**Noninvasive chimeric DNA profiling identifies tumor-originated HBV integrants contributing to viral antigen expression in liver cancer**

Wei Chen1,#, Ke Zhang2,4, #, Peiling Dong3, #, Gregory Fanning4, Chengcheng Tao1, Haikun Zhang 1,5, Shicheng Guo6, Zheng Wang3, Yaqiang Hong1,5, Xiaobo Yang7, Huiguo Ding3, Haitao Zhao7, Changqing Zeng1,\* , Ulrike Protzer2,8,\*, Dake Zhang1,\*

1Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China

2Institute of Virology, Technical University of Munich / Helmholtz Zentrum München, 81675, Munich, Germany

3Department of Hepatology, Beijing You’an Hospital Affiliated with Capital Medical University, Beijing 100069, China

4Janssen China Research & Development Center, Shanghai 201210, China

5University of Chinese Academy of Sciences, Beijing 100049, China

6Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA

7Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

8German Center for Infection Research (DZIF), Munich partner site, Germany

# Equal contribution

\*Corresponding Authors

**Correspondence:**

Changqing Zeng, Ph.D., 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, MD/Ph.D., 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](https://wx.qq.com/cgi-bin/mmwebwx-bin/webwxcheckurl?requrl=http%3A%2F%2Fprotzer%40tum.de&skey=%40crypt_f9112205_767509501ff5e792c8538f606badd7c7&deviceid=e900409222751532&pass_ticket=undefined&opcode=2&scene=1&username=@f11e8767c386bf2338a3914547cb0933)

Dake Zhang, Ph.D., 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](mailto:zhangdk@big.ac.cn).

Running title: *Chen W et al / HBV integration and liver cancer screening*

Total word counts: 5212

# Figures: 6

# Tables: 0

# Supplementary figures: 5

# Supplementary tables: 8

## Abstract

HBV integration events frequently occur in liver cells during chronic infection and can contribute to cancer development. To screen for HBV integration in HBsAg positive individuals, we developed a low-pass sequencing method and achieved the efficient enrichment of plasma cell-free chimeric DNA from integration site. This allowed us to implement a novel integration analysis strategy for inference of the whole integrant sequence, benefiting from high depth in short reads sequencing of viral and host breakpoints of individual integration event. We demonstrated that viral-host chimeric DNA fragments detected in plasma were mainly derived from integrations in tumor tissues rather than in adjacent liver tissues. In addition, we found that most viral integrations contain the complete opening reading frame of HBV surface proteins, and over 50% of them produced viral-host chimeric transcripts detected by deep RNA sequencing in paired tumor tissues. In summary, we demonstrate that chimeric DNA profiling in plasma is a promising non-invasive approach to monitor HBV integration in liver cancer development and to determine the ability of integrated sequences to express viral proteins that can be targeted e.g. by immunotherapies.

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

## Introduction

Hepatitis B virus (HBV) integration has been known to occur in hepatocellular carcinoma (HCC) and liver tissues for some time1, 2, 3, 4, 5, 6, 7, 8, 9. It has more recently been appreciated as an early event during HBV infection10. 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) genes11. A recent study has demonstrated, that liver cells harboring integrated HBV DNA sequences can express peptides that can be recognized by HBsAg specific T cells12. Unlike retroviruses13, viral integration is not required for HBV replication, and no HBV proteins are known to have integrase activity14. Double stranded linear DNA (dslDNA) seems to be preferred DNA substrate for integration15. Despite the attention that HBV integration has received in the HCC and HBV literature, the process and implication of the “side product” of infection remains largely unknown16.

Recently, the resequencing of HCC genomes has identified hundreds of integration sites in the human genome implicating many cellular genes17, 18, 19, 20. HBV integrations were more likely to occur in chromosome sites of genomic instability such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), microsatellites and telomeres16, 21. 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 the cells carrying the integrant. In particular, these integration sites were associated with structural variations in tumor cellular genomes. Further studies also showed integration sites tend to occur within boundaries of the altered copy numbers of a gene22, 23, 24, 25, 26, and may lead to genomic instability of the infected hepatocyte, thus contributing to the oncogenic role of viral integration18. It is unclear at this point in time whether dominant integration sites exist because of a chromosome structure that facilitated integration or because the integration event itself provided tumorigenic potential. Both likely play a role and as such, at a minimum, integration patterns serve as biomarkers that can detect cell clonal expansion in both tumor and non-tumor tissue10, 27.

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 application28. Minichromosomes of HBV covalently closed circular DNA (cccDNA) in liver cells are able to form nucleosomes 29, and integrated viral DNA is expected to have similar chromatin structure that is known to be nuclease-degradation proof. Thus, a plasma-derived, cell free DNA (cfDNA) pool may contain virus-host junction DNA fragments that could serve as a biomarker, and reflect a part of the genetic changes in tumor genomes28. The detection of virus-host chimeric DNA in plasma could serve to detect HBV-integration events in “liquid biopsies”, i.e. in a noninvasive fashion30.Theoretically, the analysis of the plasma cfDNA pool is not only more accessible but also more representative of the entire liver than single liver biopsies, and it can be performed upon each patient visit. The challenge is that 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 cfDNA31, and the deep sequencing of cfDNA is very costly. Therefore, we aimed at enriching viral-host fragments using capture probes prior to deep sequencing to reduce the sequencing volume and increase the sensitivity to detect HBV integration events.

In this study, we designed viral DNA probes covering the whole HBV genome so that we could enrich plasma cfDNAs for HBV containing sequences for deep sequencing. The small size of the HBV genome enabled us to achieve a deep sequencing coverage at a small sequencing volume, with a significantly increased ability to detect viral integrants. Using this technique, we analyzed tumor and adjacent non-tumor liver samples alongside cfDNA derived from either plasma or saliva sampled from the same patient. We found this to be an efficient strategy to characterize viral integration events stemming from tumors with a potential use to characterize liver cancers in patients with HBV infection in a noninvasive fashion. In addition, we use the short reads sequencing data to predict the orientation and length of integration by pairing the two ends of each integration event.

## Results

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

To enrich for viral integrants in cfDNA isolated from plasma and reduce the required amount of sequencing data, we designed DNA probes specific for HBV DNA sequences (Method). To validate the DNA probe capture of HBV DNA, we used the HepG2.2.15 cells that are known to harbor integrated copies of HBV (Stage I, Figure 1A). We found sequencing uncovered multiple 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 ends (viral breakpoints, Figure 1B) and cellular genome ends (host breakpoints, Figure 1B). In total, five integration sites in HepG2.2.15 cells were identified using our protocol. We then applied the method to 80 biopsy samples collected from 20 liver cancer patients (Study Design in Figure 1A). In total, we identified 424 integration events (Figure 2A). The number of detected integration events detected per sample varied from 2 to 82 (average: 25) among the 17 HBsAg positive patients, and none were identified in 3 HBsAg negative patients that were used as a negative control. While the numbers of integrations varied between patients, there was no significant difference in integration events between the tumor sites and paired adjacent non-tumor sites of the same individuals (t test, P>0.05, Figures S1-S2).

Genomic annotation of 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). The most commonly directly interrupted gene was *FN1*, which was detected in 8 out of 17 patients; while the most common directly 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, were also very common (23.5%, 4/17). We also observed 29.4% (5/17) patients had integration sites in the promoter region of *TERT*, which was consistent with a previous finding (23.7%,18/76) 17.

Our capture strategy achieved high sequencing coverage (number of junction reads) of integration breakpoints, with 70% (296/424) of integrations having sequencing coverage over 100 in at least one of four aliquots analyzed in the same individual. 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 recently adopted liquid biopsy solution for other cancers32, we collected the saliva samples along with plasma samples for 7 liver cancer patients (Stage III, Figure 1A; Table S4). In total, 32 integration events successfully detected in 5 patients (Table S5). Only one patient had integration detected in saliva (Table S5). Although she had five integrants with high abundance (250-881 junction read pairs) in plasma (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 were detected in the saliva samples suggesting that saliva is not a suitable sample for detection of viral integration in liver cancer.

To trace the origin of integrants in plasma cfDNA, we collected paired tumor and adjacent normal tissues from HCC patients (Stage IV, Figure 1A; Clinical information, Table S4). The cfDNA in all seven HBV positive HCC patients was analyzed (one plasma sample failed in the cfDNA extraction) (Table S4). Overall, the plasma integration events predominantly reflected the tumor tissues. First of all, we detected 29 integration events from 7 plasma cfDNAs (MaxCF>0, Table S6) and all of them (29/29, 100%) could be detected in the corresponding liver tumor samples (Figure 3B, Table S6). Notably, junction abundance for integration in plasma samples correlated with that in tumor samples (R2=0.64, P= 6.2x10-29, Table S6; Figure 3C), but not in the paired non-tumor liver tissues (R2=0.32, *P*=6.2x10-12). This also supported cell free junctional DNA were shed by the tumor tissue compared to the non-tumor tissue.

Deep RNA sequencing for the same tumor and adjacent liver aliquots were performed for four patients (c001, c002, c003, c005 in Table S6). Totally, chimeric RNA transcripts were observed for 17 integration sites, among which 13 were also seen in plasma cfDNA and the rest four were only observed in the transcriptome data. As a result, among 24 integrants in cfDNAs from these four patients, 54% (13/24) had transcription activity. Besides, the amounts of chimeric transcripts from integration sites showed a better association with DNA junction abundance in tumor tissues (R2=0.27, *P* = 1.5x10-8, Table S6), than that in adjacent liver tissues (R2=0.01, *P* = 0.15). It suggested integration events observed in cfDNAs were more likely to have transcription activity in tumor than in adjacent liver tissues, and they would contribute to viral antigen production if they kept intact ORFs of these proteins.

Taken together, we find plasma integration profile is valuable for the detection of tumor derived integration events, and RNA sequencing data not only confirmed that some of these integrations are transcriptionally active but information about tumor expression could be derived from the plasma. Then, we also sought HBV integrations in the plasma cfDNAs of 10 chronic hepatitis B patients without liver cancer (Stage V, Figure 1A; Clinical information, Table S4). No events met the integration criteria in all these patients (Methods section). 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 (Figure S1). However, their authenticity should be supported by analyzing paired tissue samples from liver biopsy. The scarcity of DNA fragments derived from chronically infected liver was likely due to a combination of limited turnover of infected cells and clonality. It would be interesting to evaluate the cfDNA from patients where there is an active HBsAg specific T cell response 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 accurately infer the HBV integration sequence and orientation will be valuable for assessing the viral protein/peptide potential of integrations. Each integration should have two viral breakpoints and two host breakpoints (Figure 3A). As a premise we assumed that two independent integration events in host genome should be far away from each other. Therefore, we applied a rule that two cellular breakpoints that occurred within 20 K base pairs (bp) were 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 breakpoints ranged between 35 bp and >4,000 bp, and mostly between 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) was relatively low. Notably, seven of the 11 integrations in the intronic region of *FN1* gene had repeated sequences at the breakpoints.

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 for individual integration, we obtained the corresponding sequences covering viral boundary of this integration event. Then, accurate mapping of the integrated viral sequence could be achieved (Figure 6). Among 218 integration events with known host breakpoints, 215 integrated sequences could be characterized including the orientation of the viral sequence. Four distinct viral sequence patterns were observed (Figures 6A-6B). Their viral breakpoints were relatively consistent, showing similar distributions between the integration patterns. The majority of integrated sequences consisted of nt 1,600-1,900 of the viral genome (64.2%, 138/215) (Figures 6C-6D). This region included 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 from 952 to3,214 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). 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) appeared to be formed by ligation of the ends of least two viral fragments in a 3’-to-3’ or 5’-to-5’ manner. In addition, most individual samples 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 observed chimeric RNA transcripts from all four viral patterns of integration sites in tissue samples with both DNA capture experiments and deep RNA sequencing. In all, 76 integrations were identified in the DNA capture experiments (Table S8). Among them, viral patterns of breakpoints were determined in 42 integrants, and each pattern had 19% integrations with transcription activity (Table S8). Therefore, there were no obvious difference in transcription activity among these four patterns of integration events.

## Discussion

In the present study, we adopted an HBV capture strategy to enrich DNA fragments covering the virus-host junctions of integration events in plasma cfDNA from patients with HBV infection. The method 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 originated from tumor clones in the liver. This highlights the exciting possibility that HBV integration profiling in plasma may be included as a biomarker or part of a panel of analyses from liquid biopsies to assist in liver cancer profiling. Inference of the whole integrant sequence may provide a new companion diagnostic for HCC immunotherapy as it helps to predict potential target antigens for T cell therapy33.

Unlike circulating tumor DNA which is confounded by the DNA released from blood cells, detection virus-host DNA junction maybe influenced by both integrated and non-integrated HBV DNA. There have been efforts to take urine as specimen source34, 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 concentration35, 36. In our approach, it however didn’t prove suitable. After enrichment procedure, we only obtained maximally one integrated fragment in 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 increases the sensitivity of integration detection, and reduced the interference from non-integrated HBV DNA in samples. Here, 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 1011 target molecules. The DNA extracted from each liver tissue consisted of 105 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 108 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 all HBV integration events. 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 individual as well as total integration events among all analyzed samples. The same, unique HBV integration site carried by many liver tumor cells was uncovered by a higher number of sequencing reads, supporting the idea that the relative abundance of HBV integrations can serve as a genetic marker for clonal selection and the expansion of affected hepatocytes10.

The results from the sequencing of liver tissue samples revealed an average of 35 non-redundant junctions for each host sequence breakpoint. This was higher than the 18 junctions reported by Zhao *et al*., using a similar strategy20. Notably, the highest sequencing read depth for the cellular sequence breakpoints from liver tissues reached 11,579 in the present study. Hence, we assume that the potential for identification of the majority of integration events in a given sample was adequately addressed in the present study. However, we cannot rule out that there is a possibility that more tissue aliquots, more probes or ultra-deep sequencing may identify additional integrations at low frequencies or increase the read depth for breakpoints than the use of relatively limited supporting reads, as used in the present study.

The detection of new integration events or changes in integration events over time may help to monitor disease progression in the liver and may have the potential to indicate metastases or secondary tumors both of which are hard to detect in particular in a cirrhotic liver. Theoretically, cfDNA in blood HBV- containing HBV-host DNA junctions should mirror integration events in the liver, since it must have been released from the liver as HBV only infects hepatocytes and only integrates into hepatocytes. The present data revealed that the HBV integration events detected in plasma cfDNA predominantly originated from liver cancer cells. In the cancer and paired non-HCC liver tissues from the patients at Stage II (Figure 1A), we observed that the integration events had much higher sequencing read numbers in non-tumor liver tissues, indicating that some non-tumor clones with integrations may already have significantly expanded before an HCC grew out (Figure 2C). Nevertheless, it was regrettable that corresponding plasma samples were not available to compare the abundance of HBV-DNA fragments in the circulation with that in liver tissue. However, investigation of the patients enrolled in Stage IV revealed that HBV integration events identified in HCC but not in those detected in non-tumor liver tissues had correspondent counter-part fragments detected in plasma (Figure 3B). Circulating cfDNA in plasma is most likely released from dying or circulating tumor cells and less likely from injured adjacent tissue. Thus, one would expect that less DNA is released from non-HCC liver tissue compared to liver tumor tissue.

Differences in integration profiles between tumor and non-tumor tissues reflect the different clone compositions of HCC and non-tumor, adjacent liver tissues, implying a divergent clone evolution of HCC an and non-tumor cells. It has been suggested that the occurrence of certain integrations is the driving force for tumorigenesis. For instance, HBV-integration may result in the upregulation of the expression of TERT, MLL4 and CCNE117. However, both HCC and non-HCC cells also shared common integration events. As these integrations occur randomly, some may contribute to the tumorigenesis process, while others do not. Different studies have demonstrated that consequences of integrations may be different in oncogenesis17, 20.

According to the paired viral breakpoints we detected, the sequences of the 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) preserved the ORF of the large surface protein, and 14 of the remaining 26 events had an intact ORF of the middle surface protein. All these integrations would of course also encode for the small envelop protein S and have the potential to secrete HBsAg. These observations support the suggestion that integrated HBV DNA provides significant additional capacity for HBsAg production besides HBV cccDNA, and represents a challenge to reduce HBsAg production 11, 37.

The present data provides evidence that the integration patterns varied among individual patients. Diverse patterns and different percentages of different integration sites in the same individual imply distinct HBV antigen expression patterns, which are expected to impact therapeutic responses to HBV treatment or efforts targeting tumor cells expressing viral proteins. The data presented demonstrate the potential clinical utility of the capture assay to monitor HBV integration events non-invasively in peripheral blood. 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. However, future prospective studies with a larger sample size are required to validate our findings.

## 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 38. 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 1A), 17 HBV-related cancer patients (12 HCC and 5 bile duct carcinoma, BDC, 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 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 probes39. 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 40, 41. First, a reference genome was constructed by adding a HBV genotype C consensus sequence into the human reference genome (hg19) as a pseudo chromosome42. Cutadpt v1.14 was used to remove low quality and adaptor-contaminated reads43. Cleaned reads were mapped on the reference genome using the BWA-MEM algorithm44. 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 Viewer45, 46.

### Breakpoint annotation and visualization

All breakpoints were annotated using SeattleSeq Annotation ([http://snp.gs.washington.edu/Seattle -SeqAnnotation138/](http://snp.gs.washington.edu/Seattle%20-SeqAnnotation138/)). The R package OmicCircos was applied to the circos map to show the distribution47, and the R package GenVisR was for the waterfall map of the recurrent integration events and the number of integrations per Mb in each patient48. Integrative Genomics Viewer (IGV) was used to visualize the read alignment in the integration region46. The human genome browser at UCSC was used for the visualization of genomic features near the integration sites49.

### 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 pennCNV50.

### 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 raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in BIG Data Center51, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA000446 that is publicly accessible at <http://bigd.big.ac.cn/gsa> .

## 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* **61**, 1808-1813 (1987).

2. Nagaya T*, et al.* The mode of hepatitis B virus DNA integration in chromosomes of human hepatocellular carcinoma. *Genes & development* **1**, 773-782 (1987).

3. Shaul Y, Garcia P, Schonberg S, Rutter W. Integration of hepatitis B virus DNA in chromosome-specific satellite sequences. *Journal of virology* **59**, 731-734 (1986).

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* **83**, 8338-8342 (1986).

5. Fowler 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* **2**, 218-229 (1986).

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* **322**, 70-72 (1986).

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* **82**, 4458-4462 (1985).

8. Rogler C*, et al.* Deletion in chromosome 11p associated with a hepatitis B integration site in hepatocellular carcinoma. *Science* **230**, 319-323 (1985).

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* **34**, 215-223 (1983).

10. Mason WS*, et al.* HBV DNA Integration and Clonal Hepatocyte Expansion in Chronic Hepatitis B Patients Considered Immune Tolerant. *Gastroenterology* **151**, 986-998 e984 (2016).

11. Wooddell CI*, 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* **9**, (2017).

12. Tan AT*, 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* **116**, 12730-12757 (2016).

14. Andrake MD, Skalka AM. Retroviral Integrase: Then and Now. *Annu Rev Virol* **2**, 241-264 (2015).

15. Yang W, Summers J. Integration of hepadnavirus DNA in infected liver: evidence for a linear precursor. *J Virol* **73**, 9710-9717 (1999).

16. Tu T, Budzinska MA, Shackel NA, Urban S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. *Viruses* **9**, (2017).

17. Sung W-K*, et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nature genetics* **44**, 765-769 (2012).

18. Jiang Z*, et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome research* **22**, 593-601 (2012).

19. Fujimoto A*, et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nature genetics* **44**, 760-764 (2012).

20. Zhao L-H*, et al.* Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nature communications* **7**, 12992 (2016).

21. Heikenwalder M, Protzer U. LINE(1)s of evidence in HBV-driven liver cancer. *Cell Host Microbe* **15**, 249-250 (2014).

22. Pineau P*, et al.* A t (3; 8) chromosomal translocation associated with hepatitis B virus intergration involves the carboxypeptidase N locus. *Journal of virology* **70**, 7280-7284 (1996).

23. Becker SA, Zhou Y-Z, Slagle BL. Frequent loss of chromosome 8p in hepatitis B virus-positive hepatocellular carcinomas from China. *Cancer research* **56**, 5092-5097 (1996).

24. 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* **15**, 665-671 (1992).

25. Tokino T, Matsubara K. Chromosomal sites for hepatitis B virus integration in human hepatocellular carcinoma. *Journal of virology* **65**, 6761-6764 (1991).

26. Wang H, Rogler C. Deletions in human chromosome arms 11p and 13q in primary hepatocellular carcinomas. *Cytogenetic and Genome Research* **48**, 72-78 (1988).

27. Dandri M, Locarnini S. New insight in the pathobiology of hepatitis B virus infection. *Gut* **61 Suppl 1**, i6-17 (2012).

28. Zhao Y*, 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* **6**, 30 (2014).

29. Shi L*, et al.* Characterization of nucleosome positioning in hepadnaviral covalently closed circular DNA minichromosomes. *J Virol* **86**, 10059-10069 (2012).

30. 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* **68**, S121-S122 (2018).

31. 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* **49**, 635-642 (2017).

32. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nat Rev Clin Oncol* **14**, 531-548 (2017).

33. Tan AT*, et al.* Use of Expression Profiles of HBV-DNA Integrated Into Genomes of Hepatocellular Carcinoma Cells to Select T Cells for Immunotherapy. *Gastroenterology* **156**, 1862-1876 e1869 (2019).

34. Lin SY*, et al.* Detection of HCC-derived major HBV integration junctions in urine and their implications for driver identification. (ed^(eds). AACR (2017).

35. Wang Y*, 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* **7**, 293ra104 (2015).

36. Hyun KA, Gwak H, Lee J, Kwak B, Jung HI. Salivary Exosome and Cell-Free DNA for Cancer Detection. *Micromachines (Basel)* **9**, (2018).

37. Hu B*, et al.* Integration of hepatitis B virus S gene impacts on hepatitis B surface antigen levels in patients with antiviral therapy. *J Gastroenterol Hepatol* **33**, 1389-1396 (2018).

38. Hou J*, et al.* Guideline of Prevention and Treatment for Chronic Hepatitis B (2015 Update). *J Clin Transl Hepatol* **5**, 297-318 (2017).

39. Rozanov M, Plikat U, Chappey C, Kochergin A, Tatusova T. A web-based genotyping resource for viral sequences. *Nucleic Acids Res* **32**, W654-659 (2004).

40. Yang L*, et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929 (2013).

41. Yang L*, et al.* Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. *Am J Hum Genet* **98**, 843-856 (2016).

42. 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* **18**, 13-20 (2008).

43. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 3 (2011).

44. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

45. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. *Cancer Res* **77**, e31-e34 (2017).

46. Robinson JT*, et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).

47. Hu Y*, et al.* OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data. *Cancer Inform* **13**, 13-20 (2014).

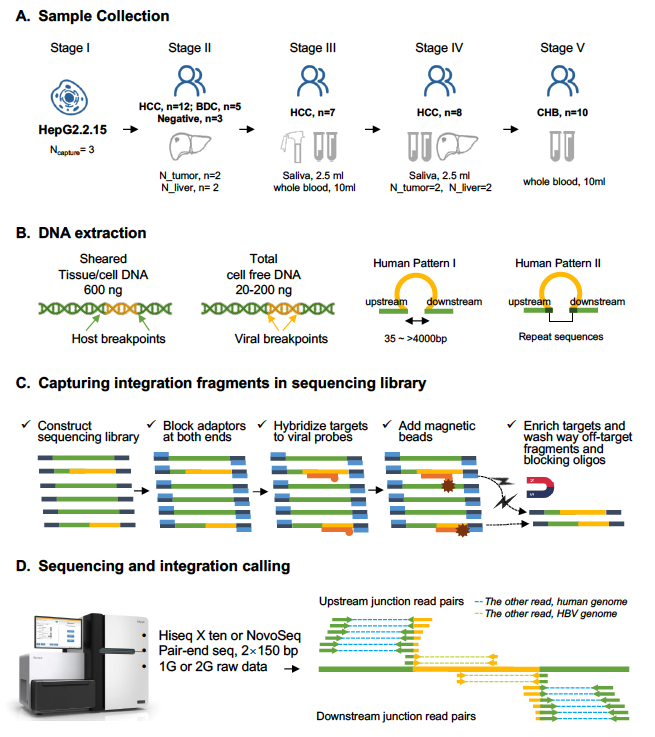
48. Skidmore ZL*, et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012-3014 (2016).

49. Kent WJ*, et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

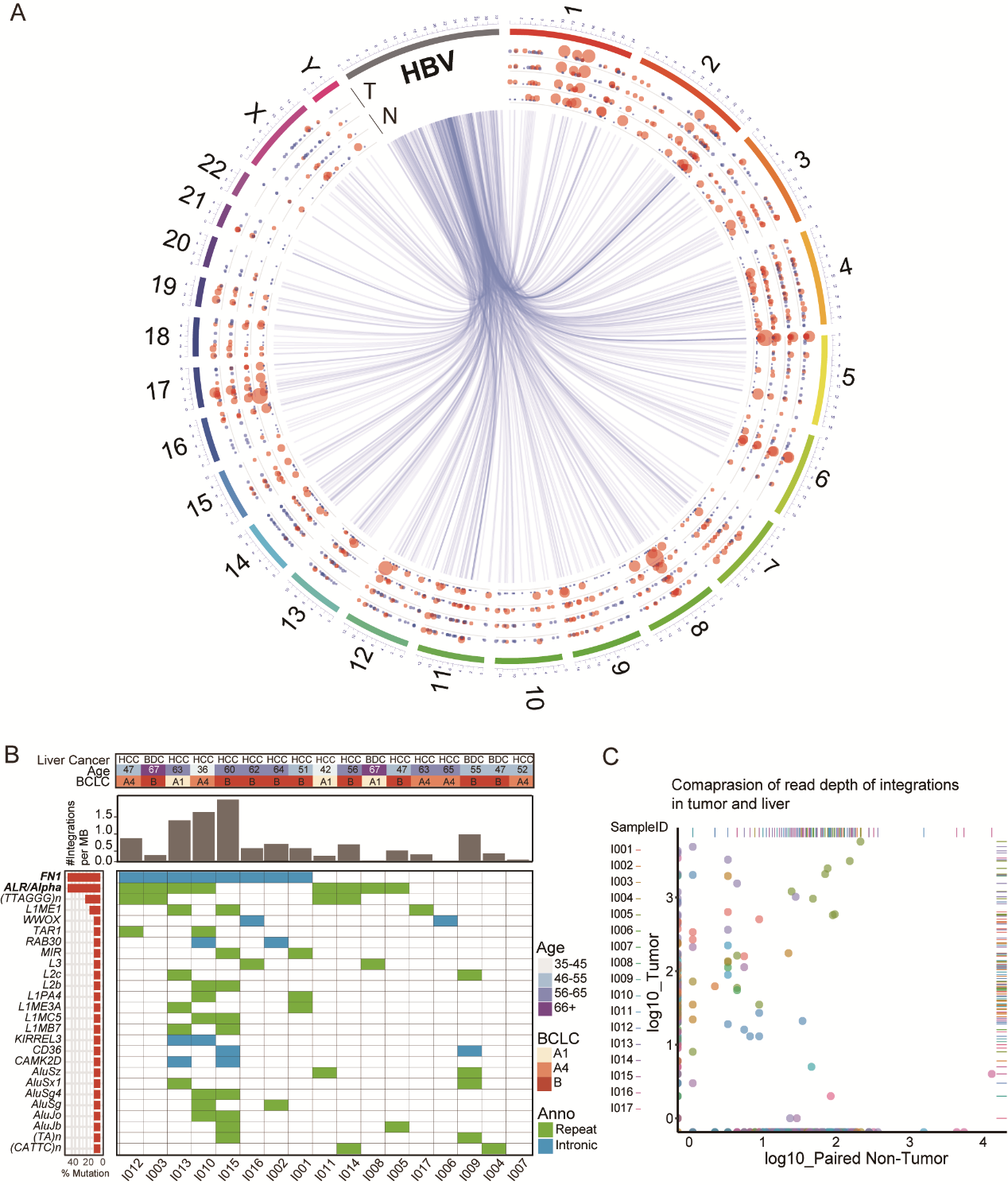
50. Wang K*, et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-1674 (2007).

51. Wang Y*, et al.* GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics* **15**, 14-18 (2017).

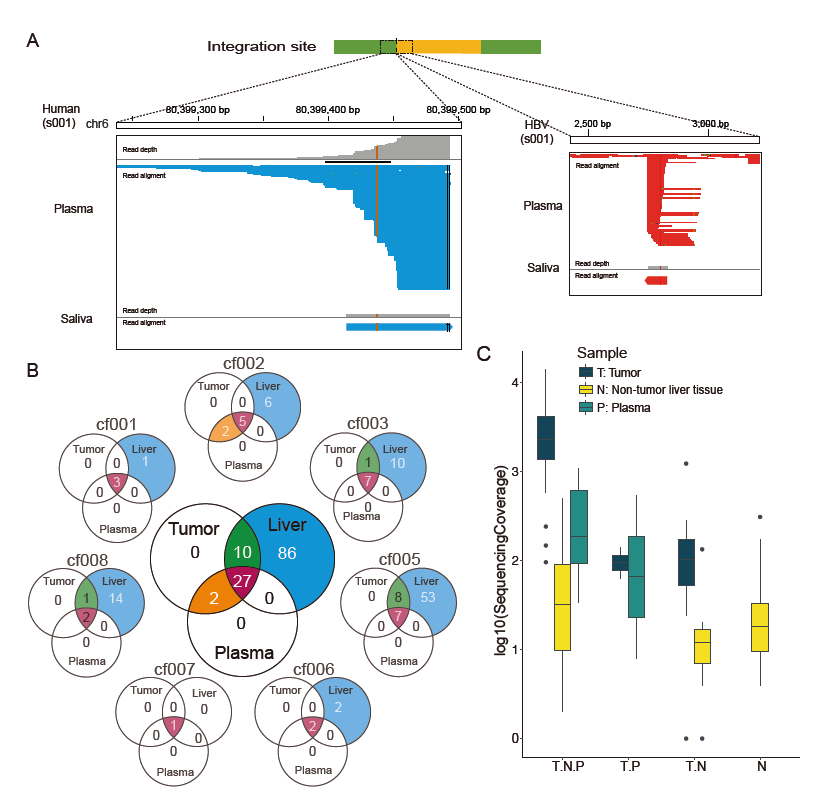
## Figure legends



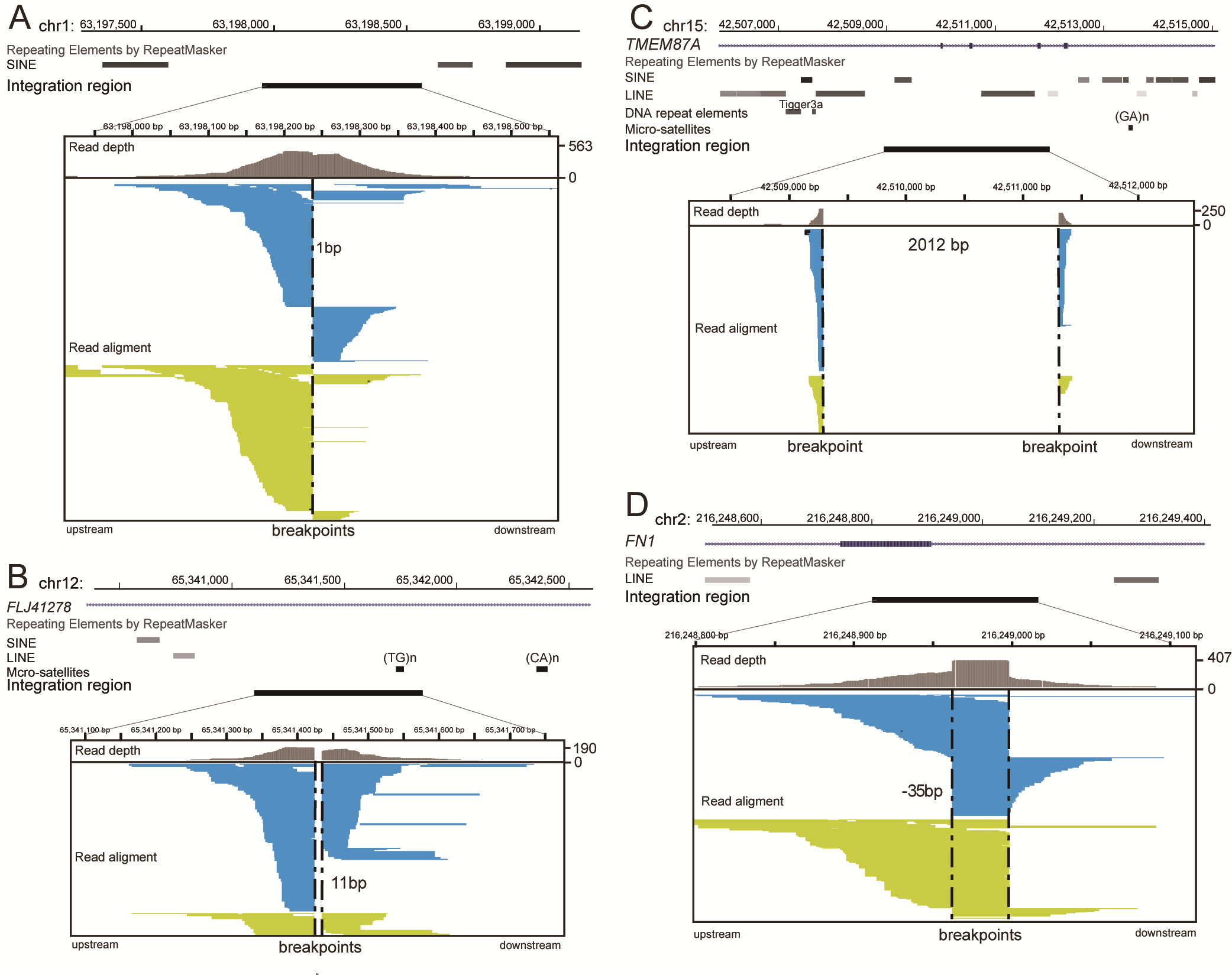
**Figure 1. Study design. A.** Sample collection in five stages. HCC, hepatocellular carcinoma; BDC, bile duct carcinoma; negative, 3 patients with liver cancer but without HBV infection. **B.** DNA amounts for sequencing library construction in tissue and plasma (Left). Integrations lead to two host 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 breakpoints 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.

****

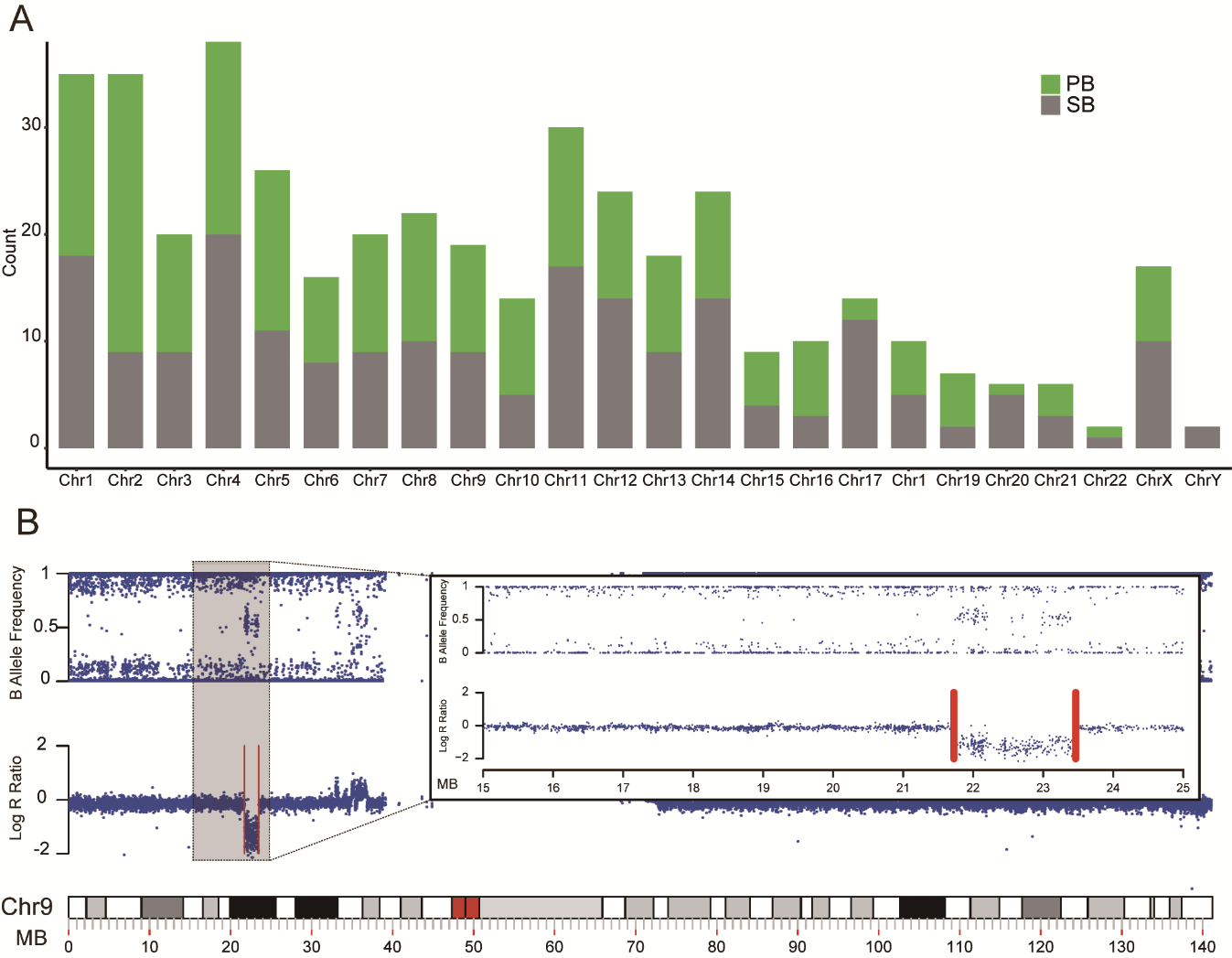
**Figure 2. (A) All integration events connecting the viral and human genomes.** Each light blue line indicates one integration event, with one end showing the breakpoint in the HBV genome and the other in human chromosomes. Each integration event was only observed in one patient. Bubbles with diverse 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.



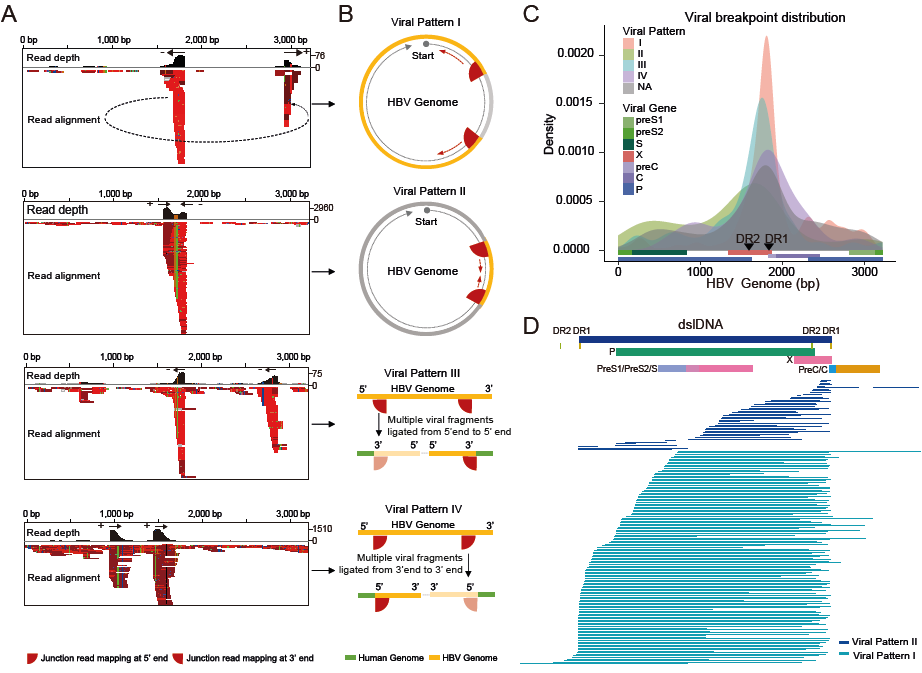
**Figure 3. Detection of integration events in saliva and plasma obtained from HCC patients. (A)** The integration event in saliva. Only one non-redundant junction read is obtained in comparison to high sequencing coverage in paired plasma sample.(B)The cfDNA was successfully extracted in seven of eight 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.