Mechanisms of disease

Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection

Norio Iizuka, Masaaki Oka, Hisafumi Yamada-Okabe, Minekatsu Nishida, Yoshitaka Maeda, Naohide Mori, Takashi Takao, Takao Tamesa, Akira Tangoku, Hisahiro Tabuchi, Kenji Hamada, Hironobu Nakayama, Hideo Ishitsuka, Takanobu Miyamoto, Akira Hirabayashi, Shunji Uchimura, Yoshihiko Hamamoto

Summary

Background Hepatocellular carcinoma has a poor prognosis because of the high intrahepatic recurrence rate. There are technological limitations to traditional methods such as TNM staging for accurate prediction of recurrence, suggesting that new techniques are needed.

Methods We investigated mRNA expression profiles in tissue specimens from a training set, comprising 33 patients with hepatocellular carcinoma, with high-density oligonucleotide microarrays representing about 6000 genes. We used this training set in a supervised learning manner to construct a predictive system, consisting of 12 genes, with the Fisher linear classifier. We then compared the predictive performance of our system with that of a predictive system with a support vector machine (SVM-based system) on a blinded set of samples from 27 newly enrolled patients.

Findings Early intrahepatic recurrence within 1 year after curative surgery occurred in 12 (36%) and eight (30%) patients in the training and blinded sets, respectively. Our system correctly predicted early intrahepatic recurrence or non-recurrence in 25 (93%) of 27 samples in the blinded set and had a positive predictive value of 88% and a negative predictive value of 95%. By contrast, the SVM-based system predicted early intrahepatic recurrence or non-recurrence correctly in only 16 (60%) individuals in the blinded set, and the result yielded a positive predictive value of only 38% and a negative predictive value of 79%.

Interpretation Our system predicted early intrahepatic recurrence or non-recurrence for patients with hepatocellular carcinoma much more accurately than the SVM-based system, suggesting that our system could serve as a new method for characterising the metastatic potential of hepatocellular carcinoma.

Lancet 2003; **361:** 923–29

Departments of Bioregulatory Function (N Iizuka MD) **and Surgery II** (M Oka MD, M Nishida MD, Y Maeda MD, N Mori MD, T Takao MD, T Tamesa MD, A Tangoku MD), **Yamaguchi University School of Medicine, Ube, Yamaguchi, Japan; Department of Computer Science and Systems Engineering, Faculty of Engineering, Yamaguchi University, Ube, Yamaguchi** (T Miyamoto PhD, A Hirabayashi PhD, S Uchimura PhD, Y Hamamoto PhD); **and Department of Oncology, Nippon Roche Research Center, Kamakura, Kanagawa** (H Yamada-Okabe PhD, H Tabuchi PhD, K Hamada PhD, H Nakayama PhD, H Ishitsuka PhD)

Correspondence to: Dr Masaaki Oka, Department of Surgery II, Yamaguchi University School of Medicine, 1-1-1 Minami-Kogushi, Ube, Yamaguchi 755-8505, Japan
(e-mail: 2geka-1@po.cc.yamaguchi-u.ac.jp)

Introduction

Hepatocellular carcinoma is a common fatal cancer worldwide.[1] A major obstacle in its treatment is INTRAHEPATIC RECURRENCE, which arises in 30–50% of patients who undergo hepatic resection.[2–4] Intrahepatic recurrence therefore limits the potential of surgery as a cure for hepatocellular carcinoma. Although the pathological TNM (pTNM) staging system has been applied clinically to patients, this system is inadequate for predicting recurrence in individuals who undergo hepatic resection.[4] Likewise, several molecules have been proposed as predictive markers for hepatocellular carcinoma; however, none has been proven clinically useful.[2,5,6] Thus, there are technological limitations for predicting accurately recurrence by traditional methods. High interpatient heterogeneity of hepatocellular carcinoma can also limit the predictive ability of tests. Our aim was, therefore, to identify a new approach for the accurate prediction of early intrahepatic recurrence of hepatocellular carcinoma.

The development of microarray technologies, which allow us to undertake parallel analyses of many genes, has led to a new era in medical science.[7,8] In particular, complementary DNA microarray analysis of gene expression of tumours has provided great insights into their properties—eg, prognosis or drug sensitivity.[9,10] SUPERVISED AND UNSUPERVISED LEARNING methods have been used widely to create various gene-expression profiles.[11–17] We created a gene-expression profile of hepatocellular carcinomas with different viral infections, using a supervised learning method.[18] Supervised learning in STATISTICAL PATTERN RECOGNITION (SSPR) has been applied to resolve various issues, including document classification, speech recognition, biometric recognition, and remote sensing.[19] In this study, we applied the SSPR method to develop a system for accurate prediction of early intrahepatic recurrence of hepatocellular carcinomas.

Methods

Patients

Between May, 1997, and August, 2000, we assessed 33 patients at Yamaguchi University Hospital who had surgical treatment for hepatocellular carcinoma with routine X-ray, ultrasonography, CT scan, MRI, and hepatic angiography before surgery. These individuals comprised the training set. We identified space-occupying lesions in the remnant liver by intraoperative ultrasonography; no distant metastases or space-occupying lesions were identified in the non-resected part of the liver—ie, putative remnant liver—of any of the individuals.

We defined curative resection as complete excision of the tumour with clear microscopic margin and no residual tumours as indicated by ultrasonography and CT scan at 1 month after surgery.[3] To assess tumour size and undertake pathological examination, we sectioned resected specimens from the slice with the largest diameter, which we then cut at intervals of 5 mm. Two experienced

GLOSSARY

A-PRIORI PROBABILITY
The a-priori probability, which is used in the Fisher criterion, is defined as the probability of each class occurring. In this article, the a-priori probability of early intrahepatic recurrence is estimated as 30%, because early intrahepatic recurrence of hepatocellular carcinoma is usually observed in about 30% of patients after curative surgery.

CROSS VALIDATION
In cross validation, the set of labelled samples are randomly split into two parts: one is used as the training set for designing a classifier, the other is used to estimate the error. By repeating this procedure many times randomly, many different training and test sets are formed.

FISHER CRITERION
The Fisher criterion measures the difference between two means normalised by the averaged variance. Unlike the Euclidean distance or a criterion based on a fold change, the Fisher criterion takes account of the variance. Moreover, the magnitude of the Fisher criterion is invariant to scale, meaning that the genes are ranked by the Fisher criterion, irrespective of the magnitude of the gene-expression level.

FISHER LINEAR CLASSIFIER
To classify a pattern, the Fisher linear classifier measures the squared Mahalanobis distance[22] from a pattern to each of the mean vectors, and assigns a pattern to the category of the nearest mean. Geometrically, a pattern falls in hyperellipsoidal clusters of equal size and shape, the cluster for the each class being centred about each mean vector.

INTRAHEPATIC RECURRENCE
Intrahepatic recurrence is a characteristic feature of recurrence of human hepatocellular carcinoma. There are two types—ie, intrahepatic metastasis and multicentric occurrence. In general, the former represents early recurrence after surgery and is correlated with poor prognosis. The latter is a de-novo primary tumour in the liver remnant caused by continuous virus infection and inflammation, and usually appears as late recurrence.

LEAVE-ONE-OUT METHOD
In this method, virtually all labelled samples are used in each training, and all labelled samples are ultimately used in the tests, though each training and test set can be regarded as independent.

STATISTICAL PATTERN RECOGNITION
A given pattern is assigned to one of some categories based on its feature values. The features are assumed to have a density function conditioned on the pattern class. Thus, a pattern vector belonging to a class is viewed as an observation drawn randomly from the class-conditional density.

SUPERVISED AND UNSUPERVISED LEARNING
Supervised learning refers to situations in which available samples are labelled. In supervised learning, a teacher provides a category label for each sample—eg, early intrahepatic recurrence or non-recurrence. Unsupervised learning refers to situations in which available samples are not labelled. In unsupervised learning or clustering there is no explicit teacher, and the system forms clusters or natural groupings of the input patterns.

pathologists independently examined residual tumours in the surgical margin, tumour differentiation, and venous invasion in all samples without any information. On the basis of these examinations, operations on all 33 patients were judged to be curative resection.

The table shows the characteristics of the training set, and the pTNM classification of the Union Internationale Contre le Cancer (UICC).[20] We classified extended lobectomy, lobectomy, and segmentectomy as major hepatectomy, and subsegmentectomy and partial hepatectomy as minor hepatectomy.

To examine the predictive performance of two oligonucleotide array-based systems, we also assessed samples (blinded set) obtained from 27 newly enrolled patients with hepatocellular carcinoma who fulfilled the above criteria (table).

The study protocol was approved by the institutional review board for the use of human subjects at the Yamaguchi University School of Medicine, and all participants provided written informed consent.

We followed up all patients at least once every 3 months after surgery, and did ultrasography, CT scan, or MRI. We also measured serum concentrations of α fetoprotein and protein induced by vitamin K absence II (PIVKA-II). When tumour recurrence was suspected, hepatic angiography was included in the follow up. Within 1 year after surgery, this technique indicated recurrent liver tumours, which were enhanced during the hepatic arterial phase, in 12 (36%) of 33 patients in the training set and in eight (30%) of 27 patients in the blinded set. Among the 20 patients with recurrence, 17 had widespread multiple or diffuse type nodules in the liver remnant, which grew rapidly. Six of these 17 patients died of recurrent disease within 1 year after surgery. Of the 20 patients, three underwent hepatectomy again, and one died of recurrent disease and underwent autopsy. Histopathological examination revealed that all recurrent tumours in the four patients had moderate or poor differentiation. Thus, on the basis of imaging technologies, clinical course, and pathological findings, we judged that novel liver tumours in the 20 patients were recurrent and caused by intrahepatic metastasis.

Development of predictive system

We divided the resected specimens that contained no necrotic tissues into two groups immediately after surgery; one was subjected to oligonucleotide array analysis, and the other was fixed in 10% formaldehyde solution and embedded in paraffin, so that we could histopathologically confirm that tissues used for oligonucleotide array lacked sizeable necroses. Sample preparation, RNA extraction, and high-density oligonucleotide array were done as previously described.[18,21] We obtained data for each pixel level by laser scanner (Affymetrix, Santa Clara, CA, USA), and calculated the degree of expression of each complementary DNA and reliability (present/absent call) with Affymetrix GeneChip (version 3.3) and Affymetrix Microarray Suite (version 4.0), respectively.

We divided the patients in the training set into two groups: those who had early intrahepatic recurrence (group A, n=12) and those who did not (group B, n=21). We constructed a predictive scoring system with the FISHER LINEAR CLASSIFIER[18] based on the SSPR method as follows (figure 1). Panel 1 shows the notations used in the report.

To design a predictive scoring system with the number of samples we had available for training, we adopted a CROSS-VALIDATION approach.[19] We divided the 33 available samples into 30 training samples and three test samples, a process that we repeated ten times (step 1 in figure 1). On the basis of the A-PRIORI PROBABILITIES, we determined a training subset of 30, consisting of 11 samples from group A and 19 samples from group B. As a result, the test set of three samples consisted of one from group A and two from group B.

We wanted to select an optimum subset of genes (d). We selected the top 50 of 6000 genes on the basis of their individual effectiveness in each trial (steps 2 and 3) as follows. First, we identified all genes that had a mean average difference greater than two fold between groups A and B. We then selected the top 50 genes by means of the FISHER CRITERION (panel 2). Finally, we did an exhaustive search to assess the effectiveness of all possible subsets of d genes from the top 50 genes. The number of subsets to be searched was equal to $50!/(50-d)!d!$, where $d!=1\times2\times \ldots \times d$. With the exhaustive search, we investigated candidate gene subsets that minimised the error rate estimated from the LEAVE-ONE-OUT METHOD (step 4).[19] We used the Fisher linear classifier (panel 2) to estimate the error rate.[22]

	Training set			Blinded set		
	Early intrahepatic recurrence (n=12)	Non-recurrence (n=21)	p	Early intrahepatic recurrence (n=8)	Non-recurrence (n=19)	p
Sex (men/women)	8/4	16/5	0·691	7/1	13/6	0·633
Age (mean, SD) (years)	63·8 (10·8)	62·6 (7·4)	0·704	60·6 (5·3)	63·7 (8·7)	0·124
Viral infection (HBV/HCV/non-B non-C)	3/8/1	4/14/3	0·839	1/5/2	3/13/3	0·848
Total bilirubin (mean, SD) (μmol/L)	14·88 (3·69)	14·19 (4·06)	0·691	15·22 (3·83)	16·25 (5·85)	0·653
Serum ALT (<50 IU/L or ⩾50 IU/L)	9/3	12/9	0·457	4/4	12/7	0·657
Serum albumin (mean, SD) (g/L)	370 (40)	380 (40)	0·706	380 (50)	370 (50)	0·701
ICG-15 (mean, SD) (%)	17·3 (7·4)	15·1 (5·2)	0·331	17·9 (7·8)	16·2 (10·0)	0·68
Ascites (no/yes)	11/1	20/1	>0·999	8/0	17/2	>0·999
Cirrhosis (no/yes)	3/9	12/9	0·145	4/4	5/14	0·375
Tumour size (mean, SD) (cm)	5·4 (3·3)	3·4 (2·0)	0·033	4·9 (2·4)	3·5 (1·8)	0·114
Primary lesion (single/multiple)	3/9	13/8	0·041	3/5	11/5	0·420
Histological grading (G1/G2/G3)*	0/9/3	2/17/2	0·416	1/4/3	2/15/2	0·277
Venous invasion (no/yes)	5/7	17/4	0·052	1/7	15/4	0·0025
Hepatectomy (minor/major)	5/7	14/7	0·162	4/4	14/5	0·375
Stage (I/II/IIIA)*	2/5/5	10/9/2	0·075	0/7/1	10/7/2	0·022

HBV=hepatitis B virus; HCV=hepatitis C virus; non-B non-C=patient negative for both HBS antigen and HCV antibody; ALT=alanine aminotransferase; ICG-15=indocyanine green retention at 15 min; G1=well differentiated; G2=moderately differentiated; G3=poorly differentiated. *Assessment based on pTNM classification of UICC.

Relation between early intrahepatic recurrence and clinical and pathological characteristics

As a cross-validation, we repeated steps 2, 3, and 4 in figure 1 independently ten times with ten different training subsets. We obtained several candidate gene subsets of size d that minimised the error rate in each trial. Finally, we selected the optimum gene subset that appeared most frequently among the candidate gene subsets, appearing more than five times throughout the ten trials (step 5, figure 1). We then calculated a score by $T(x)$ (panel 2).

In general, as the number of genes increases, the recognition rate of a classifier designed with a finite number of training samples increases at first, reaches a maximum, and then decreases. This pattern is known as the peaking phenomenon in the statistical pattern recognition field.[19] In practice, the optimum number of genes is dependent on the classifier used and the training sample size. Here, we used criterion $\mathcal{J}$ (panel 2) to identify the optimum number of genes with the training samples (step 6).

In the leave-one-out method, the value of $\mathcal{J}$ was calculated for test samples. This calculation was also done for the ten different training subsets. The mean of the $\mathcal{J}$ values rose as the number of genes increased, indicating that the separation of group A from B was greater as the number of genes increased. However, the 95% CI estimated with 12 genes was almost equivalent to that obtained with 14 genes (figure 2). When 14 genes were used, saturation of $\mathcal{J}$ was observed. When d was equal to or greater than 16, we could not identify an optimum gene subset because there was no candidate subset that appeared more than five times throughout the ten trials. We thus concluded that 12 genes is the optimum number for our scoring system. Hence, we identified an optimum subset of 12 genes from the 6000 (figure 3). Expression of 11 of the 12 genes was downregulated in group A; the mean average expression of each of these genes in group A was less than half of that in group B (figure 3). By contrast, the mRNA for *SEMA3F* (HUMLUCA19 cosmid clone LUCA19 from 3p21.3) was expressed in group A at levels more than four-fold those expressed in group B (figure 3).

For each of the ten trials, we independently assessed three test samples with the scoring system based on 30 training samples and 12 genes (steps 7 and 8 in figure 1). The scoring system accurately predicted early intrahepatic recurrence or non-recurrence in all the test samples in the ten trials (figure 4). All patients with negative $T(x)$ values had early intrahepatic recurrences, whereas all those with positive $T(x)$ values had no recurrences.

On the basis of the above findings, we developed a scoring system for early intrahepatic recurrence, using all

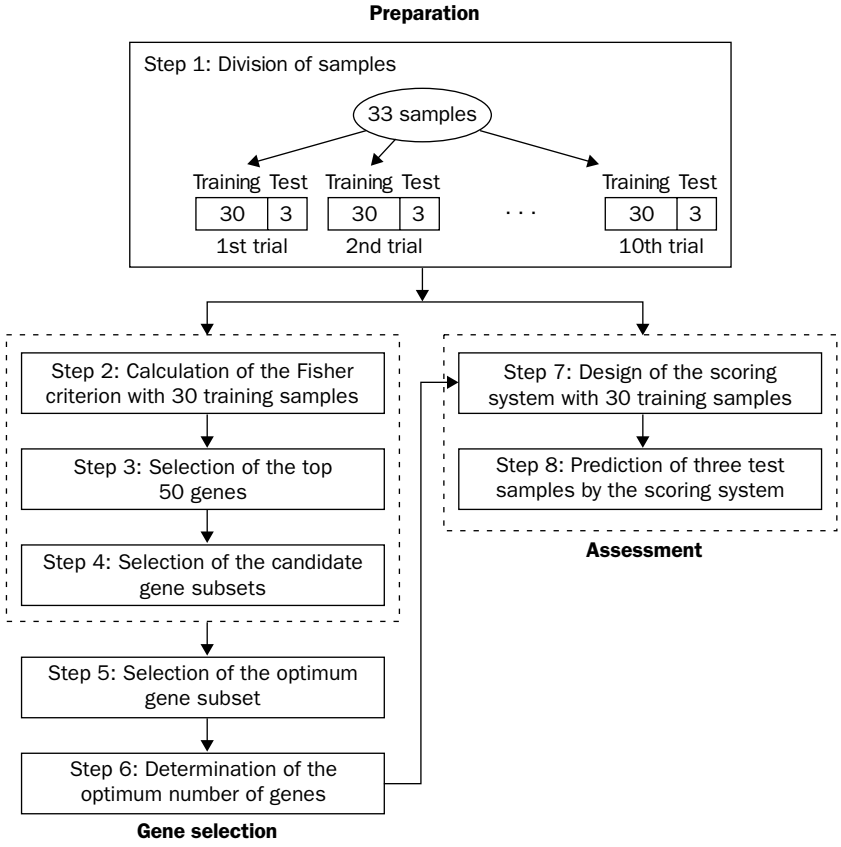

Figure 1: **Procedure for selection of optimum gene subset and system assessment**

A-priori probability $P(i)$ of group i is estimated by:

$$P(i)=\frac{n_i}{n_A+n_B}$$

where n_i is the number of training samples from group i.

The sample mean vector $\hat{\boldsymbol{\mu}}_i$ of group i is estimated by:

$$\hat{\boldsymbol{\mu}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j^i$$

where $\boldsymbol{x}_j^i$ is the jth training sample from group i.

And the sample covariance matrix $\hat{\Sigma}_i$ of group i is estimated by:

$$\hat{\Sigma}_i=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^i-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j^i-\hat{\boldsymbol{\mu}}_i)^T$$

Panel 1: **Notations used in report**
A vector is denoted in bold italic face.

33 hepatocellular carcinoma samples as training samples (panel 3).

We investigated the predictive performance of the support vector machine (SVM)-based system by increasing the number of genes from the top ten to the top 300 identified, according to a gene selection procedure described previously.[23] The SVM-based system performed best when the number of genes was 50 (data not shown).

Statistical analysis

We used the χ^2 test, Fisher's exact test, and Student's t test to assess differences in clinicopathological factors between recurrence and non-recurrence. These data were analysed with Statview (version 5.0). We did multivariate analysis to assess independent factors for early intrahepatic recurrence in the 60 samples, using the stepwise logistic regression model (SPSS, version 11.0J). Four variables (tumour size, number of primary lesion, venous invasion, and stage) were entered into a forward stepwise regression model. Each model was tested for goodness of fit by –2 log likelihood and χ^2 in each step. We also investigated independence between our predictive system ($T(x)$) and the above four variables in the blinded set by multivariate analysis. We judged a p value of less than 0·05 as significant.

The Fisher criterion (F) for a gene (j) is given by:

$$F(j)=\frac{(\hat{\mu}_j(A)-\hat{\mu}_j(B))^2}{P(A)\hat{\sigma}_j^2(A)+P(B)\hat{\sigma}_j^2(B)}$$

where $\hat{\mu}_j(i)$ is the jth component of the sample mean vector $\hat{\boldsymbol{\mu}}_i$ of group i, and $\hat{\sigma}_j^2(i)$ is the jth diagonal element of the sample covariance matrix $\hat{\Sigma}_i$ of group i.

The Fisher linear classifier

The Fisher linear classifier assigns a given $\boldsymbol{x}$ to be classified to group A if $f_A(\boldsymbol{x})<f_B(\boldsymbol{x})$ where:

$$f_A(\boldsymbol{x})=\tfrac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_A)^T[P(A)\hat{\Sigma}_A+P(B)\hat{\Sigma}_B]^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_A)-\log_e P(A)$$

Formula of $T(\boldsymbol{x})$

With use of the optimum subset of genes selected, the score is defined by:

$$T(\boldsymbol{x})=f_A(\boldsymbol{x})-f_B(\boldsymbol{x})$$

Formula of criterion J

The criterion J, which quantifies the separability of groups A and B, is defined by:

$$J=\tfrac{1}{30}\left[\sum_{\boldsymbol{x}\in B}T(\boldsymbol{x})-\sum_{\boldsymbol{x}\in A}T(\boldsymbol{x})\right]$$

where $\boldsymbol{x}$ is a test sample in the leave-one-out method.

Panel 2: **Equations used in report**
A vector is denoted in bold italic face.

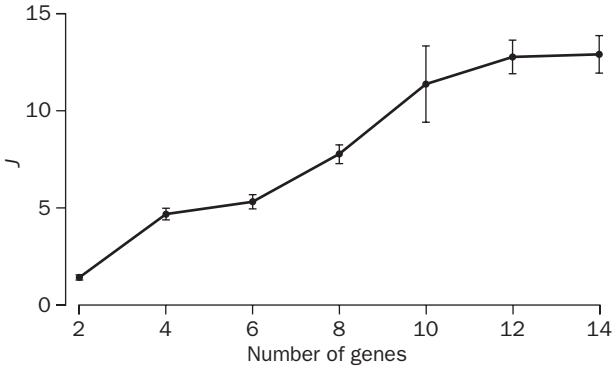

Figure 2: **Optimum number of genes for predicting early intrahepatic recurrence**
Error bars are 95% CI.

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Univariate analyses showed that early intrahepatic recurrence was associated with tumour size and the number of primary lesions in the training set, and venous invasion and stage in the blinded set (table). There was no association between early intrahepatic recurrence and other clinical or pathological factors. Multivariate analysis identified venous invasion as an independent risk factor for early intrahepatic recurrence in 60 hepatocellular carcinomas (risk ratio for venous invasion-negative cases 0·107, 95% CI 0·031–0·367; p<0·0001).

Our system correctly predicted early intrahepatic recurrence or non-recurrence in 25 of 27 blinded samples (93%; figure 5). The result also yielded a positive predictive value of 88% and a negative predictive value of 95%. Even when the most robust SVM-based system, comprising 50 genes, was applied to the blinded set, early intrahepatic recurrence or non-recurrence was predicted accurately in only 16 (60%) samples (figure 5). The SVM-based system had a positive predictive value of 38% and a negative predictive value of 79%. Thus, our system predicted early intrahepatic recurrence or non-recurrence in the blinded set much more accurately than the SVM-based system. Two of three patients with stage IIIA and four of 11 patients with venous invasion did not have recurrence within 1 year after surgery (figure 5). Thus, patients with advanced hepatocellular carcinoma did not necessarily have recurrence. Our system correctly predicted non-recurrence in those patients. Multivariate analysis showed that our system was independent from the four variables (tumour size, number of primary lesions, venous invasion, and stage) of early intrahepatic recurrence in the 27 blinded samples ($T(x)>0$ regression coefficient –4·836, SE 1·483, risk ratio 0·008 [95% CI 0·000–0·145], p=0.001)

Discussion

Although great progress has been made in the surgical treatment of hepatocellular carcinoma, recurrence of cancer

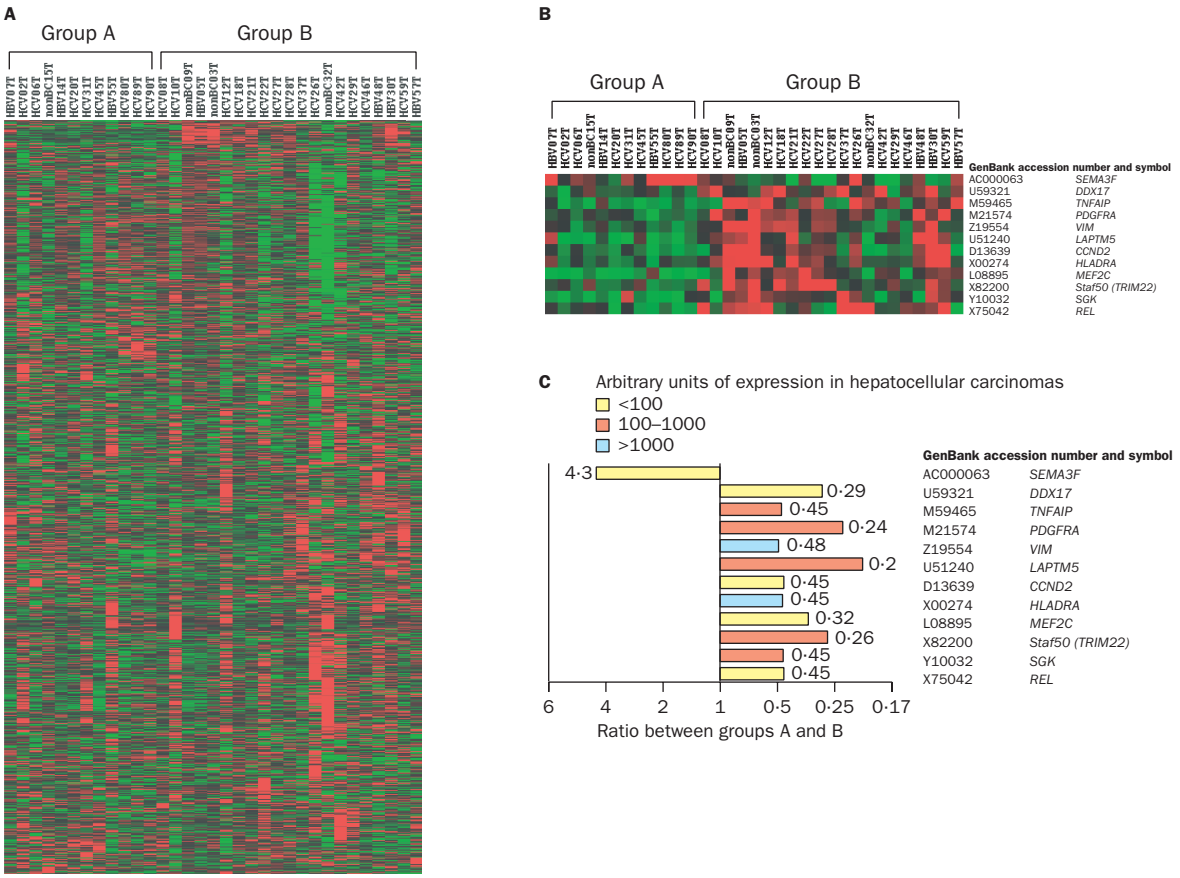

Figure 3: **Gene-expression profiles of hepatocellular carcinoma and patterns of expression of the 12 genes selected**
HBV-T=HBV-positive hepatocellular carcinoma; HCV-T=HCV-positive hepatocellular carcinoma; non BC-T=HBV-HCV-double negative hepatocellular carcinoma. A: Colour display of the expression of 6000 genes in group A and group B. The colour display was created with the EISEN program,[12] which displays the average expression of all genes (all raw data are available at http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/). B: Colour display of expression of the 12 genes used to design the scoring system. In A and B, red colour represents relatively high expression and green relatively low expression. C: Mean average differences in expression of the 12 genes between groups A and B.

after surgery as a result of either intrahepatic metastasis or multicentric occurrence remains common.[1–3] Metastases seem to arise from the early spread of tumour cells via the portal venous system within 1 year after surgery and are closely correlated with poor prognosis of hepatocellular carcinoma.[3] By contrast, multicentric occurrence involves a de-novo primary tumour and is thought to be affected by host factors rather than tumour factors.[3] We defined intrahepatic recurrence within 1 year after surgery as early intrahepatic recurrence. Consistent with a previous report,[3] we noted that early intrahepatic recurrence correlated with several tumour factors such as venous invasion and stage, but not with host factors in the training or blinded set. This finding lends support to the possibility that we can predict early intrahepatic recurrence on the basis of tumour factors. However, such factors were not necessarily consistent with early intrahepatic recurrence. In our study, venous invasion correlated most closely with early intrahepatic recurrence. Yet prediction based on venous invasion misclassified five of 27 blinded samples. This error rate is too high, and indicates that there are limitations for accurate prediction of early intrahepatic recurrence by tumour factors based on traditional methods.

Supervised learning has been introduced into gene-expression analysis over the past few years.[11,13–18] We used mRNA expression profiles at a primary site of hepatocellular carcinoma to develop a scoring system with the Fisher linear classifier for predicting early intrahepatic recurrence by the SSPR method. Our scoring system accurately predicted early intrahepatic recurrence or non-recurrence in most samples in the blinded set and was very accurate. Additionally, the performance of our system with only 12 genes was better than the best performance of the SVM-based system with 50 genes. When compared with the SVM-based system, our approach is advantageous in that it directly assesses the combination of genes. Thus, the difference in predictive performance between our system and the SVM-based system is possibly due to different gene selection procedures. Our findings show clearly that gene selection based on the SSPR method is useful for developing a reliable prediction system even when the number of available samples is small by comparison with

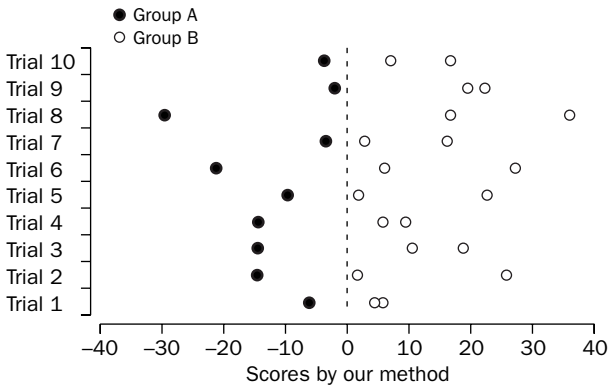

Figure 4: **Assessment of the scoring system on test samples in cross validation**

Gene (GenBank accession number)	Description	Function
$x1$ (M21574)	Platelet-derived growth factor receptor α (*PDGFRA*)	Signal transduction
$x2$ (M59465)	Tumour necrosis factor α inducible protein A20 (*TNFAIP3*)	Immune response
$x3$ (U51240)	Lysosomal-associated multitransmembrane protein (*LAPTM5*)	Protein interacting with ubiquitin
$x4$ (X00274)	HLA-DR α heavy chain (*HLADRA*)	Immune response/MHC class II antigen
$x5$ (X75042)	Rel proto-oncogene (*REL*)	Transcription/proto-oncogene
$x6$ (X82200)	Staf50 (*TRIM22*)	Transcription/interferon-inducible
$x7$ (Y10032)	Putative serine/threonine protein kinase (*SGK*)	Sodium transport/stress response
$x8$ (L08895)	MADS/MEF2-family transcription factor (*MEF2C*)	Transcription
$x9$ (AC000063)	HUMLUCA19 human cosmid clone LUCA19 from 3p21·3 (*SEMA3F*)	Embryonic development/cell motility
$x10$ (U59321)	DEAD-box protein p72 (*DDX17*)	RNA helicase/RNA processing
$x11$ (Z19554)	Vimentin (VIM)	Cytoskeleton/liver metastasis
$x12$ (D13639)	*KIAK0002* gene (*CCND2*)	Control of cell cycle

$$T(x)=0{\cdot}053862x_1+0{\cdot}038848x_2+0{\cdot}030176x_3+0{\cdot}001824x_4+0{\cdot}096997x_5+0{\cdot}017259x_6+0{\cdot}015908x_7+0{\cdot}103081x_8-0{\cdot}093746x_9+0{\cdot}024031x_{10}-0{\cdot}005417x_{11}-0{\cdot}119177x_{12}-11{\cdot}046007$$

Panel 3: **Scoring system, consisting of 12 genes to predict early intrahepatic recurrence**

the number of genes. On the basis of these findings, we suggest that the SSPR method with larger arrays would allow us to predict early intrahepatic recurrence of hepatocellular carcinoma with increased accuracy.

Our system correctly classified almost all early-stage and late-stage hepatocellular carcinomas, but was most accurate in advanced stages of disease. To address this possible deficiency of our system, we would have to study a larger number of early hepatocellular carcinomas as a training set. In the blinded set, none of the patients with stage I had recurrence. When our system was applied to patients only with stage II and stage IIIA, it correctly classified 94% of individuals. These findings suggest that, by combining our system with pTNM, we could correctly predict early intrahepatic recurrence or non-recurrence in a large number of samples.

The 12 genes used in our system seem to be involved in a wide range of biological processes. Many investigators have profiled gene expression responsible for various aspects of hepatocellular carcinoma.[24–26] Shirota and colleagues,[25] for example, showed that *PDGFRA* was downregulated in hepatocellular carcinoma tissue, but the relation of its expression to the metastatic potential of hepatocellular carcinoma was not discussed. Furthermore, the 12 genes selected in our study were not included in the list of venous invasion-related genes identified by Okabe and co-workers.[24] We selected genes by focusing on differences between hepatocellular carcinoma with early intrahepatic recurrence and that without. The discrepancy, therefore, might be in part due to differences in the focus of gene selection or the algorithm used. Among the 12 genes, two (*TNFAIP3* and *SGK*) were greatly downregulated in hepatocellular carcinoma with venous invasion (supplementary figure 1 at http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/). The levels of *SGK* transcript are altered in hepatoma cells in response to osmotic changes or cell volume changes; however, it remains unclear how *SGK* is related to venous invasion. Results of one study[27] revealed that *TNFAIP3* protects immune cells from TNF-induced apoptosis; thus downregulation might increase

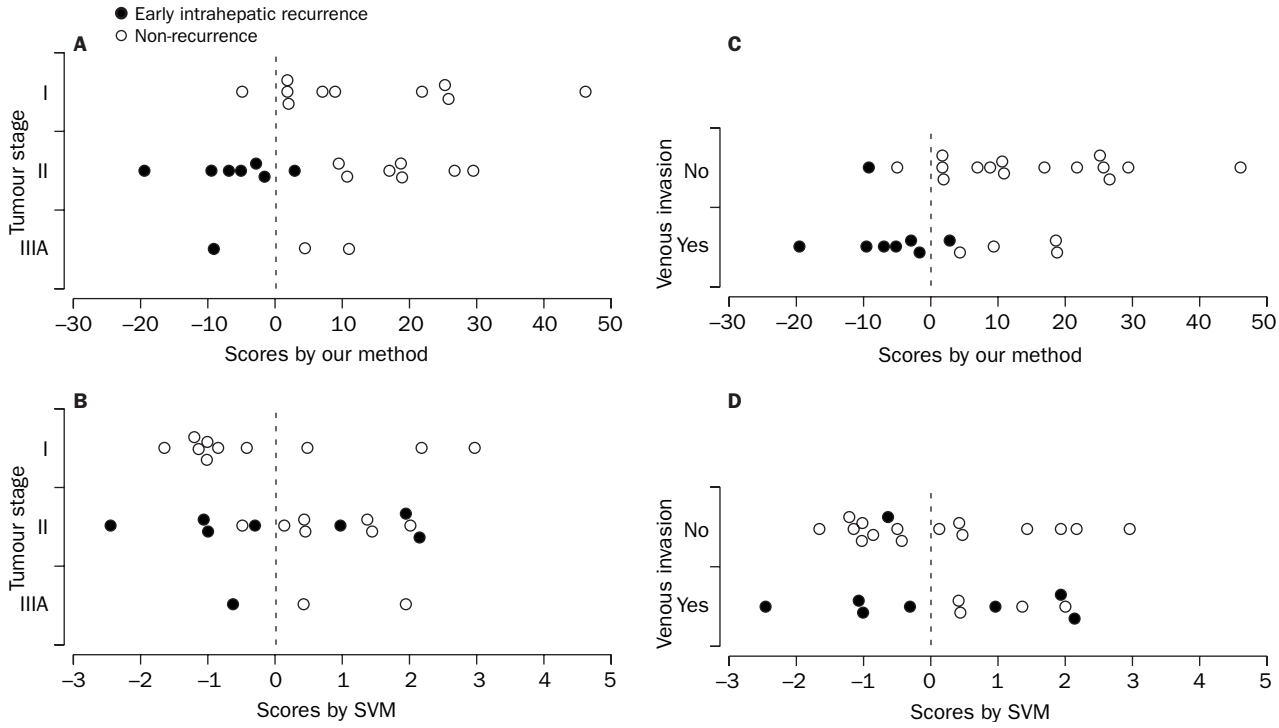

Figure 5: **Prediction of early intrahepatic recurrence in the blinded set with our system (A and C) and the SVM-based system (B and D)**

venous invasion via withdrawal of the immune system. Immune response-related genes, *HLADRA* and *TRIM22*, were also downregulated in hepatocellular carcinomas with early intrahepatic recurrence. Since *HLADRA* is thought to play an important part in the antigen-presenting system through its expression by macrophages, its downregulation in tumour tissues might permit tumour cells to escape from host immune surveillance.[28] Expression of *TRIM22* is induced by interferon.[29] Thus, downregulation of these genes suggests that a weak immune response against tumour cells exists. Additionally, Takayama and colleagues[30] showed that adopted immunotherapy can reduce recurrence and improve recurrence-free outcomes after surgery for hepatocellular carcinoma. Our system, based on the 12 genes, could be useful when such immunotherapy is being considered as a treatment option for patients with hepatocellular carcinoma undergoing hepatic resection. Thus, some of the genes selected here might be related to the immune response or metastatic potential of hepatocellular carcinoma. *KIAK0002* (*CCND2*), which is a member of the D-type cyclins, was downregulated in hepatocellular carcinomas with early intrahepatic recurrence. Loss of *CCND2* expression by promoter hypermethylation correlates with the tumour progression of breast cancer.[31] This fact suggests the possibility of DNA methylation in a cyclin D2 gene in progressed hepatocellular carcinoma and selected genes might represent tumour progression rather than metastatic potential. Immunohistochemical examination also showed that the primary source of vimentin (VIM) protein was stromal cells but not cancer cells and its expression level was reduced in hepatocellular carcinoma with early intrahepatic recurrence (supplementary figure 2 at http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/). Thus, some of the genes selected here could represent stromal change during development of hepatocellular carcinoma. Further biological studies on the 12 genes is required to gain key insights into our findings.

In conclusion, the SSPR method yielded a unique predictive system with high accuracy that was independent of known prognostic variables. Our system constructed by the SSPR method will be useful for characterising the metastatic potential of individual hepatocellular carcinomas, and the SSPR method could provide new insights into bioinformatics for microarray data.

Contributors

N Iizuka conceived the study and wrote the report. M Oka undertook the study, helped to write the report, and did hepatic resections. M Nishida, Y Maeda, N Mori, T Takao, T Tamesa, and A Tangoku helped do hepatic resections. H Yamada-Okabe, H Tabuchi, K Hamada, H Nakayama, and H Ishitsuka analysed gene expression with oligonucleotide arrays and wrote the array procedure. Y Hamamoto, T Miyamoto, S Uchimura, and A Hirabayashi undertook gene selection, developed a predictive system based on supervised learning in statistical pattern recognition, and helped to write the report.

Conflict of interest statement

None declared.

Acknowledgments

We thank F Ford for proofreading the manuscript. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, Sports, and Culture of Japan (12671230).

References

1 Okuda K. Hepatocellular carcinoma. *J Hepatol* 2000; **32:** 225–37.
2 Iizuka N, Oka M, Noma T, Nakazawa A, Hirose K, Suzuki T. NM23-H1 and NM23-H2 messenger RNA abundance in human hepatocellular carcinoma. *Cancer Res* 1995; **55:** 652–57.
3 Poon RT, Fan ST, Ng IO, Lo CM, Liu CL, Wong J. Different risk factors and prognosis for early and late intrahepatic recurrence after resection of hepatocellular carcinoma. *Cancer* 2000; **89:** 500–07.
4 Izumi R, Shimizu K, Ii T, et al. Prognostic factors of hepatocellular carcinoma in patient undergoing hepatic resection. *Gastroenterology* 1994; **106:** 720–27.
5 Hsu HC, Tseng HJ, Lai PL, Lee PH, Peng SY. Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness. *Cancer Res* 1993; **53:** 4691–94.
6 Mathew J, Hines JE, Obafunwa JO, Burr AW, Toole K, Burt AD. CD44 is expressed in hepatocellular carcinomas showing vascular invasion. *J Pathol* 1996; **179:** 74–79.
7 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270:** 467–70.
8 DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996; **14:** 457–60.
9 Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403:** 503–11.
10 Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000; **24:** 236–44.
11 Brazma A, Vilo J. Gene expression data analysis. *FEBS Lett* 2000; **480:** 17–24.
12 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95:** 14863–68.
13 Kell DB, King RD. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol* 2000; **18:** 93–98.
14 Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286:** 531–37.
15 Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000; **97:** 262–67.
16 Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001; **7:** 673–79.
17 Luo J, Duggan DJ, Chen Y, et al. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 2001; **61:** 4683–88.
18 Iizuka N, Oka M, Yamada-Okabe H, et al. Comparison of gene expression profiles between hepatitis B virus- and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data based on a supervised learning method. *Cancer Res* 2002; **62:** 3939–44.
19 Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Trans Pattern Analysis Machine Intelligence* 2000; **22:** 4–37.
20 Sobin LH, Wittekind C. TNM classification of Malignant Tumours, 6th edn. New York: UICC, Wiley-Liss, 2002: 81–83.
21 Lipshutz RL, Fodor SPA, Gingeras TR, Lockhart DJ. High density oligonucleotide arrays. *Nat Genet* 1999; **21:** 20–24.
22 Duda RO, Hart PE, Stork, DG. Pattern classification, 2nd edn. New York: John Wiley and Sons, 2001: 1–152.
23 Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000; **16:** 906–14.
24 Okabe H, Satoh S, Kato T, 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* 2001; **61:** 2129–37.
25 Shirota Y, Kaneko S, Honda M, Kawai HF, Kobayashi K. Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. *Hepatology* 2001; **33:** 832–40.
26 Xu L, Hui L, Wang S, et al. Expression profiling suggested a regulatory role of liver-enriched transcription factors in human hepatocellular carcinoma. *Cancer Res* 2001; **61:** 3176–81.
27 He KL, Ting AT. A20 inhibits tumor necrosis factor (TNF) alpha-induced apoptosis by disrupting recruitment of TRADD and RIP to the TNF receptor 1 complex in Jurkat T cells. *Mol Cell Biol* 2002; **22:** 6034–45.
28 Cabrera T, Ruiz-Cabello F, Garrido F. Biological implications of HLA-DR expression in tumours. *Scand J Immunol* 1995; **41:** 398–406.
29 Tissot C, Mechti N. Molecular cloning of a new interferon-induced factor that represses human immunodeficiency virus type 1 long terminal repeat expression. *J Biol Chem* 1995; **270:** 14891–98.
30 Takayama T, Sekine T, Makuuchi M, et al. Adoptive immunotherapy to lower postsurgical recurrence rates of hepatocellular carcinoma: a randomised trial. *Lancet* 2000; **356:** 802–07.
31 Evron EC, Umbricht D, Korz V, et al. Loss of cyclin D2 expression in the majority of breast cancers is associated with promoter hypermethylation. *Cancer Res* 2001; **61:** 2782–87.