Exome sequencing of hepatitis B virus–associated hepatocellular carcinoma

Jian Huang1,4,­ 7, Qing Deng1,2,7, Qun Wang1,2,7, Kun-Yu Li2, Ji-Hong Dai2, Niu Li1,2, Zhi-Dong Zhu3, Bo Zhou1,2, Xiao-Yan Liu1,2, Rui-Fang Liu1,2, Qian-Lan Fei2, Hui Chen2, Bing Cai5, Boping Zhou4, Hua-Sheng Xiao3, Lun-Xiu Qin6 & Ze-Guang Han1,2

Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide and shows a propensity to metastasize and infiltrate adjacent and more distant tissues1. HCC is associated with multiple risk factors, including hepatitis B virus (HBV) infection, which is especially prevalent in China. Here, we used exome sequencing to identify somatic mutations in ten HBV-positive individuals with HCC with portal vein tumor thromboses (PVTTs), intrahepatic metastases. Both C>G>A: T and T>A>A:T transversions were frequently found among the 331 non-silent mutations. Notably, ARID1A, which encodes a component of the SWI/SNF chromatin remodeling complex, was mutated in 14 of 110 (13%) HBV-­associated HCC specimens. We used RNA interference to assess the roles of 91 of the confirmed mutated genes in cellular survival. The results suggest that seven of these genes, including VCAM1 and CDK14, may confer growth and infiltration capacity to HCC cells. This study provides a view of the landscape of somatic mutations that may be implicated in advanced HCC.

Hepatocellular carcinoma is associated with infection by HBV or hepatitis C virus (HCV), alcohol consumption and aflatoxin B1 (AFB1) contamination of food. Individuals with HCC with PVTTs, which are caused by tumor invasion and metastasis via the intrahepatic portal vein, have extremely poor prognoses2,3.

To identify the genetic lesions in advanced HBV-associated HCC with intrahepatic metastasis, we employed two types of exome sequencing strategies, the Illumina/Solexa Genome Analyzer II and the ABI SOLiD platforms, to identify the somatic mutations in ten individuals with HCC with PVTTs (Supplementary Table 1). The average coverage of each base in the target regions was 28-fold and 88-fold per tumor sample (primary tumor and matched PVTT) for the Illumina and ABI platforms, respectively (Online Methods, Supplementary Fig. 1 and Supplementary Table 2). Here, we propose 475 candidate somatic single-nucleotide variant (SNV) mutations, including 331 non-silent mutations, 112 synonymous mutations and 32 mutations located in untranslated regions, in the 10 individuals with HCC (Supplementary Fig. 2 and Supplementary Table 3), which were supported by the results of both sequencing platforms. Of 224 non-silent mutations selected from these SNVs, 193 (86.2%) were confirmed in the primary tumors and/or PVTTs via conventional PCR-based Sanger sequencing (Supplementary Fig. 3 and Supplementary Table 4). In addition, 25 small somatic insertion and/or deletion (indel) mutations were identified (Supplementary Figs. 4 and 5).

A total of 356 nonsynonymous somatic mutations occurred within 347 genes (Supplementary Table 5). The average frequency of mutations was 35.6 per affected individual (Supplementary Table 6). The numbers of mutations ranged from a minimum of 8 in 1 individual with HCC (P54) to a maximum of 99 in another (P48) (Supplementary Fig. 6).

Notably, among the 331 non-silent SNV mutations, the C>G>A:T transversion was significantly enriched in 5 affected individuals, and the T>A>A:T transversion was significantly enriched in 4 affected individuals (Fig. 1). The prevalence of these transversions was significantly different from the patterns reported in HCV-associated HCCs and in other tumors whose exomes have been sequenced5,9,10. Enrichment of the C>G>A:T transversion implies that AFB1 exposure may be a critical contributor in these HBV-associated HCCs.9,11,12. However, whether AFB1 leads to T>A:A:T transversions is uncertain. We suspect that other carcinogens—such as aristolochic acid, a component of plant extracts, including herbal remedies13,14, and vinyl chloride, a gas used in the plastics industry—are involved in HBV-associated HCCs. Both aristolochic acid and vinyl chloride have been implicated in T>A:A:T transversion in some cancer-associated genes14,16.

Of the 344 nonsynonymous somatic mutations found in the 9 pairs of matched primary tumors and PVTTs (excluding 12 mutations from P47 due to unsuccessful sequencing), 324 (94.2%) were identical in these paired tumors, whereas 4 (1.1%) and 16 (4.7%) were detected only in primary tumors or PVTTs, respectively (Fig. 2a), suggesting that the PVTTs derived from malignant cells within the primary tumors that had most of the genetic lesions. Subsequently, we compared the number of sequencing reads that covered the somatic mutations in

Received 13 December 2011; accepted 1 August 2012; published online 26 August 2012; doi:10.1038/ng.2391
The primary tumors and their matched PVTTs (Supplementary Table 7) and analyzed the corresponding PCR-based Sanger sequencing data. Of a total of 65 mutations with significant differences in the number of sequencing reads between the primary tumors and the PVTTs (Fig. 2b), the mutations in KDM6A, CUL9, FGD6, AKAP3 and RNF139 were found only in the PVTTs of three individuals.

We further analyzed somatic mutations in six genes on chromosome X—FTHL17, DOCK11, TMEM35, CHRD11, ATRX and USP26—in three males with HCC. These mutations, which exhibited single sequencing peaks in the PVTTs, occurred in a heterozygous manner in the primary tumors, implying that the metastatic cancer cells within the primary tumors had been endowed with mutations in these genes.

The 347 mutated genes were significantly enriched in certain biological processes (P < 0.05), including cell adhesion, motility, migration and morphogenesis (Supplementary Fig. 7 and Supplementary Tables 8 and 9). Some of these genes could be assigned to known core cancer signaling pathways.

Ten genes carrying two confirmed non-silent mutations were found to have statistically significant differences in gene mutation rate (P < 0.01) (Table 1 and Supplementary Fig. 8). These ten genes included TP53, a known driver gene in HBV-associated HCC,\textsuperscript{17,18} ARID2, which is commonly mutated in HCV-positive HCC,\textsuperscript{19} and ARID1A, in which mutations have been found in other tumors.\textsuperscript{20-23} We further evaluated the mutation frequency of the ten candidate genes in additional paired HBV-associated HCC samples via PCR-based Sanger sequencing of their protein-coding exons. These analyses identified 33 mutations in TP53 and 21 mutations in ARID1A in 28 (28%) and 13 (13%) of the additional 100 HCC cases, respectively (Supplementary Table 10). Of the 110 paired HCC samples used in this study, 35 (27.3%) and 14 (12.7%) had TP53 and ARID1A mutations, respectively, whereas 4 HCC samples carried somatic mutations in both TP53 and ARID1A.
ARID1A (P = 0.838), implying that an association between ARID1A and TP53 mutations cannot be absolutely excluded. Moreover, the 23 mutations in ARID1A that are present in 14 cases included 15 missense, 2 nonsense and 5 frameshift indel mutations (Fig. 3a and Supplementary Table 10). In addition, we found 3 missense mutations in SAMD9L, a paralog of SAMD9 in which germline mutations have been reported in normophosphatemic familial tumoral calcinosis, in 3 (3.8%) of an additional 80 HCC cases.

To evaluate whether ARID1A mutations occur in HCCs with other etiologies, we sequenced ARID1A in 26 HCV-associated HCCs, 10 non-virus-infected HCCs and 1 HCC with both HBV and HCV infection. ARID1A mutations were found in four cases, including two HCV-associated HCCs and one each of the other subclasses (Supplementary Table 10).

Furthermore, we sequenced ARID1A in 13 HCC cell lines. We identified six candidate nonsynonymous mutations in four HCC cell lines: PLC/PRF/5, Hep3B, YY-8103 and HCC-LM6 (Fig. 3b and Supplementary Table 10). Most notably, 1 in-frame indel, 2 missense mutations and 63 synonymous mutations were found in HCC-LM6 cells, which exhibit a pronounced capacity to metastasize to the lung.

Next, we compared ARID1A sequences in the MHCC97-L, MHCC97-H and HCC-LM6 cell lines, which exhibit different levels of high metastatic potential and are all derived from the same individual with HCC, on the basis of published literature and our analysis of their SNP genotypes (Supplementary Fig. 9). However, none of the ARID1A mutations detected in the HCC-LM6 cells with the highest metastatic potential occurred in the other two cell lines (Fig. 3b).

### Table 1 Somatically mutated genes in the exome sequences of ten HBV-associated HCCs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein varianta</th>
<th>P valueb</th>
<th>Total number</th>
<th>Cases with mutations</th>
<th>Non-silent mutations</th>
<th>Synonymous mutations</th>
<th>dN/dSc</th>
<th>P valuec</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>p.Tyr220Cys p.Ile251Phe</td>
<td>1.01 × 10^{-19}</td>
<td>110</td>
<td>30</td>
<td>35</td>
<td>0</td>
<td>35.0</td>
<td>1.41 × 10^{-4}</td>
</tr>
<tr>
<td>ARID1A</td>
<td>p.Tyr1211* p.Gln1212Leu</td>
<td>2.50 × 10^{-6}</td>
<td>110</td>
<td>14</td>
<td>22</td>
<td>1</td>
<td>22.1</td>
<td>4.88 × 10^{-3}</td>
</tr>
<tr>
<td>SAMD9L</td>
<td>p.Trp495Cys p.Gly1554Ala</td>
<td>1.15 × 10^{-7}</td>
<td>90</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5.0</td>
<td>0.22</td>
</tr>
<tr>
<td>ARID2</td>
<td>p.Ile968Leu p.Pro536fs</td>
<td>1.44 × 10^{-6}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
<tr>
<td>C1orf34</td>
<td>p.Leu511Gln p.Glu642*</td>
<td>4.07 × 10^{-15}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
<tr>
<td>AMPH</td>
<td>p.Phe399Le p.Met979Ile</td>
<td>4.15 × 10^{-15}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
<tr>
<td>ELMO1</td>
<td>p.Leu659Phe p.Tyr395Phe</td>
<td>5.53 × 10^{-13}</td>
<td>50</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2.1</td>
<td>1.00</td>
</tr>
<tr>
<td>DSE</td>
<td>p.Try867Phe p.Lys514Met</td>
<td>1.55 × 10^{-12}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
<tr>
<td>SPAG17</td>
<td>p.Asp1807Val p.Ser329Thr</td>
<td>1.13 × 10^{-7}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
<tr>
<td>CSMD3</td>
<td>p.Phe2678Ile p.Ser157Cys</td>
<td>2.13 × 10^{-4}</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2.0</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The numbers include all of the HBV-associated HCC samples examined in this study. Protein variants due to mutations found in two of the ten individuals with advanced HCC or representing two mutations in the same gene. Statistical significance of the occurrence of mutations in each gene was calculated with the χ²-squared test to distinguish the gene mutation rate from the background mutation rate (Online Methods). Statistical analyses were performed with a Fisher’s exact test to distinguish between the observed results and a null model (dN/dS = 2).

**Figure 3 ARID1A mutations in HCC samples and cell lines have roles in cell proliferation and metastasis.** (a) The distribution of the ARID1A mutation types in the HCC samples and cell lines is shown on a schematic of the ARID1A protein. Symbols indicate the type and location of the mutations. (b) Schematic showing the mutational evolution of ARID1A, including stepwise increases in metastatic potential, in three HCC cell lines that were derived from the same individual with HCC. Symbols indicate the type and location of the mutations. (c) ARID1A knockdown via RNAi promoted cell migration (c) and invasion (d) in MHCC-97L and MHCC-97H cells. Cell migration and invasion were evaluated using Matrigel transwell assays; permeable cells were stained with crystal violet and counted. Cell migration and invasion assays were repeated three times. Error bars, s.d. of triplicate experiments. Results were statistically analyzed with two-tailed t tests. *P < 0.05; **P < 0.01. NC, negative control.
This observation suggests that genetic lesions in ARID1A may contribute to tumor progression. To test this hypothesis, we assessed the effect of loss of function of this gene on cellular behaviors using RNA interference (RNAi) in four HCC cell lines with wild-type ARID1A (MHCC-97L, MHCC-97H, SK-hep-1 and WRL68). Notably, ARID1A knockdown significantly promoted the proliferation, migration and invasion of these HCC cell lines but did not affect the behaviors of HCC-LM6 cells (Fig. 3c,d and Supplementary Figs. 10 and 11), supporting the notion that the emergence of loss-of-function ARID1A mutations may be a crucial event in HCC invasion and metastasis.

Driver mutations tend to be non-silent and are clustered in functionally relevant regions of cancer-associated genes.27,28 With this consideration in mind, we identified potential driver mutations within cancer-associated genes in a large number of HCC samples (85–120 cases) by sequencing partial protein-coding regions adjacent to confirmed mutations in 147 genes. We found additional non-silent mutations in several genes, including HOXA1 in two cases, as well as VCAM1, TMEM35 and PIK3AP1 in one case each (Supplementary Fig. 12 and Supplementary Table 11).

Furthermore, we employed a loss-of-function RNAi screening strategy to find potential cancer-associated genes that are likely to have crucial roles in cellular proliferation or survival (Supplementary Fig. 13a). We knocked down 91 genes (Supplementary Table 12) in 8 HCC cell lines (Fig. 4a, Supplementary Fig. 13b,c and Supplementary Tables 13 and 14). Notably, knockdown of two genes, VCAM1 and TMEM2, significantly promoted proliferation of HCC cells in four cell lines (Table 2). These independent experiments indicate that effective VCAM1 knockdown could promote colony formation and migration in multiple HCC cell lines (Fig. 4b,c and Supplementary Figs. 14, 15a,b and 16a).

In contrast, the knockdown of five genes, including CDK14, HOXA1, TMEM35, ELL and CSNK1G3, significantly suppressed cell growth in at least three HCC cell lines (Table 2). In these independent experiments, efficient knockdown of CDK14, which is known to be involved in cell cycle progression and cancer cell motility,29,30 significantly suppressed the growth and colony formation of multiple HCC cell lines (Fig. 4d and Supplementary Figs. 14, 15c and 16a), whereas the CDK14R99W mutant that was detected by exome sequencing significantly promoted colony formation in three HCC cell lines (Fig. 4e, Supplementary Figs. 15d,e and 16b).

In addition to the mutations identified in seven cancer-associated genes, VCAM1, HOXA1 and TMEM35, in a few HCC cases (such as Supplementary Table 11), the transcription of VCAM1 encoding a cell adhesion molecule was clearly downregulated in 48% of the HCC specimens (Table 2 and Supplementary Fig. 17).

Overall, the present work provides a list of genes that may be implicated in HCC pathogenesis. Of these genes, mutations in ARID1A are noteworthy, as they may indicate a potential for its use as a novel biomarker for HCC therapy and prognosis.

**Table 2.** RNAi screen for candidate cancer-associated genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of cell lines</th>
<th>Z score</th>
<th>P value</th>
<th>Expression change in HCC samples&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Upregulation (&gt;2-fold)</th>
<th>Downregulation (&gt;2-fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promotion of cell viability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCAM1</td>
<td>4</td>
<td>23.06</td>
<td>1.39 × 10⁻³</td>
<td>16% (8)</td>
<td>48% (24)</td>
<td></td>
</tr>
<tr>
<td>TMEM2</td>
<td>4</td>
<td>62.97</td>
<td>1.03 × 10⁻³</td>
<td>16% (8)</td>
<td>22% (11)</td>
<td></td>
</tr>
<tr>
<td>Inhibition of cell viability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDK14</td>
<td>4</td>
<td>–41.22</td>
<td>2.90 × 10⁻³</td>
<td>20% (10)</td>
<td>22% (11)</td>
<td></td>
</tr>
<tr>
<td>HOXA1</td>
<td>4</td>
<td>–28.14</td>
<td>2.38 × 10⁻³</td>
<td>18% (9)</td>
<td>54% (27)</td>
<td></td>
</tr>
<tr>
<td>TMEM35</td>
<td>4</td>
<td>–48.97</td>
<td>1.34 × 10⁻³</td>
<td>28% (14)</td>
<td>32% (16)</td>
<td></td>
</tr>
<tr>
<td>ELL</td>
<td>4</td>
<td>–34.75</td>
<td>1.41 × 10⁻³</td>
<td>4% (2)</td>
<td>28% (14)</td>
<td></td>
</tr>
<tr>
<td>CSNK1G3</td>
<td>3</td>
<td>–27.68</td>
<td>1.68 × 10⁻³</td>
<td>2% (1)</td>
<td>24% (12)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>The numbers of cell lines with significantly altered cell viability according to a cell-based RNAi screen in eight HCC cell lines.

<sup>b</sup>Statistical analysis (calculation of Z scores and P values) is described in the Supplementary Note. The upregulation or downregulation of the mRNA levels of these genes in a given HCC specimen was determined by comparison to levels in the matched non-tumorous liver in RT-PCR data. Numbers in parentheses represent the numbers of HCC samples.

---

**Fig. 4** Effects of mutated genes on cell behaviors as determined using cell-based RNAi. (a) Evaluation of cell viability in response to RNAi against 91 mutated genes in 8 HCC cell lines. Readout data from the cell-based RNAi screen were subjected to normalization and quality proofreading. Two statistical parameters, the Z score and the P value, were used to evaluate the effectiveness and reliability of each gene-targeting small interfering RNA (siRNA). The cutoffs for significance were Z score < −5, P value < 0.01 (red box) and Z score < 5, P value < 0.01 (purple box). (b) Anchorage-dependent (left) and anchorage-independent (right) colony formation was assessed with RNAi against VCAM1 in the Huh-7 and MHCC-97L cell lines. Negative control (NC/LUC) siRNA was directed against luciferase. (c) The effect of VCAM1 on cell migration was evaluated in MHCC-97H, HCC-LM3 and Huh-7 cells through RNAi. (d) The effect of CDK14 on colony formation was evaluated in Huh-7 and WRL68 cells through RNAi. (e) The effect of mutant CDK14R99W on colony formation was evaluated in WRL68 and MHCC-97H cell lines, with wild-type (WT) CDK14 and empty vector (V) used as controls. All experiments on VCAM1 and CDK14 were repeated three times. Error bars, s.d. of triplicate experiments. Statistical analyses were performed with a two-tailed t test.* P < 0.05; ** P < 0.01.
METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Exome sequences have been deposited at the NCBI Sequence Read Archive (SRA) under accession SRA053063.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Chinese National Key Program on Basic Research (2010CB529204 and 2010CB529206), the China National Key Projects for Infectious Disease (2012ZX10002012-008), the National Natural Science Foundation of China (81071722 and 81101875), the Shanghai Commission for Science and Technology (11ZJ1425300), the International Scientific Collaborative Project (2011ZR0001) and the National High Technology Research and Development Program of China (863 Program, 2012AA02A205 and SS2012AA020103).

AUTHOR CONTRIBUTIONS

Z.-G.H. conceived and designed the study. Q.D., J.H., Z.-D.Z. and H.-S.X. performed exome sequencing and analyzed the sequence data set. Q.D., Q.W., K.-Y.L., J.-H.D., R.-F.L. and H.C. evaluated mutations in HCC samples. Q.D. and N.L. performed large-scale RNA interference against mutated genes. Q.D., Q.W., K.-Y.L., N.L., Bo Zhou, X.-Y.L. and Q.-L.F. performed in vitro experiments on individual genes. B.C., Boping Zhou and L.-X.Q. contributed HCC samples. Z.-G.H. and Q.D. integrated, analyzed and interpreted all data. Z.-G.H. contributed to the supervision of the work. Z.-G.H. and Q.D. wrote the manuscript. All authors read and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2391. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.


ONLINE METHODS

Tissue specimens. All of the HCC specimens were obtained from individuals who underwent surgical liver tumor resection and signed informed consent forms before surgery. The primary tumors, adjacent non-tumorous livers and PVTT specimens were immediately frozen at −80 °C until DNA extraction. HCC diagnosis was confirmed by pathological examination. Protocols for investigations involving human tissues used in this project were approved by the ethics committee of the Chinese National Human Genome Center (Shanghai, China).

Liver cancer cell lines. The HCC cell lines, which included PLC/PRF/5, Sk-hep1, MHCC-97H, MHCC-97L, YY-8103, SMMC-7721, WRL68, Focus, HepG2, Hep3B, Huh-7, HCC-LM6 and HCC-LM3, were grown under standard cell culture conditions in a 5% CO₂ humidified chamber in EMEM (Sigma-Aldrich) supplemented with 10% FBS (Life Technologies), 1% l-glutamine and 1% nonessential amino acids.

Extraction of DNA. Genomic DNA was extracted from specimens using the DNeasy Tissue kit (Qiagen) according to the manufacturer’s protocols. Total DNA concentration and quantity were assessed by measuring absorbance at 260 nm with a DNA/Protein Analyzer (DU 530, Beckman).

Illumina/ Solexa sequencing. Genomic DNA was fragmented and hybridized to NimbleGen Human Exome 2.1M Arrays (Roche NimbleGen), which can enrich the exonic sequences of more than 18,000 protein-coding genes deposited in the Consensus Coding Sequence Region database (see URLs). Exome-enriched shotgun libraries were sequenced on the Illumina Genome Analyzer II platform, and single-end reads with an average size of 75 bp were generated. Image analysis and base calling were performed with Genome Analyzer Pipeline version 1.3, using default parameters. After removing reads with sequence matching the sequencing adaptors and low-quality reads with more than five unknown bases, high-quality reads were aligned to the NCBI human reference genome (hg19) using MAQ31, software that can build assemblies by mapping short reads to a reference genome, with the default parameters. To identify potential mutations, we performed local realignments of the Burrows-Wheeler Aligner (BWA)-aligned reads using the Genome Analysis Toolkit (GATK)32. The raw lists of potential somatic substitutions were called using VarScan (v2.2) on the basis of MAQ alignments33.

ABI SOLID sequencing. Following the SureSelect Target Enrichment System for the ABI SOLID System Protocol (see URLs), the genomic DNA from each individual was hybridized to the SureSelect Human All Exon Kit (38 Mb), which includes exons in NCBI CCDS, miRNAs in the Sanger V13 database and more than 300 human non-coding RNAs (for example, snoRNAs and scaRNAs). Genomic libraries were constructed using the manufacturer’s SOLID Fragment Library protocols.

Identification of candidate somatic mutations. To identify candidate somatic mutations using the two massively parallel sequencing platforms, we first used SOAPspn and BioScope software tools to identify potential SNP loci in the exomes of the tumors (primary tumors and/or PVTTs) sequenced by the Illumina/Solexa and ABI SOLID sequencing systems, respectively, using the default parameters. Next, identified candidate variation sites were filtered to remove those that did not meet the following criteria: (i) homozygous genotypes in the matched non-tumorous liver samples; (ii) potential base substitution mutations in the primary tumor and PVTT samples compared with the genotypes in the matched non-tumorous liver samples; (iii) base substitutions that had not been deposited in dbSNP or the 1000 Genomes Project database and (iv) localization to exons and splice sites.

Identification of somatic indels. First, potential indel variants were screened from ten tumor exomes using the ABI BioScope indel tool with the default parameters. Second, candidate indel variants were filtered to remove those that did not meet the following criteria: (i) presence of potential indel variants in the primary or PVTT samples; (ii) no reads harboring indel genotypes in the non-tumorous liver samples and (iii) indel variants had not been deposited in dbSNP or the 1000 Genomes Project database. Of the 622 candidate somatic indel variants that met these prerequisites, 59 were located in coding exons or splice sites and were further subjected to manual confirmation with the Integrative Genomics Viewer (IGV)34 on the basis of SOLID and/or Solexa sequencing data.

PCR-based Sanger resequencing. Primers were designed according to the target regions of interest and were used to amplify these regions by PCR for genomic resequencing. Dye terminator sequencing was followed by separation and detection on ABI 3730xl sequencers. The resulting data were screened for mutations through automated and manual steps.

Gene functional annotation. The functions of the mutated genes were annotated and analyzed using the annotation tools from DAVID Bioinformatics Resources35,36 (see URLs). Briefly, the 347 mutated genes were defined as the interrogating list, and all human genes (30,000 genes in total) were defined as the background. These sets were jointly searched against the KEGG and Gene Ontology (GO) databases. The genes that were annotated in the KEGG or GO database as being involved in signaling pathways, molecular functions or biological processes were subjected to further analysis. The threshold for the EASE score, a modified Fisher’s exact P value, was computed for gene enrichment analysis by comparing the classifications of the mutated genes with those of all human genes in the background set. If the P value was <0.05, enrichment of a given category was considered to be statistically significant.

RNA interference screen. siRNA sequences were designed against candidate genes with the siRNA Selection Program of the Whitehead Institute (see URLs). Double-stranded siRNA oligonucleotides were chemically synthesized (Shanghai GenePharma Co.), and the cell-based siRNA screen was performed in 96-well culture plates. siRNA labeled with 6-carboxyfluorescein (6-FAM) was used as a transfection control, two irrelevant siRNAs were used as negative controls and two siRNAs against AURKA were used as positive controls in each plate. Cell viability was measured using the Cell Counting Kit-8 (CCK-8, Dojindo Laboratories). Raw data from each plate were normalized and analyzed according to a previously described method37. Phenotypes were scored for their statistical significance.

ARID1A mutation screening. The coding exons of ARID1A were sequenced via PCR-based capillary Sanger sequencing. Data were analyzed with semi-automated mutation detection followed by a visual inspection of the sequenc traces. Primer sequences will be provided upon request.

ARID1A knockdown and functional assays. Two double-stranded siRNA oligonucleotides against ARID1A were designed and chemically synthesized (Shanghai GenePharma Co.) to knock down ARID1A in cell lines for functional experiments to determine growth curves, migration and invasion. These siRNAs were transfected into HCC cell lines, and cell growth was observed. Cells grown to 30–50% confluence were transfected with synthetic siRNAs at final concentrations of 50 nM, using the Lipofectamine 2000 Transfection Reagent (Invitrogen) according to the manufacturer’s instructions. Cell invasion assays were performed using 24-well transwells (8-µm pore size; BD Biosciences) coated with Matrigel (Falcon 354480, BD Biosciences). HCC cells were starved overnight in serum-free medium, trypsinized and washed three times in DMEM containing 1% FBS. A total of 1 × 10⁴ cells were then suspended in 500 µl of DMEM containing 1% FBS and added to the upper chamber, while 750 µl of DMEM containing 10% FBS and 10 µg/ml fibronectin (356008, BD Biosciences) was placed in the lower chamber. After 48 h of incubation, the Matrigel and the cells remaining in the upper chamber were removed with cotton swabs. The cells on the lower surface of the membrane were fixed in 4% paraformaldehyde and stained with 0.5% crystal violet. All experiments were independently repeated at least three times.

Statistical analysis. The statistical significance of the differences between the mutation rates of the mutated genes in HCC and the background rate was determined using the χ²-squared test or Fisher’s exact test. The gene mutation rate was calculated as the number of nonsynonymous mutations found in a given gene in the ten examined individuals with HCC divided by the total.
number of covered bases in the exons of that gene. The background mutation rate was considered to be the total number of non-silent somatic mutations identified in these primary tumors and/or PVTTs divided by the total number of covered bases of all ten HCC exomes. The statistical significance of the ratio of nonsynonymous to synonymous mutations (dN/dS) in the primary tumors and the matched PVTTs—compared with the ratio expected by chance (2:1) and based on the massively parallel sequencing data sets—was calculated using the χ²-squared test or Fisher’s exact test. The statistical significance of the results of the RNAi experiments and the cell migration and colony formation assays was determined by two-tailed t test.


