Revealing gene clusters associated with the development of cholangiocarcinoma, based on a time series analysis

JIANYU WU¹, ZHIFU XIAO¹, XIULEI ZHAO² and XIANGSONG WU¹

¹Department of General Surgery, First Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang 150001; ²Department of General Surgery, The Central Hospital of Cangzhou, Cangzhou, Hebei 061000, P.R. China

Received December 17, 2013; Accepted June 26, 2014

DOI: 10.3892/mmr.2015.3216

Abstract. Cholangiocarcinoma (CC) is a rapidly lethal malignancy and currently is considered to be incurable. Biomarkers related to the development of CC remain unclear. The present study aimed to identify differentially expressed genes (DEGs) between normal tissue and intrahepatic CC, as well as specific gene expression patterns that changed together with the development of CC. By using a two-way analysis of variance test, the biomarkers that could distinguish between normal tissue and intrahepatic CC dissected from different days were identified. A k-means cluster method was used to identify gene clusters associated with the development of CC according to their changing expression pattern. Functional enrichment analysis was used to infer the function of each of the gene sets. A time series analysis was constructed to reveal gene signatures that were associated with the development of CC based on gene expression profile changes. Genes related to CC were shown to be involved in ‘mitochondrion’ and ‘focal adhesion’. Three interesting gene groups were identified by the k-means cluster method. Gene clusters with a unique expression pattern are related with the development of CC. The data of this study will facilitate novel discoveries regarding the genetic study of CC by further work.

Introduction

Cholangiocarcinoma (CC) is an adenocarcinoma that forms glands or secretes significant amounts of mucins. These are composed of mutated epithelial cells that originate in the bile ducts, which drain bile from the liver into the small intestine (1-3). CC is currently considered to be an incurable and rapidly lethal malignancy. Currently, surgery has been shown to be the only existing curative treatment for CC (4,5). Though the incidence rate of CC has been rising over the past decades (6-8), the overall survival rate remains low, and overall survival rate has been rising over the past decades (6-8).

Both environmental and genetic factors can affect the risk of CC. One of the most common factors that cause CC in the western world is primary sclerosing cholangitis (PSC), which is an inflammatory disease closely related with ulcerative colitis. It was suggested that PSC can increase the risk of CC by ~10-15% (9). However, the molecular mechanism of PSC increasing the risk of CC remains largely unknown (10,11). Some liver diseases caused by parasitic infection have been considered causative of CC. Previous research has identified that colonization with the liver flukes Opisthorchis viverrini and Clonorchis sinensis is associated with the development of CC (12-14). Numerous studies have suggested that chronic liver disease can also increase the risk of CC, including viral hepatitis related liver disease and alcoholic liver disease (15,16). To date, numerous genetic factors have also been investigated to show evidences of being involved in the outcome of CC, including p53, cyclins, mucins and CRP (17). Aneuploidy has additionally been suggested to have a significant role in the development of CC.

Despite previous research aiming to reveal the genetic association with the risk of CC, few studies have focused on gene expression changes together with the development of CC over different days. The present study aimed to generate a time series analysis of CC at the gene expression level in order to identify novel gene patterns that are associated with the development of CC (18).

Materials and methods

Microarray data for CC. The study was approved by the ethics committee of Harbin Medical University. A gene expression profile dataset related to CC development formed over different days was obtained from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) (19) in the National Center for Biotechnology Information (NCBI), with accession number GSE15498. There were 21 samples in total, including nine normal tissues, nine intrahepatic CC tissues and three metastasis tissues. All microarray data were generated on the platform of the Affymetrix Rat Genome 230 2.0 Array (Affymetrix, Inc., Santa Clara, CA, USA).

Microarray data pre-processing. The original CEL files were downloaded from NCBI and the standard array processing procedures of the Affymetrix Rat Genome 230 2.0 Array.
were followed. The ‘affy’ package in R (20) was used to extract signal values, which were then scaled to a mean of 100, followed by log2 transformation of the array signals. A quantile normalization method (21) was used to normalize the expression value across samples. Finally, all probe set IDs were transformed to gene symbol IDs based on the documentation of the manufacturer for downstream analysis.

**Principle component analysis (PCA) and pair-wise correlation analysis.** The R (http://www.r-project.org/) ‘princomp’ function in the package ‘stats’, was used to perform the (PCA) based on the gene expression level following normalization across all 21 samples. Furthermore, a pair-wise correlation matrix was calculated among these samples. A heatmap based on the correlation matrix was constructed, as well as hierarchical clustering, by using ‘euclidean’ distance as the input.

**Differentially expressed genes (DEGs) analysis.** To detect the DEGs among normal, CC and metastasis tissue, a two-way analysis of variance (ANOVA) test was performed (22) with the gene expression profile as an input. Genes with a P<0.0001 were considered to be associated with the development of CC. To investigate whether these DEGs were significant, a permutation test to evaluate the false discovery rate (FDR) was additionally performed.

**Functional enrichment analysis.** The Database for Annotation, Visualization and Integrated Discovery (DAVID) (23) was used to analyze the enriched Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway terms (24). All expressed genes were treated as the background for this test.

**k-means cluster analysis.** To perform a comprehensive analysis of genes associated with CC genesis and identify gene clusters that may function in the development of CC, a k-means cluster method (25) was used to classify genes into different groups according to their change in expression pattern. Specifically, the mean expression value for each differentially expressed gene within each group (normal and tumor tissues from 10, 15 and 25 days, respectively, as well as metastasis tissue) was calculated. Next, a z-score transformation for the gene expression profile with mean values across sample groups described above based on a standard normalization method was performed in order to eliminate the absolute expression value effect and focus on the pattern change. Finally, a k-means cluster method was carried out for these significantly DEGs, according to their z-transformed expression profile (k=8).

**Gene module functional analysis.** For each gene module identified by the k-means cluster method, DAVID was used to analyze the enriched GO terms, as well as KEGG pathway terms. Functional terms with P<0.05 were considered to indicate statistically significant enriched functions.

**Results**

**Genes associated with CC.** There were 21 samples used in total in the present study, which were grouped into three types, including, 12 intrahepatic CC samples formed at different days ranging from 10 to 25 days, 12 normal control samples taken from the same days as a pair-wise comparison with intrahepatic CC samples and three metastasis samples used as a positive control. A PCA was conducted to demonstrate the association between samples in terms of their gene expression profile. The results showed a clear separation pattern between control samples and samples with disease (Fig. 1A). Hierarchical cluster results based on the pair-wise correlation matrix of each sample (shown in Fig. 1B) also suggest that control samples and disease samples may be successfully separated in terms of their gene expression profile. A two-way ANOVA was used to detect genes differentially expressed among these three groups of samples, as well as between the samples that derived from different days. Genes with P<0.0001 were defined as CC related genes. In total, 4,136 genes were identified that showed significantly different levels of expression between at least two types of samples. To evaluate whether these genes were significant rather than random, a permutation test was performed by shuffling the samples and repeating the ANOVA test 100,000 times. By using P<0.0001 as a cutoff, an FDR=0.03% was produced.

**Functional analysis of CC gene signatures.** In order to fully determine the association between the DEGs and the phenotype of CC, a hypergeometric test to detect the enriched GO terms was used. DAVID was used to perform this analysis. Only over-represented GO terms with P<0.05 following Bonferroni correction were considered to be significantly enriched GO terms. The top ten enriched GO terms are listed in Table I. The enriched terms, such as intracellular organelle lumen, membrane-enclosed lumen, organelle lumen have been associated with CC in several previous studies (26,27).

**Pathway analysis of CC gene signatures.** To further understand the association between the DEGs and their involvement in the development of CC, the type of molecular pathways that were significantly enriched were tested using the KEGG definition. KEGG terms with a P<0.05 following Bonferroni correction were selected (Table II). As expected, a significantly enriched pathway termed ‘mitochondrion’, was identified, which is directly linked with CC (28), suggesting that these DEGs participate in the development of CC.

**Identification of gene clusters associated with specific CC pathways.** For those genes that were differentially expressed between normal and tumor tissues at various developmental stages, it was asked what was their changing pattern as well as each pattern's biological function, in terms of their association with CC. A k-means cluster method was used to classify the genes into different modules according to their change in expression pattern. To eliminate the absolute expression level effect and only focus on the gene expression change curve difference, the gene expression profile was normalized based on the z-score transformation. Eight patterns were produced in total (Fig. 2). Genes that show a consistent changing pattern along with developmental stages of CC (such as from 10 to 15 days and to 25 days) should be considered to be associated with the chronological development of CC, since it was hypothesized that a good correlation between change in gene expression and change in disease phenotype is strong evidence of association.
Table I. Top ten enriched GO terms.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Count</th>
<th>Bonferroni</th>
<th>Fold enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005829</td>
<td>Cytosol</td>
<td>384</td>
<td>2.69x10^-23</td>
<td>3.412796</td>
</tr>
<tr>
<td>GO:0005739</td>
<td>Mitochondrion</td>
<td>409</td>
<td>2.37x10^-20</td>
<td>3.078936</td>
</tr>
<tr>
<td>GO:0031974</td>
<td>Membrane-enclosed lumen</td>
<td>405</td>
<td>7.46x10^-19</td>
<td>2.782038</td>
</tr>
<tr>
<td>GO:0043233</td>
<td>Organelle lumen</td>
<td>394</td>
<td>3.12x10^-18</td>
<td>2.423745</td>
</tr>
<tr>
<td>GO:0070013</td>
<td>Intracellular organelle lumen</td>
<td>380</td>
<td>1.98x10^-17</td>
<td>3.017707</td>
</tr>
<tr>
<td>GO:0006414</td>
<td>Translational elongation</td>
<td>61</td>
<td>3.36x10^-16</td>
<td>2.548286</td>
</tr>
<tr>
<td>GO:0022627</td>
<td>Cytosolic small ribosomal subunit</td>
<td>31</td>
<td>1.10x10^-11</td>
<td>1.543415</td>
</tr>
<tr>
<td>GO:0044429</td>
<td>Mitochondrial part</td>
<td>185</td>
<td>1.21x10^-11</td>
<td>2.26445</td>
</tr>
<tr>
<td>GO:0044445</td>
<td>Cytosolic part</td>
<td>60</td>
<td>4.53x10^-11</td>
<td>1.482382</td>
</tr>
<tr>
<td>GO:0022626</td>
<td>Cytosolic ribosome</td>
<td>37</td>
<td>4.91x10^-11</td>
<td>2.004213</td>
</tr>
</tbody>
</table>

GO, gene ontology.

Table II. Enriched Kyoto Encyclopedia of Genes and Genomes pathways.

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Term</th>
<th>Count</th>
<th>Bonferroni</th>
<th>Fold enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>rno03010</td>
<td>Ribosome</td>
<td>64</td>
<td>6.62x10^-23</td>
<td>3.412796</td>
</tr>
<tr>
<td>rno00280</td>
<td>Valine, leucine and isoleucine degradation</td>
<td>32</td>
<td>1.83x10^-08</td>
<td>3.078936</td>
</tr>
<tr>
<td>rno03030</td>
<td>DNA replication</td>
<td>22</td>
<td>3.75x10^-04</td>
<td>2.782038</td>
</tr>
<tr>
<td>rno00071</td>
<td>Fatty acid metabolism</td>
<td>23</td>
<td>0.004411</td>
<td>2.423745</td>
</tr>
<tr>
<td>rno00410</td>
<td>β-alanine metabolism</td>
<td>15</td>
<td>0.008072</td>
<td>3.017707</td>
</tr>
<tr>
<td>rno00640</td>
<td>Propanoate metabolism</td>
<td>19</td>
<td>0.012183</td>
<td>2.548286</td>
</tr>
<tr>
<td>rno04510</td>
<td>Focal adhesion</td>
<td>68</td>
<td>0.016978</td>
<td>1.543415</td>
</tr>
<tr>
<td>rno00380</td>
<td>Tryptophan metabolism</td>
<td>22</td>
<td>0.025275</td>
<td>2.26445</td>
</tr>
<tr>
<td>rno04810</td>
<td>Regulation of actin cytoskeleton</td>
<td>70</td>
<td>0.049876</td>
<td>1.482382</td>
</tr>
</tbody>
</table>

KEGG, Kyoto Encyclopedia of Genes and Genomes.

Figure 1. Principle component analysis and pair-wise correlation cluster analysis of 21 samples. (A) The principle component analysis of 21 samples. Black represents tumor samples; green represents normal samples; red represents metastasis samples. The different symbols represent the different histological days of each sample (as shown in the figure). (B) A heatmap of 21 samples showing the hierarchical clustering of all 21 samples based on the pair-wise correlation matrix. T, tumor; N, normal; M, metastasis; D, days.
Genes that show a pattern difference between tumor and normal samples across the same chronological days should also be firstly considered to have functional impact, especially for those genes that have an opposite pattern. Notably, three interesting gene clusters were identified: i) 470 downregulated genes along with the developmental stage of CC while the control curve was flat (Fig. 2C); ii) 419 downregulated genes along with the developmental stage of CC while the control curve showed a slight upregulation (Fig. 2E); iii) 491 upregulated genes along with the developmental stage of CC while the control curve was flat (Fig. 2H). These genes were possible candidate biomarkers of CC in specific pathways, which may function in the development of CC and should be studied further in future studies.

To investigate the functional importance of the gene clusters mentioned, the enriched function of all three interested gene groups was investigated. For gene cluster-3 (Fig. 2C), it was identified that genes were enriched in cell structural related functions such as ‘structural molecule activity’, ‘microtubule’, and ‘microtubule skeleton’. Genes in cluster-5 (Fig. 2E) were predominantly enriched in gene transcription and translation related pathways, suggesting that these genes may participate in facilitating translation of important genes that are associated with CC. Notably, genes in module-5 were found to be additionally enriched in a KEGG pathway termed ‘mitochondrion’, which was suggested to be a direct evidence that these genes were related with the development of CC (28). This indicated that genes downregulated with the developmental stage of CC were candidate biomarkers of CC. Genes in cluster-8 (Fig. 2H) were enriched in the GO terms ‘Focal adhesion’ and ‘Regulation of actin cytoskeleton’ (P<0.05).

Discussion

CC is considered to be an incurable cancer, as well as a rapid lethal malignancy, unless the primary tumor and any metastases can be fully removed surgically. To date, only a few genetic regions have been investigated to show an association with the development of CC, including p53, cyclins, mucins, and CRP (17). In the present study, a gene expression profile change along with the chronological development of CC, was performed in a time-series dataset of CC patient samples by...
using the same chronological samples as a control. The PCA revealed that normal samples may be obviously distinguished both from tumor and cancer samples (Fig. 1A), demonstrating that the most marked change in the gene expression level is between normal developed tissue and tumor tissue. It was additionally noticed, however, that there was a fundamental difference between tumor tissues with various developmental stages, ranging from 10 to 25 days. Notably, there was a consistent change in gene expression pattern as indicated by the PCA, in particular in the late developmental stages of tumor tissue (25 days) is closer to the stage of cancerous tissue (Fig. 1A). A heatmap based on pair-wise correlation of each sample in terms of their gene expression profile, generated the same conclusion as the PCA (Fig. 1B). These data suggested that the development of CC from early to late stage was generally a consistent process at the molecular level, reflecting that the transcriptome status is a potential molecular symbol of the developmental stage of CC. It is hypothesized that this result will facilitate the molecular diagnostic research of CC in the future.

A two-way ANOVA was performed in order to identify candidate gene markers that were differentially expressed between normal and tumor tissues, with various developmental stages. In total, 4,136 DEGs, of which most were significantly differentially expressed between normal and intrahepatic CC tissues (P<0.0001), were identified. Permutation testing provided a considerably low FDR (0.03%), suggesting these DEGs were truly significant molecular markers rather than background noise. Of these DEGs, numerous genes have been previously identified to be CC-associated genes, including GAPDH, CITED4 and CLDN4 (17,30,31). Notably, a well studied zinc finger gene, SNAI1, which is considered to be associated in the molecular pathways of CC development was identified in the DEG gene list (4.1 fold change; P=1.08e-07) (31-33). SNAI1 was shown to be upregulated in intrahepatic CC, which is consistent with the observations of the present study. The function of SNAI1 is to suppress the expression of E-cadherin that mediates cell-to-cell adhesion in order to regulate tumor progression and metastases (26,35).

GO enrichment analysis revealed that the significant DEGs were likely to be involved in functions such as intracellular organelle lumen, membrane-enclosed lumen and organelle lumen. These terms were significantly associated with the development of CC as additionally shown in previous research (26,36). This suggested that our analysis was robust and consistent with previous studies. In addition, mitochondrial functions (including mitochondrion, mitochondrial part) were also significantly enriched (Bonferroni P<cle-12), suggesting that mitochondrion-related molecular pathways were associated with the development of CC. In agreement with this, previous studies have shown that mitochondrion-mediated specific cell type apoptosis is an important cause for the development of metastases in human CC cells. Okaro et al (28) showed that a mitochondrial benzodiazepine receptor antagonist, PkII1195, can downregulate the apoptosis tolerance in human CC cells (28). Another study found that Mcl-1 can mediate apoptosis-inducing ligand resistance in human CC cells (37). Taken together, these data suggest that other genes in the gene list produced in the present study, that are associated with mitochondrion function, are potential candidate genes that should be studied in the future in the context of CC. KEGG pathway enrichment analysis provided additional information regarding the linkage between molecular pathways and the phenotype of CC (Table II). It was observed that significantly enriched terms, such as ‘Focal adhesion’ and ‘Regulation of actin cytoskeleton’ (P<0.05) were well studied pathways that have been previously associated with metastasis (38,39), thus adding to the reliability of these DEGs. In conclusion, it is considered that these DEGs detected by two-way ANOVA represent true molecular markers that can distinguish between normal intrahepatic duct cells and intrahepatic CC cells.

To further investigate genes that have different functions in the tumorigenesis and development of CC, a k-means cluster method was performed for these DEGs, in order to identify gene markers that could distinguish each other by gene expression pattern. Specifically, genes that showed a consistent change with developmental stages (from 10 to 25 days) and a different change in expression curve in that we believe a correlation between gene expression and phenotypic change of CC is always a strong evidence of association were investigated. Distinguishable patterns between tumor and normal samples should be considered to be of importance in terms of gene function in specific pathways of CC. Based on these two criteria, 3/8 clusters produced notable patterns of expression. In cluster-3, genes in normal tissues were highly expressed as compared with tumor tissues, and were expressed in a steady state across different days. In comparison the genes in the tumor tissues were downregulated together with the development of CC. Enrichment tests suggested that genes within this cluster were involved in pathways including ‘structural molecule activity’, ‘microtubule’ and ‘microtubule skeleton’, which have been linked with tumorigenesis in several previous studies (40,41). Genes in cluster-5 were also shown to have a downregulation expression pattern in the developmental stages of CC. However, genes in this group were highly expressed in tumor tissues as compared with normal tissues. Pathway enrichment analysis showed that these genes were enriched in mitochondrion related pathways, suggesting they may function in cell apoptosis-induced pathways. Interestingly, SNAI1 was consistently within this cluster, indicating that studying other genes within this cluster may provide a link between SNAI1 regulation in the development of CC. For cluster-8, genes that were upregulated across tumor samples whilst maintaining flat expression across normal samples were identified. GO enrichment analysis additionally suggested that genes within this group were related with ‘Focal adhesion’ and ‘Regulation of actin cytoskeleton’, raising the possibility that these genes may be tumorigenesis related markers associated with CC development.

In conclusion, the present study has conducted a time series analysis to reveal gene signatures that are associated with the development of CC, based on changes in the gene expression profile. Genes related to CC were shown to be involved in cancer pathways such as ‘mitochondrion’ and ‘focal adhesion’. Three interesting gene groups were identified by a k-means cluster method based on a normalized gene expression profile, of which each showed enrichment functions associated with the development of CC. It is considered that this work is of great importance and will lead to novel discoveries regarding the genetic study of CC.
References


