Title

Temporal and structural patterns of hepatitis B virus integrations in hepatocellular carcinoma.

Journal

Journal of Medical Virology, 95(10)

Authors

Ren, Haozhen

Chen, Xun

Wang, Jinglin

et al.

Publication Date

2023-10-01

DOI

10.1002/jmv.29187

Peer reviewed

HHS Public Access

Author manuscript

J Med Virol. Author manuscript; available in PMC 2024 October 01.

Published in final edited form as:

J Med Virol. 2023 October ; 95(10): e29187. doi:10.1002/jmv.29187.

Temporal and structural patterns of hepatitis B virus integrations in hepatocellular carcinoma

Haozhen Ren[1,2,#,*], **Xun Chen**[3,#,*], **Jinglin Wang**[1,2,#], **Ying Chen**[4], **Alex Hafiz**[4], **Qian Xiao**[5], **Shiwei Fu**[6], **Advaitha Madireddy**[4], **Wei Vivian Li**[6], **Xiaolei Shi**[1,2], **Jian Cao**[4,7,*]

[1]Department of Hepatobiliary Surgery, the Affiliated Drum Tower Hospital of Nanjing University Medical School, Nanjing, China

[2]Hepatobiliary Institute, Nanjing University, Nanjing, China

[3]Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan

[4]Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ

[5]Institute of Modern Biology, Nanjing University, Nanjing, China

[6]Department of Statistics, University of California, Riverside, Riverside, CA

[7]Department of Medicine, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ

Abstract

Chronic infection of hepatitis B virus (HBV) is the major cause of hepatocellular carcinoma (HCC). Notably, 90% of HBV-positive HCC cases exhibit detectable HBV integrations, hinting at the potential early entanglement of these viral integrations in tumorigenesis and their subsequent oncogenic implications. Nevertheless, the precise chronology of integration events during HCC tumorigenesis, alongside their sequential structural patterns, has remained elusive thus far. In this study, we applied whole-genome sequencing (WGS) to multiple biopsies extracted from six HBV-positive HCC cases. Through this approach, we identified point mutations and viral integrations, offering a blueprint for the intricate tumor phylogeny of these samples. The emergent narrative paints a rich tapestry of diverse evolutionary trajectories characterizing the analyzed tumors. We uncovered oncogenic integration events in some samples that appear to happen before and during the initiation stage of tumor development based on their locations in reconstituted trajectories. Furthermore, we conducted additional long-read sequencing of selected samples and unveiled integration-bridged chromosome rearrangements and tandem repeats of the HBV sequence within integrations. In summary, this study revealed premalignant oncogenic and sequential complex

[*]Correspondence to: Haozhen Ren, M.D., renhaozhen1984@163.com, Xun Chen, Ph.D, chen.xun.3r@kyoto-u.ac.jp, and Jian Cao, Ph.D, jian.cao@rutgers.edu.

[#]These authors contributed equally to this work.

Author contribution statement

J.C. and X.S. initiated the project and were responsible for overall project conception. H.R., J.W., and X.S. were responsible for tissue collection and sequencing. X.C. and Y.C. were responsible for data processing, analyses and figure generating. S.G., A.B., L.T., Q.X., and A.M contributed to figure or manuscript preparation. S.F. and V.W.L. contributed to some analyses. J.C. was responsible for data interpretation and manuscript preparation.

Conflict of Interest statement

The authors declare that there is no conflict of interest.

integrations and highlighted the contributions of HBV integrations to HCC development and genome instability.

Keywords

viral integration; hepatocellular carcinoma; multiregion sequencing; intratumor heterogeneity; clonal evolution; chromosomal translocation

Introduction

Chronic infection with the hepatitis B virus (HBV) is accountable for approximately 40% of liver cancer worldwide [1]. The integration of the viral sequence into the tumor genome is recognized as a key oncogenic mechanism of HBV [2], and it is detectable in about 90% of HBV-caused liver cancer [3]. Viral integrations start to occur shortly after HBV infection in hepatocytes and tend to follow a mostly random pattern [4, 5]. In hepatocellular carcinoma (HCC), HBV integrations are more prevalent in certain human genes, such as Telomerase Reverse Transcriptase (*TERT*) and Lysine Methyltransferase 2B (KMT2B) [3, 6], indicating their involvement in tumorigenesis. Integrated viral sequences have the potential to influence nearby genes either transcriptionally or functionally [6] and may contribute to genome instability [7]. However, our understanding of the dynamics of integrations during tumor development and clonal evolution remains limited. Multiple regional sequencing serves as a potent method to uncover the historical genetic evolution of tumors and has been applied to various cancers including lung cancer [8–10], renal cell carcinoma [11, 12], breast cancer [13], glioblastoma [14], and others. Among the major cancer types, oncoviral integrations impact only HCC and cervical cancer.

Viral integrations can be characterized through high-throughput sequencing coupled with appropriate bioinformatics analyses. Nevertheless, detecting structural changes in the human genome caused by viral integrations remains notably challenging, primarily due to the need for resolving the ambiguous mapping of short reads to both human and virus genomes simultaneously [15]. Identifying integrations with low allele frequencies and those inserting into repetitive regions proves to be particularly difficult. The relatively lower sequencing depth commonly employed in whole-genome sequencing (WGS) studies, when compared to whole-exome sequencing, along with variable tumor purity, also contributes to the potential misidentification of viral integrations. Notably, conventional paired-end sequencing can solely determine the human-virus boundaries, as the short read lengths hinder the comprehensive elucidation of internal sequential structures within integrations. Recently, new sequencing techniques such as Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing have emerged, generating reads exceeding 10 kb in length [16, 17]. These extended reads sufficiently cover entire integrations as well as adjacent human sequences, thereby enabling the reconstruction of integration structures. This advancement holds the potential to unveil novel structures that were hitherto unresolved by short-read sequencing methods.

To understand the timing and structure of HBV integrations in HCC, we conducted multiple-biopsy paired-end sequencing and PacBio long-read sequencing of HBV-positive

HCC. Our findings reveal oncogenic integration events that likely occurred before and during tumor initiation, despite the diverse evolutionary phylogeny of the analyzed tumors. Long-read sequencing discovered integrations that caused interchromosome and inversion rearrangements. This study unveils premalignant oncogenic and sequential complex integrations, emphasizing the impact of HBV integrations on HCC development and genome instability.

Star Methods

Patient Samples:

Surgically resected tumor samples and blood samples were collected from Chinese patients diagnosed with hepatocellular carcinoma (HCC) at the Affiliated Drum Tower Hospital of Medical School of Nanjing University. The inclusion criteria are: (1) positive for hepatitis B surface antigen (HBsAg); (2) categorized as having Child-Pugh (CP) Class A liver function; (3) diagnosed with a single tumor; (4) tumor size falling between 3.5 cm and 10 cm in diameter; (5) age below 80 years; (6) alpha-fetoprotein (AFP) levels exceeding 100 ng/ml. Conversely, individuals meeting any of the following conditions were excluded: (1) those who lacked surgical options due to limitations in performance status and existing comorbidities; and (2) cases presenting with coinfection involving hepatitis C virus (HCV). All six patients had no recorded antiviral treatment. The pathologies of the tissues were confirmed by the department of pathology in the hospital. The serology test were performed by the clinical laboratory in the hospital. For each tumor, five distinct biopsies that were deliberately spaced at intervals of no less than 0.5 cm from one another, without any intervening capsule, were isolated. The adjacent tissue samples 2 cm away from the nearest tumor tissues were also isolated. This study was approved by the Research Ethics Committees of the Affiliated Drum Tower Hospital of Medical School of Nanjing University (2013-081-06). The written informed consents were obtained from all patients involved in this study. The study was carried out in accordance with the approved guidelines.

WGS and PacBio sequencing:

Genomic DNA was extracted from fresh tumor tissues, tumor adjacent tissues, and blood samples using QIAGEN DNeasy Blood & Tissue Kits. The libraries were constructed with TruSeq Nano DNA LT Sample Preparation Kit (Illumina, San Diego, CA, USA) and 150 bp paired-end sequenced on the Illumina sequencing platform HiSeq X Ten platform (Illumina Inc., San Diego, CA, USA). The sequencing was conducted by OE Biotech Co., Ltd. (Shanghai, China). For PacBio sequencing, high molecular weight DNA was extracted from leaves following the ~20 kb SMRTbell Libraries Protocol and sheared to an average size of 15kb using g-TUBE (Covaris), followed by enrichment and purification of large fragments with 0.45× AMPure beads. The libraries were built as recommended by PacBio and sequenced using the P6-C4 chemistry on a PacBio Sequel II sequencing platform (PacBio) at Shanghai OE Biotech Co., Ltd (Shanghai, China).

Somatic mutations, CNVs, and mutation signature profiling with short-reads sequencing:

Paired-end reads were pre-processed and mapped to human reference genome hg38 using GATK and bwa mem following GATK best practice workflow. Somatic short variants

were called using Mutect2 with matched samples following GATK best practice workflow. Somatical copy-number alterations were identified using ascatNGS. Mutational signature profiling were carried out using SigProfiler Bioinformatic Tools [18] follwoing online protocols. Clonal evolution were inferred and visualized using PHYLogeny Inference Package with the input of non-silent mutations. Non-silent mutations include missense mutations, nonsense mutations, splicing site mutations, and indels (<50 bp).

HBV read distribution:

After the alignment against the human reference genome using BWA-MEM (v0.7.17) with the default parameters, paired-end unmapped reads were extracted using SAMtools *fastq* function (v1.10) [19]. Obtained unmapped PE reads were next aligned against the HBV reference genome, sourced from the NCBI RefSeq database (NC_003977), using BWA-MEM with the default parameters. The output SAM files were converted to BAM format and then sorted and indexed using SAMtools. DeepTools *bamCoverage* function was used to achieve the read depth with the input BAM file with the parameters "*--binSize 1 -of bedgraph*" [20]. The average read depths of every 5-bp window was computed. The HBV copy number was obtained by dividing the average read depths by the tumor purity and sequencing coverage.

HBV integrations analysis using WGS short reads:

To profile HBV integrations, we analyzed the generated WGS of each biopsy using our previous tool VIcaller [3]. After the removal of low-quality integration candidates, we used the VIcaller *calculate* function to compute its integration allele fraction (IAF). Integration allele fraction (IAF) refers to the number of reads supported by integrations versus the total number of reads supported by both viral and no integrations [3]. To predict the actual IAF, the computed IAF was normalized by the tumor purity through the formal: actual IAF = IAF × tumor purity. Some integration events were manually inspected.

HBV integration analysis using PacBio long reads:

Long reads were directly aligned against the HBV reference genome (NC_003977) using minimap2 in SAM format with the default parameters [21]. Mapped reads were extracted and converted to FASTQ format using SAMtools. We then aligned the kept reads against the human reference genome to identify integration sites. After the alignment, we measured the alignment of each long reads against both human and HBV genomes using our in-house python script. We further identified reads supporting the same HBV integrations by grouping all reads covered by the same integration breakpoints.

RNA-seq analysis:

We first trimmed the Illumina adapter sequences from the raw paired-end RNA-seq reads per sample using Trimmomatic (v.039) [22]. Secondly, we combined the hg38 and HBV sequences into a combined reference genome file and indexed it using STAR (v2.7.9) [23]. Thirdly, we aligned the clean RNA-seq reads against the combined reference genome using STAR. Lastly, the output SAM files were converted to BAM format using SAMtools (v1.17) [24]. After we sorted and indexed the BAM files, we then achieved the count per million

(CPM) values using deeptools (v3.5.1) bamCoverage function [25]. ggplot2 was used for the visualization. To detect the HBV fusion transcripts, we first aligned the RNA-seq reads against the HBV genome using bwa-mem with default parameters [26]. The aligned reads were extracted in paired-end using SAMtools and were further aligned against the human reference genome to identify HBV integration sites. Integration sites with two or more supporting reads were kept as candidates.

Results

Five biopsies were collected per tumor, with a separation margin of 0.5 cm, from six patients diagnosed with single, non-metastatic, giant HBV-positive HCC for whole genome paired-end sequencing (Table S1, Figure S1). As controls, we included four peripheral blood samples and two adjacent liver tissues, with one pair from the same patient. The median coverage was 90.2% (90.1–91.8) and the median depth was 46.6 (32.9–65.5) (Table S2). Computational estimates of tumor purity indicated a median of 62%, with the lowest being 28% (Table S3). On average, we identified 45,032 (19,214–84,942) somatic point mutations per sample, including 491 (136–1,361) non-silent mutations (Figure 1A, 1B and 1C, and Table S4). Mutational signatures, which are based on nucleotide substitution, can be mathematically calculated to infer the mutagenesis mechanisms [27]. In our samples, SBS5 (unknown) contributed to more than 50% mutations in 14 of 25 biopsies, consistent with a previous report that SBS5 is the most dominant signature in liver cancer [27]. SBS4 (tobacco), SBS20 (POLD1 mutation and mismatch repair deficiency), SBS22 (aristolochic acid exposure), SBS24 (aflatoxin exposure), and SBS25 (chemotherapy), were identified in some samples (Figure 1D), suggesting possible etiologies. Biopsies from different patients showed diverse patterns of copy number variations (CNV) and indels (Figure 1E, and Table S5 and 6). Point mutations on liver cancer related genes [28–30] were identified, including in TP53 in four patients, ALB in two patients, and KEAP1, AXIN1, ATM, FRAS1, TSC2, and OTOP1 (Table S4). In addition, CNV affected TP53, PTEN, Rb1, CDKN2A, NCOR1, CCND1, and TERT (Table S5).

We then mapped paired-end reads to the HBV reference genome. We found that tumors from patient 1 and 3 harbored approximately half and 1/5 of HBV genomes, respectively, suggesting the absence of episomal virus in these tumors (Figure 2A and Figure S2). The full length of the HBV genome were detected in the other four tumors. Notably, certain biosies, e.g. these from patient 4, displayed variable copy numbers across different parts of the HBV genome (Figure 2A and Figure S2), indicating partial viral genome integrations. Variability in HBV sequencing depths and read distribution among biopsies from the same tumors appeared minimal, largely attributed to sequencing and calling variations, with some exceptions. For example, Patient 5-biopsy 4 and Patient 6-biopsy 1 displayed lower copy numbers of HBV genome, compared to other samples from the same patients (Figure S2).

We then conducted viral integration analyses with VIcaller [3]. In total, we identified 44 unique HBV integration sites across the six patients, with 43 detected in tumors and two in adjacent tissues, one of which was shared by both tumors and matched adjacent tissue (Figure 2B and Table S7). Patient 3, biopsy 1 was the only sample without detectable integrations among the 30 tumor biopsies, although it carried an almost identical HBV

partial genome as other biopsies from the same patient (Figure S2). Thus, it is possible that this integration was missed during the sequencing or calling processes. Patients 3 and 5 exhibited only two detectable integrations in their tumor tissues, while Patients 2 and 6 carried the highest numbers of integrations (21 and 10, respectively). Of the 43 integrations detected in tumors, 26 were shared by at least four biopsies from the same tumors, with 9 found across all five biopsies. Integrations shared by more biopsies had higher allele frequencies compared to those shared by fewer biopsies (Figure 2C). Additionally, two tumor-adjacent tissues (from Patient 2 and 4) were included in this study, each carrying a single integration (Figure 2B). Notably, a TERT promoter integration, an oncogenic event [31], was detected in the adjacent and all tumor tissues from Patient 4, suggesting its occurrence prior to transformation (Figure S3). This adjacent tissue shared no non-silent mutations with any of the five tumor biopsies, which largely excluded the possibility of tumor contamination in the adjacent sample. In addition to patient 4, tumor biopsies from patient 3 also carried integrations in the *TERT* promoter (Figure S3). Thus, among the 6 analyzed tumors, 2 carried HBV integrations in the *TERT* promoter, in line with our 30% estimate for HBV-positive HCC [3]. On average, sequencing a single biopsy would miss 44% of non-silent mutations, 44% of all mutations, and 38% of viral integrations. Thus, multiple-biopsy sequencing significantly boosts sensitivity for detecting mutations and integrations.

Next, we used the identified non-silent mutations to construct phylogenetic trees for the six tumors (Figure 3). Oncogenic mutations, CNV events, and integrations were delineated onto the phylogenetic trees' trunk, branch, or private segments, corresponding to their occurrence in all biopsies, multiple biopsies, or a sole biopsy within a tumor (Figure 3). The cohort of six patients were categorized into two groups: patients 1, 2, and 3 exhibited elongated trunks in their phylogenetic trees, indicating minor distinctions among all five tumor biopsies; the remaining three patients (4, 5, and 6) displayed branching patterns in their phylogenetic trees, highlighting pronounced heterogeneity among these tumors. Notably, biopsies from patients 5 and 6 were situated within distinct branches. In both cases, one biopsy differed significantly from the other four biopsies, sharing only 3 and 46 total trunk mutations (constituting 0.01% and 0.2% of total mutations) in Patient 5 and 6, respectively. When considering exclusively non-silent mutations, the shared mutations dwindled to 0 and 1, respectively (Table S4).

In Patient 5, two HBV integrations in chromosomes 5 and 6 marked two primary branches which shared zero non-silent mutations (Figure 2B, 3E and Figure S4). These biopsies likely represented separate tumor occurrences. Patient 6, on the other hand, featured just one truncal non-silent mutation (ATP2B3(D510Y)), while all five biopsies shared a single integration (Figure 2B and 3F). Consequently, among the six examined tumors, we observed viral integration events preceding known oncogenic mutations in two instances (Patients 4 and 6), based on their locations in the predicted phylogenetic trees. One of them, HBV integration in *TERT* is a premalignant oncogenic event. The functional oncogenic potential of the integration shared by the five distinguished biopsies in Patient 6 remains uncertain. It's plausible that this integration also spurred hepatocyte clonal amplification. This could have paved the way for subsequent distinct oncogenic events in different daughter cells,

ultimately giving rise to two distinguished tumors exhibiting only one shared non-silent mutation.

Subsequently, we conducted RNA-seq on an additional tumor biopsy and an adjacent normal tissue biopsy from each tumor. Three tumors exhibited minimal expression of HBV genes, despite one of the adjacent tissues from a tumor displaying robust HBV gene expression. Moreover, three tumors demonstrated partial genome expression. Among these, two displayed consistent patterns and levels of expression when compared to their matched adjacent tissues. Notably, one exhibited higher expression than the adjacent tissue (Figure S5). We identified chimeric human-virus reads at only a few integration sites identified in WGS, suggesting that the majority of integrated HBV sequences might not be actively expressing (Table S7). We also performed immunostaining for HBsAg on the available tumor tissues. Among the five tumors tested, only two tumors that exhibited HBV gene expression at the RNA level showed weak positive results for HBsAg staining (Figure S6).

Multiple integrated viral sequences may promote chromosome rearrangement through homologous recombination. Identification of viral integration using massively parallel sequencing relies on discordant reads, which are concurrently mapped to both human and viral reference genomes. Consequently, chromosomal translocations bridged by viral integrations are typically identified as two distinct integration events, often leading to the oversight of one boundary for “each integration”. Previously, we have observed that 73% of identified integrations displayed just one of the two junctions in short-read WGS data [3]. Several factors, including limited sequencing depth, algorithmic limitations in bioinformatics, and challenges in mapping reads to repetitive sequences, could contribute to the absence of one boundary. Irrespective of these factors, they might also arise as outcomes of inter- or intra-chromosomal rearrangements facilitated by viral integrations. With sequencing involving multiple biopsies, the likelihood of capturing one end of an integration and missing the other in all biopsies is exceedingly low. However, out of the 26 integrations detected in at least four biopsies, 20 of them exhibited only one identical breakpoint across all biopsies (Figure 4A and Table S7).

We selected two biopsies (Patient 2 biopsy 3, P2-3; Patient 6 biopsy 5, P6-5) with the most detected integrations for PacBio long-read sequencing. We confirmed the majority of integrations identified in short-read sequencing (7/11 and 7/10, respectively) (Table S8), despite the lower sequencing depth in long-read sequencing compared to short-read sequencing (Table S2). Furthermore, we discovered two additional integrations in Patient 2 and four in Patient 6 (Figure 4B). Notably, all four integrations that exhibited both upstream and downstream breakpoints, as identified by short reads, were effectively validated using long reads (Figure 4B). We found two integration-bridged translocations in each sample (Figure 4C): Chr2/21 and Chr1/8 in P2-3 and Chr4/17 and Chr7/9 in P6-5. In all these instances, two chromosomes were separated by partial HBV genome. The inserted HBV fragments varied in length from 243 to 2,539 bp, making it highly improbable for a single paired-end short read to cover and identify both boundaries. Indeed, in all tumor biopsies from Patient 2, Chr7 and Chr9 were identified as two separated integration events with only one end detected in short-read sequencing (Figure 4B, 4C, 4D and Figure S7). Copy number gains based on short-read data were observed in these locations in Chr7 and Chr9,

terminating at the integration site (Figure 4F). In addition, we found one HBV integration-bridged inversion in P2-3, which was fully covered by a total of 14 long reads (Figure 4E). Notably, this inversion also manifested as dual closely spaced integrations in short-reads sequencing, with each integration having only one end identified (Figure S8).

A rearranged genome offers the potential to modify chromatin interactions in tumor cells, including hijacking enhancers [32]. Nevertheless, the question of whether HBV integration-bridged genome rearrangements also contribute to tumorigenesis needs further investigation. Furthermore, long-read sequencing has identified four integration sites harboring intricate recombination events within the inserted HBV fragments (Figure 4G). In one case, a second copy was attached to the first copy in a head-to-tail fashion, likely attributable to the circular nature of the HBV genome. In the other two cases, full or partial HBV genomes were linked together at different positions. This is believed to stem from the recombination events occurring either prior to or during the progression of HBV integration. The existing of these structurally intricate integrations underscores the complexities inherent in viral integration processes, which cannot be resolved by massively parallel sequencing.

Discussion

Cancer development is a multistep process characterized by the gradual accumulation of genetic alterations, including mutations, structure alterations, and in certain instances, viral integrations. Favorable alterations undergo selection, leading to the proliferation of specific clones. Recent advances in high-throughput sequencing techniques have enabled the exploration of genetic heterogeneity and clonal evolution within tumors. Nonetheless, conventional bulk sequencing predominantly offers a snapshot of the tumor genome. In contrast, the approach of sequencing multiple spatially distinct biopsies extracted from a single tumor presents a more direct avenue to investigate heterogeneity and extrapolate historical evolutionary trajectories. It retrospectively illuminates the intricate evolutionary pathways underpinning tumorigenesis, although it only demonstrates the genetic characteristics of tumors at the time of the tissue collection and may not capture the full breadth of temporal dynamics. Multiregional sequencing has been applied to lung cancer [10], renal cell carcinoma [12], breast cancer [13], glioblastoma [14] and other cancers. HBV positive HCC, along with cervical cancer, are the only major cancer types that are affected by oncoviral integrations. The timing of viral integrations during initiation and progression of HCC is largely unknown. Here, we applied WGS to multiple tumor biopsies and adjacent normal tissues collected from six HBV-positive HCC.

Our study revealed a noteworthy pattern wherein the majority of viral integrations were detected in most cancer biopsies obtained from the same patient (Figure 2B and Table S7). It is important to consider the inherent limitations of viral integration detection, given its notably lower sensitivity compared to the identification of single nucleotide variations [15]. For example, no integrations were found in Patient 3 Biopsy 1, which exhibited an almost identical HBV partial genome to the other tumor biopsies harboring detected integrations from patient 3 (Figure S2). This integration was potentially overlooked in the Biopsy 1. This scenario suggests that a substantial portion of integrations could indeed occur during the early phases of tumorigenesis. Significantly, a pivotal finding underscores this notion.

We identified an integration within the *TERT* promoter in the adjacent tissue, exhibiting identical integration patterns as observed in all five tumor biopsies within the same patient (Figure 2B, 3D and Table S7). This congruence strongly suggest this integration event predating the transformative process. The reactivation of TERT, caused by this integration, could potentially set the stage for the clonal expansion of the host hepatocyte. The subsequent accumulation of additional genetic changes within descendant hepatocytes then seemingly catalyzed the transformation into HCC. Patient 6 further reinforces this narrative, wherein an early-stage integration appears to have transpired before or during the transformative process, based on its location in the predicted phylogenetic trees (Figure 3F). The role of this integration in promoting the clonal expansion of hepatocytes needs further investigation.

Although we did not identify a lot of identical mutations or integrations shared by analyzed tumor-adjacent tissue pairs, we acknowledge that this absence does not negate the theory of tumor origin from normal tissue. Two factors contribute to this scenario: (1) sampling variation, our collection of tumor-adjacent tissues is limited to one spot adjacent to the tumor, potentially missing the pre-malignant tissues from which the tumor originates; and (2) proportion of ancestral cells, normal cells from the common pre-malignant ancestor of the tumor may represent a minority fraction within the adjacent tissue sample, rendering alterations shared with tumor tissues potentially below the detection threshold.

HBV, a partially double-stranded DNA virus, follows a distinct replication pathway necessitating reverse transcription. Unlike retroviruses, the integration into the host genome is not a programed stage within its life cycle. Instead, it appears to occur randomly through a viral integrase-independent mechanism, stemming from double-stranded linear DNA formed during viral replication [33]. Shortly after infection, HBV integration can be identified in a fraction (<1%) of infected hepatocytes [34]. It becomes far more prevalent in HBV+ tumors, illustrating its pivotal role in the oncogenic context. Numerous studies have been conducted to profile HBV integrations in the last decade using different approaches, including WGS and viral sequence-enriched methods (reviewed in [33, 35]). It is hypothesized that HBV integrations have the potential to play a role in tumorigenesis by affecting viral transcripts, host genes, and genome stability. Notably, two frequently targeted host genes by these integrations, namely *TERT* and *KMT2B*, have been consistently highlighted across various studies [33]. HBV integrations within the *TERT* promoter region lead to reactivation of telomerase reverse transcriptase expression, a well-recognized oncogenic driver [31]. The precise impact of HBV integrations on *KMT2B* remains largely unexplored.

Detecting viral integrations through massive parallel sequencing is challenging due to the need for simultaneous mapping of short reads to both human and viral genomes. While improved algorithms could enhance integration detection sensitivity and accuracy [15]; several factors, such as limited sequencing depth in WGS compared to whole exome sequencing (WES), variable tumor purity, cancer genome mutations, viral genome polymorphisms, and intricate low-complexity regions in the human genome, continue to pose significant challenges. Furthermore, resolving the complete sequence of integrations with short reads is complicated, as virus-only reads cannot be accurately mapped to any integrations. Typically, only the human-virus boundaries can be established, based on

chimeric or split reads. Notably, some integrations exhibit both breakpoints in the same direction in human genomes, or similarly, both breakpoints in the same direction in viral genomes [5]. These scenarios cannot be attributed to linear viral genome insertion at a single human genome location. Unfortunately, short-read sequencing is inadequate for resolving such mysteries.

Here, we applied PacBio SMRT sequencing technology to selected samples, providing reads spanning tens of kilobases. These extended reads effectively cover entire integrations and adjacent human sequences, enabling us to reconstruct integration structures, including novel arrangements unresolvable by short reads sequencing. Our findings unveiled integrations bridging chromatin rearrangements and inversions, as well as integrations containing tandem viral sequence repeats. Integration-bridged chromatin rearrangements may arise from homologous recombination involving two independent integrations. Chromatin inversions may stem from recombination facilitated by homologous sequences within integrated viral sequences, like tandem repeats.

The belief that HBV integrations foster genome instability has been long-standing. Our results provide substantial support for this notion. These chromatin rearrangement events could lead to oncogene gain, tumor suppressor gene deletion, and disruption of regulatory elements. Further exploration is essential to elucidate their role in hepatocarcinogenesis.

This study employed multiregional sequencing and long-read sequencing techniques to explore the temporal and structural characteristics of HBV integrations in HCC. Through this approach, we identified oncogenic integration events occurring before and during tumor initiation, as well as integration events that induced chromatin instability. These findings offer novel insights into the role of HBV integrations in the process of tumorigenesis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by a Start-up Fund and a New Investigator Award provided by Rutgers Cancer Institute of New Jersey (State of NJ appropriation and National Institutes of Health grant P30CA072720, both to J.C.), a Busch Biomedical Grant (to J.C.), National Institutes of Health Awards R01CA272578 (to J.C.). We would like to acknowledge Rutgers Cancer Institute of New Jersey Shared Resources Biomedical Informatics, Calcul Québec and Compute Canada, and Dr. Dawei Li, Florida Atlantic University, for access to computing resources. We would like to thank Drs. Hossein Khiabanian and Shridar Ganesan, Rutgers Cancer Institute of New Jersey, for reviewing manuscript and providing advices.

Data Availability Statement

All sequencing data of this study have been deposited at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB), under the project numbers: PRJCA019414, PRJCA019413, and PRJCA019412. The access to the data and method details are available from the corresponding authors upon reasonable request.

References

1. Zamor PJ, deLemos AS, and Russo MW, Viral hepatitis and hepatocellular carcinoma: etiology and management. J Gastrointest Oncol, 2017. 8(2): p. 229–242. [PubMed: 28480063]
2. Levrero M and Zucman-Rossi J, Mechanisms of HBV-induced hepatocellular carcinoma. J Hepatol, 2016. 64(1 Suppl): p. S84–s101. [PubMed: 27084040]
3. Chen X, et al. , A virome-wide clonal integration analysis platform for discovering cancer viral etiology. Genome Res, 2019. 29(5): p. 819–830. [PubMed: 30872350]
4. Mason WS, et al. , Clonal expansion of normal-appearing human hepatocytes during chronic hepatitis B virus infection. J Virol, 2010. 84(16): p. 8308–15. [PubMed: 20519397]
5. Tu T, et al. , Hepatitis B Virus DNA Integration Occurs Early in the Viral Life Cycle in an In Vitro Infection Model via Sodium Taurocholate Cotransporting Polypeptide-Dependent Uptake of Enveloped Virus Particles. J Virol, 2018. 92(11).
6. Sung W-K, et al. , Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nature Genetics, 2012. 44(7): p. 765–769. [PubMed: 22634754]
7. Zapatka M, et al. , The landscape of viral associations in human cancers. Nature Genetics, 2020. 52(3): p. 320–330. [PubMed: 32025001]
8. Jamal-Hanjani M, et al. , Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med, 2017. 376(22): p. 2109–2121. [PubMed: 28445112]
9. Zhang J, et al. , Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science, 2014. 346(6206): p. 256–9. [PubMed: 25301631]
10. de Bruin EC, et al. , Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science, 2014. 346(6206): p. 251–6. [PubMed: 25301630]
11. Gerlinger M, et al. , Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med, 2012. 366(10): p. 883–892. [PubMed: 22397650]
12. Turajlic S, et al. , Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. Cell, 2018. 173(3): p. 595–610.e11. [PubMed: 29656894]
13. Yates LR, et al. , Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med, 2015. 21(7): p. 751–9. [PubMed: 26099045]
14. Kim J, et al. , Spatiotemporal Evolution of the Primary Glioblastoma Genome. Cancer Cell, 2015. 28(3): p. 318–28. [PubMed: 26373279]
15. Sulovari A and Li D, VIpower: Simulation-based tool for estimating power of viral integration detection via high-throughput sequencing. Genomics, 2020. 112(1): p. 207–211. [PubMed: 30710609]
16. Clarke J, et al. , Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol, 2009. 4(4): p. 265–70. [PubMed: 19350039]
17. Koren S, et al. , Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol, 2012. 30(7): p. 693–700. [PubMed: 22750884]
18. Bergstrom EN, et al. , SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics, 2019. 20(1): p. 685. [PubMed: 31470794]
19. Li H, et al. , The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. 25(16): p. 2078–2079. [PubMed: 19505943]
20. Ramírez F, et al. , deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Research, 2014. 42(W1): p. W187–W191. [PubMed: 24799436]
21. Li H, Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018. 34(18): p. 3094–3100. [PubMed: 29750242]
22. Bolger AM, Lohse M, and Usadel B, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014. 30(15): p. 2114–20. [PubMed: 24695404]
23. Dobin A, et al. , STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. 29(1): p. 15–21. [PubMed: 23104886]
24. Li H, et al. , The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. 25(16): p. 2078–9. [PubMed: 19505943]

25. Ramírez F, et al. , deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res, 2014. 42(Web Server issue): p. W187–91. [PubMed: 24799436]

26. Jung Y and Han D, BWA-MEME: BWA-MEM emulated with a machine learning approach. Bioinformatics, 2022. 38(9): p. 2404–2413. [PubMed: 35253835]

27. Alexandrov LB, et al. , The repertoire of mutational signatures in human cancer. Nature, 2020. 578(7793): p. 94–101. [PubMed: 32025018]

28. Cancer Genome Atlas Research Network. Electronic address, w.b.e. and N. Cancer Genome Atlas Research, Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell, 2017. 169(7): p. 1327–1341.e23. [PubMed: 28622513]

29. Schulze K, et al. , Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet, 2015. 47(5): p. 505–511. [PubMed: 25822088]

30. Fujimoto A, et al. , Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet, 2012. 44(7): p. 760–4. [PubMed: 22634756]

31. Sze KM, et al. , Hepatitis B Virus-Telomerase Reverse Transcriptase Promoter Integration Harnesses Host ELF4, Resulting in Telomerase Reverse Transcriptase Gene Transcription in Hepatocellular Carcinoma. Hepatology, 2021. 73(1): p. 23–40. [PubMed: 32170761]

32. Wang X, et al. , Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. Nature Methods, 2021. 18(6): p. 661–668. [PubMed: 34092790]

33. Yeh SH, et al. , Hepatitis B Virus DNA Integration Drives Carcinogenesis and Provides a New Biomarker for HBV-related HCC. Cell Mol Gastroenterol Hepatol, 2023. 15(4): p. 921–929. [PubMed: 36690297]

34. Tu T, et al., Hepatitis B Virus DNA Integration Occurs Early in the Viral Life Cycle in an In Vitro Infection Model via Sodium Taurocholate Cotransporting Polypeptide-Dependent Uptake of Enveloped Virus Particles. 2018. 92(11): p. 10.1128/jvi.02007-17.

35. Wang G and Chen Z, HBV Genomic Integration and Hepatocellular Carcinoma. Advanced Gut & Microbiome Research, 2022. 2022: p. 2140886.

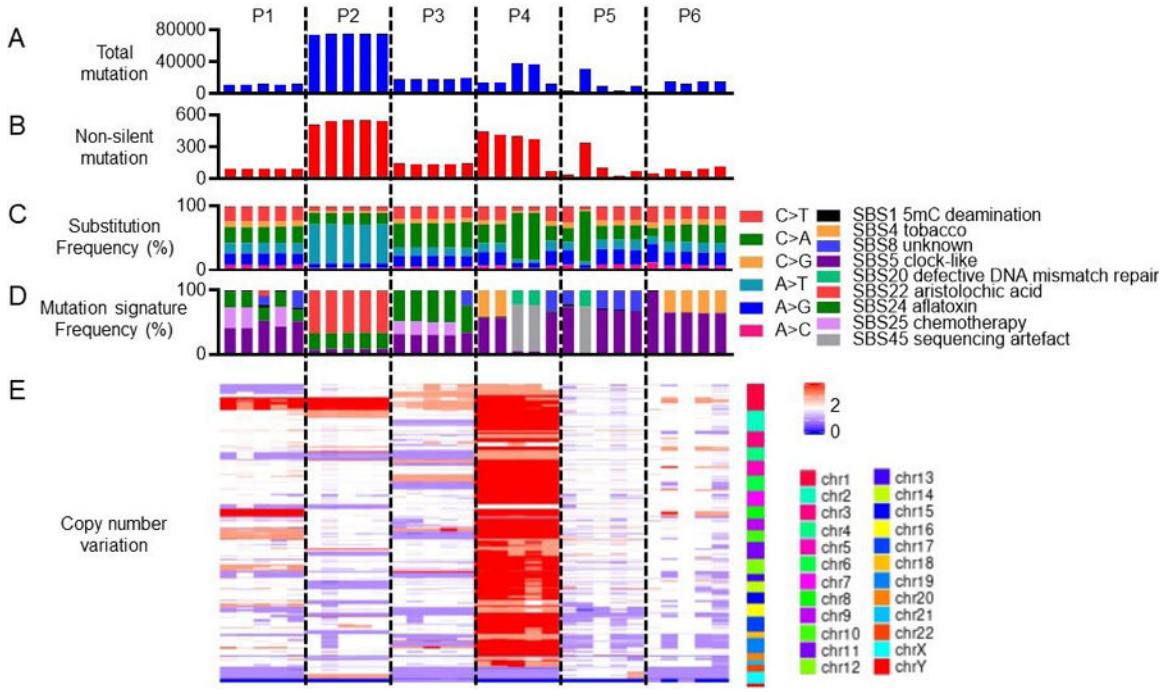

Figure 1. The genomic landscape of analyzed HCC tumor biopsies.

(A) The number of total mutations, (B) the number of non-silent mutations, (C) the nucleotide base substitution frequencies, (D) the COSMIC mutation signature compositions, and (E) the copy number variations in individual biopsies.

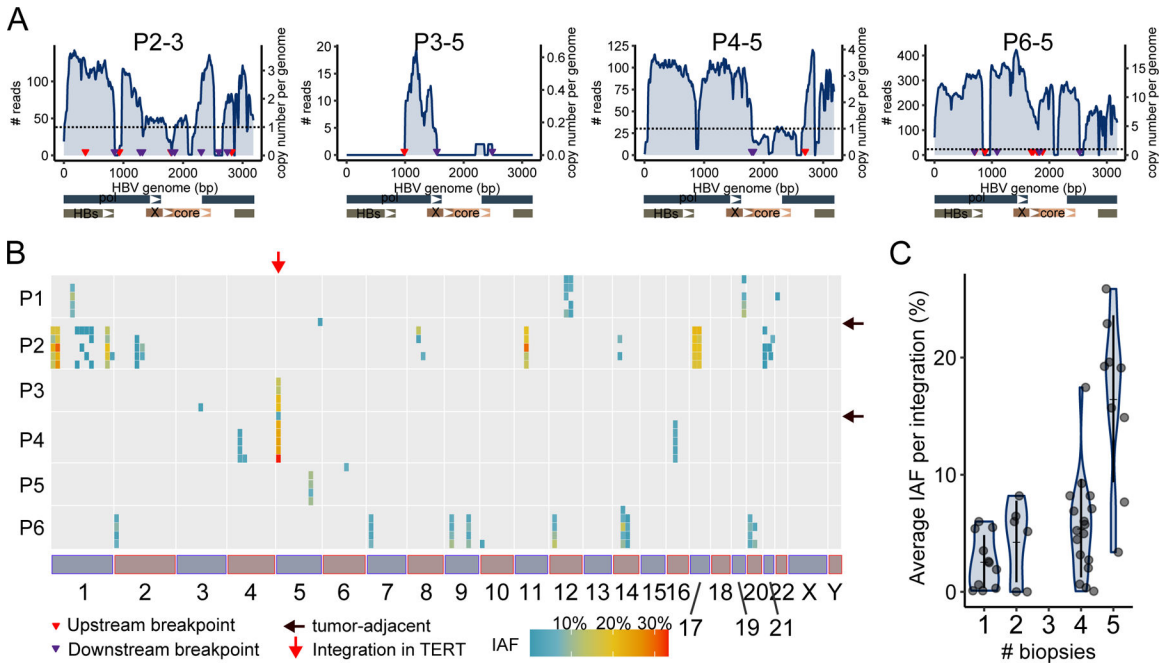

Figure 2. HBV sequence and integrations identified in analyzed tumor and tumor adjacent biopsies.

(A) The distribution of sequencing reads on the HBV genome in four selected tumor biopsies. (B) A summary of the location and allele frequency of the identified HBV integrations in all tumor and tumor adjacent biopsies. (C) The allele frequency of integrations shared by different numbers of tumor biopsies. No integration was observed in three out of five biopsies for any of the tumors analyzed in this study.

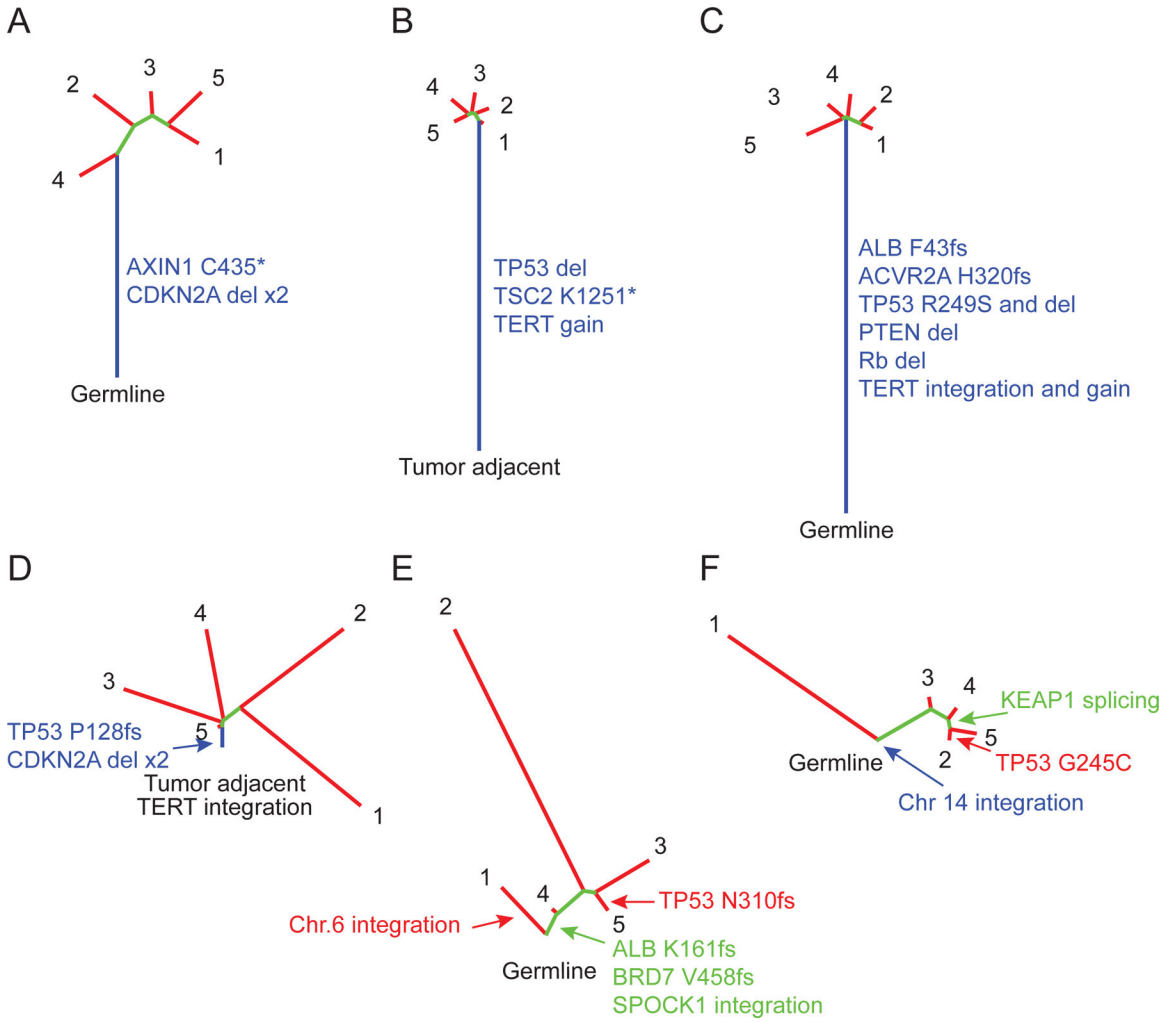

Figure 3. Phylogenetic trees of six analyzed HCC tumors.

Selected oncogenic mutations, tumor genes affected by copy number alterations, and integrations were mapped to trunks and branches, as indicated.

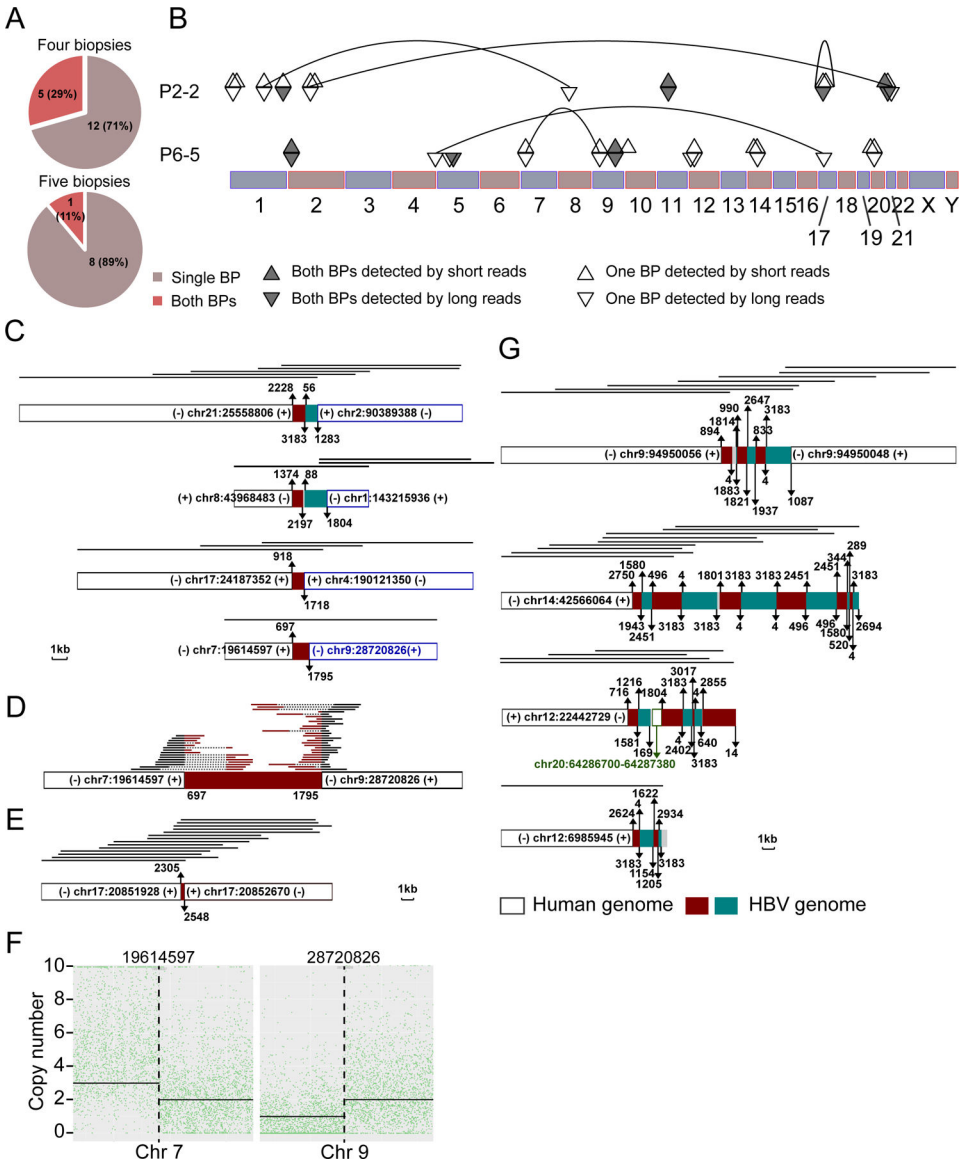

Figure 4. HBV integration-bridged chromosomal recombination identified by long-reads sequencing.

(A) The number and percentage of integrations with single or both human-virus breakpoints identified by short-reads sequencing. BP: breakpoint. (B) The location of integrations identified by short-reads and long-reads sequencing in two biopsies. Lines represent integration-bridged chromosomal translocation or inversions. (C) Structure of the four integration-bridged chromosomal translocation identified by long-reads sequencing. The lines represent single PacBio reads. (D) Mapping of supporting short-reads to the integration-bridged chromosomal translocation between chr7 and chr9 in Patient 6 Biopsy 5. (E) Mapping of supporting short-reads to the integration-bridged chromosomal inversion of chr17 and chr9 in Patient 2 Biopsy 3. (F) Log R ratio showing copy number alterations around breakpoints showing in (D). (G) Structure of the four complicated integrations identified by long-reads sequencing. The lines represent single PacBio reads.