Short Communication

Identifying Genes for Establishing a Multigenic Test for Hepatocellular Carcinoma Surveillance in Hepatitis C Virus-Positive Cirrhotic Patients

Kellie J. Archer,1,2 Valeria R. Mas,3,4 Krystle David,3 Daniel G. Maluf,3 Karen Bornstein,3 and Robert A. Fisher3

1Department of Biostatistics, 2Massey Cancer Center, 3Division of Transplantation - Department of Surgery, and 4Department of Pathology, Virginia Commonwealth University, Richmond, Virginia

Abstract

In this study, we used the Affymetrix HG-U133A version 2.0 GeneChips to identify genes capable of distinguishing cirrhotic liver tissues with and without hepatocellular carcinoma by modeling the high-dimensional dataset using an L1 penalized logistic regression model, with error estimated using N-fold cross-validation. Genes identified by gene expression microarray included those that have important links to cancer development and progression, including VAMP2, DPP4, CALR, CACNA1C, and EGR1. In addition, the selected molecular markers in the multigenic gene expression classifier were subsequently validated using reverse transcriptase-real time PCR, and an independently acquired gene expression microarray dataset was downloaded from Gene Expression Omnibus. The multigenetic classifier derived herein did similarly or better than standard abdominal ultrasonography and serum α-fetoprotein, which are currently used for hepatocellular carcinoma surveillance. Because early hepatocellular carcinoma diagnosis increases survival by increasing access to therapeutic options, these molecular markers may prove useful for early diagnosis of hepatocellular carcinoma, especially if prospectively validated and translated into gene products that can be reproducibly and reliably tested noninvasively. (Cancer Epidemiol Biomarkers Prev 2009;18(11):2929–32)

Introduction

Surveillance for hepatocellular carcinoma includes following patients with chronic hepatitis or liver cirrhosis every 6 to 12 months (1) and monitoring them with abdominal ultrasonography, serum α-fetoprotein, and/or the protein induced by vitamin K absences (PIVKA-II; ref. 2). Abdominal ultrasonography has been described as highly user-dependent (3). Although serum α-fetoprotein determination is less costly compared with ultrasonography (4), it is a nonspecific marker for hepatocellular carcinoma, especially among hepatitis C virus (HCV) cirrhotic patients. In fact, in a series of 606 hepatocellular carcinoma patients, normal serum α-fetoprotein levels (<20 ng/mL) were observed in 40.4% of patients with small hepatocellular carcinoma (≤2 cm diameter), in 24.1% of patients with tumors 2 to 3 cm in diameter, and in 27.5% of patients with 3- to 5-cm tumors (5). Nevertheless, asymptomatic patients diagnosed with hepatocellular carcinoma in a screening program that included ultrasonography and serum α-fetoprotein monitoring had significantly smaller tumors, were able to undergo treatment in a significantly increased number, and had significantly longer survival compared with patients presenting with symptomatic hepatocellular carcinoma (4).

Due to the poor clinical outcomes of patients with hepatitis-C–induced cirrhosis who are diagnosed with advanced-stage hepatocellular carcinoma, improved markers for early detection are needed. Markers useful for early diagnosis may reduce time to transplantation and thereby yield improved patient outcomes. In this study, a multigenic classifier was derived using gene expression microarray data that is capable of detecting the presence of hepatocellular carcinoma in cirrhotic tissues, as cirrhotic tissues have been described as a premalignant condition (6). Thereafter, the selected molecular markers in the multigenic gene expression classifier were validated using reverse transcriptase-real time PCR (QPCR) reactions and an independently acquired gene expression microarray dataset.

Materials and Methods

Affymetrix HG-U133A 2.0 GeneChip Arrays were available for 16 cirrhotic tissues from patients with HCV plus hepatocellular carcinoma and 47 cirrhotic tissues from HCV-positive patients who did not have concomitant hepatocellular carcinoma. The study was approved by the Institutional Review Board at Virginia Commonwealth University, and informed consent was obtained from all patients. The sample preparation protocol followed the Affymetrix GeneChip Expression Analysis Manual. Total
RNA was extracted from tissue samples using TRIzol (Life Technologies). Integrity of RNA was checked using Agilent 2100 Bioanalyzer. Briefly, total RNA was reverse-transcribed using T7-polydT primer and converted into double-stranded cDNA (One-Cycle Target Labeling and Control Reagents, Affymetrix), with templates used for an in vitro transcription reaction to yield biotin-labeled antisense cRNA. The labeled cRNA was chemically fragmented and made into the hybridization cocktail according to the Affymetrix GeneChip protocol, which was then hybridized to U133A 2.0 GeneChips. The array image was generated by the high-resolution GeneChip Scanner 3000 by Affymetrix. The data are available from Gene Expression Omnibus.

Statistical Analysis. Patient demographics were examined for each group and continuous variables were compared using a two-sample t-test, whereas categorical variables were compared using Fisher’s exact test. For the gene expression microarray data, the robust multivariate average method was used to obtain probe set expression summaries (7). Thereafter, control probe sets were removed, leaving 22,215 probe sets for statistical analysis. Ordinarily when predicting a dichotomous class, such as HCV-positive cirrhosis with and without hepatocellular carcinoma, logistic regression is commonly used. However, traditional logistic regression models cannot be estimated when the number of predictor variables (p) exceeds the sample size (n). Even if the gene expression dataset were filtered using the False Discovery Rate method, the number of predictors for this dataset would still greatly exceed the sample size, with 4,579 probe sets significantly using a false discovery rate of 10% and 2,386 probe sets significant using a false discovery rate of 5%.

Penalized methods have been effectively used when modeling microarray data to identify important genes as well as gene groups associated with survival and dichotomous outcomes (8, 9). The least absolute shrinkage and selection operator (LASSO) is a penalized method for estimating a logistic regression model when \( p > n \) and when there is collinearity among the candidate predictors (10). The LASSO model is estimated using maximum likelihood with the additional constraint that the sum of the absolute values of the regression coefficients is less than some tuning parameter, \( \lambda \), which renders a sparse solution (10). The LASSO model was fit to predict class where the final model selected was that having the minimum Akaike Information Criterion (AIC) using the glmpath package in the R programming environment. The minimum AIC was selected over the minimum Bayesian Information Criterion (BIC) based on a large simulation study in which it was concluded that although the BIC tends to select the right-sized model, the AIC more often includes a nonzero coefficient estimate for the true predictor (11). The predicted class was cirrhosis with hepatocellular carcinoma if the fitted probability was \( \geq 0.50 \) and cirrhosis without hepatocellular carcinoma otherwise. To obtain an unbiased estimate of classification error, leave-one-out cross-validation, also referred to as \( N \)-fold cross-validation where \( N \) is the total sample size, was used. All analyses were conducted in the R programming environment (12) using appropriate Bioconductor packages (13). The LASSO models were fit using the glmpath package (14).

For diagnostic purposes, to obtain a more parsimonious model, all probe sets included in the \( N \) different LASSO models from the \( N \)-fold cross-validation procedure were then subjected to a best subsets logistic regression modeling procedure. Best subsets identifies the best fitting model for each model size \( p = 1,2,...,P \), where \( p \) reflects the number of covariates in the model, by an exhaustive search whereby all possible combinations of models are fit and the optimal model for each size \( p \) is selected. The leaps package in the R programming environment was used for the best subsets procedure. Again, error was estimated using \( N \)-fold cross-validation. In addition, an independent gene expression microarray dataset was used as an independent test set for assessing error.

QPCR was used to measure gene expression of \( CACNA1C, CALR, DPP4, EGR1, \) and \( VAMP2, \) with \( GAPDH \) profiled as the internal housekeeping gene. The dataset was first restricted to samples with HCV-cirrhosis \( (n = 26), \) HCV-EtOH cirrhosis \( (n = 14), \) and HCV-cirrhosis with concomitant hepatocellular carcinoma \( (n = 23), \) and the two former groups were combined to form one group representing HCV-induced cirrhosis without hepatocellular carcinoma \( (n = 40), \) for all QPCR analyses, the quantity used in the statistical analysis was the difference between the mean cycle threshold (CT) of the gene of interest and \( CT_{\text{GAPDH}} \), or \( H_{\text{CT(Gene)}} - H_{\text{CT(GAPDH)}} \). For each of the five gene profiles using QPCR, a two-sample t-test was done to compare the mean expression between the HCV-induced cirrhosis without hepatocellular carcinoma \( (n = 40) \) and HCV-cirrhosis with hepatocellular carcinoma \( (n = 23; \) ref. 15). Thereafter, a logistic regression model was derived using a backward elimination method whereby genes remained in the model provided \( P < 0.10 \).

Results

There were no significant differences between HCV-positive patients with and without hepatocellular carcinoma when examining patient demographic characteristics with respect to patient age, gender, race, albumin, and alanine aminotransferase, although the two groups differed with respect to international normalized ratio (INR) and total bilirubin (Supplementary Table S1).

The best fitting LASSO model included 14 probe sets (Supplementary Table S2). The resubstitution error associated with this best fitting model was 1.6%, or in other words, 98.4% of the samples were correctly classified. The resubstitution error is known to be downwardly biased, therefore \( N \)-fold cross-validation was used as an additional means to assess error. The \( N \)-fold cross-validation error was 9.5%, or the classifier was 90.5% accurate. Using \( N \)-fold cross-validation (CV), five HCV plus hepatocellular carcinoma and one HCV plus cirrhosis without concomitant hepatocellular carcinoma cases were misclassified, rendering \( N \)-fold estimates for sensitivity, specificity, positive predictive value, and negative predictive value of 68.8%, 97.9%, 91.7%, and 90.2% respectively.

From the \( N \)-fold CV procedure there were 63 different LASSO models, each of which included different probe sets having nonzero coefficient estimates. In fact, considering all 63 different models, there were 62 unique probe sets included. Best subsets logistic regression was done using the 62 probe sets yielding the sequence of models listed in Supplementary Table S3. Due to collinearities,
the best fitting logistic regression model contained two genes, DPP4 and CALR. N-fold CV on this two-gene model resulted in a 6.3% error rate, corresponding to 93.7% accuracy, with two HCV plus hepatocellular carcinoma and two HCV-positive cirrhotic cases misclassified. One HCV plus hepatocellular carcinoma misclassified sample was from a patient with one 2.6-cm T2aN0M0 tumor, whereas the other HCV plus hepatocellular carcinoma misclassified sample was from a patient with four T2aN0M0 tumors of sizes 0.9, 1.0, 1.4, and 1.6 cm. This yielded N-fold CV estimates for sensitivity, specificity, positive predictive value, and negative predictive value of 87.5%, 95.7%, 87.5%, and 95.7%, respectively. A scatter-plot of CALR against DPP4 shows that the two groups are well separated by a straight line (Fig. 1).

To test the molecular classifier using an independent set of samples, data from a previous gene expression study that included 13 HCV cirrhotic nontumor liver tissues and 33 HCV plus hepatocellular carcinoma liver samples hybridized to an Affymetrix HG-U133 Plus 2.0 GeneChip were downloaded from Gene Expression Omnibus (16). All probe sets on the HG-U133A 2.0 array were also on the HG-U133 Plus 2.0 array. Among the 33 hepatocellular carcinoma tissues were 7 very early hepatocellular carcinoma, 9 early hepatocellular carcinoma, 7 advanced hepatocellular carcinoma, and 10 very advanced hepatocellular carcinoma. To mitigate any effect due to GeneChip type and center, prior to obtaining probe set expression summaries, the 63 HG-U133A 2.0 GeneChips and the 46 HG-U133 Plus 2.0 GeneChips were independently read into the R programming environment using the affy Bioconductor package (17). Thereafter, probe level data from the two GeneChip types were merged by probe sequence using mashprobes package in R (18). Subsequently, the robust multiarray average method was used to obtain probe set expression summaries (7).

The logistic regression model including DPP4 and CALR was fit to the Virginia Commonwealth University’s acquired samples and was applied to the independent set described by Wurmbach et al. (16). The sensitivity for detecting hepatocellular carcinoma in the independent test set was 72.7%, with 77.4% positive predictive value, although the specificity was only 46.2%. Among the nine hepatocellular carcinoma tissues misclassified, three were very early hepatocellular carcinoma, two were early hepatocellular carcinoma, three were advanced hepatocellular carcinoma, and one was very advanced hepatocellular carcinoma. Unfortunately, there were 38 unique HCV-infected patients included in the previously published study, but the number and sample labels of nontumor cirrhotic tissues procured from patients with concomitant hepatocellular carcinoma could not be identified. Because our model was derived using cirrhotic tissues with and without concomitant hepatocellular carcinoma, the low specificity observed in this independent test set may indicate that many samples were from patients with concomitant hepatocellular carcinoma. We do note that other biomarkers used for cancer screening have also reported low specificities, including prostate-specific antigen which is used to screen subjects for prostate cancer (19).

Genes identified by the best fitting LASSO model and those derived in the cross-validation process had important links to cancer development and progression, including VAMP2 (20), DPP4 (21), CALR (22), CACNA1C (23), and EGR1 (24). These genes were then profiled using QPCR; the interrogated sequence for Taqman overlapped the Affymetrix probe set target sequence for all five genes. P values from the two-sample t-test applied to each gene for comparing the mean expression between the HCV-induced cirrhosis without hepatocellular carcinoma (n = 40) and HCV-cirrhosis with hepatocellular carcinoma (n = 23) were $P = 0.007$ (CACNA1C), $P = 0.09$ (CALR), $P = 0.98$ (DPP4), $P = 0.0003$ (EGR1), and $P = 0.009$ (VAMP2). The final logistic regression model included DPP4, VAMP2, and EGR1; the odds ratio corresponding to a one-cycle change in the CT difference and 95% confidence intervals as well as the P values for the genes in the final model are presented in Table 1. Thus, although DPP4 was not significant univariately, when collectively considered with other genes it was important for controlling confounding (25). The sensitivity, specificity, positive predictive value, and negative predictive value for the final model using a cut point of 0.50 on the predicted probabilities are 78.3%, 87.5%, 78.3%, and 87.5%, respectively. The area under the receiver operator characteristic curve was 87.6 (Supplementary Fig. S1). To visualize how these three genes vary together by group, a conditioning plot (Fig. 2) was constructed whereby VAMP2 by EGR1 is plotted in each

Table 1. Multivariable logistic regression model predicting HCV-cirrhosis with and without concomitant hepatocellular carcinoma using reverse transcriptase-real time PCR (QPCR) data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Odds ratio (95% confidence interval)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPP4</td>
<td>3.52 (1.58, 7.85)</td>
<td>0.0007</td>
</tr>
<tr>
<td>EGR1</td>
<td>0.23 (0.10, 0.53)</td>
<td>0.0072</td>
</tr>
<tr>
<td>VAMP2</td>
<td>2.38 (1.00, 5.70)</td>
<td>0.051</td>
</tr>
</tbody>
</table>
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Acknowledgments

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

Identifying Genes for Establishing a Multigenic Test for Hepatocellular Carcinoma Surveillance in Hepatitis C Virus-Positive Cirrhotic Patients


Updated version  Access the most recent version of this article at: doi:10.1158/1055-9965.EPI-09-0767
Supplementary Material  Access the most recent supplemental material at: http://cebp.aacrjournals.org/content/suppl/2009/10/27/1055-9965.EPI-09-0767.DC1

Cited articles  This article cites 25 articles, 2 of which you can access for free at: http://cebp.aacrjournals.org/content/18/11/2929.full#ref-list-1
Citing articles  This article has been cited by 2 HighWire-hosted articles. Access the articles at: http://cebp.aacrjournals.org/content/18/11/2929.full#related-urls

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions  To request permission to re-use all or part of this article, use this link http://cebp.aacrjournals.org/content/18/11/2929. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.