelium and adenocarcinoma.

(*c*) cluster 3. Cluster 3 consists of elements that are more highly expressed in a significant subset of the normal samples than in the carcinomas or adenomas. Embedded within this cluster are products the expression of which is authentically repressed in colorectal neoplasms (such as guanylin and colon mucosa antigen) as well as transcripts obviously related to smooth muscle or connective tissue (4). Subsequently, however, ~40 genes associated with smooth muscle or connective tissue were removed from the data matrix, and the data were reclustered. This procedure did not significantly affect the phylogenetic tree (data not shown).

Neither the histological grade of the tumor sample nor the clinical stage of disease could be related to a specific pattern of gene expression.

DISCUSSION

This analysis used oligonucleotide arrays capable of detecting approximately 6600 human transcripts. Of these, ~4000 (62%) were detected at a level greater than a threshold intensity of 10, the minimum used in this experiment for discrimination from a value of 0. In tumors, 19 transcripts of 4000 (0.48%) produced 4–10-fold elevated levels of transcripts when compared with normal tissue and 47 transcripts (1.3% of 4000) produced 4–38-fold lower levels of mRNA in tumors *versus* normal tissue. Thus, under the conditions of this experiment, a very small number (~1.8% of 4000 transcripts) are differentially expressed in tumors and normal tissue. This is in good agreement with the work of Zhang *et al.* (1), who showed in their SAGE analysis that ~1.5% of detected transcripts was differentially expressed in normal and neoplastic colon. Others using various types of cDNA microarrays have reported that from 0.2 to ~10% of transcripts are differentially expressed between several types of cancer and normal tissue (9–11).

To compare the existing SAGE data in colon cancer with the results of this project, the SAGE database on deposit at National Cancer Biotechnology Information was analyzed with xProfiler,[8] using the default settings (*i.e.,* a 2-fold difference in expression). Libraries derived from two bulk normal colon tissues (SAGE NC1 and NC2) and two bulk colon tumors (Tu98 and Tu102) were used for this comparison. Of the 100 SAGE tags most likely to be differentially expressed between bulk normal and tumor samples, one (M77349; transforming growth factor-β-induced gene product, BIGH3) was identified in the Affymetrix data set as overexpressed in colon cancer, and 6 were identified as more highly expressed in normal colon tissue (see Table 2). BIGH3 is not included in Table 1, because the associated P in the present analysis was 0.003, although the magnitude of overexpression was 4.8-fold.

In addition, the 100 dysregulated SAGE tags identified with xProfiler were matched to the 6003 unique GenBank accession numbers included on the GeneChip. Thirty-two of the 100 tags were identified in the Affymetrix data set; of these, 19 changed significantly ($P < 0.01$) in the Affymetrix data set. However, even for this selected set of transcripts, the actual correlation between changes in expression in one data set with those in the other data were only fair ($r = 0.5$, $P = 0.04$, Spearman's). The list of these genes, their expression values, and a graph comparing expression of these genes in the Affymetrix and SAGE databases have been deposited with the supplementary material.[5] Despite some concordance, the SAGE and Affymetrix results are different, perhaps reflecting sample heterogeneity, the limited number of tissue samples analyzed for the SAGE databases, and gene assignment differences in mapping oligonucleotide probes or SAGE tags. Clearly, it is important to develop a better understanding of the similarities and differences produced by alternative approaches to large-scale expression monitoring.

Many of the differences in transcript level between normal and neoplastic tissue reported herein make a good deal of sense even if some of the examples were unsuspected in colon tissue. For example, MGSA is elevated in selected colon cancers (Fig. 2), and it will be of some interest to determine whether this has an adverse prognosis. MGSA is homologous with GROα and is structurally related to interleukin 8 (12). In contrast, guanylin expression is very low in adenocarcinomas and adenomas, but it is abundant in normal tissue (Fig. 2). Cohen *et al.* (13) using *in situ* hybridization observed a large decrease in epithelial guanylin expression in adenocarcinoma of the colon, and these present experiments add the information that guanylin expression is also lost in the benign precursor, the adenoma, although the biochemical link between guanylin and colorectal neoplasia remains unclear.

An important goal of expression profiling is to develop a molecular approach to classifying tumors (14, 15). The task is to develop systems that will supersede the histological schema that has existed for many years, but a first step to ensuring that this is possible is to recapitulate the present classification using global expression data. In a recent publication, Golub *et al.* (15) were able to distinguish acute myelogenous from acute lymphocytic leukemia on the basis of a self-organizing map and class predictor operating on a gene expression database that was acquired with the Affymetrix GeneChip. In previous work using cluster analysis to examine breast cancer, Perou *et al.* (14) found elevated expression of genes related to cell proliferation. Many of the genes included in cluster 2 seem related to cell growth, metabolic rate, and nucleic acid synthesis. Their elevation in adenoma and cancer may be a secondary effect of proliferation rate rather than a primary transforming event. Even so, they represent a neoplastic signature that can be used as an aggregate classifying property. Present technology permits more than an order-of-magnitude increase in the number of genes whose expression can be monitored and organized by clustering routines, self-organizing maps, and factor analysis (16, 17). This should result in a much more textured and informative expression map, such that it may soon become possible to develop models of neoplastic transformation, and even clinical predictive models, that are based on concerted patterns of gene expression. For example, Alizadeh *et al.* (18) used microarray analysis and a hierarchical clustering algorithm to identify two molecularly distinct forms of B-cell lymphoma, one associated with a significantly better prognosis than the other. Others have used expression analysis to distinguish characteristic patterns of expression in malignant melanoma and breast cancer, which seem to correlate with phenotype (19, 20). Thus, a molecular classification may differentiate hitherto unknown and clinically important subsets of cancer. This present report provides support in colon cancer as well for the concept that global transcription patterns can drive a system of classification that is fundamentally related to the underlying biological context.

ACKNOWLEDGMENTS

The Cooperative Human Tissue Network provided tissue. GeneChips were a gift of Affymetrix. Construction of the colon tissue database[5] was supported, in part, by a grant from the Merck Genome Research Institute. David Fussell performed the programming for the Internet database[5]. Suzanne Ybarra and Kurt Gish performed the RNA extraction, labeling, and hybridization. David Mack provided extensive advice. Michael Eisen and Ash Alizadeh assisted with application of clustering algorithms to the adenoma/adenocarcinoma data set. Maureen Murphy, Renbin Zhao, and Lisa Taneyhill provided helpful comments.

REFERENCES

1. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. Science (Washington DC), *276:* 1268–1272, 1997.
2. Brown, P. O., and Botstein, D. Exploring the new world of the genome with DNA microarrays. Nat. Genet., *21:* 33–37, 1999.
3. Kinzler, K. W., and Vogelstein, B. Lessons from hereditary colorectal cancer. Cell, *87:* 159–170, 1996.
4. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA, *96:* 6745–6750, 1999.

[8] Internet address: http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi.

5. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.
6. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. The transcriptional program in the response of human fibroblasts to serum, Science (Washington, DC), *283:* 83–87, 1999.
7. Forte, L. R. Guanylin regulatory peptides: structures, biological activities mediated by cyclic GMP and pathobiology. Regul. Pept., *81:* 25–39, 1999.
8. Crawford, H. C., Fingleton, B. M., Rudolph-Owen, L. A., Goss, K. J., Rubinfeld, B., Polakis, P., and Matrisian, L. M. The metalloproteinase matrilysin is a target of β-catenin transactivation in intestinal tumors. Oncogene, *18:* 2883–2891, 1999.
9. 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.
10. Gress, T. M., Muller-Pillasch, F., Geng, M., Zimmerhackl, F., Zehetner, G., Friess, H., Buchler, M., Adler, G., and Lehrach, H. A pancreatic cancer-specific expression profile. Oncogene, *13:* 1819–1830, 1996.
11. Cole, K. A., Krizman, D. B., and Emmert-Buck, M. R. The genetics of cancer–a 3D model. Nat. Genet., *21:* 38–41, 1999.
12. Geiser, T., Dewald, B., Ehrengruber, M. U., Clark-Lewis, I., and Baggiolini, M. The interleukin-8-related chemotactic cytokines GROα, GROβ, and GROγ activate human neutrophil and basophil leukocytes. J. Biol. Chem., *268:* 15419–15424, 1993.
13. Cohen, M. B., Hawkins, J. A., and Witte, D. P. Guanylin mRNA expression in human intestine and colorectal adenocarcinoma. Lab. Investig., *78:* 101–108, 1998.
14. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA, *96:* 9212–9217, 1999.
15. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Washington DC), *286:* 531–537, 1999.
16. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, *96:* 2907–2912, 1999.
17. Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. Analysis of gene expression data using self-organizing maps. FEBS Lett., *451:* 142–146, 1999.
18. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature (Lond.), *403:* 503–511, 2000.
19. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature (Lond.), *406:* 536–540, 2000.
20. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.