Cancer Genetics 208 (2015) 482–491

Cancer Genetics

ORIGINAL ARTICLE

Comparative transcriptome analysis reveals that the extracellular matrix receptor interaction contributes to the venous metastases of hepatocellular carcinoma

Hong Zhang [a,1], Junyi Ye [a,1], Xiaoling Weng [a,1], Fatao Liu [b], Lin He [a,c], Daizhan Zhou [c,*], Yun Liu [a,*]

[a] *Institutes of Biomedical Sciences, Fudan University, Shanghai, China;* [b] *Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Graduate School of the Chinese Academy of Sciences, Shanghai, China;* [c] *Bio-X Center, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China*

Hepatocellular carcinoma (HCC) is the most common type of liver cancer in the world. Portal vein tumor thrombus (PVTT) is one of the most serious complications of HCC and is strongly correlated with a poor prognosis for HCC patients. However, the detailed mechanism of PVTT development remains to be explored. In this study, we present a large-scale transcriptome analysis, by RNA sequencing, of 11 patients diagnosed with HCC with PVTT. The dysregulated genes between HCC and PVTT suggested that the extracellular matrix receptor interaction was correlated with the venous metastases of HCC. Among all of the recurrent alternative splicing events, we identified exon 6 skipping of *RPS24*, which is likely to be a cancer driver. We also identified five common fusion genes between HCC and its corresponding PVTT samples, including *ARID1A-GPATCH3*, *MDM1-NUP107*, *PTGES3-RARG*, *PRLR-TERT,* and *C9orf3-TMC1*. All of these findings broaden our knowledge of PVTT development and may also contribute to the diagnosis and treatment of HCC patients with PVTT.

Keywords Hepatocellular carcinoma, portal vein tumor thrombus, differentially expressed genes, alternative splicing, gene fusion

Hepatocellular carcinoma (HCC) is the fifth-most common cancer (1) and is the second leading cause of cancer-associated deaths worldwide (2). HCC leads to a large number of deaths in the world each year, almost half of which occur in China (3). Although cancer-associated mortality has generally been decreasing, the mortality of HCC has exhibited the fastest rate of increase in the United States (4). A great proportion of HCC patients show symptoms of vascular invasion with intrahepatic metastases, and the 5-year survival rate is only 30–40% (5). HCC tends to invade branches of the portal vein to form portal vein tumor thrombus (PVTT), which may block the portal vein and result in portal hypertension. PVTT is one of the most serious complications of HCC and is strongly correlated to the poor prognosis of HCC patients. PVTT yields a median survival of less than 6 months if left untreated (4). In previous studies, aberrant epigenetic modifications (4–7), loss of DNA repair genes (8,9), and genomic instability (10,11) were observed to contribute to HCC tumorigenesis and intrahepatic metastasis. For instance, Huang et al. used exome sequencing to screen out a set of frequently mutated genes, such as the AT rich interactive domain 1A gene (*ARID1A*), the sterile alpha motif domain containing 9-like gene (*SAMD9L*), and the homeobox A1 gene (*HOXA1*) (12). These studies suggested that somatic mutations may contribute to the development of HCC. These changes in the cancer genome can result in the reprogramming of the transcriptome, which can subsequently lead to

Received April 9, 2015; received in revised form June 8, 2015; accepted June 9, 2015.

* Corresponding authors.

E-mail addresses: zhoudaizhan@gmail.com, superliuyun@gmail.com

[1] Both authors contributed equally to this work, and both are considered first author.

2210-7762/$ - see front matter © 2015 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cancergen.2015.06.002

abnormal cellular behavior, ultimately driving cancer progression (13). Unbiased high throughput approaches for identifying expression changes between tumor and normal samples can reveal multiple diagnostic and prognostic markers (14,15). However, because of the genetic heterogeneity and complex genomic alterations shown in this type of cancer (16,17), the clear and comprehensive understanding of the development of HCC with PVTT and the proper prognosis remain to be studied in-depth. A thorough exploration of gene expression profiling for HCC patients with PVTT has not been performed, though it is common to use RNA sequencing (RNA-seq) to study various types of cancer transcriptomes (18,19).

This study explores the transcriptomic signatures of HCC with PVTT patients, which could help to elucidate the intrahepatic metastasis of HCC and may also contribute to the development of novel diagnostic and therapeutic approaches.

Materials and methods

Patients and tissue samples

Samples of PVTT, primary tumor (HCC), and adjacent nontumor (ANT) tissues from 11 patients were obtained from Shanghai Eastern Hepatobiliary Surgery Hospital. HCC diagnosis was confirmed by pathological examination. Serum alpha-fetoprotein (AFP), carcinoembryonic antigen (CEA), and carbohydrate antigen 19-9 (CA19-9) were measured using an electrochemiluminescence immunoassay on Cobas e601 (Roche, Basel, Switzerland) immunoassay analyzers. Detailed information for the discovery cohort is shown in Supplementary Table S1. Written informed consent was obtained from all patients for the research that used their tumor samples. After surgical resection, the tissue samples were immediately snap frozen in liquid nitrogen and subsequently stored at -80°C until use. Clinical and pathological information was collected from the patients' records and pathology reports. The study protocol was approved by the hospital ethics committee or the ethics committee of the Shanghai Institute for Biological Sciences.

Library construction

Total RNA was extracted from ANT, HCC, and PVTT tissues with TRIzol (Thermo Fisher Scientific, Waltham, MA). According to the Illumina (San Diego, CA) standard kit, a TruSeq RNA Sample Prep v2 Guide was used to prepare the RNA sequencing library. Details are shown in Supplementary Information: RNA-seq library construction.

Processing and mapping raw reads

According to a base-calling pipeline, the images generated by the HiSeq 2000 sequencing system (Illumina) were translated into nucleotide sequences. The raw reads were saved in FASTQ format, and Trimmomatic (20) was used to filter the raw data prior to analyzing the data. Three criteria were used: reads shorter than 36 bases were discarded, reads containing sequencing adaptors were removed, and bases with a quality score less than 15 were removed. All the subsequent analyses were based on clean and qualified data. Next, the clean sequencing reads were aligned with the UCSC Genome Browser hg19 reference genome using TopHat version 2.0.9 (21), which incorporates Bowtie version 1.0.0 (22) to perform the alignment. We built the reference index by using Bowtie with a FASTA format file for the whole hg19 genome. All the parameters were set to the default settings.

Detection of differentially expressed genes

The aligned files were subsequently processed by Cufflinks version 2.1.1 (21). The unit for transcript abundance estimation was fragments per kilobase pair of exon per million fragments mapped (FPKM). The Bayesian inference method was used to calculate confidence intervals for FPKM estimation. Two methods, the *t* test and DESeq (23), were used to detect differentially expressed genes (DEGs) among the ANT, HCC, and PVTT samples, respectively (Supplementary Figure S1). First, we integrated the data of the 11 samples and later applied the paired *t* test to test the DEGs; *P* values less than 0.05 and fold change greater than 2 were set as the thresholds of significance. Next, we used DESeq, which is a new nonparametric algorithm for the detection of DEGs; *P* values less than 0.01 and fold change greater than 2 were considered significant.

Differential exon-skipping events testing

A mixture of isoforms (MISO) (24) analysis, which adopted a Bayesian inference algorithm to calculate the probability that a read came from a specific isoform, was employed to detect exon-skipping events (25). In our study, MISO was applied to identify differentially regulated exons between samples. The read alignment BAM files produced by TopHat and the hg19 alternative events annotations downloaded from the MISO main page served as an input. The software computed the percentage of transcripts that were spliced to include exons, known as the PSI (Ψ) (26):

$$\text{Percentage Splicing In (PSI)} = \frac{\text{\# of reads supporting inclusion}}{\text{\# of reads supporting inclusion} + \text{skipping}}$$

To obtain more reliable differential exon-skipping events, we used the stringent form of MISO (Bayes factor >1,000, default value of 10, and $|\Delta PSI| > 20\%$, namely, $|\Delta\Psi| > 0.2$).

Detecting human gene fusions

We used TopHat to map the clean reads to the downloaded reference transcript sequences. Next, the mapped BAM files were passed to TopHat-Fusion (27) to detect the candidate fusions. The parameters were set to the default settings. To obtain more reliable results, we identified several filtering steps for the candidate fusions detected by the software:

- Number of reads had to be 10 or greater.
- Both of the breakpoints were known exon boundaries.
- Candidates consisting of two adjacent or overlapping genes were discarded.
- The fusion was not identified in ANT tissues.