Table 1 | Clinical and pathological features of 30 LCBs analyzed by whole-genome sequencing.

ID	Age	Gender	Viral infection	Histology*	TNM†	Tumour size (mm)	vp‡	vv§	b‖	Liver fibrosis¶	Note
RK067	89	M	HCV	cHCC/CC	T2N0M0	25	−	−	−	3	
RK069	85	F	(−)	cHCC/CC	T3N0M0	80	+	+	−	3	
RK073	62	M	HBV	ICC	T1N0M0	10	−	−	−	4	HCC MCT+#
RK084	67	F	HCV	cHCC/CC	T3N0M0	30	+	−	−	4	
RK108	74	M	HCV	cHCC/CC	T1N0M0	12	−	−	−	2	
RK109	84	M	HCV	CoCC	T2N0M0	16	−	+	−	1	
RK112	83	M	HBV	cHCC/CC	T3N0M0	25	−	−	−	1	
RK137	74	F	HCV	ICC	T3N0M0	24	−	−	+	3	HCC MCT+
RK138	75	F	(−)	ICC	T3N0M0	120	+	+	+	**0**	Hepatitis-negative**
RK142	57	M	(−)	ICC	T1N0M0	15	−	−	+	3	
RK146	57	F	HBV	ICC	T2N0M0	23	+	−	−	3	
RK166	67	M	HBV	cHCC/CC	T3N0M0	25	−	−	+	4	
RK182	65	M	HCV	ICC	T3N0M0	28	+	−	−	3	
RK184	64	M	(−)	cHCC/CC	T3N0M0	110	+	−	−	1	
RK194	67	M	(−)	ICC	T3N0M0	35	+	−	+	**0**	Hepatitis-negative
RK204	83	M	HCV	ICC	T1N0M0	12	−	−	−	4	HCC MCT+(RK209)
RK208	60	M	(−)	ICC	T3N0M0	60	−	−	+	2	
RK226	59	M	HBV	ICC	T3N0M0	45	+	+	−	2	
RK269	74	M	(−)	ICC	T1N0M0	12	−	−	−	**0**	Hepatitis-negative
RK272	78	F	(−)	ICC	T3N0M0	45	+	+	−	**0**	Hepatitis-negative
RK279	69	M	HCV	ICC	T3N0M0	35	+	+	+	3	
RK298	68	M	HCV	ICC	T3N0M0	40	+	−	−	1	
RK303	76	M	(−)	CoCC	T2N0M0	20	+	−	−	2	
RK307	61	F	(−)	ICC	T3N1M0	75	+	−	+	**0**	Hepatitis-negative
RK308	70	F	(−)	ICC	T1N0M0	14	+	+	+	**0**	Hepatitis-negative
RK309	56	M	(−)	ICC	T2N1M0	36	−	−	+	**0**	Hepatitis-negative
RK310	62	F	(−)	ICC	T3N0M0	90	+	+	+	**0**	Hepatitis-negative
RK312	66	M	HBV	ICC	T3N1M0	48	+	+	−	**0**	Hepatitis-negative
RK316	54	F	(−)	ICC	T3N3M0	54	+	+	−	**0**	Hepatitis-negative
RK317	73	M	HBV	ICC	T3N0M0	45	+	+	−	2	

*ICC (intrahepatic cholangiocellular carcinoma), CoCC (cholangiolocellular carcinoma), cHCC/CC (combined hepatocellular cholangiocellular carcinoma).
†TNM staging in UICC.
‡vp; portal vein invasion.
§vv; hepatic vein invasion.
‖b; bile duct invasion.
¶Fibrosis in non-cancerous liver tissue is determined according the New Inuyama Classification.
#MCT; multicentric tumour.
**Liver fibrosis 0 indicates hepatitis-negative LCB.

morphological classification, whereas the somatic substitution pattern in the LCBs is strongly determined by the aetiological background of chronic hepatitis.

Recurrently mutated genes and pathway analysis. We examined recurrently mutated genes in our LCB samples. RK308 had an exceptionally large number of mutations and was excluded from the subsequent analyses. Across the 29 LCB genomes, we detected 892 protein-altering mutations, including 760 nonsynonymous, 108 short coding indels and 24 splice-site mutations (Supplementary Table 3). Thirty-two genes were recurrently mutated (Fig. 3 and Supplementary Table 5): cytoskeleton genes (*XIRP2, KIF2B* and *MYO10*), a cell adhesion molecule (*CDH2*), known tumour suppressors (*TP53, PTEN* and *BAP1*), known oncogenes (*KRAS* and *PIK3CA*), chromatin regulators (*PBRM1, ARID1A* and *ARID2*), which are highly mutated in HCC and other cancers[10,12], neuron growth genes (*ODZ1, EPHA2* and *PLCO*) and tyrosine kinase receptors (*ERBB4* and *EPHA2*). To validate the frequency of the mutations, the protein-coding exonic regions of recurrently mutated genes (*BAP1, CDH2, EPHA2, KIF2B, MGAT4C, ODZ1, PBRM1, PCLO, SYT1, ARID2* and *XIRP2*) were amplified in an additional 68 LCB samples (Supplementary Table 6) and sequenced by the Illumina HiSeq2000 sequencer. In addition, as *KRAS, IDH1, IDH2* and *TERT* promoter mutations were frequently observed in ICCs[28], fluke-related ICCs[18] and HCCs[29], we sequenced exons 2 and 3 of *KRAS*, exon 4 of *IDH1* (codon 132), exon 4 of *IDH2* (codon 172) and *TERT* promoter hotspots (chr5:1,295,228 and 1,295,250) in the additional 68 samples by Sanger sequencing. *TERT* promoter hotspots were also examined in the 30 WGS samples by Sanger sequencing owing to a low depth of coverage in the WGS. This validation experiment revealed that one *TERT* promoter hotspot (chr5:1,295,228) was mutated in 14 samples (15.2%), *KRAS* and *PBRM1* in 7 samples (10.3%), *ARID2* in 5 (7.4%), *BAP1, PCLO* and *IDH1* in 4 (5.9%), *ODZ1* in three (4.4%) and *EPHA2, SYT2, CDH2, XIRP2* and *IDH2* in two samples (2.9%) (Fig. 3 and Supplementary Table 7). In the *ARID2, PBRM1* and *BAP1* genes, which encode chromatin regulators, an accumulation of loss-of-function mutations was observed, suggesting that they are likely to function as tumour suppressors in LCBs as well as HCCs[10,12]. As observed in previous HCC studies[10,12,13], more than half of LCBs had somatic mutations and rearrangements accumulated in chromatin regulators (Supplementary Fig. 8).

We then examined the frequency of gene mutation and its association with clinical information in the WGS and validation samples (Table 2 and Supplementary Table 8). The frequency of the *TERT* promoter hotspot mutations was significantly lower in LCBs than in HCCs (Fisher's exact test P-value $= 1.2 \times 10^{-9}$) (Table 2). Furthermore, the frequency was significantly higher in cHCC/CCs and HCCs than in the ICCs (Fisher's exact test ICCs versus cHCC/CC; P-value $= 6.5 \times 10^{-5}$ and ICCs versus HCC; P-value $= 2.1 \times 10^{-11}$), but no significant difference was observed between cHCC/CCs and HCCs (Table 2). The frequency

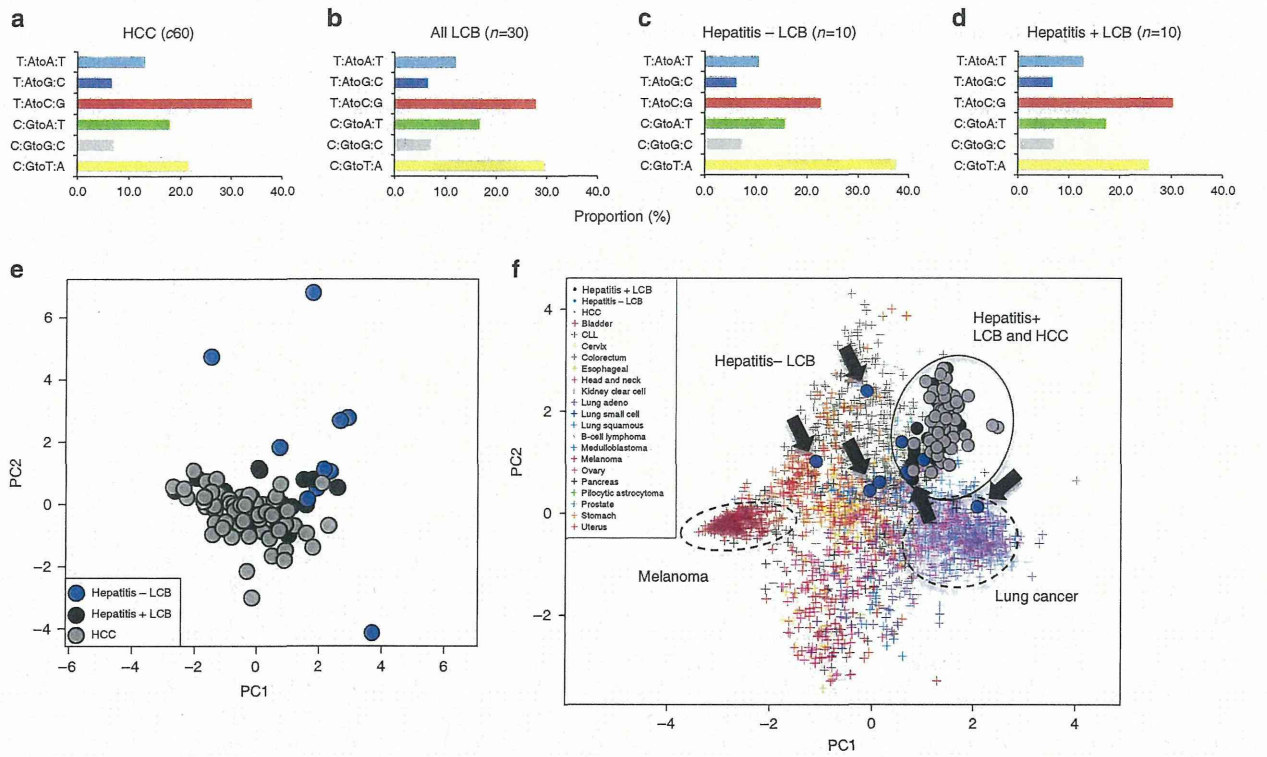

Figure 2 | Genome-wide substitution pattern on the 30 LCBs and 60 HCCs. Average proportion of somatic nucleotide substitutions for (**a**) the 60 HCCs, (**b**) the 30 LCBs, (**c**) the 10 hepatitis-negative LCBs and (**d**) the 20 hepatitis-positive LCBs. (**e**) Principal component analysis (PCA) of the whole-genome substitution patterns of the 30 LCBs and 60 HCCs. Hepatitis-positive LCBs (black dots) overlap the HCC cluster (gray). LCBs developed in livers without hepatitis (blue dots) diverged from others. (**f**) PCA of the whole-genome substitution patterns of the 30 LCBs, the 60 HCCs and other types of cancers[21]. Hepatitis-positive liver cancers (HCC and LCBs) are tightly clustered, as are melanomas, indicating that chronic hepatitis or inflammation can strongly impact the somatic mutation signature.

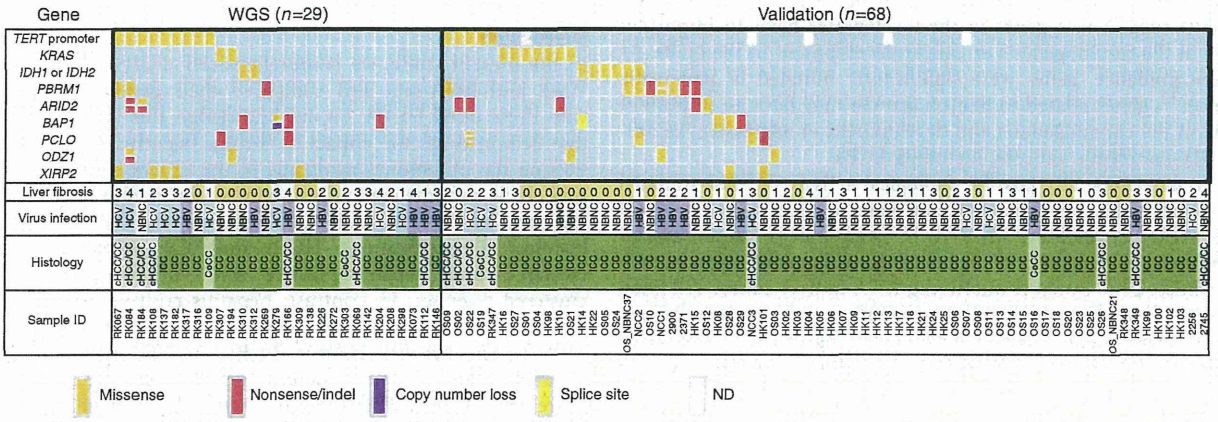

Figure 3 | Analysis of recurrently mutated genes in LCBs. Recurrently mutated genes in the 30 LCB WGS set and the 68 validation LCB set. Histological subtype of LCB: ICC; intrahepatic cholangiocarcinoma, cHCC/CC; combined hepatocellular-cholangiocarcinoma and CoCC; cholangiolocellular carcinoma. Liver fibrosis was determined according to the New Inuyama Classification (0～4). Liver fibrosis 0 indicates hepatitis-negative LCB. Virus infection status is classified as HBV, HCV or negative, which was determined by serological study.

of mutations in the *KRAS* and *IDH* gene hotspots and *PBRM1* was higher in the LCBs than in the HCCs (not significant after the Bonferroni correction) and no mutations were observed in the 60 HCCs.

When taking into account the presence of hepatitis, we found significant difference in the frequency of the *TERT* promoter hotspot mutations among the hepatitis-positive LCBs, the hepatitis-negative LCBs and the HCCs. The HCCs and hepatitis-positive LCBs had a higher frequency of the *TERT* promoter mutations than the hepatitis-negative LCBs (Table 2). In contrast, HCCs had significantly lower frequency of *KRAS* and *IDH* gene hotspot mutations than the hepatitis-negative LCBs, all of which are ICCs (Fisher's exact test *KRAS*; *P*-value = 0.0006, *IDH* genes; *P*-value = 0.0006). Hepatitis-positive LCBs shared more mutated genes with HCCs, which is consistent with the substitution pattern shown in Fig. 2f. However, the frequency of

Table 2 | Summary of mutations in WGS and validation set of LCBs.

	LCBs		HCCs	Hepatitis		Histology		Unadjusted *P*-value for comparison						
	WGS ($n=29$)	Validation ($n=68$)	WGS ($n=60$)	Hepatitis-positive LCBs ($n=62$)	Hepatitis-negative LCBs ($n=35$)	ICC ($n=82$)	cHCC/CC ($n=15$)	LCB versus HCC	Hepatitis-positive LCB versus hepatitis-negative LCB	Hepatitis-negative LCB versus HCC	Hepatitis-positive LCB versus HCC	ICC versus cHCC/CC	ICC versus HCC	cHCC/CC versus HCC
TERT Promoter	9 (30%)	5 (10%)	38 (63%)	12 (20%)	2 (6%)	4 (8%)	8 (53%)	1.2×10^{-9}*	NS	1.8×10^{-8}*	2.8×10^{-6}*	6.5×10^{-5}*	2.1×10^{-12}*	NS
KRAS	2 (7%)	7 (10%)	0 (0%)	2 (3%)	7 (20%)	9 (11%)	0 (0%)	0.013	0.010	0.00061*	NS	NS	0.010	NS
PBRM1	3 (10%)	7 (10%)	0 (0%)	8 (13%)	2 (6%)	7 (9%)	3 (20%)	0.014	NS	NS	0.0062	NS	0.021	0.0067
IDH1/IDH2	2 (7%)	6 (9%)	0 (0%)	1 (2%)	7 (20%)	8 (10%)	0 (0%)	0.024	0.0030*	0.00061*	NS	NS	0.021	NS
ARID2	2 (7%)	5 (7%)	3 (5%)	4 (6%)	3 (9%)	3 (4%)	4 (27%)	NS	NS	NS	NS	0.010	NS	0.026
BAP1	2 (7%)	4 (6%)	1 (2%)	3 (5%)	3 (9%)	6 (7%)	0 (0%)	NS	NS	NS	NS	NS	NS	NS
PCLO	2 (7%)	4 (6%)	4 (7%)	4 (6%)	2 (6%)	3 (4%)	3 (20%)	NS	NS	NS	NS	NS	NS	NS
ODZ1	2 (7%)	3 (4%)	0 (0%)	3 (5%)	2 (6%)	4 (5%)	1 (7%)	NS	NS	NS	NS	NS	NS	NS
XIRP2	5 (17%)	1 (1%)	5 (8%)	4 (6%)	2 (6%)	4 (5%)	2 (13%)	NS	NS	NS	NS	NS	NS	NS

*Significant after the Bonferroni correction.
RK308 had an exceptionally large number of mutations and was excluded from this analysis. Number and frequency (%) of mutations are shown. For *TERT* promoter, *KRAS* and *IDH* genes, hotspot mutations were counted. *TERT* promoter hotspots (chr5:1,295,228 and 1,295,250) were examined by Sanger sequencing method. CoCCs ($n=4$) were included in ICCs. *P*-values were obtained by the Fisher's exact test.

mutations in the *PBRM1* gene was different between hepatitis-positive LCBs and HCCs, which may be related to cell differentiation in the liver cancer development[30] (marginal significance after the Bonferroni correction).

Mutations in the *KRAS* gene were significantly enriched in patients with lymph node metastasis (Supplementary Table 8), and mutations in the *BAP1* gene were significantly enriched in patients with vascular or bile duct invasion (Supplementary Table 8). Mutations in *IDH* genes were associated with poor disease-free survival after adjustment for age, which is consistent with the previous study[20] (Supplementary Fig. 10).

To identify gene sets and pathways related to the LCB development, we carried out gene set enrichment analysis for all nonsynonymous mutations, short indels and rearrangements[31]. After adjustment for the multiple testing, 36 categories including 'synapse organization' and 'cytoskeleton' were significantly overrepresented (Supplementary Table 9). These results suggest that genes in these categories have an important role in the carcinogenesis or cancer development in the LCBs. As 'axon guidance' genes were significantly mutated in pancreatic cancer[32], genes related to neuron growth may have an important role in the carcinogenesis and development of aggressive hepato-biliary-pancreatic cancers, including LCBs.

To determine any possible biological activity of these mutated genes in LCBs, we examined four genes (*PCLO*, *EPHA2*, *ODZ1* and *XIRP2*), which are involved in synapse organization and/or cytoskeleton structure. We knocked down the expression of each candidate gene by short interfering RNA in liver cancer cell lines, and examined their proliferation, migration and invasion abilities. These experiments confirmed that silencing of *PCLO* promoted cell invasion in liver cancer cell lines (Supplementary Figs 11 and 12).

Genetic heterogeneity within liver tumour. A tumour is a population of heterogeneous cancer cells, and the analysis of this heterogeneity should provide us with deeper insights into tumorigenesis[12,33–36]. To examine intratumour heterogeneity, the clonal proportion of the 1,085 nonsynonymous point mutations and short indels, detected in randomly selected 15 LCBs and 10 HCCs, were sequenced to an average depth of 56,462x by ultra-deep sequencing (Supplementary Methods). Copy number alternations were adjusted for mutant-allele frequencies using allelic imbalance ratio, and proportion of mutated allele (PMA) was obtained. The distribution of PMA in the ICCs and the cHCC/CCs significantly differed (Wilcoxon's test *P*-value = 0.0047), and ICC genomes had a larger number of mutations with higher PMA (Supplementary Fig. 13). One possible explanation is that the pattern of genetic heterogeneity is different between cHCC/CC and ICC, and cHCC/CC had larger number of shared mutations in the tumour population than ICC, which is consistent with the histological diversity of cHCC/CC, showing mixed components of both hepatocellular and biliary epithelial differentiation. We then examined the clonal proportion for each gene (Supplementary Dataset 2). Genes in 15 categories, including 'replicative senescence' and 'negative regulation of DNA replication' had a higher PMAs after adjustment for multiple testing (Supplementary Table 10). All categories contained the *TP53* gene, indicating that the *TP53* gene conferred clonal advantage to cancer cells. This result is consistent with a breast cancer study[33]. Various genes showed high frequency of mutations in each tumour and would be candidates for tumour initiators (Supplementary Dataset 2).

Discussion

In the present study, we comprehensively analyzed 30 LCBs by WGS and RNA-seq, and compared their genomic landscapes with those of 60 HCCs. To our knowledge, this is the first study that demonstrates the impact of chronic hepatitis and inflammation on the mutational landscape of the cancer genome and the first whole-genome comparison between LCBs and HCCs. In our analysis, gene expression patterns are consistent with the histological classifications; HCCs and LCBs were differentially clustered and hepatitis-positive and -negative LCBs were clustered together. In contrast, hepatitis-positive and -negative LCBs differentiated in their genome-wide somatic substitution pattern. The hepatitis-positive LCBs clustered more tightly to hepatitis-related HCCs, whereas hepatitis-negative LCBs were more spread out. These results suggest that gene expression depends on the histological phenotype, whereas the somatic substitution pattern is strongly influenced by aetiological background like the occurrence of chronic hepatitis. Previous studies suggested that the expression pattern is consistent with pathological phenotype, but does not reflect tumour origin[37–39]. A mouse study on pancreatic ductal adenocarcinoma suggested that inflammation can promote neoplasia by altering cell differentiation[38], and a comparison between virus-associated ICCs and HCCs suggested that they can arise from the hepatic progenitor cells[5]. Considering these studies, the similarities between the hepatitis-positive LCBs and the HCCs in the somatic substitution pattern may indicate their same cellular origin, such as liver progenitor cell. In contrast, hepatitis-negative LCB may arise from different origins, such as cholangiocytes[40].

The frequency of some driver mutations, such as hotspot mutations in *KRAS, IDH1/2* and *TERT* promoter, differed among cancer types, hepatitis-positive and -negative LCBs. Mutations of *KRAS* and *IDH* genes were more frequent in the hepatitis-negative LCBs, and the *TERT* promoter mutation was more frequent in the cHCC/CCs and HCCs. As almost all cHCC/CC and HCCs were hepatitis-positive, it is difficult to differentiate the impact of hepatitis from that of the histology. In general, HCC and cHCC/CC, which mainly developed under a hepatitis background, had a larger frequency of *TERT* promoter mutations and a lower frequency of *KRAS* and *IDH1/2* mutations.

In the current study, we found that the occurrence of chronic hepatitis impacted the mutational landscape, discovered new driver gene and examined intratumour heterogeneity in the LCBs. Our analysis indicates that the WGS can reveal the impact of aetiological background on the genome-wide substitution pattern, and suggest that the WGS can contribute to molecular classification based on their aetiology. However, we did not find mutations in the driver gene candidates in about a half of the samples, suggesting that LCB is a highly heterogeneous cancer. Analysis of larger number of samples would be necessary for deeper understanding of LCB.

Methods

Clinical samples. The clinical and pathological features of 30 LCBs that were used in WGS analysis are in Table 1. Our pathologists evaluated hematoxylin and eosin-stained slides and diagnosed HCC, ICC and cHCC/CC according to the 2010 WHO Classification of Tumors of the Digestive System[41]. We defined ICC and cHCC/CC, both of which contain varying degrees of epithelial tubular-differentiated cells (Fig. 1a), as liver cancer displaying biliary phenotype (LCB), distinguishing them from the hepatocellular phenotype (HCC). Viral infection was defined by the presence of HB surface antigen in patient's serum, or by the presence of antibody to HCV in patient's serum. Hepatitis-negative LCB was defined as a tumour showing no sign of chronic inflammation and liver fibrosis, which was determined according the New Inuyama Classification. All subjects had undergone partial hepatectomy, and pathologists estimated the ratio of viable tumour cells in each sample. High molecular weight genomic DNA was extracted from fresh-frozen tumour specimens and blood. Non-cancerous liver tissues were used as the normal tissue for RK182, RK307, RK308, RK309 and RK310. All subjects agreed with informed consent to participate in the study following ICGC guidelines[42]. IRBs at RIKEN and all groups participating in this study approved this work.

Whole-genome sequencing. DNA was extracted from tumours and non-cancer frozen tissues, and 500 bp insert libraries were prepared according to the protocol provided by Illumina. The libraries were sequenced on HiSeq2000 platforms with paired reads of 101 bp. The mutation data for the 60 HCCs have been generated in the same way by RIKEN and deposited to the ICGC dataset version 8 released at 2012 March (http://icgc.org/).

Somatic mutation and short indel calling. Point mutations and somatic indels were identified using our in-house methods[12]. In brief, read pairs were mapped by BWA[43], and the result files were converted to pileup file by samtools[44]. After PCR duplications were removed and comparing between cancer genome sequences and non-cancer genome sequences, somatic point mutations and indels were identified by our in-house mutation caller[12]. False-negative and false-positive rates of our analysis pipeline were described previously[12]. Information for all point mutations and indels in the 30 LCBs and the 60 HCCs was deposited to the ICGC web site (http://www.icgc.org/).

Identification of rearrangements. Inconsistent read pairs which occurred within 500 bp of each other were considered to support the same rearrangement. We identified candidate rearrangements in both tumour (support read pairs ≥ 4) and normal tissue (support read pairs ≥ 1) samples, and tumour-specific rearrangement candidates were identified. To exclude mapping errors, we performed a blast search of read pairs that support rearrangements against the reference genome. If a read pair mapped with correct orientation and distance (≤ 500 bp) with an *E*-value $< 10^{-7}$, we excluded that read pair. Reads mapped with more than two mismatches were also discarded. After filtering, candidates supported by ≥ 4 read pairs and at least one perfect match pair were considered as somatic rearrangements. The candidates that the same rearrangement was found in other normal samples were filtered out. False discovery rate of this method was estimated to be 2.3% (4/176).

Statistical analysis. The random distribution was calculated by multiplying (proportion of nucleotide in the reference genome sequence) and (total number of mutations) as done in the previous study[12]. Tests for significantly mutated genes and PCA of the substitution pattern were carried out as described previously[12].

Survival analyses were done using the 'survival' package for the R programming environment (http://www.r-project.org). A Cox proportional hazards model was used to test association between disease-free survival and mutations in the genes (*TERT* promoter, *KRAS*, *XIRP2*, *ARID2*, *BAP1*, *PBRM1*, *PCLO*, *ODZ1* and *IDH* genes) and clinical factors (age, gender, virus type and liver fibrosis). Model selection was done by the stepAIC function, and the model with age and mutations in *IDH* genes was selected.

Estimation of PMA was described in the Supplementary Methods.

To test the difference of the clonal proportion of mutations among ICCs, cHCC/CCs and HCCs, we calculated PMA for each mutation, which was standardized by the maximum PMA in each sample. Then we compared the median of the distribution of PMA between ICCs, cHCC/CCs and HCCs by Wilcoxon's test.

To identify gene sets with high clonal proportion, we used 'biological process' terms with depth = 5 in the Gene Ontology (GO) database (http://www.geneontology.org). The clonal proportions of the genes within and outside the gene category were compared by Wilcoxon's test as a previous study[33]. Note that we used unadjusted clonal proportions (not PMAs) for this analysis to consider the influence of copy number changes.

Sanger sequencing and ultra-deep amplicon sequencing. Sanger sequencing of PCR products was performed on ABI 3770x. For ultra-deep sequencing of mutations, each of the 100 bp target regions was amplified and the amplicons were directly ligated with Illumina TruSeq adaptors and sequenced on HiSeq2000 platform. Mapping was done by BWA to the target region, and uniquely mapped read pairs with proper distance and orientation were selected. More than 98% of the exonic target regions were covered with a depth ≥ 100. We filtered out reads with a mapping quality < 10 and base calls with base quality < 10. Base calls with a depth ≥ 100 were used for the analysis. We identified variants with frequency ≥ 0.05. Variants found in more than one individual in the 1000 Genome database[45] were discarded. We performed Sanger sequencing verification for the predicted candidates in the both cancer and matched normal tissues.

RNA sequencing. RNA-seq was carried out for 25 LCBs and 44 HCCs for which high-quality RNA was available among the 30 LCBs and the 60 HCCs. Total RNA was extracted by Trizol from the frozen liver cancer tissues and the corresponding non-cancerous liver tissues and quality and quantity were evaluated by Bioanalyzer (Agilent). The high-quality RNA was subjected to polyA + selection and chemical fragmentation, and 100–200 base RNA fraction was used to construct complementary DNA libraries according to Illumina's protocol. RNA-seq was performed on HiSeq2000 using the standard paired-end 101 bp protocol.

Analysis of RNA sequencing data. First, all sequencing reads were aligned to the known transcript sequences of UCSC known gene database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz) using Bowtie[46], with -a --best --strata -m 20 -v 3 options, and the coordinates of the aligned reads were converted to the human reference sequence (hg19). Then, reads unaligned in the above step were aligned to the human reference sequence (hg19) and as well as HBV sequence (AP011098) using Blat[47], with -stepSize = 5 -repMatch = 2253, and aligned reads by Bowtie or Blat were combined together. For each short read, the alignment positions with the maximum number of matched bases were adopted, and mapping quality for each read was assigned to as follows: for a location a, let $B(a)$ denote the number of matched bases and let a_{best} denote the best location selected arbitrarily from those with the maximum number of matched bases.

$$\min\left(100 - 10\times\log10\left(1 - \frac{1}{\sum_{a} 0.02^{B(a_{\mathrm{best}}) - B(a)}}\right)\right)$$

Finally, sorting and PCR duplicate removal of short reads were performed by using Picard (http://picard.sourceforge.net/). For quantification of expression values, we used a slightly modified version of RKPM (reads per kilobase of exon per million mapped reads) measures[48]. After removing improperly aligned or low-quality (mapping quality < 60) sequencing reads, the number of bases on each exonic region for each refSeq genes (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz) were counted. Then, the numbers of bases were normalized as per kilobase of exon and per 100 million of aligned bases. Finally, the expression value of each gene was determined by choosing the maximum of multiple refSeq genes, if any, corresponding to the gene symbol.

References

1. Blechacz, B. & Gores, G. J. Cholangiocarcinoma: advances in pathogenesis, diagnosis, and treatment. *Hepatology* **48,** 308–321 (2008).
2. Shin, H. R. *et al.* Epidemiology of cholangiocarcinoma: an update focusing on risk factors. *Cancer Sci.* **101,** 579–585 (2010).

3. Sia, D., Tovar, V., Moeini, A. & Llovet, J. M. Intrahepatic cholangiocarcinoma: pathogenesis and rationale for molecular therapies. *Oncogene* **32,** 4861–4870 (2013).
4. Sripa, B. *et al.* Liver fluke induces cholangiocarcinoma. *PLoS Med.* **4,** e201 (2007).
5. Lee, C. H. *et al.* Viral hepatitis-associated intrahepatic cholangiocarcinoma shares common disease processes with hepatocellular carcinoma. *Br. J. Cancer* **100,** 1765–1770 (2009).
6. Sekiya, S. & Suzuki, A. Intrahepatic cholangiocarcinoma can arise from Notch-mediated conversion of hepatocytes. *J. Clin. Invest.* **122,** 3914–3918 (2012).
7. Xu, J. *et al.* Intrahepatic cholangiocarcinoma arising in chronic advanced liver disease and the cholangiocarcinomatous component of hepatocellular cholangiocarcinoma share common phenotypes and cholangiocarcinogenesis. *Histopathology* **59,** 1090–1099 (2011).
8. Coulouarn, C. *et al.* Combined hepatocellular-cholangiocarcinomas exhibit progenitor features and activation of Wnt and TGFbeta signaling pathways. *Carcinogenesis* **33,** 1791–1796 (2012).
9. Guest, R. V. *et al.* Cell lineage tracing reveals a biliary origin of intrahepatic cholangiocarcinoma. *Cancer Res.* **74,** 1005–1010 (2013).
10. Li, M. *et al.* Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.* **43,** 828–829 (2011).
11. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.* **43,** 464–469 (2011).
12. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44,** 760–764 (2012).
13. Guichard, C. *et al.* Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.* **44,** 694–698 (2012).
14. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22,** 593–601 (2012).
15. Sung, W. K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44,** 765–769 (2012).
16. Huang, J. *et al.* Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat. Genet.* **44,** 1117–1121 (2012).
17. Kan, Z. *et al.* Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.* **23,** 1422–1433 (2013).
18. Ong, C. K. *et al.* Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nat. Genet.* **44,** 690–693 (2012).
19. Chan-On, W. *et al.* Exome sequencing identifies distinct mutational patterns in liver fluke-related and non-infection-related bile duct cancers. *Nat. Genet.* **45,** 1474–1478 (2013).
20. Jiao, Y. *et al.* Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. *Nat. Genet.* **45,** 1470–1473 (2013).
21. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
22. Gammie, A. E. *et al.* Functional characterization of pathogenic human MSH2 missense mutations in Saccharomyces cerevisiae. *Genetics* **177,** 707–721 (2007).
23. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
24. Raimondo, G., Pollicino, T., Cacciola, I. & Squadrito, G. Occult hepatitis B virus infection. *J. Hepatol.* **46,** 160–170 (2007).
25. Tsai, W. L. & Chung, R. T. Viral hepatocarcinogenesis. *Oncogene* **29,** 2309–2324 (2010).
26. Komuta, M. *et al.* Clinicopathological study on cholagiolocellular carcinoma suggesting hepatic progenitor cell origin. *Hepatology* **47,** 1544–1566 (2008).
27. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14,** R39 (2013).
28. Borger, D. R. *et al.* Frequent mutation of isocitrate dehydrogenase (IDH)1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping. *Oncologist* **17,** 72–79 (2011).
29. Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat. Commun.* **4,** 2218 (2013).
30. Romero, O. A. & Sanchez-Cespedes, M. The SWI/SNF genetic blockade: effects in cell differentiation, cancer and developmental diseases. *Oncogene* **33,** 2681–2689 (2014).
31. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
32. Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491,** 399–405 (2012).
33. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
34. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
35. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481,** 506–510 (2012).
36. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).
37. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15,** 907–913 (2009).
38. Gidekel Friedlander, S. Y. *et al.* Context-dependent transformation of adult pancreatic cells by oncogenic K-Ras. *Cancer Cell* **16,** 379–389 (2009).
39. Goldstein, A. S., Huang, J., Guo, C., Garraway, I. P. & Witte, O. N. Identification of a cell of origin for human prostate cancer. *Science* **329,** 568–571 (2010).
40. Komuta, M. *et al.* Histological diversity in cholangiocellular carcinoma reflects the different cholangiocyte phenotypes. *Hepatology* **55,** 1876–1888 (2012).
41. Bosman, F. T. *et al. WHO Classification of Tumors of the Digestive System* 4th edn, 225–227 (IARC, 2010).
42. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
45. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
46. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
47. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
48. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

Acknowledgements

The super-computing resource 'SHIROKANE' was provided by Human Genome Center, The University of Tokyo (http://sc.hgc.jp/shirokane.html). This work was supported partially by RIKEN President's Fund 2011, the Princess Takamatsu Cancer Research Fund and Takeda Science Foundation.

Author contributions

A.F., Y.Shiraishi, M.F., D.S., T.A., K.A.B., T.T. and H.N. performed data analyses. M.F., K.N., A.S., R.K. and H.N. performed WGS and validation sequencing study. M.F. A.F. and H.H.N. performed functional experiments. H.T., T.Shibuya and S.Miyano operated the super-computer system. Y.K., M.U., K.G., S.A., T.N., T.Shibata, K.A., H.O., K.S., Y.Shigekawa, S.Maruhashi, T.Y., O.I., H.A., H.O., S.H., M.Y., H.Y. and K.C. collected clinical samples and cell lines. A.F., M.F., Y.Shiraishi, K.A.B. and H.N. wrote the manuscript. H.N. conceived the study and led the design of the experiments. A.F., M.K. and H.N. contributed to the funding for this study.

Additional information

Accession codes: WGS data have been deposited in the EGA under the accession codes: EGAN00001187542, EGAN00001187543, EGAN00001187546, EGAN00001187547, EGAN00001187552, EGAN00001187553, EGAN00001187568, EGAN00001187569, EGAN00001187612, EGAN00001187613, EGAN00001187614, EGAN00001187615, EGAN00001187618, EGAN00001187619, EGAN00001187650, EGAN00001187651, EGAN00001187652, EGAN00001187653, EGAN00001187658, EGAN00001187659, EGAN00001187664, EGAN00001187665, EGAN00001187690, EGAN00001187691, EGAN00001187707, EGAN00001187708, EGAN00001187709, EGAN00001187710, EGAN00001187712, EGAN00001187713, EGAN00001187714, EGAN00001187715, EGAN00001187716, EGAN00001187717, EGAN00001187719, EGAN00001187720, EGAN00001187721, EGAN00001187722, EGAN00001187723, EGAN00001187724, EGAN00001187725, EGAN00001187726, EGAN00001187727, EGAN00001187728, EGAN00001187729, EGAN00001187730, EGAN00001187731, EGAN00001187732, EGAN00001187733, EGAN00001187734, EGAN00001187735, EGAN00001187736, EGAN00001187737, EGAN00001187738, EGAN00001187739, EGAN00001187740, EGAN00001187741, EGAN00001187742, EGAN00001187744 and EGAN00001187743 (summarized in Supplementary Table 2).

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interest.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Fujimoto, A. *et al.* Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. *Nat. Commun.* 6:6120 doi: 10.1038/ncomms7120 (2015).

Exploration of liver cancer genomes

Tatsuhiro Shibata and Hiroyuki Aburatani

Abstract | Liver cancer is the third leading cause of cancer-related death worldwide. Advances in sequencing technologies have enabled the examination of liver cancer genomes at high resolution; somatic mutations, structural alterations, HBV integration, RNA editing and retrotransposon changes have been comprehensively identified. Furthermore, integrated analyses of trans-omics data (genome, transcriptome and methylome data) have identified multiple critical genes and pathways implicated in hepatocarcinogenesis. These analyses have uncovered potential therapeutic targets, including growth factor signalling, WNT signalling, the NFE2L2-mediated oxidative pathway and chromatin modifying factors, and paved the way for new molecular classifications for clinical application. The aetiological factors associated with liver cancer are well understood; however, their effects on the accumulation of somatic changes and the influence of ethnic variation in risk factors still remain unknown. The international collaborations of cancer genome sequencing projects are expected to contribute to an improved understanding of risk evaluation, diagnosis and therapy for this cancer.

Shibata, T. & Aburatani, H. *Nat. Rev. Gastroenterol. Hepatol.* advance online publication 28 January 2014; doi:10.1038/nrgastro.2014.6

Introduction

Liver cancer is the third leading cause of cancer-related death worldwide.[1] Hepatocellular carcinoma (HCC) is the most common form of liver cancer, followed by intrahepatic cholangiocarcinoma (IHCC).[1] Chronic liver damage, such as that caused by chronic hepatitis, liver cirrhosis and fatty liver disease, is closely associated with the occurrence of liver cancers. Hepatitis virus infection (for example HBV, HCV and others), aflatoxin B exposure, alcohol intake, and other metabolic diseases (such as obesity, diabetes mellitus and haemochromatosis) are well-known risk factors for liver cancer.[2–4] In addition, parasites such as liver fluke are associated with IHCC in Southeast Asian countries.[5,6]

The incidence of liver cancer is high in East Asian and African countries.[1–3,5] HBV infection is more prevalent in Africa and Asian countries (except Japan) than other regions of the world.[3] However, the number of patients infected with HCV has been rapidly increasing in Japan and Western countries, especially in the USA where viral hepatitis infection is partly mediated through drug abuse.[2,3] In this Review, we mainly focus on HCC, as HCC and IHCC showed distinctive genomic alterations and fairly little is known about the IHCC genome alterations at present.

Somatic alterations in the liver cancer genome

The liver cancer genome contains multiple types of somatic alterations, including mutations (such as single nucleotide substitutions, and small insertions and deletions), changes of gene copy numbers (copy number loss, gain and amplification), and intra-chromosomal and inter-chromosomal rearrangements (large deletion, inversion, tandem duplication and translocation).

Genome-wide copy number analysis

Copy number changes in human cancers have been analysed mainly by array-based comparative genome hybridization methods. Bacterial artificial chromosome (BAC) clone DNA or oligonucleotide probe arrays (microarray-based comparative genomic hybridization) have been used in a number of studies to search for copy number changes in liver cancer.[7–20] Table 1 summarizes recurrent copy number alterations in HCC. In addition to well-known oncogenes, such as *MYC* and *CCND1*, and tumour suppressor genes, such as *TP53* and *RB*, liver cancers harbour multiple chromosomal amplifications and deletions.

The identification of target genes solely by copy number data has been challenging. Therefore, strategies based on integrative analysis of genetic alterations, gene expression profiling and oncogenic function of candidate genes might be an effective approach. Zender *et al.*[21] selected potential tumour suppressor genes using data from copy number analyses of human HCC, and functionally identified novel tumour suppressor genes, including *XPO4*, by *in vivo* short hairpin RNA screening in a mosaic mouse model. Sawey *et al.*[22] extracted genes located in chromosomal regions of recurrent focal amplification in human HCC and tested their oncogenic activity using a mouse hepatoblast model. These authors identified 18 tumour-promoting genes, including *FGF19*, which is located next to the *CCND1* gene on 11q13.3. FGF19 and CCND1 cooperatively promote tumour formation through the CTNNB1 pathway.[22]

Katoh *et al.*[13] attempted to define a molecular classification of HCC on the basis of the copy number

Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Tokyo 104-0045, Japan (**T. Shibata**). Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan (**H. Aburatani**).

Correspondence to:
T. Shibata
tashibat@ncc.go.jp

Competing interests
The authors declare no competing interests.

Key points

- Whole-exome and whole-genome sequencing have provided a comprehensive and high-resolution view of somatic genomic alterations in liver cancer
- Global epigenetic analyses have further identified both unique and complementary molecular alterations in liver cancer
- Somatic mutational signatures of the liver cancer genome are complex and tend to be associated with epidemiological backgrounds
- Integration of genetic and epigenetic alteration profiles has elucidated core oncogenic pathways, potential therapeutic targets and new molecular classifications in liver cancer

Table 1 | Amplified and deleted genes in HCC

Gene name	Locus	Function
Recurrently amplified genes in HCC		
MDM4	1q32.1	p53 pathway
BCL9	1q21.1	WNT pathway
ARNT	1q21.2	Xenobiotics metabolism
ABL2	1q25.2	Proliferation
MET	7q31.2	Proliferation
COPS5	8q13.1	Proteolysis
MTDH	8q22.1	Metastasis
COX6C	8q22.2	Mitochondria
MYC	8q24.21	Proliferation
CCND1	11q13.2	Proliferation
FGF19	11q13.2	WNT pathway
RPS6KB1	17q23.1	Proliferation
EEF1A2	20q13.33	Translation
Recurrently deleted genes in HCC		
TNFRSF14	1p36.33	Immune response
CDKN2C	1p36.11	Cell cycle
ARID1A	1p36.11	Chromatin remodelling
TNFAIP3	6q26	NF-κB pathway
CSMD1	8p23.2	Immune response
DLC1	8p22	Small GTPase
SORBS3	8p21.3	Migration
WRN	8p21.3	DNA repair
SH2D4A	8p21.2	Proliferation
PROSC	8p11.2	Unknown
CDKN2A	9p21.3	Cell cycle
CDKN2B	9p21.3	Cell cycle
PTEN	10q23.31	Proliferation
SPRY2	13q31.1	Proliferation
BRCA2	13q13.1	DNA repair
RB1	13q14.3	Cell cycle
XPO4	13q11	Nuclear export
SMAD4	18q21.31	TGF-β signalling

alteration profiles of 87 HCC tumours, including HBV-associated and HCV-associated cases. Two molecular subgroups were identified that are associated with virus status, the presence of intrahepatic metastasis and patient prognosis. The researchers also reported six distinctive combinations of copy number alterations in HCC.[13] In another study, copy number changes in 63 HCCs of various aetiologies (viral and nonviral) were analysed and 8q24 copy number gains associated with *MYC* overexpression were identified that were unique to viral and alcohol-related HCCs.[15] Amplification of *MDM4* (1q32.1) and copy number gain of *EEF1A2* (20q13.33) were shown to be frequent and aetiology-independent molecular events in HCC.[15] A meta-analysis of four independent microarray comparative genomic hybridization datasets, including 169 samples, identified chromosomal gains in five broad (1q, 6p, 8q, 17q, and 10q) and two narrow (5p15.33 and 9q34.2–34.3) regions, and 88 significant losses frequently present in 4q, 6q, 8p, 9p, 13q, 14q, 16q and 17p.[18] Wang *et al.*[19] reported the results of copy number analysis of 286 HCC tumours by single nucleotide polymorphism array, which identified 29 recurrently amplified and 22 recurrently deleted regions, as well as *BCL9* and *MTDH* as novel amplified oncogenes in HCC.[19]

Whole-exome sequencing

Innovations in sequencing technologies have enabled researchers to explore the liver cancer genome in more depth. The capture or enrichment of DNA fragments containing the exonic region followed by massively parallel sequencing can determine somatic mutations in the whole exon domain (exome).[23,24] This approach enables the comprehensive detection of somatic alterations in the protein-coding region, and has led to the discovery of many novel genes implicated in liver cancer. Exomic sequencing of 10 HCV-positive HCCs and subsequent analysis of an additional tumour cohort of various aetiological backgrounds identified recurrent inactivating mutations of the *ARID2* gene in 18.2% of HCV-associated HCCs.[25] Guichard *et al.*[26] performed copy number analysis of 125 HCC cases and whole-exome sequencing of 24 of these cases and found new recurrent alterations in four genes (*ARID1A*, *RPS6KA3*, *NFE2L2* and *IRF2*). Huang *et al.*[27] performed whole-exome sequencing of nine pairs of HCCs and their intrahepatic metastases. Although most substitutions (94.2%) were common in both primary and metastatic tumours, a fraction of mutations were only detected in primary (1.1%) or metastatic (4.7%) tumours. Among them, *KDM6A*, *CUL9*, *FGD6*, *AKAP4* and *RNF139* were found only in the metastatic tumours of three individuals.

Using whole-exome sequencing of 87 HCC cases, Cleary *et al.*[28] identified recurrent alterations in the *NFE2L2–KEAP1* and *KMT2A* (also known as *MLL*) pathways, and other genes (*C16orf62* and *RAC2*) with lower mutation frequencies. Eight fluke-associated cholangiocarcinomas (the predominant type of liver cancer in northern Thailand and neighbouring countries) were analysed, and showed that the number of coding mutations per tumour ranged from 19 to 34, with an average of 26 mutations per sample.[29] In addition to *TP53* and *KRAS*, recurrent inactivating mutations in the *MLL3*, *ROBO2*, *RNF43* and *PEG3* genes were identified, and activating mutations were found in the *GNAS* gene.

Whole-genome sequencing

Several research groups have sequenced the full liver cancer genome in further attempts to identify all somatic driver events related to hepatocarcinogenesis, including substitutions in noncoding regions, structural rearrangements, and viral genome integration. Totoki *et al.*[30] first performed whole-genome sequencing of one HCV-associated HCC case (tumour genome and corresponding normal genome) and identified >16,000 somatic mutations and 26 intra-chromosomal and inter-chromosomal rearrangements generating four fusion transcripts. Among them, one in-frame fusion transcript (*BCORL1-ELF4*), generated by a small inversion on the X chromosome, showed reduced transcriptional repression activity compared to wild-type *BCORL1*, which encodes a tumour suppressor gene.

Fujimoto *et al.*[31] reported the results of whole-genome sequencing of 27 HCCs and matched normal genomes, 25 of which were associated with HBV or HCV infection. The average number of somatic point mutations at the whole-genome level was 4.2 per Mb. One tumour that contained an exceptionally large number of somatic mutations (24,147 substitutions) showed a DNA mismatch-repair defect caused by a somatic nonsense mutation in the *MLH1* gene. Furthermore, mutations in several chromatin regulators, including *ARID1A*, *ARID1B*, *ARID2*, *KMT2A* and *MLL3*, were detected in ~50% of the tumours.

Whole-genome sequencing of 88 HCC tumour and normal tissue pairs, including 81 HBV-positive and no HCV-positive cases showed an average somatic mutation rate of 3.69 per Mb and a mean protein-altering mutation rate of 1.8 per Mb, which are mid-range among different cancer types.[32] In this study, the WNT/CTNNB1 and JAK/STAT pathways were shown to be major oncogenic drivers in HCC and activating *JAK1* mutations were identified in 9.1% of total cases, suggesting that these pathways could be novel therapeutic targets in HCC.

HBV genome integrations in the host genome

HBV is a DNA virus whose genome is integrated into the host genome. The integration of the viral genome affects host gene expression near the integration site and its effect on the integrity of the host genome is associated with virus-mediated hapatocarcinogenesis.[33] In the past, Southern blot analysis or inverse PCR was applied to identify viral genome integration sites. However, current genome sequencing technology can detect virus integration events more comprehensively and at higher resolution than previously.

Jiang *et al.*[34] performed high-depth (>80× and 240× coverage of the genome, two or three times more than that used for conventional whole-genome sequencing) whole-genome and transcriptome sequencing of four pairs of HBV-positive HCCs and identified 225 HBV genome integration sites by taking advantage of paired reads mapping to both human and viral genomes. A variety of genomic aberrations near viral integration sites were found, including direct gene disruption, viral promoter-driven gene transcription, viral–human transcript fusion, and DNA copy number alterations. Frequent HBV integration in *TERT* and *MLL4* loci has also been reported.[31] Sung *et al.*[35] conducted whole-genome sequencing, at >30× coverage on average, of 81 HBV-positive and seven HBV-negative HCC samples. Analysis of HBV integration sites identified 399 integration breakpoints (4.9 per case). Frequent HBV integration breakpoints were observed in the *TERT*, *KMT2D* (also known as *MLL4*), *CCNE1* and *FN1* genes.

Somatic change of retrotransposons in HCC

The human genome contains a variety of repetitive sequences, including tandem repeats (such as satellite DNA and microsatellite DNA) and retrotransposons, such as short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs). In the human genome, Alu and LINE-1 are major forms of SINEs and LINEs, respectively. Given that current massive parallel sequencing technologies can produce only short reads (~200 bp), repetitive sequences, which constitute ~20% of the human genome, remain to be explored in genome sequencing.[36]

Retrotransposon capture sequencing applied to HCC samples revealed two LINE-1-mediated somatic changes associated with liver tumorigenesis.[37] One was a germline retrotransposon insertion in the *MCC* gene, a tumour suppressor gene that is known to be mutated in colorectal cancers. This retrotransposon insertion was found to downregulate *MCC* expression and activate the WNT/CTNNB1 pathway. The other event, a tumour-specific LINE-1 insertion, activates a potential oncogene, *ST18*, in liver tumours.

Mutation signatures and aetiological factors

There are six patterns of somatic substitution (C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G and T>G/A>C) in the cancer genome and they are affected by exogenous or endogenous mutagens, such as oxidative stress, exposure to chemicals or UV, and defects in the DNA repair machinery.[38] Whole-genome sequencing in cancer can identify large numbers of neutral mutations and is more appropriate for the analysis of mutation signatures in an unbiased manner than is whole-exome sequencing.

The first whole-genome sequencing study of a Japanese HCV-positive HCC case showed a distinct mutation signature (dominance in C>T/G>A and T>C/A>G) in the liver cancer genome.[30] A similar substitution pattern was also reported in Asian HBV-positive HCC cases.[32,33] Guichard *et al.*[26] reported the over-representation of C>A/G>T substitutions in HCC in a Western population with multiple aetiological backgrounds, although their data was obtained using whole-exome sequencing.[26] Using whole-genome sequencing, a study of 27 HCC cases of different aetiological backgrounds demonstrated a dominance of T>C/A>G transitions as well as C>A/G>T transversions and C>T/G>A transitions, particularly at CpG sites.[31] As C>T/G>A transitions are commonly found in other cancers, T>C/A>G transitions and C>A/G>T transversions could be characteristic mutational signatures

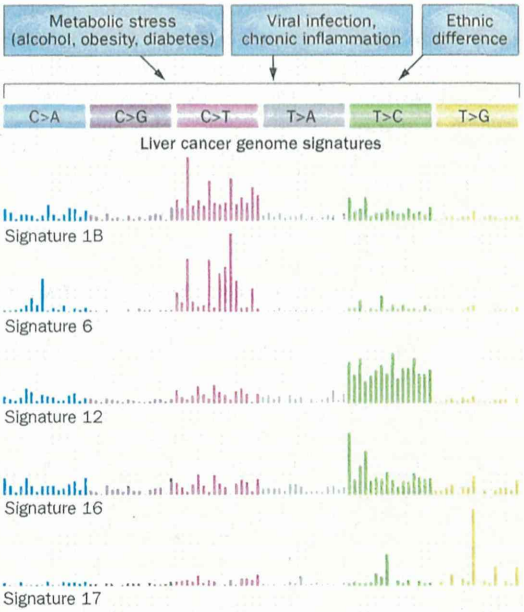

Figure 1 | Multiple aetiological factors and ethnic differences affect somatic mutation signatures in liver cancer. Five characteristic mutation signatures identified in the liver cancer genome are shown.[39] Permission obtained from Nature Publishing Group © Alexandrov, L. B. *Nature* **500**, 415–421 (2013).

of HCC genomes. Habitual alcohol drinking and the occurrence of synchronous or metachronous multiple liver nodules were significantly associated with the principal components of the somatic substitution patterns.[31] Somatic substitutions in IHCC associated with liver fluke are predominantly C>T/G>A transitions, the majority of which are identified in the context of a CpG-to-TpG change as the result of 5-methylcytosine deamination.[29]

In addition to six different substitutions, information on the bases immediately 5' and 3' to each mutation has been used to identify context-dependent mutation patterns in a wide range of cancers. Among 22 mutational signatures identified by this cross-tumour analysis, HCC contained five distinct signatures, which was the highest number among the 30 tumour types and indicates that a complex mutagenesis process operates in this tumour (Figure 1).[39]

Epigenetic alterations in HCC

HCC is a heterogeneous disease in terms of aetiology and cell of origin.[40] Various environmental agents and lifestyles known to be risk factors for HCC are suspected to promote its development by eliciting epigenetic changes, which have a key role in a wide range of human malignancies.[41]

DNA methylation in HCC

Altered DNA methylation is an early event in HCC development. Global hypomethylation mainly affects intergenic regions of the genome and has a critical role in increasing chromosomal instability.[42] DNA methylation of gene promoters, which is important in transcriptional regulation and the cellular differentiation process,[43] is a common mechanism of gene silencing in cancer cells. Furthermore, CpG island hypermethylation phenotypes have been reported in various types of cancers, such as colorectal,[44] uterine,[45] glioma,[46] and renal[47] cancers. However, the presence of such phenotypes is still controversial in HCC.[48,49]

A molecular mechanism of active DNA demethylation has been identified and shown to be involved in tumorigenesis,[50] particularly in glioma and haematological malignancies. Hydroxymethylcytosine is present at a considerable level in normal adult liver tissues and is often decreased in tumour tissues;[51] however, its role in liver carcinogenesis remains unknown. *IDH1* and *IDH2* mutations are frequent in IHCCs and have been detected in 34 of 326 cases (10%).[52] Tumours containing mutations in *IDH1* or *IDH2* had lower 5-hydroxymethylcytosine and higher 5-methylcytosine levels compared with those without mutations, and 50% of hypermethylated genes overlapped with DNA hypermethylation in *IDH1*-mutant glioblastomas.[51]

To investigate DNA methylation patterns comprehensively, aberrantly methylated genes are identified by methylated DNA immunoprecipitation (meDIP) followed by tiling array[53] or next-generation sequencing. Deng *et al.*[54] applied the meDIP-chip method to identify 15 genes preferentially methylated in HCV-related HCCs. Alternatively, a genome-wide DNA methylation assay that was developed on Beadchip™ (Illumina Inc., San Diego, CA) technology[55] can measure methylation levels quantitatively at single CpG sites, and yield largely comparable results to meDIP sequencing[56] and whole-genome bisulphite sequencing. This assay has been applied to methylation profiling in various cancers and in the cancer genome atlas project.[57] A few distinct epigenetic subtypes identified on the basis of the methylation pattern have been detected in HCCs and will be integrated with genetic alteration data.

Shen *et al.*[58] used a 27K Infinium™ array (Illumina) to analyse 62 HCC cases and identified 2,324 differentially methylated CpG sites, of which 684 hypermethylation markers could be utilized for plasma DNA diagnostics. They also analysed 66 HCC cases using a 450K array in which the top 500 significant CpG sites that were differentially methylated were able to distinguish HCC from adjacent tissues.[59] Meanwhile, Tao *et al.*[60] analysed non-cancerous tissues of HBV-associated HCC on a 27K array and identified hypermethylated genes. Accumulation of such methylations would form "an epigenetic field for cancerization".[61]

An early study[62] showed that extensive methylation is associated with *CTNNB1* mutations, while HCC with a *TP53* mutation is often characterized by chromosomal instability. Given that *CTNNB1* and *TP53* mutations are mutually exclusive in HCCs, such distinct methylation patterns could be associated with particular genetic alterations.

Promoter CpG islands of the *CDKN2A* and *CDKN2B* tumour suppressor genes are frequently hypermethylated, leading to inactivation of the *RB* pathway.[63] Methylation of the *CDKN2A* gene promoter occurs in 73% of HCC

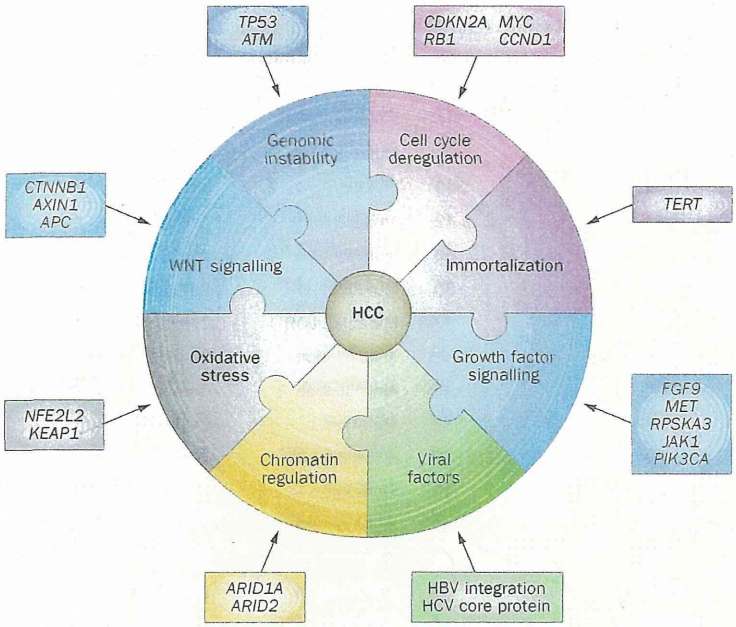

Figure 2 | Core oncogenic pathways in hepatocarcinogenesis. Representative genes involved in each pathway are indicated.

tissues,[64] 56% of HBV-related HCC, and 84% of HCV-related HCC.[65] *RASSF1A* is methylated in up to 85% of HCCs,[66] *GSTP1* in 50–90%,[67,68] and *MGMT* in 40%.[69]

Transcriptome analysis and beyond

RNA sequencing technology has enabled not only transcriptomic profiling, but also the identification of rearranged transcripts, such as translocations and inversions, and tumour-specific expression of noncoding RNAs, although the latter analysis requires deep coverage of sequencing reads. No recurrent fusion genes have been reported in HCC to date.

Classification based on gene expression, copy number and DNA methylation profiling data would help elucidate the correlation between mutation profiles and molecular subclasses.[70,71] Gene expression profiles in cancer are the result of genetic and epigenetic alterations. Therefore, an integrated genomic analysis is necessary to determine how these genetic and epigenetic alterations affect cancer phenotypes, because the combination of somatic mutations, promoter methylation, and chromosomal loss might lead to gene inactivation.[72]

Vetter *et al.*[73] reported on the association between the increase of splicing variants of the *KLF6* gene and increased hepatocarcinogenesis. Splicing variants in HCCs have been reported in several genes,[74] including *LLGL1* (also known as *HUGL1*),[75] *TCF4*,[76] and *p73*.[77] Transcriptome sequencing of HCC samples combined with genotyping validation identified a frequent adenosine-to-inosine RNA editing event in the *AZIN1* gene in HCC.[78,79] This editing induces a serine-to-glycine amino acid change that confers gain-of-function activity and a stronger affinity of the edited protein to antizyme. Increased AZIN1 (antizyme inhibitor 1) protein stability could promote cell proliferation, presumably through the neutralization of ornithine decarboylase (ODC) and G1/S-specific cyclin-D1 (CCND1) degradation mediated by antizyme. Adenosine-to-inosine RNA editing will contribute to more transcriptome diversity and liver carcinogenesis.

Core liver cancer genes and pathways

Comprehensive analyses of the liver cancer genome have demonstrated that multiple cancer genes and molecular pathways are recurrently altered and have pivotal roles in hepatocarcinogenesis (Figure 2). Table 2 summarizes important mutated genes in liver cancer.

TP53 pathway

TP53 is the top gene among recurrently mutated genes in HCC, and its mutation frequency varies between 18% and 35.2% (25.9% on average) of HCCs.[80] Alterations of other genes located upstream and downstream on the TP53 pathway, such as recurrent mutations of the *ATM* (an upstream regulator of TP53 activation[81]) and *CDKN1A* (a target of TP53[82]) genes, have also been reported. Moreover, mutations of the *IRF2* gene, which encodes a positive regulator of TP53 protein expression, are mutually exclusive to the TP53 mutation in a cohort of patients with HCC.[26]

Cell cycle regulation pathway

The G1/S cell cycle checkpoint and cell senescence are regulated by *RB* and *CDKN2A*. Inactivation of the *RB* and *CDKN2A* genes by homozygous deletion and promoter CpG hypermethylation or point mutations has been reported in HCC.[83,84] The tumour suppressing activity of *RB* in the liver was evaluated in a mouse model, and RB inactivation was found to be associated with both increased cell proliferation and chromosomal instability.[85]

TERT pathway

Activation of telomerase (encoded by the *TERT* gene), which is physiologically silenced in most normal cells, is required for infinite replication in cancer cells.[86] Somatic mutations in the *TERT* gene promoter have been shown to promote *TERT* gene expression in melanoma.[87,88] Killela *et al.*[89] screened these mutations in >1,000 tumours of various organs and reported that 27% of HCC cases harboured these alterations. Nault *et al.* reported *TERT* promoter mutations in 54% of human HCCs and 25% of cirrhotic preneoplastic nodules, suggesting that this alteration could be the earliest recurrent genetic event in hepatocarcinogenesis.[90]

WNT pathway

Aberrant activation of WNT signalling is a driving molecular event in a wide range of tumours, including liver cancers.[91] Somatically acquired missense mutations in exon 3 of the *CTNNB1* gene are frequently reported in HCC (10.0–32.8% in genome-wide sequencing studies).[92] In addition to *CTNNB1*, alterations of *APC* and *AXIN1*, which are tumour suppressor genes that negatively regulate catenin β-1 (CTNNB1) protein levels in a post-transcriptional manner, have been recurrently reported