Non-invasive profiling of tumor-originated viral integrants in patients with HBV infection and integrant prediction revealing their abilities in producing HBsAg transcripts.

Wei Chen[1,#], Ke Zhang[2,4,#], Peiling Dong[3,#], Gregory Fanning[4], Chengcheng Tao[1], Haikun Zhang [1,6], Zheng Wang[3], Yaqiang Hong[1,6], Shicheng Guo[7], Xiaobo Yang[5], Huiguo Ding[3], Haitao Zhao[5], Changqing Zeng[1,*] , Ulrike Protzer[2,*], Dake Zhang[1,*]

[1], Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China
[2], Institute of Virology, Technical University of Munich / Helmholtz Zentrum München, 81675, Munich, Germany
[3], Department of Hepatology, Beijing You'an Hospital Affiliated with Capital Medical University, Beijing 100069, China
[4], Janssen China Research & Development Center, Shanghai 201210, China
[5], Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[6], University of Chinese Academy of Sciences, Beijing 100049, China
[7]Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA

[#] Equal contribution
* Corresponding Authors

Correspondence: Changqing Zeng, Beijing Institute of Genomics, Chinese Academy of Sciences, NO.1 Beichen West Road, Chaoyang District, Beijing 100101, China. Phone: +86-10-84097818; Fax: +86-10-84097706; Email: czeng@big.ac.cn.
Ulrike Protzer, Institute of Virology, Technische Universität München / Helmholtz Zentrum München, Trogerstrasse 30, 81675 Munich, Germany. Phone: +49 89 41406886; Fax: +49 89 41406823; Email: protzer@tum.de
Dake Zhang, Beijing Institute of Genomics, Chinese Academy of Sciences, NO.1 Beichen West Road, Chaoyang District, Beijing 100101, China. Phone: +86-10-84097566; Fax: +86-10-84097706; Email: zhangdk@big.ac.cn.

Running title: *Chen W et al / HBV integration and tumor screening*
Total word counts: 4971
Figures: 6
Tables: 0
Supplementary figures: 5
Supplementary tables: 8

Abstract

HBV-host integration events that occur in liver cells during chronic infection are thought to contribute to cancer development. In this study, we developed a low pass integrant sequencing method and achieved high resolution of integrations by capture-enriching HBV-human junctions in plasma to screen for liver cancer in HBsAg positive individuals, and proposed a novel strategy for integrant prediction based on short reads sequencing in which required an extensive reduced sequencing volume to detect HBV integration events. Particularly, we demonstrated viral integration events detected in plasma were mainly derived from tumor tissues rather than from adjacent liver tissues. In addition, we found most of viral integrations might contain complete opening reading frame of HBV surface proteins, about 50% of which produced viral-human chimeric transcripts detected by deep RNA sequencing in paired tumor tissues. In summary, we demonstrate integration profiling in plasma is the promising non-invasive approach to monitor the liver cancer development and to evaluate the viral-protein coding ability of integrations in tumor cells.

Keywords: Circulating Cell-free DNA; Hepatocellular carcinoma; viral integration; repeat elements.

Introduction

Hepatitis B virus (HBV) integration has long been noticed in hepatocellular carcinoma (HCC) and liver tissues[1-9]. It has more recently been appreciated as an early event during HBV infection[10]. The integrated HBV DNA may preserve an intact open reading frame (ORF) of envelope proteins, and serve as an additional template for transcribing hepatitis B surface antigen (HBsAg) genes[11]. A latest study has pointed out, liver cells harboring integrated HBV DNA sequences can express chimera peptides that can be recognized by T cells[12]. Unlike retroviruses[13], viral integration is not required for HBV replication, and no HBV proteins are known to have integrase activity[14]. Experimental evidence has indicated that double stranded linear DNA (dslDNA) is the preferred DNA substrate for integration[15]. Despite the attention that HBV integration has received in HCC and HBV virology investigations, the process and implication of the "side product" of infection remains largely unknown[16]. During the 1980s, HBV integration events in HCC liver tissues were assessed with a combination of restriction enzyme digestion and southern hybridization, and the isolated HBV-human DNA junctions were subjected to Sanger sequencing[1-9]. However, these procedures are difficult to scale up for clinical studies. In 1995, Minami *et al.* developed an Alu-PCR strategy, which enabled for the more practical screening of integrations[17,18]. However, this method fell short of the full evaluation of all integration sites due to the selective amplification of known viral fragments and host genome regions, and the added technical complication of template switching during amplification, which created artifact chimeric products[19,20]. Recently, the resequencing of HCC genomes has identified >100 viral integration breakpoints that implicated many cellular genes[21-24]. Among these, *TERT*, *MLL4* and *CCNE1* were the most frequently observed[21-25]. HBV integrations were more likely to occur in chromosome sites which are exhibiting genomic instability featured with repetitive regions, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and microsatellites and telomeres[16,26]. A causal impact of HBV integration on tumorigenesis has always been asserted, since these integrations may interrupt functions of cellular genes around the integration sites, and become dominant in tumor cell populations after clonal expansion, possibly conferring growth and survival advantages to affected cells. Although HBV integration has been considered to randomly occur, enriched viral integrations in specific chromosomes 5, 16, 17 and 19 have been reported[24]. In particular, these integration sites were associated with structural variations in tumor cellular genomes[27-31]. Evidences also showed integration sites tend to occur within boundaries of the altered copy numbers of a gene, and may lead to genomic instability of the infected hepatocyte, which is one of the oncogenic roles of viral integration[22]. In addition, integration patterns serve as biomarkers that can detect cell clonal expansion in both tumor and virus biology studies[10,32].

A major obstacle for using viral integration as a tumor marker is that it is liver biopsy dependent, and liver tissue availability becomes a bottleneck in clinical application[33]. Integrated viral DNA is expected to form nucleosomes that are nuclease-degradation proof and resembles the minichromosomes of HBV covalently closed circular DNA (cccDNA) in liver cells[34]. Thus, a plasma cell free DNA (cfDNA) pool may contain virus-host junction fragments that may serve as a biomarker, and reflect a part of the genetic changes in tumor genomes[33]. The detection of plasma integrants was dubbed as liquid biopsy without invasiveness[35].Theoretically, the analysis of the plasma cfDNA pool is a more accessible and reproducible procedure that can be performed upon each patient visit, resulting in the timely detection of viral integrations that may be significant for tumorigenesis. However, the level of virus-host junctions is expected to be similar to the amount of circulating tumor DNA, accounting for approximately 1% of the total cfDNA[36], and the deep sequencing of cfDNA is cost prohibitive. Therefore, viral-host fragments enrichment in the plasma could effectively reduce the sequencing volume and increase the detection sensitivity.

Here we designed viral DNA probes to capture the whole HBV genome so that we could enrich HBV-host integrants in human circulating cell-free DNAs for deep sequencing. The small size of HBV genome enables us to achieve a deep sequencing coverage at small sequencing volume, with significantly increased ability to detect viral integrants. In this study, we analyzed the integrated fragments in the paired samples from two types of body fluids, tumor and adjacent liver tissues in the same individual to explore the representation of plasma integrants in the integration profiles observed in tumor and adjacent non-tumor liver tissues. Our study showed this approach can be an efficient strategy to characterize the viral integrants in plamsa to screen for liver cancer screening in patients with HBV infection, and short reads sequencing data can also be used to predict the formats of integration by pairing the two ends of each integrant.

Results

Landscape of HBV integration in cancer and adjacent non-cancer liver tissues

To enrich the integrants in circulation and reduce the sequencing volume, we designed viral probes to target HBV DNA sequences (Method). To validate proposed capture-enrichment assay, we used a HBV stable transfected HepG2.2.15 cells and performed capture-sequencing to identify HBV-host integrants (Stage I, Study Design, Figure 1A). We found sequencing uncovered various fragments containing virus-cell DNA junctions, and evidence of HBV DNA integration (Table S1). The virus-cell DNA junctions in the integrant fragments consisted of both the viral genome end (viral breakpoint, Figure 1B) and cellular genome end (host breakpoint, Figure 1B). In total, five integration sites in HepG2.2.15 were identified using our protocol. In Stage II, we then applied the method to 80 samples collected from 20 liver cancer patients (Study Design, Figure 1A). We identified 424 integration events totally (Figure 2A). The number of detected integration events varied from 2 and 82 (average: 25) among the 17 HBsAg positive patients, and none were identified in 3 HBsAg negative patients as negative control. While the numbers of integration varied between patients, there was no significant difference in integration amount between the tumor sites and paired adjacent non-tumor sites of the same individuals (t test, P>0.05, Figures S1 and S2). Analysis of the genomic sites of integration (host breakpoints) specific to tumor or non-tumor tissues revealed no significant differences in cellular genome locations of the integration sites in these two types of tissues, which can be either between genic and intergenic regions (Chi-square test, P=0.9; Table S3), or between repeated and non-repeated regions (Chi-square test, P=0.09; Table S3). The most common gene interrupted by the integration was *FN1*, detected in 8 of 17 patients, while the most common interrupted repeat sequence was ALR/Alpha, which was also found in 8 patients (Figure 2B). Furthermore, integrations in the telomeres of chromosomes, characterized by the repeat sequence of $(TTAGGG)_n$ (Figure 2B), were also very common (23.5%,4/17). In addition to those interruptions in introns of genes or directly interrupted genomic elements, we also observed 29.4% (5/17) patients had integration sites in the promoter region of *TERT*, consistent with previous observation (23.7%,18/76) [21]. Our capture strategy achieved high sequencing coverage (number of junction reads) of integration breakpoints, and 70% (296/424) integrations had sequencing coverage over 100 in at least one of four aliquots analyzed in the same individual (MaxCoverage, Table S3) and the highest one even reached 11,579. The sequencing coverage of a breakpoint could indicate the abundance of the integrant, reflecting the size of the clone carrying the corresponding integration. Integration events in tumor samples had sequencing coverage around 10-fold higher than that in adjacent non-tumor samples (Figures 2C and Figure 3B), indicating significant expansion of tumor clones. In conclusion, our assay can be applied to efficiently capture and characterize integration events.

Capturing the HBV integrations in body fluids

In order to examine the suitability of saliva, a new source of body fluids recently adopted in liquid biopsy solutions for other cancers[37], we collected the saliva samples along with plasma samples for 7 liver cancer patients (Stage III, Study Design, Figure 1A; Table S4). Totally, 32 integration events successfully detected in 5 patients (Table S5). One patient had integration detected in saliva (Table S5). Although she had five integrants with high abundance (250-881 junction read pairs) in plasma integration profile (Table S5), only one integrant (279 junction read pairs in plasma) was seen in saliva supported by one non-redundant junction (Figure 3A). Therefore, only 1 of 32 (3%) integrations can be detected in the saliva samples which showed that saliva was unsuitable for liquid biopsy using viral integrant as a liver cancer biomarker. To trace the origin of integrants in plasma cfDNA, we collected paired tumor and adjacent normal tissues from 8 HCC patients (Stage IV, Study Design, Figure 1A; Table S4). The cfDNA in all seven HBV positive HCC patients was analyzed (one plasma sample failed in the cfDNA extraction) (Table S4). In addition, deep transcriptome sequencing was performed for 4 paired HCC liver tissues and adjacent normal liver tissues.

Overall, the plasma integration events predominantly reflected the counterparts in tumor tissues, most likely with chimeric RNA transcripts observed. First of all, we detected 29 integration events from 7 plasma samples and all of them could be detected in the corresponding liver tumor samples (Figure 3B, Table S6). Notably, junction abundance for integration in plasma samples significantly corrected with that in tumor samples (R^2=0.64, P= 6.2×10^{-29}, Table S6), rather than that in the paired non-tumor liver tissues (R^2=0.32, *P*=6.2×10^{-12}, Table S6). Deep RNA sequencing for the same aliquot found 17 integration sites with chimeric RNA transcripts, 76.5% (13/17) were seen in plasma; meanwhile, 45% (13/29) integration events observed in cfDNA had transcription in the tissue samples. This conforms the popular observation that tumor-derived fragmented DNA in circulation has a better representation of tumor genomic diversity compared to a single tumor aliquot. Although it was not likely to predict transcription level of integration sites according to the junction abundance in DNA capture assay, the prediction performance was much better in tumor tissues (R^2=0.27, *P* = 1.5×10^{-8}, Table S6), in comparison with adjacent liver tissues (R^2=0.01, *P* = 0.15, Table S6).

Secondly, missing of integrations specific to adjacent liver tissues in plasma was unlikely due to the scarcity of original cell clones carrying them. Both DNA and RNA sequencing showed that non-tumor clones carrying those integrations, may have size rival to tumor clones harboring integrants detected in plasma, according to the abundance of HBV-human junctions and chimera RNA transcripts (Table S7, Figure 3C). Taken together, we find plasma integration profile is valuable for the detection of tumor derived integration events, and RNA sequencing data proved the transcription activity of these genomic integration sites in tumor cells.

Besides, HBV integrants were also sought in the sera of 10 chronic hepatitis B patients without liver cancer (Stage V, Study Design, Figure 1A; Table S4). No events met the integration criteria in all these patients (Methods section). However, we did observe a lot of single junction reads indicating the existence of integration events, and breakpoint distribution of them was consistent with that of breakpoints found in tumor and liver tissues col (Figure S1). However, their authenticity should be supported by analyzing paired tissue samples from liver biopsy. The scarcity of DNA fragments from the integration sites was likely due to the limited death of affected cells, or the number of which was relatively small possibly without obvious clone expansion process. Another exception could be in a scenario whereby an HBsAg specific T cell response was active in the liver where selective killing could increase the overall number of HBV integrations detected.

HBV integrant prediction: sequence boundary of a single integration in human genome.

To predict the HBV integrant will be crucial to evaluate the viral protein-coding ability of integrations. Each integration should have two viral breakpoints and two host breakpoints (Figure 3A). Diverse integrated viral fragments may coexist in one patients, and complexity of overlapping among them make it impossible to pair the two viral breakpoints within the viral genome at the length of only 3K bp. In contrast, two independent integration events in host genome should be far away from each other. Therefore, we attempted to link two cellular breakpoints that occurred within 20 K base pairs (bp) as a single integration event (Figure 4). Among the total 424 integrations observed in paired tumor and adjacent non-tumor tissues, we were able to map 218 of these accurately at each end of the integrated sequence. The genomic distance between paired breakpoints were between 35 bp and >4,000 bp, and the most common distance was approximately 0-50 bp (87%, 189/218; Figures 4A-4B, Figure S3). The frequency of large deletions (>1,000 bp, 5%, 11/218; Figure 4C) and redundant human DNA fragments (13%, 30/218; Figure 4D) around the host breakpoints was relatively low. Notably, seven of all 11 integrations in the intronic region of *FN1* gene had repeated sequences at the breakpoints, although the biological explanation is unclear now. Obviously, genomic structure variation may influence the pairing analysis of the remaining 206 integrations (Figure 5A). To test this hypothesis, the whole genome genotyping of tumor tissues from four HCC patients was performed. Among the 19 unpaired breakpoints identified in these patients, nine breakpoints were located at the telomere or centromere regions, while 10 host breakpoints were located at the boundary of large structure variations (SV in Table S6 and Figures S4A-S4H). Particularly, two sites in chromosome 9 were separated by 1.7 M bp, and each was located at one end of a same length deletion region in the human genome (Figure 5B). Thus, the alterations in tumor genomes and the inaccurate mapping of junction reads in repetitive sequences were the two major reasons for the inability to pair some host breakpoints.

HBV integrant prediction: four patterns of integrated viral fragments

After pairing the host breakpoints, we obtained the corresponding viral sequencing reads for paired viral breakpoints. Then, a clear view of the integrated viral sequence could be achieved (Figure 6). Among 218 integration events with host breakpoints successfully paired, 215 integrated sequences could be characterized including the orientation of the viral sequence and four distinct viral sequence patterns were proposed (Figures 6A-6B). Their viral breakpoints were relatively consistent, showing similar distributions between the integration patterns. The majority of integrated sequences the viral breakpoints were identified between nt 1,600-1,900 of the viral genome (64.2%, 138/215) (Figures 6C-6D). This region well overlapped with the cohesive ends of DR1 and DR2, which also are features of dslDNA ends. Almost all Pattern I integrations had viral ends consistent with the ends of the dslDNA, and the viral segments in this group were shorter than the full-length HBV genome, ranging between 952-3,214 bp. Interestingly, viral breakpoints in viral pattern II were located more common between nt 1-1,000 than those in viral pattern I (P = 3.1x10^{-7}, *t*-test). Besides, the viral segments in viral pattern II (21.4%, 46/215) were shorter than pattern I, ranging from 32 to 1,584bp. Viral integrants in viral pattern III (10.2%, 22/215) and IV (4.2%, 9/215) seemed to be formed by ligating at the ends of least two viral fragments in a 3'-to-3' or 5'-to-5' manner. In addition, most individuals contained all four patterns (Figure S5). We acknowledged here that the method infers patterns by assembling sequence data from multiple 150 bp reads. Clearly, a direct verification would require a sequence of long DNA segments without fragmentation.

Furthermore, we explored the chimera RNA transcripts from all four patterns of integration sites in tissue samples with deep RNA sequencing. In all, DNA capture experiments found 76 integration sites (Table S8). Among them, viral patterns of breakpoints were determined in 42 events, and each pattern had 19% events with transcription activity (Table S8). Therefore, there were no obvious difference in transcription activity among diverse types of integration events.

Discussion

In the present study, we adopted the HBV capture strategy to enrich DNA fragments covering the virus-host junctions of integrants in plasma cfDNA from patients with HBV infection. It enables us to decrease the sequencing volume to only 2G raw data in next generation sequencing platform with high resolution of integration profiling. The potential for this method was explored in tumor, adjacent non-tumor tissue, paired blood and saliva samples, revealing the unexpected observation that integrated fragments captured in plasma cfDNA exclusively oriented from tumor clones in the liver. This highlights the exciting possibility that HBV integration profile in plasma may be included as a solo biomarker or part of a panel of analyses from liquid biopsies to assist in liver cancer screening. Furthermore, HBV integrant prediction may also be a crucial analysis to make plasma viral integration proofing as a new companion diagnostic for HCC immunotherapy using T cells targeting viral proteins encoded by integrations[38].

It has been suspected that there will be a sensitive tool to detect the integration fragments in the blood. Unlike the circulating tumor DNA which is confounded by the DNA released from blood cells, virus-host junction detection is influenced by both integrated viral DNA and non-integrated HBV DNA. There have been efforts to take urine as specimen source[39], and we have evaluated both saliva and plasma. Recently, saliva has attracted a lot of interests in the research field of liquid biopsy, and diverse studies have reported circulating tumor DNA in saliva at low concentration[40,41]. After enrichment procedure, we only obtained one integrated fragment in all saliva samples, which shows the limitation of integration detection using saliva as a liquid biopsy. By contrast, enriched viral fragments from plasma samples were adequate for further analysis.

The capture enrichment strategy has increased the sensitivity of integration detection, and effectively reduced the inference from non-integrated HBV DNA in samples. Here, to 200 ng of probes were applied for each sample to ensure that all viral fragments would be captured. Theoretically, a 200-ng probe scan can capture at least 10^{11} target molecules. The DNA extracted from each liver tissue consisted of 10^5 cells that resulted in 600 ng of double-stranded DNA. Each HBV infected cell can contain up to 1,000 copies of replicative intermediates, resulting in approximately 10^8 copies of non-integrated HBV DNA, which can only consume a maximal of 1/1,000 of input probes. Thus, there were sufficient probes to capture the HBV integrants. Although this assay was not designed to perform a quantitative analysis on integrations, the same number of viral probes and a relatively equal amount of input DNA enabled the comparison of the relative abundance of individuals and the total integration events among all analyzed samples. The same unique integration events carried by many liver tumor cells, which was uncovered by a higher sequencing depth, supported the idea that the relative abundance of HBV integrations is a genetic marker for clonal selection and the expansion of affected hepatocytes[10].

The results from the sequencing of liver tissue samples revealed an average of 35 non-redundant junctions for each host sequence breakpoint, and this was higher than 18, when compared to that reported by Zhao *et al*., in which a similar strategy was used[24]. Notably, the highest sequencing read depth for the cellular sequence breakpoints from liver tissues reached 11,579 in the present study. Hence, it was considered that this potential for the identification of most integration events in each sample was adequately explored in the present experiment. However, it could not be ruled out that there is a possibility that more tissue aliquots, more probes or ultra-deep sequencing may identify more integrations at low frequencies or increase the read depth for breakpoints with relatively limited supporting reads, as shown in the present results.

The kinetic detection of integration events and changes in integration events over time may help monitor disease progression in the liver. Theoretically, the cfDNA in blood may contain integration events that mirror the counterparts in the liver, since these were released from the liver. After all, it is easier to obtain blood samples than a liver biopsy. The present data revealed that the detected plasma HBV integrants predominantly originated from liver tumor cells. In the tumor and paired non-HCC liver tissues from the

patients at Stage II (Figure 1A), it was observed that the integration events had extremely higher read depths in non-tumor liver tissues, indicating that some non-tumor clones with integrations may already have great expansion before tumorigenesis (Figure 2C). Nevertheless, it was regrettable that the corresponding plasma was not collected to investigate the abundance of integrant fragments in the circulation. However, the detection of HBV integrants in the corresponding cfDNA pool in the patients enrolled in Stage IV revealed that all integrations identified in non-tumor liver tissues had no fragments detected in plasma (Figure 3B).

Circulated tumor DNA in plasma was released from injured and dead tumor cells, as well as from dead adjacent cells or blood cells. It was possible that the DNA level released from non-HCC liver tissues was less abundant, when compared to liver tumor tissues. The present data revealed the potential clinic utility of the capture assay to monitor plasma integration events. Future studies with a larger sample size are required to validate these present findings. The HBV integration detected in the plasma cfDNA pool may potentially become a new plasma biomarker that could complement present biomarkers to monitor HBV related to liver disease stage, including liver occurrence.

Differences in integration profiles between tumor and non-tumor tissues may reflect the different clone compositions of HCC and non-HCC liver tissues, implying a divergent clone evolution from non-HCC cells during tumorigenesis. It has been suggested that the occurrence of certain integrations is the driving force for tumorigenesis. For instance, these lead to the upregulation of the expression of TERT, MLL4 and CCNE1 [21]. However, both HCC and non-HCC cells also share common integration events, but with different abundances, implying distinct clone expansions they may undergo. As these integrations occur randomly, some may contribute to the tumorigenesis process, while others do not. Diverse studies have demonstrated that consequences caused by integrations may be different in oncogenesis[21,24].

According to the paired viral breakpoints, the sequences of the detected integrated viral segments for viral pattern I and II were assembled (Figure 6D). It was considered that a majority of viral pattern I events (81.2%, 112/138) reserved an ORF of large surface protein, and 14 of the remaining 26 events had an intact ORF of middle surface protein. These observations support the recent suggestion that integrated HBV DNA provides additional capacity for HBsAg production, and represents a challenge to reduce HBsAg production [11,42]. The present data provides evidence that the integration patterns varied among individual patients. Diverse patterns and different percentages of different integrants in the same individual simply distinct HBV antigen expression patterns, which are expected to impact therapeutic responses to HBV treatment or efforts targeting tumor cells expressing viral proteins. This would be one of important factors considered in selecting patients for these clinical trials.

Materials and methods

Patients and samples

The present study was conducted in You'an Hospital (Beijing, China). A total of 42 patients were enrolled in the stages of sample collection (Figure 1A). Among these patients, 27 patients had HCC, 5 patients had BDC (all HBV positive), and 10 patients had chronic hepatitis B. Blood samples from HCC and BDC patients were collected before surgery, and the corresponding liver tissues were obtained afterwards. A total of four samples, which included two tumor sites and two adjacent non-tumors, were used for the analysis. Chronic hepatitis B patients only provided blood samples for analysis. The diagnosis was made according to the guidelines for the prevention and treatment of chronic hepatitis B: a 2015 update [43]. The BCLC staging criteria were used to classify HCC patients. The laboratory findings are summarized in Tables S2 and Table S4. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Ethics Committee of You'an Hospital. An informed consent was obtained from all patients. In Stage II of sample collection (Figure 1), 17 HBV-related cancer patients (12 HCC and 5 cholangiocarcinoma, CHOL, Tables S2-S3). Besides, three patients without HBV infection, including two HCV-related HCC patients, and one BDC patient negative for both HBsAg and HCV, served as negative controls.

DNA sample of HepG2.2.15 cell line was provided by Beijing Tricision BioTherapeutics Inc, and cell line authentication was examined by Guardian Technology Co. Ltd. using short tandem repeat loci. We performed three capture experiments (replicate 1-3) using HepG2.2.15 DNA samples following the below procedure, and obtained 1G, 1.5G, and 2G raw sequencing data respectively (Table S1).

Both cancer and adjacent liver tissues obtained from each HCC or BDC patient were collected and stored at -80°C until analyzed. Total DNA was extracted from the paired tumor and normal adjacent samples using a QIAamp DNA Mini Kit (Qiagen, Valencia, CA), according to manufacturer's instructions. Ten milliliters (ml) of whole blood was collected from each patient in Streck Cell-Free DNA BCT® tubes (Streck, Omaha, NE) at approximately one week before surgery. The blood collected in Streck BCT tubes were immediately centrifuged at 3,000 × g for 15 minutes at 4°C within two hours. 2.5ml Saliva samples were collected using using Oragene OG-500 collection kits (DNA Genotek, Inc., Ottawa, ON, Canada). Then, the plasma or saliva was carefully transferred into a fresh microcentrifuge tube, followed by a 2nd centrifugation at 16,000 × g for 10 minutes at room temperature. Five ml of resultant plasma or two ml of resultant saliva was used for DNA extraction using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Valencia, CA). After extraction, total DNA was quantified using a Qubit dsDNA HS Assay kit (Life technologies, Grand Island, NY, USA). For two failed samples in Stage III (Figure 1), one (s006) had high amount of DNA possibly confounded by DNA from blood cells. All DNA samples were stored at -80°C before the capture experiment.

Viral capture design and experiment

Viral probes (baits) for liquid capture were obtained from iGeneTech Bioscience (design.igenetech.com, China), and the design adopted the tilling strategy across the whole HBV genome. HBV subtype (A-H) reference sequences from the NCBI genotyping tool were used to design the probes[44]. Then, the probes were filtered to remove redundancy and ensure non-complementarity to the human genome sequences (hg19). The length of the probes ranged from 80 to 150 base pairs (bp), and the DNA samples were sheared into fragments at 150-200 bp before the following experiments. Capture assay utilized an input of 600 ng of tissue DNA or total cfDNA (range from 20 ng to 200 ng) from each sample and 200 ng of HBV probes per standard capture protocol, which were included in the TargetSeq Enrichment Kit following the instructions (Target DNA Capture, iGeneTech Bioscience).

Briefly, the DNA templates were sheared into fragments using a Covaris focused-ultrasonicator (Covaris, Inc. MA, USA). The resultant fragment (150-200 bp) were evaluated using an Agilent 2100 bioanalyzer (Agilent Technologies, CA, USA). Then illumina adaptors were ligated to the DNA fragments after the end repair and A-tailing procedure. After pre-PCR enrichment, the products were purified using Agencourt AMPure XP beads (Beckman Coulter, MA, USA) before hybridization, which was performed by mixing and incubating the pre-PCR products first with blocking oligos in PCR tubes at 65°C for 5 minutes then with viral probes at 65°C for 16 hours. Dynabeads MyOne Streptavidin T1 (Invitrogen, NY, USA) were used to immobilize the hybridized products at room temperature. These immobilized products were washed with wash buffer I at room temperature once, and subsequently with wash buffer II at 65°C for three times. Then, the 2nd round of PCR (post-PCR) was performed, and the products were purified before the sequencing procedure.

Sequencing experiments and Integration calling

For DNA samples, each sequencing library was established by performing paired-end sequencing (2 × 150 bp) on an HiSeq X Ten sequencer (Illumina Inc., San Diego, CA, USA), which comprised of 1G (saliva and plasma pairs) or 2G (plasma and tissue pairs) raw data. For transcriptome sequencing, total RNA of 10-20 mg tumor or liver tissue was extracted using RNeasy (Qiagen, Valencia, CA). RNA samples were treated with DNase I (Ambion, Austin, TX) and rRNA was removed from total RNA using Epicentre's Ribo-Zero rRNA Removal kit (Epicentre Technologies, Madison, WI). Next, 30-100 ng Ribo-Zero RNA was used for the construction of the library using the Illumina TruSeq™ RNA Sample Prep Kit and followed the manufacturer's instruction. Paired end sequencing (2 × 150 bp) was performed for each sequencing library on a NovoSeq sequencer (Illumina Inc., San Diego, CA, USA), which generated 32G raw data on average.

Meerkat, a discordant read-pair and split reads mapping based structure variation detection pipeline, was used to identify the HBV integration events [45,46]. First, a reference genome was constructed by adding a HBV genotype C consensus sequence into the human reference genome (hg19) as a pseudo chromosome[47]. Cutadpt v1.14 was used to remove low quality and adaptor-contaminated reads[48]. Cleaned reads were mapped on the reference genome using the BWA-MEM algorithm[49]. Duplicated read-pairs after the generation of bam files were labeled using Picard tools (2.7.1) and removed using the MarkDuplicates function. Then, Meerkat was used to detect and annotate the structure variations. Inter-chromosomal translocations between the pseudo HBV chromosome and human chromosome were extracted as candidate integration events. Default function parameters were used in all aforementioned

software tools. The breakpoints of these events were further filtered using the following criteria: (1) supported by at least two non-redundant split reads; (2) <2 mismatches in the human genome side; (3) <5 mismatches in the HBV side. Paired breakpoints were defined as two breakpoints located within 20 kb in the human genome with opposite orientation. In addition, all the reported virus-host junctions were manually reviewed using the Integrative Genomics Viewer[50,51].

Breakpoint annotation and visualization

All breakpoints were annotated using SeattleSeq Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). The R package OmicCircos was applied to the circos map to show the distribution [52], and the R package GenVisR was for the waterfall map of the recurrent integration events and the number of integrations per Mb in each patient[53]. Integrative Genomics Viewer (IGV) was used to visualize the read alignment in the integration region[51]. The human genome browser at UCSC was used for the visualization of genomic features near the integration sites[54].

DNA Microarray experiment for structure variation analysis

The genome wide genotyping of tumor tissues was performed using the HumanCoreExome-24 BeadChip (Illumina Inc.), which was scanned by the iScan Reader (Illumina Inc). The LogR ratio (LRR) and B allele frequency (BAF) of each genotyped locus were extracted by Illumina GenomeStudio 2011. Variations in copy numbers by each sample were determined by pennCNV[55].

Statistical analysis

Statistical analysis including Chi-square test, t-test, and linear regression, was performed using R packages (https://www.r-project.org/).

Data availability

The sequencing data generated by the present study was deposited in the BIG Data Center (http://bigd.big.ac.cn/bioproject/) under the BioProject accession code PRJCA000834.

Author Contributions

The concept and study design was provided by D.Z., K.Z, W.C., C.Z. and U.P.; patient enrollment, the sample and clinic data collection were performed by P.D., Z.W. X.Y. and H.D.; the sequencing experiments were performed by C.T., and D.Z.; the data analysis and interpretation were performed by W.C., D.Z., K.Z. P.D., Y.H., S.G, H.Z., U.P. and G.F.; the manuscript preparation was performed by: D.Z., K.Z., G.F. and U.P.; overall responsibility was given to D.Z., W.C., K.Z, P.D., U.P. and C.Z

Conflicts of interests

Dr. Dake Zhang has a patent pending for the probe-based HBV DNA capture in plasma as a liquid biopsy to monitor HCC development. The authors declare no other potential conflicts of interests.

Acknowledgements: This work was presented in the International Liver Congress™ 2018 at Paris in the late breaker section (LBP-029). This manuscript has been edited and proofread by Medjaden Bioscience Limited. This project was partially supported by the Innovation Promotion Association CAS (2016098), the National Natural Science Foundation of China (81201700), the Major State Basic Research Development Program (2014CB542006), the Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L14), Beijing Natural Science Foundation (7192158), the Fundamental Research Funds for the Central Universities (3332018032) and the Capital's Funds for Health Improvement and Research (2018-1-1151). The sponsor or funding organization had no role in the design or conduction of this research.

Reference

1. Yaginuma K, Kobayashi H, Kobayashi M, Morishima T, Matsuyama K, Koike K. Multiple integration site of hepatitis B virus DNA in hepatocellular carcinoma and chronic active hepatitis tissues from children. J Virol 1987;61:1808-13.

2. Nagaya T, Nakamura T, Tokino T, et al. The mode of hepatitis B virus DNA integration in chromosomes of human hepatocellular carcinoma. Genes & development 1987;1:773-82.

3. Shaul Y, Garcia P, Schonberg S, Rutter W. Integration of hepatitis B virus DNA in chromosome-specific satellite sequences. Journal of virology 1986;59:731-4.

4. Hino O, Shows TB, Rogler CE. Hepatitis B virus integration site in hepatocellular carcinoma at chromosome 17; 18 translocation. Proceedings of the National Academy of Sciences 1986;83:8338-42.

5. Fowler M, Greenfield C, Chu C-M, et al. Integration of HBV-DNA may not be a prerequisite for the maintenance of the state of malignant transformation: an analysis of 110 liver biopsies. Journal of hepatology 1986;2:218-29.
6. Dejean A, Bougueleret L, Grzeschik K-H, Tiollais P. Hepatitis B virus DNA integration in a sequence homologous to v-erb-A and steroid receptor genes in a hepatocellular carcinoma. Nature 1986;322:70-2.
7. Yaginuma K, Kobayashi M, Yoshida E, Koike K. Hepatitis B virus integration in hepatocellular carcinoma DNA: duplication of cellular flanking sequences at the integration site. Proceedings of the National Academy of Sciences 1985;82:4458-62.
8. Rogler C, Sherman M, Su C, et al. Deletion in chromosome 11p associated with a hepatitis B integration site in hepatocellular carcinoma. Science 1985;230:319-23.
9. Koshy R, Koch S, Von Loringhoven AF, Kahmann R, Murray K, Hofschneider P. Integration of hepatitis B virus DNA: evidence for integration in the single-stranded gap. Cell 1983;34:215-23.
10. Mason WS, Gill US, Litwin S, et al. HBV DNA Integration and Clonal Hepatocyte Expansion in Chronic Hepatitis B Patients Considered Immune Tolerant. Gastroenterology 2016;151:986-98 e4.
11. Wooddell CI, Yuen MF, Chan HL, et al. RNAi-based treatment of chronically infected patients and chimpanzees reveals that integrated hepatitis B virus DNA is a source of HBsAg. Sci Transl Med 2017;9.
12. Tan AT, Yang N, Lee Krishnamoorthy T, et al. Use of Expression Profiles of HBV-DNA Integrated Into Genomes of Hepatocellular Carcinoma Cells to Select T Cells for Immunotherapy. Gastroenterology 2019.
13. Lesbats P, Engelman AN, Cherepanov P. Retroviral DNA Integration. Chem Rev 2016;116:12730-57.
14. Andrake MD, Skalka AM. Retroviral Integrase: Then and Now. Annu Rev Virol 2015;2:241-64.
15. Yang W, Summers J. Integration of hepadnavirus DNA in infected liver: evidence for a linear precursor. J Virol 1999;73:9710-7.
16. Tu T, Budzinska MA, Shackel NA, Urban S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. Viruses 2017;9.
17. Minami M, Poussin K, Brechot C, Paterlini P. A Novel PCR Technique UsingAlu-Specific Primers to Identify Unknown Flanking Sequences from the Human Genome. Genomics 1995;29:403-8.
18. Murakami Y, Saigo K, Takashima H, et al. Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. Gut 2005;54:1162-8.
19. Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 2003;96:317-23.
20. Judo MS, Wedel AB, Wilson C. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 1998;26:1819-25.
21. Sung W-K, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nature genetics 2012;44:765-9.
22. Jiang Z, Jhunjhunwala S, Liu J, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome research 2012;22:593-601.
23. Fujimoto A, Totoki Y, Abe T, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nature genetics 2012;44:760-4.
24. Zhao L-H, Liu X, Yan H-X, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. Nature communications 2016;7:12992.
25. Furuta M, Tanaka H, Shiraishi Y, et al. Characterization of HBV integration patterns and timing in liver cancer and HBV-infected livers. Oncotarget 2018;9:25075-88.
26. Heikenwalder M, Protzer U. LINE(1)s of evidence in HBV-driven liver cancer. Cell Host Microbe 2014;15:249-50.
27. Pineau P, Marchio A, Terris B, et al. A t (3; 8) chromosomal translocation associated with hepatitis B virus intergration involves the carboxypeptidase N locus. Journal of virology 1996;70:7280-4.
28. Becker SA, Zhou Y-Z, Slagle BL. Frequent loss of chromosome 8p in hepatitis B virus-positive hepatocellular carcinomas from China. Cancer research 1996;56:5092-7.
29. Meyer M, Wiedorn KH, Hofschneider PH, Koshy R, Caselmann WH. A chromosome 17: 7 translocation is associated with a hepatitis B virus DNA integration in human hepatocellular carcinoma DNA. Hepatology 1992;15:665-71.
30. Tokino T, Matsubara K. Chromosomal sites for hepatitis B virus integration in human hepatocellular carcinoma. Journal of

virology 1991;65:6761-4.
31. Wang H, Rogler C. Deletions in human chromosome arms 11p and 13q in primary hepatocellular carcinomas. Cytogenetic and Genome Research 1988;48:72-8.
32. Dandri M, Locarnini S. New insight in the pathobiology of hepatitis B virus infection. Gut 2012;61 Suppl 1:i6-17.
33. Zhao Y, Xue F, Sun J, et al. Genome-wide methylation profiling of the different stages of hepatitis B virus-related hepatocellular carcinoma development in plasma cell-free DNA reveals potential biomarkers for early detection and high-risk monitoring of hepatocellular carcinoma. Clinical epigenetics 2014;6:30.
34. Shi L, Li S, Shen F, et al. Characterization of nucleosome positioning in hepadnaviral covalently closed circular DNA minichromosomes. J Virol 2012;86:10059-69.
35. Zhang D, Chen W, Zhang K, Dong P, Protzer U, Zeng C. Viral integration profiles in the plasma cell-free DNA from patients with HBV infection well represent tumor clone compositions during HCC development. J Hepatol 2018;68:S121-S2.
36. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nature genetics 2017;49:635-42.
37. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol 2017;14:531-48.
38. Tan AT, Yang N, Lee Krishnamoorthy T, et al. Use of Expression Profiles of HBV-DNA Integrated Into Genomes of Hepatocellular Carcinoma Cells to Select T Cells for Immunotherapy. Gastroenterology 2019;156:1862-76 e9.
39. Lin SY, Steffen JD, Su Y-P, et al. Detection of HCC-derived major HBV integration junctions in urine and their implications for driver identification. AACR; 2017.
40. Wang Y, Springer S, Mulvey CL, et al. Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas. Sci Transl Med 2015;7:293ra104.
41. Hyun KA, Gwak H, Lee J, Kwak B, Jung HI. Salivary Exosome and Cell-Free DNA for Cancer Detection. Micromachines (Basel) 2018;9.
42. Hu B, Wang R, Fu J, et al. Integration of hepatitis B virus S gene impacts on hepatitis B surface antigen levels in patients with antiviral therapy. J Gastroenterol Hepatol 2018;33:1389-96.
43. Hou J, Wang G, Wang F, et al. Guideline of Prevention and Treatment for Chronic Hepatitis B (2015 Update). J Clin Transl Hepatol 2017;5:297-318.
44. Rozanov M, Plikat U, Chappey C, Kochergin A, Tatusova T. A web-based genotyping resource for viral sequences. Nucleic Acids Res 2004;32:W654-9.
45. Yang L, Luquette LJ, Gehlenborg N, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 2013;153:919-29.
46. Yang L, Lee MS, Lu H, et al. Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. Am J Hum Genet 2016;98:843-56.
47. Wu GH, Ding HG, Zeng CQ. Overview of HBV whole genome data in public repositories and the Chinese HBV reference sequences. Prog Nat Sci-Mater 2008;18:13-20.
48. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17:3.
49. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.
50. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res 2017;77:e31-e4.
51. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24-6.
52. Hu Y, Yan C, Hsu CH, et al. OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data. Cancer Inform 2014;13:13-20.
53. Skidmore ZL, Wagner AH, Lesurf R, et al. GenVisR: Genomic Visualizations in R. Bioinformatics 2016;32:3012-4.
54. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res 2002;12:996-1006.
55. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17:1665-74.

1
2

Figure legends

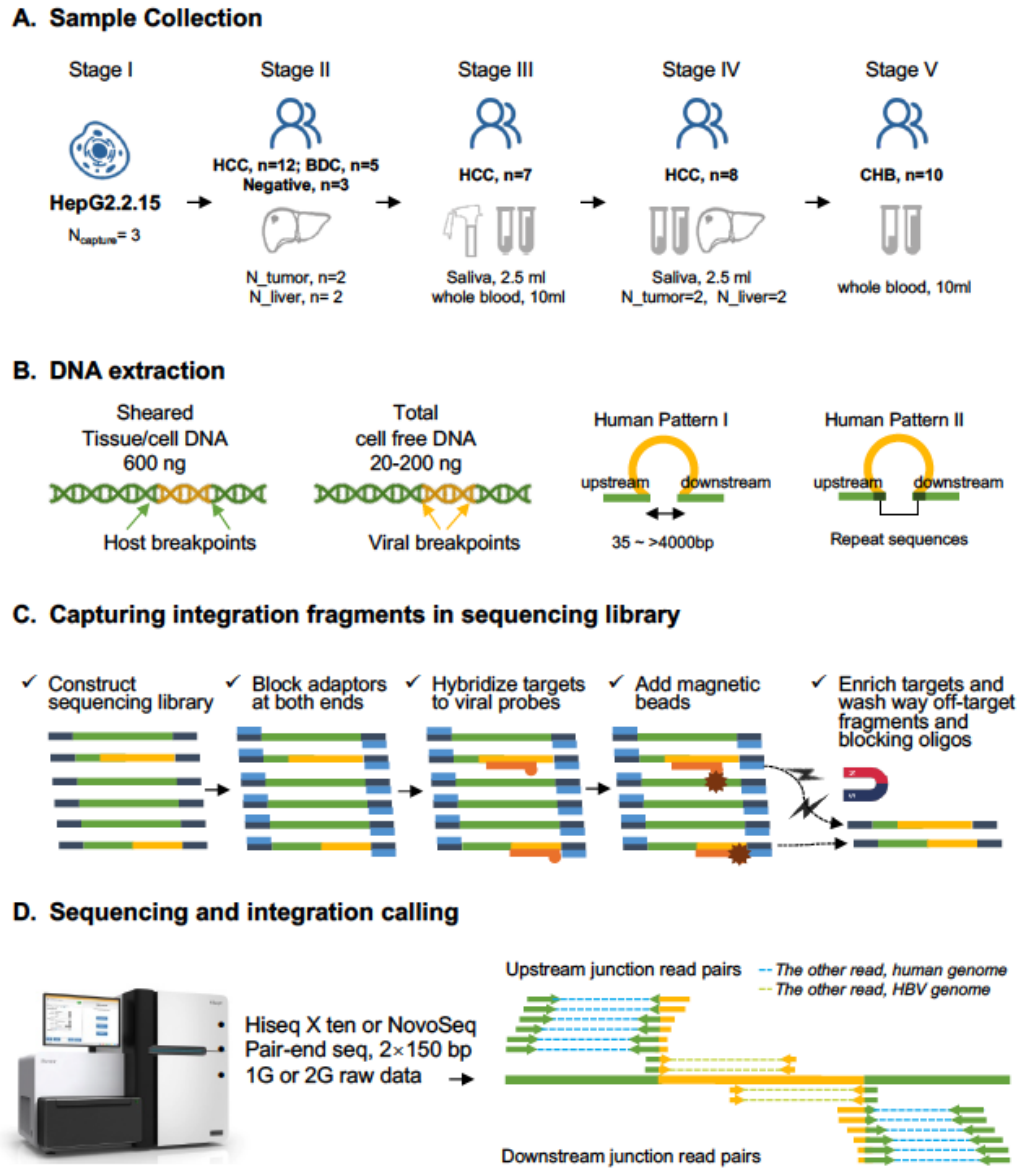

Figure 1. Study design. A. Sample collection in five stages. HCC, hepatocellular carcinoma; BDC, bile duct carcinoma; negative, 3 patients with liver cancer but without HBV infection. **B.** DNA amounts for sequencing library construction in tissue and plasma (Left). Integrations lead to two host breakpoints and two viral breakpoints in the human genome and HBV genome, respectively. Two host breakpoints are located at upstream and downstream of the integrated viral fragment. Most of integration sites have deletions in human genome, leading two 35- <4000 base pairs in distance between two host breakpoints (Pattern I). In some cases (Pattern II), sequences of both breakpoint are consistent. **C.** Experiment work flow for the capture assay. **D.** Sequencing volumes for captured fragments (Left), and junction read mapping to the reference region of integration sites. Human fragments in the virus-host junction reads can be mapped to either the upstream or downstream of the breakpoints. Theoretically, each integration event should be supported by these four types of junction read pairs with adequate read depth at both upstream and downstream breakpoints. In pair-end sequencing, at most, one read in a read pair would represent the junction read covering the integration boundary. The other read would either be a host fragment (read pairs with a dashed line in blue) or a viral fragment (read pairs with a dashed line in light green), and the read alignment shows the mapping of these two groups of read pairs in the corresponding color.

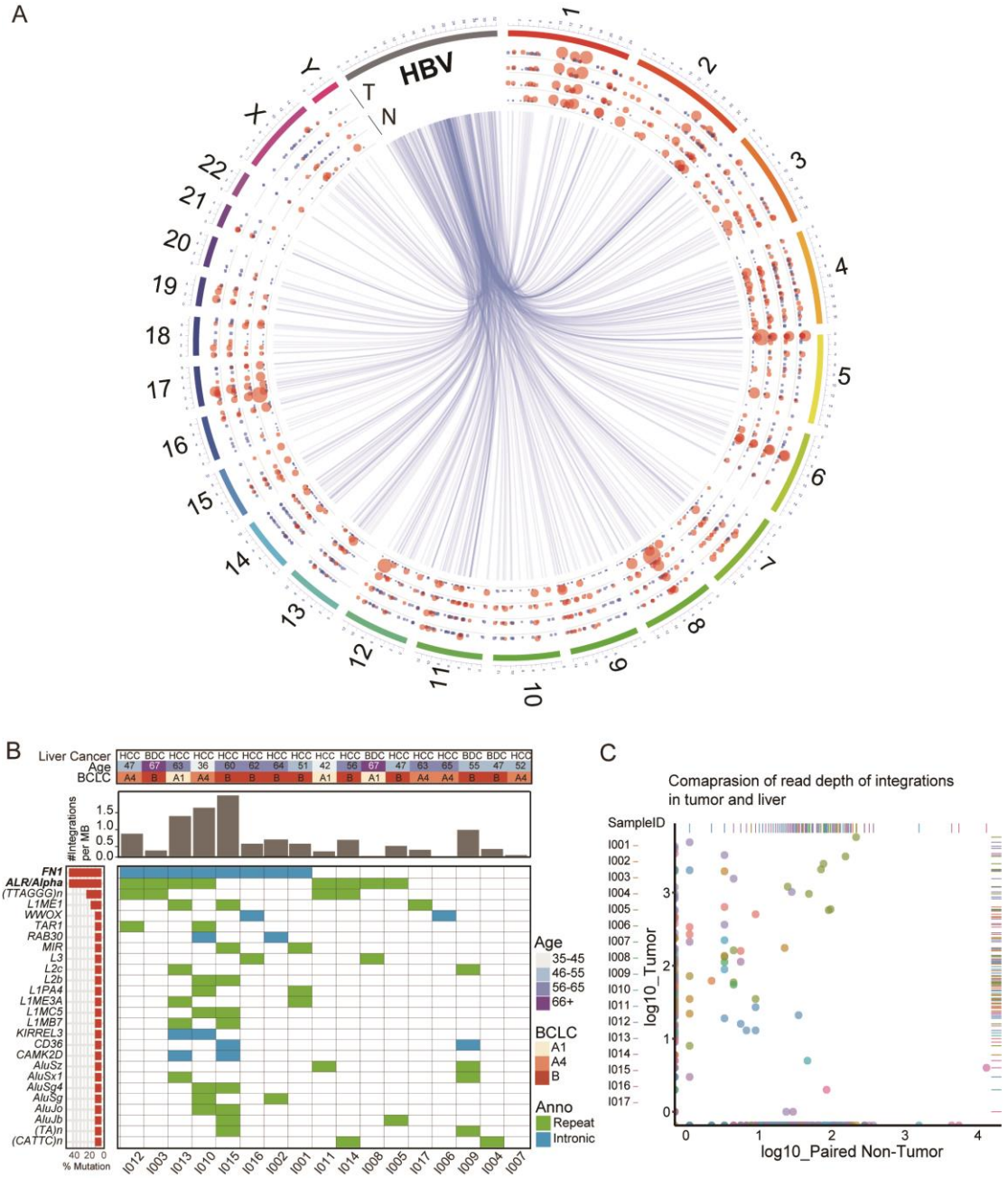

Figure 2 (A) All integration events connecting the viral and human genomes. Each light blue line indicates one integration event, with one end showing the breakpoint in the HBV genome and the other in human chromosomes. Each integration event was only observed in one patient. Bubbles with diverse diameters, between chromosomes and central connections, illustrate the sequencing read depth in multiple samples from each patient. T: two samples from tumor tissues; N: adjacent non-HCC tissues. The hotspot for viral breakpoints locates at approximately nt 1,600-1,900. **(B)** Integration events with the same sequence features at the disrupted human genome regions and integration burden in all patients. The top panel shows the diagnosis of liver cancer, the age of the patient and the Barcelona clinic liver cancer (BCLC) stages. The middle panel provides the number of integration events per MB according the total events observed in all four solid tissue samples obtained from each individual. At the bottom panel, patients were listed by frequencies (left part) of the 26 types of integrations observed in >2 patients (right part). **(C)** Comparison of sequencing read depths of integrations in tumors and adjacent liver tissues. The values for the sequencing depths were log transformed. Each dot indicates the sequencing read depth of an integration in tumor (y axis) and non-HCC liver (x axis). A higher read depth in two sites of each sample was used for plotting. Diverse colors indicate the different patients.

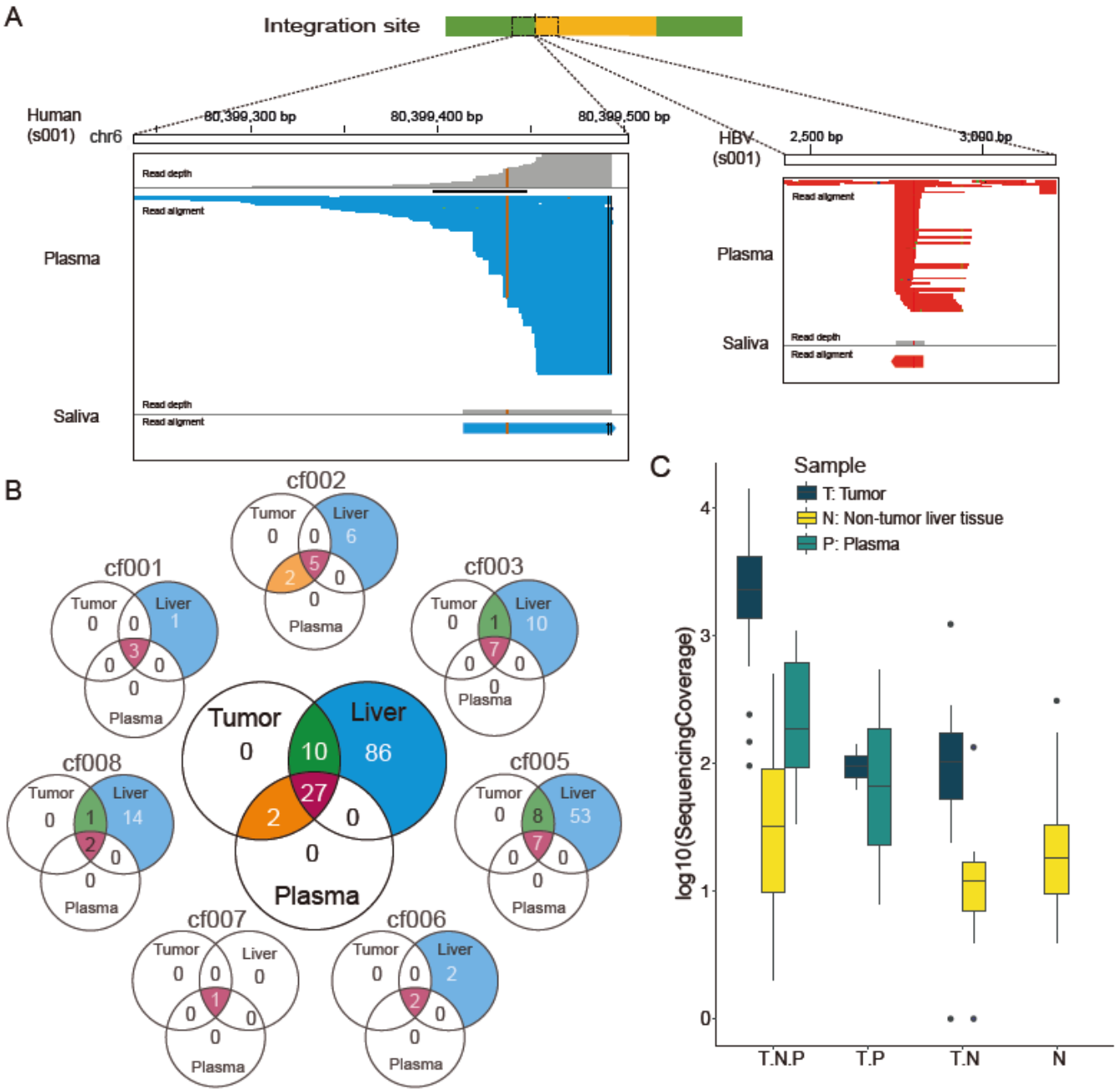

Figure 3 Detection of integration events in saliva and plasma obtained from HCC patients. (A) The integration event in saliva. Only one non-redundant junction read is obtained in comparison to high sequencing coverage in paired plasma sample. (B) The cfDNA was successfully extracted in seven of eight patients, and viral integration events were detected in all seven plasma samples. More integration events were identified in paired liver tissues. The limited integration events were shared between tumors and paired non-HCC liver samples (green). In particular, the detected integrations in plasma well-reflected the counterparts uncovered in tumor tissues (orange and purple), and the integration events specific to paired non-HCC liver tissues (blue) were not observed in the corresponding plasma samples. **(C)** The read depths of integration events in tumors, non-tumor liver tissues and plasma samples. For those detected in all three tissues (T.N.P.).

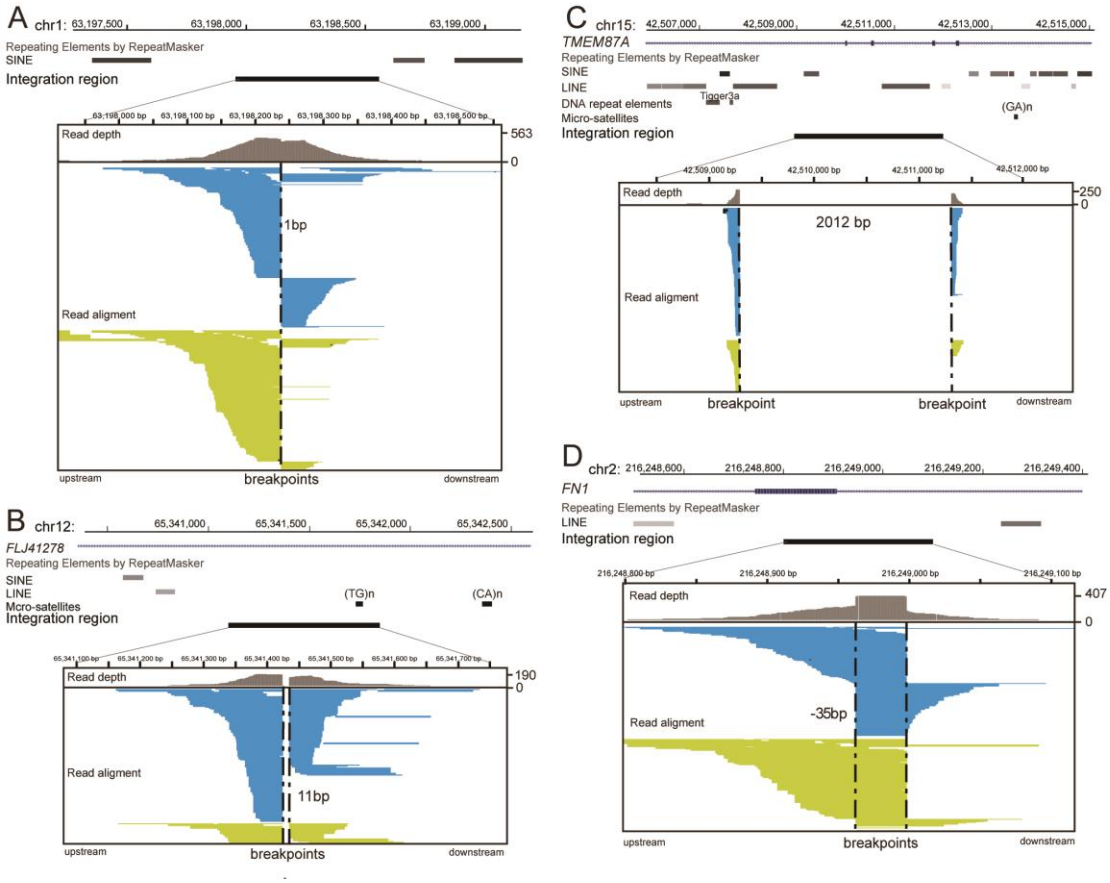

Figure 4 Patterns of integration events according to the read alignment features of host breakpoints in the human genome. At the integration sites of the human genome, there can be lost sequences with diverse length (d, the position of the downstream breakpoint minus the upstream one). There can also be no loss of host sequences in the integration sites, and the d should be 1 bp **(A)**. Deletions with a diverse length can also be observed at the integration site of the human genome. For instance, a 11 bp deletion **(B)** and a 2,012 bp deletion **(C).** For host pattern II, the longest redundant sequence is 35 bp **(D)**. For these integrations, the top panel shows the transcripts of genes and repeat sequences near the corresponding regions in the human genome according to the online UCSC genome browser (hg19). The shading of repeat elements reflects the amount of sequence variations associated with the repeat element. The darker it is, the fewer such variations are observed.

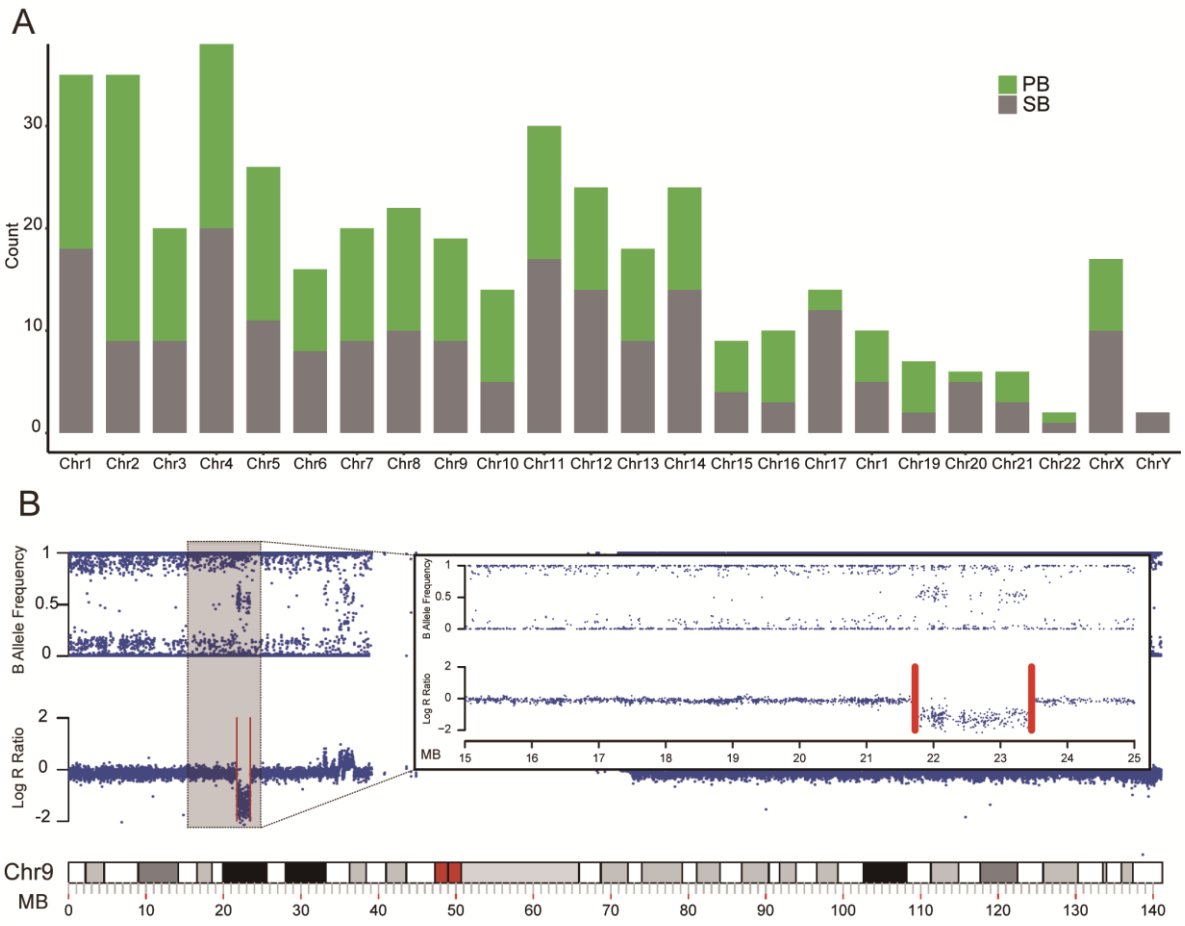

Figure 5 (A) Chromosome locations for all integration events. Events with identified paired boundaries (PB) are presented in green, while single boundaries (SB) are presented in grey. **(B)** The influence of structure variations on the boundary or breakpoint pairing analysis. An example was provided to show that two breakpoints at chromosome 9, which were separated by 1.7 M bp, were located exactly at the two boundaries of a same length genomic deletion.

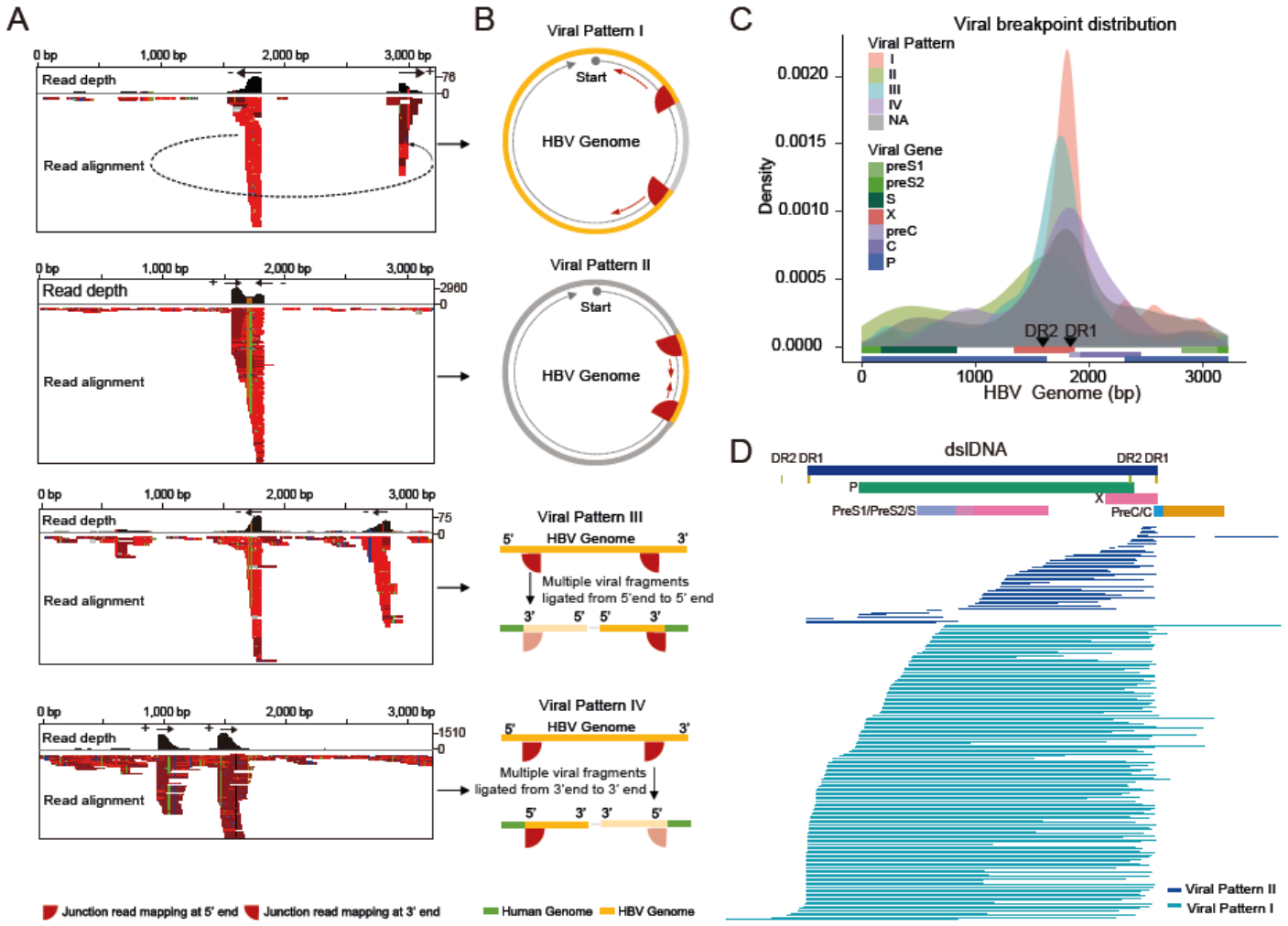

Figure 6. Patterns of integration events according to the features of paired breakpoints in the HBV genome (A). The mapping of viral fragments in virus-human junctions to the HBV genome is shown. The reads mapped to the plus strand are in red, and those to the minus strand are in dark red. The read depth shows the sequencing coverage of each base along the viral genome, and the read alignment demonstrates the mapping of all reads to the genome. The arrows illustrate the extending direction according to the tail of the peak at the breakpoints. The integrants were predicted according to the directions at two boundaries of the integration. All four combinations of two directions at breakpoints, and their schematic diagram are illustrated in **(B)**. The red fans summarize the features of the junction read mapping at the breakpoints of both 5' and 3' ends . The curved edge indicates the inconsistent ending of the reads, and the vertical edge indicates the consistent boundary. The yellow parts of the circle indicate the estimated integrants. Viral Patterns III and IV seem to have multiple viral fragments firstly ligated in different ways before integrating into the host genome. **(C)** The breakpoint distribution across the HBV genome is shown**.** The distributions of the four viral patterns (I-IV) and the unpaired breakpoints (NA) in different colors, as well as the hotspot for viral breakpoints located around the DR1-DR2 region, are shown. **(D)** All inferred integrants for viral pattern I and II are shown using the dslDNA format of the HBV genome as a reference sequence.