Cancer Science
The official journal of the Japanese Cancer Association

Epithelial–mesenchymal transition gene signature to predict clinical outcome of hepatocellular carcinoma

Jongmin Kim,[1] Seok Joo Hong,[1] Jin Young Park,[1] Jun Ho Park,[1] Yun-Suk Yu,[1] Sun Young Park,[1] Eun Kyung Lim,[1] Kwan Yong Choi,[2] Eun Kyu Lee,[3] Seung Sam Paik,[4] Kyeong Geun Lee,[5] Hee Jung Wang,[6] In-Gu Do,[7] Jae-Won Joh[8] and Dae Shick Kim[7,9], on behalf of the Korea Cancer Biomarker Consortium

[1]CbsBioscience, Inc., Daejeon; [2]Department of Life Science, Pohang University of Science and Technology, Pohang; [3]College of BioNano Technology, Kyungwon University, Seongnam; Departments of [4]Pathology and [5]Surgery, Hanyang University School of Medicine, Seoul; [6]Department of Surgery, Ajou University School of Medicine, Suwon; Departments of [7]Pathology and [8]Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

(Received November 24, 2009/Revised February 5, 2010/Accepted February 14, 2010/Accepted manuscript online February 18, 2010/Article first published online March 16, 2010)

Hepatocellular carcinoma is one of the most lethal cancers worldwide. More accurate stratification of patients at risk is necessary to improve its clinical management. As epithelial–mesenchymal transition is critical for the invasiveness and metastasis of human cancers, we investigated expression profiles of 12 genes related to epithelial–mesenchymal transition through a real-time polymerase chain reaction. From a univariate Cox analysis for a training cohort of 128 hepatocellular carcinoma patients, four candidate genes (E-cadherin [*CDH1*], inhibitor of DNA binding 2 [*ID2*], matrix metalloproteinase 9 [*MMP9*], and transcription factor 3 [*TCF3*]) with significant prognostic values were selected to develop a risk score of patient survival. Patients with high risk scores calculated from the four-gene signature showed significantly shorter overall survival times. Moreover, the multivariate Cox analysis revealed that four-gene signature ($P = 0.0026$) and tumor stage ($P = 0.0023$) were independent prognostic factors for overall survival. Subsequently, the four-gene signature was validated in an independent cohort of 231 patients from three institutions, in which high risk score was significantly correlated with shorter overall survival ($P = 0.00011$) and disease-free survival ($P = 0.00038$). When the risk score was entered in a multivariate Cox analysis with tumor stage only, both the risk score ($P = 0.0046$) and tumor stage ($P = 2.6 \times 10^{-9}$) emerged as independent prognostic factors. In conclusion, we suggest that the proposed gene signature may improve the prediction accuracy for survival of hepatocellular carcinoma patients, and complement prognostic assessment based on important clinicopathologic parameters such as tumor stage. (*Cancer Sci* 2010; 101: 1521–1528)

Hepatocellular carcinoma (HCC) is the fifth most common cancer worldwide and the most common primary hepatic malignancy, being responsible for 80% of malignant tumors in adult livers. Moreover, its mortality is third among all cancers, behind only lung and colon cancer.[1] HCC is known for its endemic prevalence in Asia and Africa, and the incidence of HCC has doubled in the USA and Europe in the past four decades.[1–3] HCC is resistant to conventional chemotherapy and is rarely amenable to radiotherapy,[4] leaving this disease with no effective therapeutic options and a very poor prognosis. Although the major etiological agents have been identified, the molecular pathogenesis of HCC remains unclear.[5] It is therefore important to identify molecular targets to develop novel diagnostic, therapeutic, and preventive strategies.

Epithelial–mesenchymal transition (EMT) is a key step during embryogenesis but also plays a critical role in cancer progression, through which epithelial cancers invade and metastasize.[6] Therefore, EMT-related pathways have been studied in relation to cancer management and drug resistance, for instance in breast cancer[7] and ovarian cancer.[8] The existence of EMT *in vivo* has been controversial due to its spatial and temporal heterogeneity that complicates a direct observation in clinic.[6] Nevertheless, several EMT markers have been analyzed in clinical specimens and cell lines *in vitro*.[9–11] Meta-analysis of gene expression profiles in HCC revealed three robust subclasses of HCC.[12] Interestingly, one of the subgroups was characterized by overexpression of transforming growth factor-β (TGF-β) target gene sets including genes involved in EMT, and this subgroup was correlated with early recurrence. In another recent study analyzing EMT markers in HCC, the protein expression levels of E-cadherin, Snail (SNAI1), Slug, and Twist were evaluated by immunohistochemistry in 123 HCC samples and a significant association of Snail and Twist on prognosis was revealed.[13] Thus, we hypothesized that the gene expression profiling of EMT markers in a large number of HCC patients could provide a basis for prognostic predictors of patient outcomes.

In the present study, to construct a reliable prognostic gene signature that could identify HCC patients with a high risk of death, we examined the expression of twelve genes related to EMT by quantitative real-time polymerase chain reaction (PCR). Four genes (E-cadherin [*CDH1*], inhibitor of DNA binding 2 [*ID2*], matrix metalloproteinase 9 [*MMP9*], and transcription factor 3 [*TCF3*]) were selected as highly predictive of survival in the training cohort of 128 patients. The four-gene signature was positively validated in an independent cohort of 231 patients from three institutions. Thus, the novel four-gene signature may be useful to refine a patient's prognosis and improve clinical management.

Materials and Methods

Patients and tissue samples. The study comprised patient cohorts from three medical institutions. The training cohort included 128 randomly selected patients who underwent curative hepatectomy for primary HCC between 2001 and 2005 in the Department of Surgery, Samsung Medical Center (SMC), Korea. The validation cohort comprised three patient cohorts from three medical centers: 104 additional independent cases randomly selected from patients who underwent curative hepatectomy for primary HCC between 2001 and 2005 at the SMC, 94 randomly selected cases from patients who underwent curative hepatectomy for primary HCC between 1995 and 2004 at Ajou University Medical Center (AMC), and 33 randomly selected cases from patients who underwent curative hepatectomy for primary HCC between 2001 and 2004 at Hanyang

[9]To whom correspondence should be addressed. E-mail: oncorkim@skku.edu

doi: 10.1111/j.1349-7006.2010.01536.x

Table 1. Clinical characteristics of the training and validation cohorts (*N* = 359)

Clinicopathologic parameters	Training cohort, SMC (*n* = 128)	Validation cohort, SMC (*n* = 104)	Validation cohort, AMC (*n* = 94)	Validation cohort, HMC (*n* = 33)
Age				
<55 years	80	81	55	18
≥55 years	48	23	39	15
Gender				
Male	104	85	67	30
Female	24	19	27	3
HBV				
Absent	32	13	20	4
Present	96	91	74	29
HCV				
Absent	122	103	86	31
Present	6	1	6	2
Liver cirrhosis				
Absent	68	51	18	11
Present	60	53	73	22
Tumor stage				
I	48	34	1	10
II	56	49	55	11
III	23	20	27	9
IV	1	1	11	3
AFP level				
<100 ng/mL	64	60	41	15
≥100 ng/mL	64	44	52	18
Vascular invasion				
Absent	49	36	39	17
Present	79	68	55	16
Tumor number				
Single	102	79	67	18
Multiple	26	25	27	15
Tumor size				
<5 cm	75	63	40	16
≥5 cm	53	41	54	17
Edmondson grade				
I	10	5	13	5
II	106	89	30	14
III	12	9	47	14
IV	0	1	4	0
Follow-up period, months				
Median	54.3	50.4	60.8	32.0
Range	3.4–93.8	3.6–95.6	2.0–153.0	2.0–100.2

There are two patients with unknown hepatitis C virus (HCV) infection, one patient with unknown alpha fetoprotein (AFP) level, and three patients with unknown status of liver cirrhosis from Ajou University Medical Center (AMC). HBV, hepatitis B virus; HMC, Hanyang University Medical Center; SMC, Samsung Medical Center.

University Medical Center (HMC). Patient characteristics for patient cohorts are summarized in Table 1. The study protocol was approved by the Institutional Review Boards of SMC, AMC, and HMC. Complete clinical data were available in all cases, except for two patients with unknown hepatitis C virus (HCV) infection, one patient with unknown alpha fetoprotein (AFP) level, and three patients with unknown status of liver cirrhosis from AMC. All patients had adequate liver function reserve, and had survived for at least 2 months after hepatectomy. Recurrence or death was evaluated from medical records of patients. We defined the recurrence as evidence of an overt new growing mass in the remaining liver or as distant metastasis in radiologic studies including computed tomography or magnetic resonance imaging. None of the patients had received treatment prior to surgery such as transarterial chemoembolization or radiofrequency ablation. Immediately after hepatectomy, fresh tumors and background livers were partly snap-frozen in liquid nitrogen and stored at −80°C and were partly embedded in paraffin after fixation in 10% formalin for histological diagnosis. All available hematoxylin–eosin stained slides were reviewed. The tumor grading was based on the criteria proposed by Edmondson and Steiner (I, well differentiated; II, moderately differentiated; III, poorly differentiated; IV, undifferentiated).[14] The conventional TNM system outlined in the cancer staging manual (6th ed.) by the American Joint Committee on Cancer was used in tumor staging. The tumor size was obtained from the pathology reports.

RNA extraction and cDNA synthesis. RNA extraction and cDNA synthesis were carried out as described previously.[15,16] Briefly, total RNA was extracted from cancerous and surrounding non-cancerous frozen tissues using an RNeasy minikit (Qiagen, Hilden, Germany). The integrity of all tested total RNA samples was verified using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). DNase I treatment was routinely included in the extraction step. Samples containing 4 μg of total RNA were incubated with 2 μL of 10 μM oligo $d(T)_{18}$ primer (Genotech, Daejeon, Korea) at 70°C for 7 min and cooled on ice for 5 min. After adding the enzyme mix to the annealed total RNA sample, the reaction was incubated for 90 min at 42°C prior to heat inactivation of reverse-transcriptase at 80°C for 10 min. The cDNA samples were brought up to a final volume of 400 μL with the addition of diethylpyrocarbonate (DEPC)-treated water.

Quantitative real-time PCR. Real-time PCR amplifications were carried out as described previously.[15,16] Briefly, using Applied Biosystems Prism 7900HT instruments (Applied Biosystems, Foster City, CA, USA), the real-time PCR analysis was performed in a total volume of 10 μL with the amplification steps described below: an initial activation step at 95°C for 10 min which was followed by 45 cycles of denaturation at 95°C for 15 s and elongation at 60°C for 1 min. The primer and probe sequences were designed using Primer Express 3.0 software (Applied Biosystems), and all probe sequences were labeled with FAM at the 5′ end and with TAMRA at the 3′ end (Table 2). The mRNA levels of target genes (*CDH1*, *CDH2*, *ID2*, *MMP2*, *MMP9*, *TCF3*, *TWIST1*, vascular endothelial growth factor A [*VEGFA*], *SNAI1*, *SNAI2*, zinc finger E-box binding homeobox 1 [*ZEB1*], and *ZEB2*) were measured (the threshold cycle, C_T) in triplicate and were then normalized relative to a set of reference genes (beta-2-microglobulin [*B2M*], *GAPDH*, hydroxymethylbilane synthase [*HMBS*], hypoxanthine phosphoribosyltransferase 1 [*HPRT1*], and succinate dehydrogenase complex, subunit A, flavoprotein [*SDHA*]) by subtracting the average of the expression of the five reference genes as an internal control.[17] Using the ΔC_T values (target gene C_T – average C_T of reference genes), the mRNA copy number ratio was calculated as $2^{-\Delta Ct}$. Standard curves were constructed from the results of simultaneous amplifications of serial dilutions of the cDNA samples.

Statistical analysis. Clinicopathologic variables of the training and validation cohorts were evaluated using a χ^2-test and Fisher's exact test (HCV infection). The gene expression data were normalized by means of the $\log_2$ transform. After transformation, results for each gene were centered and scaled to an average of 0 and an SD of 1. For the training cohort, univariate Cox regression analyses were run for each gene, as summarized in Table 3. Genes achieving *P*-values of <0.05 in the univariate Cox analyses were then entered as potential predictors of patient risk. The risk score was derived by the summation of each gene expression level multiplied by its corresponding regression coefficient.[18] The classification accuracy was measured by the area under the curve (AUC)

doi: 10.1111/j.1349-7006.2010.01536.x

Table 2. Oligonucleotide sequences of PCR primers and probes

Gene	Sequences
B2M	Forward: 5′-CAT TCG GGC CGA GAT GTC T-3′
	Reverse: 5′-CTC CAG GCC AGA AAG AGA GAG TAG-3′
	Probe: 5′-CCG TGG CCT TAG CTG TGC TCG C-3′
GAPDH	Forward: 5′-CAC ATG GCC TCC AAG GAG TAA-3′
	Reverse: 5′-TGA GGG TCT CTC TCT TCC TCT TGT-3′
	Probe: 5′-CTG GAC CAC CAG CCC CAG CAA G-3′
HMBS	Forward: 5′-CCA GGG ATT TGC CTC ACC TT-3′
	Reverse: 5′-AAA GAG ATG AAG CCC CCA CAT-3′
	Probe: 5′-CCT TGA TGA CTG CCT TGC CTC CTC AG-3′
HPRT1	Forward: 5′-GCT CGA GAT GTG ATG AAG GAG AT-3′
	Reverse: 5′-CCA GCA GGT CAG CAA AGA ATT-3′
	Probe: 5′-CCA TCA CAT TGT AGC CCT CTG TGT GCT C-3′
SDHA	Forward: 5′-CAC CTA GTG GCT GGG AGC TT-3′
	Reverse: 5′-GCC CAG TTT TAT CAT CTC ACA AGA-3′
	Probe: 5′-TGG CAC TTA CCT TTG TCC CTT GCT TCA-3′
CDH1	Forward: 5′-AAA TCT GAA AGC GGC TGA TAC TG-3′
	Reverse: 5′-CGG AAC CGC TTC CTT CAT AG-3′
	Probe: 5′-CCC CAC AGC CCC GCC TTA TGA-3′
CDH2	Forward: 5′-GCA CCC GCC TCA GTC AAC T-3′
	Reverse: 5′-CAA CAT GGT ACC GGC ATG AA-3′
	Probe: 5′-AAT GAA AAC CCT TAT TTT GCC CCC AAT CC-3′
ID2	Forward: 5′-AAC GAC TGC TAC TCC AAG CTC AA-3′
	Reverse: 5′-GGA TTT CCA TCT TGC TCA CCT T-3′
	Probe: 5′-TGC CCA GCA TCC CCC AGA ACA A-3′
MMP2	Forward: 5′-GGT TGT CTG AAG TCA CTG CAC AGT-3′
	Reverse: 5′-CTC GGT AGG GAC ATG CTA AGT AGA G-3′
	Probe: 5′-CAT CTC AGC CCA CAT AGT GAT GGT TCC C-3′
MMP9	Forward: 5′-GGG CTC CCG TCC TGC TT-3′
	Reverse: 5′-ACT CCT CCC TTT CCT CCA GAA C-3′
	Probe: 5′-TGC CAT GTA AAT CCC CAC TGG GAC C-3′
TCF3	Forward: 5′-GCT GCC TTT GGT CTC TGG TTT-3′
	Reverse: 5′-AGA AAT GCA ATG CTC AGT CTA GGA-3′
	Probe: 5′-AGT CCC GTG TCT CTC GCT ATT TCT GCT G-3′
TWIST1	Forward: 5′-TCC GCG TCC CAC TAG CA-3′
	Reverse: 5′-AGT TAT CCA GCT CCA GAG TCT CTA GAC-3′
	Probe: 5′-CAC CCC CTC AGC AGG GCC G-3′
VEGFA	Forward: 5′-GCT TAC TCT CAC CTG CTT CTG AGT T-3′
	Reverse: 5′-TGG GCT GCT TCT TCC AAC A-3′
	Probe: 5′-AGA CCA CTG GCA GAT GTC CCG GC-3′
SNAI1	Forward: 5′-CCC AGT GCC TCG ACC ACT AT-3′
	Reverse: 5′-TCC TGC AGC TCG CTG TAG TTA G-3′
	Probe: 5′-CCG CGC TCT TTC CTC GTC AGG A-3′
SNAI2	Forward: 5′-TGC TCC AAA ACC TTC TCC AGA A-3′
	Reverse: 5′-GCG TCA CTC AGT GTG CTA CAC A-3′
	Probe: 5′-TCT CCT GCA CAA ACA TGA GGA ATC TGG C-3′
ZEB1	Forward: 5′-GCC CAG TTA CCC ACA ATC GT-3′
	Reverse: 5′-TGA CCG TAG TTG AGT AGG TGT ATG C-3′
	Probe: 5′-TTC CAT GCT TAA GAG CGC TAG CTG CCA-3′
ZEB2	Forward: 5′-TTT TCC TGC CCT CTC TGT AGC T-3′
	Reverse: 5′-GGT TAG CAT TTG GTG CTG ATC TG-3′
	Probe: 5′-ACG TTT GCC TAC CGC ACC CAG CT-3′

B2M, beta-2-microglobulin; *CDH1*, E-cadherin; *HMBS*, hydroxymethylbilane synthase; *HPRT1*, hypoxanthine phosphoribosyltransferase 1; *ID2*, inhibitor of DNA binding 2; *MMP2*, matrix metalloproteinase 2; *SDHA*, succinate dehydrogenase complex, subunit A, flavoprotein; *SNAI1*, Snail 1; *TCF3*, transcription factor 3; *VEGFA*, vascular endothelial growth factor A; *ZEB1*, zinc finger E-box binding homeobox 1.

of the receiver–operator curves (ROC). A multivariate Cox proportional hazard model was then used to identify independent prognostic factors for overall survival (OS). Kaplan–Meier survival curves were calculated using tumor recurrence (defined as the first appearance of a tumor at any site following definitive treatment) or death as the end points. The differences in OS curve or disease-free survival (DFS) curve were examined by log-rank test. Significant differences between gene expression levels for HCC and non-cancerous tissues were evaluated by a Student's *t*-test. A two-tailed *P*-value test was used, with a *P*-value of <0.05 considered statistically significant. All statistical analyses were done with the open source statistical programming environment R.

Table 3. Univariate Cox regression analysis of OS according to gene expression in the training cohort

Variable	RC	HR (95% CI)	*P*-values
CDH1	−0.333	0.72 (0.54–0.95)	0.020
CDH2	−0.216	0.81 (0.63–1.04)	0.094
ID2	−0.400	0.67 (0.50–0.90)	0.0068
MMP2	−0.061	0.94 (0.69–1.29)	0.70
MMP9	0.339	1.40 (1.05–1.88)	0.023
TCF3	0.387	1.47 (1.08–2.01)	0.015
TWIST1	0.056	1.06 (0.77–1.45)	0.73
VEGFA	0.135	1.14 (0.84–1.56)	0.40
SNAI1	−0.100	0.91 (0.66–1.25)	0.54
SNAI2	−0.207	0.81 (0.59–1.13)	0.22
ZEB1	−0.171	0.84 (0.62–1.15)	0.28
ZEB2	−0.323	0.72 (0.51–1.03)	0.073

CDH1, E-cadherin; CI, confidence interval; HR, hazard ratio; *ID2*, inhibitor of DNA binding 2; *MMP2*, matrix metalloproteinase 2; OS, overall survival; RC, regression coefficient; *SNAI1*, Snail 1; *TCF3*, transcription factor 3; *VEGFA*, vascular endothelial growth factor A; *ZEB1*, zinc finger E-box binding homeobox 1.

Results

We performed quantitative real-time PCR for 12 genes (*CDH1*, *CDH2*, *ID2*, *MMP2*, *MMP9*, *TCF3*, *TWIST1*, *VEGFA*, *SNAI1*, *SNAI2*, *ZEB1*, and *ZEB2*) related to the EMT process from frozen paired samples derived from a training cohort of 128 patients with HCC. Expressions of these 12 genes were measured in triplicate and were then normalized relative to the expression of a set of reference genes (*B2M*, *GAPDH*, *HMBS*, *HPRT1*, and *SDHA*) as an internal control.[17] The $\log_2$-transformed gene expression levels were centered and scaled to an average of 0 and an SD of 1. Gene expression levels were correlated with OS by univariate Cox regression analysis and the genes were ranked according to their effect on OS (Table 3). Four genes were significantly correlated with the OS of patients: *CDH1* and *ID2* were protective genes (associated with a hazard ratio of less than 1), and *MMP9* and *TCF3* were risk genes (associated with a hazard ratio of more than 1). These genes were then profiled in 40 paired noncancerous hepatic samples from the training cohort. Protective genes were down-regulated and risk genes were up-regulated in HCC compared to noncancerous livers (Fig. 1).

A patient's risk score was derived by the summation of each gene expression level multiplied by its corresponding coefficient, as follows: risk score = $(-0.333 \times CDH1) + (-0.400 \times ID2) + (0.339 \times MMP9) + (0.387 \times TCF3)$, wherein *CDH1*, *ID2*, *MMP9*, and *TCF3* refer to the $\log_2$-transformed and normalized results for each gene. The AUC of ROC showing prediction of patient survival by risk score was 0.772 (Fig. 2a). The cut-off value of the risk score ($\theta = 0.303$) was determined from the ROC of patient survival in the training cohort with 62.5% sensitivity, 80.6% specificity, and 75.0% accuracy (Fig. 2a). The risk score was used to classify patients into high (>0.303) or low (<0.303) risk groups, where high risk indicated poor survival. At the 5- and 7-year follow-up, approximately 80% and 80% of the low-risk group survived, whereas 52% and 36% of the high-risk group survived,

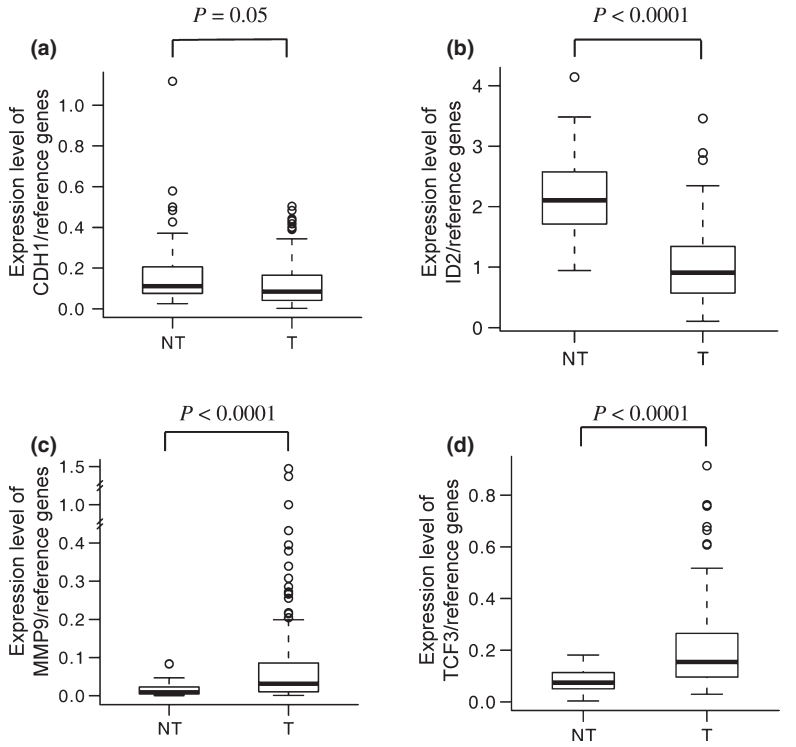

Fig. 1. Box and whiskers plot for (a) E-cadherin (*CDH1*), (b) inhibitor of DNA binding 2 (*ID2*), (c) matrix metalloproteinase 9 (*MMP9*), and (d) transcription factor 3 (*TCF3*) mRNA levels in 40 noncancerous liver tissues (NT) and 128 hepatocellular carcinoma (HCC) tissues (T) of the training cohort determined by real-time RT-PCR. The box is marked by the first and third quartile with the median marked by a thick line. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

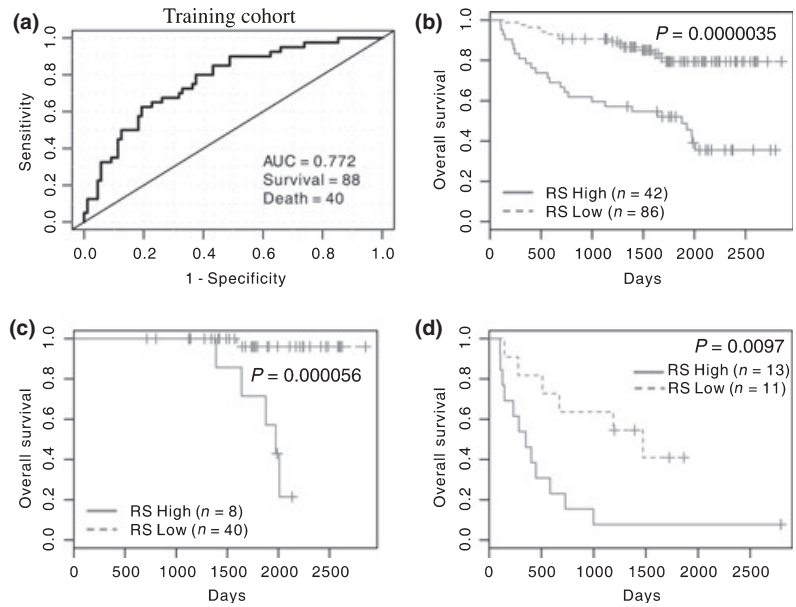

Fig. 2. Receiver–operator curves (ROC) and Kaplan–Meier curves for the overall survival (OS) of patients in the training cohort. (a) Area under the curve (AUC) of ROC showing prediction of patient survival in the training cohort by four-gene risk score was 0.772. (b) patients with high risk scores (>0.303) had a significantly shorter OS time ($P = 3.5 \times 10^{-6}$). (c) Stage I tumor patients with high risk scores had a significantly shorter OS time ($P = 5.6 \times 10^{-5}$). (d) Advanced stage tumor patients with high risk scores had a significantly shorter OS time ($P = 0.0097$). RS, risk score.

respectively (Fig. 2b). The log-rank test showed that patients with a high risk score had a significantly shorter OS time ($P = 3.5 \times 10^{-6}$). Univariate Cox analysis of clinicopathologic parameters revealed that tumor grade ($P = 0.0042$), AFP level ($P = 0.0052$), tumor size ($P = 0.00058$), tumor stage ($P = 3.1 \times 10^{-10}$), vascular invasion ($P = 0.00077$), and tumor number ($P = 8.9 \times 10^{-7}$) were significant prognostic factors for OS. The risk score was included in a multivariate Cox regression analysis with clinicopathologic parameters. The risk score ($P = 0.0026$) and tumor stage ($P = 0.0023$) emerged as independent prognostic factors (Table 4). When patients were divided into subgroups according to tumor stage, patients with a high risk score had a significantly shorter OS time for both stage I ($P = 5.6 \times 10^{-5}$) and stage III/IV tumors ($P = 0.0097$, log-rank test; Fig. 2c,d). Among the stage I patients, five out of eight high-risk patients died within the follow-up period, whereas one out of 40 low-risk patients died during follow-up, resulting in an accuracy of 92% for OS.

Next, we tested the prognostic value of the four-gene risk score in independent cohorts of 104 cases from SMC, 94 cases from AMC, and 33 cases from HMC. The validation cohort from SMC contained a higher proportion of patients younger than 55 years ($P = 0.017$) and patients with hepatitis B virus (HBV) ($P = 0.026$) compared to the training cohort (Table 1). The validation cohort from AMC contained a higher proportion of patients with high tumor grade ($P < 0.001$), liver cirrhosis

doi: 10.1111/j.1349-7006.2010.01536.x

Table 4. Multivariate Cox regression analysis for OS in the training and validation cohort

Variable	Training cohort (n = 128)		Validation cohort (n = 231)	
	HR (95% CI)	P-values	HR (95% CI)	P-values
Risk score (low *vs* high)	3.09 (1.48–6.42)	0.0026	1.75 (1.19–2.57)	0.0046
Tumor stage (I–II *vs* III–IV)	5.53 (1.84–16.6)	0.0023	3.25 (2.20–4.78)	2.6×10^{-9}
Edmondson grade (I–II *vs* III–IV)	1.76 (0.75–4.12)	0.19	NA	NA
AFP level (<100 ng/mL *vs* ≥100 ng/mL)	0.94 (0.42–2.10)	0.88	NA	NA
Vascular invasion (absent *vs* present)	1.91 (0.71–5.09)	0.20	NA	NA
Tumor number (single *vs* multiple)	0.99 (0.38–2.63)	0.99	NA	NA
Tumor size (<5 cm *vs* ≥5 cm)	1.08 (0.45–2.56)	0.86	NA	NA

AFP, alpha fetoprotein; CI, confidence interval; HR, hazard ratio; OS, overall survival. NA, not applicable.

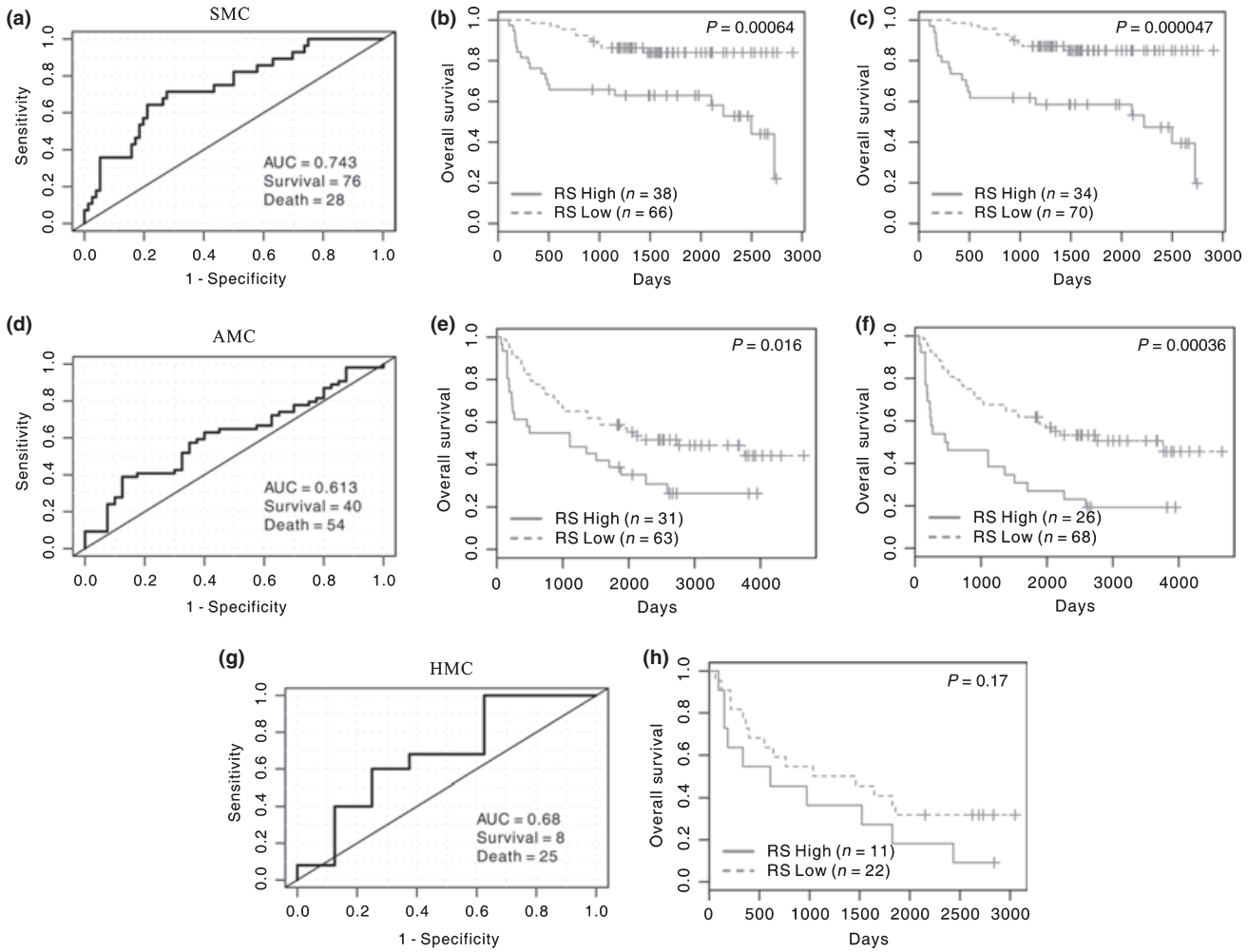

Fig. 3. Receiver–operator curves (ROC) and Kaplan–Meier curves for the overall survival (OS) of patients in the validation cohort from the Samsung Medical Center (SMC) (a–c), Ajou University Medical Center (AMC) (d–f), and Hanyang University Medical Center (HMC) (g,h). (a) Area under the curve (AUC) of ROC showing prediction of patient survival in the validation cohort of SMC by four-gene risk score was 0.743. (b,c) patients with high risk scores had a significantly shorter OS (P = 0.00064 for θ = 0.303, $P = 4.7 \times 10^{-5}$ for θ = 0.388). (d) AUC of ROC for the validation cohort of AMC was 0.613. (e,f) Patients with high risk scores had a significantly shorter OS (P = 0.016 for θ = 0.303, P = 0.00036 for θ = 0.385). (g) AUC of ROC for the validation cohort of HMC was 0.68. (h) Patients with high risk scores (>0.303) tended to have a shorter OS (P = 0.17). RS, risk score.

($P < 0.001$), high tumor stage ($P < 0.001$), and large tumor size ($P = 0.026$) compared to the training cohort. The validation cohort from HMC contained a higher proportion of patients with high tumor grade ($P < 0.001$), high tumor stage ($P = 0.023$), and multiple tumors ($P = 0.0063$) compared to the training cohort. Consequently, the validation cohorts from AMC and HMC had worse prognosis compared to the training cohort from SMC. In the validation cohort, both the regression coefficients of risk score and the cut-off value derived from the training cohort were applied directly. For the SMC cohort, AUC of ROC

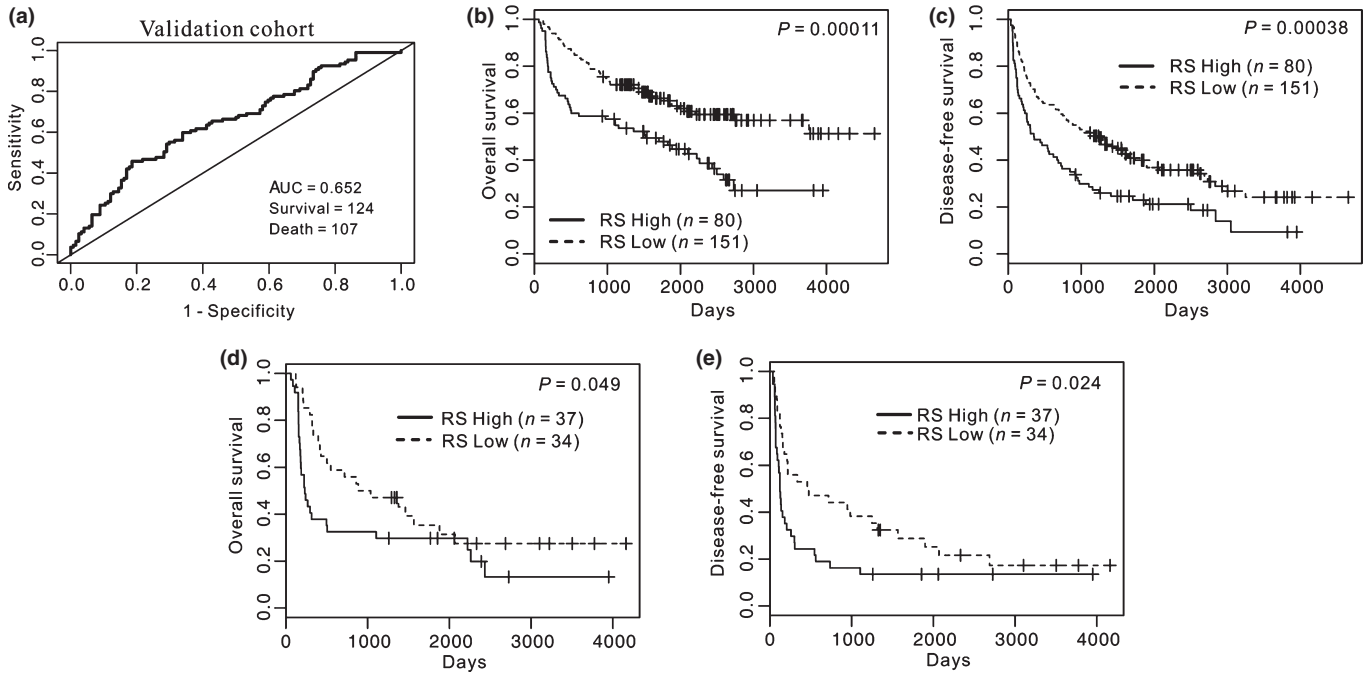

Fig. 4. Receiver–operator curves (ROC) and Kaplan–Meier curves for the overall survival (OS) and disease-free survival (DFS) of patients in the validation cohort. (a) Area under the curve (AUC) of ROC showing prediction of patient survival in the validation cohort by four-gene risk score was 0.652. (b,c) Patients with high risk scores (>0.303) had a significantly shorter OS ($P = 0.00011$) and DFS ($P = 0.00038$). (d,e) Advanced stage tumor patients with high risk scores had a significantly shorter OS ($P = 0.049$) and DFS ($P = 0.024$). RS, risk score.

showing prediction of patient survival by risk score was 0.743 (Fig. 3a) and patients with a high risk score (>0.303) had a significantly shorter OS time ($P = 0.00064$, log-rank test; Fig. 3b). The difference in OS time remained significant for a slightly higher cut-off value ($P = 4.7 \times 10^{-5}$, log-rank test; Fig. 3c). For the AMC cohort, AUC of ROC was 0.613 (Fig. 3d) and patients with a high risk score (>0.303) had a significantly shorter OS time ($P = 0.016$, log-rank test; Fig. 3e). The difference in OS time remained significant for a slightly higher cut-off value ($P = 0.00036$, log-rank test; Fig. 3f). However, for the HMC cohort, AUC of ROC was 0.68 (Fig. 3g) and patients with a high risk score (>0.303) had a shorter OS time, but the difference was not statistically significant ($P = 0.17$, log-rank test; Fig. 3h). For the whole validation cohorts ($n = 231$), AUC of ROC showing prediction of patient survival by risk score was 0.652 (Fig. 4a) and patients with a high risk score (>0.303) had a significantly shorter OS time ($P = 0.00011$, log-rank test; Fig. 4b). In addition, patients with a high risk score had a significantly shorter DFS time ($P = 0.00038$, log-rank test; Fig. 4c). Univariate Cox analysis of clinicopathologic parameters revealed that tumor grade ($P = 7.1 \times 10^{-6}$), AFP level ($P = 2.1 \times 10^{-5}$), liver cirrhosis ($P = 0.0064$), tumor size ($P = 0.00017$), tumor stage ($P = 7.8 \times 10^{-11}$), vascular invasion ($P = 1.1 \times 10^{-6}$), and tumor number ($P = 3.6 \times 10^{-10}$) were significant prognostic factors for OS. However, the risk score was not an independent prognostic factor in a multivariate Cox analysis with all the important clinicopathologic parameters. When the risk score was entered in a multivariate Cox analysis with tumor stage only, both the risk score ($P = 0.0046$) and tumor stage ($P = 2.6 \times 10^{-9}$) emerged as independent prognostic factors (Table 4). On the other hand, in a multivariate Cox analysis for the risk score treated as a continuous variable, the risk score ($P = 0.012$), liver cirrhosis ($P = 0.0056$), tumor number ($P = 0.00094$), and vascular invasion ($P = 0.0073$) emerged as independent prognostic factors (data not shown). When patients were further stratified into subgroups according to tumor stage, patients with a high risk score had a significantly shorter OS time ($P = 0.049$) and DFS time ($P = 0.024$, log-rank test) for stage III–IV tumors (Fig. 4d,e).

Discussion

HCC is a highly heterogeneous disease, and even in patients with similar clinical and pathological features, the outcome varies. Staging systems for HCC that are based on clinical and pathological findings can be complemented by molecular methods that add more predictive power in patient outcomes. Gene-expression profiling with the use of microarrays or real-time PCR has been utilized to identify molecular classifications of patients with HCC.[19] However, the use of microarrays in clinical practice is limited by the large number of genes and relatively complex methodology involved.[20,21] On the other hand, quantitative real-time PCR involving a small number of genes allows for accurate and reproducible quantification of RNA obtained from both frozen tissues and paraffin-embedded tissues.[22,23] Thus, a gene signature based on real-time PCR may offer a more convenient clinical application.

Currently, there is no clear molecular classification of HCC.[19] In a study utilizing 91 HCC samples, a 406-gene signature could classify patients with significant differences in survival.[24] This gene signature revealed that transcripts related to cell proliferation, apoptosis, histone modification, and ubiquitination were important discriminators of patient survival. Subsequently, a subpopulation of patients with progenitor cell characteristics was found to be correlated with poor prognosis.[25] Another study utilized a 153-gene signature generated from 40 HCC patients to discriminate patients with different risk levels of death.[26] In addition, multiple gene signatures have been proposed to predict recurrence in HCC (12,[27] 20,[28] and 57 genes[29]). Gene expression signatures for predicting HCC prognosis may not be unique. Similarly,

doi: 10.1111/j.1349-7006.2010.01536.x

multiple gene expression signatures were developed for predicting prognosis of breast cancers including 21-gene,[30] 70-gene,[31] and 76-gene signatures.[32] While these gene signatures contained largely non-overlapping genes, the prognostic values were significant.

In this study, we evaluated 12 genes related to EMT processes and constructed a prognostic four-gene signature (*CDH1*, *ID2*, *MMP9*, and *TCF3*) for HCC. Not surprisingly, the prediction accuracy of the four-gene signature was best when applied to the validation cohort from SMC which was most similar in patient characteristics compared to the training cohort. The AUC of ROC were smaller in validation cohorts from AMC and HMC, and the four-gene risk score did not achieve statistically significant classification at the designated score threshold for the validation cohort from HMC. However, the prognostic value of the gene-expression signature was positively validated in the total validation cohort from three institutions. Multivariate analysis further strengthened the finding that the four-gene signature was an independent prognostic factor along with tumor stage, thus complementing traditional clinicopathologic parameters.

The four genes in our model are closely related to tumor invasion and metastasis. E-cadherin encoded by *CDH1* is the most prominent epithelial marker as the main molecule of adherent junctions.[6] A decreased expression of E-cadherin in HCC has been reported[33,34] and correlated with poor prognosis.[13] ID2 encoded by the *ID2* gene belongs to a helix-loop-helix family of proteins and represses EMT induced by TGF-β in epithelial cells.[35] Decreased *ID2* expression was correlated with shorter DFS in HCV-related HCC patients.[36] *ID2* was also found in the 57-gene signature for predicting HCC recurrence.[29] At the protein level, decreased ID2 expression was correlated with de-differentiation of HCC.[37] MMPs have been found to be up-regulated in EMT cells[38] but are also capable of inducing EMT.[39] *MMP9* overexpression has been linked to the growth of small HCC[40,41] and elevated plasma MMP9 levels have been observed in patients with HCC.[42] Overexpression of MMP9 protein has reported to be correlated with poor prognosis of HCC patients.[43] Interestingly, the expression level of *TCF3* was significantly associated with prognosis in our analysis, yet little is known about its relation to HCC. E12/E47 encoded by *TCF3* and Twist encoded by *TWIST1* are potent repressors of E-cadherin expression.[44,45] Expression of Twist has been reported to be significantly correlated with prognosis in HCC.[46] In another recent study analyzing EMT markers in HCC, a significant association of Snail and Twist on prognosis was revealed.[13] It is not clear why *SNAI1* and *TWIST1* were not a significant prognostic factor in our patient cohort. We hypothesize that *TCF3* may play a regulatory role similar to *TWIST1*, as shown by its close correlation with prognosis in our patient cohort.

In conclusion, we found that the novel four-gene expression signature was associated with the prognosis of HCC patients. This signature could be useful in stratifying patients according to risk beyond traditional clinicopathologic parameters. Moreover, a quantitative real-time PCR assay is convenient in terms of the work load and is applicable for routine clinical use. Therefore, this new gene expression signature merits further study as a basis for selecting high-risk HCC patients.

Acknowledgments

This work was supported by intramural research funds from CbsBioscience, Inc (CBS-08-71).

Disclosure Statement

The authors have no conflict of interest.

References

1 Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005; **55**: 74–108.
2 Bosch FX, Ribes J, Diaz M, Cleries R. Primary liver cancer: Worldwide incidence and trends. *Gastroenterology* 2004; **127**: S5–16.
3 El-Serag HB. Hepatocellular carcinoma: recent trends in the United States. *Gastroenterology* 2004; **127**: S27–34.
4 Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. *Lancet* 2003; **362**: 1907–17.
5 Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. *Nat Genet* 2002; **31**: 339–46.
6 Voulgari A, Pintzas A. Epithelial-mesenchymal transition in cancer metastasis: mechanisms, markers and strategies to overcome drug resistance in the clinic. *Biochim Biophys Acta* 2009; **1796**: 75–90.
7 Konecny G, Venkatesan N, Yang G *et al.* Activity of lapatinib a novel HER2 and EGFR dual kinase inhibitor in human endometrial cancer cells. *Br J Cancer* 2008; **98**: 1076–84.
8 Kajiyama H, Shibata K, Terauchi M *et al.* Chemoresistance to paclitaxel induces epithelial-mesenchymal transition and enhances metastatic potential for epithelial ovarian carcinoma cells. *Int J Oncol* 2007; **31**: 277–83.
9 Impola U, Uitto V, Hietanen J *et al.* Differential expression of matrilysin-1 (MMP-7), 92 kD gelatinase (MMP-9), and metalloelastase (MMP-12) in oral verrucous and squamous cell cancer. *J Pathol* 2004; **202**: 14–22.
10 Hotz B, Arndt M, Dullat S, Bhargava S, Buhr H-J, Hotz HG. Epithelial to mesenchymal transition: expression of the regulators snail, slug, and twist in pancreatic cancer. *Clin Cancer Res* 2007; **13**: 4769–76.
11 El-Bahrawy MA, Poulsom R, Jeffery R, Talbot I, Alison MR. The expression of E-cadherin and catenins in sporadic colorectal carcinoma. *Hum Pathol* 2001; **32**: 1216–24.
12 Hoshida Y, Nijman SMB, Kobayashi M *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res* 2009; **69**: 7385–92.
13 Yang M-H, Chen C-L, Chau G-Y *et al.* Comprehensive analysis of the independent effect of twist and snail in promoting metastasis of hepatocellular carcinoma. *Hepatology* 2009; **50**: 1464–74.
14 Edmondson H, Steiner P. Primary carcinoma of the liver: a study of 100 cases among 48,900 necropsies. *Cancer* 1954; **7**: 462–503.
15 Kim J, Hong SJ, Park JH *et al.* Real-time reverse transcription PCR analysis for validation of transketolase gene in hepatocellular carcinoma tissues. *Biochip J* 2009; **3**: 130–8.
16 Kim J, Hong SJ, Lim EK *et al.* Expression of nicotinamide N-methyltransferase in hepatocellular carcinoma is associated with poor prognosis. *J Exp Clin Cancer Res* 2009; **28**: 20.
17 Vandesompele J, Preter KD, Pattyn F *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; Jun 18; **3**(7): RESEARCH 0034.
18 Chen H-Y, Yu S-L, Chen C-H *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; **356**: 11–20.
19 Villanueva A, Newell P, Chiang DY, Friedman SL, Llovet JM. Genomics and signaling pathways in hepatocellular carcinoma. *Semin Liver Dis* 2007; **27**: 55–76.
20 Ramaswamy S. Translating cancer genomics into clinical oncology. *N Engl J Med* 2004; **350**: 1814–6.
21 Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007; **99**: 147–57.
22 Cronin M, Pho M, Dutta D *et al.* Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. *Am J Pathol* 2004; **164**: 35–42.
23 Bast RCJ, Hortobagyi GN. Individualized care for patients with cancer – a work in progress. *N Engl J Med* 2004; **351**: 2865–7.
24 Lee J-S, Chu I-S, Heo J *et al.* Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* 2004; **40**: 667–76.
25 Lee J-S, Heo J, Libbrecht L *et al.* A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nat Med* 2006; **12**: 410–6.
26 Ye Q-H, Qin L-X, Forgues M *et al.* Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat Med* 2003; **9**: 416–23.

27 Iizuka N, Oka M, Yamada-Okabe H *et al.* Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003; **361**: 923–9.
28 Kurokawa Y, Matoba R, Takemasa I *et al.* Molecular-based prediction of early recurrence in hepatocellular carcinoma. *J Hepatol* 2004; **41**: 284–91.
29 Wang SM, Ooi LLPJ, Hui KM. Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. *Clin Cancer Res* 2007; **13**: 6275–83.
30 Paik S, Shak S, Tang G *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004; **351**: 2817–26.
31 Buyse M, Loi S, van't Veer L *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006; **98**: 1183–92.
32 Wang Y, Klijn JG, Zhang Y *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005; **365**: 671–9.
33 Prange W, Breuhahn K, Fischer F *et al.* Beta-catenin accumulation in the progression of human hepatocarcinogenesis correlates with loss of E-cadherin and accumulation of p53, but not with expression of conventional WNT-1 target genes. *J Pathol* 2003; **201**: 250–9.
34 Kanai Y, Ushijima S, Hui A-M *et al.* The E-cadherin gene is silenced by CpG methylation in human hepatocellular carcinomas. *Int J Cancer* 1997; **71**: 355–9.
35 Kowanetz M, Valcourt U, Bergström R, Heldin C-H, Moustakas A. Id2 and Id3 define the potency of cell proliferation and differentiation responses to transforming growth factor beta and bone morphogenetic protein. *Mol Cell Biol* 2004; **24**: 4241–54.
36 Tsunedomi R, Iizuka N, Tamesa T *et al.* Decreased ID2 promotes metastatic potentials of hepatocellular carcinoma by altering secretion of vascular endothelial growth factor. *Clin Cancer Res* 2008; **14**: 1025–31.
37 Damdinsuren B, Nagano H, Kondo M *et al.* Expression of Id proteins in human hepatocellular carcinoma: relevance to tumor dedifferentiation. *Int J Oncol* 2005; **26**: 319–27.
38 Mohamed MM, Sloane BF. Cysteine cathepsins: multifunctional enzymes in cancer. *Nat Rev Cancer* 2006; **6**: 764–75.
39 Przybylo JA, Radisky DC. Matrix metalloproteinase-induced epithelial-mesenchymal transition: tumor progression at Snail's pace. *Int J Biochem Cell Biol* 2007; **39**: 1082–8.
40 Sakamoto Y, Mafune K, Mori M *et al.* Overexpression of MMP-9 correlates with growth of small hepatocellular carcinoma. *Int J Oncol* 2000; **17**: 237–43.
41 Qin L-X, Tang Z-Y. The prognostic molecular markers in hepatocellular carcinoma. *World J Gastroenterol* 2002; **8**: 385–92.
42 Hayasaka A, Suzuki N, Fujimoto N *et al.* Elevated plasma levels of matrix metalloproteinase-9 (92-kd type IV collagenase/gelatinase B) in hepatocellular carcinoma. *Hepatology* 1996; **24**: 1058–62.
43 Chen Z-B, Shen S-Q, Ding Y-M *et al.* The angiogenic and prognostic implications of VEGF, Ang-1, Ang-2, and MMP-9 for hepatocellular carcinoma with background of hepatitis B virus. *Med Oncol* 2009; **26**: 365–71.
44 Pérez-Moreno MA, Locascio A, Rodrigo I *et al.* A new role for E12/E47 in the repression of E-cadherin expression and epithelial-mesenchymal transitions. *J Biol Chem* 2001; **276**: 27424–31.
45 Yang J, Mani SA, Donaher JL *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* 2004; **117**: 927–39.
46 Lee TK, Poon RTP, Yuen AP *et al.* Twist overexpression correlates with hepatocellular carcinoma metastasis through induction of epithelial-mesenchymal transition. *Clin Cancer Res* 2006; **12**: 5369–76.

doi: 10.1111/j.1349-7006.2010.01536.x