Available online at www.sciencedirect.com

SCIENCE DIRECT

Biochimica et Biophysica Acta 1739 (2004) 50–61

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

http://www.elsevier.com/locate/bba

Feature genes of hepatitis B virus-positive hepatocellular carcinoma, established by its molecular discrimination approach using prediction analysis of microarray

Bu-Yeo Kim[a,1], Je-Geun Lee[a,1], Sunhoo Park[b], Jae-Yeon Ahn[a], Yeun-Jin Ju[a], Jin-Haeng Chung[b], Chul Ju Han[a,c], Sook-Hyang Jeong[a,c], Young Il Yeom[e], Sangsoo Kim[e], Yong-Sung Lee[f], Chang-Min Kim[g], Eun-Mi Eom[h], Dong-Hee Lee[h], Kang-Yell Choi[i], Myung-Haing Cho[j], Kyung-Suk Suh[k], Dong-Wook Choi[a,d], Kee-Ho Lee[a,*]

[a]*Laboratory of Molecular Oncology, Korea Institute of Radiological and Medical Sciences, 215-4, Gongneung-Dong, Nowon-Ku, Seoul, 139-706, Korea*
[b]*Department of Pathology, Korea Institute of Radiological and Medical Sciences, Seoul, Korea*
[c]*Department of Internal Medicine, Korea Institute of Radiological and Medical Sciences, Seoul, Korea*
[d]*Department of Surgery, Korea Institute of Radiological and Medical Sciences, Seoul, Korea*
[e]*Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea*
[f]*Department of Biochemistry, Han Yang University School of Medicine, Seoul, Korea*
[g]*National Cancer Center, Gyeonggi-do, Korea*
[h]*Department of Life Sciences, Ewha Womans University, Seoul, Korea*
[i]*Department of Biotechnology, College of Engineering, Yonsei University, Seoul, Korea*
[j]*Laboratory of Toxicology, College of Veterinary Medicine and School of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea*
[k]*Department of Surgery, Seoul National University, College of Medicine, Seoul, Korea*

Received 20 April 2004; received in revised form 12 July 2004; accepted 20 July 2004
Available online 7 October 2004

Abstract

Recent introduction of a learning algorithm for cDNA microarray analysis has permitted to select feature set to accurately distinguish human cancers according to their pathological judgments. Here, we demonstrate that hepatitis B virus-positive hepatocellular carcinoma (HCC) could successfully be identified from non-tumor liver tissues by supervised learning analysis of gene expression profiling. Through learning and cross-validating HCC sample set, we could identify an optimized set of 44 genes to discriminate the status of HCC from non-tumor liver tissues. In an analysis of other blind-tested HCC sample sets, this feature set was found to be statistically significant, indicating the reproducibility of our molecular discrimination approach with the defined genes. One prominent finding was an asymmetrical distribution pattern of expression profiling in HCC, in which the number of down-regulated genes was greater than that of up-regulated genes. In conclusion, the present findings indicate that application of learning algorithm to HCC may establish a reliable feature set of genes to be useful for therapeutic target of HCC, and that the asymmetric expression pattern may emphasize the importance of suppressed genes in HCC.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Hepatocellular carcinoma; Microarray; Learning; Prediction

1. Introduction

Hepatocellular carcinoma (HCC) is the most fatal cancer worldwide, particularly in Asia and sub-Saharan Africa, accounting for an annual half million death [1]. Although the risk factor of HCC is chronic infection with hepatitis B virus (HBV) or hepatitis C virus (HCV) [1], their exact

Abbreviations: HCC, hepatocellular carcinoma; HBV, hepatitis B virus; HCV, hepatitis C virus; SAM, significance analysis of microarrays; PAM, prediction analysis of microarrays; RT, reverse transcription; PCR, polymerase chain reaction; FDR, false discovery rate

* Corresponding author. Tel.: +82 2 970 1312; fax: +82 2 977 0381.
E-mail address: khlee@kcch.re.kr (K.-H. Lee).

[1] These two authors equally contributed to this work.

0925-4439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbadis.2004.07.004

molecular mechanisms leading to HCC have not yet been clearly defined. Recent studies have revealed the involvement of altered p53 [2], h-catenin [3], AXIN [3,4], $p21^{WAF1/CIP1}$ [5] and p27 Kip1 [5,6] in conjunction with hepatitis virus infections. However, these individual genetic changes do not precisely reflect the heterogenous nature and clinical characteristics of HCC patients. The cellular and molecular diversities within individual HCC emphasize the importance of studying multiple genetic alterations.

Current cancer diagnosis and classification depend entirely on subjective judgments of tumor histology by pathologists. However, recent application of cDNA microarray technology has made it possible at molecular level to identify a large number of genes actively participating in the development of human cancers, including HCC, through simultaneous analysis of genome wide expression [7,8]. Although the most precise diagnosis of HCC is possible by pathological findings, the correct differentiation of HCC from non-tumor liver tissues by the differences between their expression profiles is important, because it can verify an accurate set of genes involved in HCC evolution, thereby identifying appropriate target for therapies and enhancing therapeutic efficacy. The majority of earlier studies to identify genes involved in human cancers, including HCC, exploited unsupervised clustering algorithms, which are very useful in revealing class discrimination with overall pattern of gene expression [7–9]. However, this unsupervised analysis is inadequate to extract an accurate set of genes differentially expressed from large expressional information of the given classes and to provide accurate class discrimination. Therefore, various supervised learning algorithms have been successfully introduced to classify and predict samples according to their characteristics and clinical outcomes such as recurrence and survival with the robust set of genes displaying minimum statistical error [10–15].

In an effort to obtain reliable information on HCC which is associated with chronic HBV infection, the present study was undertaken to investigate gene expression pattern of B-type HCC and their corresponding non-tumor liver tissues, using cDNA microarray together with supervised learning approaches. Notably, the accuracy of HCC discrimination with feature set, which was selected by training and cross-validating a HCC sample set, was evaluated with the application to independent HCC sample sets. The resultant expression profiles would provide an informed speculation of tumor suppression, drug detoxification and lipid signaling during the development of liver cancer, and offer appropriate therapeutic targets for treatment of HCC.

2. Materials and methods

2.1. Patients and tissue samples

Surgical specimens were obtained from 68 patients who underwent surgical treatment for HCC at Korea Cancer Center Hospital. Informed consent was obtained from patients for the use of their operated specimens and clinicopathological data for research purpose. All cases were HBV surface antigen-positive, but not HCV-positive. The clinicopathological features of each patient are presented in Supplementary Tables 1 and 2. All tissues were surgically resected, immediately snap frozen, and stored in liquid nitrogen. Both tumor and adjacent non-tumor liver tissues were collected. Total RNA was isolated from frozen tissues by phenol–chloroform extraction, according to the manufacturer's instructions (Trizol, Gibco/BRL), and quality of the total RNA was judged based on the ratio of 28S/18S RNA after agarose gel electrophoresis.

2.2. cDNA microarray and sequence verification

Human cDNA clones were obtained from Incyte Genomics (Fremont, CA), and their identities were verified by sequencing at our facility. A total sequence set of ~7400 cDNA (~5700 genes) was generated after quality control processes [16]. These sequence-verified clones were printed onto glass microscope slides. The probe preparation and hybridization were performed using NEN Micromax kit according to the manufacturer's protocol. Briefly, 100 Ag of each experimental HCC or non-tumor liver RNA and 100 Ag of common reference RNA obtained from a mixture of five different person's placentas were labeled with Cy5-dUTP and Cy3-dUTP, respectively, by incubation with reverse transcriptase for 2 h. The two probes were purified by passing through a Microcon 30 columns (Millipore, Bedford, MA) and hybridized in Express Hybk solution (BD Biosciences, Palo Alto, CA) to microarray overnight at 65 8C. The arrays were washed and scanned by ScanArray scanner (Perkin-Elmer, Boston, MA). Placental reference RNA was verified to detect majority (over 85%) of cDNA spots of microarray.

2.3. Data analysis

Primary data from image files were obtained using IMAGENE 4.0 (Bio-discovery, Marina del Rey, CA) and then normalized using intensity and spatially dependent method, as previously described [17]. The primary data can be obtained at the website of Human Genome Project database in Korea (http://chips.kribb.re.kr/~khleekcch). All array elements, for which the fluorescent intensity in each channel was greater than 1.4 times the local background, were considered well measured. Genes that were not well measured on at least 80% of the samples were excluded. The expressional ratios were hierarchically clustered by using CLUSTER and then visualized using TREEVIEW (M.B. Eisen, http://rana.lbl.gov) [18]. To filter genes that were differentially expressed in two sample sets, we used two kinds of analysis: significance analysis of microarrays (SAM) [19] and prediction analysis of microarrays (PAM) [12]. In SAM analysis, sample labels were randomly permutated

1000 times between HCC and non-tumor liver tissues to calculate the estimated percentage of genes selected by chance. The error estimation in PAM, which uses nearest shrunken centroids algorithm, was computed by adopting 10-fold cross-validation between HCC and non-tumor liver tissues. Briefly, by dividing HCC and non-tumor liver tissue samples into 10 nearly equal-sized parts, we established the model of HCC gene expression on 90% of the samples and then predicted the class labels of the remaining 10%. This procedure was repeated 10 times with each part playing the role as test samples, and the errors on all 10 parts were added together to compute the overall error rate. Then, we selected subset of genes that exhibited minimum error of cross-validation. The *P* value for the prediction of tissue samples (HCC vs. non-tumor liver tissues) was calculated, using the proportional chance probability as described [20].

2.4. Semi-quantitative reverse transcription-polymerase chain reaction (RT-PCR)

One microgram of total RNA isolated from HCC or corresponding non-tumor liver tissues was used for reverse transcription reaction. The number of PCR cycles was optimized to represent the product quantities within the linear range of amplification. The primer sets used in this study were as follows: diubiquitin, 5V-GACCATCCACCTTACCCTGA-3V (forward) and 5V-CTGCCATCATCTTCCCATCT-3V (reverse); hepatocyte nuclear factor 3 alpha, 5V-GACTCCAGCCTCCTCAACTG-3V (forward) and 5V-CGAGGAGGACATGAGGTTGT-3V (reverse); desmoglein 2, 5V-CCACCTGAAGACAAGGTGGT-3V (forward) and 5V-GCAGGCTTCTGTGTTCTTCC-3V (reverse); stromal cell derived factor 1, 5V-CTAGTCAAGTGCGTCCACGA-3V (forward) and 5V-GGACACACCACAGCACAAAC-3V (reverse); glutathione *S*-transferase A3, 5V-GATGCCAAGATTGCCTTGAT-3V (forward) and 5V-GCAGGTTGCTGATTCTGGTT-3V (reverse); $p21^{WAF1/CIP1}$, 5V-CGAAGTCAGTTCCTTGTGGAGC-3V (forward) and 5V-GCGGATTAGGGCTTCCTCTTGG-3V (reverse). PCR products were separated by electrophoresis on 1.5% agarose gel and visualized under UV light after ethidium bromide staining. h-Actin was used as an internal control to confirm equal amount of templates.

3. Results

3.1. Asymmetric gene expression pattern of HCC revealed by supervised clustering analysis

To verify genes abnormally expressed in HCC which is associated with HBV, we compared gene expression patterns of B-type HCC (*n*=33) and their corresponding non-tumor liver tissues (*n*=23) using cDNA microarrays containing ~7400 cDNA. The HCC samples were chosen to include mostly moderately differentiated tumors which account for Edmonson grade II and III (Supplementary Table 1). To obtain significant and consistent expression changes associated with HCC, we explored two different approaches with different supervised algorithm in the analysis of cDNA microarray data: SAM [19] and PAM [12]. A distinct feature of SAM is to identify the genes expressed abnormally in HCC, compared to non-tumor liver tissues, with statistically significant changes by assigning the percentage of genes identified by chance, the false discovery rate (FDR). This error rate is calculated by random permutations of the repeated measurements for each gene. In comparison, PAM allows us to obtain a subset of genes that best characterize HCC with minimum error rate in training sample set. This subset of HCC genes is further validated by predicting unknown test HCC sample set. The schematic illustration of analytical procedure is presented in Fig. 1. Initially, our effort was focused on SAM to investigate the overall pattern of genes differentially expressed among HCC (*n*=33) and non-tumor liver tissues (*n*=23). As shown in Fig. 2A, the most prominent feature was that, compared to non-tumor liver tissues, the number of genes induced or repressed in HCC was asymmetrically distributed, where more genes were down-regulated than up-regulated in HCC. The estimated number of significant genes at each stringency level with FDR is presented in Fig. 2A. Specifically, by applying delta value of 1.3, we could evaluate 4 up-regulated genes and 287 down-regulated genes with 0.23% FDR (Fig. 2B and Supplementary Fig. 1). The full list of these genes and supplementary data can also be found at the website (http://www.chips.kribb.re.kr/~khleekcch).

cDNA chip contents, experimental approach, and analytical method on cDNA chip technology might influence our observed asymmetric pattern of HCC gene expression profiling. Therefore, we analyzed two different HCC databases, which had been earlier reported, to find out whether they might also exhibit similar pattern to our finding. The initial application was carried out with HCC cDNA microarray database of Chen et al. [9] who performed an assay with cDNA chips containing ~17,000 genes, only 22.5% of which were found to overlap with ours. When this database was applied to the algorithm of SAM, it also displayed asymmetric expression patterns in which more genes were down-regulated than up-regulated in this HCC sample set, similar to our present result (Supplementary Fig. 2A and B). The intermediation of reference RNAs used in the above and our present HCC databases might also lead to an overall bias in interpreting expression ratios of HCC versus their corresponding non-tumor liver tissues during the normalization processes which adjust the variations of inter- and intra-chip expression intensities. To exclude such possibility, therefore, we analyzed another available HCC database [21] which was obtained using cDNA chips printed with Incyte clones which contained 88.7% of ours, but approached differently from our experimental procedures,

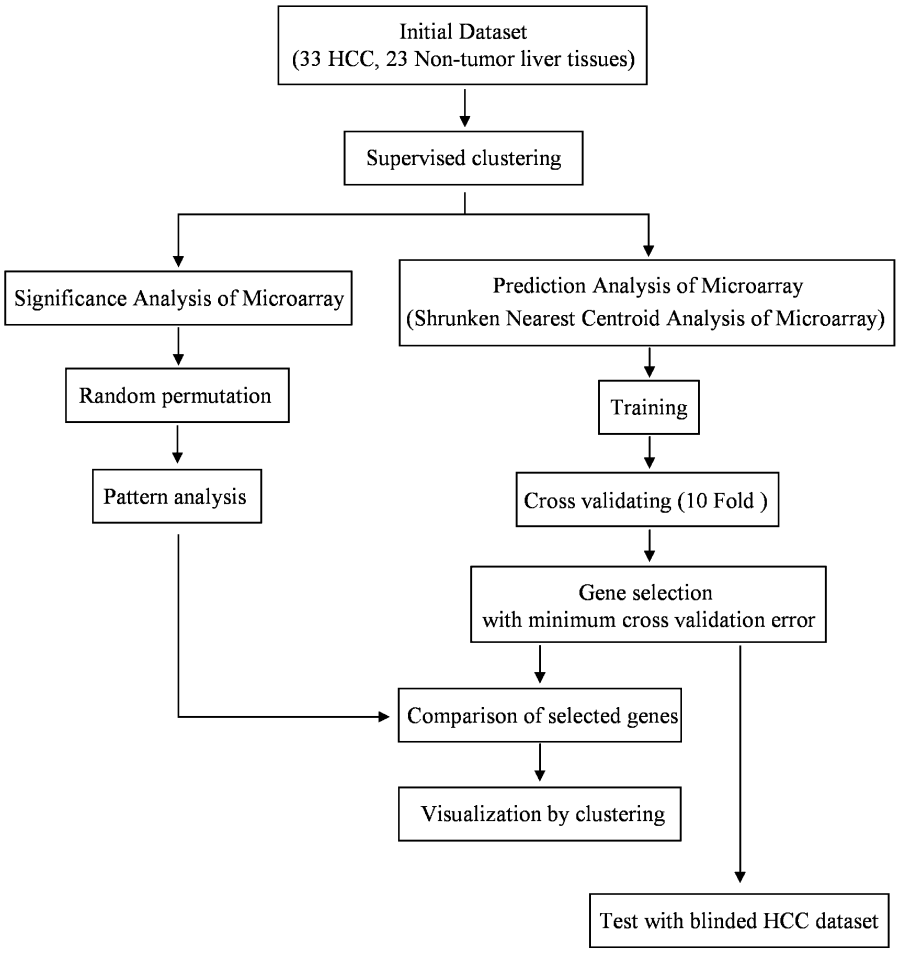

Fig. 1. Schematic illustration of the analysis strategy of HCC cDNA microarray. Initial data set, consisting of 33 HCC and 23 non-tumor liver tissues, was used to define a feature set of genes that can discriminate the status of tissues. The overall variations of gene expressions could be analyzed and visualized by applying SAM to this sample set. Based on nearest shrunken centroids algorithm, a feature set, showing minimum error of cross-validation, could be selected using PAM. Then, the efficacy of this feature set was tested in blinded HCC sample set with the evaluation of the error probability.

in that gene expression between HCC and their corresponding non-tumor liver tissues was directly compared in the same chip without reference RNA. In this set of HCC, it is also evident that more genes were down-regulated than up-regulated (Supplementary Table 3). Specifically, comparing genes changed commonly in more than 30% of HCC, we estimated that the numbers of genes overexpressed more than threefold and down-regulated less than one-third in HCC than their non-tumor liver tissues were 36 and 150, respectively, thus confirming our observation. In addition to these various sources of HCC gene expression pattern based on cDNA microarray, the asymmetric expression pattern was also confirmed in our serial analysis of gene expression (SAGE) of HCC (data not shown). The presently observed asymmetric expression patterns in our HCC sample set, together with those of the previously reported two different HCC databases, therefore, indicate that down-regulated genes on expression profile are more common in HCC than up-regulated genes, thus carrying out epigenetic silencing as seen in tumor suppressor gene [22].

To validate our results of cDNA-microarray, we first performed RT-PCR analysis of the genes obtained by SAM (Fig. 3), which showed relatively high or low expressional ratio in HCC samples compared to non-tumor liver tissues. The results of semi-quantitative RT-PCR of eight pair-matched HCC and non-tumor liver tissues for up-regulated genes (diubiquitin, hepatocyte nuclear factor 3 alpha, and desmoglein 2) or down-regulated genes (stromal cell derived factor 1, glutathione *S*-transferase A3 and $p21^{WAF1/CIP1}$) in HCC confirmed the expression data of cDNA microarray. Comparison of the expressional ratio of cDNA microarray with that of SAGE further confirms that our microarray analysis is credible enough to reproduce the result obtained by the conventional method (data not shown).

3.2. Prediction analysis for discrimination of HCC from non-tumor liver tissues

Class prediction in cDNA microarrays has recently received a great deal of attention, and various supervised

(A)

Parameter	Numbers called significant	Up	Down	FDR(%)
D = 0.8	555	34	521	4.0
D = 0.9	458	10	448	2.5
D = 1.0	398	6	392	1.5
D = 1.3	**291**	**4**	**287**	**0.23**
D = 1.5	232	2	230	0.28

(B)

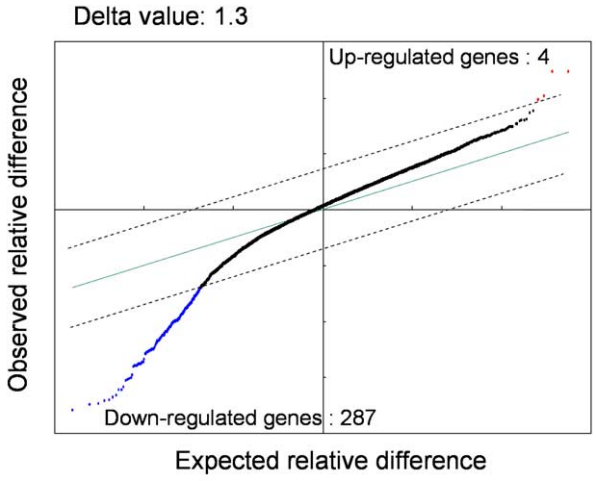

Fig. 2. Significance analysis of gene expression. (A) The number of significant genes was estimated at each stringency level. FDR was defined as the percentage of falsely significant genes compared to the genes called significant. (B) Scatter plot of the observed relative difference versus the expected relative difference was obtained at specific stringency level of 1.3. The up-regulated or down-regulated genes are indicated with red or blue colors, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

learning algorithms, including shrunken centroids [12], weighted voting [10] and *k*-nearest neighbor [13], have been introduced to predict the test samples. A simple supervised clustering such as SAM reveals information on a group of genes related to HCC and class discrimination with relatively low error ratio. However, to obtain an optimized set of genes to discriminate HCC from non-tumor liver tissues, the same sample set used for SAM was trained and cross-validated, using PAM algorithm as a training set (Supplementary Table 1). The error plot for different values of the shrinkage parameter is shown in Fig. 4A. Then, the best fit set of genes, consisting of 41 down-regulated and 3 up-regulated genes with minimum misclassification error (amount of shrinkage=3.32), was selected (Table 1). As presented in Supplementary Fig. 1, these selected genes were found to also rank high in the significant gene list extracted from SAM algorithm. In the expression profile shown in Fig. 4B, these genes distinguished tissues with only two misclassified samples (T43 and T10) (Pb0.0001). The cross-validation, using this feature set, resulted in the statistical probabilities of HCC and non-tumor liver tissues' status (Fig. 4C). Of the 56 samples, 50 (without mark on their circle symbols) could clearly be identified to be correct with over 80% high probabilities. Among the remaining six samples, three non-tumor liver tissue samples, marked with N14, N53 and N63, could be identified as ambiguously correct with less probability of 70.2%, 54.7%, and 72.6%, respectively, whereas two HCC samples, T10 and T43, were cross-validated diametrically opposite to their status and, one HCC sample, T35, was ambiguously misclassified with 38.5% probability (Fig. 4C). In spite of the misclassification of a few samples, it should be noted that the resultant overall error ratio of 0.053 (3 in 56 samples) was statistically

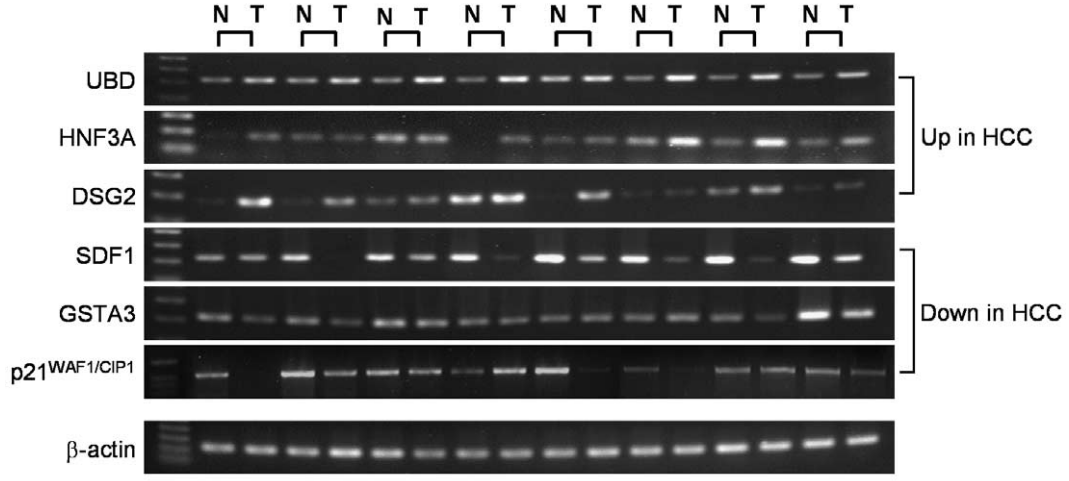

Fig. 3. Validation analysis of gene expression profiling. cDNA microarray results were confirmed by semi-quantitative RT-PCR in eight paired HCC (T) and corresponding non-tumor liver tissues (N) which showed multiple expression changes in the selected genes. The abbreviations used are: UBD, diubiquitin; HNF3A, hepatocyte nuclear factor 3 alpha; DSG2, desmoglein 2; SDF1, stromal cell derived factor 1; GSTA3, glutathione *S*-transferase A3.

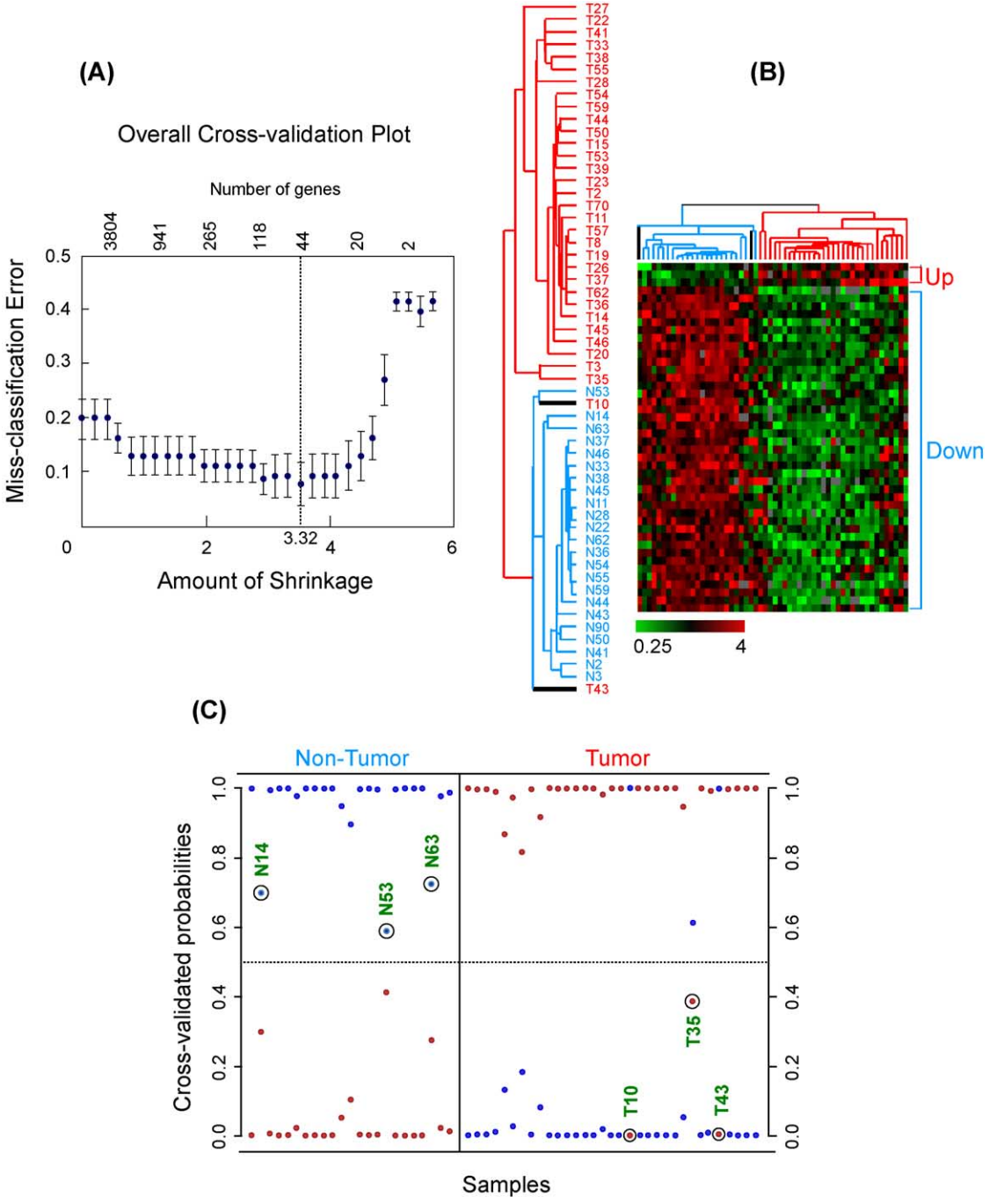

Fig. 4. Prediction analysis of gene expression variation. (A) Cross-validation error is shown as a function of the shrinkage parameter. The shrinkage value of 3.32 with minimum error probability was chosen for display in B and C. (B) Gene expression profile of 44 genes, chosen on the basis of cross-validation plot, was ordered by hierarchical clustering, where 3 genes were up-regulated and 41 genes were down-regulated in HCC. The tumor samples (T) or non-tumor liver samples (N) are colored as red or blue, respectively. Misclassified samples are highlighted with black colors. (C) Estimated probabilities for the sample data set using the selected 44 genes. All 56 samples were 10-fold cross-validated, and miss or ambiguously classified samples are marked with black circle.

significant in proportional chance statistical analysis (Pb0.0001). Similar prediction results were also reproduced, when weighted voting and k-nearest neighbor learning algorithms were applied to the same sample sets (data not shown), thus indicating that the selected feature set could be a good candidate to distinguish the status between HCC and non-tumor liver tissues.

To explore the possibility of whether the defined feature set could significantly predict other HCC sample set, we attempted to validate these selected genes in an independent HCC test sample set (Supplementary Table 2). As shown in Fig. 5, most test set samples displayed good separation with four misclassifications (four in non-tumor liver tissues and none in HCC tissues, Pb0.0001) out of 40 samples (21 non-