Cancer-Associated Molecular Signature in the Tissue Samples of Patients With Cirrhosis

Jin Woo Kim,[1] Qinghai Ye,[2] Marshonna Forgues,[1] Yidong Chen,[3] Anuradha Budhu,[1] Jessica Sime,[1] Lorne J. Hofseth,[1] Rashmi Kaul,[4] and Xin Wei Wang[1]

Several types of aggressive cancers, including hepatocellular carcinoma (HCC), often arise as a multifocal primary tumor. This suggests a high rate of premalignant changes in noncancerous tissue before the formation of a solitary tumor. Examination of the messenger RNA expression profiles of tissue samples derived from patients with cirrhosis of various etiologies by complementary DNA (cDNA) microarray indicated that they can be grossly separated into two main groups. One group included hepatitis B and C virus infections, hemochromatosis, and Wilson's disease. The other group contained mainly alcoholic liver disease, autoimmune hepatitis, and primary biliary cirrhosis. Analysis of these two groups by the cross-validated leave-one-out machine-learning algorithms revealed a molecular signature containing 556 discriminative genes ($P < .001$). It is noteworthy that 273 genes in this signature (49%) were also significantly altered in HCC ($P < .001$). Many genes were previously known to be related to HCC. The 273-gene signature was validated as cancer-associated genes by matching this set to additional independent tumor tissue samples from 163 patients with HCC, 56 patients with lung carcinoma, and 38 patients with breast carcinoma. From this signature, 30 genes were altered most significantly in tissue samples from high-risk individuals with cirrhosis and from patients with HCC. Among them, 12 genes encoded secretory proteins found in sera. In conclusion, we identified a unique gene signature in the tissue samples of patients with cirrhosis, which may be used as candidate markers for diagnosing the early onset of HCC in high-risk populations and may guide new strategies for chemoprevention. ***Supplementary material for this article can be found on the HEPATOLOGY website (http://www.interscience.wiley.com/jpages/0270-9139/suppmat/index.html).*** **(HEPATOLOGY 2004;39:518–527.)**

Primary liver carcinoma is the fifth most frequent cancer in the world (there were an estimated 548,554 deaths in 2000). There has been a sharp increase in the incidence of primary liver carcinoma in the United States during the last decade.[1,2] Hepatocellular carcinoma (HCC) is the major type of primary liver carcinoma. Currently, survival remains poor for most patients with HCC, which is due to the aggressiveness of the lesions at the time of diagnosis and the lack of effective therapy. Although routine screening of individuals who are at risk for developing HCC may lead to early diagnosis and extend survival, most patients are still diagnosed with advanced HCC. The prognosis for these patients is poor.[3] Surgical resection of small HCCs diagnosed in the early stage may be potentially effective. However, 70% of these patients develop recurrent tumors after 5 years.[4] Therefore, current diagnostic and therapeutic approaches are inadequate.

The current dogma for tumor evolution is that a tumor is initiated from clonal expansion of an initiated cell with a mutation either in a tumor suppressor gene or oncogene, followed by an acquisition of sequential multiple genetic changes.[5] Similarly, HCC development has been speculated to be a multistage process because of its progressively

Abbreviations: HCC, hepatocellular carcinoma; cDNA, complementary DNA; CLD, chronic liver disease; HBV, hepatitis B virus; HCV, hepatitis C virus; HHC, hemochromatosis; WD, Wilson's disease; ALD, alcoholic liver disease; AIH, autoimmune hepatitis; PBC, primary biliary cirrhosis; LTPADS, Liver Tissue Procurement and Distribution System; KNN, k-nearest neighbor; SVM, support vector machine.

From the [1]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD; [2]Liver Cancer Institute and Zhongshan Hospital, Fudan University, Shanghai, China; [3]Cancer Genetics Branch, National Human Genome Research Institute, Bethesda, MD; and [4]Division of Pediatric Gastroenterology and Nutrition, University of Minnesota, Minneapolis, MN.

Received August 15, 2003; accepted November 2, 2003.

Supported, in part, by the Intramural Research Program of the Center for Cancer Research, National Cancer Institute.

Address reprint requests to: Dr. Xin Wei Wang, Laboratory of Human Carcinogenesis, CCR, NCI, NIH, 37 Convent Drive, Building 37, Room 3044A, Bethesda, MD 20892. E-mail: xw3u@nih.gov; fax: 301-496-0497.

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/hep.20053

pathologic morphology.[6] However, the stepwise progress in human HCC is ill defined and specific genetic changes associated with HCC progression remain vague.[7] This has been explained by the finding that most HCC tissue specimens used for analysis may be at an advanced stage and most of the changes found in these lesions can be secondary. Multiple genetic and epigenetic changes have been found in morphologically altered hepatocytes from patients with HCC.[7] Several studies have focused on the utilization of the complementary DNA (cDNA) microarray approach to identify unique gene sets that are abnormally expressed in HCC.[8–11] Such an approach has yielded a wealth of information about new potential tumor markers that may be useful for diagnosing the onset of HCC. However, many of the markers can be late events due to the secondary effect of tumor progression and the advanced stage of the tumors. One of the common features for HCC is that it often presents as a multifocal primary tumor.[3] This feature also is associated frequently with other types of aggressive tumors in the oral cavity, breast, skin, and aerodigestive tract.[12,13] Furthermore, most HCCs developed in patients with chronic liver diseases (CLD).[3] These observations suggest that a high degree of premalignant changes may take place in noncancerous tissue before clonal expansion of a clinical tumor mass. Identification of these premalignant changes may be useful for early cancer detection.

Accordingly, a microarray-based strategy was followed by focusing on preneoplastic CLD to search for genes that were altered in the early stage of HCC (Fig. 1). This strategy was based on a unique feature associated with HCC development. Most patients with HCC have accompanying CLD with underlying hepatitis, fibrosis, and cirrhosis.[3,14] Predisposing factors associated with HCC, such as viral hepatitis infection, alcohol abuse, metabolic disorders, and other environmental agents, also induce cirrhosis.[15,16] However, it is not clear whether these factors induce HCC directly or whether they act indirectly by producing chronic liver injury and regeneration. Epidemiologic data indicate that patients with chronic hepatitis B (HBV) or C virus (HCV) infection seem to be at an extremely high risk for developing HCC.[17] Clinically, HCC is mainly a viral hepatitis-associated cancer, because greater than 85% of HCCs worldwide retain markers of HBV and/or HCV.[18] Similarly, patients with genetic disorders, such as hemochromatosis (HHC) and type I tyrosinemia, who develop CLD are at a high risk for developing HCC.[14] In contrast, patients with other CLD including Wilson's disease (WD), alcoholic liver disease (ALD), primary biliary cirrhosis (PBC), and autoimmune hepatitis (AIH) may have a relatively low risk for developing HCC.[14,16,19–22] However, all of these CLD share

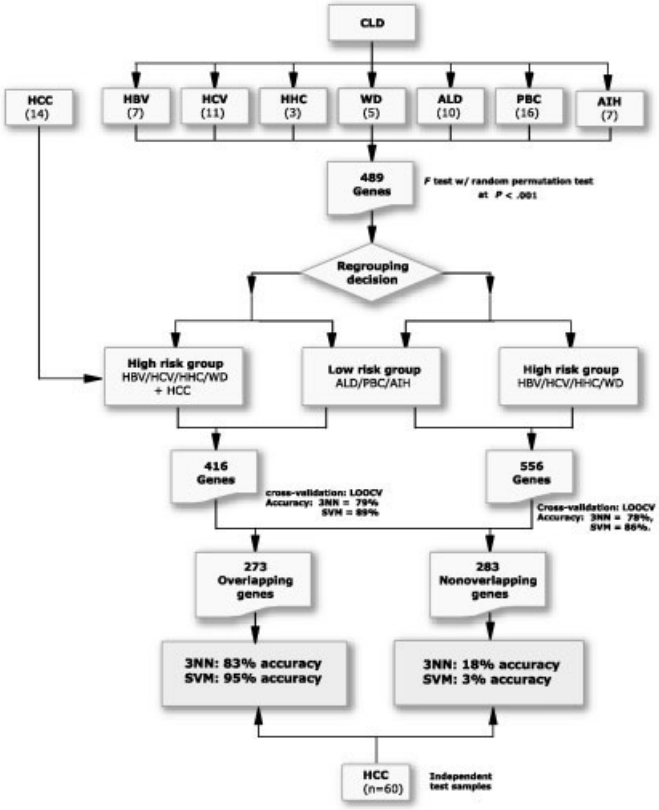

Fig. 1. Schematic illustration of the analysis strategy and outcomes. Initial dataset consisted of 59 experiments for the seven CLD groups. The *F* test was used to define a gene set that can discriminate these etiologies. Based on the hierarchical clustering analysis, a regrouping decision was made to divide these samples into two groups that reflect the risks of patients to develop HCC. Two models were built using both leave-one-out cross-validated (LOOCV) KNN and SVM algorithms. One model was derived from a comparison between high-risk and low-risk groups and the second model was generated by comparing the high-risk group plus HCC to the low-risk group. These models were further calibrated by comparing the two gene sets to define two new models, one derived from 273 overlapping genes (mostly enriched with HCC-associated genes) and the other from 283 nonoverlapping genes (mostly reflecting the etiologies of CLD). The new models were then validated with 60 independent HCC tissue samples as testing sets and the high-risk and low-risk groups as training sets using both KNN and SVM predictor algorithms. Numeric values in parentheses indicate the number of patients.

common features including liver inflammation, lymphocytic infiltration, liver regeneration, fibrosis, and cirrhosis.[14] We hypothesized that changes in gene expression specific to HCC may occur in preneoplastic CLD. Therefore, tissue specimens from 59 patients with cirrhosis who received liver transplantations were used in the current study. To validate potential HCC-associated genes, we also included HCC surgery specimens from three independent cohorts (14 U.S. patients, 60 Shanghai patients, and 103 Hong Kong patients). Microarrays containing greater than 9,000 human genes were employed. Supervised machine-learning algorithms with cross-validation

were used to compare high-risk patients with CLD with low-risk patients with CLD. This approach allowed us to identify misregulated genes that are commonly associated with the high-risk CLD group and HCC.

Patients and Methods

Patients and Tissue Samples. Surgical tissue specimens were collected after informed consent was obtained from the subjects. The protocols were approved by the institutional review board of the University of Minnesota (Minneapolis, MN) and the National Institutes of Health (Bethesda, MD). Tissue samples were obtained from 59 patients with end-stage CLD who received liver transplantation during 1995 to 2001. Tissue samples from eight disease-free normal liver donors were used as controls. Tissue sample collection was managed through the Liver Tissue Procurement and Distribution System (LTPADS) at the University of Minnesota. Tumor and matched nontumor tissue samples from 74 patients were obtained through either the LTPADS program or the Liver Cancer Institute at Fudan University (Shanghai, China). Initial clinical diagnosis and laboratory tests to define various etiologies were performed by primary physicians who contribute to the LTPADS program. Pathologic diagnosis was performed independently by two pathologists. Detailed histories for each tissue sample, such as age, sex, ethnicity, initial clinical diagnosis, history of alcohol use, and viral hepatitis status are available as Supplemental Table 1 on the HEPATOLOGY website (http://www.interscience.wiley.com/jpages/0270-9139/suppmat/index.html) and in Ye et al.[23] Total RNA samples were extracted from snap-frozen tissue sections using Trizol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol. Total RNA samples from eight control tissue samples were combined and used as a common reference pool. These control tissue samples were negative for HBV or HCV markers and presented normal histologic features (Supplemental Table 1 and data not shown).

cDNA Microarray. cDNA microarrays were generated by the National Cancer Institute (Bethesda, MD) microarray facility at the Advanced Technology Center. The array was based on the Incyte human UniGem version 2.0 platform containing 9,180 cDNA clones that map to 8,281 unique UniGene clusters and 122 EST clones from Incyte Genomics (Palo Alto, CA). For each experiment, fluorescent cDNA probes were prepared from a common reference pool RNA (Cy3) and a disease sample total RNA (Cy5). Detailed microarray platform, hybridization, quality control, data acquisition, and data filtering were performed essentially as previously described.[23]

Analysis and Statistics. All analyses were performed using the BRB ArrayTools (version 3.0).[23] This Excel-based platform contains several statistical tools including hierarchical clustering, class comparison, and class prediction. The class comparison tool used the F test to compute the number of genes that were differentially expressed among different etiologic groups at a statistically significant level ($P < .001$) with median-centered log-ratios expression data. The permutation distribution of the F statistic, based on 2,000 random permutations, was used to confirm statistical significance. Two machine-learning class prediction tools, namely, the k-nearest neighbor predictor (K = 3; KNN) and the support vector machine predictor (SVM), were applied. The nearest neighbor predictor was based on determining which expression profile in the training set was most similar to the expression profile of the specimen whose class was to be predicted. Euclidean distance was used as the distance metric for the nearest neighbor predictor. Once the nearest neighbor in the training set of the test specimen was determined, the class of that nearest neighbor was used as the prediction of the class of the test specimen. In the current study, the expression profile of the test specimen was compared with the expression profiles of all specimens in the training set and the three specimens in the training set most similar to the expression profile of the test specimen were determined. The distance metric was again Euclidean distance with regard to the genes that were univariately, significantly, and differentially expressed between the two classes at $P < .001$. Once the three nearest specimens were identified, their classes vote and the majority class among the three was the class predicted for the test specimen. SVM is a machine-learning algorithm that has the potential to include collective and nonlinear effects among the genes.[24] However, our SVM algorithm is based on linear kernel functions as previous experience has been that more complex SVMs perform less well for this application (R. Simon, BRB Array Tools manual). Therefore, our SVM predictor was a linear function of the log ratios or the log intensities that best separated the data subject to penalty costs on the number of specimens misclassified. Both class predictors were based on a leave-one-out cross-validation test and on 2,000 random permutations of the class labels using CLD as training samples to generate weights for predicting independent tumor samples. Averaged gene expression data from duplicate samples were included in the analysis.

To test our signature genes with publicly available microarray data, we downloaded the raw intensity data from http://genome-www5.stanford.edu/MicroArray/SMD/ and converted these data to match our gene ID. Because these datasets utilized a universal reference RNA as their

control, the data were converted by dividing the intensity channels of the tumor samples by the intensity channels of the noncancerous tissue samples. The gene intensities from each tumor sample were normalized against the average gene intensities derived from several available corresponding noncancerous tissue samples. The expression ratios were then converted to the log2 base and were normalized by median centering. Once converted, the discriminatory weights of the specific gene sets from the comparison of the high-risk and low-risk groups were then applied to these samples to provide a final vote. A correct prediction was an indicator of a similar gene expression profile between the high-risk samples and tumor samples with a given signature. Only the breast carcinoma, HCC, and lung carcinoma datasets were chosen because they had appropriate noncancerous tissue samples included and had sufficient matched genes for our signatures.

Results

Gene Expression Profiles of Liver Disease With Various Etiologies. To search for genes abnormally expressed in both HCC and CLD, we initially compared the gene expression profiles of cirrhotic tissue samples obtained from 59 patients with end-stage CLD and tissue samples from 14 patients with HCC with a pool of eight normal tissue samples by microarray containing 9,180 human cDNA clones (Fig. 1). The CLD samples were obtained from patients with HBV (n = 7) and HCV (n = 11) infection, HHC (n = 3), WD (n = 5), ALD (n = 10), PBC (n = 16), and AIH (n = 7). Because the global expression profile of preneoplastic liver tissue samples did not satisfactorily separate these samples based on their etiologies (Supplemental S-1), we used a supervised univariate *F* test algorithm to search for genes that can discriminate these seven CLD groups. This analysis yielded a total of 489 discriminative genes ($P < .0005$). Hierarchical clustering analysis[25] of the 489 genes revealed that these seven groups were separated into two major groups, one consisting mostly of tissue samples from patients with HBV infection, HCV infection, HHC, and WD and the other containing mainly tissue samples from patients with PBC, ALD, and AIH (Fig. 2A). These results indicate that HBV infection, HCV infection, HHC, and WD are related more closely to each other than PBC, ALD, and AIH. The segregation of these tissue samples by a molecular signature specifically reflecting their etiologies correlates with the risk of these patients to develop HCC, with the exception of WD tissue samples (Fig. 2A).

Defining HCC-Associated Genes in Cirrhotic Tissue Samples. We hypothesized that genes that were commonly misregulated in HBV/HCV/HHC/WD tissue samples, but not ALD/PBC/AIH, would more closely resemble the molecular signature of HCC. Therefore, the decision to regroup was made to include HBV infection, HCV infection, HHC, and WD as the high-risk group and ALD, PBC, and AIH as the low-risk group. To globally search for such a gene set, we applied KNN (K = 3) and SVM algorithms to the high-risk (HBV/HCV/HHC/WD) and low-risk (ALD/PBC/AIH) groups, which is a computation strategy similar to that recently published.[23] This analysis yielded a composite classifier containing 556 discriminative genes ($P < .001$), which separated these two groups very well. It provided a significant class prediction among these groups with an overall cross-validation accuracy of 78% by KNN and 86% by SVM. The cross-validated misclassification rates were significantly lower than expected by chance ($P < .0005$; Table 1). Using SVM, five tissue samples from the high-risk group (n = 26) and three tissue samples from the low-risk group (n = 33) were misclassified. Because of the limitation of the computation model used, it is unclear whether the misclassified tissue samples represent a meaningful class or just simply background noise. In contrast, random grouping of these tissue samples yielded statistically insignificant classification (Supplemental Table 2).

Many genes in the 556 gene set were found in the HCC tissue samples (Fig. 2B). To search for genes that were commonly misregulated in the high-risk group and in HCC, we included 14 HCC tissue samples (all from U.S. patients; two were HBV positive, seven were HCV positive, and five had unknown etiology) together with the high-risk group and compared this with the low-risk group using both KNN and SVM. This analysis yielded 416 discriminative genes ($P < .001$), of which 273 genes were found in the 556 gene set (49% overlap). The analysis also resulted in an overall cross-validation accuracy of 79% by KNN and 89% by SVM. The cross-validated misclassification rates were significantly lower than expected by chance (KNN, $P < .001$; SVM, $P < .0005$; Fig. 1 and Supplemental Table 3). These results indicate that approximately one-half of the signature genes that can discriminate between the high-risk and the low-risk groups may also be misregulated in HCC tissue samples.

To determine whether the 273 gene set was a common signature for tumors, we applied this set to 60 independent HCC tissue samples (from the Shanghai patients; all HBV positive) described recently,[23] using both KNN and SVM predictors. The 273 gene signature provided an increased fitness by SVM as an indicator to match 95% of the new HCC tissue samples (Fig. 3), which was an im-

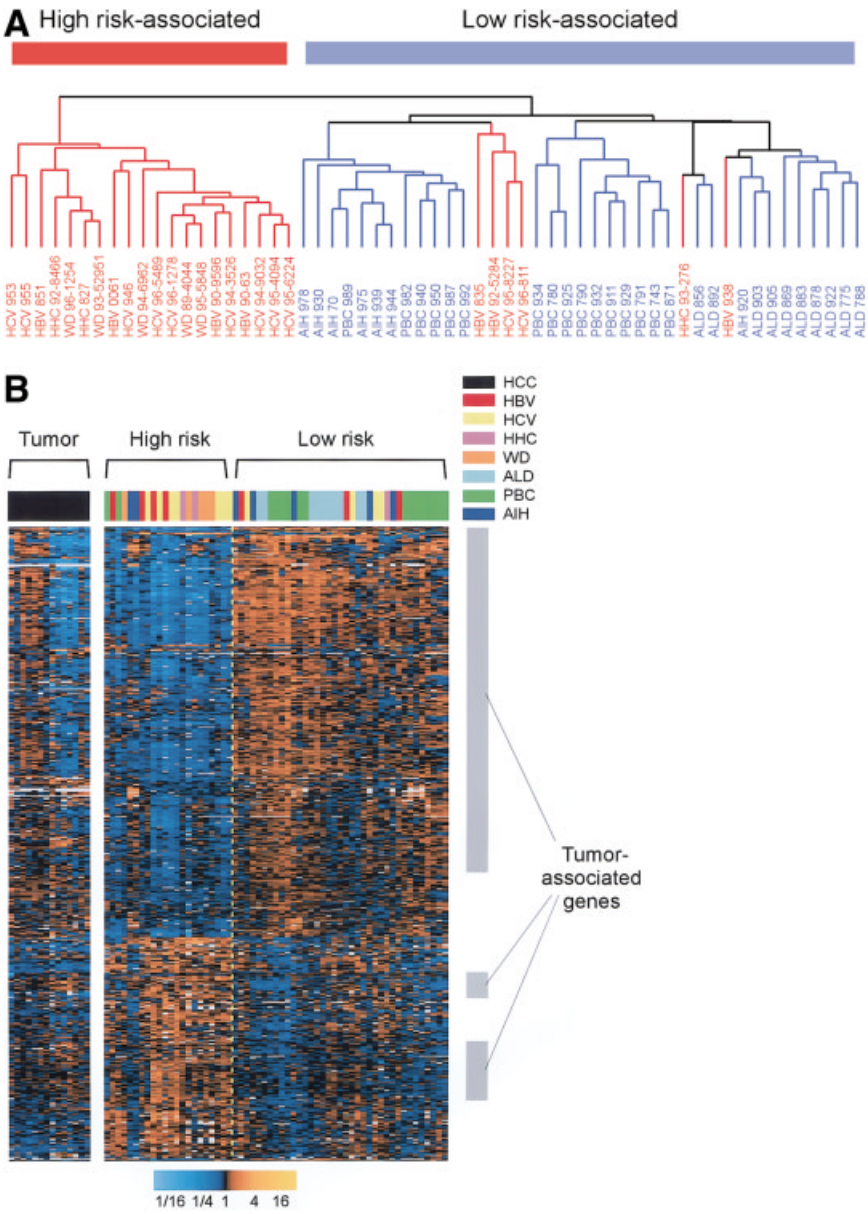

Fig. 2. Classification of CLD with various underlying etiologies by gene expression. (**A**) Hierarchical clustering of 59 CLD tissue samples using 489 significant genes ($P < .0005$) derived from supervised class comparison. Dendrogram has two large branches. The high-risk group tissue samples are labeled in red and the low-risk group tissue samples are labeled in blue. (**B**) Hierarchical clustering of 556 genes that separate the high-risk group from the low-risk group. Each row represents an individual gene and each column represents an individual tissue sample. Genes were ordered by Euclidean distance and average linkage according to the ratios of their abundance in each tissue sample when compared with a normal tissue sample pool, which were normalized to the median abundance of all genes. Pseudocolors indicate differential expression (blue, transcript levels below the median; black, transcript levels equal to the median; yellow, transcript levels greater than the median; gray, missing data). The scale represents the gene expression ratios from 1/16 to 16 in log-base two units.

provement in overall performance when compared with the 556 gene set. Therefore, we referred to the 273 gene set as the HCC-associated signature (see Supplemental Table 4 for the detailed gene list). In contrast, the nonoverlapping 283 gene set did not provide any satisfactory performance (Fig. 3). Because we eliminated most of the HCC-associated genes in the nonoverlapping gene set, most of these genes may belong to the signatures separating the etiologies. We referred to this gene set as the etiology-associated signature (see Supplemental Table 5 for the detailed gene list). The etiology-associated signature serves as a good negative control to provide the confidence of the test sets.

To minimize the number of genes in the model, we selected the 30 biologically most significant genes by filtering genes with lower P values ($P > .0003$) and larger t values from the HCC-associated signature (see Supplemental Table 6 for filtering criteria). The 30 gene set also performed well as an HCC indicator (Fig. 3). Filtering genes using these approaches did not alter the classification accuracy to the original 59 CLD training samples because the 273, 283, and the 30 gene set yielded a comparable cross-validation accuracy to classify these samples when compared with the 556 gene set (Supplemental Table 7). Among the proteins encoded by the 30 gene set, 12 proteins were identified as being secretory and five proteins as being cell surface associated (Fig. 4). Only three genes were up-regulated, whereas the remaining 27 genes were down-regulated. Table 2 provides a detailed description of these genes.

Table 1. Performances of the KNN and SVM Classifiers on Premalignant Liver Samples in Relation to Their Potential Risks to Develop HCC

Predefined Clinical Group	Potential Risks	Cases Correctly Predicted (%) KNN	SVM	n
HBV/HCV/HHC/WD	High	73	81	26
ALD/PBC/AIH	Low	82	91	33
Overall accuracy		78	86	59

NOTE. Analyses were based on leave-one-out cross-validated classification with 2,000 random permutations at the significant level of $P = .001$. The analyses yielded a composite classifier containing 556 genes. The cross-validated misclassification rate was significantly lower than expected by chance ($P < .0005$).

Abbreviations: KNN, K-nearest neighbor; SVM, support vector machine; HCC, hepatocellular carcinoma; HBV, hepatitis B virus; HCV, hepatitis C virus; HHC, hemochromatosis; WD, Wilson's disease; ALD, alcoholic liver disease; PBC, primary biliary cirrhosis; AIH, autoimmune hepatitis.

HCC-Associated Signature Shares Common Features With Other Solid Tumors. To evaluate whether the HCC-associated signature is common in other human tumors, we used SVM and applied the gene parameters from this signature to breast carcinoma (n = 38), lung carcinoma (n = 56), and HCC (n =103 Hong Kong patients: 89 HBV, four HCV, 10 others), from three publicly available microarray datasets.[26–28] These three datasets were chosen because they had a reasonable number of available matching genes for the signatures (Fig. 5). Additional HCC tissue samples from a totally different array platform were intended to be used as a further confirmation for the 273 gene set as the HCC-associated signature. Although the HCC-associated signature consistently performed well with an additional 83% of HCC tissue samples, it also matched 89% of the breast carcinoma cases and 96% of the lung carcinoma cases (Fig. 5). As a control, the etiology-associated signature did not provide a satisfying prediction of these samples (Fig. 5). Therefore, the HCC-associated signature contains genes that are commonly misregulated in HCC, breast carcinoma, and lung carcinoma.

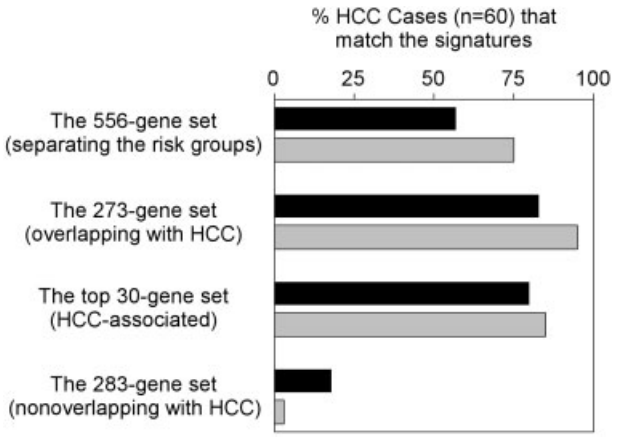

Fig. 3. Predicting independent HCC tissue samples with the HCC-associated signature. Sixty independent HCC tissue samples were used for testing various gene sets for their fitness to the relevance of HCC development. The simulated tests were based on a training of the high- and low-risk groups from 59 patients with CLD with the weights derived from the specified gene sets as indicated using the KNN (k = 3) or SVM predictor. Black bars, KNN predictor; gray bars, SVM predictor.

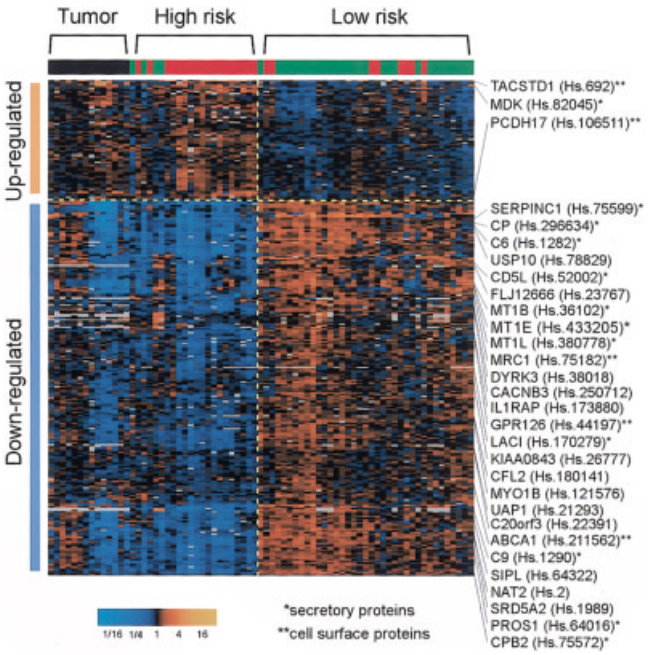

Fig. 4. The expanded view of 273 HCC-associated genes in tissue samples from 14 patients with HCC and from 59 patients with CLD defined by hierarchical clustering. Data are presented in the same format as essentially described in the Fig. 2 legend. The colored bars above the image plot represent sample categories (red, high-risk CLD; green, low-risk CLD; black, HCC). The top 30 biologically significant genes are indicated. See supplemental information for full data.

Discussion

Identification of patients at an early stage of cancer progression may provide a window of opportunity to intervene with an effective therapy. However, early diagnosis of patients with HCC has been hampered by the lack of reliable tumor markers for HCC development. In the current study, we used a cDNA microarray-based strategy to identify genes that are consistently altered both in cirrhotic tissue samples before the formation of a solitary tumor and in HCC. We hypothesized that many of these genes may serve as an early diagnostic marker for the onset of HCC. By analyzing gene expression profiles of several types of CLD liver tissue samples with various etiologies, we found a unique signature containing 273 genes that are abnormally expressed both in premaligant CLD and HCC tissue samples. This signature was further validated as a

Table 2. The Top 30 Most Significant Genes That can Discriminate the High-Risk Group from the Low-Risk Group

Gene Symbol*	UniGene Cluster	Description	Mean Ratios in the High-Risk Group	Mean Ratios in the Low-Risk Group	*P* Value
Cellular transporter					
MT1B	Hs.36102	Metallothionein 1B	0.5	1.0	.000001
MT1E	Hs.433205	Metallothionein 1E	0.5	1.1	<.000001
MT1L	Hs.380778	Metallothionein 1L	0.5	0.9	.000005
CP	Hs.296634	Ceruloplasmin	0.5	1.2	<.000001
ABCA1	Hs.211562	ATP-binding cassette 1	0.5	0.7	<.000001
CACNB3	Hs.250712	Calcium channel beta 3 subunit	0.6	1.2	<.000001
Blood coagulation					
PROS1	Hs.64016	Protein S	0.5	0.9	.000003
CPB2	Hs.75572	Carboxypeptidase B2	0.5	0.9	.000002
SERPINC1	Hs.75599	Antithrombin III	0.5	1.1	.000005
LACI	Hs.170279	Tissue factor inhibitor	0.7	1.0	.000003
Immune response					
MDK	Hs.82045	Midkine	2.1	1.5	.000259
C6	Hs.1282	Complement component 6	0.4	0.7	.000001
C9	Hs.1290	Complement component 9	0.3	0.8	<.000001
CD5L	Hs.52002	CD5 antigen-like	0.7	1.2	.000001
Signaling pathway					
GPR126	Hs.44197	G protein-coupled receptor 126	0.7	1.1	.000006
IL1RAP	Hs.173880	IL-1 receptor accessory protein	0.6	1.0	.000004
DYRK3	Hs.38018	Protein kinase, similar to sc. YAK1	0.6	1.1	<.000001
Cell adhesion					
PCDH17	Hs.106511	Protocadherin 17	1.4	1.0	<.000001
TACSTD1	Hs.692	EpCAM	3.6	1.7	.000016
Miscellaneous					
UAP1	Hs.21293	Sperm-associated antigen 2	0.7	1.2	<.000001
C20orf3	Hs.22391	Unknown	0.5	0.7	<.000001
CFL2	Hs.180141	Cofilin 2	0.6	0.9	<.000001
FLJ12666	Hs.23767	Hypothetical protein FLJ12666	0.6	1.0	.000006
KIAA0843	Hs.26777	KIAA0843 protein	0.7	0.9	<.000001
MRC1	Hs.75182	Mannose receptor, C type 1	0.7	1.4	.000002
NAT2	Hs.2	N-acetyltransferase 2	0.5	0.8	.000001
MYO1B	Hs.121576	Myosin IB	0.7	1.1	<.000001
SIPL	Hs.64322	SIPL protein	0.7	1.1	<.000001
SRD5A2	Hs.1989	Steroid-5-alpha-reductase 2	0.5	0.9	.000004
USP10	Hs.78829	Ubiquitin-specific protease 10	0.4	0.7	.000003

Abbreviations: ATP, adenosine triphosphate; IL-1, interleukin-1; ORF, open reading frame.

* Entries with HUGO-approved symbols.

cancer-associated gene set by applying the gene expression parameters to independent microarray datasets from two additional HCC cohorts (the 60 Shanghai patients with HCC and the 103 Hong Kong patients with HCC), a lung carcinoma cohort (n = 56), and a breast carcinoma cohort (n = 38). We found that the HCC-associated signature (273 genes) matches very well to both the Shanghai patients with HCC and the Hong Kong patients with HCC. The common signature shared by CLD and HCC indicate that they may serve as early markers in diagnosing the onset of a tumor. Consistently, some of the genes in the HCC-associated signature were known previously to be related to HCC. For example, the two most significantly up-regulated genes, that is, *TACSTD1* and *MDK*, were identified previously to be up-regulated in HCC and other solid tumors, suggesting that these proteins may play a role in HCC development.[29,30] In addition, five of the top 30 gene sets (*i.e.*, *C9*, *CD5L*, *CPB2*, *IL1RAP*, and *MT1B*), or 25 of the 273 gene set (*i.e.*, *ALDH8A1*, *ANXA7*, *C8A*, *C9*, *CD5L*, *CD163*, *CG018*, *CPB2*, *CRHBP*, *CYP2C9*, *ELF3*, *F11*, *FGB*, *FOXO1A*, *GHR*, *IL1RAP*, *MT1B*, *MTHFD1*, *NAT1*, *ORM1*, *PGRMC1*, *PLG*, *PCK2*, *QDPR*, and *TRIP13*), have been described as candidate HCC markers in several other microarray-based studies.[8,11,28] These provide a further indirect validation for the relevance of this signature to predict HCC. Therefore, the identification of such a unique HCC-associated molecular signature in premalignant CLD tissue samples may help in the future to classify patients with CLD at risk for developing HCC. However, it should be emphasized that we do not have sufficient follow-up data regarding the patients with CLD in our cohort to confirm that

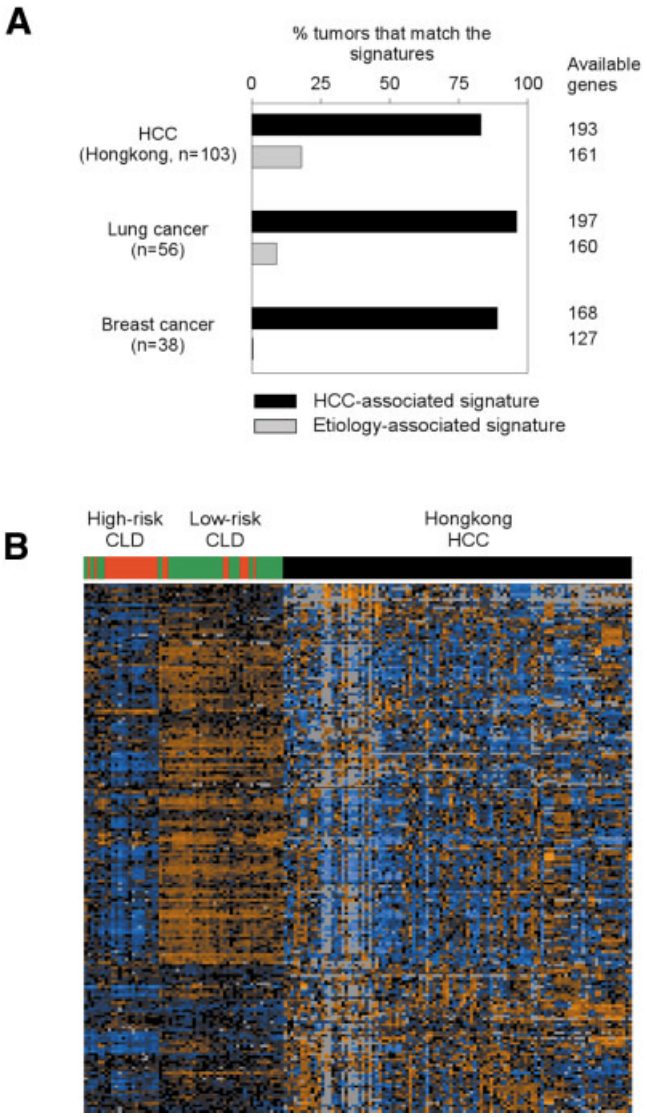

Fig. 5. The HCC-associated signature shares common features in human tumors. (**A**) The weights of the HCC-associated signature (273 genes) or the etiology-associated signature (283 genes) generated by the SVM predictor algorithm were applied to three test sample sets including breast carcinoma, lung carcinoma, and HCC from patients in Hong Kong to seek their matches. The simulated tests were based on a training set of 59 CLD tissue samples (high-risk and low-risk groups) with the defined gene sets using leave-one-out-cross-validated SVM at $P < .001$. (**B**) The hierarchical clustering results of the available 193 genes of the HCC-associated signature in training (CLD) and testing samples (HCC) are shown. Data are presented in the same format as essentially described in the Fig. 2 legend. The colored bars above the image plot represent sample categories (red, high-risk CLD; green, low-risk CLD; black, HCC). The expression ratios of this plot are on the same scale as Fig. 4.

they developed HCC. Therefore, this molecular signature will require a prospective study that includes a large number of patients with CLD to validate its clinical usage as a diagnostic tool.

In the current study, we found that a large number of HCC-associated genes were misregulated in tissue samples from patients with cirrhosis before any tumor development. It remains to be determined whether any of these genes contributed to the initiation of HCC. One argument against this possibility is that many of the expression changes that are common to CLD and HCC may be simply due to physical disruption of epithelial-mesenchymal signaling interactions that regulate epithelial cell function. Chronic liver damage and subsequent cirrhosis may lead to hyperproliferation, which may be sufficient to induce the disruption that leads to these large-scale gene expression changes. However, this is an unlikely consequence. First, all of our liver tissue samples were obtained from patients with end-stage CLD who had underlying cirrhosis. Therefore, they may have similar degrees of physical disruption. Second, the HCC-associated signature separates very well premalignant CLD in patients with HBV infection, HCV infection, HHC, and WD from CLD in patients with AIH, PBC, and ALD. Most of the genes in this signature are commonly misregulated in patients with HBV infection, HCV infection, HHC, and WD, with a much less degree of misregulation in patients with AIH, PBC, and ALD. This coincides with the clinical experiences of patients with cirrhosis and HBV infection, HCV infection, and HHC who have an extremely high risk of developing HCC whereas patients with cirrhosis and AIH, PBC, and ALD have a relatively low risk of developing HCC.[14] Moreover, although we identified a molecular signature consisting of greater than 500 genes in tissue samples that can determine the patients with CLD at risk of developing HCC, about one-half of the genes in the signature were also consistently misregulated in tissue samples obtained from patients with HCC. It is difficult to reconcile the finding that such a large number of genes occurring both in high-risk patients with CLD and in HCC are attributed to liver damage. It is likely that many of these genes may act as procarcinogenic genes that contribute to hepatocarcinogenesis. Our results are consistent with the hypothesis that many procarcinogenic genes may be activated/inactivated in cirrhosis before the formation of a solitary tumor. Consistently, silencing of *TACSTD1*, a lead gene in the HCC-associated signature, by a small interfering RNA approach resulted in growth inhibition of HCC cells (unpublished data). The model outlined in the current study demonstrates that HCC development is not a rare event in high-risk patients. This view is consistent with our observation that a high percentage of patients present with multiple aggressive HCC lesions at the time of diagnosis and that no single genetic change can be identified as a dominant event for HCC progression. Our results support the "field cancerization" model to explain multifocal tumors including HCC, head and neck carcinoma, skin carcinoma, breast carcinoma,

and lung carcinoma.[12,13,31] This model suggests that repeated exposure of a "preconditioned" organ to carcinogens leads to multiple and simultaneous changes in different areas of the organ, which eventually lead to multiple malignant primary tumors. Identification of these changes may help to develop a strategy in the future to prevent cancer development.

Although our gene expression profiling suggests that patients with WD are at a high risk of developing HCC, the cancer incidence is generally low.[14] The nature of this discrepancy is unclear. A similar procarcinogenic indicator such as p53 mutations has been detected in cirrhotic tissue samples from patients with HHC and WD,[32] presumably because these patients have a similar degree of procarcinogenic activity in the liver.[33] It is reasonable to speculate that patients with WD have the same incidence of HCC as patients with HHC. The reason for the reported low incidence of HCC may be due to the finding that these patients generally do not live beyond adolescence.

Among the top 30 biologically most significant genes in the classifier, 12 genes encode secretory proteins and five genes encode cell surface proteins (Fig. 4, Table 2). These findings provide feasibility for future diagnosis of patients with HCC using only the sera samples. Moreover, SERPINC1, LACI, PROS1, and CPB2 are involved in the clotting-fibrinolytic pathway and abnormal expression of these proteins has been linked to thromboembolism and cancer.[34] C6 and C9 are components of the complement system and complement deficiency may lead to immune system abnormality. In addition, four proteins (CP, MT1B, MT1E, and MT1L) are involved in iron transport and abnormal expression of these proteins leads to iron accumulation in hepatocytes, a condition associated with liver carcinoma. It is significant that many of these abnormally regulated genes are clustered into several known functional pathways. The most noticeable pathways include metal transport, blood coagulation, and immune response (Table 2). Identification of gene clusters that are functionally related indicates that these pathways may be significant in early-stage liver carcinogenesis.

A detailed analysis of the HCC-associated signature revealed that *TACSTD1* was a lead gene in this signature with an average of a 3.6-fold increase in the high-risk group (Table 2). It is known that TACSTD1 is a tumor antigen with an increased expression in many tumors with epithelial origin including major gastrointestinal tumors.[35,36] Mature hepatocytes do not express the *TACSTD1* gene. It may function as a cell adhesion molecule to promote tumor cell growth.[35,36] Our data suggest that *TACSTD1* may also be involved in hepatocarcinogenesis. Studies are under way to further characterize this biologically significant gene as well as others in the HCC-associated signature for their roles in HCC initiation.

Acknowledgment: The authors thank C.C. Harris, A. Robles, and S.S. Thorgeirsson for their invaluable comments; R. Simon for advice on bioinformatics; E. Asaki, D. Peterson, J. Powell, and members of the National Cancer Institute microarray team at the Advanced Technology Center for technical support; P. He for pathologic diagnosis; R. Markin, H. Sharp, and members of the LTPADS program for archived tissue samples; and D. Dudek and the National Cancer Institute-Center for Cancer Research Fellows editorial board for editorial assistance.

References

1. El-Serag HB, Mason AC. Rising incidence of hepatocellular carcinoma in the United States. N Engl J Med 1999;340:745–750.
2. Parkin DM, Pisani P, Ferlay J. Global cancer statistics. CA Cancer J Clin 1999;49:33–64.
3. Carr BI, Flickinger JC, Lotze MT. Hepatobiliary cancers: cancer of the liver. In: DeVita Jr VT, Hellman S, Rosenberg SA, eds. Cancer Principles and Practice of Oncology. Philadelphia: Lippincott-Raven, 1997:1087–1114.
4. Nakakura EK, Choti MA. Management of hepatocellular carcinoma. Oncology (Huntingt) 2000;14:1085–1098.
5. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
6. Kirimlioglu H, Dvorchick I, Ruppert K, Finkelstein S, Marsh JW, Iwatsuki S, Bonham A, et al. Hepatocellular carcinomas in native livers from patients treated with orthotopic liver transplantation: biologic and therapeutic implications. HEPATOLOGY 2001;34:502–510.
7. Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. Nat Genet 2002;31:339–346.
8. Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, Tsunoda 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;61:2129–2137.
9. Xu XR, Huang J, Xu ZG, Qian BZ, Zhu ZD, Yan Q, Cai T, 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 U S A 2001;98:15089–15094.
10. Shirota Y, Kaneko S, Honda M, Kawai HF, Kobayashi K. Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. HEPATOLOGY 2001;33:832–840.
11. Smith MW, Yue ZN, Geiss GK, Sadovnikova NY, Carter VS, Boix L, Lazaro CA, et al. Identification of novel tumor markers in hepatitis C virus-associated hepatocellular carcinoma. Cancer Res 2003;63:859–864.
12. Vauthey JN, Walsh GL, Vlastos G, Lauwers GY. Importance of field cancerisation in clinical oncology. Lancet Oncol 2000;1:15–16.
13. Raasch BA, Buettner PG. Multiple nonmelanoma skin cancer in an exposed Australian population. Int J Dermatol 2002;41:652–658.
14. Craig JR. Tumors of the liver. In: Zakim D, Boyer TD, eds. Hepatology: a Textbook of Liver Disease. Philadelphia: Saunders, 2003:1355–1370.
15. Tiribelli C, Melato M, Croce LS, Giarelli L, Okuda K, Ohnishi K. Prevalence of hepatocellular carcinoma and relation to cirrhosis: comparison of two different cities of the world—Trieste, Italy, and Chiba, Japan. HEPATOLOGY 1989;10:998–1002.
16. del Olmo JA, Serra MA, Rodriguez F, Escudero A, Gilabert S, Rodrigo JM. Incidence and risk factors for hepatocellular carcinoma in 967 patients with cirrhosis. J Cancer Res Clin Oncol 1998;124:560–564.

17. El Serag HB. Hepatocellular carcinoma: an epidemiologic view. J Clin Gastroenterol 2002;35:S72–S78.
18. Wang XW, Hussain SP, Huo TI, Wu CG, Forgues M, Hofseth LJ, Brechot C, et al. Molecular pathogenesis of human hepatocellular carcinoma. Toxicology 2002;181/182:43–47.
19. Ikeda K, Saitoh S, Koida I, Arase Y, Tsubota A, Chayama K, Kumada H, et al. A multivariate analysis of risk factors for hepatocellular carcinogenesis: a prospective observation of 795 patients with viral and alcoholic cirrhosis. HEPATOLOGY 1993;18:47–53.
20. Kaplan MM. Primary biliary cirrhosis. N Engl J Med 1996;335:1570–1580.
21. Park SZ, Nagorney DM, Czaja AJ. Hepatocellular carcinoma in autoimmune hepatitis. Dig Dis Sci 2000;45:1944–1948.
22. Riordan SM, Williams R. The Wilson's disease gene and phenotypic diversity. J Hepatol 2001;34:165–171.
23. Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9:417–423.
24. 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;16:906–914.
25. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–14868.
26. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, Van De RM, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–13789.
27. Perou CM, Sorlie T, Eisen MB, Van De RM, Jeffrey SS, Rees CA, Pollack JR, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–752.
28. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13:1929–1939.
29. Koide N, Hada H, Shinji T, Ujike K, Hirasaki S, Yumoto Y, Hanafusa T, et al. Expression of the midkine gene in human hepatocellular carcinomas. Hepatogastroenterology 1999;46:3189–3196.
30. Ruck P, Wichert G, Handgretinger R, Kaiserling E. Ep-CAM in malignant liver tumours. J Pathol 2000;191:102–103.
31. Slaughter DP, Southwick HW, Smejkel W. 'Field cancerization' in oral stratified squamous epithelium: clinical implications of multicentric origin. Cancer 1953;6:963–968.
32. Hussain SP, Raja K, Amstad PA, Sawyer M, Trudel LJ, Wogan GN, Hofseth LJ, et al. Increased p53 mutation load in nontumorous human liver of Wilson disease and hemochromatosis: oxyradical overload diseases. Proc Natl Acad Sci U S A 2000;97:12770–12775.
33. Hussain SP, Hofseth LJ, Harris CC. Radical causes of cancer. Nat Rev Cancer 2003;3:276–285.
34. Lip GY, Chin BS, Blann AD. Cancer and the prothrombotic state. Lancet Oncol 2002;3:27–34.
35. Balzar M, Winter MJ, de Boer CJ, Litvinov SV. The biology of the 17-1A antigen (Ep-CAM). J Mol Med 1999;77:699–712.
36. Litvinov SV, Balzar M, Winter MJ, Bakker HA, Briaire-de Bruijn IH, Prins F, Fleuren GJ, et al. Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. J Cell Biol 1997;139:1337–1348.