Network analysis of HBV- and HCV-induced hepatocellular carcinoma based on Random Forest and Monte Carlo cross-validation

SHAN-NA ZHAO¹*, LING-LING LIU²*, ZHI-PING LV¹, XIAO-HUA WANG¹ and CHENG-HONG WANG¹

¹Department of Clinical Laboratory, Yantaishan Hospital, Yantai, Shandong 264008; ²Department of Clinical Laboratory, Shandong Provincial Hospital Affiliated Shandong University, Jinan, Shandong 250021, P.R. China

Received May 5, 2016; Accepted March 28, 2017

DOI: 10.3892/mmr.2017.6861

Abstract. Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-associated mortality worldwide. Hepatitis B virus (HBV) and hepatitis C virus (HCV) are two common risk factors for HCC. The majority of patients with HCC present at an advanced stage and are refractory to therapy. It is important to identify a method for efficient diagnosis at early stage. In the present study gene expression profile data, generated from microarray data, were pretreated according to the annotation files. The genes were mapped to pathways of Ingenuity Pathways Analysis. Dysregulated pathways and dysregulated pathway pairs were identified and constructed into individual networks, and a main network was constructed from individual networks with several edges. Random Forest (RF) classification was introduced to calculate the area under the curve (AUC) value of this network. Subsequently, 50 runs of Monte Carlo cross-validation were used to screen the optimal main network. The results indicated that a total of 4,929 genes were identified in the pathways and gene expression profile. By combining dysregulated pathways with Z<0.05 and dysregulated pathway pairs with Z<0.2, individual networks were constructed. The optimal main network with the highest AUC value was identified. In the HCV group, the network was identified with an AUC value of 0.98, including 41 pairs of pathways, and in the HBV group, the network was identified with an AUC value of 0.94, including eight pairs of pathways. In addition, four pairs were identified in both groups. In conclusion, the optimal networks of HCV and HBV groups were identified with the highest AUC values. The use of these networks is expected to assist in diagnosing patients effectively at an early stage.

Introduction

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-associated mortality worldwide, particularly in men, being the second most frequent cause of cancer-associated mortality (1). Viral hepatitis, including hepatitis B virus (HBV) and hepatitis C virus (HCV), is a common risk factor for HCC (2,3). The majority of patients with HCC present at an advanced stage, and are refractory to chemotherapy and radiotherapy (4,5). Understanding the mechanism of carcinogenesis is pivotal for the diagnosis of HCC (6). It has been shown that using screening procedures for diagnosing HCC results in improved survival rates for patients, compared with those with symptomatic disease at presentation (7).

Several biomarkers have been investigated for the diagnosis of HCC, including clinical, radiological and laboratory modalities, with or without liver biopsy. The majority of patients with HCC are usually asymptomatic until the late stages. Generally, it is difficult to make a clinical diagnosis of HCC at an early stage due to the deep position of the liver within the body. The most frequently used HCC marker for diagnosis is the serum concentration level of α fetoprotein (AFP), the level of which is increased in patients with HCC (8). However, inconsistent sensitivity and specificity have been reported in AFP-L3 in previous studies (9), therefore, it may not be a valuable prognostic biomarker in patients with known HCC.

There has been significant progress in understanding the mechanism of HCC based on microarray technology, including differentially expressed genes, which can overcome the limitations of clinical prognostic factors (10). Several growth factors, including epidermal growth factor receptor, hepatoma-derived growth factor and insulin-like growth factor, have been reported to be involved in the progression of HCC (11,12). In addition, apoptotic genes, anti-apoptotic genes and tumor suppressor genes in have been identified in hepatocarcinogenesis (13,14). However, results have not been uniform and share only a limited number of potential genes.

To overcome the limitations of a gene-based approach, pathway analysis was introduced, which provides biological
information to facilitate characterization of the functional network and the associations between selected significant genes (15). Until now, a number of methods have been suggested to identify core pathways for diagnosis, including individual pathway aberrance score (iPAS) analysis (16) and principal component analysis (17). Although HBV- and HCV-infected HCC are not distinguishable in histological and clinical evaluations, microarray analyses have shown that different molecular mechanisms underlie the development of HBV- and HCV-positive HCC (15,17,18).

In the present study, pathway-based iPAS analysis with Random Forest (RF) classification and Monte Carlo cross-validation were used to identify pathway-based networks of HBV- and HCV-positive HCC. The disease datasets of HBV and HCV were compared with healthy data, generating different networks following 50 runs of Monte Carlo cross-validation. This method aimed to provide clinical molecular insights into the mechanism of HBV- and HCV-positive HCC.

Materials and methods

Gene expression data. Microarray gene expression data of E-MTAB-950 and its annotation files were downloaded from the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) (19). This included 42 normal samples, 149 HCV samples and 8 HBV samples. All software and hardware analyses were provided by Honghui Biotech Co., Ltd. (Jinan, China). The gene expression profile data was generated from the A-AFFY-44-Affymetrix GeneChip Human Genome U133 Plus 2.0 platform (Affymetrix, Inc., Santa Clara, CA, USA), and the platform title was ‘Transcription profiling by array of human normal liver, HBV, HCV liver samples’. According to the platform annotation files, the probes were mapped to gene symbols. If more than one probe was mapped to a single gene, the average level of the probes was used as the final gene expression value. In total, 20,545 genes were obtained, and the gene expression levels were normalized with quantile normalization using the preprocess Core package version 1.36.0 (https://github.com/bmbolstad/preprocessCore) (20).

Pathway data. Ingenuity Pathways Analysis (IPA) is a pathway database (http://www.ingenuity.com/), and pathways were downloaded from this database in the present study. In total, 589 biological pathways were obtained, including 5,169 genes. The genes in the expression profile were mapped to the IPA pathways, from which 4,929 genes were obtained. With gene expression values in each pathway, Fisher’s exact test was applied to evaluate the enrichment. P-values of the pathways were corrected by the false discovery rate of Benjamini-Hochberg (21).

Gene level statistics. The mean and standard deviation of gene expression levels in the normal group were calculated following normalization (20). For the disease group, the genes in the pathways were normalized using the quantile normalization method on combining a single disease case with all normal samples. The gene level statistics of each gene in a disease case were standardized as the mean ± standard deviation.

Pathway level statistics. For each pathway, the gene level statistics of all the genes in the pathway were extracted. The mean gene level statistic was considered the pathway level statistic. The following formula was used, in which n represents the number of genes belonging to the pathway:

\[ iPAS = \frac{\sum z_i}{n} \]

Assessment of significance. To assess the dysregulated pathways associated with the lesioned liver, significance was assessed. The mean ± standard deviation of the pathway level statistics in the normal samples was calculated. The significance was determined to assess the pathway level statistics in the disease group, with the normal group as a reference. The Z value of a specific pathway in each sample was obtained. Z<0.05 was considered to indicate a dysregulated pathway.

Discriminating score (DS) of pathway pairs. To evaluate the associations between pathway pairs, aDS (22,23) was introduced. The mean ± standard deviation of the expression level of genes belonging to a pathway in each sample was calculated. A pathway was randomly combined with another pathway, generating a pathway pair.

Significance assessment of pathway pairs. For each pathway pair in the normal sample, the mean ± standard deviation of the DS values was calculated. Significance was determined to assess differential pathway pairs following combining a disease sample with all normal samples, generating a Z value of each pathway pair in an individual sample. Z<0.2 was considered to indicate a differential pathway pair.

Network construction. Individual networks were constructed by combining dysregulated pathways and dysregulated pathway pairs in individual samples. The main network of HCV was constructed from individual networks in which edges appeared >5 times. The main network of HBV was constructed from individual networks, the edges of which appeared >1.

RF classification. In machine learning, RF is an important classifier, which contains a multi-decision-making tree. RF is an extension of a classification and regression tree, which builds a class prediction model using class-labeled input samples (24), and calculates a ranking of input variables ordered by the extent of association with classification (25). In the present study, the RF algorithm was introduced to evaluate the performance of the DS value in the main network. In this classification, two parameters were used: Number of variables randomly sampled as candidates at each split (mtry)= and number of trees grown (ntree)=500. The DS values of the pathway pairs in the main network were randomly divided into a training set and test set, followed by RF classification. The area under the curve (AUC) was estimated using a 10-fold cross-validation method.

Monte Carlo cross-validation. To obtain the optimal network, Monte Carlo cross-validation (26) was performed using the expression profile data. Data were randomly divided into the training set and test set. The validation was processed from
procedure 2.2 to 2.5 and repeated 50 times. For each validation, the differential pathways and differential pathway pairs were identified, which were constructed into a new individual network and a main network, in addition to the AUC. When all 50 runs were completed, the main networks were ranked by the values of the AUC. The network with the highest AUC value was considered the optimal network. The optimal networks of HBV and HCV underwent contrastive analysis, and an intersection network was obtained.

Results

Pathway enrichment analysis. There were 20,545 genes in the gene expression profile. When the genes were mapped to 589 IPA pathways, a total of 4,929 genes were identified in the pathways and gene expression profile. By ranking pathways with gene numbers, the top 4 pathways with the highest numbers of genes were determined, as listed in Table I.

Network analysis. By combining the dysregulated pathways with Z<0.05 and dysregulated pathway pairs with Z>0.2, individual networks were constructed. The occurrence number of pathway pairs in the individual networks was recorded in 50 runs of Monte Carlo cross-validation. In the HCV group, 15 pathway pairs occurred >40 times with 5 pathway pairs occurring 50 times, as shown in Fig. 1. In the HBV group, 8 pathway pairs occurred >40 times, with 1 pair occurring 50 times, as shown in Fig. 2.

A main network of HCV was constructed from the individual networks whose edges appeared >5 times, and a main network of HBV was constructed from individual networks whose edges appeared >1. Following 50 runs of Monte Carlo cross-validation, 50 networks were constructed. By ranking the networks by their AUC values, the highest was considered the optimum main network.

In the HCV group, the optimal network contained 41 pairs of pathways, as shown in Fig. 3, and the AUC value was 0.98. In the HBV group, the optimal network contained 8 pairs of pathways, as shown in Fig. 4, and the AUC value was 0.94. Following comparative analysis, 4 pathway pairs were identified in the two groups, which are listed in Table II.

Discussion

Microarray analyses have indicated that different molecular mechanisms underlie the development of HBV- and HCV-positive HCC. To identify biomarkers in HBV- and HCV-positive HCC in the present study, a pathway-based approach was applied by combining iPAS analysis and Monte Carlo cross-validation.

The data from the present study indicated a similar mechanism in HCV and HBV. In the results of the Monte Carlo cross-validation, among pathway pairs, which occurred 50 times, the pair ‘acetone degradation I’ and ‘Bupropion degradation’ was identified in both the HCV and HBV group. In the optimal main network, ranked by the AUC values, 4 pathway pairs were present in the two groups.

There were more differences between HBV- and HCV-positive HCC at the pathway level. In the heatmap of pathway pairs, 4 more pairs were identified with an occurrence of 50 times in the HCV group. In the optimal main networks, with the exception of the 4 identical pathway pairs, 38 additional pathway pairs were identified in the HCV group and 4 additional pathway pairs were identified in the HBV group.

In the HCV group, compared with the normal control, HCV induced dysregulated biological pathways, including DNA damage response, hormone degradation, carbohydrate degradation, cell cycle, cholesterol biosynthesis, amino acid degradation and biosynthesis and signaling pathways. The main network with the highest AUC was considered optimal for diagnosing HCV-positive HCC. In the main network, the pathways connected with several other pathways were considered important. It is generally accepted that cholesterol biosynthesis is important in the production of HCV, and the targeting enzyme has been suggested as a potential antiviral strategy against HCV (27). The pathway of adenine and adenosine salvage VI was connected to 5 other pathways. A cluster of 10 adenines encodes a core protein from HCV, which is considered to direct programmed ribosomal frame shifting (28). In 2008, the involvement of transcriptional slip-page in this recoding event was first demonstrated (29). The pathway of Melatonin degradation I was linked to 4 pathways. It has been demonstrated that melatonin has proapoptotic and antiangiogenic properties in HepG2 liver tumor cells (30), through a molecular mechanism involving the upregulation of TIMP metalloproteinase inhibitor 1, and attenuation of the expression and activity of matrix metalloproteinase-9 via nuclear factor-κB signal pathway inhibition. Therefore, the present study hypothesized that, in patients with HCV-positive HCC, its degradation-related pathway is downregulated. The pathway of bupropion degradation was also found to link with 4 other pathways. Bupropion is a commonly used antidepressant. Previously it was found to exert marked anti-inflammatory effects via downregulating tumor necrosis factor-α, interleukin-1β and interferon (IFN)-γ (31). It was demonstrated that bupropion was effective in treating IFN-α-induced depressive and somatic symptoms in patients with HCV infection (32).

In the present study, the optimal main network was screened in the HBV group, which included 8 pathway pairs, of which 4 pairs were also present in the HCV group. In the remaining pathway pairs, the lipoate salvage and modification pathway had a higher number of edges, compared with the other pathways. It is generally accepted that α-lipoic acid has antioxidant and redox-regulatory properties. Due to these characteristics, it has been used for the treatment of liver disease (33). In addition, it has been shown to suppresses the proliferation of different types of tumor cell through facilitating apoptosis in breast cancer (34). In the present study, the lipoate-related pathway was closely connected with HBV, which provided novel insight into the therapy of HBV-positive HCC. In the networks, certain pathways have been investigated in previous studies, whereas others have received minimal attention and require further investigations for verification. In addition, with the exception of HBV and HCV diseases, network analysis for other liver associated diseases, including hemochromatosis on a background of cirrhosis and HCC liver samples, require clarification in future investigations.

In conclusion, the present study combined iPAS analysis with the RF classification and Monte Carlo cross-validation to
identify pathway-based networks of HBV- and HCV-positive HCC. The optimal main networks of the HCV and HBV groups were identified with the highest AUC values. With these networks, it may be possible to diagnose patients effectively at an early stage.

Table 1. Ingenuity pathways analysis pathways overlapped with genes. Common genes indicated the number of genes belonging to the pathway and gene expression profile.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Genes in pathway (n)</th>
<th>P-value</th>
<th>FDR</th>
<th>Common genes (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axonal guidance signaling</td>
<td>421</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>408</td>
</tr>
<tr>
<td>Protein kinase a signaling</td>
<td>365</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>354</td>
</tr>
<tr>
<td>Molecular mechanisms of cancer</td>
<td>335</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>331</td>
</tr>
<tr>
<td>G-protein coupled receptor signaling</td>
<td>254</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>248</td>
</tr>
</tbody>
</table>

Common genes indicate the number of genes belonging to the pathway and gene expression profile. FDR, false discovery rate.
**Table II. Common pathway pairs in the main networks of HBC and HCV.**

<table>
<thead>
<tr>
<th>Pathway A</th>
<th>Pathway B</th>
<th>Weight (HCV)</th>
<th>Weight (HBV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salvage pathways of pyrimidine ribonucleotides</td>
<td>Pyridoxal 5'-phosphate salvage pathway</td>
<td>0.203549</td>
<td>0.19776</td>
</tr>
<tr>
<td>ATM signaling</td>
<td>Cyclins and cell cycle regulation</td>
<td>0.205993</td>
<td>0.205265</td>
</tr>
<tr>
<td>Alanine degradation III</td>
<td>Alanine biosynthesis II</td>
<td>0.189711</td>
<td>0.191768</td>
</tr>
<tr>
<td>4-hydroxybenzoate biosynthesis</td>
<td>4-hydroxyphenylpyruvate biosynthesis</td>
<td>0.189711</td>
<td>0.191768</td>
</tr>
</tbody>
</table>

Weight score was calculated by the Z score method with units of 1. HBV, hepatitis B virus; HCV, hepatitis C virus.

---

**Figure 3.** In the hepatitis C virus group, the optimal main network contained 41 pairs of pathways. Lines represent the pathway pairs. Blue nodes represent all the pathways in the validation procedure. The yellow nodes represent common pathway pairs in the main networks of hepatitis B and hepatitis C virus.

**Figure 4.** In the hepatitis B virus group, the optimal main network contained eight pairs of pathways. Lines represent the pathway pairs. Blue nodes represent all the pathways in the validation procedure. The yellow nodes represent common pathway pairs in the main networks of hepatitis B and hepatitis C virus.
et al. BMC Genomics 17: 14.