Molecular Biology of the Cell
Vol. 13, 1929–1939, June 2002

# Gene Expression Patterns in Human Liver Cancers

**Xin Chen,*†‡ Siu Tim Cheung,†§ Samuel So,‖ Sheung Tat Fan,§ Christopher Barry,‖ John Higgins,¶ Kin-Man Lai,‖ Jiafu Ji,# Sandrine Dudoit,* Irene O.L. Ng,§§ Matt van de Rijn,¶ David Botstein,§§** and Patrick O. Brown‡****

Departments of *Biochemistry, ‖Surgery, ¶Pathology, and @Genetics, and ‡Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305; #Department of Surgery, Beijing Cancer Hospital, Beijing, China; and Departments of §Surgery and §§Pathology, The University of Hong Kong, Hong Kong, China

Submitted November 26, 2001; Revised February 19, 2002; Accepted March 18, 2002
Monitoring Editor: Keith R. Yamamoto

Hepatocellular carcinoma (HCC) is a leading cause of death worldwide. Using cDNA microarrays to characterize patterns of gene expression in HCC, we found consistent differences between the expression patterns in HCC compared with those seen in nontumor liver tissues. The expression patterns in HCC were also readily distinguished from those associated with tumors metastatic to liver. The global gene expression patterns intrinsic to each tumor were sufficiently distinctive that multiple tumor nodules from the same patient could usually be recognized and distinguished from all the others in the large sample set on the basis of their gene expression patterns alone. The distinctive gene expression patterns are characteristic of the tumors and not the patient; the expression programs seen in clonally independent tumor nodules in the same patient were no more similar than those in tumors from different patients. Moreover, clonally related tumor masses that showed distinct expression profiles were also distinguished by genotypic differences. Some features of the gene expression patterns were associated with specific phenotypic and genotypic characteristics of the tumors, including growth rate, vascular invasion, and p53 overexpression.

## INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common liver malignancy and among the five leading causes of cancer death in the world. Virtually all HCCs are associated with chronic hepatitis B virus (HBV) or hepatitis C virus infections (Beasley, 1988; Hasan *et al.*, 1990), but the molecular nature of this association is poorly understood. HCC treatment options remain limited. Surgical resection is considered the only "curative treatment" (Lin *et al.*, 1987), but >80% of patients have widespread HCC at the time of diagnosis and are not candidates for surgical treatment. Among patients with localized HCC who undergo surgery, 50% suffer a recurrence (Okuda *et al.*, 1984). Standard clinical pathological classification of HCC has limited value in predicting the outcome of treatment. Clearly, molecular markers for early and accurate diagnosis and classification of HCC would address an important medical need.

The phenotypic diversity of cancer is accompanied by a corresponding diversity in gene expression patterns (Perou *et al.*, 1999; Alizadeh *et al.*, 2000; Perou *et al.*, 2000; Ross *et al.*, 2000; Welsh *et al.*, 2001). Herein, we describe a systematic characterization of gene expression patterns in human liver cancers. We used cDNA microarrays containing 23,000 clones, representing ~17,400 human genes, to study tumor and nontumor liver tissues from HCC patients. Our aim was to characterize the gene expression programs associated with HCC as a step toward a better understanding of the molecular pathophysiology, and better methods for detection, diagnosis, and classification of HCC.

## MATERIALS AND METHODS

### *Tissues and RNA Isolation*

All patients participating in this study gave informed consent before surgery. All tissues were surgically resected, snap frozen in liquid nitrogen within 0.5 h after the resection, and stored at −80°C. In most cases, both tumor and adjacent nontumor tissues were collected. A portion of each specimen, ~0.5–1 $cm^3$, was sampled. Each sample was dissected into three equal slices. One was used for RNA extraction, one for genomic DNA iolsation, and the other processed

Article published online ahead of print. Mol. Biol. Cell 10.1091/mbc.02–02–0023. Article and publication date are at www.molbiolcell.org/cgi/doi/10.1091/mbc.02–02–0023.

** Corresponding authors. E-mail addresses: pbrown@cmgm.stanford.edu or botstein@genome.stanford.edu.
† X.C. and S.T.C. contributed equally to this work.

Abbreviations used: HCC, hepatocellular carcinoma; p53, tumor protein p53 (Li-Fraumeni syndrome).

for histological examination. Each histological slide was independently reviewed by two pathologists. Total RNA was extracted with RNeasy kit (QIAGEN, Valenica, CA), and mRNA was isolated from total RNA by using FastTrack (Invitrogen, Carlsbad, CA) or Poly-(A)Pure (Ambion, Austin, TX) mRNA purification kit. mRNA from cell lines was purified directly with FastTrack (Invitrogen) kit. We combined mRNA from the following cells, in equal quantities, to make the reference pool: HepG2, SNU398, SNU1, Jurkat, RPMI, and CCD-1070SK.

### *Microarray Procedure*

cDNA clones (23,075), representing ~17,400 genes, were mechanically printed onto treated glass microscope slides, as described previously (Perou *et al.*, 2000) (http://cmgm.stanford.edu/pbrown/array.html). Approximately 18,700 of the clones were obtained from Research Genetics, and 4,300 clones were obtained directly from Cancer Genome Anatomy Project (http://www.ncbi.nlm.nih.gov/ncicgap/). The hybridizations were performed as described previously (Alizadeh *et al.*, 2000). A detailed protocol is available at http://cmgm.stanford.edu/pbrown/protocols/5_hyb_human.html. In brief, 2 $\mu$g of sample mRNA and 2 $\mu$g of reference mRNA were labeled with Cy5-dUTP and Cy3-dUTP (Amersham Biosciences, Piscataway, NJ), respectively, by using Reverse Transcriptase (Invitrogen) for 2 h at 42°C. The two labeled cDNA probes were separated from unincorporated nucleotides by filtration, mixed, and hybridized to microarray at 65°C overnight. After hybridization, each microarray was washed with 2× SSC, 0.03%SDS for 5 min at 65°C then with 1× SSC for 5 min and 0.1× SSC for 5 min, both at room temperature. The array was then scanned using GenePix 4000A microarray scanner (Axon Instruments, Union City, CA). Array CGH was performed as described previously (Pollack *et al.*, 1999). The detailed protocol is available at http://cmgm.stanford.edu/pbrown/protocols/4_genomic.html.

### *Data Analysis*

Primary data collection and analysis were carried out using GenePix Pro 3.0 (Axon Instruments). Areas of the array with obvious blemishes were manually flagged and excluded from subsequent analysis. The raw data were deposited into Stanford Microarray Database (Sherlock *et al.*, 2001) at http://genome-www4.stanford.edu/MicroArray/SMD/index.html. All nonflagged array elements for which the fluorescent intensity in each channel was >1.5 times the local background were considered well measured. Genes for which <75% of measurements across all the samples in this study met this standard were excluded from further analysis. We selected for further analysis genes whose expression level differed by at least threefold, in at least four samples, from their mean expression level across all samples. We applied a hierarchical clustering algorithm both to the genes and arrays by using the Pearson r as the measure of similarity, and average linkage clustering, as described previously (Eisen *et al.*, 1998). The results were visualized and analyzed with TreeView (M. Eisen; http://rana.lbl.gov). We used two-sample Welch *t* statistics (allowing for unequal variances) to identify genes that were differentially expressed in two sets of samples. The statistical significance of the differential expression of any gene was assessed by computing a p value for each gene, representing the chance of observing a test statistic at least as large (in absolute value) as the value actually obtained. No specific parametric form was assumed for the distribution of the test statistics. To determine the p value, we used a permutation procedure in which the class labels of the samples were permuted 500,000 times, and for each permutation, two-sample *t* statistics were computed for each gene. The permutation p value for a particular gene is the proportion of the permutations (out of 500,000) in which the permuted test statistic exceeds the observed test statistic in absolute values. Any gene for which this p value was <0.001 was considered to be differentially expressed. The corresponding "per-family type 1 error rate" (PFER), that is, the expected number of false positives for such a multiple test procedure is PFER = number of genes × 0.001. Alternatively, the Benjamin & Hochberg procedure was applied to control the "false discovery rate" (FDR), or expected proportion of false positive among the genes declared differentially expressed.

### *Immunohistochemistry*

Immunohistochemistry was performed as described previously (Perou *et al.*, 1999). Antibody MY10, for CD34, was used at 1:10 dilution (BD Biosciences, San Jose, CA). Antibody DO-7, for p53, was used at 1:100 dilution (DAKO, Carpinteria, CA).

### *Southern Analysis*

Genomic DNA was extracted using the DNeasy extraction kit (QIAGEN), digested overnight with *Hin*dIII or *Eco*RI, and resolved by electrophoresis through a 1% agarose gel. After depurination, denaturation, and neutralization, the gel was transferred overnight with 10× SSC to a nylon Hybond-$N^+$ membrane (Applied Biosystems, Foster City, CA). HBV-specific sequences were detected by hybridization with a polymerase chain reaction-amplified copy of the complete HBV genome, from blood of a patient with HBV infection, labeled with fluorescein by random primed labeling (Applied Biosystems). Detection was performed with anti-fluorescein antibody conjugated with horseradish peroxidase, followed with enhanced chemiluminescence development (Applied Biosciences) and exposure to x-ray film (Eastman Kodak, Rochester, NY).

## RESULTS

### *Gene Expression in HCC and Nontumor Liver Tissues*

We characterized genomic expression patterns in >200 samples, including 102 primary HCC (from 82 patients), 74 nontumor liver tissues (from 72 patients), seven benign liver tumor samples (three adenoma and four FNH), 10 metastatic cancers, and 10 HCC cell lines. The complete data are available at http://genome-www.stanford.edu/hcc/Figures/ArrayInformation.htm.

As a first step to organize the results for visual display and for further analysis, we used a hierarchical clustering algorithm (Eisen *et al.*, 1998) to group the genes, as well as the samples, on the basis of similarity in their expression pattern. The results for the 3180 genes (represented by 3964 cDNAs) with the greatest variations in expression in 82 HCC and 74 nontumor liver tissues samples are displayed in Figure 1. To help provide a framework for interpretation of the expression patterns observed in the clinical samples, we compared these results with the gene expression patterns in 10 HCC cell lines. The expression data for the cell lines are displayed to the left of the main panel in Figure 1.

Several features of the gene expression patterns are evident in Figure 1. First, based solely on their gene expression patterns, the clinical samples could be divided into two major clusters, one representing HCC samples, and the other, with a few exceptions, representing nontumor liver tissues. Second, expression patterns varied significantly among the HCC and nontumor liver samples. Third, samples from HBV-infected, hepatitis C virus-infected, and noninfected individuals were interspersed in the HCC branch.

Close-up views of clusters of genes whose expression covaries in this set of samples are shown in Figure 2. One cluster of genes was highly expressed in HCC samples compared with nontumor liver tissues. It includes the

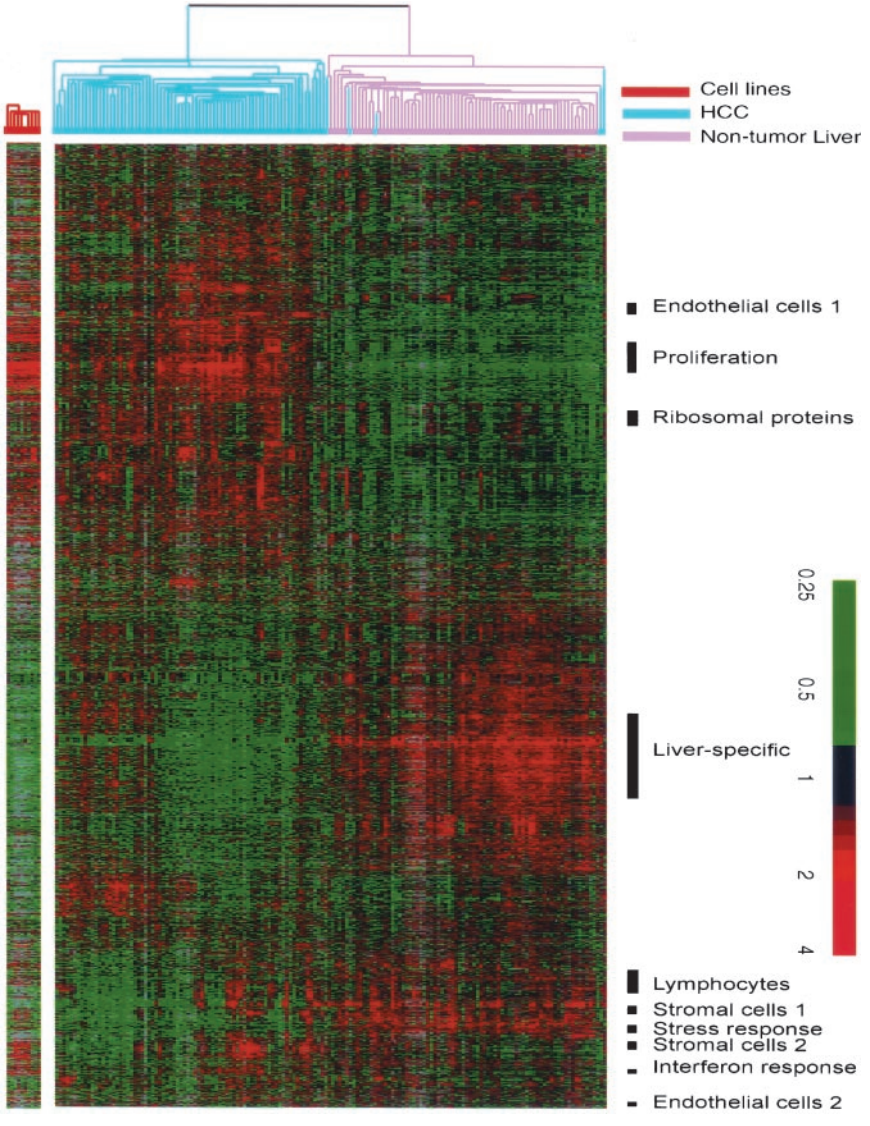


**Figure 1.** Hierarchical clustering of the patterns of variation in expression of 3180 genes (represented by 3964 cDNA), in 156 liver tissues (74 nontumor liver and 82 HCC). The data are shown in a table format, in which rows represent individual genes and columns represent individual tissue or cell sample. The color in each cell reflects the expression level of the corresponding gene in the corresponding tissue, relative to its mean expression level across the entire set of tissue samples. The scale extends from fluorescence ratios of 0.25–4 relative to the mean level for all samples. Gray indicates missing or excluded data. Expression of the same genes in 10 HCC cell lines is similarly represented in the panel to the left of the main panel. See supplementary information for the full data, including sample names.

"proliferation cluster" (Figure 2A), comprised of genes whose functions are required for cell cycle progression and whose expression levels correlate with cellular proliferation rates. Most of the genes in this cluster are specifically expressed in G2/M phase (Cho *et al.*, 2001). As expected, liver cell lines, continuously proliferating in culture, also expressed the genes in this cluster at high levels. Genes encoding ribosomal proteins were also relatively highly expressed in HCC, an expression pattern characteristically observed in growing cells. Other genes not known to be related to cell proliferation or translation also showed consistently elevated expression in HCC. These genes are implicated in a variety of cellular process, including cell signaling, transcriptional regulation, RNA splicing, protein degradation, and cell adhesion. The functional significance of their elevated expression in HCC remains to be elucidated.

Among the genes that were expressed at lower levels in HCC than in nontumor liver tissues, most seemed to be genes specifically expressed in differentiated hepatocytes (Figure 2B), including genes encoding liver-specific metabolic enzymes and many plasma proteins, including clotting factors, apolipoproteins, and complement proteins. When expanded in vitro, HCC cell lines have been found to lose expression of most "liver-specific markers." These data confirm that many of the characteristic molecular features of normal hepatocytes are clearly also deficient in the tumor themselves.

Both normal liver and liver tumors are complex tissues composed of diverse specialized cells. Distinct patterns of gene expression seemed to provide molecular signatures of several specific cell types. Expression of two clusters of genes associated with T and B lymphocytes, respectively, presumably reflects lymphocytic infiltration into liver tissues (Figure 2, C and D). In nontumor liver tissues, their expression correlated with viral infection, perhaps reflecting the chronic inflammatory response to the infection.

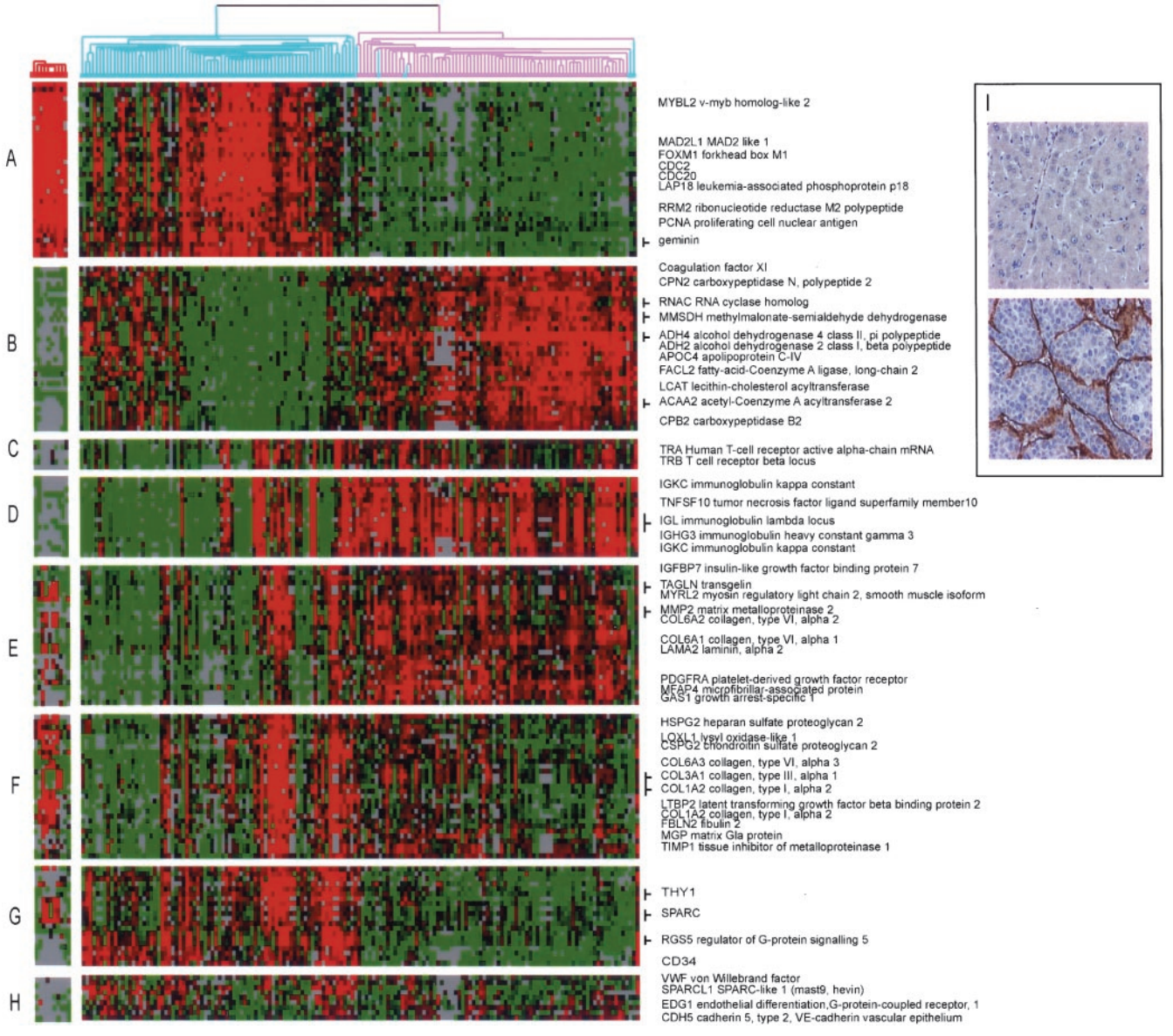


**Figure 2.** Features of the variation in gene expression patterns can be related to specific physiological or histological features of the samples. Data are the same as in Figure 1. Proliferation cluster (A), liver-specific cluster (B), T-lymphocytes cluster (C), B-lymphocytes cluster (D), stromal cell cluster 1 (E), stromal cell cluster 2 (F), endothelial cell cluster 1 (G); endothelial cell cluster 2 (H), immunohistochemistry staining of CD34 on nontumor liver (upper) and HCC (lower) tissues (I). Due to limited space, only a few selected gene names are shown. Only a portion of the proliferation cluster and the liver-specific cluster are shown. See supplementary information for the full data.

The expression patterns of two distinct clusters of genes seem to reflect variation in the density and composition of stromal cells (Figure 2, E and F). Several of the genes in the second group, including LTBP2, CSPG2, and TIMP1, have been shown to be expressed in activated stellate cells in response to liver injury (Gressner *et al*., 1994). Variation in expression of this cluster of genes may therefore reflect variation in the distribution of activated stellate cells in the tissues.

A cluster that included several genes typically expressed in endothelial cells, including CD34, RGS5, and THY1, was expressed at a higher level in HCC than in nontumor liver tissues (Figure 2G). CD34 is expressed in endothelial cells in veins and arteries but not in the endothelial cells of the sinusoids in nontumor liver. However, it is strongly expressed in the endothelial cells that line the sinusoid-like vessels in the HCC samples (Figure 2I), presumably reflecting disruption in HCC of the molecular program that normally regulates blood vessel morphogenesis in the liver. A second cluster of genes characteristically expressed in endothelial cells, including VWF, hevin, and VE-cadherin (Figure 2H), showed variable expression among the tissue samples

and no systematic difference in expression between HCC and nontumor liver tissues. These two groups of genes may therefore represent two distinct types of endothelial-like cells in liver.

### *Clonal and Genotypic Variation in Gene Expression Patterns in HCC Samples*

Previous work has suggested that the gene expression pattern of a tumor can provide a distinctive molecular portrait recognizable in successive samples over time and in metastases (Perou *et al.*, 2000). How distinctive and consistent are the gene expression patterns in individual hepatocellular carcinomas? We investigated whether samples from multiple sites in a single HCC tumor, or multiple separate tumor nodules in one patient, would share a recognizable gene expression signature. We analyzed 102 HCC samples from 82 patients by hierarchical clustering based on the overall similarity in their pattern of expression of 2640 genes (represented by 3271 cDNA clones) (Figure 3A). With a few instructive exceptions, all the tumors samples from each patient clustered together.

To further examine the relationships among multiple tumor samples from individual patients, we calculated the pairwise r for all pairs of samples and displayed the results in Figure 3B. For SF1, SF34, SF35, and HK62, each primary tumor was sampled multiple times. For each of these tumors, the pattern of gene expression in a given tumor sample was more highly correlated with the pattern seen in the other samples from the sample patient than with any of the other tumors we analyzed. Every tumor therefore had a distinctive and characteristic gene expression pattern, recognizable in all samples taken from different areas of the same tumor.

Multiple discrete tumor masses were obtained from six patients. In three of these patients, HK63, HK64, and HK66, the multiple tumors shared a distinctive gene expression pattern. In the three other patients, HK65, HK67, and HK85, the expression patterns varied between tumor nodules, and the difference provides new insights into the sources of variation in the molecular and biological characteristics of cancers.

Recent studies have shown that the gene expression patterns in samples of a given tumor taken at different times or from different sites are typically much more similar to one another than are the expression patterns observed in tumors of the same type in different patients (Perou *et al.*, 2000). Do the distinctive expression patterns characteristic of each tumor reflect the individuality of the tumor, or are they determined by the patient in whom the tumor arose? Analysis of the multifocal hepacellular carcinomas provides an opportunity to address this important question.

The expression patterns observed in the two tumor nodules from patient HK85 were not significantly more similar than those of an arbitrary pair of tumors from different patients. These two tumor nodules were each 2 cm in diameter and were separated by a distance of 7 cm. Southern analysis of the HBV integration sites showed that T1 and T2 had distinct integration patterns (see supplementary information), strongly suggesting that they were clonally independent tumors.

The tumor nodules from patient HK65 were immediately adjacent, but grossly separated foci. Each tumor mass was ~2 cm in diameter. The gene expression patterns observed in tumor nodules HK65-T2 and HK65-T4 were more similar to each other than either was to the pattern observed in HK65-T1. The gene expression patterns in the tumor nodules in this patient showed an intriguing relationship to p53 activity. In normal cells, p53 is so rare that the protein usually cannot be detected by standard immunohistochemical staining. Positive p53 staining generally indicates a mutant form of p53 gene in HCC tissues (Hsu *et al.*, 1993), probably because the mutated p53 protein is more stable and accumulates in the nuclei. p53 was undetectable by immunohistochemical staining in T2 and T4, but it was readily detected in the nuclei of tumor cells in T1 (Figure 3C). To investigate the clonality of these tumor nodules, we performed Southern analysis of the HBV integration site on T1 and T2 (see supplementary information). The HBV integration sites in these tumors seemed indistinguishable, suggesting they arose from the same clone. We used microarray CGH to characterize the patterns of chromosomal amplifications and deletions in T1 and T2 (see supplementary information). Although the chromosomal changes in T1 seemed, by this assay, generally similar to those seen in T2, T1 showed some distinct patterns, including partial or complete loss of a copy of chromosomes 5q, 9p, 12, and 22. All these data suggest that despite their common clonal origin, the genome of T1 is distinctly different from that of T2 and T4, presumably accounting for the differences in gene expression pattern.

In patient HK67, T1 was the main tumor mass, measuring 8 cm in diameter. T2 and T3, measuring 1 and 3 cm, respectively, were satellite nodules in a different liver lobe. expression patterns of T2 and T3 were remarkably similar, but only distantly related to the pattern observed in T1. Although p53 was uniformly detectable by immunohistochemistry in the nuclei of tumors cells in T2 and T3, the immunostaining pattern observed in T1 was heterogeneous. In most areas, p53 staining was undetectable, but in some patches of this tumor, p53 staining was readily detected in >50% of the tumor cell nuclei (Figure 3D). Genes associated with proliferation were expressed at significantly higher levels in T2 and T3 than in T1. Array CGH analysis revealed very similar patterns of chromosomal amplifications and deletions in T1 and T2, including rare chromosomal abnormalities (Wong *et al.*, 1999; Shiraishi *et al.*, 2001) such as amplification of 19q, and deletion of 15q and the centromere region of 22 (see supplementary information), suggesting a common clonal origin. However, T2 was distinguished from T1 by an additional deletion of chromosome 13. The data suggest that at least some of the tumor cells of T1 are genotypically unstable, resulting in genotypic heterogeneity among the cells in this tumor. In view of their smaller size, faster proliferation, and homogeneity with respect to p53 staining, T2 and T3 are probably subclones derived from T1. The p53 mutation and other genetic changes apparently provided a growth, survival, or migration advantage that enabled these clones to outgrow and metastasize.

Taken together, these results suggest that each independently arising tumor is distinguished from other tumors of the same pathological type, whether they arise in the same patient or different patients, by a distinctive gene expression program that reflects the cell of origin and the unique sequence of genetic events. Moreover, multiple clonally re-

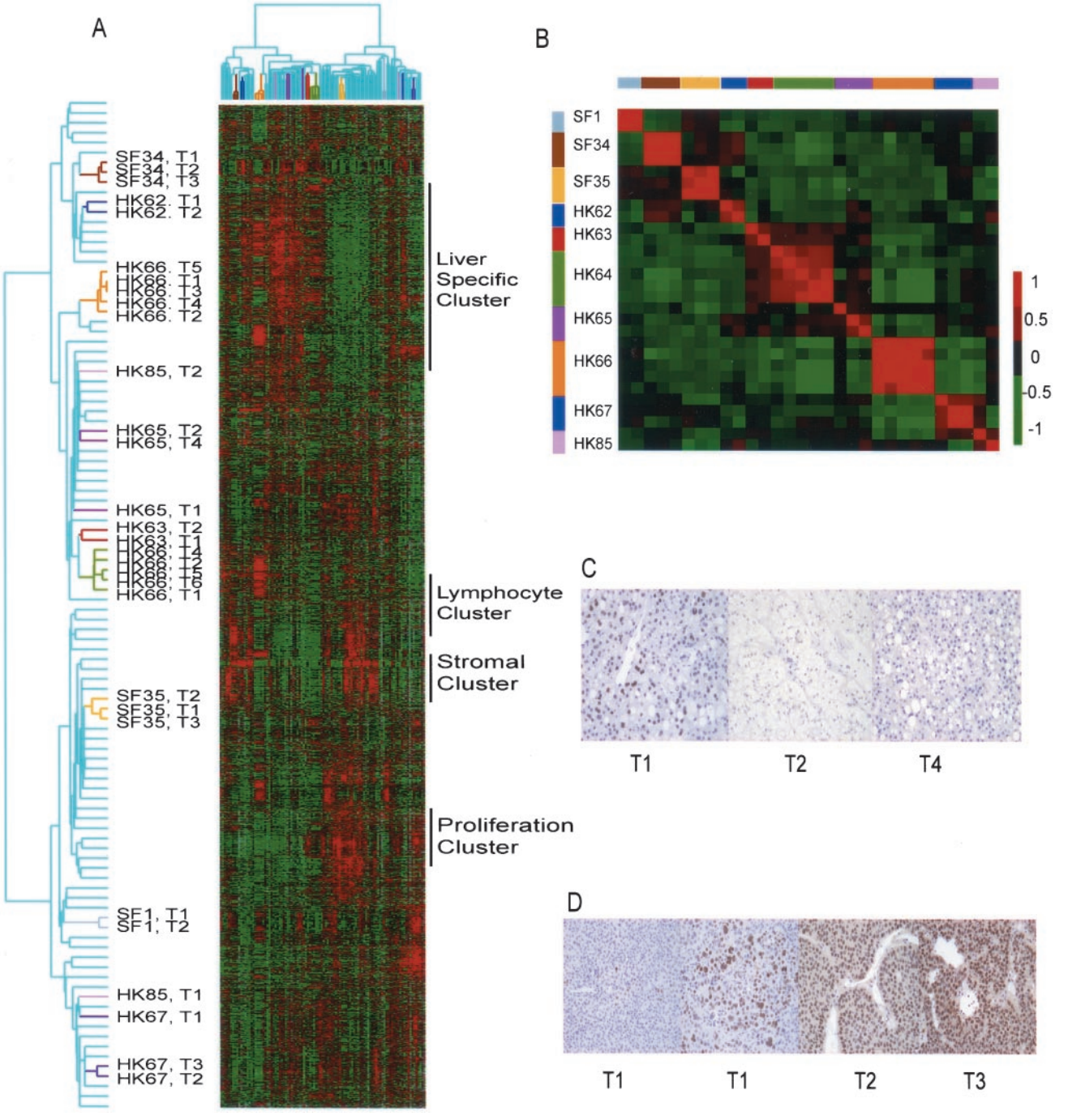


**Figure 3.** Gene expression profiles in HCC samples. (A) Hierarchical clustering of 3271 clones representing ~2640 different genes and 102 HCC samples (from 82 patients) based on similarity in gene expression patterns. The scale is the same as Figure 1. The dendrogram on the left shows the samples analyzed. Branches are colored to highlight groups of samples from the same patient. Each blue line represents a patient from whom only a single tumor sample was analyzed. Each of the other colors represents multiple tumor samples from a single patient. (B) r values between multiple samples. Using all the genes for which technically adequate measurements could be obtained, we calculated the r values for the expression patterns of each pair of samples. In this image, each cell represents the r for one pair of samples, using the color key indicated to the right of the panel. (C) Immunohistochemical staining of tumors staining HK65 T1, T2, and T4 for p53. (D) Immunohistochemical staining of tumors HK67 T1 (left two) and T2 and T3 (right two) for p53.

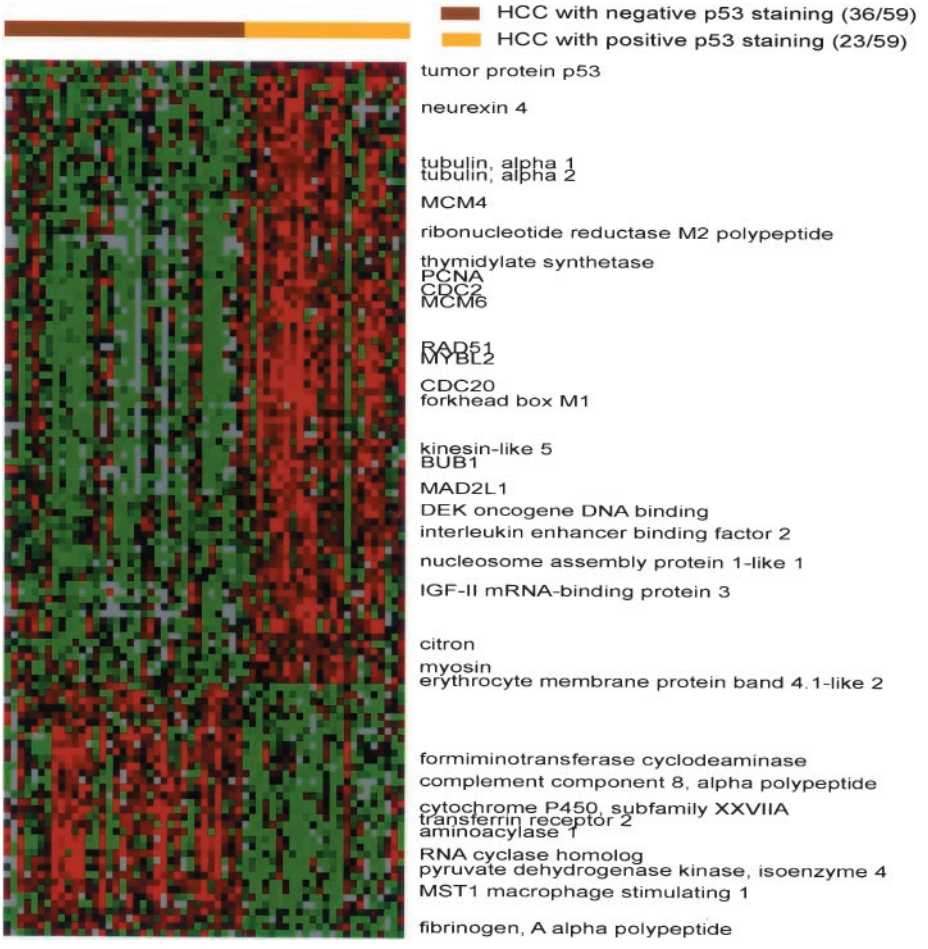


**Figure 4.** Hierarchical clustering of 121 genes whose expression level was significantly correlated with the presence or absence of nuclear p53 immunostaining in HCC (permutation p value $<$ 0.001; corresponding to a PFER of 4.043 and a nominal FDR of $\sim$0.035). The scale is the same as Figure 1. The colored bars above the panel indicate p53 positive (yellow) or negative tumor samples (brown). Due to space limitation, only some of the genes are labeled. The complete data with clone ID and gene names are available at the supplementary information.

lated tumor samples in the same patients can show different gene expression patterns due to divergent histories of mutations or chromosomal alterations.

### *Correlation between p53 Accumulation and Cell Proliferation*

Mutations in the p53 gene are a common finding in HCC and may play a key role in pathogenesis. To investigate the relationship between p53 mutations and the gene expression program in HCC, 59 HCC specimens were examined by immunohistochemical staining for p53 protein. Positive p53 staining, which has been noted to correlate with p53 mutation or inflammation in HCC (Hsu *et al.*, 1993), was found in 23 of the tumors analyzed. We found no apparent correlation between p53 staining and histological evidence of inflammation in our series.

Among the 59 samples we analyzed, we found characteristic differences in gene expression patterns between the tumors with positive p53 staining and those with negative p53 staining. We identified 121 genes whose expression level was correlated with p53 staining, with $p < 0.001$ by two-sample Welch $t$ statistics (Figure 4). Of these 121 genes, 86 were more highly expressed in tumors with positive immunostaining for p53. Most of these 86 genes belonged to the proliferation cluster, i.e., their expression was strongly associated with proliferation, suggesting that mutation of p53 might be a key pathogenetic event leading to accelerated cell proliferation during HCC tumorgenesis.

Most of the 35 genes that were characteristically expressed at lower levels in the HCC samples with positive p53 staining are noteworthy for their specific expression in differentiated hepatocytes, presumably reflecting a tendency for tumors cells with p53 mutations to be poorly differentiated (Caruso and Valentini, 1999). Interestingly, one of these genes, MST, also called hepatocyte growth factor-like protein, has been reported to be induced by p53 expression (Zhao *et al.*, 2000), raising the possibility of a role for this putative growth factor in p53-dependent regulation of hepatocyte differentiation.

### *Vascular Invasion and Gene Expression Pattern in HCC*

Vascular invasion seems to have an important role in tumor spread and metastasis. All tumor samples were classified by histopathological evaluation as either positive or negative for vascular invasion. We identified 91 genes whose expression levels were significantly correlated with the presence or

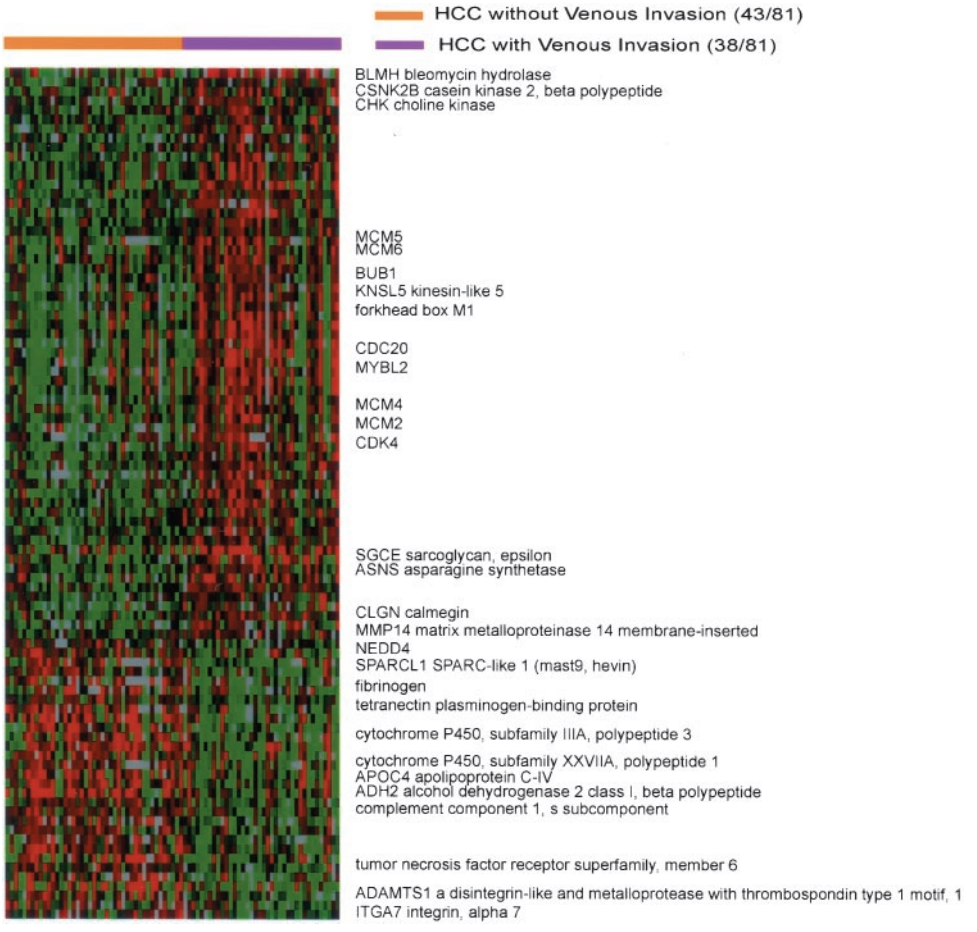


**Figure 5.** Hierarchical clustering of 91 genes whose expression was significantly correlated with vascular invasion in HCC (permutation p value $< 0.001$; corresponding to a PFER of 4.043 and a nominal FDR of $\sim 0.042$). The scale is the same as Figure 1. The colored bars above the panel indicate tumors with vascular invasion (purple) or without vascular invasion (orange). See supplementary information for full data.

absence of vascular invasion by two-sample Welch *t* statistics ($p < 0.001$) (Figure 5).

Sixty-one of these genes were expressed at higher levels in tumors with vascular invasion than in tumors without vascular invasion. Most of these genes had functions and expression patterns associated with cell proliferation. One of the few genes in this set that seems to be unrelated to cell proliferation encodes matrix metalloproteinase (MMP) 14. MMP family members are involved in the breakdown of extracellular matrix and may play important roles in invasion and metastasis (Liotta *et al.*, 1980; Arii *et al.*, 1996; Hayasaka *et al.*, 1996). MMP14 may have a direct role in activation of MMP2 (Sato *et al.*, 1994), which in turn has been shown to be related to an invasive phenotype and poor prognosis in HCC (Murakami *et al.*, 1999). The association of MMP14 expression with vascular invasion highlights the possible importance of these MMPs in the progression of HCC and underscores their potential as therapeutic targets.

Most of the genes that were expressed at lower levels in the tumors with vascular invasion were "liver specific," consistent with the classical pathological observation that poorly differentiated HCC tumors tend to be more aggressive and invasive (Ng *et al.*, 1995). One of the few genes in this group that did not belong to the liver-specific cluster encodes the metalloprotease ADAMTS1, which was recently shown to inhibit endothelial cell proliferation and to have antiangiogenic activity (Vazquez *et al.*, 1999). Although vascular invasion and angiogenesis have traditionally been viewed as independent properties of a tumor, a recent study has suggested an association between vascular invasion and intratumoral angiogenesis (Maehara *et al.*, 2000). The function and regulation of ADAMTS1, and its possible role in suppressing vascular invasion and angiogenesis, warrant further investigation.

### *Expression Patterns Distinguish Cancers Metastatic to Liver from Primary HCC*

Many tumors, particularly those arising in the gastrointestinal tract, have a propensity to metastasize to the liver. We examined several metastatic lesions in the liver. The expression patterns of 10 randomly selected HCC samples and 10 liver metastases of other cancers were analyzed by hierarchical clustering as shown in Figure 6. The HCC samples and the metastatic cancers clustered into two distinct groups, based on differences in their patterns of gene expression. Although some of the HCC samples were poorly differentiated and expressed the genes of the liver-specific cluster at very low levels compared with either normal liver or well-differentiated HCC, the genes of the liver-specific cluster were consistently expressed at higher levels in HCC than in tumors of nonliver origin. Metastatic cancers origi-

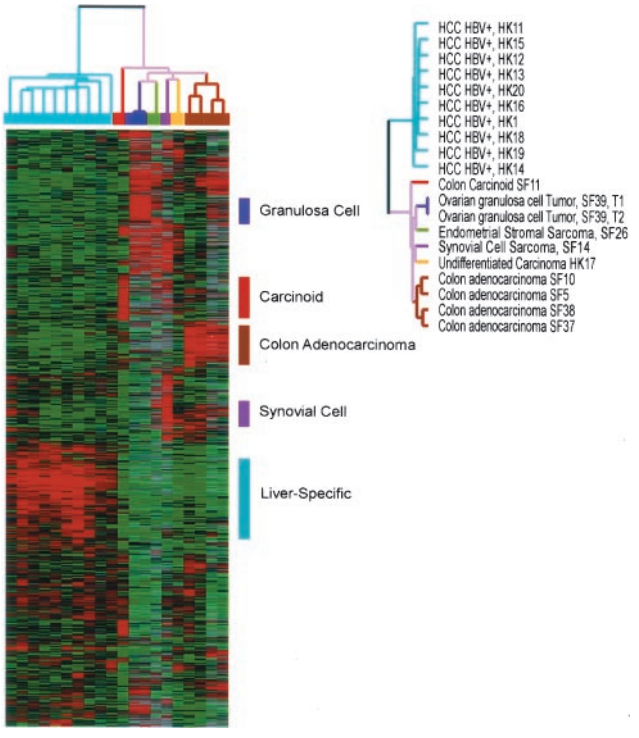


**Figure 6.** Comparison of gene expression patterns in HCC and tumors metastatic to liver. Expression data for 3474 cDNA clones representing ~2780 different genes, in 10 HCC and 10 metastatic tumor samples, were analyzed by hierarchical clustering. The scale is the same as Figure 1. Differently colored branches in the dendrogram represent different histologically defined tumor types. See supplementary information for full data.

nating from the same tissue typically clustered together, expressing genes characteristic of the cell type of origin. For example, the four samples of metastatic colon adenocarcinoma clustered together, distinguished in part by their abundant expression of the keratin 17, keratin 19, and mucin 1 genes. In contrast, a metastatic carcinoid tumor from the colon expressed genes commonly associated with neurons and neuroendocrine cells, including syntaxin 1A and syntaxin binding protein 1. The sample labeled HK17, diagnosed as undifferentiated adenocarcinoma, expressed a distinctive set of genes. Due in part to the small number and limited diversity of the non-HCC tumors in this study we cannot pinpoint the site from which this tumor arose. We are optimistic, however, that systematic comparison of the gene expression pattern in a metastatic tumor of uncertain origin with the expression patterns observed in a large, diverse sample of tumors and normal tissues will allow reliable recognition of the primary tumor of origin.

## DISCUSSION

The molecular pathogenesis of HCC is still poorly understood, and its clinical course can vary widely. Systematic analysis of global gene expression programs in human cancers can lead to new insights into pathogenetic mechanisms and improved prediction of clinical behavior (Golub *et al.*, 1999; Alizadeh *et al.*, 2000; Perou *et al.*, 2000). A recent report analyzed gene expression patterns in 20 HCC samples (Okabe *et al.*, 2001), by directly comparing tumor to nontumor liver tissue from the same patient. The studies noted the elevated expression of proliferation-associated genes in the tumors, and suggested that expression of specific genes could be associated with the viral etiology of the tumor, vascular invasion, and other features. The interpretation of these results was complicated, however, by the fact that the gene expression patterns in the nontumor liver samples from different patients can vary significantly, affected by the viral infection and degree of cirrhosis (Figure 1). Thus, the tumor-specific variation in the expression patterns could not be distinguished from variation due to differences in the corresponding nontumor liver samples.

In this study, we used a common reference RNA sample as an internal standard for measurement of expression in each clinical specimen, allowing systematic comparisons to be made among all of our tumor and nontumor tissues and cell lines. HCC and nontumor liver samples were readily divided into two separate major branches when these tissues were hierarchically clustered based on their gene expression patterns, reflecting the consistent differences between the gene expression program in HCC and nontumor liver tissues. Of the 3180 genes that showed the greatest variation in expression among all the samples we studied (Figure 1), ~1640 genes were differentially expressed in HCC vs. nontumor liver samples ($p < 0.01$ by Student's $t$ test with Bonferroni correction; see supplementary information).

Primary HCC was readily distinguished from tumors metastatic to liver based on differences in global gene expression patterns. Metastatic tumors from the same primary site shared distinctive gene expression patterns that seemed to be related to their normal cellular progenitor. Metastatic tumors of unknown primary origin are not rare. Classification of these tumors according to the primary tumor of origin has important implications for treatment. Our preliminary results provide encouraging support for the hypothesis that tumors of unknown primary could be identified by comparing their gene expression patterns with the profiles of diverse malignant and normal tissues.

As observed for breast cancers (Perou *et al.*, 2000), lymphomas (Alizadeh *et al.*, 2000), and other cancers (our unpublished data), the gene expression pattern of each HCC seems to provide a distinctive molecular portrait of that tumor, several features of which are statistically associated with specific phenotypic features of the tumors. Multiple tumor samples from the same patient typically share recognizable and distinctive features in their gene expression patterns. An important question raised by this recurrent observation is whether the differences in gene expression patterns reflect the individuality of the tumors or the patients in whom the tumors arise. The results of this study provide a partial answer. In a few cases, separate tumors from a single patient showed dramatic differences in their gene expression patterns. We found that the differences are likely due either to independent clonal origins of separate HCCs in the same patient (HK85), or to divergent genotypic evolution of clonally related tumors (HK65 and HK67). Therefore, the differences that distinguish the gene expression programs of individual tumors of the same histological

type seem primarily to arise during the development and progression of the tumors, rather than reflecting underlying differences between the patients in whom the tumors arose.

Extensive variation in histology and cell morphology is a frequent feature of cancer. The microscopic heterogeneity in p53 protein levels observed in one of the tumors were analyzed, HK67, provides an informative example (Figure 3D). The RNA samples were analyzed in this study were isolated from macroscopic tumor samples ~1 $cm^3$ in size, a scale too large to allow resolution of the differences in gene expression to accompany the differences in p53 activity (and presumably other genotypic variation) represented in this heterogeneous tumor sample. However, the differences in the expression patterns observed between the largest tumor mass in patient HK67, tumor T1, and tumor nodules T2 and T3, presumed to be intrahepatic metastases from T1, support the hypothesis that the cellular heterogeneity observed at the histological level is accompanied by corresponding heterogeneity in the gene expression patterns. Clearly, this important phenomenon warrants further study by using methods such as laser-capture microdissection and immunhistochemistry, which can allow molecular variation at the level of single cells to be characterized in greater detail.

Among the genes that were characteristically highly expressed in HCC, those whose products are membrane associated or secreted are of particular interest for their potential as therapeutic targets or as serological markers for early detection. $\alpha$-Fetoprotein has been widely used as the serum marker for HCC diagnosis and follow-up. However, it is elevated in only 50% of HCC (Johnson, 2001) (Figure 1). Additional, and better, serological markers are clearly needed. In this study, we were able to identify hundreds of genes whose expression was more consistently elevated in HCC than was $\alpha$-fetoprotein (see supplementary information). Using a DNA microarray method for identification of membrane-associated transcripts (Diehn *et al.*, 2000) and sequence analysis for transmembrane domain or signal peptides (Kyte and Doolittle, 1982; Nielsen *et al.*, 1997), we were able to distinguish a subset of these genes whose products are likely to be either membrane bound or secreted (Diehn, unpublished data). By raising antibodies against the products of these genes, we hope to identify new serum markers for detection and diagnosis of HCC, and perhaps new candidate targets for treatment.

Supplementary information is available on the authors' World-Wide Web site (http://genome-www.stanford.edu/hcc).

## ACKNOWLEDGMENTS

We thank the members of the Stanford Microarray Facility and Stanford Microarray Database, especially Mike Fero, Gavin Sherlock and Tina Hernandez-Boussard. We thank B. Strausberg at NCI for CGAP clones. We also thank the members of the Asian Liver Center at Stanford, especially Wijan Prapong and Ann Vu-Nguyen. We are grateful for the members of the Brown and Botstein laboratories for helpful discussions and comments on the manuscript. Research at Stanford was supported by grants from the National Cancer Institute to P.O.B and D.B., by a grant from the H. M. Lui Foundation to S.S., and by the Howard Hughes Medical Institute. The work of S.T.C. and S.T.F. was supported in part by a grant from the Sun Chieh-Yeh Research Foundation for Hepatobiliary and Pancreatic Surgery, HKU. X.C. is a Howard Hughes fellow of the Life Science Research Foundation. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

## REFERENCES

Alizadeh, A.A., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Arii, S., Mise, M., Harada, T., Furutani, M., Ishigami, S., Niwano, M., Mizumoto, M., Fukumoto, M., and Imamura, M. (1996). Overexpression of matrix metalloproteinase 9 gene in hepatocellular carcinoma with invasive potential. Hepatology *24*, 316–322.

Beasley, R.P. (1988). Hepatitis B virus. The major etiology of hepatocellular carcinoma. Cancer *61*, 1942–1956.

Caruso, M.L., and Valentini, A.M. (1999). Overexpression of p53 in a large series of patients with hepatocellular carcinoma: a clinicopathological correlation. Anticancer Res. *19*, 3853–3856.

Cho, R.J., Huang, M., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S.J., Davis, R.W., Lockhart, D.J., and Campbell, M.J. (2001). Transcriptional regulation and function during the human cell cycle. Nat. Genet. *27*, 48–54.

Diehn, M., Eisen, M.B., Botstein, D., and Brown, P.O. (2000). Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat. Genet. *25*, 58–62.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14688.

Golub, T.R., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Gressner, A.M., Krull, N., and Bachem, M.G. (1994). Regulation of proteoglycan expression in fibrotic liver and cultured fat-storing cells. Pathol. Res. Pract. *190*, 864–882.

Hasan, F., Jeffers, L.J., De Medina, M., Reddy, K.R., Parker, T., Schiff, E.R., Houghton, M., Choo, Q.L., and Kuo, G. (1990). Hepatitis C-associated hepatocellular carcinoma. Hepatology *12*, 589–591.

Hayasaka, A., Suzuki, N., Fujimoto, N., Iwama, S., Fukuyama, E., Kanda, Y., and Saisho, H. (1996). Elevated plasma levels of matrix metalloproteinase-9 (92-kd type IV collagenase/gelatinase B) in hepatocellular carcinoma. Hepatology *24*, 1058–1062.

Hsu, H.C., Tseng, H.J., Lai, P.L., Lee, P.H., and Peng, S.Y. (1993). Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness. Cancer Res. *53*, 4691–4694.

Johnson, P.J. (2001). The role of serum alpha-fetoprotein estimation in the diagnosis and management of hepatocellular carcinoma. Clin. Liver Dis. *5*, 145–159.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lin, T.Y., Lee, C.S., Chen, K.M., and Chen, C.C. (1987). Role of surgery in the treatment of primary carcinoma of the liver: a 31-year experience. Br. J. Surg. *74*, 839–842.

Liotta, L.A., Tryggvason, K., Garbisa, S., Hart, I., Foltz, C.M., and Shafie, S. (1980). Metastatic potential correlates with enzymatic degradation of basement membrane collagen. Nature *284*, 67–68.

Maehara, Y., Kabashima, A., Koga, T., Tokunaga, E., Takeuchi, H., Kakeji, Y., and Sugimachi, K. (2000). Vascular invasion and potential for tumor angiogenesis and metastasis in gastric carcinoma. Surgery *128*, 408–416.

Murakami, K., Sakukawa, R., Ikeda, T., Matsuura, T., Hasumura, S., Nagamori, S., Yamada, Y., and Saiki, I. (1999). Invasiveness of

hepatocellular carcinoma cell lines: contribution of membrane-type 1 matrix metalloproteinase. Neoplasia *1*, 424–430.

Ng, I.O., Lai, E.C., Chan, A.S., and So, M.K. (1995). Overexpression of p53 in hepatocellular carcinomas: a clinicopathological and prognostic correlation. *J. Gastroenterol. Hepatol. 10*, 250–255.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. *10*, 1–6.

Okabe, H., Satoh, S., Kato, T., Kitahara, O., Yanagawa, R., Yamaoka, Y., Tsunoda, T., Furukawa, Y., and Nakamura, Y. (2001). Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res. *61*, 2129–2137.

Okuda, K., Obata, H., Nakajima, Y., Ohtsuki, T., Okazaki, N., and Ohnishi, K. (1984). Prognosis of primary hepatocellular carcinoma. Hepatology *4*, 3S–6S.

Perou, C.M., *et al.* (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA *96*, 9212–9217.

Perou, C.M., *et al.* (2000). Molecular portraits of human breast tumors. Nature *406*, 747–752.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet. *23*, 41–46.

Ross, D.T., *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Sato, H., Takino, T., Okada, Y., Cao, J., Shinagawa, A., Yamamoto, E., and Seiki, M. (1994). A matrix metalloproteinase expressed on the surface of invasive tumor cells. Nature *370*, 61–65.

Sherlock, G., *et al.* (2001). The Stanford Microarray Database. Nucleic Acids Res. *29*, 152–155.

Shiraishi, K., *et al.* (2001). A comparison of DNA copy number changes detected by comparative genomic hybridization in malignancies of the liver, biliary tract and pancreas. Oncology *60*, 151–161.

Vazquez, F., Hastings, G., Ortega, M.A., Lane, T.F., Oikemus, S., Lombardo, M., and Iruela-Arispe, M.L. (1999). METH-1, a human ortholog of ADAMTS-1, and METH-2 are members of a new family of proteins with angio-inhibitory activity. J. Biol. Chem. *274*, 23349–23357.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA *98*, 1176–1181.

Wong, N., Lai, P., Lee, S.W., Fan, S., Pang, E., Liew, C.T., Sheng, Z., Lau, J.W., and Johnson, P.J. (1999). Assessment of genetic changes in hepatocellular carcinoma by comparative genomic hybridization analysis: relationship to disease stage, tumor size, and cirrhosis. Am. J. Pathol. *154*, 37–43.

Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W.H., Tom, E., Mack, D.H., and Levine, A.J. (2000). Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev. *14*, 981–993.