Oncogene (2005) 24, 3737–3747

www.nature.com/onc

An integrated data analysis approach to characterize genes highly expressed in hepatocellular carcinoma

Mohini A Patil[1,6], Mei-Sze Chua[2,6], Kuang-Hung Pan[3,6], Richard Lin[3], Chih-Jian Lih[3], Siu-Tim Cheung[4], Coral Ho[1], Rui Li[2], Sheung-Tat Fan[4], Stanley N Cohen[3], Xin Chen[1,5] and Samuel So[2]

[1]Department of Biopharmaceutical Sciences, University of California, San Francisco, CA 94143, USA; [2]Department of Surgery and Asian Liver Center, Stanford University, Stanford, CA 94305, USA; [3]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [4]Department of Surgery and Center for the Study of Liver Disease, University of Hong Kong, Hong Kong, China; [5]Liver Center, University of California, San Francisco, CA 94143, USA

Hepatocellular carcinoma (HCC) is one of the major causes of cancer deaths worldwide. New diagnostic and therapeutic options are needed for more effective and early detection and treatment of this malignancy. We identified 703 genes that are highly expressed in HCC using DNA microarrays, and further characterized them in order to uncover novel tumor markers, oncogenes, and therapeutic targets for HCC. Using Gene Ontology annotations, genes with functions related to cell proliferation and cell cycle, chromatin, repair, and transcription were found to be significantly enriched in this list of highly expressed genes. We also identified a set of genes that encode secreted (e.g. GPC3, LCN2, and DKK1) or membrane-bound proteins (e.g. GPC3, IGSF1, and PSK-1), which may be attractive candidates for the diagnosis of HCC. A significant enrichment of genes highly expressed in HCC was found on chromosomes 1q, 6p, 8q, and 20q, and we also identified chromosomal clusters of genes highly expressed in HCC. The microarray analyses were validated by RT–PCR and PCR. This approach of integrating other biological information with gene expression in the analysis helps select aberrantly expressed genes in HCC that may be further studied for their diagnostic or therapeutic utility.
Oncogene (2005) **24,** 3737–3747. doi:10.1038/sj.onc.1208479
Published online 21 February 2005

Keywords: HCC; microarrays; gene expression; tumor markers

Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer, and the fourth leading cause of cancer deaths worldwide (Parkin, 2001; Parkin *et al*., 2001). Epidemiological and molecular genetic studies have demonstrated that the development of HCC spans several decades, often starting with hepatitis B virus (HBV) or hepatitis C virus (HCV) infections. Chronic carriers of HBV or HCV are at much higher risk of developing HCC, especially when infection has been accompanied by liver cirrhosis (El-Serag, H., 2001; El-Serag, H.B., 2002). The options for effective treatment of HCC are limited to either local surgical resection or liver transplantation (Lin *et al*., 1987; Mor *et al*., 1998; Helton *et al*., 2003). However, as early HCC lesions typically are associated with few symptoms, most patients present clinically with advanced stages of HCC, precluding surgical treatment and thus contributing partly to the dismal 5- year survival rate for HCC (approximately 7% in the US) (http://www.cancer.org/). Currently, the most commonly used method for screening high-risk populations, serum determination of alpha fetoprotein (AFP), fails to identify a significant portion of HCC patients, especially those in the early stages of the disease (Nguyen and Keeffe, 2002). Thus, the development of new approaches for better detection and treatment of HCC in its early stages is critical to improving the rate of survival from this type of cancer.

DNA microarray technology provides a high-throughput means of identifying genes that may be highly expressed in tumor cells by surveying the expression levels of tens of thousands of genes on an unbiased basis. Using this technology, we and others have reported the expression profiles of liver cancer cell lines and human samples (Okabe *et al*., 2001; Shirota *et al*., 2001; Chen *et al*., 2002; Lee and Thorgeirsson, 2002; Smith *et al*., 2003; Ye *et al*., 2003; Lee *et al*., 2004; Neo *et al*., 2004). These studies have mainly focused on specific aspects of HCC development, for example, genes that may be related to HBV or HCV infection or genes that may predict HCC outcome. In an earlier study, we used cDNA microarrays containing 17 000 unique human genes to study the gene expression profiles in over 200 liver tissue samples, including 102 primary HCC (from 82 patients) and 74 nontumor liver tissues (Chen *et al*., 2002; Cheung *et al*., 2002). Consistent differences were found between the

Correspondence: Professor Samuel So, Department of Surgery and the Asian Liver Center, 300 Pasteur Drive, H3680, Stanford University, Stanford, CA 94305, USA; E-mail: samso@stanford.edu or X Chen, Department of Biopharmaceutical Sciences, 513 Parnassus Avenue, University of California, San Francisco, CA 94143-0446, USA; E-mail: xinchen@itsa.ucsf.edu
[6]These authors contributed equally to this work and are listed in random order
Received 22 October 2004; revised 14 December 2004; accepted 27 December 2004; published online 21 February 2005

expression patterns in HCC compared with those seen in nontumor liver tissues. Moreover, specific phenotypic and genotypic characteristics of the tumors, such as growth rate, vascular invasion, and p53 overexpression, were found to be associated with some features of the gene expression patterns. Here, to form a rational basis on which to select genes that have greater likelihoods of becoming clinically useful diagnostic markers or therapeutic targets of HCC, we conduct new analysis of these microarray data to identify genes overexpressed in HCC liver compared to nontumor liver. We employed an integrated analytical approach that encompasses Gene Ontology, chromosomal locations, and DNA copy number assessments to provide information for identifying and characterizing potential diagnostic markers and therapeutic targets.

Results

Expression profiles of genes highly expressed in HCC

Global expression profiles of more than 200 samples, including 102 primary HCC tissues (from 82 patients), 74 nontumor liver tissues, seven benign liver tumor samples, 10 metastatic cancers, and 10 HCC cell lines have been previously reported (Chen *et al*., 2002). We retrieved the previously obtained gene expression profiles of 156 samples, including 82 HCC (where there were multiple samples from different tumor nodules of the same patient, only one representative sample was used) and 74 nontumor liver tissues from Stanford Microarray Database (SMD) (http://smd.stanford.edu//). Data for 4863 cDNA clones with the most varying expression among the tumor and nontumor liver samples were further selected for analysis. We applied Significance Analysis of Microarrays (SAM) (Tusher *et al*., 2001) to the data set and identified 1946 DNA clones with statistically significant changes in expression between HCC and nontumor liver. The median number of false significance for the SAM analysis was less than 1 and false discovery rate was less than 0.1%. Among the 1946 clones, 866 cDNA clones, representing 703 unique genes, were highly expressed in HCC samples compared to nontumor liver tissues (Supplemental Table 1). Among this list of 703 unique genes (which we designate as the 'overexpressed in HCC' gene list), approximately 570 were named genes, whereas 133 represented ESTs (expressed sequence tags) and unnamed genes such as hypothetical protein.

As a preliminary analysis of the 'overexpressed in HCC' gene list, we applied hierarchical clustering to arrange the genes according to the similarity of their expression profiles (Figure 1a). We identified several groups of genes based on the delineation by the hierarchical clustering dendrogram (Figure 1b–h). These groupings confirm previously identified groupings of hepatocellular carcinoma gene expression in earlier publications (Iyer *et al*., 1999; Chen *et al*., 2002, 2004; Whitfield *et al*., 2002), as well as suggest some new clusters such as the histone cluster (Figure 1b), the ribosomal protein cluster (Figure 1d), and clusters that contain several genes on chromosome 1q (Figure 1f and h) and a cluster that contains several genes on chromosome 8q (Figure 1e). The discovery of these gene groupings suggested the need for a more extensive and rigorous analysis using GABRIEL (Pan *et al*., 2002), a supervised analysis approach that incorporates other

Table 1 Biological processes (A), cellular components (B), and molecular functions (C) nonrandomly enriched with the products of the 703 overexpressed genes in HCC-bearing gene ontology annotations

Category	*NumGenes*	*P-value*
(A) Biological process		
Physiological process	311	1.26E-19
Cellular process	222	1.58E-14
Cell growth and/or maintenance	153	6.72E-12
Cellular physiological process	166	2.00E-11
Cell cycle	55	6.53E-10
Metabolism	214	1.31E-09
Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism	108	2.01E-09
Cell proliferation	71	2.58E-09
M phase	25	2.43E-07
Mitotic cell cycle	26	2.83E-07
DNA metabolism	39	3.59E-07
M phase of mitotic cell cycle	21	4.40E-07
DNA replication and chromosome cycle	22	5.37E-06
Mitosis	19	8.31E-06
Regulation of mitosis	10	3.24E-05
Nuclear division	20	0.00012
Regulation of cell cycle	31	0.00016
DNA replication	16	0.000909
Protein metabolism	81	0.006386
Microtubule-based process	10	0.010424
Nucleosome assembly	9	0.012154
Cell cycle checkpoint	7	0.026506
Cytokinesis	12	0.042423
(B) Cellular component		
Intracellular	220	4.34E-17
Cell	282	9.27E-16
Nucleus	122	4.90E-11
Chromosome	22	5.57E-06
microtubule cytoskeleton	18	4.15E-05
Spindle	11	0.000268
Cytoplasm	106	0.001639
Microtubule organizing center	8	0.003069
Nucleosome	7	0.016334
Spindle pole	7	0.034906
(C) Molecular function		
Binding	228	2.34E-10
Nucleic acid binding	92	1.17E-05
Catalytic activity	144	0.003056
Nucleotide binding	62	0.005417
DNA binding	65	0.007914
Hydrolase activity, acting on acid anhydrides	25	0.012467
Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	25	0.012467
Purine nucleotide binding	60	0.014064
ATP binding	50	0.01633
Adenyl nucleotide binding	50	0.026607

NumGenes: number of genes. *P*-value: corrected *P*-value, calculated by the GO-TermFinder program. A *P*-value threshold of 0.05 is used

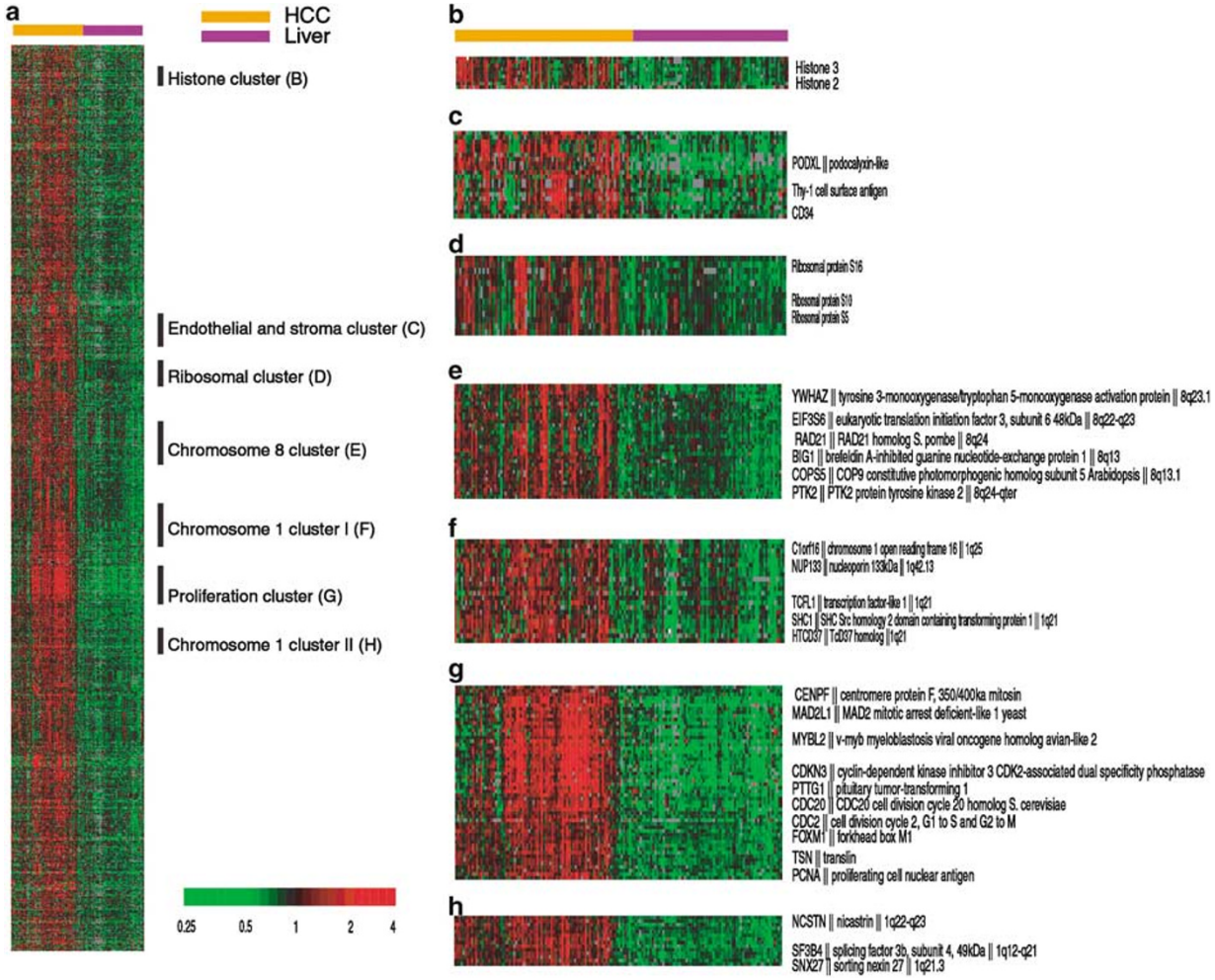

Figure 1 (**a**) Hierarchical clustering of 866 cDNA clones, representing 703 unique genes that were found to be highly expressed in HCC samples compared with nontumor liver tissues. Rows represent individual genes and columns represent individual tissue samples. In each liver tissue sample, the ratio of abundance of transcripts of each gene relative to its mean abundance across all tissue samples is depicted according to the color scale shown at the bottom. Gray indicates missing or excluded data. (**b**) to (**h**) Expanded views of specific clusters within the genes that are highly expressed in HCC. (**b**) histone cluster; (**c**) endothelial/stromal gene cluster; (**d**) ribosomal gene cluster; (**e**) chromosome 8 gene cluster; (**f**) chromosome 1 gene cluster I; (**g**) proliferation cluster; (**h**) chromosome 1 gene cluster II. Only selected gene names are labeled. See Supplementary Table 1 of the full list of genes

pieces of biological information in the interpretation of gene expression profiles.

Gene functional category analysis

To investigate the biological functions involved in the genes overexpressed in HCC, we analysed the Gene Ontology category for the overexpressed genes in HCC gene list. We identified nonrandom enrichment of a variety of biological process categories, including cell growth and maintenance, cell cycle, metabolism, and cell proliferation (Table 1). Such nonrandom enrichment suggests that metabolism, cell cycle, growth, and proliferation may be involved in HCC development. When we compared the expression pattern of these 703 genes with those expressed during the cell cycle in HeLa cells (Whitfield *et al*., 2002), we found that 149 unique genes, or 21.1% of the 703 genes, had previously shown fluctuating expression during the cell cycle. The cell cycle-regulated gene signature included genes that are found to peak during the G2 phase (TOP2A, CDC2, and CCNA2); genes involved in DNA replication (DDX11 and geminin); genes involved in DNA packaging (CHAF1A), DNA repair (FEN1 and PCNA), and cell cycle control (E2F1 and CDC25A) (Whitfield *et al*., 2002). In addition, overexpressed in HCC genes also showed enrichment for annotations related to the cellular components including nucleus and cytoplasm, and the molecular functions such as binding and catalytic activity (Table 1). These biological gene categories enriched in HCC may provide directions for future research of the molecular mechanisms of HCC.

Secreted and membrane-bound proteins as potential tumor markers

Secreted and membrane proteins, although not significantly over-represented in the list of HCC overexpressed

Table 2 Top 10 genes in the 'overexpressed in HCC' gene list that are designated as (A) secretory or extracelluar, and (B) membrane by Gene Ontology

Gene symbol	*Unigene cluster ID*	*Fold change*
(A) Secretory or extracelluar proteins		
AFP	Hs.155421	61.54
GPC3	Hs.119651	45.19
LCN2	Hs.204238	8.07
DKK1	Hs.40499	6.73
CPLX2	Hs.193235	4.97
STC2	Hs.155223	2.65
COL7A1	Hs.1640	2.56
GABRE	Hs.22785	2.35
STC1	Hs.25590	2.10
SPARC	Hs.111779	2.02
(B) Membrane proteins		
GPC3	Hs.119651	45.19
IGSF1	Hs.22111	6.98
PSK-1	Hs.6314	6.03
FADS2	Hs.388164	4.37
CD24	Hs.375108	4.12
CAP2	Hs.296341	3.98
C1orf2	Hs.348308	3.37
CD34	Hs.374990	3.26
SCD	Hs.119597	3.17
OPN3	Hs.170129	3.11

See Supplementary Table 2 for the complete list of genes

genes, were further examined because of their important potential clinical value. Upregulated genes encoding secreted proteins are of particular interest because of their potential diagnostic (noninvasive detection in peripheral blood) value. Of the 703 up-regulated genes in tumor samples, nine genes were designated to be in secretory pathways or secretion, and 28 genes were annotated with the cellular component of the extracellular or soluble fraction (Table 2A and Supplemental Table 2A). Importantly, AFP, a widely accepted clinical blood marker of HCC, was among the elevated expression genes designated as extracellular, thereby increasing our confidence that new diagnostic markers may be discovered through this analysis.

Upregulated genes encoding membrane-bound proteins present opportunities for tumor targeting for both diagnostic (histochemical) as well as therapeutic (targeting of conventional chemotherapeutic agents) applications. Among the 703 upregulated genes, 98 were designated as membrane associated (Table 2B and Supplemental Table 2B). Of potential biological significance is our observation that seven members (SLC38A6, SLC1A4, SLC39A10, SLC26A2, SLC36A1, SLC29A1, SLC26A6) of different solute carrier families are present in this list, raising the prospect that these membrane transporters may play important functional roles in the pathophysiology of hepatoma cells, probably by enhancing the transport of essential nutrients needed to support the rapidly growing tumor cells. Several genes are annotated as being both extracellular as well as membrane-associated (ATRN, CKLF, GABRE, GPC3, RTN3), suggesting that these genes may function as both peripheral and histochemical markers of HCC.

RT–PCR validation of microrarray analysis

Owing to the considerable extent of variation routinely observed in microarray data (Nielsen *et al*., 2002; Bohen *et al*., 2003), we confirmed our microarry results using RT–PCR as an independent method of analysis. To accomplish this, we studied a separate set of 16 samples (frozen HCC and adjacent nontumor liver tissue from eight patients) by RT–PCR. A total of 11 genes (HBI, NCSTN, PTTG1, SXN27, ESP8R3, LCN2, PISK1, MYBL2, GPC3, CENPF, and E2F1) were randomly selected from the 'overexpressed in HCC' gene list and their expression was quantitated by RT–PCR (Figure 2a). Overall, the RNA expression patterns of these genes were similar regardless of the technique: GPC3 and NCSTN were found to be highly expressed in all eight HCC samples compared to the corresponding nontumor liver samples; PTTG1 and E2F1 were highly expressed in seven of the HCC samples; HBI, SXN27, MYBL2, and CENP were highly expressed in 6 HCC samples; and LCN2 and PISK1 were highly expressed in five and four HCC samples, respectively. In addition, the expression patterns of five genes, which belong to the 'endothelial/stromal cell cluster', have also been validated to be highly expressed in HCC (Chen *et al*., 2004).

To further verify the expression profile of genes identified as potential candidates for novel secretory or membrane-associated tumor markers, we performed real-time quantitative RT–PCR on eight genes selected from Tables 2A and B (six encoding for secretory and two for membrane proteins) in 16 liver samples. In all cases, we found that these genes are statistically significantly highly expressed in HCC samples compared with non-neoplastic liver tissues.

In situ *hybridization study of cellular origin of candidate genes*

Normal and neoplastic liver are complex and heterogeneous tissues composed of diverse specialized cell types such as sinusoid endothelial cells, stromal cells, infiltrating T cells, B cells, and Kupffers cells. The microarray study used total RNA that was extracted from whole tissues, thus generating gene expression profiles that are a composite of gene expression signatures from different specialized cell types within the tissues. We hypothesized that identifying the cells in which the genes of interest are expressed in HCC would help in dissecting their functional contributions to the pathophysiology of HCC.

We observed a cluster of genes comprising markers specific to endothelial cells and stromal cells (Figure 1c). Among these genes, we have previously shown (using *in situ* hybridization and immunohistochemistry) that RGS5, PODXL, and CD34 were highly expressed in sinusoidal endothelial cells in HCC samples, whereas COL4A2, SPARC1, and THY1 were highly expressed in

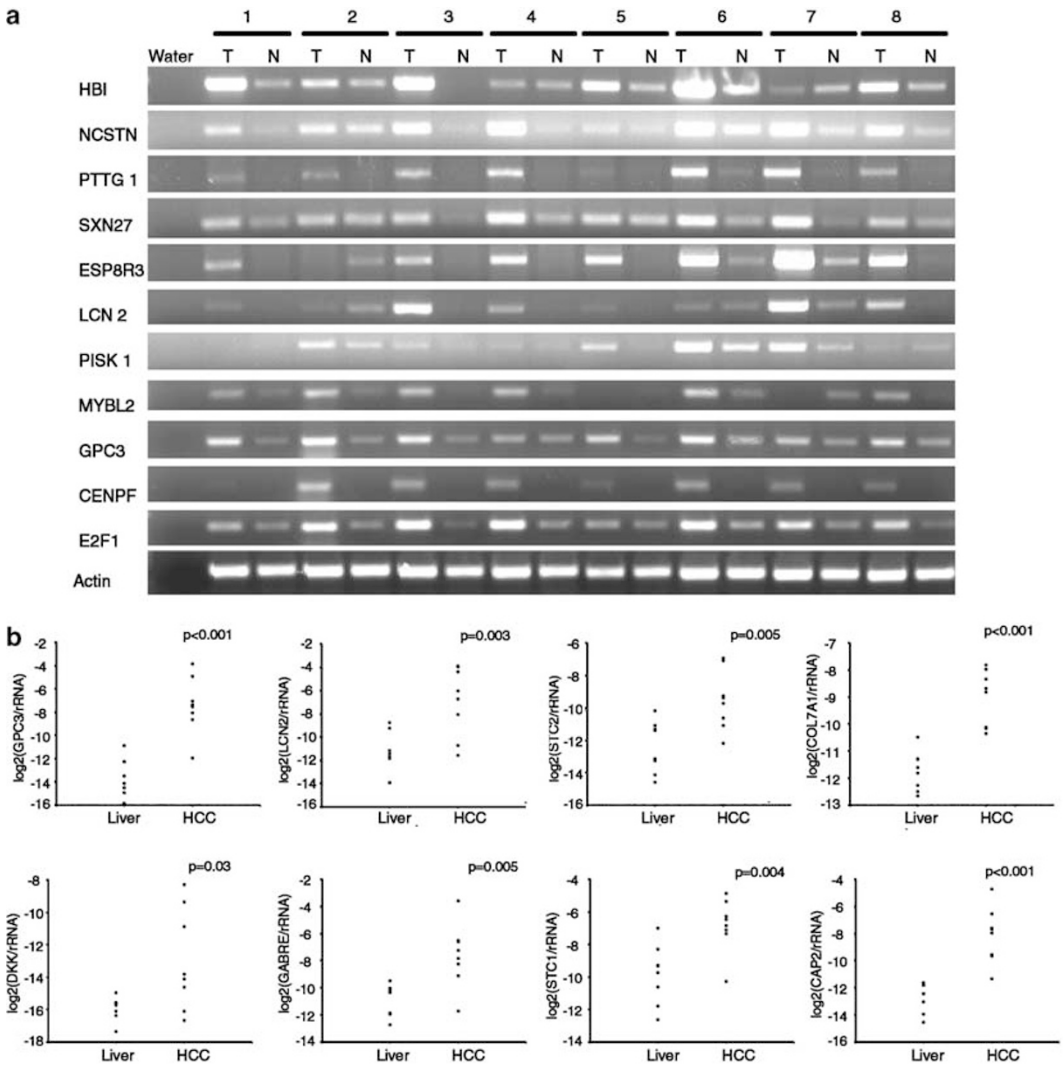

Figure 2 (**a**) Validation by RT–PCR of some genes identified from the microarray data. RT–PCR was used to analyse eight independent pairs of nontumor liver (N) and HCC (T) samples for 11 genes. Actin was used as a loading control. (**b**) Validation of eight genes that encode secreted or membrane-bound proteins

both endothelial and stromal cells in HCC samples (Chen *et al*., 2004). However, in our Gene Ontology analysis, we did not find any significant cell-type-specific gene categories related to T cells, B cells, or endothelial cells, suggesting that the gene overexpression we observe is likely from the malignant tumor cells rather than due to contaminating cells.

We used *in situ* hybridization to study the cellular localization of six genes: GPC3, NCSTN, PLK, LCN2, PSK1, and PTTG1. All six genes were found to be expressed in malignant hepatocytes as seen by the brown grain staining of each gene-specific antisense probe (Figure 3). Sense probes used as a control showed no signal, demonstrating the specificity of the antisense signals. Negative staining was observed for NCSTN, PLK, LCN2, PSK1, and PTTG1 and a weak staining was observed for GPC3 in non-neoplastic liver tissues (Supplementary Figure 1). If this observation holds true for the other overexpressed genes, the fact that these genes are expressed within the malignant tumor cells rather than in contaminating nontumor cells would suggest that some of these genes may have the potential to be new oncogenes, tumor markers, or therapeutic targets of HCC.

Chromosomal enrichment and clustering of genes highly expressed in HCC

Genomic DNA copy number gains or amplifications are shown to play important roles during development of HCC (Nishida *et al*., 2003). Interestingly, based on chromosomal location information retrieved from public databases, we observed that genes in the 'overexpressed in HCC' gene list were statistically enriched on chromosomes 1q, 6p, 8q, and 20q (Figure 4). This suggests that there might be common DNA copy number gains of HCC among chromosomes 1q, 6p, 8q, and 20q, which contribute to the high

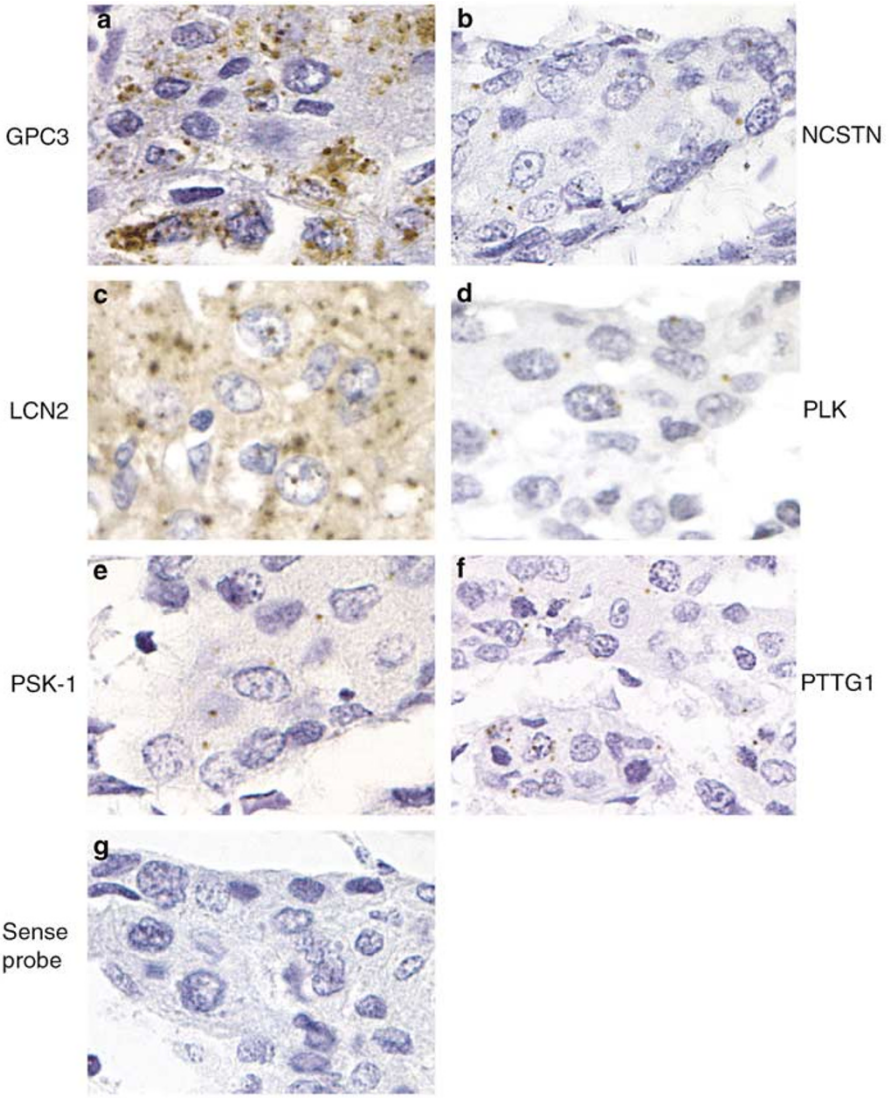

Figure 3 *In situ* hybridization analysis of the expression of candidate genes in HCC paraffin sections. (**a**)–(**f**) Antisense riboprobes against GPC3, NCSTN, LCN2, PLK, PSK-1, PTTG1 were used; brown grains denote positive signal. (**g**) A representative sense control probe shows no signal

expression of the genes located on these chromosomes. The coordinate expression of these genes indicated that DNA copy number gains within specific chromosomal regions or the regulation of the chromatin domains may contribute significantly to the expression programs of these genes. Corroborating this observation, several groups have shown recurrent chromosomal copy gains at 1q, 6p, 8q, and 20q by comparative genomic hybridization (Nishida *et al*., 2003).

To further investigate the relationship between gene expression and gene location, we looked for the existence of chromosomal clustering of overexpressed genes. A chromosomal cluster was defined as two or more genes located within a specified linear distance on the chromosome. We observed significant chromosomal clustering of genes when using distance thresholds between 20–200 kb (1 kb = 1000 base pairs) (Table 3). For example, using 200 kb as the distance threshold defining cluster boundaries, 166 of the 703 genes were found to be chromosomally clustered (false positive rate < 0.01). Using this method, we also observed that genes overexpressed in HCC were statistically significantly clustered on chromosomes 1q, 6p, 8q, 20p, 20q, and Xq (Supplementary Figure 1) with distance threshold of 200 kb. Some of these chromosomal clusters may point out directions for studying oncogenes or markers. For example, C1orf2 (chromosome 1 open reading frame 2), CLK2 (CDC-like kinase 2), FDPS (farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase, dimethylallyltranstransferase, geranyltranstransferase)), and RUSC1 (RUN and SH3 domain containing 1) form a cluster (within 100 kb apart) on chromosome 1q22, among which C1orf2 has a membrane bound protein product. MCP (membrane cofactor protein) and CD34

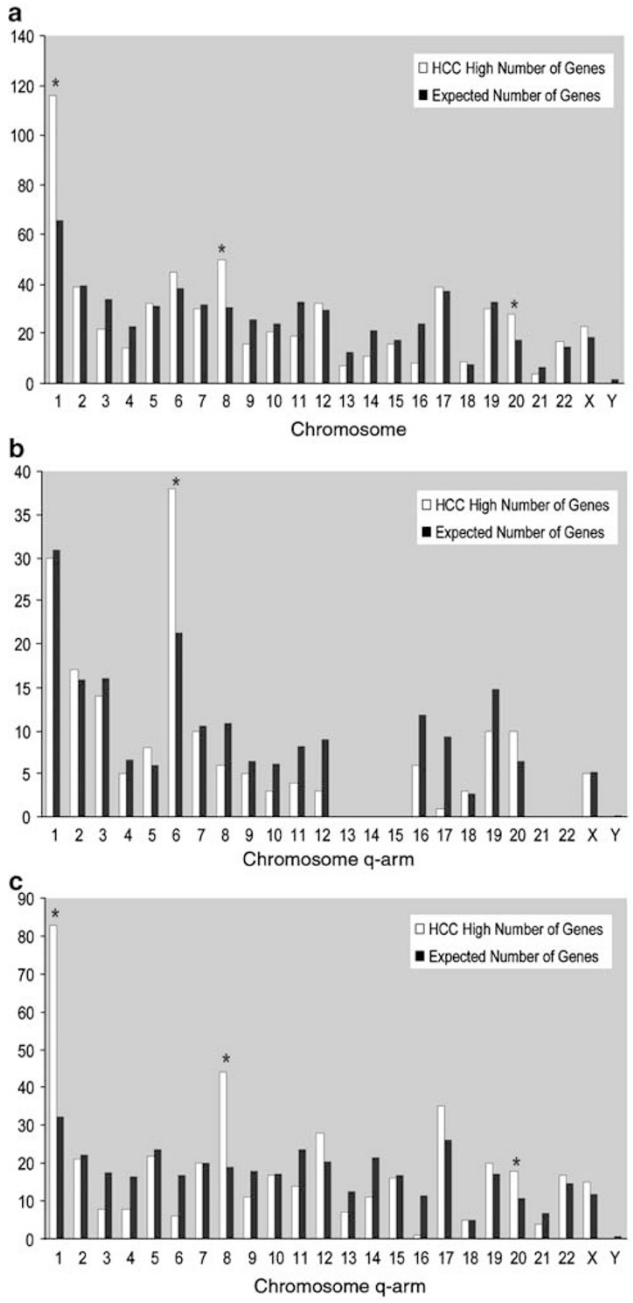

Figure 4 Chromosomal enrichment of genes overexpressed in HCC. The number of genes on each chromosome (**a**) or on the p-arm (**b**) or q-arm (**c**) of the chromosomes was compared with the average values estimated from randomly selected lists containing the same number of genes. Asterisks indicate statistical significance ($P<0.01$)

(CD34 antigen), two membrane bound proteins, form a cluster (within 100 kb apart) on chromosome 1q32 (Supplemental Table 3).

In addition to identifying chromosomal arms that have significant clustering of HCC overexpressed genes, we also observed specific chromosomal clusters on other chromosomes that may be of interest. For example, we found that COL7A1 (collagen, type VII, alpha 1) and

Table 3 Effect of distance threshold on the determination of chromosomal clustering for 703 genes that are upregulated in HCC

Distance threshold (Kb)	*Number of clusters*	*Average number of false clusters*	*False positive rate*
1	0	0.04	0.04
2	1	0.46	0.04
5	2	1.77	0.24
10	6	3.41	0.04
20	15	7.15	<0.01
50	35	16.65	<0.01
100	52	29.49	<0.01
200	74	49.51	<0.01
500	104	89.47	0.02
1000	129	123.44	0.18
2000	147	153.63	0.8
5000	145	166.32	1

Significant clustering of closely located genes was observed when using distance thresholds of 20, 50, 100, and 200 kb. A chromosomal cluster is defined as two or more genes located within the specified distance. The false positive rate indicates the probability that the difference between the observed number of clusters (or genes) and the average number of false clusters (or genes) (estimated by bootstrapping algorithm) is due to random chance. A false positive rate of <0.01 is considered significant

SLC26A6 (solute carrier family 26, member 6) form a chromosomal cluster (~60 kb apart) on chromosome 3p21, even though chromosome 3p was not previously observed to be overexpressed in HCC. COL7A1 is involved in secretory pathways, while SLC26A6 is a membrane bound protein. In addition, MAZ (MYC-associated zinc-finger protein), an oncogene-associated transcription factor, and PSK-1 (type I transmembrane receptor), a membrane bound protein, form a cluster (~60 kb apart) on chromosome 16q. Furthermore, there are a number of cell cycle genes clustered in chromosome 17q21: CDC6 (CDC6 cell division cycle 6 homolog), TOP2A (topoisomerase (DNA) II alpha), TUBG1 (tubulin, gamma 1), and TUBG2 (tubulin, gamma 2). These chromosomal clusters suggest chromosomal regions that may regulate coordinated gene expression in HCC, and potential sources for searching for diagnostic markers and therapeutic targets of HCC (Supplemental Table 3).

To investigate whether overexpression of genes in HCC could be associated with DNA amplification, we assessed the DNA copy numbers of four overexpressed genes located at chromosome 1q: SHC1 (Src homology 2 domain-containing transforming protein 1) at 1q22 (152.15–152.16 Mb); YAP (YY1-associated protein) at 1q22 (152.84–152.87 Mb); PRCC (papillary renal cell carcinoma gene 1) at 1q23.1 (153.95–153.98 Mb); and KIAA0205 at 1q32.3 (208.99–209.08 Mb). DNA copy number variations were determined using real-time quantitative PCR with primers and probes specific for intron sequences of each of these genes. Human GAPDH was used as a standard control as it is located at 12p13.31, a region with little variation among HCC samples as assayed by array-based comparative genomic hybridization (Patil *et al*, unpublished data). In the four nontumor samples and 20 HCC samples asssayed, the

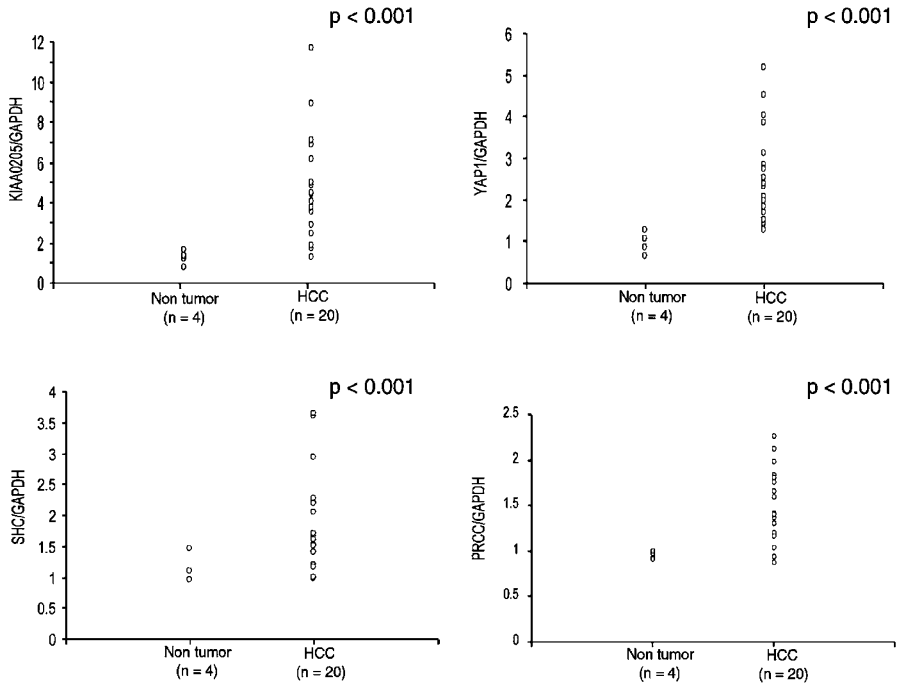

Figure 5 DNA copy number of selected genes on chromosome 1 in HCC and nontumor liver tissues analysed using real-time quantitative PCR. The *P*-values were calculated using Student's *t*-test

average DNA copy number in the four nontumor liver samples for all four genes was about 1 (Figure 5), suggesting little DNA copy number variations in these nontumor liver tissues at chromosome 1q. DNA copy numbers of all four genes tested were significantly higher in HCC samples than nontumor liver samples (Figure 5) and confirmed that there are indeed recurrent DNA copy number gains at chromosome 1q in HCC samples.

Discussion

The molecular genetics underlying the development of HCC remains largely unknown. Currently, only a handful of genes are known to be highly expressed in HCC (Feitelson *et al*., 2002). Expanding this pool of genes highly expressed in HCC and understanding any role they may possibly have in the pathogenesis of HCC is important, both for clinical and basic research. Clinically, among the genes highly expressed in HCC, the ones that encode secreted molecules potentially may be employed as diagnostic or disease-progression markers, whereas genes expressed on the cell surface may be useful as targets for HCC treatment. For basic research, the identification of candidate oncogenes may assist in elucidation of the pathogenesis of liver tumors. Identification of signaling molecules that are enriched among the products of highly expressed genes in HCC may also shed light on the critical signaling pathways involved in HCC tumorigenesis. Finally, the mapping of highly expressed genes to their chromosomal locations may identify hot spots of chromosomal aberrations, characterize chromosomal regulation of gene expression in HCC, or enable correlation of genomic DNA copy number variations with altered gene expression in HCC. In summary, the knowledge gleaned from this comprehensive analysis of genes that are highly expressed in HCC provides guidelines for the selection of potentially novel diagnostic markers and therapeutic targets of HCC for further validation and functional evaluation.

We report in this paper a comprehensive analysis of 703 genes that global gene expression profiling has shown to be highly overexpressed in HCC compared to nontumor liver. Interestingly, we found that while all these genes showed statistically significant elevation of expression in HCC, and uniformly showed comparatively low expression in nontumor liver tissues, gene expression varied greatly among different HCC samples, that is, some HCCs showed high expression of certain genes, while expression in other HCCs was similar to what was observed in nontumor liver tissue (Figure 1). This observation suggests genetic heterogeneity among HCC samples, which could result from differences in the oncogenic pathways that lead to the development of HCC. The clinical implication of this observed heterogeneity among HCC samples is that altered expression of an individual gene may not be sufficient as a tumor marker, and that most likely a set of markers will need to be tested for their ability to accurately diagnose or monitor the progression of this malignancy.

Heterogeneity among HCC samples may explain why AFP alone is not a sufficiently reliable diagnostic marker for HCC, and underscores the importance of finding new markers that might be used in combination

with AFP to increase the accuracy of HCC diagnosis. Our Gene Ontology based analysis for identifying potential diagnostic markers appears to be statistical reliable, since our gene set includes the currently accepted marker AFP, as well as others that have been reported to have diagnostic values in HCC. Most notably, among the genes designated to be secretory or extracellular, AFP showed the greatest fold change, providing retrospective rationale for its clinical use as the current standard for HCC diagnosis. Additionally, GPC3 has recently been detected in both blood and tissue sections of HCC patients (Capurro *et al*., 2003; Hippo *et al*., 2004), and PGCP has been suggested to be a potential marker of HCC through microarray analyses of hepatitis C virus-associated HCC (Smith *et al*., 2003). We are in the process of systematically validating these potential new markers to confirm the relative abundance of their encoded proteins in serum of HCC patients compared to healthy individuals. This will hopefully lead to the identification of clinically useful markers that can be used either alone or in conjunction with AFP to increase the sensitivity and specificity of diagnosing early HCC. Additionally, since specificity for the target organ is an important criterion in the selection of tumor markers, we also intend to further study the expression of our candidate membrane-bound markers in organs other than the liver.

Carcinogenesis is a multistep process arising from a combination of multiple genetic and epigenetic events. Genetic changes such as amplification or gain of oncogenes and deletion or mutation of tumor suppressor genes lead to genomic instability, which subsequently allows for the selection of genetic traits that confer a growth advantage for tumor progression (Jallepalli and Lengauer, 2001; Gollin, 2004). Using GABRIEL, we identified an enrichment of genes on chromosomes 1q, 6p, 8q, and 20q. Genes that are located close to each other on the chromosome may be coregulated by similar mechanisms. We found that physical clusters of two or more genes within predefined distance thresholds were nonrandomly present on chromosomes 1q, 6p, 8q, 20p, 20q, and Xq. A previous study that mapped genes overexpressed in HCC to chromosomal locations found that certain chromosomal arms had a disproportionately high number of such genes (Crawley and Furge, 2003). Our investigation indicates that HCC-related genes are physically clustered at specific chromosomal locations. Some of the HCC-related gene clusters we identified are present on chromosomal arms that showed overall enrichment for such genes in the Crawley and Furge study, while others are located on chromosomal arms that show no such enrichment.

Real-time quantitative PCR confirmed the increase in DNA copy numbers of four overexpressed genes on chromosome 1q, suggesting that there may be common chromosomal gains in HCC. Studies of DNA copy number variations of HCC on a genomic scale using array-based CGH (aCGH) would also be valuable; this can be carried out using BAC clones or cDNA clones to give high-resolution mapping of chromosomal aberrations (Pinkel *et al*., 1998; Pollack *et al*., 1999). Combined knowledge from these data will help to better define new candidate oncogenes and therapeutic targets, as well as to increase our understanding of the molecular genetics of the development of HCC.

Materials and methods

cDNA microarrays and data analysis

The method of cDNA microarray analysis, including liver tissue acquisition, RNA extraction, and microarray hybridization has been described previously (Chen *et al*., 2002).

For data analysis, the raw data for 156 arrays (representing 82 HCC and 74 nontumor liver tissues) were downloaded from the SMD (http://smd.stanford.edu//). Data were filtered using the following criteria: genes with 80% of spots with >2.5-fold intensities over background in either channel, 80% good data, and genes whose log(base2) of Red/Green normalized ratio (mean) was greater than three fold for at least three arrays. We retrieved 4863 cDNA clones corresponding to approximately 3718 unique Unigene clusters. We applied the SAM method (Tusher *et al*., 2001) to identify genes that were significantly differentially expressed in HCC and nontumor liver tissues. We also applied a hierarchical clustering algorithm and average linkage clustering to the selected genes using the Pearson correlation as a measure of similarity as previously described (Eisen *et al*., 1998). The results were visualized and analysed using Tree View.

Chromosomal location and clustering analyses were carried out using the GABRIEL microarray analysis platform (Pan *et al*., 2002). Chromosomal locations for all IMAGE clones were retrieved from the human genome working draft (http://genome.ucsc.edu) using their accession numbers. Analyses were carried out as previously described (Zhang *et al*., 2003). Gene ontology categories were analysed by the GO-TermFinder Program (Boyle *et al*., 2004). Genes were classified according to their annotated role in biological processes, molecular function, and cellular components from Gene Ontology (The Gene Ontology Consortium). The human Gene Ontology annotation was downloaded from the European Bioinformatics Institute (http://www.ebi.ac.uk/GOA/). The GO::TermFinder program identifies genes belonging to different Gene Ontology categories. We identified genes with secreted protein products by using GO::TermFinder to search for genes with the cellular component annotated as secretory pathways, secretion, extracelluar or soluble fraction, as well as genes with membrane-associated protein products. GO::TermFinder also calculates the statistical significance of nonrandom representation, that is, enrichment, of a Gene Ontology category among the genes under investigation by hypergeometric distribution (Boyle *et al*., 2004). Genes with multiple clones on the microarray are counted only once in the Gene Ontology analysis using the Unigene symbol (Unigene). We use the 3718 unique Unigenes that were retrieved from the data set according to the filtering criteria as the background population in the hypergeometric calculation of statistical significance of nonrandom representation of a Gene Ontology category. This analysis also provides the list of genes in the secretory or membrane categories.

RT–PCR

RT–PCR was performed using the ThermoScript RT–PCR system (Invitrogen). In brief, $1\,\mu$g of total RNA was used in $20\,\mu$l reverse transcription assay. In total, $2\,\mu$l of the reverse

transcription product was subsequently used in a 25 μl PCR reaction. All PCR was performed at 52–55°C annealing temperature for 25–32 cycles. The primer sequences for PCR for each of the genes are available in Supplemental Table 4A.

Quantitative real-time RT–PCR

Real-time quantitative RT–PCR was performed using SYBR Green kit and ABI 7900HT Sequence Detection System (Applied Biosystems). The primer sequences for each gene tested are listed in Supplemental Table 4B. Human 18 s rRNA detection kit was obtained from Applied Biosystems and was used as a loading control. Calibration curves were generated for each gene and validated using linear regression analysis ($r^2 \geqslant 0.98$). Expression levels of each gene was performed in triplicate for every sample and reported relative to 18 s rRNA.

Nonradioactive in situ *hybridization of paraffin sections*

Nonradioactive *in situ* hybridization was performed as described (Chen *et al*., 2004). Digoxigenin (DIG)-labeled sense and antisense RNA probes were generated by PCR amplification of 400–600 bp products with the T7 promote incorporated into the primers. *In vitro* transcription was performed with DIG RNA labeling kit and T7 polymerase according to the manufacturer's protocol (Roche Diagnostics, Indianopolis, IN, USA). Sections (5 μm thick) were cut from the paraffin blocks, deparaffinized in xylene, and hydrated in graded concentrations of ethanol for 5 min each. Tissue sections were then incubated with 1% hydrogen peroxide, followed by digestion in 10 μg/ml of proteinase K at 37°C for 30 min. They were then hybridized overnight at 55°C with either sense or antisense riboprobes at 200 ng/ml dilution in mRNA hybridization buffer (Dako). The following day, hybridized sections were washed in 2 × SSC and incubated with 1 : 35 dilution of RNase A cocktail (Ambion, Austin, TX, USA) in 2 × SSC for 30 min at 37°C. Next, they were stringently washed in 2 × SSC/50% formamide twice, followed by one wash at 0.08 × SSC at 50°C. Biotin blocking reagents (Dako) were applied to the section to block the endogenous biotin. For signal amplification, a HRP-conjugated rabbit anti-DIG antibody (Dako) was used to catalyse the deposition of biotinyl-tyramide, followed by secondary streptavidin complex (GenPoint kit; Dako). The final signal was developed with DAB (GenPoint kit; Dako), and the tissues were counterstained in hematoxylin for 15 s. The primer sequences used for the generation of probes are available in Supplemental Table 4C.

Quantitative real-time PCR

The DNA copy number variations of four genes located on chromosomal 1q were determined by quantitative PCR carried out on the ABI Prism® 7900HT Sequence Detection System (Applied Biosystems). The primer and probe sequences for each gene tested are listed in Supplemental Table 4D. The housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as a control for PCR analysis as it is located at chromosome 12p, a region with little genomic DNA variations in HCC (Patil *et al*, unpublished data). Calibration curves were generated for each gene using normal male or female genomic DNA and validated using linear regression analysis ($r^2 \geqslant 0.98$). DNA copy number quantification was performed in triplicate for every sample and reported relative to GAPDH.

Acknowledgements
We are grateful to Stanford Functional Genomic Center and Stanford Microarray database for their support. This work is supported by a grant to the Asian Liver Center at Stanford University by the HM Lui Foundation (to M-SC, RL, and SS), the UCSF Liver Center (DK26743-22) pilot/feasibility project award and NCI K01 award (to XC), and grants from the National Foundation for Cancer Research and the Defense Advanced Projects Research Agency (DARPA) (to SNC). K-H Pan is supported by a Stanford Graduate Fellowship.

References

Bohen SP, Troyanskaya OG, Alter O, Warnke R, Botstein D, Brown PO and Levy R. (2003). *Proc. Natl. Acad. Sci. USA*, **100,** 1926–1930.
Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM and Sherlock G. (2004). *Bioinformatics*, **20,** 3710–3715.
Capurro M, Wanless IR, Sherman M, Deboer G, Shi W, Miyoshi E and Filmus J. (2003). *Gastroenterology*, **125,** 89–97.
Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D and Brown PO. (2002). *Mol. Biol. Cell*, **13,** 1929–1939.
Chen X, Higgins J, Cheung ST, Li R, Mason V, Montgomery K, Fan ST, Rijn Mv M and So S. (2004). *Mod. Pathol.*, **17,** 1198–1210.
Cheung ST, Chen X, Guan XY, Wong SY, Tai LS, Ng IO, So S and Fan ST. (2002). *Cancer Res.*, **62,** 4711–4721.
Crawley JJ and Furge KA. (2003). *Genome Biol.*, **3** research 0075.1-0075.8.
Eisen MB, Spellman PT, Brown PO and Botstein D. (1998). *Proc. Natl. Acad. Sci. USA*, **95,** 14863–14868.
El-Serag H. (2001). *Clin. Liver Dis.*, **5,** 87–107.
El-Serag HB. (2002). *J. Clin. Gastroenterol.*, **35,** S72–S78.
Feitelson MA, Sun B, Satiroglu Tufan NL, Liu J, Pan J and Lian Z. (2002). *Oncogene*, **21,** 2593–2604.
Gollin SM. (2004). *Curr. Opin. Oncol.*, **16,** 25–31.
Helton WS, Di Bisceglie A, Chari R, Schwartz M and Bruix J. (2003). *J. Gastrointest. Surg.*, **7,** 401–411.
Hippo Y, Watanabe K, Watanabe A, Midorikawa Y, Yamamoto S, Ihara S, Tokita S, Iwanari H, Ito Y, Nakano K, Nezu J, Tsunoda H, Yoshino T, Ohizumi I, Tsuchiya M, Ohnishi S, Makuuchi M, Hamakubo T, Kodama T and Aburatani H. (2004). *Cancer Res.*, **64,** 2418–2423.
Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson Jr J, Boguski MS, Lashkari D, Shalon D, Botstein D and Brown PO. (1999). *Science*, **283,** 83–87.
Jallepalli PV and Lengauer C. (2001). *Nat. Rev. Cancer*, **1,** 109–117.
Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, Durnez A, Demetris AJ and Thorgeirsson SS. (2004). *Hepatology*, **40,** 667–676.
Lee JS and Thorgeirsson SS. (2002). *Hepatology*, **35,** 1134–1143.
Lin TY, Lee CS, Chen KM and Chen CC. (1987). *Br. J. Surg.*, **74,** 839–842.
Mor E, Kaspa RT, Sheiner P and Schwartz M. (1998). *Ann. Intern. Med.*, **129,** 643–653.
Neo SY, Leow CK, Vega VB, Long PM, Islam AF, Lai PB, Liu ET and Ren EC. (2004). *Hepatology*, **39,** 944–953.
Nguyen MH and Keeffe EB. (2002). *J. Clin. Gastroenterol.*, **35,** S86–S91.
Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR,

Brown PO, Botstein D and van de Rijn M. (2002). *Lancet*, **359,** 1301–1307.
Nishida N, Nishimura T, Ito T, Komeda T, Fukuda Y and Nakao K. (2003). *Histol. Histopathol.*, **18,** 897–909.
Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, Tsunoda T, Furukawa Y and Nakamura Y. (2001). *Cancer Res.*, **61,** 2129–2137.
Pan K-H, Lih C-J and Cohen SN. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 2118–2123.
Parkin DM. (2001). *Lancet Oncol.*, **2,** 533–543.
Parkin DM, Bray F, Ferlay J and Pisani P. (2001). *Int. J. Cancer*, **94,** 153–156.
Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW and Albertson DG. (1998). *Nat. Genet.*, **20,** 207–211.
Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO. (1999). *Nat. Genet.*, **23,** 41–46.
Shirota Y, Kaneko S, Honda M, Kawai HF and Kobayashi K. (2001). *Hepatology*, **33,** 832–840.
Smith MW, Yue ZN, Geiss GK, Sadovnikova NY, Carter VS, Boix L, Lazaro CA, Rosenberg GB, Bumgarner RE, Fausto N, Bruix J and Katze MG. (2003). *Cancer Res.*, **63,** 859–864.
Tusher VG, Tibshirani R and Chu G. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 5116–5121.
Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO and Botstein D. (2002). *Mol. Biol. Cell*, **13,** 1977–2000.
Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, Ma ZC, Wu ZQ, Ye SL, Liu YK, Tang ZY and Wang XW. (2003). *Nat. Med.*, **9,** 416–423.
Zhang H, Pan KH and Cohen SN. (2003). *Proc. Natl. Acad. Sci. USA*, **100,** 3251–3256.

Supplementary Information accompanies the paper on Oncogene website (http://www.nature.com/onc)