nature genetics

ARTICLES

Trans-ancestry mutational landscape of hepatocellular carcinoma genomes

Yasushi Totoki[1,14], Kenji Tatsuno[2,14], Kyle R Covington[3,14], Hiroki Ueda[2], Chad J Creighton[3,4], Mamoru Kato[1], Shingo Tsuji[2], Lawrence A Donehower[5], Betty L Slagle[5], Hiromi Nakamura[1], Shogo Yamamoto[2], Eve Shinbrot[3], Natsuko Hama[1], Megan Lehmkuhl[3], Fumie Hosoda[1], Yasuhito Arai[1], Kim Walker[3], Mahmoud Dahdouli[3], Kengo Gotoh[2], Genta Nagae[2], Marie-Claude Gingras[3], Donna M Muzny[3], Hidenori Ojima[6], Kazuaki Shimada[7], Yutaka Midorikawa[8], John A Goss[9], Ronald Cotton[9], Akimasa Hayashi[2,10], Junji Shibahara[10], Shumpei Ishikawa[10], Jacfranz Guiteau[9], Mariko Tanaka[10], Tomoko Urushidate[1], Shoko Ohashi[1], Naoko Okada[1], Harsha Doddapaneni[3], Min Wang[3], Yiming Zhu[3], Huyen Dinh[3], Takuji Okusaka[11], Norihiro Kokudo[12], Tomoo Kosuge[7], Tadatoshi Takayama[8], Masashi Fukayama[10], Richard A Gibbs[3], David A Wheeler[3], Hiroyuki Aburatani[2] & Tatsuhiro Shibata[1,13]

Diverse epidemiological factors are associated with hepatocellular carcinoma (HCC) prevalence in different populations. However, the global landscape of the genetic changes in HCC genomes underpinning different epidemiological and ancestral backgrounds still remains uncharted. Here a collection of data from 503 liver cancer genomes from different populations uncovered 30 candidate driver genes and 11 core pathway modules. Furthermore, a collaboration of two large-scale cancer genome projects comparatively analyzed the trans-ancestry substitution signatures in 608 liver cancer cases and identified unique mutational signatures that predominantly contribute to Asian cases. This work elucidates previously unexplored ancestry-associated mutational processes in HCC development. A combination of hotspot *TERT* promoter mutation, *TERT* focal amplification and viral genome integration occurs in more than 68% of cases, implicating *TERT* as a central and ancestry-independent node of hepatocarcinogenesis. Newly identified alterations in genes encoding metabolic enzymes, chromatin remodelers and a high proportion of mTOR pathway activations offer potential therapeutic and diagnostic opportunities.

HCC is the third leading cause of cancer deaths worldwide[1,2]. Epidemiologically, the incidence of HCC shows marked variance across geographical regions and ancestry groups and between the sexes[3]. HCC incidence predominates in East Asia and Africa, and rapid increases in prevalence have occurred in Western countries[2]. Multiple etiological cofactors are associated with liver cancer, and their contributions might additionally differ according to ancestry. Hepatitis B virus (HBV) infection is dominant in East Asia and Africa, whereas hepatitis C virus (HCV) infection among HCC cases is frequent in Japan. Aflatoxin B1 exposure is a strong risk factor of HCC in China and Africa, whereas alcohol intake is a major etiological factor for HCC in Western countries[3–5]. The average male/female ratio for HCC incidence is greater than two, which could be owing to different environmental exposures or hormone levels[6]. Overlapping but partially distinctive epidemiological backgrounds, such as liver fluke infection, were associated with intrahepatic cholangiocarcinoma (IHCC), another type of liver cancer[5]. Here we conducted the first trans-ancestry HCC genome sequencing research under the umbrella of the International Cancer Genome Consortium (ICGC)[7] and The Cancer Genome Atlas (TCGA)[8]. Thus far, this study represents the largest genomic profiling of liver cancers (608 cases) and compares ancestry groups (Japanese, Asian and European) with distinctive etiological cofactors. This genome data set also uncovers an extensive landscape of driver genetic alterations in HCC.

RESULTS

Whole-exome and oncovirome sequencing of liver cancers

As an ICGC liver cancer project, we collected 503 pairs (413 cases in the Japanese cohort and 90 cases in the US cohort) of liver cancers (488 HCC and 15 IHCC) and matched non-cancerous liver tissues

[1]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [2]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [4]Department of Medicine, Baylor College of Medicine, Houston, Texas, USA. [5]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA. [6]Division of Molecular Pathology, National Cancer Center Research Institute, Tokyo, Japan. [7]Hepatobiliary and Pancreatic Surgery Division, National Cancer Center Hospital, Tokyo, Japan. [8]Department of Digestive Surgery, Nihon University School of Medicine, Tokyo, Japan. [9]Department of Surgery, Baylor College of Medicine, Houston, Texas, USA. [10]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [11]Hepatobiliary and Pancreatic Oncology Division, National Cancer Center Hospital, Tokyo, Japan. [12]Hepato-Biliary-Pancreatic Surgery Division, Department of Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [13]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [14]These authors contributed equally to this work. Correspondence should be addressed to D.A.W. (wheeler@bcm.edu), H.A. (haburata-tky@umin.ac.jp) or T.S. (tashibat@ncc.go.jp).

Received 31 December 2013; accepted 3 October 2014; published online 2 November 2014; doi:10.1038/ng.3126

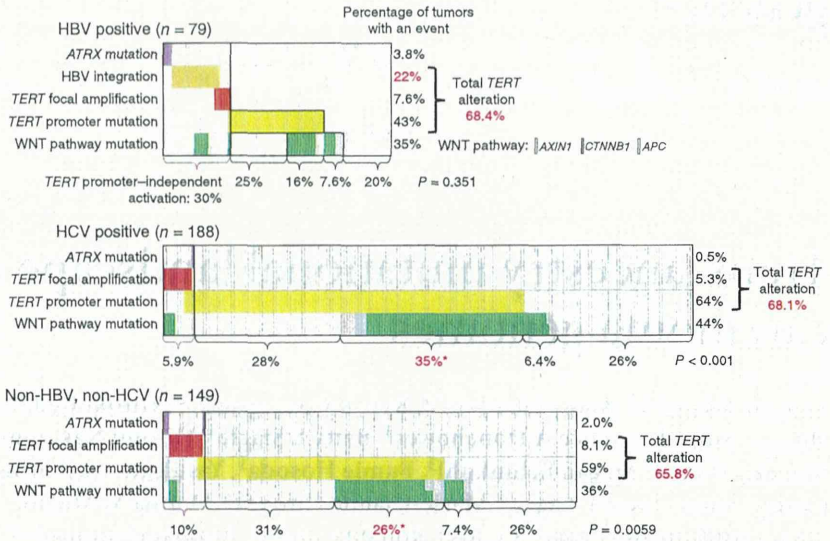

Figure 1 Multiple types of *TERT* alterations in HCC. Mutual exclusivity of HBV genome integration at the *TERT* locus, *TERT* focal amplification and *TERT* promoter mutation in HBV-positive (top), HCV-positive (middle) and non-HBV, non-HCV (bottom) cases. *AXIN1*, *CTNNB1* and *APC* mutations were included as WNT pathway mutations. *TERT* promoter mutation significantly co-occurred with WNT pathway mutation in HBV-negative cases (*$P < 0.001$, χ^2 test). HBV-positive cases without virus capture analysis (41 samples) were excluded (**Supplementary Table 28**).

or blood. This cohort contained 212 HCV-positive, 117 HBV-positive and 150 non-virus cases. The US cohort contained European-ancestry (55%), Asian (defined as US-Asian hereafter; 16%) and African-American (12%) cases. The clinical backgrounds for this cohort are shown in **Supplementary Table 1**.

The exons and surrounding noncoding genomic regions of protein-coding genes were captured in 452 pairs of tumor and non-cancerous liver tissues. Oncoviral genomes, including for HBV, human papillomavirus (HPV-16 and HPV-18) and human T-lymphotrophic virus 1 (HTLV1) (91 kb in total; **Supplementary Table 2**), were also captured in 198 cases. Whole-genome sequencing was conducted in 22 HCC pairs, including 9 exome-sequenced cases, and targeted resequencing of liver cancer genes was carried out for 38 cases. To minimize multicenter study bias due to differences in exome sequencing platform or data analysis pipeline, we optimized the somatic mutation detection algorithms and filtering conditions for three centers using Japanese cohort samples. High concordance (>87%) with a validation rate of >97% in somatic mutation detection was achieved, and substitution patterns among the three centers were consistent (**Supplementary Figs. 1** and **2**). We also confirmed that similar mutation spectra were observed in the same cases in whole-genome sequence and whole-exome sequence (**Supplementary Fig. 3**).

The average mutation rate was 2.8 mutations per megabase, and T>C and C>T substitutions were dominant in this cohort (**Supplementary Fig. 4**). Eight (1.7%) outlier tumors harboring more than 4.3 mutations per megabase showed substitution patterns distinctive from those of other cases and had somatic nonsense or missense mutations in mismatch repair (*MSH3*, *MSH4*, *MSH5* and *MSH6*), DNA polymerase (*POLA1*, *POLK*, *POLE* and *POLL*) or nucleotide excision repair (*ERCC1* and *ERCC2*) genes (**Supplementary Fig. 5**).

Panoramic view of ploidy, copy number and virus integration

We evaluated copy number alteration (CNA) by comparing the sequence depth for paired samples and allelic imbalance in the captured area (**Supplementary Fig. 6**). This digital assessment of CNA and allelic imbalance was consistent with SNP array data in cases analyzed by both methods (**Supplementary Fig. 7**). We also imputed deviation in the allele frequency of heterozygous single-nucleotide variation to predict the tumor purity and ploidy for each sample (H.U., S.Y., K.T. and H.A., unpublished data). A large fraction of cases (28.9%) represented whole-genome duplication with gross chromosomal loss (average ploidy was 3.87, and the average number of CNAs was 11.58) (**Supplementary Fig. 8**), whereas the remainder showed more stable copy number status (average ploidy was 2.08, and the average number of CNAs was 7.56). Tetraploidy was more frequently observed in higher-grade tumors ($P = 0.039$, Fisher's exact test; **Supplementary Fig. 9**).

We observed recurrent arm-level gains (1q, 5p, 6p and 8q) and losses (1p, 4q, 6q, 8p and 17p), as previously described for HCC[9] (**Supplementary Fig. 10**). Recurrent focal amplifications were detected in 25% of cases, including for *TERT* and *CCND1-FGF19*. Homozygous deletions were less frequent events (detected in 17.4% of cases). Recurrent homozygous deletion was observed for 28 genes, including *CDKN2A-CDKN2B*, *MAP2K3* and *PTEN* (**Supplementary Figs. 11** and **12**).

Using paired-end reads mapped to the HBV viral and human genomes, respectively, we detected 628 HBV virus integrations in 68 HBV-positive cases from which viral genomes were captured (9.2 integrations per case) (**Supplementary Table 3**), reflecting a detection rate that was 2–4 times more sensitive than in previous whole-genome sequencing studies[10,11]. Genes close to (less than 10 kb away from) the recurrent HBV integrations included *TERT* ($n = 17$ cases), *KMT2B* (*MLL4*; $n = 6$ cases), and *ALOX5*, *ZFPM2*, *SENP5*, *MYO19* and *RGS22* ($n = 2$ cases each). Recurrent non-genic HBV integrations were observed near the centromere, especially on chromosomes 1p, 8p and 10q. A significant fraction of HBV integrations were colocalized with (less than 500 kb away from) DNA copy number breakpoints (10.7%; $P < 1 \times 10^{-5}$, randomization test) (**Supplementary Figs. 13** and **14**). Despite intimate association between HBV genome integration and CNA breakpoints, the frequency of CNA was not different among the viral subtypes ($P = 0.29$, ANOVA test; **Supplementary Fig. 15** and **Supplementary Table 4**).

Multiple types of *TERT* genetic alteration in HCC

Somatic mutations in the transcriptional regulatory region of the *TERT* gene have been reported in a range of cancers, including HCC[12,13]. By combining captured noncoding sequence data with capillary sequencing validation, we detected *TERT* promoter mutations in 254 cases of the 469 cases analyzed (54% in total). The frequency of these mutations was highest in HCV-positive cases (121/188; 64%), with lower frequencies in non-viral cases (88/149; 59%) and HBV-positive cases (44/120; 37%) (**Supplementary Table 5**). As reported[13], the mutation located 124 bp upstream of the ATG start site (c.–124C>T, on the opposite strand; 93%) was more frequent than the c.–146C>T (4.3%) and c.–57A>C (1.6%) mutations (**Supplementary Table 6**).

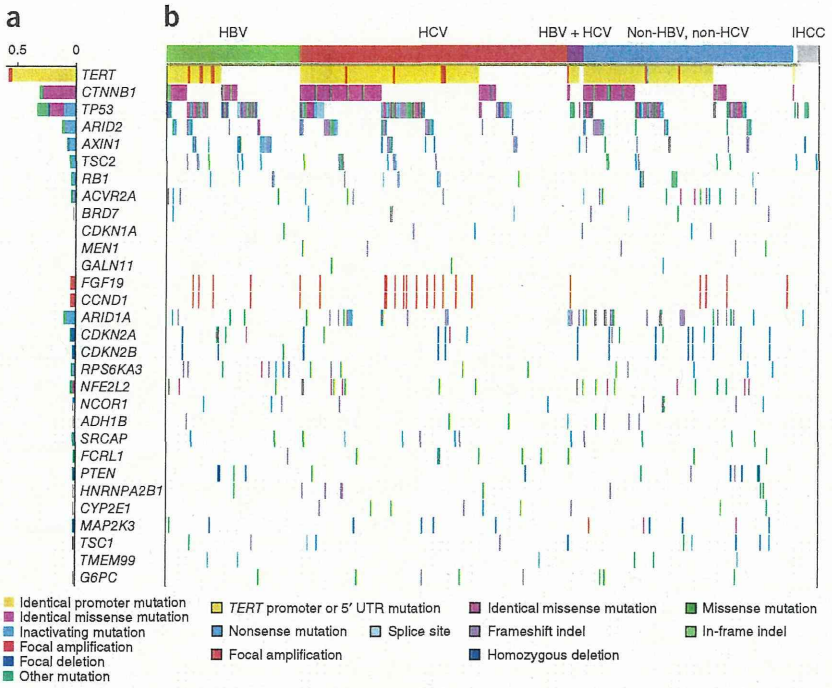

Figure 2 Significant cancer driver genes in HCC. An overview of significant driver genes in HCC. Shown are genes with statistically significant mutations or focal CNAs (**a**) and their alterations in each sample classified by the status of hepatitis virus infection (**b**). Genes were sorted by significant *q* value (**Supplementary Note**).

Additionally, *TERT* focal amplification was detected in 6.7% of the cases in total, and integration of the HBV genome in the *TERT* locus was observed in 22% of HBV-positive samples for which integration was analyzed. *TERT* promoter mutations were mutually exclusive with HBV genome integration in the *TERT* locus in integration-analyzed HBV-positive samples and were almost mutually exclusive with *TERT* focal amplifications, both of which were considered to cause higher *TERT* expression[14] (**Fig. 1**). Alterations of *ATRX* have also been reported to induce telomerase-independent telomere maintenance[15], Altogether, more than 68% of the HCC cases had alterations in either *TERT* or *ATRX*, representing the most frequent molecular event reported (**Supplementary Table 5**). In contrast, no *TERT* promoter mutations were detected in 13 IHCC cases (**Fig. 2**). *TERT* promoter mutations significantly co-occurred with WNT pathway gene alterations, such as *CTNNB1*, *AXIN1* or *APC*, in HCV-positive and non-virus cases, suggesting a cooperative oncogenic activity between *TERT* promoter mutation and the WNT pathway[16] in these subgroups (**Fig. 1**).

Significantly altered genes in HCC

To identify significantly altered genes in HCC, we used a combination of MutSigCV[17], an aggregated somatic alteration method that aggregates somatic substitutions, short indels, homozygous deletions and focal amplifications, and an inactivation bias method that calculates inactivating mutation bias (**Supplementary Fig. 16**, **Supplementary Tables 7–10** and **Supplementary Note**). Furthermore, we eliminated mutated genes that exhibited sequencing center bias and subclone bias as sources of possible false discovery (**Supplementary Tables 11** and **12**). These steps led to a final list of 30 candidate driver genes (**Fig. 2**, **Supplementary Fig. 17** and **Supplementary Tables 13–15**), including 13 that were not recurrently mutated in previous cohorts[18–20] (**Supplementary Table 16**). These 13 genes included *BRD7*, a component of the SWI/SNF nucleosome-remodeling machinery, and *MEN1*, a putative tumor suppressor somatically mutated in neuroendocrine tumors—neither of which has been reported in HCC. Mutations in *TSC2*, *SRCAP* and *NCOR1* have been reported as singletons in other

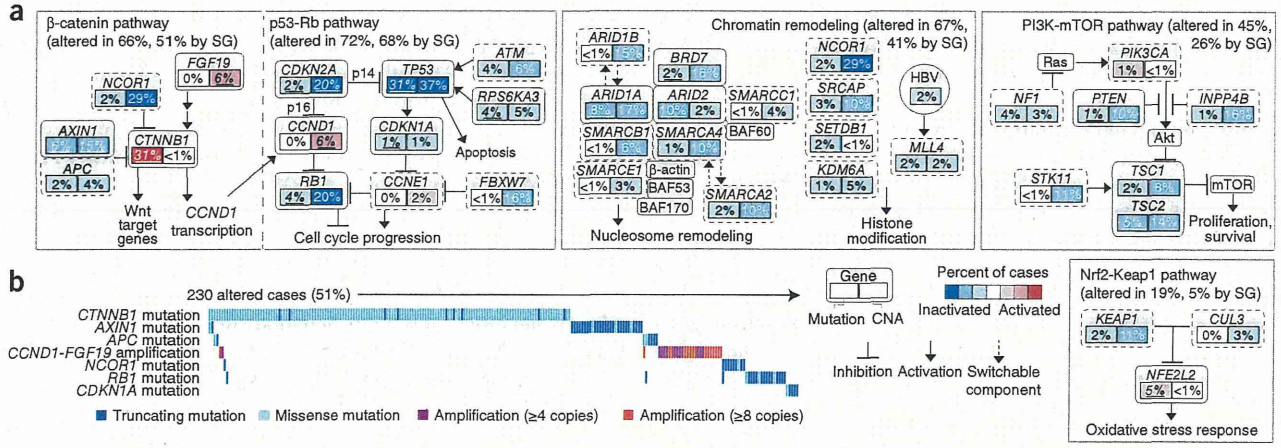

Figure 3 Oncogenic network in HCC. (**a**) Major signaling pathways involving genetic alterations in HCC. Key genes in each pathway are indicated by rectangles, with the percentages of somatic mutations and CNAs shown in the left and right portions of each rectangle, respectively. Significantly altered genes (SG; MutSigCV, $P < 0.05$ or GISTIC, q value < 0.1; percentages are underlined for alterations meeting either criterion) are bounded by solid lines, whereas other key genes in each pathway are bounded by dashed lines. (**b**) Mutual exclusivity plot of genes relevant to the WNT signaling pathway. The plot indicates that somatic mutations in WNT-related genes might contribute to the activation of WNT signaling in over half of all HCCs.

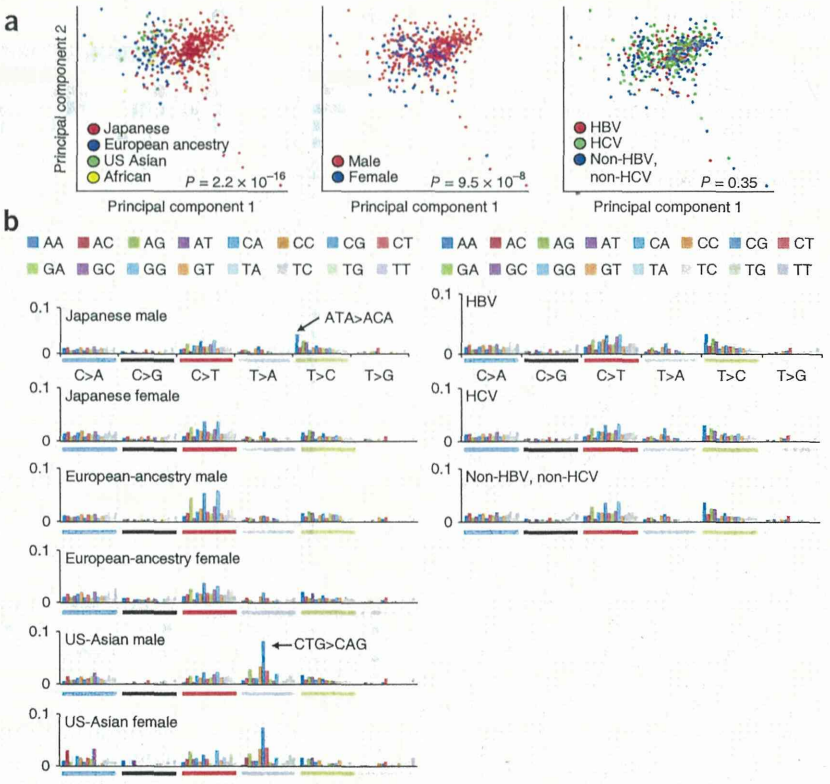

Figure 4 Somatic substitution patterns were associated with ancestry. (**a**) Principal-component analysis of the 96 substitution patterns in the HCC genome by ancestry group (left), sex (middle) and hepatitis virus group (right). (**b**) Average frequency of the 96 substitution patterns in each sample group (ancestry group, sex and virus group). The top legend shows the bases immediately 5′ and 3′ to each substitution. The *y* axis indicates the frequency of the 96 substitution patterns.

studies, but these genes were shown here to be significantly mutated. Some of the difference in results might be attributed to the greatly increased statistical power with our 503-case population, but some of the difference might also reflect contribution from the ancestry composition of the cohorts in this study. Several genes demonstrated differences in mutational frequency among virus subtypes (**Fig. 2b** and **Supplementary Table 17**). *AXIN1* was more frequently mutated in HBV-positive cases in comparison with HCV-positive and non-virus HCC ($P = 0.0055$, Fisher's exact test), indicating that different viral etiologies might activate WNT signaling in distinct ways. *ARID1A* was more frequently altered in non-virus cases ($P = 0.009$).

Alterations of drug target kinases were rarely found in HCC; low-level recurrent mutations of *FGFR2* (mutated in 1.8% of cases), *KIT* (1.3%), *FGFR3* (0.9%), *FGFR1* (0.9%), *JAK1* (0.9%) and *EGFR* (0.4%) and focal amplification of *MET* (0.5%) were detected. The specific mutations in these receptor tyrosine kinases were not generally observed in other cancers, with the exception of two *JAK1* mutations (encoding p.Ser703Ile and p.Leu910Pro substitutions), which were previously observed in a liver cancer sequencing study[20]. The liver has a central role in many metabolic processes. Our study identified recurrent mutations of metabolic enzyme genes in HCC (**Fig. 2b** and **Supplementary Tables 7** and **13**). These included *CYP2E1* (2.0%); *ADH1B* (1.8%), encoding alcohol dehydrogenase 1B; and *G6PC* (1.8%), encoding a glucose-6-phophatase catalytic subunit, whose aberrations could be linked to metabolomic changes in HCC.

Significant oncogenic pathways in HCC

Oncogenic pathways were further explored by aggregating the alterations of each gene within a particular pathway (**Fig. 3a**).

TP53-RB pathway. Inactivation of the tumor-suppressor TP53-RB pathway was a consistent theme in HCC. *TP53* mutations were observed in 31% of tumors, and two genes encoding p53-activating kinases, *ATM* and *RPS6KA3*, were also recurrently mutated. The *RB1* gene was mutated in 4.4% of cases. The *CDKN2A* gene encoding the RB regulator $p16^{INK4A}$ was subject to frequent focal homozygous deletion, and the p53 target and RB regulator *CDKN1A* (encoding $p21^{CIP1}$) was significantly mutated. Overall, 72% of cases had alterations in component genes of one or both of these pathways.

WNT pathway. In addition to activating *CTNNB1* mutations, inactivating mutations were frequently observed in WNT regulators, including *AXIN1* and *APC*. *CCND1* is a key downstream target of WNT signaling[21], and *FGF19* has been shown to activate *CTNNB1* transcriptional functions[22]. Mutual exclusivity of *CTNNB1*, *AXIN1* and *APC* mutations and *CCND1-FGF19* amplification supports the functional role of these genes in altering WNT signaling (**Fig. 3b**). Overall, 66% of HCCs showed WNT pathway–related alterations.

Chromatin and transcription modulators. A large proportion of the genes on the list of significantly mutated genes encoded chromatin modulators or transcriptional regulators. Frequent alterations in *NFE2L2*, encoding a transcriptional regulator that activates antioxidant and cytoprotective target genes[23], and its negative regulators *KEAP1* and *CUL3* (ref. 24) were noted. Also mutated were the nucleosome remodelers *ARID1A*, *ARID2* and *BRD7*, with CNAs and mutations in six additional members of the SWI/SNF complex (**Fig. 3a**), *SRCAP* and the transcriptional corepressor *NCOR1*, both of which have roles in steroid receptor–mediated transcription. These genes displayed primarily inactivating frameshift and nonsense mutations that suggest a tumor-suppressor gene function in HCC (**Supplementary Fig. 18** and **Supplementary Table 9**). *NCOR1* has been shown to directly suppress *CTNNB1* function[25] and exhibits mutual exclusivity for mutations with other WNT pathway genes (**Fig. 3b**). *SRCAP* encodes an Snf2-related CREBBP activator in several pathways, including NOTCH[26] and steroid receptors[27]. Truncating *SRCAP* mutations cause a rare hereditary disease with developmental defects and early-onset tumor formation[28,29], highlighting its potential function as a tumor-suppressor gene.

mTOR-PIK3CA pathway. Recurrent inactivating mutations in *TSC1-TSC2* and activating mutations and copy gain in *PIK3CA* were observed (**Fig. 3a**). Other modulators involved with this pathway, such as *NF1*, *PTEN*, *INPP4B* and *STK11*, were also affected, and, in total, 45% of cases had alterations in the mTOR-PIK3CA pathway. Somatic *TSC1* mutation was reported as a potential predictive biomarker of an mTOR inhibitor[30], and *TSC1*-mutated HCC cell lines showed

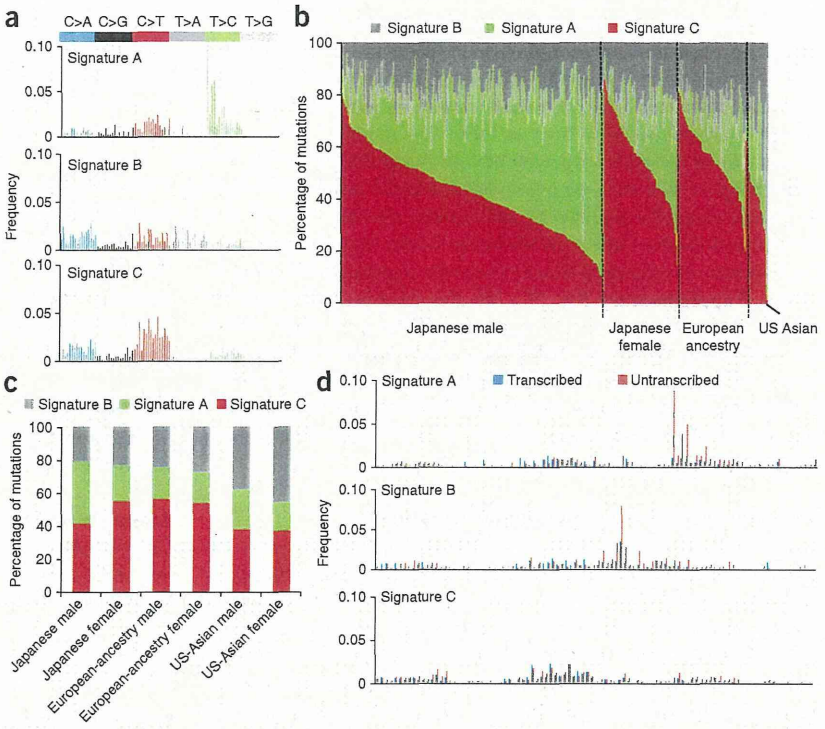

Figure 5 Ancestry-specific mutational signatures with transcriptional strand bias in the HCC genome. (**a**) The 3 mutational signatures in the HCC genome are shown according to the frequencies of 96 substitution types. The *y* axis indicates the frequency of each of the 96 substitution patterns. (**b**) Contribution of the three mutational signatures to each tumor. The *y* axis indicates the percentage of mutations comprised in each signature. The *x* axis indicates tumors classified in each ancestry group and by sex. (**c**) Contribution of the three mutational signatures to tumors from each ancestry group and sex. The *y* axis indicates the percentage of mutations comprised in each signature. (**d**) Transcriptional strand bias in mutational signatures. Each signature is displayed with 192 mutation patterns based on the 96 substitution types with transcriptional strand information. The mutation types are shown on the *x* axis, and the *y* axis indicates the frequency of each of the 192 mutation types contributing to each signature.

higher sensitivity to an mTOR kinase inhibitor (BEZ235) in comparison to cell lines with wild-type *TSC1* (**Supplementary Fig. 19**).

To identify networking among the oncogenic pathways in HCC, we developed a pathway compression algorithm and applied it to the significantly altered genes. We identified 11 core oncogenic network modules in HCC (**Supplementary Table 18**). To visualize these modules in the context of a biological network, we constructed a schematic view of the modules and the additional nodes that can connect them (**Supplementary Fig. 20**). The nodes were typically classified into two types; one type was closely connected to neighboring nodes (with higher value for centrality; **Supplementary Table 19**) and the other type had long-range edges that reached distant nodes, which can be used to measure the effect of each module alteration on the total network. Further comparison of the association between these module alterations and background clinical factors showed that the mTOR module was significantly different ($P < 0.05$, Cochran-Mantel-Haenszel test) in Asian and European-ancestry populations with respect to mutational frequencies (**Supplementary Fig. 21**).

Ancestry-dependent diversity in HCC mutation signatures

Somatic mutation patterns in human cancer are closely associated with epidemiological factors[31–34]; however, their association with ancestry remains unexplored. We integrated genomic data from an additional 105 HCC cases sequenced by TCGA along with the 503 cases sequenced by us (**Supplementary Table 1**) and compared somatic substitution patterns according to epidemiological data and ancestry group. Because mutation patterns in hypermutated cases and IHCC were distinctive (**Supplementary Figs. 4** and **22**), these two groups were excluded from further mutation pattern analysis.

Principal-component analysis of the 96 possible nucleotide triplets, dependent on the bases immediately 5′ and 3′ to each substitution, showed that the constitution of substitution patterns with these triplets was significantly different by ancestry group (Japanese, US Asian and European ancestry; $P = 2.2 \times 10^{-16}$, Wilks' test) and by sex ($P = 9.5 \times 10^{-8}$) (**Fig. 4a**). Notably, substitution patterns were not significantly associated with viral status (HBV, HCV and non-viral, $P = 0.35$; **Fig. 4a** and **Supplementary Fig. 23**). T>C substitutions, particularly in an ATA context, were specifically increased in Japanese male samples, and T>A substitutions (most frequently in a CTG context) were specifically increased in US-Asian male and female samples. The distributions of the frequencies for the 96 substitution types were similar among Japanese female samples and European-ancestry male and female samples (**Fig. 4b**).

We applied non-negative matrix factorization (NMF) analysis to the 96-substitution pattern[33] and identified 3 mutation signatures (HCC signatures A–C; **Fig. 5a** and **Supplementary Fig. 24**). Each signature was composed of context-specific substitutions: HCC signature A was characterized by dominant T>C mutations, especially in an AT(A/G/T) context, whereas HCC signature B contained dominant T>A mutations, with a sharp increase in frequency for a CTG context. HCC signature C contained dominant C>T mutations, especially in an (A/C/G)CG context. The distribution of these signatures was associated with ancestry and sex but not with the virus status (**Supplementary Table 20**). Among the different ancestry groups, HCC signatures A and B more frequently contributed to Japanese male (odds ratio (OR) = 2.2; $P = 0.0025$, Fisher's exact test) and US-Asian (OR = 2.5; $P = 0.00036$) cases, respectively, whereas HCC signature C was common across all ancestry groups and in both sexes (**Fig. 5b,c** and **Supplementary Fig. 25**). Remarkable differences in mutation prevalence between the transcribed and untranscribed strands were observed for T>C substitutions, especially in an AT(A/G/T) context ($P = 7.4 \times 10^{-152}$, χ^2 test), in HCC signature A and for T>A substitutions, especially in a CTG context ($P = 3.3 \times 10^{-8}$), in HCC signature B (**Fig. 5d**). These significant strand biases imply the involvement of transcription-coupled repair, which is tightly associated with known carcinogens in other tumor types[31–34]. There was no significant association between the signature distribution and the *ALDH2* SNP rs671, which is associated with alcohol metabolism and is a more frequent genotype in the Asian population[35] (**Supplementary Table 21**).

To collect large amounts of cancer genome data from different ancestry groups and epidemiological backgrounds, we currently need to combine data from multiple institutes that apply individual analytical platforms. An important caveat in multicenter trans-ancestry analysis has been the possibility that ancestry-specific signatures can be biased by experimental or analytical differences. To avoid this potential bias, we processed the DNA from 99 Japanese HCC cases using the sequencing and analysis pipeline at the United States–based Baylor College of Medicine. Using this data set from a single center, we replicated exactly the same signatures in each population (**Supplementary Fig. 26**). We also examined the distribution of signatures among three centers using Japanese male samples and confirmed that similar distributions were seen among the three centers (**Supplementary Fig. 27**). Furthermore, we analyzed whole-genome sequencing data for 88 Chinese HCC samples[19] and successfully identified HCC signatures B and C in this independent data set (**Supplementary Fig. 28**).

Outcome analysis from mutational signatures

We analyzed the derived NMF signatures to determine whether any signature or signature component was associated with differences in outcome in the HCC cohort. NMF signature values were merged with annotated clinical data. We performed calculations using standardized signature values to control for differences in the mutation rate between the subjects. Multivariate analysis with the Cox proportional hazards model (**Supplementary Fig. 29** and **Supplementary Tables 22–26**) indicated that histological grade, HCC signature B and the interaction with HCC signature A (but not with HCC signature C) were significant predictors of outcome.

DISCUSSION

The present trans-ancestry liver cancer genome study first identified mutational signatures that are independent of hepatitis virus infection and contribute more to the Asian cases than to ones of European ancestry (**Supplementary Tables 27**). One signature, characterized by AT>AC mutations, was predominant in Japanese males, whereas the other, featuring CTG>CAG mutations, was found more frequently in tumors from Asians living in the United States. These correlations may highlight deeper intra-ancestry diversity and/or environmental contributions, and sex bias might further affect downstream target genes and molecular features in HCC[36]. As several genetic loci are associated with individual HCC risk together with HBV and/or HCV infection[37,38], somatic and germline genome interaction might also be important to consider. Notably, these signatures were not evident in IHCC for Japanese cases (data not shown), suggesting that they are unique properties of HCC. The causes of these signatures remain unknown, but skewed transcriptional strand biases in characteristic sequence contexts strongly imply the presence of specific, previously unexplored mutational processes, which profoundly influence tumor genome constitution and behavior.

With 503 cases, this study is the largest liver cancer genome analysis thus far, enabling the formation of a more thorough picture of the mutational landscape of HCC than ever before. In addition to identifying a large number of significantly mutated genes, we have also identified recurrent alterations of 9 of the 14 core genes making up the SWI/SNF complex. We also find a combination of hotspot *TERT* promoter and *ATRX* mutations, along with focal amplification and virus genome integration in the *TERT* locus, in more than 68% of HCC cases regardless of virus subtype. These findings show that *TERT* is a central driver gene and a promising molecular target[39] in HCC. The targeting of high-prevalence mTOR-PIK3CA pathway activation and antiproliferative activity in HCC cells by chemical inhibition should also offer new therapeutic opportunities. In addition, newly identified alterations in the chromatin-remodeling complex and metabolic enzymes are expected to be associated with cancer-specific epigenetic and metabolomic features.

URLs. DNAcopy, http://www.bioconductor.org/packages/2.13/bioc/html/DNAcopy.html; R software, http://www.R-project.org/; R survival package, http://CRAN.R-project.org/package=survival/; HGSC Mercury analysis pipeline, https://www.hgsc.bcm.edu/software/mercury; GRCh38 human reference genome, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/; BWA2, http://bio-bwa.sourceforge.net/; GATK4, http://www.broadinstitute.org/gatk/.

METHODS

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

Accession codes. Sequence data have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00001000389, the ICGC database (http://www.icgc.org/) and the database of Genotypes and Phenotypes (dbGaP) under accession phs000509.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This study was supported by Grants-in-Aid from the Ministry of Health, Labour and Welfare of Japan for the third-term Comprehensive 10-Year Strategy for Cancer Control, grants from the US National Human Genome Research Institute (NHGRI; 5U54HG003273) and National Cancer Institute (NCI; HHSN261201000053C and P30 CA125123), the Program for Promotion of Fundamental Studies in Health Sciences from the National Institute of Biomedical Innovation (NIBIO, Japan) and the National Cancer Center Research and Development Funds (23-A-8, Japan). The National Cancer Center Biobank is supported by the National Cancer Center Research and Development Fund, Japan. The supercomputing resource SHIROKANE was provided by the Human Genome Center at the University of Tokyo (http://sc.hgc.jp/shirokane.html).

AUTHOR CONTRIBUTIONS

Study design: Y.T., K.T., K.R.C., H.U., M.K., D.A.W., H.A. and T.S. Sequencing data generation: K.T., D.M.M., F.H., H. Doddapaneni, H. Dinh, Y.A., K.G., K.W., M.-C.G., T.U., S.O., N.O., M.W. and Y.Z. Data analysis: Y.T., K.T., K.R.C., H.U., M.K., S.T., L.A.D., B.L.S., E.S., S.Y., H.N., M.L., N.H., K.W., K.G., M.D., G.N., D.A.W. and T.S. Statistical analysis: Y.T., K.R.C., H.U., K.T., C.J.C., M.K., S.T. and S.Y. Molecular analysis: Y.A. and T.S. Sample acquisition and clinical data collection: M.-C.G., K.S., Y.M., J.A.G., H.O., A.H., J.S., R.C., J.G., S.I., M.T., T.O., N.K., T.K., T.T. and M.F. Manuscript writing: Y.T., K.T., K.R.C., H.U., C.J.C., L.A.D., B.L.S., M.K., D.A.W., H.A. and T.S. Project oversight: D.A.W., R.A.G., H.A. and T.S.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).
2. Forner, A., Llovet, J.M. & Bruix, J. Hepatocellular carcinoma. *Lancet* **379**, 1245–1255 (2012).
3. El-Serag, H.B. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology* **142**, 1264–1273 (2012).
4. Yu, J., Shen, J., Sun, T.T., Zhang, X. & Wong, N. Obesity, insulin resistance, NASH and hepatocellular carcinoma. *Semin. Cancer Biol.* **23**, 483–491 (2013).
5. Augustine, M.M. & Fong, Y. Epidemiology and risk factors of biliary tract and primary liver tumors. *Surg. Oncol. Clin. N. Am.* **23**, 171–188 (2014).
6. Tanaka, K., Sakai, H., Hashizume, M. & Hirohata, T. Serum testosterone:estradiol ratio and the development of hepatocellular carcinoma among male cirrhotic patients. *Cancer Res.* **60**, 5106–5110 (2000).
7. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

8. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
9. Wang, K. *et al.* Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology* **58**, 706–717 (2013).
10. Sung, W.K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44**, 765–769 (2012).
11. 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).
12. Killela, P.J. *et al. TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. USA* **110**, 6021–6026 (2013).
13. 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).
14. Li, Y. & Tergaonkar, V. Noncanonical functions of telomerase: implications in telomerase-targeted cancer therapies. *Cancer Res.* **74**, 1639–1644 (2014).
15. Heaphy, C.M. *et al.* Altered telomeres in tumors with *ATRX* and *DAXX* mutations. *Science* **333**, 425 (2011).
16. Hoffmeyer, K. *et al.* Wnt/β-catenin signaling regulates telomerase in stem cells and cancer cells. *Science* **336**, 1549–1554 (2012).
17. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
18. Li, M. *et al.* Inactivating mutations of the chromatin remodeling gene *ARID2* in hepatocellular carcinoma. *Nat. Genet.* **43**, 828–829 (2011).
19. 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).
20. Kan, Z. *et al.* Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.* **23**, 1422–1433 (2013).
21. Tetsu, O. & McCormick, F. β-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature* **398**, 422–426 (1999).
22. Pai, R. *et al.* Inhibition of fibroblast growth factor 19 reduces tumor growth by modulating β-catenin signaling. *Cancer Res.* **68**, 5086–5095 (2008).
23. Motohashi, H. & Yamamoto, M. Nrf2-Keap1 defines a physiologically important stress response mechanism. *Trends Mol. Med.* **10**, 549–557 (2004).
24. Zhang, D.D., Lo, S.C., Cross, J.V., Templeton, D.J. & Hannink, M. Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. *Mol. Cell. Biol.* **24**, 10941–10953 (2004).
25. Song, L.N. & Gelmann, E.P. Silencing mediator for retinoid and thyroid hormone receptor and nuclear receptor corepressor attenuate transcriptional activation by the β-catenin–TCF4 complex. *J. Biol. Chem.* **283**, 25988–25999 (2008).
26. Eissenberg, J.C., Wong, M. & Chrivia, J.C. Human SRCAP and *Drosophila melanogaster* DOM are homologs that function in the Notch signaling pathway. *Mol. Cell. Biol.* **25**, 6559–6569 (2005).
27. Monroy, M.A. *et al.* SNF2-related CBP activator protein (SRCAP) functions as a coactivator of steroid receptor–mediated transcription through synergistic interactions with CARM-1 and GRIP-1. *Mol. Endocrinol.* **17**, 2519–2528 (2003).
28. Hood, R.L. *et al.* Mutations in *SRCAP*, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *Am. J. Hum. Genet.* **90**, 308–313 (2012).
29. Nelson, R.A. *et al.* Floating-Harbor syndrome and intramedullary spinal cord ganglioglioma: case report and observations from the literature. *Am. J. Med. Genet. A.* **149A**, 2265–2269 (2009).
30. Iyer, G. *et al.* Genome sequencing identifies a basis for everolimus sensitivity. *Science* **338**, 221 (2012).
31. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
32. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
33. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
34. Poon, S.L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
35. Goedde, H.W. *et al.* Population genetic studies on aldehyde dehydrogenase isozyme deficiency and alcohol sensitivity. *Am. J. Hum. Genet.* **35**, 769–772 (1983).
36. Keng, V.W. *et al.* Sex bias occurrence of hepatocellular carcinoma in Poly7 molecular subclass is associated with *EGFR*. *Hepatology* **57**, 120–130 (2013).
37. Zhang, H. *et al.* Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. *Nat. Genet.* **42**, 755–758 (2010).
38. Kumar, V. *et al.* Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat. Genet.* **43**, 455–458 (2011).
39. Harley, C.B. Telomerase and cancer therapeutics. *Nat. Rev. Cancer* **8**, 167–179 (2008).

ONLINE METHODS

DNA preparation, DNA capture and sequencing. The tissues and clinical information used in this study were obtained under informed consent and approval of the institutional review boards of each institute. DNA was extracted from liver cancer tissue and matched non-cancerous liver tissues or blood using a general protocol for genome sequencing. Exome capture was carried out using the SureSelect Human All Exon V3 or V4 plus kit depending on the samples (**Supplementary Table 28**). Preparation of sequencing libraries, DNA capture methods and Illumina sequencing were carried out as described in the **Supplementary Note**.

Mutation calling. *Mutation calling (National Cancer Center Research Institute).* Paired-end reads were aligned to the human reference genome (GRCh37) using the Burrows-Wheeler Aligner (BWA)[40] for both tumor and normal samples. Probable PCR duplications, for which paired-end reads aligned to the same genomic position, were removed, and pileup files were generated using SAMtools[41] and a program developed in house. Details on our filtering conditions are provided in **Supplementary Tables 29** and **30**.

Mutation calling (Research Center for Advanced Science and Technology). Next-generation sequencing reads were mapped to the human genome (hg19) using BWA and Novoalign independently. Reads with a minimal editing distance to the reference genome were taken to represent optimal alignments. Then, bam files were locally realigned with SRMA. Normal-tumor pair bam files were processed using an in-house genotyper (karkinos), with the variants further filtered to remove all variants observed fewer than four times or present at an allele frequency of less than 0.12 after adjustment for tumor sample purity. The variants also had to have a score of greater than Q20 (representing the root mean square of mapping quality). In addition, reads harboring the variant had to be observed in both forward and reverse orientation. If a variant was present in reads of only one orientation, we checked for strand bias using a *t* test comparing these reads to the reads without the variant, and variants with a *P* value of <0.03 for strand bias were rejected. Variants also had to be called in different sequence cycles and have at least one call that was outside of 3% of read ends. Variants could not be located within 5 bp of an indel call, and variants where the mean base quality of the supporting reads was lower than 10 on the Phred scale were removed. Germline variants having an allelic frequency of greater than 0.1 were collected for 50 normal liver exome samples and used as the panel of normal variants. Any variant that was observed in this panel with a population frequency of greater than 2% was filtered out. Finally, variants also observed in the paired normal sample with an allelic frequency of greater than 3% and sites registered in dbSNP Build 134 with validated status were removed.

Mutation calling (Baylor College of Medicine). Initial sequence analysis was performed using the Human Genome Sequencing Center (HGSC) Mercury analysis pipeline. First, the primary analysis software on the instrument produced bcl files that were transferred off the instrument to the HGSC analysis infrastructure by the HiSeq Real-Time Analysis module. Once each run was complete and all bcl files were transferred, Mercury ran the vendor's primary analysis software (CASAVA), which demultiplexed pooled samples and generated sequence reads and base call confidence values (qualities). In the next step, reads were mapped to the GRCh37 human reference genome using BWA (BWA2), producing a bam3 (binary alignment/map) file. The third step involved quality recalibration (using GATK4) and, where necessary, the merging of bam files for separate sequence events into a single sample-level bam file. Sorting of bam files, duplicate read marking and realignment to improve indel discovery all occurred at this step.

Processing the significantly mutated genes. The significantly mutated genes for this study were identified through three separate tests as described below (an aggregated somatic alteration method, MutSigCV[42] and an inactivation bias method), and the resulting gene lists were combined in a final table of significantly mutated genes (**Supplementary Table 13**).We also developed two tests to detect bias in the mutation list that could be a source of artifact (K.R.C., E.S., L.A.D. and D.A.W., unpublished data). One of these tests examined sequencing center bias, and the other examined bias in mutation allele fraction, which if consistently low would suggest that a gene was a passenger rather than a driver. Genes in the final combined table that failed these bias tests were removed from the final list of significantly mutated genes. Data from each process are shown in **Supplementary Tables 7–12**, and the steps are shown schematically in **Supplementary Figure 16**.

Aggregated somatic alteration method. We identified significantly altered genes by aggregating somatic substitutions, short indels, homozygous deletions and focal amplifications. We initially estimated the expected number of each alteration in each gene as follows.

First, the substitution rate was estimated by dividing the number of synonymous mutations in a sample by the number of synonymous sites in the genome. For each gene, the expected number of substitutions was calculated by multiplying the substitution rate by the number of nonsynonymous sites and splice sites in the gene. Because the substitution rate at CpG sites was much higher than that in other regions, the substitution rates and expected numbers of substitutions at CpG and non-CpG sites were estimated separately using the following equation:

$$\mathrm{EN} = \sum_{i=1}^{n}\left(\frac{M_{\mathrm{CG}_i} \times N_{\mathrm{CG}}}{S_{\mathrm{CG}} \times C_i} + \frac{M_{\mathrm{NCG}_i} \times N_{\mathrm{NCG}}}{S_{\mathrm{NCG}} \times C_i}\right)$$

where n is the number of samples, M_{CG_i} is the number of synonymous mutations at CpG sites in the ith sample, M_{NCG_i} is the number of synonymous mutations in non-CpG sites in the ith sample, S_{CG} is the number of synonymous sites at CpG sites in the genome, S_{NCG} is the number of synonymous sites at non-CpG sites in the genome, N_{CG} is the number of nonsynonymous sites and splice sites at CpG sites in a gene, N_{NCG} is the number of nonsynonymous sites and splice sites at non-CpG sites in a gene, C_i is the fraction of sequence coverage in the genome in the ith sample (usually the fraction of coding regions that have more than 20× sequence depth for whole-exome sequencing) and EN is the expected number of nonsynonymous and splice-site substitutions in a gene.

Second, the coding indel rate was estimated by dividing the number of coding indels in a sample by the number of coding sites in the genome. For each gene, the expected number was calculated by multiplying the coding indel rate by the coding length of a gene as follows:

$$\mathrm{EI} = \sum_{i=1}^{n}\frac{I_i \times L}{S \times C_i}$$

where I_i is the number of coding indels in the ith sample, S is the number of coding sites in the genome, L is the coding length of the gene and EI is the expected number of coding indels in a gene.

Third, as regions of focal amplification and homozygous deletion are much broader than gene regions, the number of focal amplifications and homozygous deletions affecting a gene in a sample is 0 or 1 and is not influenced by gene length. Therefore, the expected number of these events is the same for all genes. The expected numbers of focal amplifications and homozygous deletions were estimated separately by dividing the total length of the focal amplification or homozygous deletion region in a sample by the length of the genome as follows:

$$\mathrm{EA} = \sum_{i=1}^{n}\frac{A_i}{G \times C_i}$$

$$\mathrm{ED} = \sum_{i=1}^{n}\frac{D_i}{G \times C_i}$$

where A_i is the total length of focal amplifications in the ith sample, D_i is the total length of homozygous deletions in the ith sample, G is the length of the genome, EA is the expected number of focal amplifications in the gene and ED is the expected number of homozygous deletions in the gene.

Fourth, the expected number of protein-altering mutations was calculated by aggregating the expected numbers of nonsynonymous and splice-site substitutions in CpG and non-CpG sites, coding indels, focal amplifications and homozygous deletions as follows:

$$\mathrm{E} = \mathrm{EN} + \mathrm{EI} + \mathrm{EA} + \mathrm{ED}$$

where E is the expected number of protein-altering mutations in a gene.

doi:10.1038/ng.3126

Fifth, tests of the significance of each gene were performed by assuming a Poisson distribution of mutation number. Adjustment for multiple testing was performed using the Benjamini-Hochberg method[8].

Inactivation bias method. The number of missense mutations was compared to the number of inactivating mutations (nonsense, frameshift and splice site) using a χ^2 test.

Analysis of sequencing center bias. Because multiple centers participated in this study, we sought to control for the influence of differences in mutation calling strategy, which might promote a gene to significance merely because of a bias in the variant callers used. Many studies do not use multiple callers and therefore have no way to control for these biases. For each gene with more than five variants, we counted the number of subjects for whom the gene was called for each center. These counts were compared to the total number of subjects using the χ^2 test. The results of the analysis for center bias are presented in **Supplementary Table 11**.

Analysis of subclone bias. Oncogenic driver events in a given tumor should exhibit allele fractions that are roughly the same as the mean allele fraction for the entire sample for any given subject. We separated oncogenic (driver) events from recurrent passenger events by comparing the allele fraction of mutations in candidate genes to the matched mean allele fraction of the sample, across all samples in the cohort. First, the mean somatic allele fraction was calculated for each subject (AF*s*). Next, for each variant in each gene, the allele fraction for the variant (AF*g*) was compared to the AF*s* in the respective subject. We calculated the fraction of events where AF*g* was less than AF*s* and generated a *P* value using a one-sided pairwise Wilcoxon test where the alternative hypothesis was that AF*g* was less than AF*s* (always with respect to the relevant subject). The histogram of all allele fraction biases (sum(AF*g* < AF*s*)/*n*, where *n* is variant count) is shown in **Supplementary Figure 30**. Selected significantly mutated genes are plotted individually to show how known drivers are distributed by this test. Note that several tumor-suppressor genes exhibited enrichment above the average allele fractions (for example, *RB1* and *TP53*). In these cases, the genes were typically both mutated and underwent loss of heterozygosity (LOH) for the wild-type allele. The results of subclone bias testing for all genes with more than five mutations are presented in **Supplementary Table 12**.

Copy number analysis, tumor purity and adjustment of mutated allele frequency. Initial copy number estimates were obtained by comparing read depth information for tumor and normal samples using VarScan2 (ref. 43). Depth estimates were then segmented using circular binary segmentation (CBS) as implemented in the DNAcopy package in R[44]. We used the JISTIC[45] program to generate a combined copy number matrix file. The VCRome2.1 probe locations were used as marker positions for copy number analysis. We then used JISTIC to calculate the significance for copy number gains and losses. Focal amplification at the *TERT* locus was determined using the average read depth of each captured target region.

Evaluation of tumor ploidy and purity. Using bam files from normal and tumor samples, read depth was calculated for each captured target region. After normalization by the number of total reads and GC content using regression analysis, the tumor/normal depth ratio was calculated, and values were smoothed using the moving average. Copy number peaks were then estimated using wavelet analysis, and each peak was approximated using Gaussian models. Hidden Markov models (HMMs) with the calculated Gaussian peaks were constructed, and copy number peaks were linked to genomic regions. The allelic imbalance for each copy number peak was calculated on the basis of heterozygous SNPs within the assigned region, and imbalance information and peak distances were further analyzed by model fitting where the optimal solution for a copy number peak was determined using vector matching, yielding estimated copy number and tumor purity and ploidy data simultaneously. Detailed algorithms will be described elsewhere (H.U., S.Y., K.T. and H.A., unpublished data).

HBV integration analysis. *HBV integration detection.* Viral genomes (HBV, NC_003977.1; HPV-16, NC_001526; HPV-18, NC_001357; HTLV-1, NC_001436) were downloaded from NCBI and included in the reference files when reads were mapped by BWA. No read was mapped to a virus other than HBV. To achieve more precise HBV mapping, we mapped all reads to the HBV reference sequence using the *q*-gram and Smith-Waterman method. An 11-mer *q*-gram was first applied to both strands of the HBV reference, and reads with 15 or more hits were subjected to Smith-Waterman alignment. The other end of each read was mapped to the hg19 human sequence using BWA. Finally, HBV integration sites were clustered by genomic position with a window size of 300 bp (approximately equal to the library fragment size), and sites with more than three supporting reads were used in the analysis.

Randomization test of HBV integration and copy number breakpoints. The 7,891 copy number breakpoints and 1,039 HBV integration sites were detected in 70 HBV-positive samples. Coexistence of the copy number breakpoints and HBV integration sites was examined using a 500-kb window size. To show statistical significance, we performed a randomization test by switching the position of the HBV integration sites to the same number of integration sites observed in the normal sample of other cases. We repeated this switching 100,000 times to yield distributions and estimated the *P* value.

Verification of single-nucleotide variation. We validated our mutation calls for frequently mutated genes (**Supplementary Table 31**) by resequencing samples using the Ion Proton sequencer (Life Technologies). Details are provided in the **Supplementary Note**.

Sanger sequencing of the *TERT* promoter. Bidirectional sequencing of the *TERT* promoter region was completed for 519 HCC samples. PCR runs were set up using 20 ng of genomic DNA, 10 μM manually designed primers (**Supplementary Table 32**) and KAPA HiFi DNA polymerase (Kapa Biosystems, KK2612). Touchdown PCR was performed with the following parameters: an initial denaturation at 98 °C for 5 min followed by 10 cycles of 98 °C for 30 s, 72 °C for 30 s and 72 °C for 1 min (decreasing the annealing temperature by 1 °C per cycle). The reaction then continued with 30 cycles of 98 °C for 30 s, 63 °C for 30 s and 72 °C for 1 min followed by a final extension at 72 °C for 5 min. The PCR products were purified with a 1:15 dilution of Exo-SAP, diluted by 0.6× and cycle sequenced for 25 cycles using a 1:64 dilution of BigDye Terminator v3.1 reaction mix (Applied Biosystems, 4337456). Finally, reactions were precipitated with ethanol, resuspended in 0.1 mM EDTA and analyzed on ABI 3730xl sequencing instruments using the Rapid36 run module and 3xx base-caller. SNPs were identified using SNP Detector software and were validated visually with Consed.

Analysis of mutation patterns and signatures. Mutation patterns for cases with hypermutation and IHCC cases were distinct from those for HCC cases (**Supplementary Figs. 4** and **21**), and cases with a small number of mutations cannot accurately represent the frequency of mutational patterns; therefore, cases with hypermutation, IHCC cases and cases with fewer than 40 mutations were excluded from further mutation pattern analysis.

The number of each of 96 possible somatic substitution types, 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 with the bases immediately 5′ and 3′ to each substitution in coding regions, was counted for each sample. The frequency of each of these substitutions was determined by dividing each count by the total number of substitutions, and the resulting frequencies were used for principal-component analysis. Principal-component analysis was implemented using the R command prcomp with the scaling option on. We used Wilks' λ test to evaluate the significance of the mean vector differences in different populations. We applied NMF to the 96-substitution pattern using published software[13], running 1,000 iterations of NMF with each NMF run iterated until convergence was achieved (10,000 iterations without change) or until the maximum number of 1,000,000 iterations was reached. We used another published software package[14] for model selection in NMF (selecting the input number of mutational signatures). Details on model selection for our NMF analysis are provided in the **Supplementary Note** and **Supplementary Figures 31–35**.

Pathway analysis. We used gene sets from MsigDB C2.all as pathway data sets. To assign *P* values representing the enrichment of mutations in pathways, we first checked whether a gene had at least one non-silent mutation or overlapped with focal CNAs for each sample in a given pathway (gene set). If so, we referred to such a gene as a 'mutated gene' for a sample. We then computed a population frequency for pathways with at least one mutated gene in the given

pathway and divided the frequency by the total length of the unioned exons of all genes in the pathway to correct for the greater number of mutations in longer genes. This quotient was used as a test statistic. We used a bootstrapping approach to calculate *P* values. In the bootstrapping approach, we randomly selected as many genes as in the given pathway from all genes in the genome and then calculated the statistic. We repeated this sampling 2,000 times, calculating a fraction corresponding to the number of sampling results in which a statistic value was greater than or equal to the value in the observed data. This fraction was used as a *P* value.

To find intensively mutated gene modules in liver cancer tissue using the identified significantly mutated gene sets from MsigDB analysis, we used Pathway Commons[15] data for the whole unbiased human gene network and integrated the gene sets into this network. All pairs of gene relationships were weighted by how many mutated genes were shared by the two genes (shared ratio). These gene relationships constituted the gene network. The whole network was split into one large connected network and some isolated small networks. To extract gene modules, we recursively eliminated edges with low shared ratio values and distinguished into the smaller modules. Although the recursive edge elimination procedure gradually clarifies tightly connected gene modules, gene modules were rarely isolated from the whole network. Using this compression process and some additional manual curation, we finally selected ten representative modules that were intensively mutated in liver cancer tissues. We took essentially the same approach as described above to calculate *P* values for mutation enrichment and mutual exclusivity for a gene pair or combination of modules. For mutation enrichment, we used all genes in a pair of modules. For mutual exclusivity, if a module had at least one mutated gene, we referred to such a module as an 'impaired module' and computed a frequency of impaired modules for each sample.

Outcome analysis from non-negative matrix factorization signatures. NMF signature values were merged with annotated clinical data for our cohort. We performed calculations using standardized signature values to control for differences in mutational rate among the subjects. For the standardized data, the contributions of each signature within a subject summed to 1. We performed Cox proportional hazards analysis[46] using the R[44] survival package, factoring in all three signature components (signature A, signature B and signature C), age at diagnosis and histological tumor grade.

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
43. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
44. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2010).
45. Sanchez-Garcia, F., Akavia, U.D., Mozes, E. & Pe'er, D. JISTIC: identification of significant targets in cancer. *BMC Bioinformatics* **11**, 189 (2010).
46. Cox, D.R. & Oakes, D. *Analysis of Survival Data* (Chapman & Hall/CRC, Boca Raton, FL, 1984).

doi:10.1038/ng.3126