# Long read sequencing reveals sequential complex rearrangements driven by Hepatitis B virus integration

Songbo Wang[1,2], Jiadong Lin[1,2], Xiaofei Yang[2,3], Zihang Li[4], Xu Tun[1,2], Tingjie Wang[1,2], Bo Wang[5], Liangshuo Hu[5], Kai Ye[1,2,4,6,7*]

[1]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

[3]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[4]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[5]The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China.

[6]Faculty of Science, Leiden University, Leiden, The Netherlands

[7]Genome Institute, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]Correspondence to Kai Ye, E-mail: kaiye@xjtu.edu.cn

**Integration of Hepatitis B (HBV) virus into human genome disrupts genetic structures and cellular functions. Here, we conducted multiplatform long read sequencing on two cell lines and five clinical samples of HBV-induced hepatocellular carcinomas (HCC). We resolved two types of complex viral integration induced genome rearrangements and established a Time-phased Integration and Rearrangement Model (TIRM) to depict their formation progress by differentiating inserted HBV copies with HiFi long reads. We showed that the two complex types were initialized from focal replacement and the fragile virus-human junctions triggered subsequent rearrangement. We further revealed that these rearrangements promoted a prevalent loss-of-heterozygosity at chr4q, accounting for 19.5% of HCC samples in ICGC cohort and contributing to immune and metabolic dysfunction. Overall, our long read based analysis reveals a novel sequential rearrangement progress driven by HBV integration, hinting the structural and functional implications on human genomes.**

Persistent hepatitis B virus (HBV) infection is reported to be responsible for >50% of all hepatocellular carcinoma (HCC) worldwide[1,2]. Integration of HBV fragments into human genome, which may activate host genes and cause genome instability, happens in more than half of liver cancers[3] and is considered as a causal mechanism that lead to carcinogenesis[4].

Several studies utilized short read sequencing (SRS) to survey the genomic and oncogenic preference of HBV integrations sites[5-7]. However, due to the short read length (typically 100-200bp), SRS detects single human-virus junction but fails to reveal the full-picture patterns. With the advantages of long read sequencing (LRS) such as CLR and ONT, the genomic complexity caused by HBV integration has been classified into either intra- or inter-chromosome alterations. These alterations can directly affect cancer driver genes such as TERT and MYC[1,8] or induce focal gain, deletion and dicentric chromosome by translocations[1,9,10]. Nevertheless, these studies have shown the association between viral integration and human genome structural rearrangements, we still lack a full picture of how virus disrupt the genome due to small sample number and the high error-prone CLR or ONT, which often lead to detection artifacts from studies with low-coverage data (less than 20X)[9,10]. Besides, long and accurate HiFi long read sequencing would be ideal to characterize the complex structure of inserted HBV DNAs[10] and their underlying mechanisms, for better diagnosis and prognosis.

In this study, we applied multi-platform deep LRS on HBV-infected HCC samples, focused on identifying comprehensive and novel HBV integration patterns as well as their underlying mechanisms and impacts on human genome. In total 7 samples, consisting of 2 cell lines (SNU-182 and SNU-387) and 5 clinical samples, were sequenced with the most advanced LRS technologies such as CLR, ONT and CCS **(Supplementary Table1).** Collectively, we resolved14 full-view HBV integration events **(Methods and Supplementary Fig.1)**, which were classified into three types (**Figure 1a)**. Type I was characterized as viral insertion into human genome through focal replacement. Type II involved complex and integration-induced unbalanced translocations while Type III referred to the most complex event, chromothripsis. Specifically, the formation mechanisms of two complex types were introduced and we proposed a Time-phased Integration and Rearrangement Model (TIRM for short) to describe their formation progress from Type I (**Figure 1b)**. Finally, we utilized HCC WGS and RNA-seq data from ICGC cohort to explore the prevalence of our discoveries and their implications on immune and metabolic functions.

In Type I integration mode (n=2), these integrations focally disrupted human genomes by replacing short human sequence with two HBV fragments, composing a nearly complete HBV copy. For instance, two HBV fragments replaced about 70bp of intergenic region (chr22:40,044,376-40,044,446) in cell line SNU-387 and therefore, we could observe two human-virus junctions in one single LRS read simultaneously **(Supplementary Fig.2a)**. The breakpoints on HBV genome accorded with previously reported hotspots around 1800nt and 0nt[6]. Based on the above features observed in LRS data, we deduced 3 potential Type I events **(Supplementary Fig.2b)** from published single human-virus junctions in ICGC HCC dataset[3]. All 5 replacements located at intergenic or intron regions and were different from promoter- or exon-preferred integrations at TERT or KMT2B[1], which caused significant overexpression or altered transcript. Therefore, these Type I replacement events might not directly contribute carcinogenesis but led to instable genome, especially fragile at the human-virus junctions. As a result, the initial step of our TIRM was accomplished and instable genome promoted subsequent rearrangements.

The Type II complex events (n=6) caused large scale copy number variations (CNVs) through unbalanced-translocations. As a result, integrated HBV fragments fused two chromosomes together. Both two virus-human junctions linking two chromosomes were right at the boundary of significant CNAs, with one gain of heterozygosity (GOH) and the other on loss of heterozygosity

(LOH). We also identified a recurrent structure of inserted HBV fragments, in which two distinct copies of HBV fragments harboring different single-nucleotide-variants (SNVs) were inserted between two human chromosomes with different orientations. For instance, in Event3 from SNU-182, chromosome 4 and 8 were linked together by four HBV fragments, two of which headed forward while the other two headed reverse. Both forward and reversed fragments could form a nearly complete HBV genome **(Figure 2a)**. Meanwhile, HiFi long reads facilitated the identification of SNVs, differentiating forward and reversed fragments**,** indicating that the different-oriented fragments were originated from two independent integration events during long-term and persistent infection.

Next, we built a phylogeny tree with all HBV fragments in SNU-182 **(Methods and Figure 2b)**. We noticed that HBV fragments from different integration events located at different branches and nodes in the tree, indicating multiple different integration events in one sample occurred at different time points. Moreover, different-oriented fragments connecting two chromosomes (e.g. chromosome 4 and 8 in Event3) even located on different branches, confirming our hypothesis that current Event3 were initialized by two independent HBV integrations events and the inter-translocation happened as a subsequent rearrangement.

Given the evidences above, we proposed the Time-phased Integration and Rearrangement Model (TIRM for short, need rename) to illustrate the potential formation mechanism of Type II **(Figure 2c)**. Initially, during the persistent infection, HBV integrated into two different chromosomes at two different time points through two separate DNA double-strand-breaks (DSB) and its subsequent repair pathways, such as microhomology-mediated end joining (MMEJ)[11]. This process would form two individual Type I replacement events. Later the two fragile human-virus junctions broke again and left two exposed DSBs. Finally, under the micro-homology mediated break-induced replication (MMBIR)[12], the replication fork jumped from one DSB to the other to form a new chromosome, resulting in GOH and LOH simultaneously on two different chromosomes. In general, the TIRM well established the process of the initial integration and the following translocation step by step.

The most complex Type III (n=2), integration-induced chromothripsis, was identified from our clinical samples. For instance, in Event 4, we observed chromosome-wide CNAs and 30 genomic rearrangements localized at chr14 and chr16 **(Methods and Figure 3a)**. Those rearrangements frequently linked together copy number neutral, loss or amplified regions, indicating that chr14

and chr16 underwent a chromothripsis event. Moreover, HBV integrated into the boundary of the rightmost amplified region and formed a symmetrical fold-back-inversion structure, which was a strong evidence for breakage-fusion-bridge (BFB) mechanism[13]. We also found another fold-back-inversion structure at the boundary of the highest amplified region **(Supplementary Fig. 4)**. We could infer that the second BFB happened later than the integration-induced BFB by comparing the CNA around the breakpoints because the CN always doubled after a BFB event.

The above evidences also implied TIRM, starting with focal integration and resulting in chromothripsis events **(Fig. 3b)**. Specifically, at the beginning time point, HBV integrated into chr14 by replacement as well. Later, at the integration site, telomeric DSB happened and the unprotected ends was fused by non-homologous end-joining (NHEJ) during DNA replication. Later, the formed dicentric chromosome was cut off during mitosis and in the next cell cycle, the second BFB happened and the chromosome became so unstable that chromothripsis eventually happened. Similar TIRM was also identified in Patient1 on chr7, where the chromothripsis occurred after the HBV-induced BFB **(Supplementary Fig. 5)**.

The whole-genome deep LRS provided us the opportunity to investigate the impactions of the two complex rearrangements on human genome. Notably, we observed a frequent LOH at chr4's long arm (chr4q-LOH for short), presenting in 5 out of 7 in our data (**Supplementary Fig**). To reveal its prevalence in HCC, we collected HCC samples (n=303) across 2 projects (LIRI-JP and LIHC-US) in ICGC cohort **(Supplementary Table2)**. We found ~19.5% HCC samples (n=59) included chr4q-LOH, exceeding that in ovarian cancer (11%). When considering viral conditions, ~12% of the NBNC-HCCs while ~32% of HBV-HCCs showed chr4q-LOH, suggesting HBV as a promotion factor of chr4q-LOH. We also examined the correlation between HBV integration and these LOHs. Although not all these chr4q-LOHs were directly caused by HBV integration (3 out of 5 in our data, 5 out of 59 in ICGC), we considered HBV could promoted chr4q-LOH by integration and subsequent rearrangement.

Next, we examined the hotspots of these chr4q-LOH breakpoints. 16 out of 59 concentrated at chr4q33-34 (Cluster1), leading to LOH at chr4q35. Previous reports found that genes located at chr4q35, such as *IRF2* and *ING2*, suffered frequent LOH and considered them as potential HCC suppressor genes[14]. However, compared with chr4q-neutral samples, two novel genes located at chr4q35, *ACSL1* and *F11*, were observed significantly down-regulated ($P$=0.0012 and 0.0035, respectively) **(Methods and Fig. 4b)**. The roles of *ACSL1* and *F11* have been associated with

immunological processes[15-17] and low expression was observed after tumorigenesis **(Fig. 4b)**. Therefore, down-regulation of the two genes implied that the host immune system might under a disorder, which could also be proved by other 14 down-regulated genes enriched in KEGG item complement and coagulation cascades. **(Figure 4d, Supplementary Table3)**. Since HBV has been reported to inhibit the expression of complement *C3* and *C4* to cause immune dysfunction[18], these expression and pathway results suggested that the immune dysfunction was worsen in the effort of HBV integration and the subsequent LOH at chr4q, which was beneficial to carcinoma progress. In addition to chr4q35-LOH, we discovered another enriched cluster at chr4q12-13 region, leading a larger LOH from chr4q13 to 3' telomere (37 out of 59, Cluster2). Except the two immune-related genes mentioned above, metabolic genes such as *UGT2B* genes[19] and *ADH1B*[20] that located after chr4q13 were significantly down-regulated **(Figure 4c)**. As a consequence, 22 metabolism pathways (12 in Cluster1) were enriched **(Supplementary Fig 6b)**. Since metabolic alteration clearly characterized HCC[21], HBV-integration worsen the aberrant metabolism in the same way as that in immune dysfunction.

Functional analysis among Cluster1 and Cluster2 suggested that HBV integration could indirectly disturb genes and affect immune and metabolic pathways through subsequent rearrangements. Although, it has been known that HBV integration was a random event[22], integrating into specific regions, such as chr4q12-13 and chr4q33-34, would promote genome rearrangements and accelerate tumorigenesis. In general, the initial integration might not be deadly (not in exons or promoters), but the subsequent rearrangement described in TIRM could alter gene copies and expressions, granting cells with selective advantages to HCC.

In this study, utilizing deep LRS on HBV-HCC samples, we proposed the Time-phased Integration and Rearrangement Model (TIRM) to demonstrate the formation progress of the two types of integration-induced sequential complex rearrangements. DNA repair mechanisms such as MMEJ, MMBIR and BFB, were involved during the progress. Moreover, we showed two widespread clusters of LOH at chr4q, promoted by HBV integration and resulting in aberrant immune and metabolism.

Although we tried to conclude as many integration types as TIRM could, there were still exceptions (n=4). For instance, we found an integration-induced tandem duplication. We assumed that starting from focal replacement, the tandem duplication was later formed through MMBIR **(Supplement Fig. 7a)**. We also found there was a special unbalanced translocation, where three enhancer-

enriched segments from chr3, chr5 and chr12 were fused with 2 integrated HBV fragments including open reading frames of HBx and Core **(Supplement Fig. 7b)**. We assumed it as a novel HBV-host DNA contact by hijacking human enhancers to form super enhancers. We expected to carry out following studies to decipher their formation progress.

HBV integrations have been discussed and explored with SRS and LRS in the past decade. We demonstrate that LRS especially the HiFi LRS can comprehensively characterize the complex integration patterns, mechanisms and impacts. Our work lays down an important foundation for further large-scale HBV integration studies with LRS, making up the disadvantages of SRS and investigating the full spectrum of HBV integrations.

# MAIN FIGURES

## Figure 1

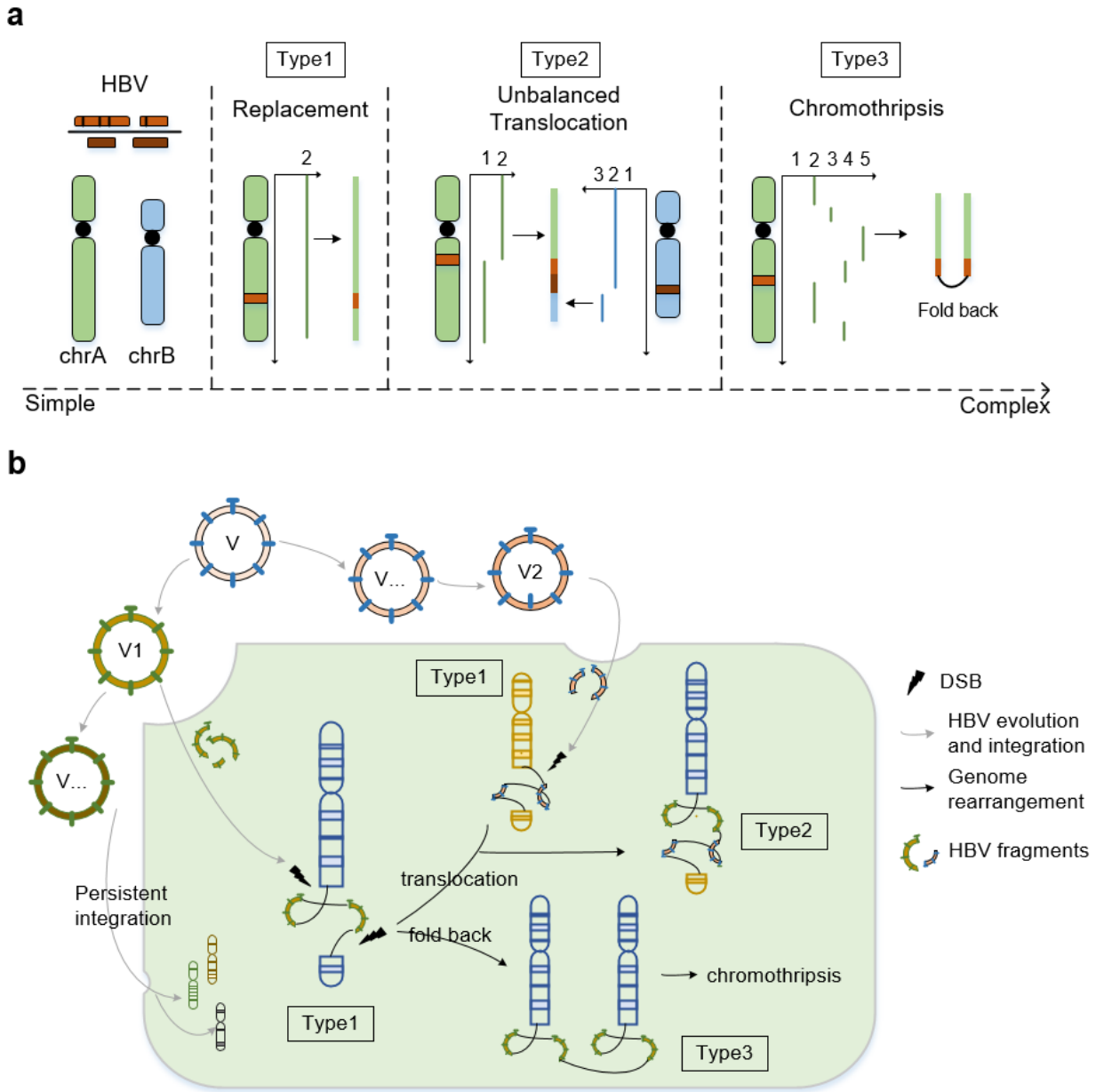


**Fig. 1 | Introduction of novel discoveries. a**, Classification of found integration events. 3 major types of integration have been identified and their corresponding CNAs are shown in the line-plots. The 3 types are ranged from left to right, which accorded with their complexity from simple to complex. The black arrows indicated the formation results. **b**, The Time-phased Integration and Rearrangement Model (TIRM for short). During persistent infection, HBV evolved and might integrated into human genome by replacement at any time point (outside the

box). Later, through a new DSB, the integration site broke again and subsequent rearrangement happened to form the complex Type2 and Type3 (inside the box).

# Figure 2

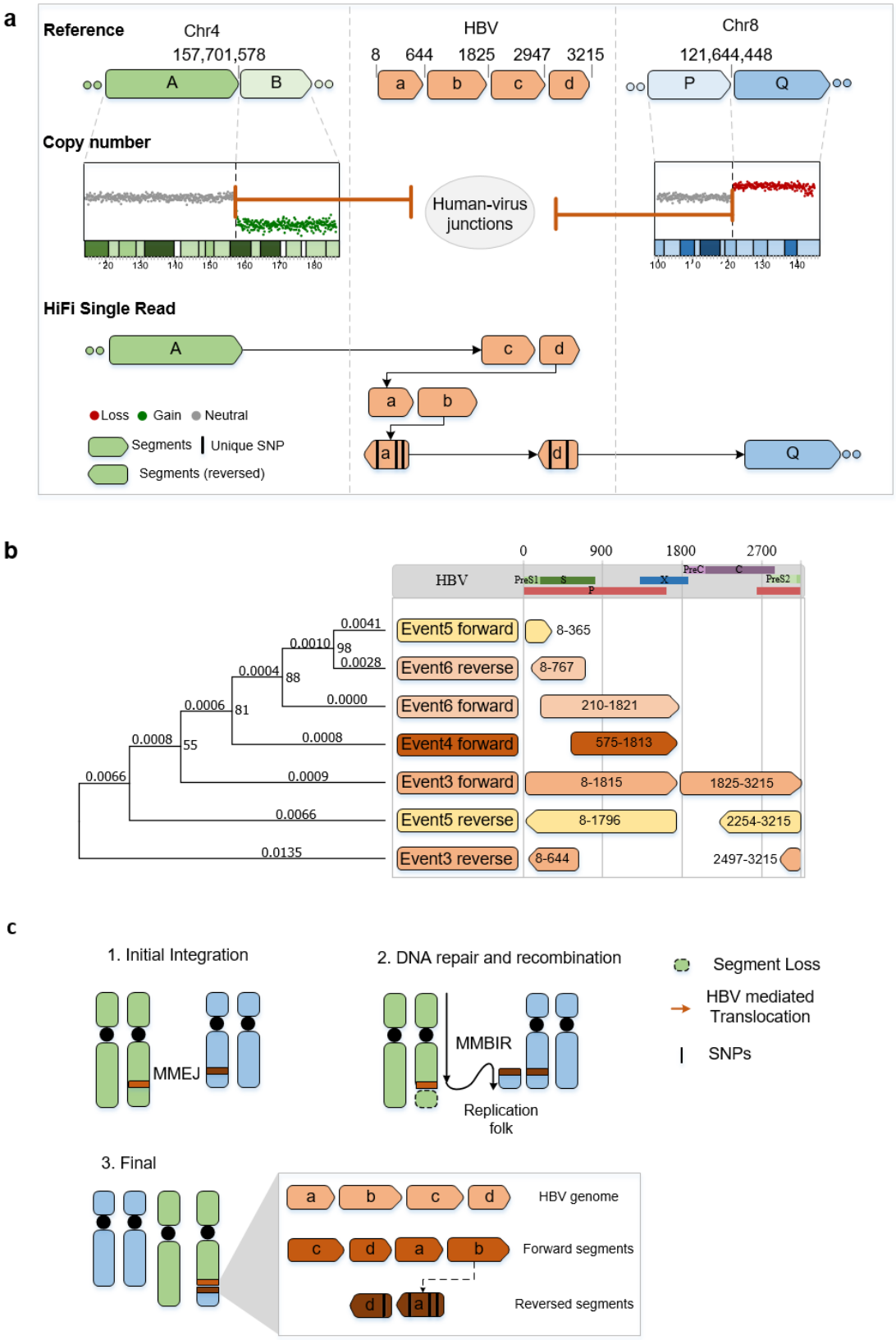
a
Reference
Chr4
157,701,578
A
B
HBV
8
644
1825
2947
3215
a
b
c
d
Chr8
121,644,448
P
Q
Copy number
Human-virus junctions
HiFi Single Read
Loss
Gain
Neutral
Segments
Unique SNP
Segments (reversed)
b
HBV
0
900
1800
2700
PreS1
S
P
X
PreC
C
PreS2
Event5 forward
8-365
Event6 reverse
8-767
Event6 forward
210-1821
Event4 forward
575-1813
Event3 forward
8-1815
1825-3215
Event5 reverse
8-1796
2254-3215
Event3 reverse
8-644
2497-3215
0.0041
0.0010
98
0.0028
0.0004
88
0.0000
0.0006
81
0.0008
0.0008
55
0.0009
0.0066
0.0066
0.0135
c
1. Initial Integration
MMEJ
2. DNA repair and recombination
MMBIR
Replication folk
Segment Loss
HBV mediated Translocation
SNPs
3. Final
HBV genome
Forward segments
Reversed segments

**Fig 2. | HBV-induced unbalanced translocation. a**, Observed evidence from SNU-182-Event3, including CNA and direct single read that covered the entire event. **b**, Phylogeny tree built with all HBV fragments from SNU-182. Fragments from same event but with opposite orientations were marked in same color. Fragments' coordinates were indicated by colored arrows. **c**, Proposed TRIM to explain the formation of unbalanced translocation. First, two independent replacement integrations happened with unique HBV fragments. Then, under the MMBIR mechanism, the replication fork jump from one human-virus junction to another. Finally, the unbalanced translocation was formed. As a result, blue and green chromosomes were linked together by unique HBV fragments. The green chromosome suffered a LOH while the blue chromosome suffered a GOH.

## Figure 3

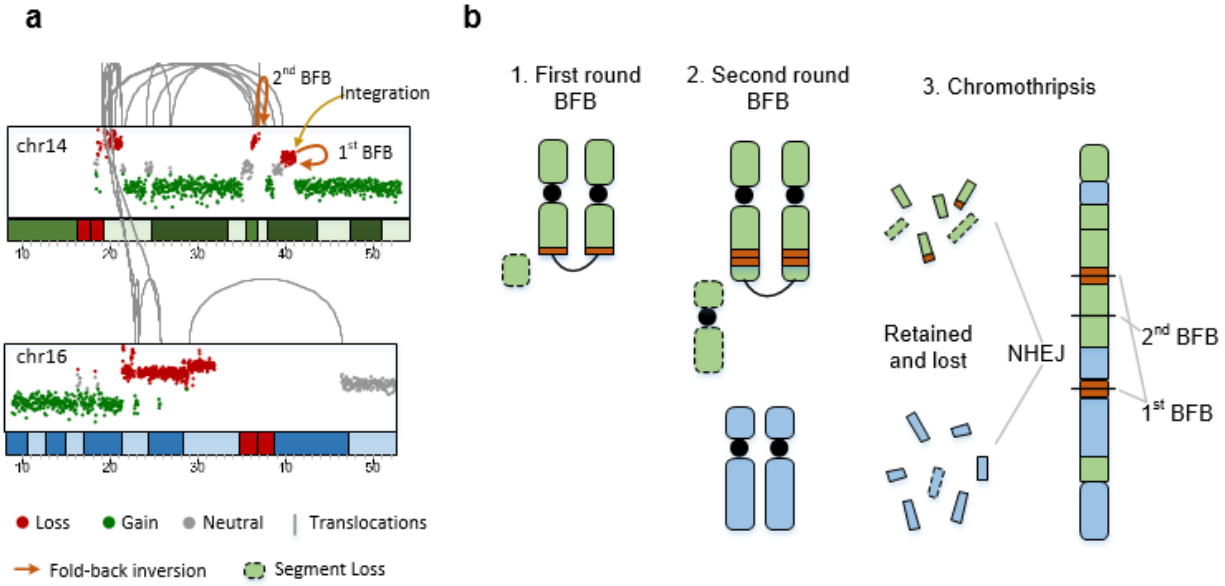


**Fig 3 | HBV-induced chromothripsis**. **a**, CNA and inter/intro-chromosome translocations in Patient2. Two BFBs and HBV integration site were marked as non-gray arrows. **b**, Proposed TIRM to explain the formation of chromothripsis. The two BFBs happened at human-virus junction and human genome, respectively. Then the chromothripsis happened and the retained fragments are re-connected by NHEJ to form the final new chromosome.

# Figure 4

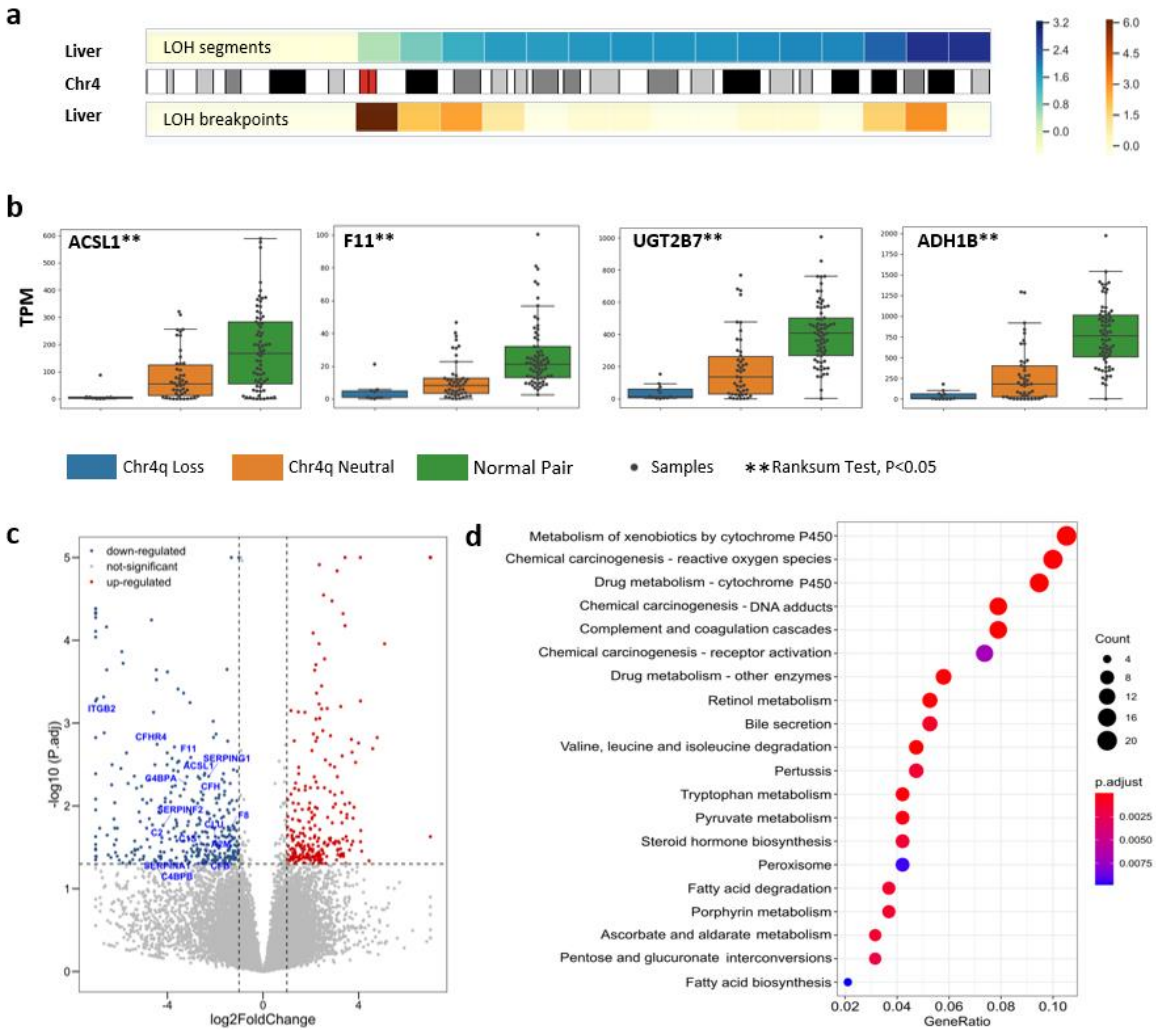


**Fig. 4 | Analysis results from ICGC**. **a**, Enrichment of LOH segments and breakpoints across ICGC HCC cohort. Liver tumors (HCC) had two significantly enriched breakpoints regions. **b**, Influence of LOH on ACSL1, F11, UGT2B7 and ADH1B. **c**, Volcano plot for differential gene expression in Cluster1. Immune-related genes were labeled. **d**, Scatter plot of KEGG pathway enrichment statistics. Pathways terms were ranged by gene ratio and top 20 were displayed in the figure.

# Methods

## Cell lines

SNU-182 and SNU-387 were acquired from the National Collection of Authenticated Cell Cultures. The SNU-387 were sequenced on the PacBio Sequel machine, producing the CLR reads while SNU-182 were sequenced on the PacBio Sequel II sequencer, generating HiFi reads. Briefly, libraries for single molecule real-time (SMRT) PacBio genome sequencing were constructed following the standard protocols. The high molecule genomic DNA was sheared to ~20Kbp targeted size, followed by damage repair and end repair, blunt-end adaptor ligation and size selection. The average CLR sequencing coverage was 45 reads per bp while that forThe average HiFi sequencing coverage was 24 reads per bp.

Additionally, the RNA sequencing of SNU-182 and SNU-387 were produced. A total amount of 1 g RNA per sample was used for library preparation, and the library was generated using NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, USA), following the manufacturer's recommendations. Then, the prepared library was sequenced on the Illumina Novaseq platform, generating the 150bp paired-end reads.

## Clinical samples

Five primary HCC samples were collected from the First Affiliated Hospital of Xi'an Jiaotong University. Five collected tumor tissues were sequenced on the Nanopore PremethION sequencer according to the manufactures' instructions. In general, the long DNA fragments were qualified and selected using the PippinHT system (Sage Science, USA), and the A-ligation reaction were conducted with NEBNext Ultra II End Repair/dA-tailing Kit. The adapter in the SQK-LSK109 (Oxford Nanopore Technologies, UK) was used for further ligation reaction and DNA library was measured by Qubit 4.0 Fluorometer (Invitrogen, USA). Finally, about 700ng DNA library was constructed for each tumor tissue and sequenced with PremethION. The average ONT sequencing coverage was 45 reads per bp.

## ICGC cohort

Project LIRI-JP and LIHC-US were selected from ICGC cohort (n=303). Filtering out those in black list, we total obtained 303 HCC samples, 118 of which have RNA-seq. The corresponding normal paired data were also collected. We utilized the corresponding profiles of CNA and

transcript read count aggregated by PCAWG working group. The viral condition for each sample was also obtained from previous studies as well[3,23,24].

**Characterization of HBV integration with LRS.** HBV integration pattern characterization was performed by a self-developed pipeline, including: (1) read mapping (2) candidate reads selection; (3) cluster reads to events; and (4) classification of integration events

(1) Read mapping.

Raw reads from the LRS technologies were aligned to GRCh38 with HBV genome via Minimap2[25]. The preset parameter of Minimap2, 'map-pb' and 'map-ont' were activated for PacBio and Oxford Nanopore (ONT) reads, respectively. Besides soft-clip and MD tag were set by –Y and –MD. Parameter –a and --eqx was set to output SAM format with =/X CIGAR operators. Then we used SAMtools[26] to sort and index the output.

(2) Candidate reads selection.

We selected reads that span both human and viral genomes by traversing reads that partially aligned on HBV genome. Minimum map quality was set to 20 to filter imprecise reads. For each candidate read $r_i$, we collect all its primary and supplementary alignments and sorted them by read coordinates. Therefore, $r_i$ can be represented as all its alignments: $r_i =< (\mathrm{chr1}, \mathrm{start1}, \mathrm{end1}, \mathrm{T}), (\mathrm{HBV}, \mathrm{start2}, \mathrm{end2}, \mathrm{F}), \ldots (\mathrm{chr3}, \mathrm{start2}, \mathrm{end2}, \mathrm{T}) >$, where T and F denoted the mapping direction.

(3) Cluster reads to events.

We cluster candidate reads to integration events. Different reads but with similar alignments were merged together based on their coordinates. We set the maximum difference between two alignments to 50bp. When comparing two reads, we considered their orientations and toke the minus strand into account. After merging, integration events with less than 5 supporting reads were filtered out. We totally obtained 17 integration events, 3 of which had centromere alignments and were excluded.

(4) Classification of integration events.

Integration events were classified to 3 types, including focal replacement, unbalanced translocation and chromothripsis. Reads that supporting focal replacement usually included 4 alignments: the middle two aligned to HBV while the left and right segments aligned focally and closely on the same chromosome. The unbalanced translocations featured as at least 3

alignments, two of which aligned to different chromosomes and the left aligned to HBV. The chromothripsis always included fold back inversions formed by HBV fragments. As a result, we could observed a nearly symmetrical arrangement of alignments. Accompanying with the CNA and inter-/intra-translocations, we determined the chromothripsis event.

**Phylogeny tree.** We used HiFi sequenced SNU-182 to build a phylogeny tree. Firstly, we separated apart the aligned HBV fragments by the four events. We also separated the different oriented segments in one event. Because HBV always fragmented and some sequences might lost, we then fixed the gap region with reference genome sequences. As a result, we obtained 7 individual HBV mutated genomes. We used MEGA-X[27] (GUI mode) to build a neighbor-joining tree using maximum composite likelihood method. Both transitions and transversions of substitutions were included. The test of phylogeny ran bootstrap method 500 replications.

**Copy number variation analysis.** For LRS, genome-wide profile of CNA was performed by CNVkit[28] without paired control samples. We ran CNVkit under the default settings except the bin size was set to 100Kbp, and copy number plots were created with the scatter option. In addition, we applied the IGVtools[29] to all chromosomes at 100Kbp resolution, generating the binned whole genome coverage. The binned coverage was saved as TDF format for further visualization. For SRS (ICGC samples), CNV profiles of all samples were obtained from PWAWG working group. The published results included total CN, major CN and minor CN of each chromosome segment. We further implemented a simple script to visualize the results for better understanding.
LOH or GOH in ICGC samples were determined by relative change between two neighboring CN segments. CN segments with length less that 10Kbp were excluded to avoid sequence coverage bias. To avoid aneuploidies which were common in cancer, we strictly required that the CN change must start from neutral CN (CN=2). Therefore, LOH meant CN change from 2 to 1 while GOH meant CN change from 2 to 3. Besides, we required LOH or GOH to extend to the telomere. Samples whose chr4 satisfied the above conditions were kept and the corresponding change breakpoints were kept as well. Chr4-neutral samples' chr4 kept CN=2.

**Chromothripsis detection.** Chromothripsis featured as massive CNVs and inter-/intra-chromosome translocations. CNV breakpoints of our samples were obtained from CNVkit

mentioned above. Translocation detection for chromothripsis followed a similar pipeline as that for HBV integration. We collected all abnormal reads with two alignments aligned in different chromosomes (inter-translocation) or in same chromosome but aligned two distant region (> 1Mb, intra-translocation). After clustering single reads and filtering by minimum supporting reads (default 5), we assigned each translocation breakpoint with CNV breakpoint by imitating the coordinate's differences within 50bp. Those inter-/intra- translocations without CNV were excluded. Finally, we visualize CNV and translocations together utilizing karyoploteR[30].

**Gene expression and pathway enrichment.** ICGC HCC samples were contributed into 4 groups. Group1 and group2 were samples from chr4q-LOH breakpoint Custer1 and Cluter2, respectively. Group3 included all chr4-netural HCC samples. Group4 was the set of all normal-paired samples of group 1, 2 and 3. We performed two rounds of analysis: <group1, group3> and <group2, group3> to reveal the impaction of different chr4q-LOH clusters on genes. Besides, we also performed <group3, group4> as a control to understand the gene expression alteration after tumorigenesis. Gene transcript read count matrix was obtained from PWAWG working group[31]. Differential transcript expression between each two groups was analyzed by DESeq2[32]. Genes with P.adjust < 0.05 were considered as significantly altered. Besides, genes with log2FoldChange > 1 or <-1 were set to 'UP' and 'DOWN', respectively for further pathway enrichment. For genes with more than one transcript, we chose the most significant one and removed the left. The visualization (volcano plot) was done by R package ggpubr (https://rpkgs.datanovia.com/ggpubr).
KEGG and Go pathway enrichments were perform with clusterProfiler[33]. Up- and down-regulated genes were enriched separately. The p and q value cutoff were both set to 0.05. Pathway items that satisfied the p and q value cutoff and had more than 10 genes enriched were kept and visualize by R package enrichplot (https://yulab-smu.top/biomedical-knowledge-mining-book/).

**Acknowledgements**
K.Y. and X.Y. are supported by the National Science Foundation of China (32125009, 32070663, 62172325,), the Key Construction Program of the National '985' Project, the World-Class Universities (Disciplines), and the Characteristic Development Guidance Funds for the Central Universities and the Fundamental Research Funds for the Central Universities.

**Author contributions**

K.Y. designed and supervised research. S.W. and J.L performed HBV integration analysis. S.W., T.X. and T.W performed the gene expression analysis. B.W. and L.H. performed clinical surgeries and offered the clinical HCC samples. Z.L. performed Sanger analysis of HBV-human junctions. K.Y., S.W., J.L. and X.Y. wrote the paper. All authors read and proved the final version of this manuscript.

**Competing interesting**

The authors declare no competing interests.

## REFERENCES


1. Peneau, C. *et al.* Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. *Gut* (2021).
2. Guthle, M. & Dollinger, M.M. Epidemiology and risk factors of hepatocellular carcinoma. *Radiologe* **54**, 654-+ (2014).
3. Zapatka, M. *et al.* The landscape of viral associations in human cancers. *Nature Genetics* **52**, 320-+ (2020).
4. Tarocchi, M., Polvani, S., Marroncini, G. & Galli, A. Molecular mechanism of hepatitis B virus-induced hepatocarcinogenesis. *World Journal of Gastroenterology* **20**, 11630-11640 (2014).
5. Sung, W.K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nature Genetics* **44**, 765-U188 (2012).
6. Zhao, L.H. *et al.* Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nature Communications* **7**(2016).
7. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res* **22**, 593-601 (2012).
8. Li, C.L. *et al.* Androgen Receptor Enhances Hepatic Telomerase Reverse Transcriptase Gene Transcription After Hepatitis B Virus Integration or Point Mutation in Promoter Region. *Hepatology* **69**, 498-512 (2019).
9. Zhuo, Z. *et al.* Long-read sequencing reveals the structural complexity of genomic integration of HBV DNA in hepatocellular carcinoma. *NPJ Genom Med* **6**, 84 (2021).
10. Álvarez, E.G. *et al.* Aberrant integration of Hepatitis B virus DNA promotes major restructuring of human hepatocellular carcinoma genome architecture. *Nature Communications* **12**(2021).
11. Tu, T., Budzinska, M.A., Shackel, N.A. & Urban, S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. *Viruses* **9**(2017).
12. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**, 849-53 (2009).
13. Li, Y.L. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98-+ (2014).
14. Zhang, H. *et al.* Analysis of loss of heterozygosity on chromosome 4q in hepatocellular carcinoma using high-throughput SNP array. *Oncol Rep* **23**, 445-55 (2010).
15. Kalugotla, G. *et al.* Frontline Science: Acyl-CoA synthetase 1 exacerbates lipotoxic inflammasome activation in primary macrophages. *Journal of Leukocyte Biology* **106**, 803-814 (2019).
16. Roelands, J. *et al.* Long-Chain Acyl-CoA Synthetase 1 Role in Sepsis and Immunity: Perspectives From a Parallel Review of Public Transcriptome Datasets and of the Literature. *Frontiers in Immunology* **10**(2019).

402 17. Deng, Y.B. *et al.* Long non-coding RNA F11-AS1 inhibits HBV-related
403 hepatocellular carcinoma progression by regulating NR1I3 via binding to
404 microRNA-211-5p. *Journal of Cellular and Molecular Medicine* **24**, 1848-1865
405 (2020).
406 18. Zhu, C. *et al.* Hepatitis B virus inhibits the expression of complement C3 and C4, in
407 vitro and in vivo. *Oncol Lett* **15**, 7459-7463 (2018).
408 19. Lu, L.L. *et al.* Drug-Metabolizing Activity, Protein and Gene Expression of UDP-
409 Glucuronosyltransferases Are Significantly Altered in Hepatocellular Carcinoma
410 Patients. *Plos One* **10**(2015).
411 20. Shih, S., Huang, Y.T. & Yang, H.I. A multiple mediator analysis approach to quantify
412 the effects of the ADH1B and ALDH2 genes on hepatocellular carcinoma risk.
413 *Genetic Epidemiology* **42**, 394-404 (2018).
414 21. De Matteis, S. *et al.* Aberrant Metabolism in Hepatocellular Carcinoma Provides
415 Diagnostic and Therapeutic Opportunities. *Oxid Med Cell Longev* **2018**, 7512159
416 (2018).
417 22. Budzinska, M.A., Shackel, N.A., Urban, S. & Tu, T. Cellular Genomic Sites of Hepatitis
418 B Virus DNA Integration. *Genes* **9**(2018).
419 23. Lian, Q.Y. *et al.* HCCDB: A Database of Hepatocellular Carcinoma Expression Atlas.
420 *Genomics Proteomics & Bioinformatics* **16**, 269-275 (2018).
421 24. Furuta, M. *et al.* Characterization of HBV integration patterns and timing in liver
422 cancer and HBV-infected livers. *Oncotarget* **9**, 25075-25088 (2018).
423 25. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
424 3094-3100 (2018).
425 26. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
426 27. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular
427 Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and*
428 *Evolution* **35**, 1547-1549 (2018).
429 28. Talevich, E., Shain, A.H., Botton, T. & Bastian, B.C. CNVkit: Genome-Wide Copy
430 Number Detection and Visualization from Targeted DNA Sequencing. *Plos*
431 *Computational Biology* **12**(2016).
432 29. Robinson, J.T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26
433 (2011).
434 30. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable
435 genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).
436 31. Campbell, P.J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82-+
437 (2020).
438 32. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and
439 dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**(2014).

33. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* **2**, 100141 (2021).