HBV integrations reshaping genomic structures promote hepatocellular carcinoma

Zhaoyang Qian,1,2 Junbo Liang,3 Rong Huang,3,4 Wei Song,3 Jianming Ying,5 Xinyu Bi,1 Jianjun Zhao,1 Zhenyu Shi,2,6,7 Wenjie Liu,1 Jianmei Liu,1 Zhiyu Li,1 Jianguo Zhou,1 Zhen Huang,1 Yefan Zhang,1 Dongbing Zhao,8 Jianxiong Wu,1 Liming Wang,1 Xiaochen,9 Rui Mao,1 Yanchi Zhou,1 Lei Guo,5 Hanjie Hu,1 Dazhuang Ge,1 Xingchen Li,1 Zhiwen Luo,1 Jinjie Yao,1 Tengyan Li,1 Qichen Chen,1 Bingzhi Wang,5 Zhewen Wei,1 Kun Chen,9 Chunfeng Qu,9 Jianqiang Cai,1,10,11 Yuchen Jiao,11 Li Bao,2,6,7 Hong Zhao1,10,11

ABSTRACT

Objective Hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC), mostly characterised by HBV integrations, is prevalent worldwide. Previous HBV studies mainly focused on a few hotspot integrations. However, the oncogenic role of the other HBV integrations remains unclear. This study aimed to elucidate HBV integration-induced tumourigenesis further.

Design Here, we illuminate the genomic structures encompassing HBV integrations in 124 HCCs across ages using whole genome sequencing and Nanopore long reads. We classified a repertoire of integration patterns featured by complex genomic rearrangement. We also conducted a clustered regularly interspaced short palindromic repeat (CRISPR)-based gain-of-function genetic screen in mouse hepatocytes. We individually activated each candidate gene in the mouse model to uncover HBV integration-mediated oncogenic aberration that elicits tumourigenesis in mice.

Results These HBV-mediated rearrangements are significantly enriched in a bridge-fusion-bridge pattern and interchromosomal translocations, and frequently led to a wide range of aberrations including driver copy number variations in chr4q, 5p (TERT), 6q, 8p, 16q, 9p (CDKN2A/B), 17p (TP53) and 13q (RB1), and particularly, ultra-early amplifications in chr8q. Integrated HBV frequently contains complex structures correlated with the translocation distance. Paired breakpoints within each integration event usually exhibit different microhomology, likely mediated by different DNA repair mechanisms. HBV-mediated rearrangements significantly correlated with young age, higher HBV DNA level and TP53 mutations but were less prevalent in the patients subjected to prior antiviral therapies. Finally, we recapitulated the TONSL and TMEM655 amplification in chr8q led by HBV integration using CRISPR/Cas9 editing and demonstrated their tumorigenetic potentials.

Conclusion HBV integrations extensively reshape genomic structures and promote hepatocarcinogenesis (graphical abstract), which may occur early in a patient’s life.
to the expressions of viral proteins, viral DNA integrations can result in aberrant expressions of oncoproteins such as TERT and/or tumour suppressor genes, both of which contribute to HBV-associated carcinogenesis. Moreover, HBV integrations often correlate with chromosomal abnormalities in HCCs, suggesting that the genomic instability contributes to HBV integration-associated carcinogenesis. However, the molecular insights into this process still need to be improved. Notably, recurrent HBV integrations appear to be restricted to a small cohort of genes (TERT, MLL4 and CCNE1). In contrast, most integrations impose minimal functional impact on the host genes.

The disruption by canonical HBV integrations in the host genome is generally less than 200bp. In contrast, non-canonical HBV integrations which have been reported decades ago can lead to a much larger scale of alterations in the genome and megabase-size fragment copy number variations (CNVs). However, the information on complete genomic reorganisation from an integration is fairly limited due to the bias in next-generation sequencing (NGS) towards shorter reads.

To shed new light on the mechanisms by which HBV integrations fuel the tumourigenesis of HCC, we comprehensively reconstructed the genomic landscape of HBV integrations by combining the whole genome sequencing (WGS), transcriptome and third-generation sequencing data sets. As a result, we identified a collection of genomic alterations associated with dysregulated gene expression. Furthermore, we recapitulated the corresponding integration profiles using DNA editing mediated by CRISPR-Cas9 and confirmed the oncogenic role of these genomic changes.

MATERIALS AND METHODS

WGS of genomic DNAs was performed on HiSeq X Ten (Illumina; San Diego, California, USA) and Oxford Nanopore Technology (ONT) platforms (Oxford, UK). Among the 50 ONT samples, 26 were randomly selected from 100 NGS samples, and another 24 were from a novel cohort of young HCC patients (age <35). The RNA-seq aims to investigate the impact of non-canonical integrations on gene expression, so we select 25 in 100 NGS samples with a relatively high number of non-canonical integrations and simultaneously have sufficient remaining tissue for RNA sequencing. One aim of this study is to investigate why some young people could develop HCC, so we included a substantial proportion (54/124) of tumours from young patients (age <35) in our cohort (further details are provided in online supplemental materials and methods).

RESULTS

Common genomic aberrations and HBV integrations in HCC

With WGS, we analysed the tumour DNA samples derived from HCC patients (n=100; online supplemental table 1), in which small mutations (online supplemental table 2) and structural variations (SVs) were identified. In 62% of cases, we confirmed mutations and SVs in TP53 that accounted for the most frequently mutated gene. At the same time, the hotspot mutations (C228T and C250T) were also accumulated at the TERT promoter (n=21). Additional genetic alteration occurred in AXIN1 (n=31), CTNNB1 (n=13), ARID1A (n=11), RB1 (n=9) and SETD2 (n=8). We found recurrent gains on chromosome arms 1q, 5p, 6p, 8q, 17q and 20q, consistent with the The Cancer Genome Atlas (TCGA) HCC dataset, whereas loss of heterozygosity was mostly associated with chromosome arms 1p, 4q, 8p, 9p, 13q, 16q and 17p (online supplemental figure 1).

The HBV genotype was determined using the reference HBV genomes of different genotypes (A, B, C, D, E, F, G and H). Among all samples, 87 were positive for HBV DNA (online supplemental methods), and genotypes C (n=84) and B (n=13) were the dominant types (figure 1A). We further investigated the integration of HBV, and there were 482 HBV integrations identified totally in 84 clinical samples with a median of 5 in each sample (online supplemental table 3), in which 88.8% were supported by at least eight reads after removing the duplications, indicating robust callings. In interrogating the HBV integrations affecting exons (1.9%), introns (26.8%) and the upstream genomic regions (<10kb) the HBV integrations were discovered as the recurrent gene loci, in line with previous studies.

Non-canonical HBV integration sites are generally CNV break points

To understand the biological role of the dominant non-hotspot integrations, we analysed the distribution of 482 integration sites together with CNV segmentations. In total, 71 integration sites were mapped to the non-unique genomic regions such as telomeres and centromeric satellite DNA, thus obscuring the precision of location. However, we were able to determine the genomic locus for the rest sites. There were 190 integration sites within 10kb of CNV breakpoints, of which 140 were within 1kb from a CNV breakpoint, much more prominent than the random simulation (44.9% vs 0.023%, p<10−10). We used a piecewise least square fitting algorithm to resolve the CNV flanking the integration sites at a high-precision level (online supplemental methods). Further, we confirmed that 192 sites adjacent to a CNV edge were bona fide CNV breakpoints despite the position deviation due to the minor inaccuracy of the primary CNV algorithm (online supplemental figure 2). Additional 31 integration sites distant (>300kb) from any CNV edge were characterised as breakpoints of small focal CNVs, which were initially missed by CNV calling. In total, 223 of 411 integration sites were bona fide CNV breakpoints.

We categorised these integration sites into two groups (denoted as α and β sites) based on the direction of host-viral fusions along the reference strand (hg19) (online supplemental figure 3A). A canonical HBV integration features a fragment of HBV DNA inserted into host DNA, resulting in an α-β pair of integration sites. In 411 precisely located integration sites, 174 were classified as canonical, giving rise to 87 α-β pairs. The distance from α to β sites ranged from −84 to 230 kb, with the majority between −20bp and 100bp (76.5%, 65/87). Small deletions in the human genome between α and β may be due to resection during double-strand breakage (DSB) repair. In 16 cases, the β site is located several bases upstream of α, likely due to the fill-in of 5' overhangs at the host DSBs during DSB repair. Despite the deletion between the α and β sites, 93.1% (162/174) of these canonical integrations were distant (>300kb) from any other CNV edges (figure 1B). Additionally, 237 integrations identified as non-canonical did not accord to the canonical pattern, most (61.6%) have no neighbouring integrations within 10Mb. In contrast to the canonical integrations, 90.7% of non-canonical (215/237) were likewise CNV breakpoints (figure 1B), and half of those non-canonical integrations (48.8%, 105/215) were breakpoints of large CNVs (>10Mb) (online supplemental tables 3 and 4). Some oddity lies in a small group of non-canonical integrations (15/237), which each exhibits cross-over with a nearby (<3kb) host SV to form a balanced translocation, thus do not match...
Non-canonical integrations occur in specific genomic regions

Opposed to the canonical counterparts, non-canonical integrations were less frequently located in gene promoters and coding regions (41.6% vs 60.6%, \( p = 0.0034 \), Fisher’s exact test) (figure 1C) or target hotspot genes such as \( TERT \), \( MLL4 \) and \( CCNE1 \) (8.3% vs 24.4%, \( p = 0.0025 \), Fisher’s exact test) (online supplemental table 3). Non-canonical integrations were found to be more enriched in chr8, chr11 and chr17, whereas canonical ones in chr2 and chr19 (\( p < 0.05 \); Fisher’s exact test) (figure 1D–F). Both \( \alpha \) and \( \beta \) integration sites bear a hotspot around/in the \( TERT \) promoter, where \( \alpha \) sites appear dominant in the non-canonicals, whereas canonical integrations are generally \( \alpha \)-\( \beta \) pairs. Unique to non-canonicals is a hotspot region in the short arm of chr17 close to the centromere (13.0–22.5 Mb), which were primarily \( \beta \) sites (online supplemental figure 3B,C, online supplemental table 3).

Figure 1 Overview of HBV infection and integrations in 100 HCCs. (A) Heatmap depicting HBV genotype and HBV DNA load in 87 HBV-positive samples. (B) Detailed classification of 482 HBV integrations. (C) Functional region distribution of canonical and non-canonical integrations. (D) Distribution of canonical HBV integrations across the human genome, with \( \alpha \) site (strand+) and \( \beta \) site (strand−) shown in different colours. (E) Distribution of non-canonical HBV integrations across the human genome, with colours same as D. (F) Comparison of chromosome distribution between canonical and non-canonical integrations. Significance are presented as *\( p < 0.05 \), **\( p < 0.01 \). (G) Distribution of canonical HBV integrations across the HBV genome. The integrations with a 5’ end viral DNA at the same strand (strand+) or the complementary strand (strand−) of HBV reference are shown in different colours. (H) Distributing non-canonical HBV integrations across the HBV genome, with colours the same as G. Cen, centromere; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; NUM, non-unique mapping; Tel, telomere.
and often resulted in the loss of chr17p, including the TP53 gene. The β sites also dominate non-canonical integrations in chr8, which lead to chr8p deletion and/or chr8q amplification. In addition, both canonical (32.4%) and non-canonical (33.7%) integrations bear the breakpoints enriched in HBV X gene 3' (1600–1900 bp) (figure 1G,H), particularly in the ones spanning the promoter or coding regions (39.2%). Furthermore, in contrast to a lower frequency of canonical integrations (5.7%), there were substantial cases (19.4%) where the non-canonical were incorporated into complex genome rearrangements such as chromoplexy and chromothripsis (online supplemental figure 3D,E, online supplemental method) with a 10.8-fold enrichment compared with the random distribution.

Characterisation of the HBV integrations with third-generation sequencing

Using third-generation sequencing (ONT), we identified 298 integration sites in 50 tumour samples. In 26 samples sequenced with both NGS and ONT methods, 91.2% (134/147) of the total integration sites were supported by reads from both platforms (online supplemental table 3). We obtained the full length of integrated HBV DNA based on the long reads (median ≈10kb) collected and matched integration sites into pairs which indicated a complete HBV integration event (online supplemental methods). We systematically examined 154 such events for the two integration sites’ relative position, strand direction and the neighbouring copy number profiles (online supplemental table 3). Interestingly, 27 such events existed as a fold-back inversion, with 2 identical host-viral junction sites along with the integration of a long symmetric palindrome HBV sequence (figure 2A, online supplemental figure 4A). The two breakpoints were indistinguishable by shot Illumina reads, thus only counted once in the NGS analysis. Altogether, five types of integration events were characterised (figure 2A, online supplemental table 3, online supplemental methods). All type I events, 30 in total, were characterised as α–β pairs, consistent with the canonical pattern. Other integration types identified, in line with the definition of non-canonical integrations, include larger duplications (type II), inverted fusions (type IIIa and IIIb) and remote translocations (type IV) (figure 2A,C). A β–α pair of integrations featured the 10 type II events, with the β site located upstream (2.7 kb to 1.6 Mb).

Figure 2 Five classes of integrations were detected with third-generation sequencing. (A) Diagram of five types of HBV integrations, including types I, II, IIIa, IIIb and IV. The top panels of each subfigure depict the structure of HBV insertion into the human genome, including the relative position of the two breakpoints and the strand in the host genome. The middle panels of each subfigure show the induced CNV pattern corresponding to each integration type. The bottom panels show the junction model of DNA fragments, including strands (arrows) in each integration type. (B) The distance between the two breakpoints of each integration event shown in (A). The two breakpoints were identical for every type IIIa integration; thus, the distance was not displayed. (C) Examples of complex integrated HBV sequences constructed by ONT long reads, including examples representing different integration types. For each case, the multiple HBV fragments are presented in different colours with a strand (arrow), and their arrangements are shown. (D) The segments of simple integrated HBV sequences. Each line depicts the integrated HBV sequence in each event, and colours represent the integration type. (E) Barplot and boxplot display the length distribution of integrated HBV DNA for the five integration types. Simple and complex HBV inserts are presented in different colours in the barplot. CNV, copy number variation; HBV, hepatitis B virus; ONT, Oxford Nanopore Technology.
of the \( \alpha \) site. The genome region from \( \beta \) to \( \alpha \) was duplicated (figure 2B), and the HBV DNA was inserted between two duplicated copies (figure 2A). Type IIIa (n=27), as described above, harboured two identical integration sites as the result of fold-back inversion of two allelic chromosomes (figure 2A, online supplemental figure 4A), whereas type IIIb (n=27), though also as fold-back inversion, was differentiated from type IIIa by two different breakpoints and a non-palindromic inserted HBV sequence, that is, either an \( \alpha-\alpha \) pair or \( \beta-\beta \) pair with a distance of up to 3 Mb in-between (figure 2A and B, online supplemental figure 4B). Both types (IIIa and IIIb) led to a copy number gain/loss on either or both sides (figure 2A). Type IV (n=59) fused two different (distance >10Mb) host chromosomal regions, analogous remote translocations and are mostly interchromosome events (91.5%, 54/59), leading to either a gain or loss in the host genome at large scale (figure 2A). Collectively, a broader spectrum of integration events was revealed from this analysis.

**Complex rearrangements of the integrated HBV DNAs**

Overall, integrations among all types, including type I (5/30), II (4/10), IIIa (27/27), IIIb (6/27) and IV (25/59) exhibited intra-HBV translocations (online supplemental figures 4C–7), suggesting prevalent rearrangements of integrated HBV DNA. Particularly, each type IIIa integration harboured minimal one intra-HBV inversion. In the 26 sequences by both methods, 93% of the intra-HBV rearrangements were confirmed by Illumina short reads. Multiple intra-HBV translocations co-occupy the same integration event, likely resulting from a patchwork in which a series of viral–viral fusion events occurred sequentially via DSB repair (figure 2C and online supplemental figures 4C–7). A higher proportion of intra-HBV translocations were associated with type IV integrations (25/59) than type I and IIIb (11/57) \( (p=0.009, \text{ Fisher’s exact test}) \). It could be explained by the longer primary physical distance between two breakpoints in type IV that might lead to a longer latency of DSB rejoining. During the latency, other free viral DNA fragments could be subsequently ligated to the unsolved DSB before eventually makeup of the complete integration.

Similarly, a positive correlation was established between the complexity of the HBV DNA makeup and the length of the integrated HBV sequence, with a median of 594bp (117–2768bp) and 4119bp (375–38171bp) for simple and complex integrated HBV DNA, respectively (figure 2D,E). For instance, the longest stretch of HBV DNA was associated with type IIIa integrations that all contain complex integrated HBV sequences (online supplemental figure 6).

**Characterisation of integration-mediated SVs**

The interesting structure of type IIIa indicates the occurrence of bridge-fusion-bridge (BFB) cycle, which is characterised by fold-back inversion result from sister-chromatids produced after replication of a broken chromatid that fuse at the location of the break.\(^{21}\) However, a subset of type IIIb (16/27) with short distance (<5kb) between two breakpoints also accords with the BFB pattern.\(^{24}\) Specifically, in a type IIIa integration, one end of an HBV DNA ligated to the DSB of a broken chromatid precedes the S phase of the cell cycle, and the unsolved viral DSB is duplicated in the S phase and then fused (online supplemental figure 4A). In contrast, in type IIIb, an unsolved host DSB replicated during the S phase, and then an HBV DNA ligated the two DSBs (online supplemental figure 4B). The intervals between two host breakpoints in these type IIIb integrations (44–4727bp, median=726bp) were likely created by a long resection through alternative nonhomologous end-joining (alt-NHEJ).\(^{21}\) Consistently, we identified a prevalent asymmetric sequence (15–210bp, median=900bp) at the fold-back point of palindromic HBV sequence in type IIIb, which also suggested a resection before end-joining of replicated viral DSBs (online supplemental figure 6).

We further elucidated the HBV-mediated SV patterns with reference to host SVs of the International Cancer Genome Consortium (ICGC) and TCGA datasets\(^ {25} \) and our HCC samples (online supplemental methods). Only those aberrations affecting the human genome longer than 1 kb were analysed. In contrast to the host SVs of all three datasets, the HBV-mediated SVs exhibited a significant enrichment \( (p<10^{-10}) \) of the BFB pattern and interchromosome translocations (figure 3A), which tends to alter the large scale of the human genome. The enrichment may suggest a positive selection of these HBV-mediated SVs. However, viral integration-mediated SVs require at least one extra step of DSB-rejoining, which usually takes several hours.\(^ {26}\)

This delay suppresses locally resolving the primary host DSB pair due to temperature-dependent Brownian chromatin motion.\(^ {27}\)

We analysed the microhomologies at junction sites through locally assembling reads around breakpoints of SVs and HBV integrations (online supplemental methods). There were significant proportions of integrations with and without microhomology (\( \geq 2 \)bp) junctions, suggesting both enrollment of classical NHEJ (c-NHEJ) and alt-NHEJ (online supplemental figure 8),\(^ {25,28}\) which is error-prone, cell cycle-dependent, with delayed activity, and favours chromosome translocations at high frequency.\(^ {21,26,28,29}\)

The scarcity of integrations with microhomology longer than 5 bp possibly reflects the absence of single-stranded annealing and homologous recombination (HR).\(^ {30}\) We compared the microhomology feature between HBV integrations and diverse types of human SVs. We observed high similarity in microhomology features between the canonical and non-canonical integrations and the host intrachromosome translocations with distance >1 Mb (figure 3B and online supplemental figure 8). Finally, we investigated those integrations that can be paired into a single integration event by nanopore long reads or a canonical pattern.

We observed that 70.7% (87/123) pairs possessed at least one microhomologous integration site. Interestingly, in these pairs, the cases with microhomology at both sites were more scarce than expected (13.8% vs 32.4%; \( p=0.0064 \)), suggesting favour of different DSB-repair mechanisms for the first and second step of end-joining during viral integration (figure 3C,D).

**Functional and clinical relevance of the non-canonical integrations**

Non-canonical (types II–IV) integrations induced gene amplification in TERT (\( n=16 \) samples), CCNE1 (\( n=2 \)), CCND1 (\( n=2 \)) and MLL4 (\( n=1 \)) (figure 4A,C), accompanied by the loss of chr17p and deletion of TP53 (\( n=15 \) ) that mostly result from type IV integrations (figure 4B,D). Integration-induced CNVs also occurred at large scales across the genome, which, in addition to the ones in chr5 and chr17, frequently gave rise to the amplifications of chr8q (\( n=14 \)), 19q13.42 (\( n=7 \)), 1q (\( n=6 \)), 7p (\( n=4 \)), 6p (\( n=4 \)) and 20q13.3 (\( n=4 \)), as well as the deletions of chr8p (\( n=11 \)), 4q (\( n=11 \)), 16q (\( n=8 \)), 6q (\( n=7 \)), 9p (\( n=6 \)), 13q (\( n=6 \)) and 1p (\( n=5 \)) (figures 4 and 5A). Importantly, such genomic alterations appear prevalent in HCC and contain critical tumour driver genes such as ARID1A, CDKN2A and RB1. We reconstructed the possible evolutionary scenario leading to the integration-induced amplification in chr8q, for which the changes in the number of the clock-like mutations (COSMIC...
Hepatology

Figure 3 Characterisation of integration-mediated structural variations. (A) Comparison of patterns between human SVs of three datasets and HBV-mediated SVs. Human SV patterns including LocalINDEL (duplications and deletions with length 1 kb–1 Mb), LocalINV (Inverted fusions with two breakpoints distant 5 kb–1 Mb), IntraChr (all other intrachromosome translocations with two breakpoints distant >1 Mb), BFB (Inverted fusions with two breakpoints distant <5 kb) and InterChr (interchromosome translocations). (B) Heatmap of microhomology-signature correlation matrix between different types of human SVs and HBV-mediated SVs. The details of each SV type are described in online supplemental figure 8. (C) Microhomology at the junction sites of paired integrations (site1 and site 2, corresponding to the same row). The y-axis of the bars denotes the length of microhomology at each site. Colours highlighted microhomology status. (D) Examples of microhomology at the paired integration sites corresponding to C. BFB, bridge-fusion-bridge; HBV, hepatitis B virus; SVs, structural variations.

SBS signature 1 and 5) before and after amplification were used as the reference for timing (online supplemental methods). Most of these amplifications (7/8) likely occurred decades before the clinical diagnoses and even ultra-early (< 10 years old) in the lifetime of patients (5/8) (figure 4E).

We compared the CNVs of TCGA HCC samples (n=361) between groups of HBsAg positive (n=106) and HBsAg negative (n=245) to investigate if the HBV-induced CNVs were over-represented in HBV+ HCCs. The results showed that HBsAg+ samples have significantly higher (p<0.02) frequencies of chr8q
amplification and chr4q, chr16q and chr17p deletions (online supplemental figure 9A), which are among the top rank of HBV-induced CNVs (figure 4A,B). Additionally, we combined the 124 HCCs in our study with a novel cohort of 221 shallow WGS sequencing data of HCCs and compared the CNV spectrum between HBsAg+ (n=248) and HBsAg− (n=97) samples. Although there are some discrepancies, the results are similar to the TCGA analysis in that enrichment of chr8q amplification and chr4q, chr16q and chr17p deletions were observed (online supplemental figure 9B), which suggests a contribution of integration-related CNVs in HBV-positive HCCs.

A substantial proportion (54/124) of the recruited patients were aged under 35 (figure 5A, online supplemental table 1), which allowed us to establish a negative correlation between patients’ age and the frequency of non-canonical integrations (p<10^-5), in contrast to the number of canonical integrations (p>0.05) (figure 5B,C), suggesting that non-canonical integrations may play specific carcinogenic roles in young patients, accelerating tumourigenesis. Higher levels of HBV DNA load and TP53 mutations were significantly associated with the number of non-canonical integrations (p<10^-5; Students’ t-test) (figure 5D–F) but not the canonical integrations (p>0.05), suggesting both genomic instability and the concentration of intracellular free HBV DNA determine the incidence of non-canonical integrations. In addition, fewer integrations were confirmed in the patients previously subjective to antiviral therapies (p=0.055, Wilcoxon rank-sum test) (figure 5G) or with a negative hepatitis B envelope antigen (HBeAg) score (p=0.022, Wilcoxon rank-sum test) (figure 5H), highlighting the importance of infection control and antiviral therapies in lowering the risk of pathogenic HBV integration and HCC.

We did not observe a significant correlation between the prognosis and the number of canonical or non-canonical integrations (online supplemental figure 10). However, these integrated HBV DNA is exogenous DNA of human cells and frequently related to driver gene activation such as TERT; it can potentially be a therapeutic target by state-of-art technologies (eg, CRISPR).

HBV integration-associated CNVs mediate dysregulated gene expression

We compared the transcriptomes between tumours and the adjacent normal tissues for 25 cases. Among 143 unique-mapping integration sites in these 25 cases, we identified 23

Figure 4  HBV integration-related CNVs. (A, B) Distribution of HBV integration-induced CNVs, including amplifications (A) and deletions (B) across the human genome. (C) Heatmap depicting integration-induced amplifications in chr5p increased the copy number of TERT. (D) Heatmap depicting integration-induced deletions in chr17p, which decreased the copy number of TP53. (E) Timing of HBV integration-induced chr8q amplification. For HCC08, HCC73 and HCC82, the chr8q amplified repeatedly at different times. CNVs, copy number variations; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.
integration-associated overexpression events (p<0.01, Student’s t-test), in most cases of which (19/23), the integration occurred with the HBV enhancer inserted into the upstream/promoter/intronic regions on the ORF-located strand of the target gene (online supplemental figure 11A), suggesting the HBV core promoter as the main driver for these activation events. Specifically, TERT and CCNE1 in chr5p and chr19q dominate the 23 activation events (14/23) (online supplemental figure 11C,E and F). Additional identification, including FLT4 and FGF18, has been associated with cancer development.32 33

In the TERT promoter region, we discovered 19 aberrations, including type I integrations (n=7), types II–IV integrations (n=6) and human SVs (n=2), and hotspot point mutations (C228T and C250T) (n=4) as mutually exclusive events (online supplemental table 5). The wild-type TERT accounted for all the adjacent normal tissues and the remaining six tumour samples, they were almost absent in TERT expression. In contrast, type I integrations gave rise to higher levels of TERT expression than the mutations and SVs in the promoter region (p=0.008, Wilcoxon rank-sum test). Types II–IV integrations generally led to copy number gain and increases in TERT expression (online supplemental figure 11E). For instance, in HCC15, a 4-fold amplification of TERT induced by a type IV integration enhanced the gene expression of 5.4-fold. The HBV promoter is in the opposite strand of ORF, thus suggesting a CNV-dependent activation mechanism. Types III–IV integration-induced gene

Figure 5 Landscape of genomic aberrations and clinical relevance associated with HBV integrations in 124 HCCs (A) Mutations and SVs in driver genes and HBV-induced CNVs are presented. The top and middle panels display patients’ age and clinical characteristics. (B–E) Association between age and the number of (B) non-canonical and (C) canonical integrations, and the association between HBV DNA load and the number of (D) non-canonical and (E) canonical integrations. (F) Comparison of numbers of non-canonical HBV integrations between TP53 mutant and wild-type samples. (G) Comparison of non-canonical integration numbers between samples with a positive or negative HBeAg. CNVs, copy number variations; HBeAg, hepatitis B envelope antigen; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; SVs, structural variations.
copy number changes altered gene expressions in large chromosome arms (online supplemental figure 11G), less likely due to the HBV integration per se.

We also identified the host-viral fusion transcripts in 30.1% (43/143) integration sites independent of the integration types (canonical or non-canonical; 27.4% vs 32.1%; p=0.26, Fisher's exact test), including alternative splicing events in TERT and CCNE1. The detection was more evident when the viral breakpoint was at the 3’end of the HBV X gene (online supplemental figure 11B,H), which allows the generation of fusion transcripts initiated from the HBV enhancer to proceed through the HBV X gene and further into the host genome. In addition, approximately 65.1% (28/43) of fusion transcripts were mapped to the human intergenic regions rather than ORFs (online supplemental figure 11D, online supplemental table 6).

**TONSL and TMEM65 promote hepatocarcinogenesis identified by in vivo CRISPRa screening**

The genes amplified at the HBV-integration sites in chr8q may play important roles at the ultra-early stages of HCC progression. We, thus, conducted a CRISPR-based gain-of-function genetic screen in mouse fetal hepatocyte BNL-CL.2, which is not tumorigenic in the immunosuppressed mice, to investigate the contribution of candidate genes identified in commonly amplified regions of chr8q by HBV integrations. We used a murine lentiviral library containing 388 sgRNAs targeting 55 amplified candidates in chr8q and three in other chromosomes (Ccncl1, Ccncl 1 or Tert) and 10% non-targeting controls (NTCs) (online supplemental table 7). The BNL-CL.2 cells (expressing dCas9-VP64) were transduced with the lentiviral sgRNA library and then subjected to 3D culture for 2 weeks before the subcutaneous implantation into the recipient mice that were immune incompetent (figure 6A). The engrafts were collected at the eighth week after transplantation for sgRNA abundance evaluation and histopathological examination (figure 6B). Importantly, these engrafts developed histological features reminiscent of human HCC, such as irregular contour, high nuclear-to-cytoplasmic ratio, and frequent mitotic and apoptotic features, which were also validated using the corresponding functional markers, including glypican-3 (Gpc3) for early HCC marker, cytokeratin 8 (Ck8) for hepatocyte differentiation and Ki-67 for proliferation (figure 6C). Among all screened genes, we uncovered Tmem65, Tonsl and Myc as the top candidates (figure 6D, online supplemental table 7). We then used the CRISPRa system to generate the BNL-CL.2 lines mimicking the corresponding pathogenic genetic alterations by individually activated three top candidates (online supplemental figure 12A). After 2 weeks of 3D growth, the designated line of cells and the NTC cells were subcutaneously engrafted into immunosuppressive mice (online supplemental figure 12B), after which we started to monitor the tumour formation. At day 44, contrary to other groups, all the mice engrafted 1×10⁶ cells developed detectable adenoma and carcinoma neoplasm features in NIH/3T3 (figure 6E). These engrafts showed remarkable differences in size and weight (figure 6F). From day 11 onwards to day 104, more mice developed visible subcutaneous tumours with either the Tmem65 (7/7) or Tonsl (4/7) overexpression (online supplemental figure 16C,D). By day 118, all the remaining mice (3/7) developed visible subcutaneous tumours with Tonsl overexpression (data not shown), whereas all mice (n=7) in the control group remained tumour-free. All the tumours developed were characterised as HCC (online supplemental figure 16E). These findings indicated that TONSL and TMEM65 activation or overexpression elicits carcinogenesis in mice.

**Potentially tumorigenic mechanisms of TONSL and TMEM65**

We sequenced and compared the transcriptomes between 12 NIH/3T3 allograft tumours (3 NTCs, 3 MYC-overexpressed, 3 TONSL-overexpressed and 3 TMEM65-overexpressed). A total of 444 differently expressed genes (q<0.1) were identified and grouped into 5 clusters (figure 7A, online supplemental table 8). The upregulated genes on TMEM65 overexpression fell into cluster1. The overexpression of both TMEM65 and MYC led to the downregulation of the genes in cluster 2 and, more significantly, in cluster 5; cluster 3 included the genes downregulated in response to MYC overexpression; cluster 4 featured the genes upregulated with the enhanced activity of TONSL and MYC, and TMEM65 with a less extent.

A systematic reprogramming of the pathways, including glycolysis (p<10⁻¹⁰), TGFβ1 targets (p<10⁻⁹), hypoxia (p<10⁻⁸) and vasculature development (p<10⁻⁴) were enriched in cluster 1, displaying a close association between TMEM65 and hypoxia response (figure 7B). Cluster 1 involves the enzymes that catalyse the classical 10 steps of glycolysis and further downstream procedures, including Tpi1 (step 5), Pfk1 (step 7), Pgami1 (step 8), Eno1 (step 9), Pkm (step 10), Pdk1 and Ldha (figure 7A). The other glycolysis enzymes, including Gpi1 (step 2; p=0.0066), Fpk1 (step 3; p=0.0057) and Gapdh (step 6; p=0.022), though less significant (0.1<q<1), could also be upregulated. Using the genomic SNPs to verify the origin of single cells, all the hepatocytes accorded with the cell line lineage, and all the microenvironment cells accorded with the mouse lineage (online supplemental figure 13A,B). We inferred CNVs of all hepatocytes in the five samples and did not observe significant CNVs and subclones (online supplemental figure 14).

Next, we investigated whether TONSL and TMEM65 protein could transform NIH/3T3 mouse fibroblasts that were commonly used to analyse the oncogenic potential, and another mouse hepatocyte AML-12. To this end, NIH/3T3 and AML-12 cells were individually transduced with lentivirus expressing TONSL, TMEM65 or MYC, known to possess transforming potential. Revealed by qRT-PCR and immunoblot analysis, the transduced cells stably overexpressed TONSL, TMEM65 or MYC (online supplemental figures 15 and 16A,B). All transduced NIH/3T3 cells showed increased colony formation compared with the mock control (figure 6H,I). We inoculated those transduced cells into the flank of nude mice to confirm the transforming capacity in vivo. For NIH/3T3, all the nude mice (10 out of 10 injections for each group) of TONSL, TMEM65 or MYC group developed subcutaneous tumours. In contrast, the mock transduced cells showed weaker tumourigenicity (6 of 10 injections), reflected by a significant decrease in tumour size and weight (figure 6J-L). Further histopathological analysis revealed adenoma and carcinoma neoplasm features in NIH/3T3 engrafts of TONSL, TMEM65 or MYC rather than that of Mock (figure 6M). For AML-12, tumour lungs with Myc overexpression (n=7) became detectable from day 55, and by day 76 reached 500 mm³ in volume. From day 97 onwards to day 104, more mice developed visible subcutaneous tumours with either the Tmem65 (7/7) or Tonsl (4/7) overexpression (online supplemental figure 16C,D). By day 118, all the remaining mice (3/7) developed visible subcutaneous tumours with Tonsl overexpression (data not shown), whereas all mice (n=7) in the control group remained tumour-free. All the tumours developed were characterised as HCC (online supplemental figure 16E). These findings indicated that TONSL and TMEM65 activation or overexpression elicits carcinogenesis in mice.

Extended into upregulating genes in response to TMEM65 overexpression. Importantly, Pdk1 and Ldha were critical in promoting anaerobic glycolysis and inhibiting aerobic glycolysis by blocking the flow of pyruvate into the tricarboxylic acid cycle. Furthermore, cluster 1 involves the angiogenic factor Pdgfa and cellular oxygen sensors Egln1 and Egln3, which catalyses the post-transcriptional modification of hypoxia-inducible factor alpha (HIF-α) proteins (figure 7A).

Cluster 3 was enriched on inflammatory response (p<10^-6). In contrast, cluster 2 and cluster 5 were enriched on extracellular...
matrix structures which closely relate to epithelial-mesenchymal transition (EMT) \( (p<10^{-10}) \) (figure 7B), suggesting the potential contribution of MYC and TMEM65 overexpression to EMT.

Cluster 4, featured by the apoptotic regulator Bcl2l1 (bcl-X) and cell-cycle controller E2f1 (figure 7A), was more involved \( (p<10^{-5}) \) in the regulation of cell death and apoptotic process (figure 7B).

The results were further validated by analysing transcriptomes of AML-12 allografts established as above. On Tmem65 overexpression, there was a significant increase in the expressions of a set of genes encoding the glycolysis enzymes including Eno1, Pgk1, Pgam1, Tpi1 and Pkm, in addition to the ones involved in the hypoxia-response such as Hif1a, and in angiogenesis like Pdgfa and Vegfa. In contrast, Tonsl overexpression enhanced the expression of Bcl2l1 (online supplemental figure 17). Similar results were obtained from the single-cell analysis of the BNLC2.2-derived allografts. Eno1, Pgk1, Pgam1, Ldh2 and Pdgfa were upregulated to greater extents \( (p<10^{-4}) \) in the Tmem65-activated tumour cells in comparison to the Tonsl activation and the original BNLC2.2 (figure 7C), whereas Tonsl was mainly expressed in cycling hepatocytes (online supplemental figure 13C). For the cancer hallmark pathway, we discovered that angiogenesis, hypoxia, glycolysis and EMT were particularly enhanced by the Tmem65 activation \( (p<10^{-10}) \) (figure 7D). In contrast, there was an upregulation of the G2M checkpoint and E2F targets in addition to a twofold proportion \( (6.7\% \text{ vs } 3.5\% \) online supplemental figure 13D) for cycling hepatocytes in Tonsl-activated allografts in comparison to the Tonsl activation and the original BNLC2.2 (figure 7D), in line with functional analysis of TONSL which maintain genomic stability during DNA replication, and TMEM65 which regulate mitochondrial dynamics. The pseudotime analysis of hepatocytes in five samples suggests the evolution initiated from the original cell line (state1) and processed into two branch states (state 2 and state 3) present in four engrafts (online supplemental figure 13E,F). State 3, but not state 2, corresponds to a cluster of cells with highly expressed EMT hallmark genes (online supplemental figure 13G–I).

Our analysis of TCGA pan-cancer data indicated that TMEM65 and TONSL were upregulated in most cancer types (online supplemental figure 18A,B) in addition to HCC. The chr8q24 amplification covering the TMEM65, TONSL and MYC loci was prevalent among diverse cancer types (online supplemental figure...
DISCUSSION

Individuals with HBV infection are prone to the development of HCC. HBV integration patterns are rare in HPV integrations, suggesting different underlying integration mechanisms or positive selections. However, based on our analysis, the enrichment of only one-end microhomology of paired HBV integration sites suggests an integration model that a first host-viral DSB ligation occasionally occurred by c-NHEJ, while the other end of HBV DSB remained unsolved and may relegate to other free HBV DNA, a novel host DSB, or a replicated copy of itself, mainly employing alt-NHEJ. In this scenario, after the occasional first host-viral DSB fusion, the remaining ends of viral and host DSBs were likely with structures (neither as blunt ends nor as complementary overhangs) that cannot be ligated directly, thus preclude rapid and precise c-NHEJ and instead favouring repair by alt-NHEJ.

In addressing the contribution of gene amplification in chr8q associated with the HBV integration to hepatocarcinogenesis, we found that TONSL and TMEM65 can elicit hepatocarcinogenesis individually in mice. TONSL comprises multiple structural elements involved in protein–protein interaction, functioning as a scaffold for assembling the MMS22L-TONSL complex essential for HR in response to replication stress and DNA DSB repair in cell cycle regulation. Localised in the mitochondrial inner membrane, TMEM65 is required for mitochondrial functions and is involved in the pathophysiology of Barth syndrome (BTHS) and cardiac conduction. c-Myc regulates glycolysis under normoxia conditions by directly activating LDH and other glycolytic enzymes. However, our results suggested a greater contribution of TMEM65 to cancer energy metabolism reprogramming. It appeared that the upregulation of TMEM65 promotes tumourigenesis via shifting the functions of hypoxia response, glycolysis, angiogenesis and EMT, while TONSL activation exerts more impact on cancer cell proliferation. This differential contribution to tumourigenesis helps to understand the role of chr8q24 amplification as a prevalent genomic alteration in cancer that is also preferentially aberrated by non-canonical HBV integrations in ultra-early stages of HCC development.

The young HCC patients were absent of TERT promoter mutations (1/54; 1.9%) and CTNNB1 mutations (2/54; 3.7%) (figure 5A), compared with a much higher mutation rate in non-young patients (30% and 17% for TERT-promoter and CTNNB1, respectively), and those in HBV-related HCCs of other datasets including TCGA (24.5% for CTNNB1) and the Japanese cohort (37% and 26% for TERT-promoter and CTNNB1, respectively). However, the young patients have more non-canonical integrations, suggesting the different aetiology in young and old HCC patients. In addition, we also identified an over-representation of TP53 mutations in young patients (56.7%; 17/30) and the patients (83.3%; 25/30) with an aflatoxin exposure-related mutational signature (Cosmic signature SBS24), compared with the HBV-related HCC in TCGA (34.0%) and the Japanese cohort (40.2%).

In summary, as revealed in this work, the mechanisms underlying the genomic instability introduced by HBV integrations can provide more insights into the relationship between pathogenic viral infections and related tumourigenesis.

Author affiliations

1Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
2Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin’s Clinical Research Center for Cancer, Tianjin, China
3Department of Biochemistry and Molecular Biology, State Key Laboratory of Common Mechanism Research for Major Diseases, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China
4Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China
5Department of Pathology, State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
6Key Laboratory of Cancer Prevention and Therapy, Tianjin, China
7Liver Cancer Center, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China
8Department of Pancreatic and Gastric Surgical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China


Hepatology


