Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning

Qing-Hai Ye[1], Lun-Xiu Qin[1], Marshonna Forgues[2], Ping He[2], Jin Woo Kim[2], Amy C. Peng[3,4], Richard Simon[3], Yan Li[1], Ana I. Robles[2], Yidong Chen[5], Zeng-Chen Ma[1], Zhi-Quan Wu[1], Sheng-Long Ye[1], Yin-Kun Liu[1], Zhao-You Tang[1] & Xin Wei Wang[2]

[1]*Liver Cancer Institute and Zhongshan Hospital, Fudan University, Shanghai, China*
[2]*Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA*
[3]*Biometrics Research Branch, National Cancer Institute, Rockville, Maryland, USA*
[4]*EMMES Corp., Rockville, Maryland, USA*
[5]*Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland, USA*
Correspondence should be addressed to X.W.W.; e-mail: xw3u@nih.gov

Published online 17 March 2003; doi:10.1038/nm843

Hepatocellular carcinoma (HCC) is one of the most common and aggressive human malignancies. Its high mortality rate is mainly a result of intra-hepatic metastases. We analyzed the expression profiles of HCC samples without or with intra-hepatic metastases. Using a supervised machine-learning algorithm, we generated for the first time a molecular signature that can classify metastatic HCC patients and identified genes that were relevant to metastasis and patient survival. We found that the gene expression signature of primary HCCs with accompanying metastasis was very similar to that of their corresponding metastases, implying that genes favoring metastasis progression were initiated in the primary tumors. Osteopontin, which was identified as a lead gene in the signature, was over-expressed in metastatic HCC; an osteopontin-specific antibody effectively blocked HCC cell invasion *in vitro* and inhibited pulmonary metastasis of HCC cells in nude mice. Thus, osteopontin acts as both a diagnostic marker and a potential therapeutic target for metastatic HCC.

HCC is a common and aggressive malignant tumor with especially high prevalence in Asia and Africa and relatively low prevalence in Europe and North America[1,2]. Recent studies indicate that the incidence of HCC in the US and UK has increased substantially over the last two decades[3,4]. Although routine screening of individuals at risk for developing HCC may extend the life of some patients, many are still diagnosed with advanced HCC and have little chance of survival[5–9]. A small subset of HCC patients qualifies for surgical intervention, but the consequent improvement in long-term survival is only modest[10,11]. The extremely poor prognosis of HCC is largely the result of a high rate of recurrence after surgery or of intra-hepatic metastases that develop through invasion of the portal vein or spread to other parts of the liver; extra-hepatic metastases are less common[12,13]. These data indicate that the liver is the main target organ of HCC metastasis. The portal vein is the main route for intra-hepatic metastases of HCC cells in animal model systems and in human patients[14–16]. This feature of HCC underscores the need to develop an accurate molecular profiling model to improve diagnosis and identify therapeutic targets for the treatment of HCC patients with intra-hepatic metastases.

Current studies have been largely focused on individual candidate genes[17–19], an approach that may be insufficient to precisely define the genetic basis of metastatic HCC. Microarray technology allows us to examine disease-related gene expression on a global genome scale[20]. This approach has resulted in successful molecular classification of several human malignant tumors with respect to their stage, prognostic outcome or response to therapy[21–26]. A few reports have dealt with the gene expression profiles of primary HCC samples[27,28], but little is known about the molecular signature associated with a poor prognosis for metastatic HCC.

We applied cDNA microarray–based gene expression profiling to investigate the global changes associated with HCC metastasis. Our initial goal was to identify genes that distinguish primary tumors from their matched intra-hepatic metastatic lesions. Unexpectedly, we found that the intra-hepatic metastatic lesions were indistinguishable from their primary tumors, regardless of tumor size, encapsulation and age of patient. Primary metastasis-free HCC was distinct from primary HCC with metastasis. These data indicate that changes favoring intra-hepatic metastasis are initiated in the primary HCC. In addition, we showed that osteopontin, a secreted phosphoprotein, is a significant factor in HCC metastasis. Osteopontin over-expression correlated with metastatic potential of primary HCC and with invasiveness of liver tumor–derived cell lines *in vitro*. An osteopontin-neutralizing antibody efficiently blocked *in vitro* invasion and *in vivo* pulmonary metastasis of HCC cells. Our studies

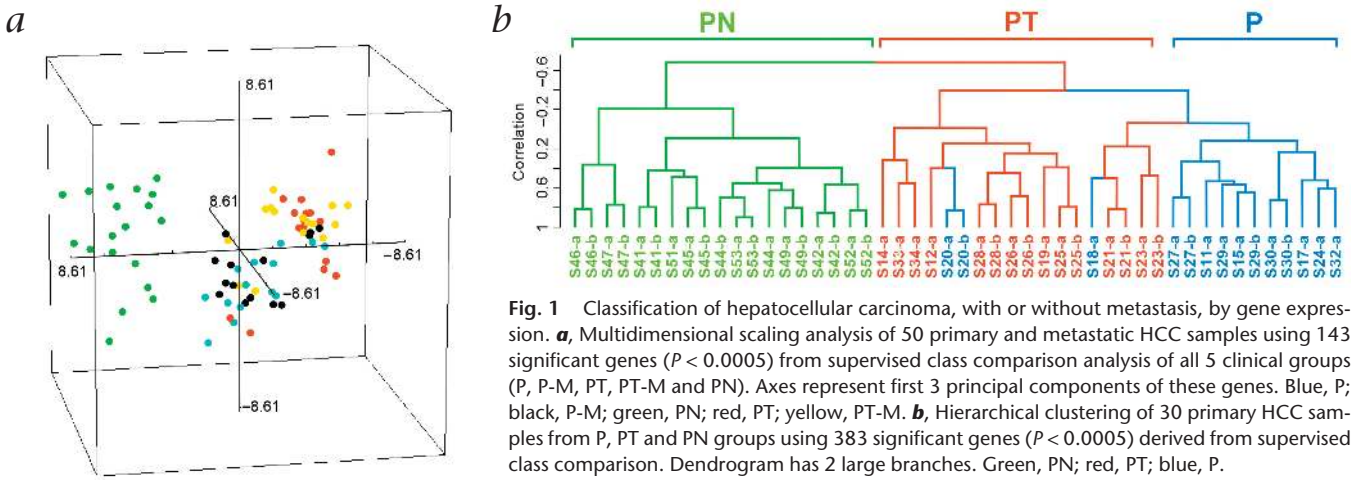

Fig. 1 Classification of hepatocellular carcinoma, with or without metastasis, by gene expression. ***a***, Multidimensional scaling analysis of 50 primary and metastatic HCC samples using 143 significant genes ($P < 0.0005$) from supervised class comparison analysis of all 5 clinical groups (P, P-M, PT, PT-M and PN). Axes represent first 3 principal components of these genes. Blue, P; black, P-M; green, PN; red, PT; yellow, PT-M. ***b***, Hierarchical clustering of 30 primary HCC samples from P, PT and PN groups using 383 significant genes ($P < 0.0005$) derived from supervised class comparison. Dendrogram has 2 large branches. Green, PN; red, PT; blue, P.

identify osteopontin as both a molecular marker for defining HCC patients with metastatic potential and a potential therapeutic target for metastatic HCC.

Metastatic lesions are identical to primary tumors

To define the specific changes associated with metastatic progression, we compared gene expression profiles of 67 primary and metastatic HCC samples from 40 patients. We assessed intra-hepatic spread (group P) or tumor thrombi in the portal vein (group PT), as well as matched metastatic lesions (P-M or PT-M, respectively) and corresponding non-cancerous liver tissues. Initial analysis included 50 primary and metastatic tumor samples from 30 randomly selected patients (ten in metastasis-free HCC group PN, ten in group PT and ten in group P; see Supplementary Note online). In contrast, the unsupervised hierarchical clustering algorithm (which used either all 9,180 genes or ~2,487 genes after those not significantly more variable than the median at $P < 0.01$ were excluded) did not yield any meaningful classification that corresponded to predefined clinical groups (Supplementary Fig. 1 and data not shown). Similar results were obtained with a 107-gene set from a two-fold cutoff filter (see Supplementary Fig. 2 online). These results imply that primary and metastatic HCC differ by a relatively small subset of genes, but the gene clustering algorithm may be dominated by variations among many other genes, thus hindering classification.

To search for such small differences, we applied a supervised class comparison analysis with univariate *F*-tests and a global permutation test to define genes that were differentially expressed among predefined clinical groups. A comparison of five clinical groups (P, P-M, PT, PT-M and PN) yielded a total of 143 significant genes ($P < 0.0005$). Multidimensional scaling analysis, based on the first three principal components of these genes, showed that the PN samples were distinct from the remaining samples, whereas the P, P-M, PT and PT-M samples were indistinguishable (Fig. 1*a*). Unexpectedly, the gene expression profiles of primary and matched metastatic HCC tumors were not significantly different. Similar results were obtained when a class comparison was applied only to primary HCC samples with 383 significant genes ($P < 0.0005$; Fig. 1*b*). Thus, primary metastasis-free HCC has a gene expres-

Table 1 Performance of classifier during 'leave-one-out' cross-validation[a]

Classifier category	Clinical groups	Total number of cases	Number of cases misclassified	Classifier *P* value	Number of genes in the classifiers
PN vs. PT	PT	10	0	<0.0005	153
	PN	10	0		
PN vs. P	PN	10	1	<0.0005	157
	P	10	0		
PN vs. P and PT	PN	10	2	<0.001	256
	P and PT	20	0		
P vs. PT	P	10	3	0.216	20
	PT	10	4		
PT vs. PT-M	Paired samples	10	3	0.296	1
P and PT vs. P-M and PT-M	Paired samples	20	5	0.132	7
P vs. PT-M	P	10	4	0.248	14
	PT-M	10	3		
PT vs. P-M	PT	10	2	0.163	9
	P-M	10	4		
Tumor diameter	>5 cm	16	7	0.234	7
	≤5 cm	14	4		
Ages	>45 years	17	5	0.334	4
	≤45 years	13	7		
Tumor encapsulation	Presence	9	2	0.037	13
	Absence	21	4		
Cirrhosis	Presence	14	7	0.798	1
	Absence	6	6		

[a]CCP was used to classify various clinical groups with a total of 9,180 genes at a significance level of $P = 0.001$. Classifier was based on 2,000 random permutations. Expected number of false-positive genes in classifier is 10.

sion profile markedly different from that of primary HCC with metastatic lesions.

To further define a gene set that could accurately classify metastatic HCC, we used a supervised machine-learning classification algorithm known as compound co-variate predictor (CCP). This algorithm includes a 'leave-one-out' cross-validation test to avoid the statistical problem of over-estimating prediction accuracy that occurs when a model is trained and evaluated with the same samples[29]. This analysis also creates a multivariate predictor for determining which one of the two classes a given sample belongs to, and a gene list that is univariately significant at a given level of statistical significance. We applied CCP to various pairs of 50 HCC samples from 30 patients (Table 1). Again, we found no significant difference between primary HCCs and their matched metastatic lesions. Gene expression profiles in P and PT samples were almost identical to their paired metastatic P-M and PT-M samples. Similarly, no significant difference could be found between P-M and PT-M (see Supplementary Table 1 online).

We accurately classified primary tumors (100%) from ten PN and ten PT samples with a total of 153 significant genes in the classifier. The cross-validated misclassification rates were significantly lower than expected by chance ($P < 0.0005$; Table 1). Similar results were obtained when PN was compared with P and when PN was compared with P and PT (Table 1). Again, no significant difference was observed among P, PT, PT-M and P-M. Moreover, no significant difference was found with age, tumor size, tumor encapsulation or cirrhosis (Table 1 and Supplementary Table 1 online). Thus, it appears that primary and metastatic tumors have a similar gene expression signature whereas metastasis-free primary HCC is distinct from metastatic primary HCC.

Predicting metastatic HCC samples

CCP analysis of PN and PT samples generated a classifier containing 153 genes with weights that can be used to predict new samples (see Supplementary Table 2 online). We applied these weights to a test set containing 20 primary HCCs (15 P, 2 PT and 3 PN patients). Calculated 'weighted voting' of L values with metastatic samples yielded negative values; non-metastatic samples yielded positive values (Fig. 2). All test samples, with the exception of one P sample (patient S29), were classified in the metastatic group (Fig. 2*a*). However, patient S29 was more similar to the P and PT groups than to that of the PN group by

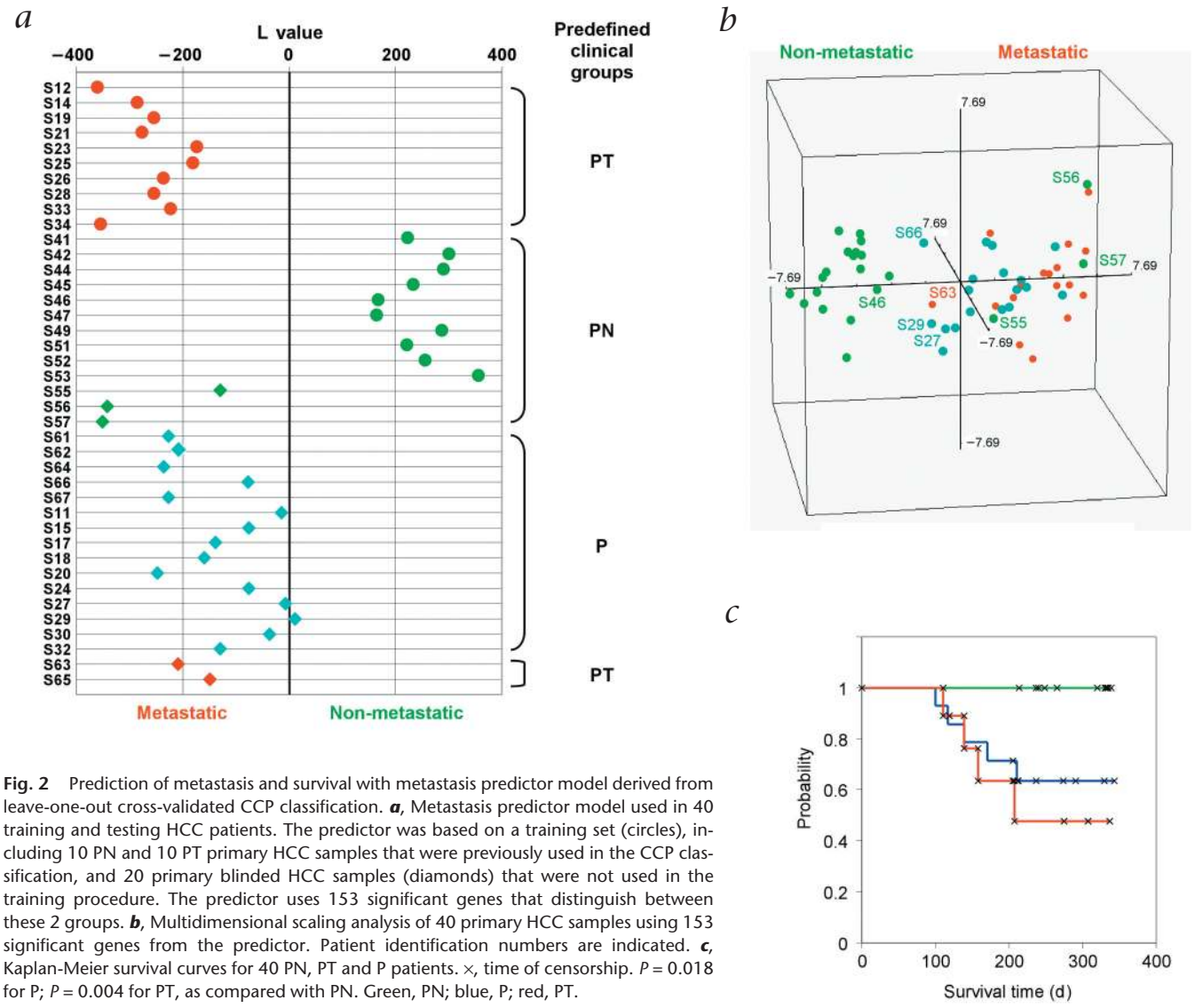

Fig. 2 Prediction of metastasis and survival with metastasis predictor model derived from leave-one-out cross-validated CCP classification. ***a***, Metastasis predictor model used in 40 training and testing HCC patients. The predictor was based on a training set (circles), including 10 PN and 10 PT primary HCC samples that were previously used in the CCP classification, and 20 primary blinded HCC samples (diamonds) that were not used in the training procedure. The predictor uses 153 significant genes that distinguish between these 2 groups. ***b***, Multidimensional scaling analysis of 40 primary HCC samples using 153 significant genes from the predictor. Patient identification numbers are indicated. ***c***, Kaplan-Meier survival curves for 40 PN, PT and P patients. ×, time of censorship. $P = 0.018$ for P; $P = 0.004$ for PT, as compared with PN. Green, PN; blue, P; red, PT.

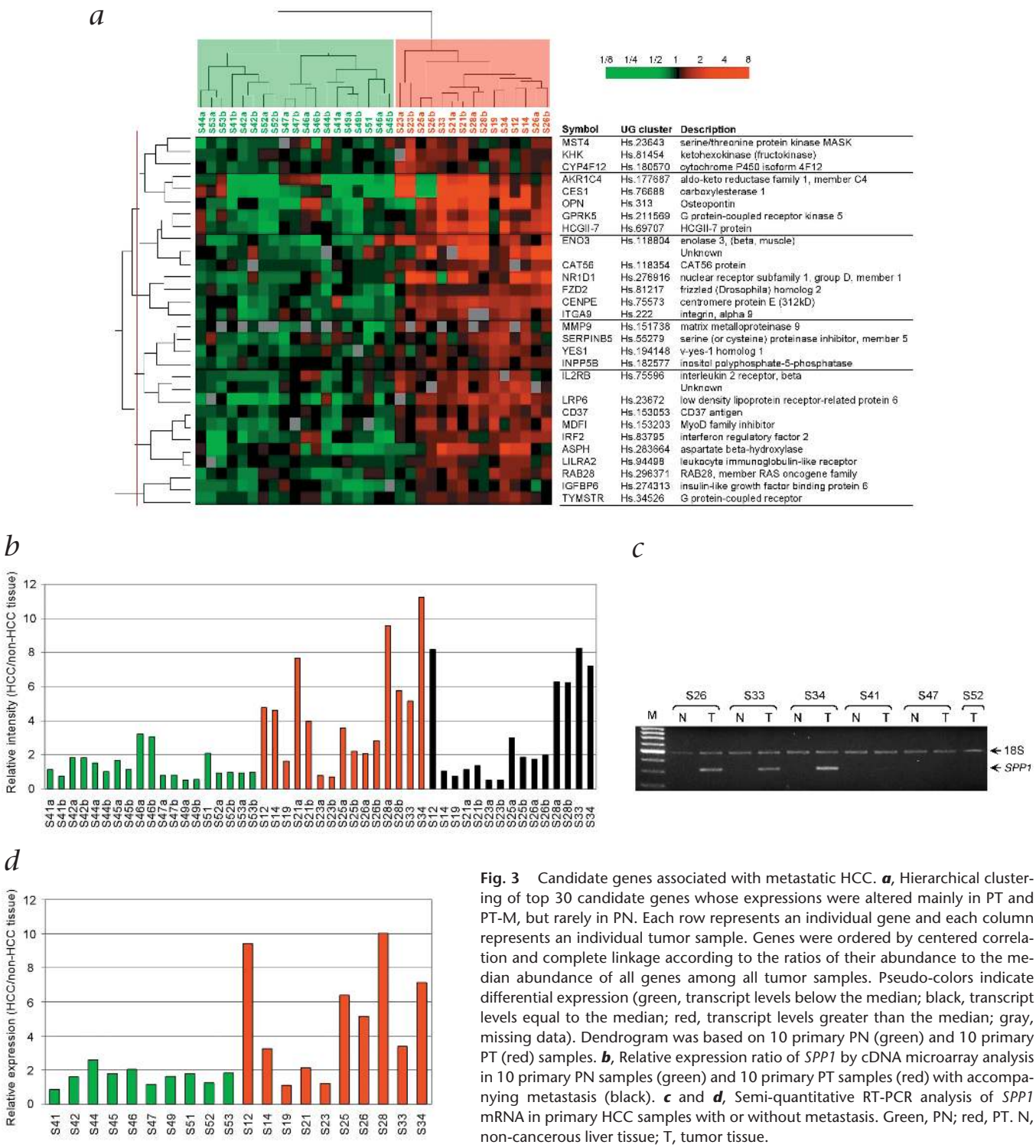

Fig. 3 Candidate genes associated with metastatic HCC. ***a***, Hierarchical clustering of top 30 candidate genes whose expressions were altered mainly in PT and PT-M, but rarely in PN. Each row represents an individual gene and each column represents an individual tumor sample. Genes were ordered by centered correlation and complete linkage according to the ratios of their abundance to the median abundance of all genes among all tumor samples. Pseudo-colors indicate differential expression (green, transcript levels below the median; black, transcript levels equal to the median; red, transcript levels greater than the median; gray, missing data). Dendrogram was based on 10 primary PN (green) and 10 primary PT (red) samples. ***b***, Relative expression ratio of *SPP1* by cDNA microarray analysis in 10 primary PN samples (green) and 10 primary PT samples (red) with accompanying metastasis (black). ***c*** and ***d***, Semi-quantitative RT-PCR analysis of *SPP1* mRNA in primary HCC samples with or without metastasis. Green, PN; red, PT. N, non-cancerous liver tissue; T, tumor tissue.

multidimensional scaling analysis (Fig. 2*b*). Although this model misclassified all three metastasis-free PN patients, follow-up data indicated that patient S56 developed lung metastases eight months after surgery, and patient S57 did not respond to the follow-up request. In addition, CCP analysis of all 40 patients correctly predicted 23 of 28 patients (82%) with metastatic potential and 8 of 12 metastasis-free patients (67%), with an overall accuracy of 78% (see Supplementary Table 3 online). Similar results were obtained with three additional class prediction algorithms (k-nearest neighbor, nearest centroid, and support vector machine; see Supplementary Table 3 online). These algorithms yielded a composition of classifiers containing 85 genes with weights, which can in principle be used to classify future samples.

The above predictors separated 40 patients into metastatic and non-metastatic groups. Kaplan-Meier survival data indicate that

patients who were predicted to be metastatic had substantially shorter survival than metastasis-free HCC patients (Fig. 2*c*). Archived hospital records for deceased HCC patients showed minimum losses of liver function after liver resection (see Supplementary Table 4 online); these patients died at least 111 d after surgery. These findings indicated a low likelihood that patient death resulted from surgical complications. Because the mortality of HCC patients relies largely on whether they develop intra-hepatic metastases, our results indicate that the classifier may provide a signature reflecting HCC metastasis and survival at least for this cohort.

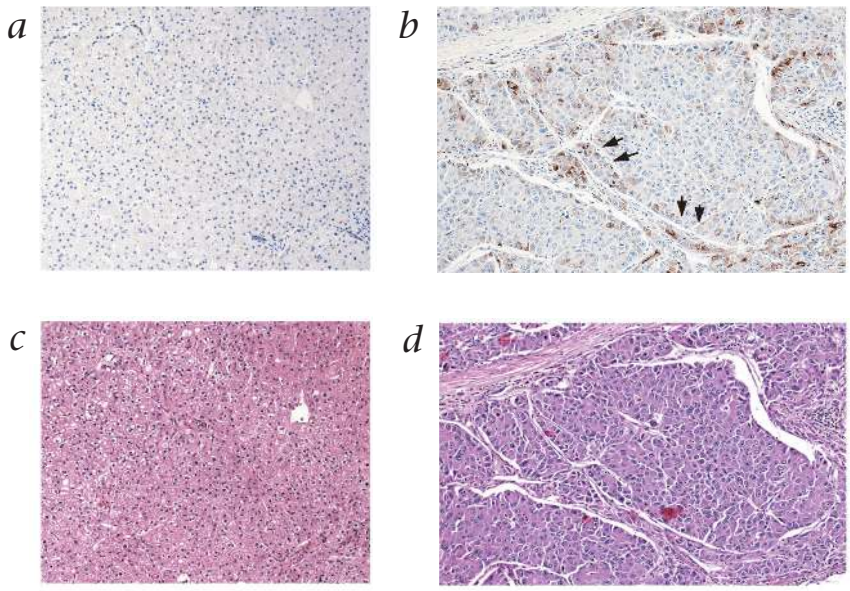

Fig. 4 Immunohistochemical analysis of osteopontin in normal liver and hepatocellular carcinoma. ***a–d***, Primary tumor cells (from tumor S30) show cytoplasmic osteopontin immunoreactivity, especially in the area with a high density of vasculature (*b* (arrows) and *d*), but fibrous septa regions near tumor cells (*b* and *d*; upper left area) or normal liver parenchyma cells (*a* and *c*; from normal liver 914) show no reactivity. H&E stain. Magnification, ×50.

Osteopontin may promote HCC metastasis

The above study indicates that the genes necessary for intra-hepatic metastasis should be included in the prediction model. To broaden our search, we performed univariate *F*-tests at $P < 0.002$ on ten PN and ten PT HCC samples, which yielded 224 significant genes. We selected the top 30 genes whose expression was altered largely in PT and PT-M, but rarely in PN (Fig. 3*a*). A gene with an average of three-fold over-expression in PT, but not in PN, was identified as osteopontin (*SPP1*; Fig. 3*b*), a secreted phosphoprotein highly expressed in patients with metastatic breast tumors and malignant lung, colon and prostate cancers[30,31]. Over-expression of *SPP1* was confirmed using RT-PCR (Fig. 3*c* and *d*). Immunohistochemical analysis of SPP1 was also performed on 29 primary HCC samples (including 16 new HCC cases) and 8 normal livers from healthy organ donors. The immunoreactivity of SPP1 on these samples was evaluated in blind experiments. Only metastatic tumors were positive for cytoplasmic SPP1 staining, especially in the area with a high density of vasculature (Fig. 4*a–d*). The immunohistochemical analysis results mostly agreed with microarray and RT-PCR data (61% positive cases; 11 of 18 metastatic HCC; see Supplementary Table 5 online). Taken together, these studies indicate that *SPP1* has good diagnostic value for metastatic HCC patients.

To determine the role of *SPP1* in metastasis, we used western blotting to compare SPP1 expression in human HCC cell lines and Matrigel assays to compare their *in vitro* invasiveness. SPP1 expression was high in SK-Hep-1, intermediate in Hep3B and low in CCL13 cells (Fig. 5*a*), which coincided with their invasiveness (Fig. 5*b*). An SPP1-neutralizing antibody significantly blocked invasion of SK-Hep-1 ($P < 0.001$) and Hep3B cells ($P < 0.04$). Recombinant mouse Spp1, however, did not show any statistically significant stimulation ($P > 0.05$) of Hep3B and Sk-Hep-1 cells, implying either that SPP1 produced by tumor cells is sufficient for maintaining an invasive phenotype or that the lesser effect of mouse Spp1 is due to species variation. Similar results were obtained with five additional HCC cell lines (Fig. 5*c*). The neutralizing antibody had little effect on cell viability and migration (Fig. 5*c*, right panel).

To extend the above findings, we examined the role of SPP1 in pulmonary metastasis of HCC cells in nude mice. The HCCLM3 cell line is a human clone derived from MHCC97 cells with a high degree of pulmonary metastasis after subcutaneous injection[32]. Consistent with our other data, 100% tumorigenicity was achieved one week after subcutaneous injection. There was no significant difference between the sizes of primary tumors in control and SPP1-neutralizing antibody groups (Fig. 5*e*), which is consistent with our *in vitro* observations that SPP1-specific antibody does not affect HCC cell growth. In the fifth week, pulmonary metastatic lesions were detected in every mouse in the control group, with mainly grade I–II and some grade III–IV tumor clusters (Fig. 5*d* and *f*). The control mice had an average of 11.1 ± 2.9 tumor clusters per lung. In contrast, only about half the mice in the antibody group developed lung metastases; with mostly grade I tumor clusters with a combined average of 2.6 ± 1.0 tumor clusters per lung. This effect was statistically significant ($P < 0.01$). Thus, SPP1-specific antibody significantly inhibits lung metastasis of HCCLM3 cells.

Discussion

HCC patients have a poor prognostic outcome; the major reason is intra-hepatic metastasis that includes tumor thrombi in the portal vein (group PT) and intra-hepatic spread (group P). A clinical challenge is to be able to identify these patients in advance and to identify a therapeutic target for successful intervention. Using gene expression profiling and supervised machine learning, we have developed a strategy to classify HCC patients with intra-hepatic metastasis. Although our model was based on a relatively small set of samples, the 153-gene model provided a robust signature that correctly classified 100% of the training samples during cross-validation. Although it did not predict 3 non-metastatic HCC patients with a testing sample set, the model correctly predicted 17 of 20 new metastatic HCC patients (85%). Similar results were obtained with additional independent prediction algorithms, with an overall predictive value close to 85%. The prediction outcome seemed to correlate with patient survival. Therefore, at least in principle, this model or a similar one may be used to predict metastatic HCC patients.

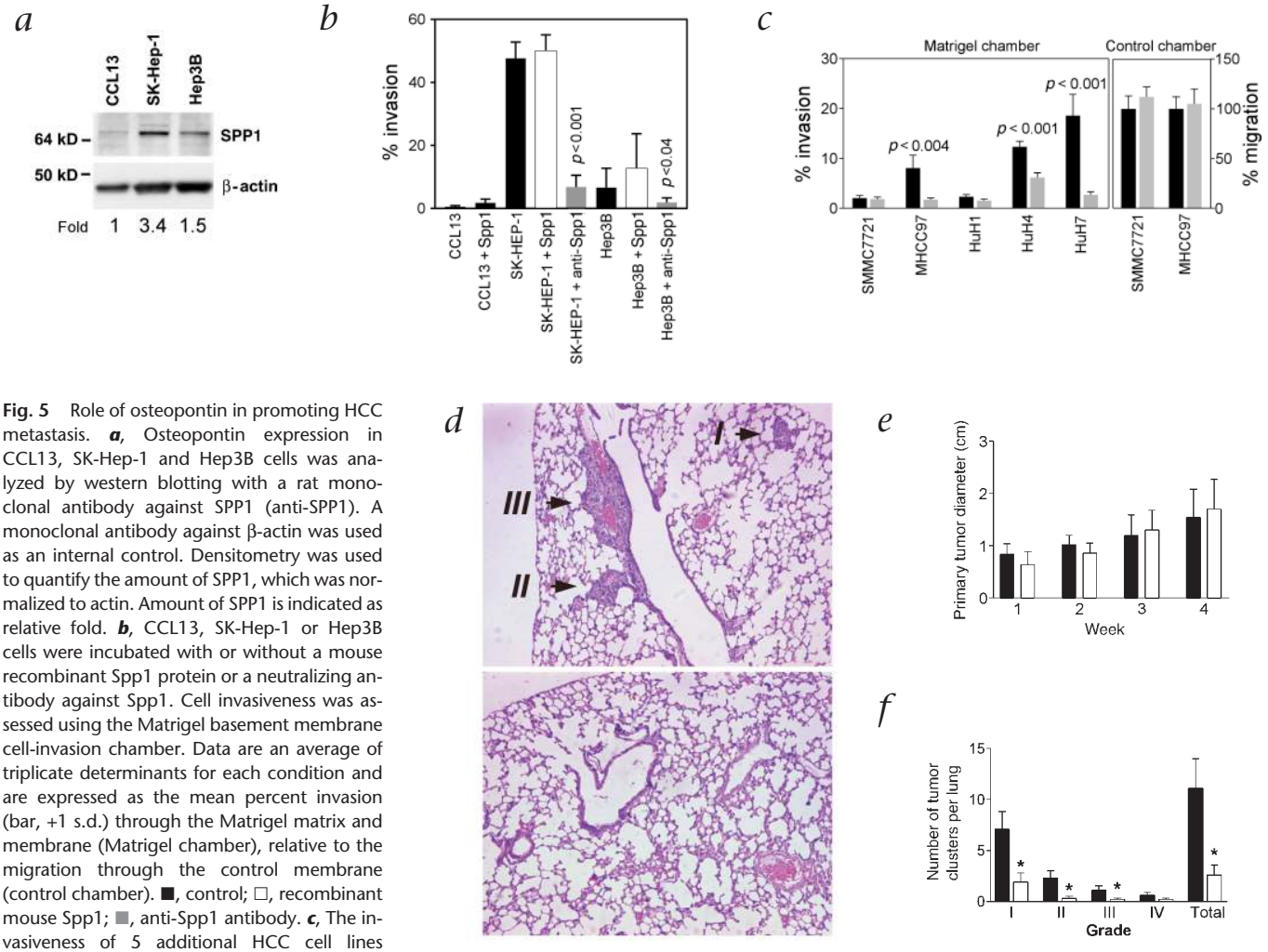

Fig. 5 Role of osteopontin in promoting HCC metastasis. ***a***, Osteopontin expression in CCL13, SK-Hep-1 and Hep3B cells was analyzed by western blotting with a rat monoclonal antibody against SPP1 (anti-SPP1). A monoclonal antibody against β-actin was used as an internal control. Densitometry was used to quantify the amount of SPP1, which was normalized to actin. Amount of SPP1 is indicated as relative fold. ***b***, CCL13, SK-Hep-1 or Hep3B cells were incubated with or without a mouse recombinant Spp1 protein or a neutralizing antibody against Spp1. Cell invasiveness was assessed using the Matrigel basement membrane cell-invasion chamber. Data are an average of triplicate determinants for each condition and are expressed as the mean percent invasion (bar, +1 s.d.) through the Matrigel matrix and membrane (Matrigel chamber), relative to the migration through the control membrane (control chamber). ■, control; □, recombinant mouse Spp1; ■, anti-Spp1 antibody. ***c***, The invasiveness of 5 additional HCC cell lines (SMMC7721, MHCC97, HuH1, HuH4 and HuH7) through Matrigel matrix in the presence or absence of Spp1-neutralizing antibody was assessed as in (*b*) ■, control; ■, anti-Spp1 antibody. ***d***, Representative lung tissue sections (H&E stain; magnification, ×100) from mice 35 d after subcutaneous injection of HCCLM3 cells without (top) or with (bottom) Spp1-neutralizing antibody. Arrows indicate tumor grades. ***e***, Primary tumor diameter was monitored at various times after subcutaneous injection of HCCLM3 cells into nude mice. Data are an average of 10 mice. ***f***, The formation of pulmonary metastases in nude mice was assessed 35 d after subcutaneous injection of HCCLM3 cells with or without Spp1-neutralizing antibody. The number of metastatic foci was quantified based on their grades. Data are an average of 10 mice per group. *, $P < 0.05$. Black bars, control; white bars, Spp1-neutralizing antibody (*e* and *f*).

Because of the small cohort used for this study, however, and because three PN samples were misclassified, this predictor is only suggestive. It has not been entirely validated and awaits confirmation from larger independent data sets to refine and validate its clinical usefulness. In addition, all HCC samples in this study were obtained from hepatitis B virus–positive Chinese patients. It remains to be determined whether this model also can be applied to other populations, including those with hepatitis C virus–related HCC.

We have identified genes relevant to primary HCC with accompanying intra-hepatic metastasis. We were not, however, able to identify any gene distinguishing primary HCC from its metastases. Our data indicate that the changes favoring metastasis may occur in primary HCC, and that primary HCC with metastatic potential may be evolutionarily distinct from metastasis-free primary HCC. These results are consistent with our findings that the gene expression signature that is relevant to metastasis is independent of tumor size, tumor encapsulation and patient age. Additional subtle genetic changes may occur during the progression of primary HCC to metastasis, such as loss of chromosome 8p in PT patients[33]. Such small changes may be undetectable by our methods. We propose three models to explain the progression of HCC metastasis: (i) metastasis is an acquired activity resulting from genetic changes during the switch from primary to metastatic lesions; (ii) the changes favoring metastasis occur in primary tumors at a very early stage of tumor development; or (iii) primary tumors may acquire a genetic susceptibility to metastasis promoters. We refer to the third model as an 'epigenetic switch' mechanism. Our current data are consistent with the last two models.

The molecular program associated with intra-hepatic metastases may be unique to HCC patients, as patients with other types of solid malignant tumors usually develop distant metastasis, such as colon cancer with liver or lung metastases, or breast cancer with lymph node metastasis, rather than local invasion. Similarities between gene expression profiles of a metastasis and

its primary tumor were also found in studies of gene expression profiles of breast cancer[23]. The authors reported that a metastasis and its primary tumor were as similar in their overall pattern of gene expression as were repeated samplings of the same primary tumor, and suggested that the molecular program of a primary tumor may generally be retained in its metastases[23]. Our findings support this hypothesis that a molecular program associated with metastatic progression is initiated in primary tumors with metastatic potential. There is a need to re-examine the current views on both metastatic progression and approaches to rational therapy. In addition, our findings may also explain our current clinical experience with high incidence of HCC recurrence after surgical intervention of HCC patients diagnosed by routine screening. This emphasizes the need to identify new key targets for treating metastatic HCC patients.

Alterations in cell adhesion molecules and changes in genes that control matrix degradation are the two main acquired capabilities that allow primary tumors to invade tissue and metastasize[34]. Consistent with this view, several genes belonging to those categories were identified in our classifier, including genes encoding Spp1, α_9-integrin, interleukin-2 receptor, serine proteinase inhibitor member-5, matrix metalloproteinase-9, leukocyte immunoglobulin-like receptor subfamily A member-2 and CD37 antigen (Fig. 3*a*). SPP1 expression increased the most in primary HCC with accompanying metastasis, and SPP1 expression correlated with the invasiveness of HCC cells in tissue culture. In addition, a neutralizing antibody against SPP1 was able to block *in vitro* invasion of highly metastatic HCC cells and *in vivo* lung metastasis of HCC tumors. These data indicate that SPP1 may be necessary to support metastasis in primary HCC. Our results are consistent with recent findings of SPP1 over-expression in highly metastatic tumor cell lines[35,36] and elevated plasma SPP1 in metastatic breast cancer patients[30]. Moreover, high-level SPP1 expression can confer a metastatic phenotype on benign tumor cells[37,38]. SPP1 is a glycosylated phosphoprotein that acts as a cytokine and binds receptors, including several integrins, to deliver signals to cells[39,40]. A recent study indicates that SPP1 is a major transcription target induced by hepatocyte growth factor and may contribute to hepatocyte growth factor–mediated cell-cell dissociation and cell growth and invasiveness[41]. SPP1 is an ideal diagnostic marker because it can be found in all bodily fluids and because elevated plasma SPP1 can be found in patients with malignant tumors. In addition, our data indicate that SPP1 is a potential target for therapy of HCC patients with metastatic potential. Because SPP1 is an extracellular cytokine ligand, its interaction with receptors is more readily accessible to pharmaceuticals than intracellular targets. Studies are under way to further characterize all the genes in the classifier and to refine our predictor model with a larger patient cohort, with the ultimate goal of decreasing HCC aggressiveness and increasing patient survival.

Methods

HCC samples. All HCC samples were obtained with informed consent from patients who underwent curative resection at the Liver Cancer Institute and Zhongshan Hospital (Fudan University, Shanghai, China). The 107 paired samples, including primary HCC, metastatic HCC and corresponding adjacent non-tumor liver tissue, were derived from 40 predominantly male and hepatitis B–positive Chinese patients with an average age of 50. All samples were histopathologically diagnosed as HCC according to Edmonson's classification[42]. Primary HCC samples ranged from 1.3 cm to 17.5 cm in diameter with a median diameter of 7.2 cm. The majority of the cases (65%) were larger than 5 cm, which indicates that they were not early cases (see Supplementary Table 6 online for detailed patient profiles). This study was approved by the Institutional Review Board of the Liver Cancer Institute. Total RNA was extracted from each sample using TRIzol (Invitrogen, Carlsbad, California) according to the manufacturer's instructions.

cDNA microarrays. The cDNA microarrays were prepared at the Advanced Technology Center of the National Cancer Institute. Each array contained 9,180 cDNA clones with 7,102 'named' genes, 1,179 expressed sequence tag clones and 122 clones from Incyte (Palo Alto, California). Preparation of fluorescent cDNA targets by a direct labeling approach and cDNA microarray hybridization were previously described[43]. Most of the tumor tissues were sampled twice using 2 independent cDNA hybridizations, with tumor samples labeled in red (Cy5) and non-cancerous tissues labeled in green (Cy3).

Analysis and statistics. Unsupervised hierarchical clustering analysis was done by the CLUSTER and TREEVIEW software[44] using median-centered correlation and complete linkage. We also used the BRB-ArrayTools software for both unsupervised and supervised analyses. This is an integrated package for the visualization and statistical analysis of cDNA microarray gene expression data, developed by the Biometric Research Branch of the National Cancer Institute. We used the Class Comparison Tool based on univariate *F*-tests to find genes differentially expressed between predefined clinical groups. The permutation distribution of the *F*-statistic, based on 2,000 random permutations, was also used to confirm statistical significance. In comparing primary with metastatic tumors of the same patient, a paired-value *t*-statistic was used in the same manner. We also used the CCP Tool with a leave-one-out cross-validation test based on a weighted linear combination of gene expression variables that were univariately significant in the training set, with the weights being the corresponding *t*-statistics[29]. Averaged gene expression data from duplicate samples were included for the analysis. The misclassification rate was determined by leave-one-out cross-validation. For each step of the cross-validation in which one sample was left out, the selection of informative genes and the creation of the multi-gene classifier was repeated from scratch. The probability of obtaining a small cross-validated misclassification rate by chance was obtained by repeating the entire cross-validation procedure using 2,000 random permutations of the class labels for the clinical criteria being evaluated; this gave rise to a classifier *P*. To generate a prediction model to classify HCC with metastasis potential, we used the linear combination

$$L = \Sigma_i\, t_i \times (x_i - m_i)$$

where t_i = *t*-value for gene *i* in the classifier, x_i = log-ratio of gene *i* in the new sample to be classified and m_i = midpoint between the PN and PT groups for gene *i* (see Supplementary Table 2 online). The Kaplan-Meier survival analysis was used to compare patient survival, using Excel-based WinSTAT software (http://www.winstat.com). The statistical *P* value was generated by the Cox-Mantel log-rank test when PN was compared to P or PT.

RNA and protein analysis. Relative quantitative RT-PCR of *SPP1* was done according to the manufacturer's instructions for QuantumRNA 18S Internal Standards (Ambion, Austin, Texas) with *SPP1*-specific primer pairs. Western blotting was done as previously described[43].

Cell lines and mouse model. We used 7 human hepatoma–derived cell lines (HuH1, HuH4, HuH7, MHCC97, SMMC7721, SK-Hep-1 and Hep3B) with different metastatic potential and one non-transformed liver cell line, CCL13 (Chang liver cells), to determine the functional association of SPP1 with metastatic potential. We used the BioCoat Matrigel Invasion Chamber (Becton Dickinson, Bedford, Massachusetts) according to the manufacturer's instructions. Cells were routinely maintained as previously described[43]. We used a well-established nude mouse model of pulmonary metastasis to examine the role of SPP1 in the metastatic potential of HCC cells[32]. The study protocol was approved by the Shanghai Medical Experimental Animal Care Commission.

Additional microarray information. The description of this microarray study followed the MIAME guidelines issued by the Microarray Gene

Expression Data group[45]. The original data will be available in the NCBI's Gene Expression Omnibus public database (http://www.ncbi.nlm.nih.gov/geo/) at a later date. Information is available from the authors on request.

Note: Supplementary information is available at the Nature Medicine *website.*

Acknowledgments

We thank C.C. Harris and L. Varticovski for comments; D. Dudek and K. MacPherson for editorial assistance; D. Petersen, J. Powell and members of the National Cancer Institute microarray team at the Advanced Technology Center for technical support; C. Drachenberg for pathological diagnosis; and J. Fan and X.D. Zhou for help in preparing human tissues. This work was supported in part by the Intramural Research Program of the US National Cancer Institute. Q.H.Y., L.X.Q., Z.C.M., Z.Q.W., S.L.Y., Y.K.L. and Z.Y.T. were supported by research grants from the State Key Basic Research Program of China (No. G1998051210) and from the key project of the Ministry of Education of China.

Competing interests statement

The authors declare that they have no competing financial interests.

RECEIVED 24 DECEMBER 2002; ACCEPTED 25 FEBRUARY 2003

1. Parkin, D.M., Pisani, P. & Ferlay, J. Global cancer statistics. *CA Cancer J. Clin.* **49**, 33–64 (1999).
2. Pisani, P., Parkin, D.M., Bray, F. & Ferlay, J. Estimates of the worldwide mortality from 25 cancers in 1990. *Int. J. Cancer* **83**, 18–29 (1999).
3. El-Serag, H.B. & Mason, A.C. Rising incidence of hepatocellular carcinoma in the United States. *N. Engl. J. Med.* **340**, 745–750 (1999).
4. Taylor-Robinson, S.D., Foster, G.R., Arora, S., Hargreaves, S. & Thomas, H.C. Increase in primary liver cancer in the UK, 1979-94. *Lancet* **350**, 1142–1143 (1997).
5. Curley, S.A. *et al.* Identification and screening of 416 patients with chronic hepatitis at high risk to develop hepatocellular cancer. *Ann. Surg.* **222**, 375–380 (1995).
6. Larcos, G., Sorokopud, H., Berry, G. & Farrell, G.C. Sonographic screening for hepatocellular carcinoma in patients with chronic hepatitis or cirrhosis: an evaluation. *AJR Am. J. Roentgenol.* **171**, 433–435 (1998).
7. Yang, B. *et al.* Prospective study of early detection for primary liver cancer. *J. Cancer Res. Clin. Oncol.* **123**, 357–360 (1997).
8. Izzo, F. *et al.* Outcome of 67 patients with hepatocellular cancer detected during screening of 1125 patients with chronic hepatitis. *Ann. Surg.* **227**, 513–518 (1998).
9. Bolondi, L. *et al.* Surveillance programme of cirrhotic patients for early diagnosis and treatment of hepatocellular carcinoma: a cost effectiveness analysis. *Gut* **48**, 251–259 (2001).
10. Tang, Z.Y. Hepatocellular carcinoma-cause, treatment and metastasis. *World J. Gastroenterol.* **7**, 445–454 (2001).
11. Zhou, X.D. *et al.* Experience of 1000 patients who underwent hepatectomy for small hepatocellular carcinoma. *Cancer* **91**, 1479–1486 (2001).
12. Yuki, K., Hirohashi, S., Sakamoto, M., Kanai, T. & Shimosato, Y. Growth and spread of hepatocellular carcinoma. A review of 240 consecutive autopsy cases. *Cancer* **66**, 2174–2179 (1990).
13. Genda, T. *et al.* Cell motility mediated by rho and rho-associated protein kinase plays a critical role in intrahepatic metastasis of human hepatocellular carcinoma. *Hepatology* **30**, 1027–1036 (1999).
14. Mitsunobu, M., Toyosaka, A., Oriyama, T., Okamoto, E. & Nakao, N. Intrahepatic metastases in hepatocellular carcinoma: the role of the portal vein as an efferent vessel. *Clin. Exp. Metastasis* **14**, 520–529 (1996).
15. Lindsay, C.K., Sinha, C.C. & Thorgeirsson, U.P. Morphological study of vascular dissemination in a metastatic hepatocellular carcinoma model in the monkey. *Hepatology* **26**, 1209–1215 (1997).
16. Kuriyama, S. *et al.* Analysis of intrahepatic invasion of hepatocellular carcinoma using fluorescent dye-labeled cells in mice. *Anticancer Res.* **18**, 4181–4188 (1998).
17. Osada, T. *et al.* E-cadherin is involved in the intrahepatic metastasis of hepatocellular carcinoma. *Hepatology* **24**, 1460–1467 (1996).
18. Guo, X.Z. *et al.* KAI1, a new metastasis suppressor gene, is reduced in metastatic hepatocellular carcinoma. *Hepatology* **28**, 1481–1488 (1998).
19. Hui, A.M., Li, X., Makuuchi, M., Takayama, T. & Kubota, K. Over-expression and lack of retinoblastoma protein are associated with tumor progression and metastasis in hepatocellular carcinoma. *Int. J. Cancer* **84**, 604–608 (1999).
20. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
21. Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
22. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
23. Perou, C.M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
24. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
25. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74 (2002).
26. Khan, J. *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**, 673–679 (2001).
27. Okabe, H. *et al.* 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 (2001).
28. Xu, X.R. *et al.* Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. *Proc. Natl. Acad. Sci. USA* **98**, 15089–15094 (2001).
29. Radmacher, M.D., McShane, L.M. & Simon, R. A paradigm for class prediction using gene expression profiles. *J. Comput. Biol.* **9**, 505–511 (2002).
30. Singhal, H. *et al.* Elevated plasma osteopontin in metastatic breast cancer associated with increased tumor burden and decreased survival. *Clin. Cancer Res.* **3**, 605–611 (1997).
31. Fedarko, N.S., Jain, A., Karadag, A., Van Eman, M.R. & Fisher, L.W. Elevated serum bone sialoprotein and osteopontin in colon, breast, prostate, and lung cancer. *Clin. Cancer Res.* **7**, 4060–4066 (2001).
32. Li, Y. *et al.* Establishment of a hepatocellular carcinoma cell line with unique metastatic characteristics through in vivo selection and screening for metastasis-related genes through cDNA microarray. *J. Cancer Res. Clin. Oncol.* (doi 10.1007/s00432-002-0396-4).
33. Qin, L.X. *et al.* The association of chromosome 8p deletion and tumor metastasis in human hepatocellular carcinoma. *Cancer Res.* **59**, 5662–5665 (1999).
34. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
35. Sharp, J.A., Sung, V., Slavin, J., Thompson, E.W. & Henderson, M.A. Tumor cells are the source of osteopontin and bone sialoprotein expression in human breast cancer. *Lab. Invest.* **79**, 869–877 (1999).
36. Urquidi, V. *et al.* Contrasting expression of thrombospondin-1 and osteopontin correlates with absence or presence of metastatic phenotype in an isogenic model of spontaneous human breast cancer metastasis. *Clin. Cancer Res.* **8**, 61–74 (2002).
37. Chen, H., Ke, Y., Oates, A.J., Barraclough, R. & Rudland, P.S. Isolation of and effector for metastasis-inducing DNAs from a human metastatic carcinoma cell line. *Oncogene* **14**, 1581–1588 (1997).
38. Oates, A.J., Barraclough, R. & Rudland, P.S. The identification of osteopontin as a metastasis-related gene product in a rodent mammary tumour model. *Oncogene* **13**, 97–104 (1996).
39. Denhardt, D.T., Giachelli, C.M. & Rittling, S.R. Role of osteopontin in cellular signaling and toxicant injury. *Annu. Rev. Pharmacol. Toxicol.* **41**, 723–749 (2001).
40. Weber, G.F. The metastasis gene osteopontin: a candidate target for cancer therapy. *Biochim. Biophys. Acta* **1552**, 61–85 (2001).
41. Medico, E. *et al.* Osteopontin is an autocrine mediator of hepatocyte growth factor-induced invasive growth. *Cancer Res.* **61**, 5861–5868 (2001).
42. Edmondson, H.A. & Steiner, P.E. Primary carcinoma of the liver. A study of 100 cases among 48900 necropsies. *Cancer* **7**, 462–503 (1954).
43. Wu, C.G. *et al.* Distinctive gene expression profiles associated with hepatitis B virus x protein. *Oncogene* **20**, 3674–3682 (2001).
44. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
45. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).