nature genetics

Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma

Wing-Kin Sung[1–4,16], Hancheng Zheng[5,16], Shuyu Li[6,16], Ronghua Chen[7,16], Xiao Liu[5,16], Yingrui Li[5], Nikki P Lee[1], Wah H Lee[4], Pramila N Ariyaratne[4], Chandana Tennakoon[2,3], Fabianus H Mulawadi[4], Kwong F Wong[1,8–10], Angela M Liu[1,8–10], Ronnie T Poon[1], Sheung Tat Fan[1], Kwong L Chan[1], Zhuolin Gong[5], Yujie Hu[5], Zhao Lin[5], Guan Wang[5], Qinghui Zhang[5], Thomas D Barber[6], Wen-Chi Chou[6], Amit Aggarwal[6], Ke Hao[7], Wei Zhou[7], Chunsheng Zhang[7], James Hardwick[7,11], Carolyn Buser[7], Jiangchun Xu[12], Zhengyan Kan[12], Hongyue Dai[7], Mao Mao[11,12], Christoph Reinhard[6], Jun Wang[5,13,14] & John M Luk[1,8–10,15]

To survey hepatitis B virus (HBV) integration in liver cancer genomes, we conducted massively parallel sequencing of 81 HBV-positive and 7 HBV-negative hepatocellular carcinomas (HCCs) and adjacent normal tissues. We found that HBV integration is observed more frequently in the tumors (86.4%) than in adjacent liver tissues (30.7%). Copy-number variations (CNVs) were significantly increased at HBV breakpoint locations where chromosomal instability was likely induced. Approximately 40% of HBV breakpoints within the HBV genome were located within a 1,800-bp region where the viral enhancer, X gene and core gene are located. We also identified recurrent HBV integration events (in ≥4 HCCs) that were validated by RNA sequencing (RNA-seq) and Sanger sequencing at the known and putative cancer-related *TERT*, *MLL4* and *CCNE1* genes, which showed upregulated gene expression in tumor versus normal tissue. We also report evidence that suggests that the number of HBV integrations is associated with patient survival.

Hepatocellular carcinoma is a common solid tumor worldwide and represents the third leading cause of cancer deaths[1,2]. HBV is a major etiologic agent that is endemic in China, Southeast Asia and sub-Saharan Africa. Individuals with chronic HBV infection are at increased risk of developing HCC, especially those with chronic liver disease and cirrhosis[3,4]. There are three reported mechanisms by HBV promotes carcinogenesis[5,6]: (i) expression of viral protein, in particular, from viral gene X (HBx), to modulate cell proliferation and viability, (ii) integration of HBV DNA into the host genome to alter the function of endogenous genes or induce chromosomal instability and (iii) accumulation of genetic damage due to hepatic inflammation mediated by virus-specific T cells. The present study focuses on the events of HBV integration and their effects on the HCC genome using whole-genome sequencing and integrated expression profiling analyses.

The presence of integrated HBV DNA sequences in cellular DNA from human HCCs was first reported in the early 1980s[7–10]. Afterwards, many studies were carried out to further investigate HBV integration[11–16]. Most observed HBV integrations are not recurrent (they were each only identified in one sample). The first reported recurrent HBV integration event was found to be located at the human *TERT* gene in two liver tumor samples[13,14]. Subsequently, five more recurrent integrations at *FAR2*, *ITPR1* (also known as *IP3R1*), *IRAK2*, *MAPK1*, *MLL2* and *MLL4* genes were identified[12–15]. Nevertheless, these studies were limited to small sample sizes with no clinical annotation, and the HBV integration events were largely restricted to the sequences in proximity to Alu repeats (due to technical limitations on long-range PCR; see Online Methods). We therefore conducted whole-genome sequencing and analyzed 88 Chinese individuals with HCC from an HBV-endemic area. Our work provides the first unbiased, genome-wide, single-base resolution HBV integration map in HCC, revealing new recurrent HBV integration sites and molecular mechanisms. We also show that HBV integration alters chromosome stability and gene expression levels and shortens the overall survival of affected individuals.

This study surveyed tumor and adjacent non-tumor liver genomes extracted from 81 HBV-positive and 7 HBV-negative HCC samples from individuals who underwent curative primary hepatectomy or liver transplantation (Online Methods). Demographic and clinicopathological

[1]Department of Surgery, University of Hong Kong, Hong Kong. [2]School of Computing, National University of Singapore (NUS), Singapore. [3]NUS Graduate School for Integrative Sciences & Engineering, National University of Singapore, Singapore. [4]Department of Computational & Systems Biology, Genome Institute of Singapore, Singapore. [5]Beijing Genomics Institute, Shenzhen, China. [6]Eli Lilly & Co., Indianapolis, Indiana, USA. [7]Merck Research Laboratories, Boston, Massachusetts, USA. [8]Department of Pharmacology, NUS, Singapore. [9]Department of Surgery, NUS, Singapore. [10]Cancer Science Institute, NUS, Singapore. [11]Asian Cancer Research Group, Inc., Wilmington, Delaware, USA. [12]Pfizer Oncology, San Diego, California, USA. [13]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [14]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [15]Present address: Department of Oncology, Roche R&D Center (China) Ltd., Shanghai, China. [16]These authors contributed equally to this work. Correspondence should be addressed to C.R. (reinhard_christoph@lilly.com), J.W. (wangj@genomics.org.cn) or J.M.L. (john.luk@roche.com).

Received 23 January; accepted 30 April; published online 27 May; doi:10.1038/ng.2295

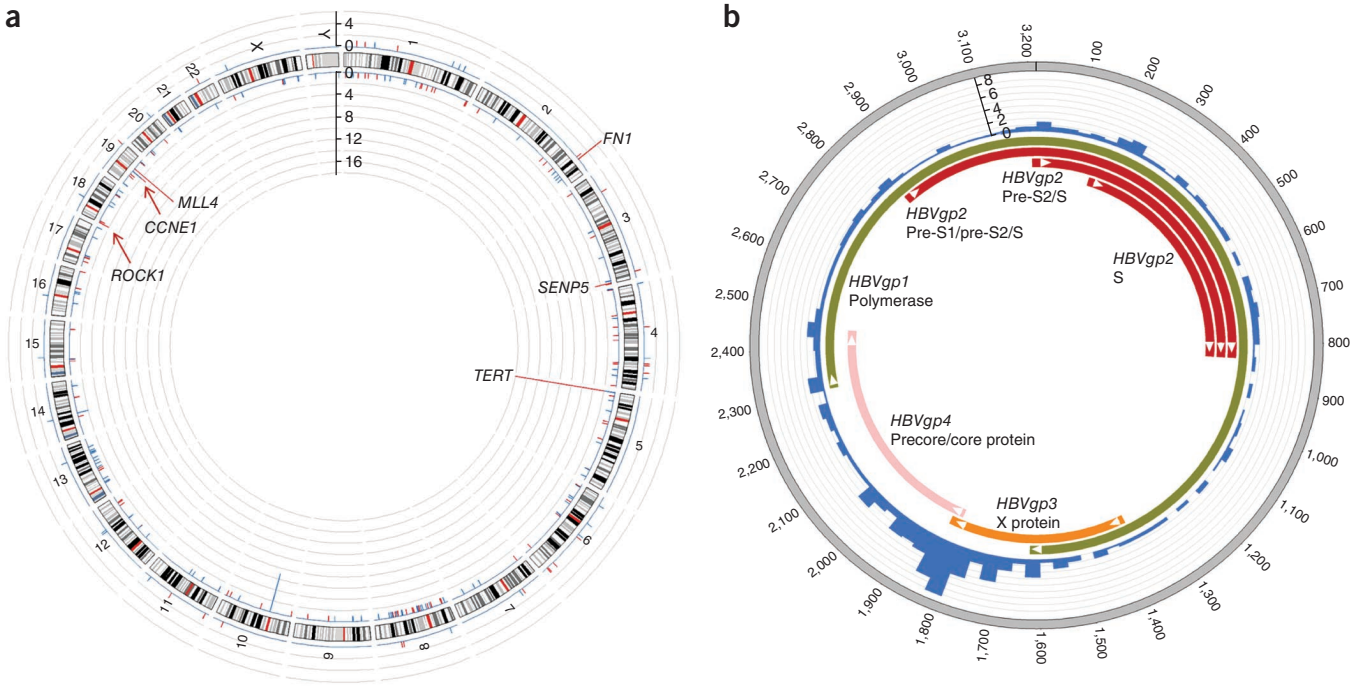

Figure 1 Visualization of HBV integration breakpoints in the HCC and HBV genomes. (**a**) Each bar represents the frequency of HBV integration breakpoints at a particular locus in the human genome (hg19). Red and blue bars correspond to the HBV integration events at the RefGene (exons, introns and promoter) and intergenic regions, respectively. Tumor and adjacent, non-tumor samples with HBV integrations are shown on the inner and outer circles, respectively. The scale bar indicates the number of tumors or non-tumor tissues. Chromosome numbers are shown. (**b**) The frequency of integration breakpoints at different loci in the HBV genome (NC_003977) is shown as a blue histogram. The locations of the genes encoding HBV polymerase (green), core (pink), S (red) and X (orange) proteins are shown. Genomic positions are numbered.

data for these individuals are summarized (**Supplementary Table 1**). The sequencing depth and coverage of all 176 (2 × 88) libraries are described (**Supplementary Table 2**). A total of 399 HBV integration breakpoints were discovered, with each supported by at least 2 paired-end reads (**Supplementary Table 3**). To confirm the newly discovered recurrent events, we randomly selected 32 breakpoints at the 6 affected genes for PCR analysis in 22 samples and successfully validated 82% of these integration sites (**Supplementary Table 4**). We further examined 8 HBV breakpoints found only in tumors in 8 pairs of HCC and normal tissue samples; all 8 did not occur in the paired non-tumor tissues, signifying that these integrations are somatic events (**Supplementary Fig. 1**).

The 399 detected HBV integration breakpoints were shown to be randomly distributed across the whole genome with a handful of hotspots (**Fig. 1a**). Moreover, in concordance with the current paradigm that HBV is a major etiologic agent in HCC development, HBV integration breakpoints occurred more frequently in tumor (344 events) than in normal (55 events) samples ($P = 5.8 \times 10^{-12}$ by Wilcoxon rank-sum test), representing an average 6.3-fold increase. An important unknown is the frequency of alleles with HBV integration within each individual tumor sample (Online Methods). We estimated that 163 of the 344 integration breakpoints in tumor samples had an integration allele frequency of at least 50%. For the five genes with recurrent HBV integration in tumors, the percentage increased to 67.3% (35/52), with these genes having integration allele frequencies of at least 50% (**Supplementary Fig. 2**).

In an attempt to infer the virus integration mechanism to understand the association between HBV integration and HCC, we surveyed the breakpoints on the HBV genome (**Fig. 1b**). Notably, approximately 40% of breakpoints observed were restricted to the 1,800-bp region of the HBV genome where the viral enhancer, X gene and core gene are located. This strategic viral breakpoint usage may facilitate HBV insertions to form chimeric human fusion genes, corrupt tumor suppressors or impose *cis*-regulatory effects on the expression of downstream genes, thereby dysregulating the transcription network in HCC.

Next, we investigated the prevalence of HBV integrations in HCC. Out of 81 HBV surface antigen (HBsAg)-positive individuals, 75 (92.6%) had HBV integrations: 48 were only in the tumor, 1 was only in the normal tissue, and 26 were in both tumor and normal tissues (**Fig. 2a**). Although HCC tumors tended to have more HBV integration breakpoints than their corresponding non-tumor liver tissues

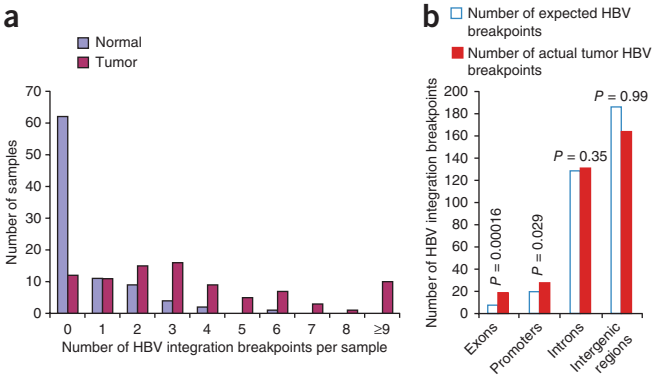

Figure 2 Overall statistics for HBV integration events in HCC tumors. (**a**) Distribution of the number of integration breakpoints per sample. Results are shown for tumors and non-tumor tissues. (**b**) Histogram of the HBV integration breakpoints according to location. The expected (assuming uniform, random distribution) and actual numbers of HBV integration breakpoints are shown. *P* values were computed assuming a binomial distribution.

and integration events were notably large in some samples (up to 17 breakpoints, as observed in sample 75T), the presence of HBV integrations in non-tumor tissues underscores the complexity of the relationship between HBV and carcinogenesis. Among the 26 samples having HBV breakpoints in both tumor and normal tissues, only 1 breakpoint was shared between the tumor and non-tumor samples (in sample 262). Thus, HBV integration patterns are different in the tumor and normal samples.

We then annotated the HBV integration breakpoints to examine their distribution in distinct genomic elements (**Fig. 2b** and **Supplementary Fig. 3**). Most HBV breakpoints in HCC were found near coding genes (209 of 399 breakpoints; $P = 0.003$), suggesting that these gene regions are more likely exposed to the open chromatin where HBV may integrate most efficiently. Of the 344 HBV breakpoints in tumor samples, 179 were located in known coding genes, and these breakpoints were significantly over-represented in exon and promoter (defined as 0 to −5 kb relative to the transcriptional start site) regions, with $P = 0.00016$ and 0.029, respectively. In contrast, for the 33 of 55 HBV breakpoints in non-tumor samples that were located close to genes, breakpoints were mainly located in introns. Such integration site bias suggests that HBV is under positive selective pressure to integrate into exons and promoters in HCC tumors.

In addition, there were few common genes affected by HBV in both normal and tumor tissues. Only two were identified in our cohort, and they affected different individuals through integrations in *HRSP12* (in samples 272T and 276N) and *INPP4B* (in samples 70T and 98N). Notably, there was one gene, *FN1*, that was recurrently affected by HBV integration in five different adjacent, non-tumor samples. These recurring integrations imply that HBV integration events did occur in normal liver tissues but are not directly associated with HCC.

The comprehensive whole-genome deep sequencing in this large-sample cohort has provided us an opportunity to investigate the frequency of recurring tumor-associated integrations in genes. Apart from *FN1*, in which integrations recur in normal samples only (**Table 1**), there were three genes recurrently affected ($n \geq 4$) by HBV integration specifically in the HCC tumor samples—*TERT* ($n = 18$), *MLL4* ($n = 9$) and *CCNE1* ($n = 4$)—accounting for 40.8% (31/76) of the tumor samples with HBV integration. The genomic locations of the tumor-specific HBV integration sites in the three recurrently affected genes are shown (**Fig. 3**). To validate the HBV integration breakpoints in these three genes, we performed PCR and Sanger sequencing at the predicted regions. RNA sequencing (RNA-seq) data, whenever available, were provided to further support the presence of integration events (**Supplementary Fig. 4**).

Next, we examined the effect of HBV integration at the gene expression level by querying gene expression array data generated on the paired tumor and normal samples from the same HCC cohort. Regardless of whether HBV integration was at a promoter, intron or exon, all three genes recurrently affected in the tumor samples showed elevated expression relative to the non-tumor samples (**Fig. 4a**). Comparative analysis of gene expression levels indicated that samples with HBV integration had significantly higher expression of *TERT*, *MLL4* and *CCNE1* than those tumors not harboring HBV DNA integration (**Fig. 4b**). *FN1*, in which HBV integrations were found only in the normal samples, showed reduced expression in tumor samples. The statistically significant recurrence of integrations in the three HBV-affected genes and resultant expression level changes indicate that they may have important roles in HCC.

In addition to integration at *CCNE1*, we discovered two new recurrent HBV integrations at *SENP5* and *ROCK1* (**Fig. 3**), which were identified in three and two samples, respectively. Gene expression arrays showed that both *SENP5* and *ROCK1* were overexpressed with HBV integration, although the *P* values did not indicate that these differences were significant (**Fig. 4a**). The five genes that harbor recurrent HBV integrations may be implicated in tumor development (**Supplementary Note**).

Table 1 Characterization of recurrent genes with HBV integration breakpoints in HCC

Human genes	Integration locations	Affected HBV proteins	Affected samples
TERT	Intron 2	Polymerase, pre-S2/S, S	13T, 268T
	Intron 6	Precore/core protein, core and e antigen	198T
	Promoter	X protein	14T, 22T, 38T, 60T, 64T, 81T
		X protein, precore/core protein	65T, 73T, 261T
		Polymerase	266T
		Precore/core protein	63T
		Precore/core protein, core and e antigen	22T, 266T
		Pre-S1/pre-S2/S	58T
		Pre-S2/S	34T, 261T, 92T
		Pre-S2/S, S	46T
MLL4	Exon 3	Precore/core protein, core and e antigen, X protein	70T, 95T
	Exon 6	Precore/core protein, X protein	186T
	Intron 3	Polymerase, X protein	116T, 159T
		X protein, precore/core protein	159T
		X protein	49T, 53T, 62T
		Precore/core protein, core and e antigen	62T
		Pre-S2/S, S, X protein	71T
CCNE1	Exon 12	X protein	200T
	Intron 2	X protein, precore/core protein	145T
	Intron 4	X protein	106T
	Promoter	Precore/core protein, core and e antigen	268T
SENP5	Intron 2	X protein, precore/core protein	177T
		Precore/core protein, core and e antigen	43T
		Pre-S2/S, S	71T
ROCK1	Intron 1	X protein	203T
	Promoter	Pre-S2/S, S	26T
FN1	Intron 1	Precore/core protein, core and e antigen	95N
	Intron 2	Precore/core protein, core and e antigen	182N
	Intron 3	X protein, precore/core protein	126N
	Intron 5	Polymerase	117N
	Intron 12	Precore/core protein	180N

It has long been hypothesized that integration of viral DNA into the host genome may induce genome instability, thereby acting as one of the molecular mechanisms leading to tumorigenesis[17]. To test whether HBV integration is associated with increased genome aberrations, we examined somatic CNVs in the vicinity of the 344 HBV integration sites in tumors. For the 648 somatic CNVs identified from whole-genome sequencing data, we found that somatic CNVs were positively correlated with the number of observed HBV integration reads at those sites ($P = 6.9 \times 10^{-15}$) (Online Methods and **Supplementary Fig. 5**). This suggested that HBV integration might alter chromosomal stability and cause changes in copy number. Our analysis also indicated that 101 of the integration sites (29%) were located inside somatic CNV regions, 54 sites (16%) were located within 1 Mb of somatic CNV regions and the rest of the sites (189, 55%) were located more than 1 Mb away from any somatic CNV region (**Supplementary Fig. 6**).

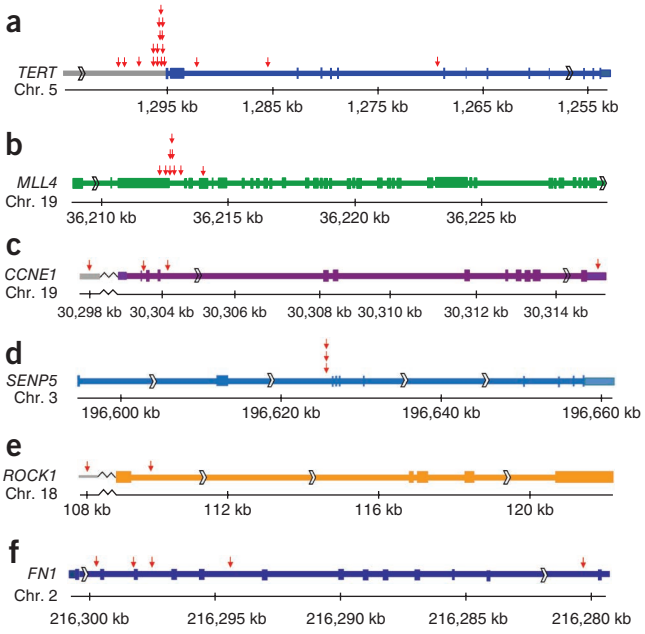

Figure 3 Mapping of HBV breakpoint integration sites. (**a**–**f**) The HBV breakpoint sites on the recurrently affected genes *TERT* (**a**), *MLL4* (**b**), *CCNE1* (**c**), *SENP5* (**d**), *ROCK1* (**e**) and *FN1* (**f**), which were mapped to the human hg19 reference sequence. Each red arrow represents the location of an HBV breakpoint identified from one clinical sample in this study. Chr., chromosome. Boxes represent exons, and the open arrows show the orientation of the genes.

Despite a lack of convincing published data, it is widely believed that HBV integration events may worsen the prognosis of HCC patients. Herein, we also examined the association between different clinical parameters and the number of HBV integrations per tumor (denoted as N_{HBV}). A high number of HBV integration events was defined as $N_{HBV} \geq 3$. We chose this cutoff because it divided the cohort into two relatively balanced groups and achieved optimal statistical power. We observed that individuals with large numbers of HBV integration events were positively associated with serum HBsAg ($P = 0.023$) and α-fetoprotein (AFP) levels ($P = 0.028$; **Fig. 5a**). Individuals with HBV integration seemed to develop HCC at younger ages (**Fig. 5b**). Most notably, individuals with large numbers of HBV integrations ($N_{HBV} \geq 3$) in the tumor survived a significantly shorter time than those with no or low numbers of HBV integrations ($N_{HBV} < 3$) ($P = 0.037$; **Fig. 5c**). Using the Cox proportional hazards model (with N_{HBV} as a continuous variable), we found that the number of HBV integrations was significantly associated with survival ($P = 0.0011$). The AFP levels and tumor size were also significantly associated with survival ($P = 0.0019$ and 0.0009, respectively). The associations between clinical parameters and the number of HBV integrations were also observed using alternative choices of N_{HBV} cutoffs (**Supplementary Table 5**).

Recently, there were reports of exome sequencing and a high-resolution genome map of HCC derived from hepatitis C virus (HCV)-associated tumors[18,19]. The molecular details and clinical impact of HBV integration on the HCC genome, however, remain elusive. Notably, we have shown that HBV integration is significantly associated with the occurrence of HCC at younger ages, which has not been reported before. Previous studies have shown different clinicopathological features in younger and older individuals with HCC and have suggested different mechanisms of hepatocarcinogenesis[20,21]. Younger individuals with HBV-related HCC have a lower incidence of cirrhosis relative to older affected individuals; that is, they tend to develop HCC without undergoing cirrhosis. Therefore, HBV integration at critical oncogenes or tumor suppressor genes may be an important mechanism of hepatocarcinogenesis in the non-cirrhotic livers of younger individuals. Furthermore, the association of greater numbers of HBV integrations with poor survival and high AFP levels is also consistent with findings of more aggressive cancer and higher AFP levels in younger persons with HCC.

We find that HBV integrations are relatively frequent and were present in 76 of 88 samples examined (86.4%). Most HBV integrations seem to be neutral and can occur in non-tumor liver tissues. It is worth noting that, in tumor samples, most genes directly disrupted by HBV integrations are not annotated as cancer related by the

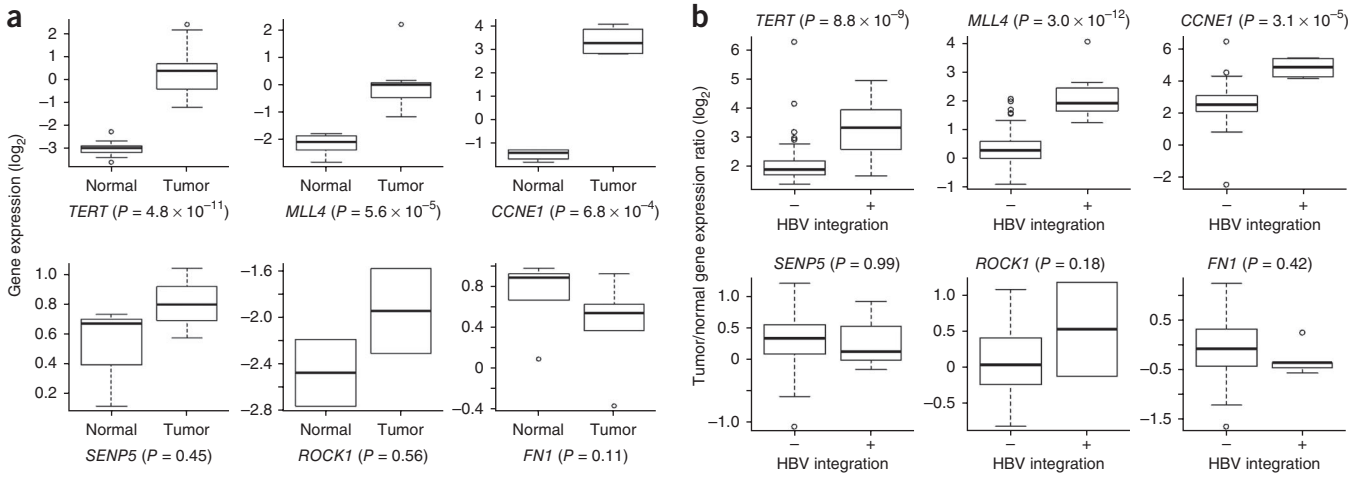

Figure 4 Influence of HBV integration on gene expression in HCC. (**a**) Microarray-derived gene expression levels in tumors versus adjacent, normal liver tissues in cases affected by recurrent HBV-integrated genes. The number of paired samples surveyed in *TERT* ($n = 18$), *MLL4* ($n = 9$), *CCNE1* ($n = 4$), *SENP5* ($n = 3$), *ROCK1* ($n = 2$) and *FN1* ($n = 5$) and their associated *P* values calculated by paired Student's *t* tests are shown. (**b**) Expression of recurrent HBV-integrated genes in HCC samples with or without HBV integration events. Gene expression was normalized by the corresponding adjacent, normal control and therefore is represented as the tumor/normal gene expression ratio. *P* values of unpaired Student's *t* tests are shown. In the box plots, the median (50th percentile) is the middle line, with the bottom and top of the box representing the 25th and 75th percentiles of the data, respectively. The ends of the whiskers represent the lowest and highest data within the 1.5 interquartile range (IQR). IQR was defined as the distance between the lower and upper quartiles of the data.

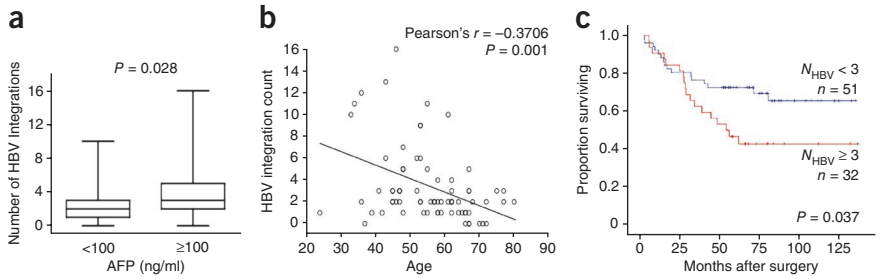

Figure 5 Clinical correlation analysis of HBV integration in HCC. (**a**) The number of HBV integration sites versus serum AFP levels (cutoff at 100 ng/ml). The box plots show the median (horizontal bar), 25th and 75th percentiles, and the whiskers of the plots show the smallest and largest values. The *P* values from unpaired Student's *t* tests are shown. (**b**) Correlation analysis of HBV integrations with age of the affected individuals. (**c**) Kaplan-Meier survival curves for individuals with high ($N_{HBV} \geq 3$) versus low ($N_{HBV} < 3$) numbers of HBV integration breakpoints by log-rank test. Those for whom liver transplantation was the primary treatment were excluded from this analysis ($n = 5$). Herein, we used N_{HBV} of ≥3 as a cutoff to divide the affected individuals into two relatively balanced groups. In addition to Pearson's correlation analysis, we also tried two additional methods (known to be insensitive to outliers) to rank the correlation of HBV integration with age. For Spearman rank correction, $\rho = -0.3378$ and $P = 0.0013$; for Kendall rank correlation, $\tau = -0.2437$ and $P = 0.0019$.

Cancer Gene Census. In contrast, all genes affected by recurrent HBV integrations are known or putative oncogenes and tumor suppressor genes. HBV integrations have certain characteristics that may help the virus control the affected host genes.

In this analysis, over 40% of HBV genomes were cleaved at approximately 1,800 bp and were integrated into the human genome. This bias may be due to the fact that the HBV enhancer and the ORF replication sites are located near this position. We also identify three genes as hotspots of recurrent integrations that are found in 40.8% of the HBV-positive tumor samples. On the basis of Sanger sequencing, many of these breakpoints seemed to form chimeric human-HBV in-frame fusion genes. This may be a mechanism by which the expression of some oncogenes or tumor suppressor genes is affected. Finally, we find that HBV integration breakpoints seem to be associated with increased copy-number variation. This association provides evidence that the chromosomal instability of the HCC genome may originate from HBV integration.

METHODS

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

Accession codes. Whole-genome sequence data have been deposited in the European Genome-phenome Archive (EGA) under the accession ERP001196. Microarray data have been deposited in the Gene Expression Omnibus (GEO) database under accession GSE25097.

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

ACKNOWLEDGMENTS

We gratefully acknowledge Y.-K. Mak and the clinical team of the Division of Hepatobiliary and Pancreatic Surgery (HBP) at Queen Mary Hospital. This study was funded by the Asian Cancer Research Group (ACRG), a not-for-profit organization formed by Eli Lilly, Merck and Pfizer. We thank S. Friend and G. Jin for initiating the establishment of ACRG. We are grateful to former and present members of ACRG, especially K. Blanchard, Y. Turpaz, J. Sedgwick, G. Tucker-Kellogg, G. Gilliland, P. Shaw, N. Gibson and S. Adams.

AUTHORS CONTRIBUTIONS

Clinical tissues were collected by R.T.P. and S.T.F. Clinical annotations and correlation analysis were performed by N.P.L., K.L.C., A.M.L. and K.F.W. DNA and RNA isolation, library construction and sequencing were performed by Y.H. and Z.L. Mapping assembly was conducted by H.Z. and Y.L. HBV-hg19 paired-end analysis was performed by W.-K.S., W.H.L., P.N.A., C.T. and F.H.M. Sanger sequencing was performed by X.L., G.W. and Q.Z. RNA-seq analyses were conducted by W.-K.S. and R.C. CNV analysis was performed by S.L., H.Z. and W.-C.C. Gene expression and/or genotyping analyses were performed by S.L., K.H. and H.D. Biostatistics and bioinformatics analyses were performed by W.Z., H.Z., Y.L., K.H., C.Z. and H.D. Experimental validation was conducted by G.W. and Q.Z. Writing and revisions were carried out by J.M.L., W.-K.S., S.L., R.C., A.A., T.D.B., Y.L., J.X. and Z.K. The project was initiated by J.M.L., W.-K.S., J.W., M.M., C.B., C.R. and Y.L. and was coordinated by Z.G., Y.H. and J.H. M.M. is the consortium leader.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2295.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Center, M.M. & Jemal, A. International trends in liver cancer incidence rates. *Cancer Epidemiol. Biomarkers Prev.* **20**, 2362–2368 (2011).
2. But, D.Y., Lai, C.L. & Yuen, M.F. Natural history of hepatitis-related hepatocellular carcinoma. *World J. Gastroenterol.* **14**, 1652–1656 (2008).
3. Ishikawa, T. Clinical features of hepatitis B virus–related hepatocellular carcinoma. *World J. Gastroenterol.* **16**, 2463–2467 (2010).
4. Bréchot, C., Gozuacik, D., Murakami, Y. & Paterlini-Brechot, P. Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). *Semin. Cancer Biol.* **10**, 211–231 (2000).
5. Zucman-Rossi, J. & Laurent-Puig, P. Genetic diversity of hepatocellular carcinomas and its potential impact on targeted therapies. *Pharmacogenomics* **8**, 997–1003 (2007).
6. Gehring, A.J. *et al.* Profile of tumor antigen–specific CD8 T cells in patients with hepatitis B virus–related hepatocellular carcinoma. *Gastroenterology* **137**, 682–690 (2009).
7. Shafritz, D.A., Shouval, D., Sherman, H.I., Hadziyannis, S.J. & Kew, M.C. Integration of hepatitis B virus DNA into the genome of liver cells in chronic liver disease and hepatocellular carcinoma. Studies in percutaneous liver biopsies and post-mortem tissue specimens. *N. Engl. J. Med.* **305**, 1067–1073 (1981).
8. Koshy, R. *et al.* Integration of hepatitis B virus DNA: evidence for integration in the single-stranded gap. *Cell* **34**, 215–223 (1983).
9. Brechot, C., Pourcel, C., Louise, A., Rain, B. & Tiollais, P. Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma. *Nature* **286**, 533–535 (1980).
10. Chakraborty, P.R., Ruiz-Opazo, N., Shouval, D. & Shafritz, D.A. Identification of integrated hepatitis B virus DNA and expression of viral RNA in an HBsAg-producing human hepatocellular carcinoma cell line. *Nature* **286**, 531–533 (1980).
11. Bonilla Guerrero, R. & Roberts, L.R. The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma. *J. Hepatol.* **42**, 760–777 (2005).
12. Murakami, Y. *et al.* Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. *Gut* **54**, 1162–1168 (2005).
13. Paterlini-Bréchot, P. *et al.* Hepatitis B virus–related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. *Oncogene* **22**, 3911–3916 (2003).
14. Gozuacik, D. *et al.* Identification of human cancer-related genes by naturally occurring Hepatitis B Virus DNA tagging. *Oncogene* **20**, 6233–6240 (2001).
15. Saigo, K. *et al.* Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (*MLL4*) gene and rearrangements of *MLL4* in human hepatocellular carcinoma. *Hum. Mutat.* **29**, 703–708 (2008).
16. Ferber, M.J. *et al.* Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. *Oncogene* **22**, 3813–3820 (2003).
17. Neuveut, C., Wei, Y. & Buendia, M.A. Mechanisms of HBV-related hepatocarcinogenesis. *J. Hepatol.* **52**, 594–604 (2010).
18. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.* **43**, 464–469 (2011).
19. Li, M. *et al.* Inactivating mutations of the chromatin remodeling gene *ARID2* in hepatocellular carcinoma. *Nat. Genet.* **43**, 828–829 (2011).
20. Furuta, T. *et al.* Clinicopathologic features of hepatocellular carcinoma in young patients. *Cancer* **66**, 2395–2398 (1990).
21. Kim, J.H. *et al.* Clinical features and prognosis of hepatocellular carcinoma in young patients from a hepatitis B–endemic area. *J. Gastroenterol. Hepatol.* **21**, 588–594 (2006).

ONLINE METHODS

Samples and data preparation. We studied 88 Chinese individuals diagnosed with HCC who underwent curative primary hepatectomy or liver transplantation at Queen Mary Hospital. All subjects gave written informed consents to use both tumor and non-tumor liver tissues collected between 1980 and 2006 for the study, as previously described[22]. Genomic DNA was purified for at least 30-fold coverage paired-end sequencing, according to our previously reported method[23]. The paired-end reads were mapped on the human reference genome (UCSC build hg19) and the HBV genome (NC_003977)[24]. If a cluster of multiple (>2) read pairs was identified with close mapping positions linking an end of hg19 to an end of HBV, it was considered to be a candidate HBV integration breakpoint. The cutoff was set at two breakpoints, as all seven HBV-negative non-tumor samples did not have any hg19-HBV breakpoint signals that were supported by at least two paired-end reads. We also performed microarray profiling of total RNA from all samples to investigate the potential biological impact of HBV integration. Transcriptome sequencing was applied to nine tumor and normal tissue pairs for validation of some potential breakpoints.

Library preparation and sequencing. Two sequencing libraries with different insert sizes were constructed for each genomic DNA sample. DNA was fragmented with the Covaris E-210 ultrasonicator. By adjusting to the relevant optimal shearing parameters, DNA fragments were set to be concentrated in 170-bp and 800-bp peaks for the relevant libraries. These fragments were purified, and we blunted the ends, added A tails and ligated them with adaptors. After size selection in a gel, 10 to 12 cycles of PCR were performed. The concentrations of the libraries were quantified by Bioanalyzer (Agilent Technologies) and quantitative PCR methods, using the ABI StepOne Plus Real-Time PCR system (Life Technologies). To obtain an optimal cluster number in the flow cell and accurate signal capture, 170-bp and 800-bp libraries were loaded on the flow cell in a 2:3 ratio. Paired-end 90-bp read length sequencing was performed on the HiSeq 2000 sequencer according to the manufacturer's instructions (Illumina).

Mapping and analysis of HBV integration sites. The analysis workflow is shown in **Supplementary Figure 7**. Both 170-bp and 800-bp paired-end fragment libraries (read length of 90 bp) were mapped to the human reference genome (hg19) and the HBV genome (NC_003977). If a paired-end read uniquely mapped with one end to hg19 and with the other end to HBV, it was reported. Due to the short fragment length of the 170-bp library and its relatively long read length, we trimmed all unmapped reads from this library to 40 bp and remapped them. All mapping locations were subjected to a filtering process to remove possible PCR duplicates. Specifically, if there were two or more paired reads that mapped to near-identical locations (± 2 bp), only one of the reads was considered. At this stage, we pooled all hg19-HBV pairs for each sample (170-bp fragment, 800-bp fragment and 170-bp fragments trimmed to 40-bp reads) and clustered them according to their forward and reverse tag mapping locations. Furthermore, to determine the exact fusion point between hg19 and HBV, we extracted all hg19-HBV and HBV-HBV mapped reads from trimmed 170-bp libraries and aligned each read entirely (90 bp) on hg19 and HBV using Blat (–minScore 25, –minIdentity 85). The fusion point was determined by analyzing cases where part of the read aligned to HBV and the other part aligned to hg19. These locations were crosschecked with the clusters of paired-end reads for consistency.

To determine the frequency of the integration events within an individual tumor sample, we estimated the percentage of HBV integration sites by computing the HBV-integrated allele frequency. The HBV-integrated allele frequency is defined as $(a - b)/b$, where a is the average number of paired-end reads covering the sites upstream of the integration site and b is the number of paired-end reads covering the integration site (**Supplementary Fig. 2**).

Unlike the widely used method of HBV-Alu PCR, our approach is unbiased and precise. We can recover HBV integration breakpoints that are far away from Alu repeats. In fact, 7.8% of the discovered HBV integration breakpoints were more than 10,000 bp away from Alu repeats (**Supplementary Fig. 8**). One example is the recurrent HBV-*CCNE1* integration breakpoint that is 2.5 Mb away from Alu repeats. These breakpoints are unlikely to be detected by the HBV-Alu PCR method.

Processing RNA-seq data. RNA-seq libraries were sequenced as paired-end 90-bp sequence tags using the standard Solexa pipeline. Sequence tags were then mapped to hg19 with TopHat software using default parameters[25]. Reads with paired-end tags that mapped uniquely to the genome were assigned to their site of origin. Reads with only one tag that mapped unambiguously to the genome were rescued and assigned to their most probable site of origin. We then quantified the transcript levels of each gene in reads per kilobase of exon model per million mapped reads (RPKM)[26].

PCR and Sanger sequencing validation. PCR primers giving products of approximately 200 bp were designed to capture the breakpoints of each HBV integration site. To investigate structural variation in the integrated HBV sequence, we also designed a pair of primers to amplify the entire integrated sequence. All PCR experiments were performed with both tumors and matched adjacent normal liver tissues on a GeneAmp PCR System 9700 thermal cycler (Life Technologies). For each PCR product, we used Applied Biosystems 3730× DNA analyzers to perform dye terminator Sanger sequencing.

Gene expression microarray analysis. Microarray gene expression profiling and subsequent data processing were carried out as previously described[27]. Differential gene expression in tumors relative to their matched adjacent normal tissues was analyzed by a pairwise two-tailed t test. To assess the statistical significance of overexpression of genes affected by HBV integration, tumor/adjacent normal tissue gene expression ratios were compared between the samples with or without HBV integration by two-tailed t tests.

Integrative analysis of HBV integration and CNVs. CNV analysis was performed following the SegSeq method[28]. Focal CNVs were further identified using Genomic Identification of Significant Targets in Cancer (GISTIC2.0)[29]. The concordance between CNVs identified from whole-genome sequencing and from microarrays was evaluated. The gene- and sample-level copy-number values from array-based and WGS-based data were compared across all protein-coding genes (17,720) and the entire cohort. The resulting correlation was highly significant (Pearson's correlation coefficient = 0.727; $P < 2.2 \times 10^{-16}$). We also computed the copy-gain and copy-loss frequencies (percent of samples) for each gene across all samples (copy gain > 2.5, copy loss < 1.5). Copy-gain and copy-loss frequencies between the two data sets were also highly correlated (Pearson's correlation coefficient = 0.948 for copy gain and 0.957 for copy loss; $P < 2.2 \times 10^{-16}$ for both). For each of the 344 HBV integration sites in tumors, the distance between the HBV integration site and the nearest CNV region was computed. To test the statistical significance of HBV integration–associated CNV events, we randomized the sample IDs, either for the HBV integration sites or for the CNV regions, and repeated the analysis 100,000 times. The distribution of the number of HBV integration sites colocalizing with CNV regions was used to determine the probability of the observation of such sites.

22. Hao, K. *et al.* Predicting prognosis in hepatocellular carcinoma after curative surgery with common clinicopathologic parameters. *BMC Cancer* **9**, 389 (2009).
23. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29**, 735–741 (2011).
24. Chen, Y. *et al.* Complete genome sequence of hepatitis B virus (HBV) from a patient with fulminant hepatitis without precore and core promoter mutations: comparison with HBV from a patient with acute hepatitis infected from the same infectious source. *J. Hepatol.* **38**, 84–90 (2003).
25. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
26. 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).
27. Lamb, J.R. *et al.* Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PLoS ONE* **6**, e20090 (2011).
28. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
29. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

doi:10.1038/ng.2295