Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Multisite long-read sequencing reveals the early contribution of somatic structural variations to HBV-related hepatocellular carcinoma tumorigenesis

Tianfu Zeng, Haotian Liao, Lin Xia, et al.

Genome Res. published online March 4, 2025
Access the most recent version at doi:10.1101/gr.279617.124

P<P Published online March 4, 2025 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Multisite Long-Read Sequencing Reveals the Early Contribution of Somatic Structural Variations to HBV-related Hepatocellular Carcinoma Tumorigenesis

Tianfu Zeng[1#], Haotian Liao[2#], Lin Xia[1#], Siyao You[1], Yanqun Huang[1], Jiaxun Zhang[1], Yahui Liu[1], Xuyan Liu[1], Dan Xie[1]*

1. Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China

2. Division of Liver Surgery, Department of General Surgery and Laboratory of Liver Surgery, and State Key Laboratory of Biotherapy and Collaborative Innovation Center of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China.

# These authors equally contributed to this work.

* Corresponding author

Dan Xie

Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China.

Email: danxie@scu.edu.cn

Running title

Somatic Variations in HCC by Long-Read Sequencing

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Somatic structural variations (SVs) represent a critical category of genomic mutations in hepatocellular carcinoma (HCC). However, the accurate identification of somatic SVs using short-read high-throughput sequencing (HTS) is challenging. Here, we applied long-read nanopore sequencing and multisite sampling in a cohort of 42 samples from five patients. We discovered a prominent presence of somatic SVs in adjacent nontumor tissues, which significantly differed from somatic single nucleotide variants (SNVs) and copy number variations (CNVs). The types of SVs were markedly different between adjacent nontumor and tumor tissues, with somatic insertions (INSs) and deletions (DELs) serving as early genomic alterations associated with HCC. Notably, hepatitis B virus (HBV) DNA integration frequently resulted in the generation of somatic SVs, particularly inducing inter-chromosomal translocations. While HBV DNA integration into the liver genome occurs randomly, multisite shared HBV-induced SVs are implicated as early driving events in the pathogenesis of HCC. Long-read RNA sequencing revealed that some HBV-induced SVs impact cancer-associated genes, with translocations being capable of inducing the formation of fusion genes. These findings enhance our understanding of somatic SVs in HCC and their role in early tumorigenesis.

Introduction

As the most prevalent form of primary liver cancer, hepatocellular carcinoma (HCC), which mostly forms after years of chronic liver disease, against a background of severe liver scarring and typically cirrhosis, has been ranked as one of the leading causes of cancer-related death worldwide (Villanueva 2019; Muller et al. 2020). More than half of HCC cases globally occur in China, where chronic hepatitis B virus (HBV) infection is the major etiological factor, accounting for over 60% of cases (Llovet et al. 2021; Bray et al. 2024; Chen et al. 2024a). Somatic DNA alterations are known as pivotal drivers of HCC tumorigenesis and progression. Mutations in the

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

TERT promoter are the most prevalent (~60%) genetic alterations in HCC(Schulze et al. 2016), and the TERT promoter is a recurrent insertion site of HBV DNA. Other recurrently mutated genes primarily harbor somatic single nucleotide variants (SNVs) within their coding regions, including TP53 (~30%), CTNNB1 (~30%), and ARID1A (~10%), which affect the cell cycle, WNT signaling, or chromatin remodeling (Khemlina et al. 2017). Nevertheless, these most prevalent mutations remain undruggable at present (Zucman-Rossi et al. 2015), and there are no mutations available in clinical practice to predict therapeutic response, highlighting the incomplete understanding of the mutational landscape of HCC.

Structural variants (SVs) are large genomic alterations (>50 bp), distinct from small variants like SNVs and short insertions and deletions (indels), as they often arise from different mechanisms (Abyzov et al. 2015). SVs are generally defined as insertions (INS), deletions (DEL), duplications (DUP), inversions (INV), and translocations (TRA), and are known to play an important role in cancer pathogenesis(George et al. 2015; Waddell et al. 2015). Over the past decade, utilizing short-read high-throughput sequencing (HTS) technologies, studies have revealed that somatic SVs can drive malignant phenotypes by altering the expression or function of oncogenes (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Cosenza et al. 2022). Additionally, certain somatic SVs within tumors have been used to make therapeutic decisions, such as the structural rearrangements involving the ALK gene in non-small cell lung cancer and the translocation leading to BCR-ABL1 fusion in chronic myeloid leukemia (Soda et al. 2007; Hochhaus et al. 2017a; Hochhaus et al. 2017b). However, limited by the short read lengths of HTS, the current understanding of somatic SVs in HCC is still incomplete.

Long-read sequencing (LRS) technology, represented by Oxford Nanopore Technologies (ONT) and Pacific Bioscience (PacBio), can produce continuous reads longer than 10,000 base pairs. These reads can completely span across genomic repetitive segments and complex genomic variant regions. Thus, long-read sequencing is progressively being applied to studies on genomic

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

SVs in tumors. For instance, utilizing nanopore sequencing, a complex SV named “cancerous local copy-number lesions” (CLCLs) has been identified within lung cancer, and the precise junctions of these complex SVs are difficult to identify using short-read sequencing (Sakamoto et al. 2020). While previous studies have revealed distinct mechanisms in the generation of germline and somatic SVs in liver cancer by analyzing complete breakpoint sequences derived from long reads, the genome-wide landscape of somatic SV characteristics within HCC has yet to be fully elucidated (Fujimoto et al. 2021). A recent study revealed notable occurrences of somatic SVs within cirrhotic tissues (Brunner et al. 2019). Given that cirrhosis from any etiology is the strongest risk factor for HCC, this further underscores the importance of deciphering somatic SVs in elucidating the pathogenesis of HCC.

Results

Detection of somatic SVs using long-read sequencing.

We performed multisite sampling of tumor tissues and adjacent nontumor tissues from five patients diagnosed with HBV-positive HCC. The nontumor tissues were categorized as proximal nontumor (located within proximal 1 cm of the tumor) and distal nontumor (>1 cm distal to the tumor), with both exhibiting fibrosis. Blood samples collected before surgical resection were used as matched normal controls. In total, twenty-seven tumor tissue samples, 10 matched nontumor liver samples (comprising 6 proximal and 4 distal), and 5 matched blood samples were obtained from the 5 patients (Fig. 1A). Detailed clinical information of included patients is presented in Supplemental Table S1.

We generated long-read whole-genome sequencing data from all samples using the PromethION (Oxford Nanopore Technologies) platform. To obtain high quality sequencing data, we excluded reads with lengths less than 1 kilobase (kb) and mean base quality lower than 7. The mean N50 of long reads was 22.41 kb (19.22 kb ~ 31.02 kb) after quality control (Supplemental

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Fig. S1A, Supplemental Table S2). The average sequencing depth reached 36.72X× (30.50× ~ 44.22×) for tumor and nontumor tissues, 22.19× (19.84× ~ 25.34×) for blood samples (Supplemental Fig. S1B). Considering the relatively high single-base error rate of long-read sequencing, short-read whole-genome sequencing using the Illumina NGS platform was additionally conducted on all samples to facilitate the detection of somatic SNVs.

Utilizing long-read sequencing, we developed a rigorous in-house bioinformatics pipeline to identify somatic SVs and HBV DNA integration events, facilitating the characterization of SV types associated with HBV integration (Supplemental Fig. S1C). Clean long reads were aligned against a custom reference genome consisting of the human genome and 21 HBV genomes using minimap2. Subsequently, the alignment results from paired samples were merged. The widely used variant caller Sniffles2 was used to cluster the long reads supporting the same SV. In instances where supporting reads for SVs span both the HBV and human genomes in alignments, SV types were re-inferred based on the alignment patterns of human sequences flanking the breakpoints. After conducting quality control and manual inspection, we identified an average of 253 somatic SVs across all samples (range from 122 to 405, a median of 252, Fig. 1B). This count notably exceeds previous reports based on short-read sequencing, where only approximately 20% of HCC cases displayed somatic SV counts greater than 200 (Fujimoto et al. 2016; Li et al. 2020b). These results underscore the superior sensitivity of long-read sequencing in the detection of SVs (Aganezov et al. 2020; Xu et al. 2023).

The number of somatic SVs rapidly decreases as the SV length increases, consistent with previous studies (Beyter et al. 2021). We identified two distinct peaks, located at approximately 170 bp and 320 bp (Fig. 1C). The predominant types of SVs at these peaks were INS and DEL, accounting for 98.47% and 99.17%, respectively. To further characterize these sequences, we performed sequence annotation using RepeatMasker. The SVs around 170 bp were predominantly composed of satellite elements (Supplemental Fig. S2A), which are involved in the maintenance

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

of chromosomal integrity and have been found to be overexpressed in various cancer types (Ting et al. 2011; Ho et al. 2017). In contrast, the sequences near 320 bp are primarily composed of SINEs (Supplemental Fig. S2B), which are associated with SV mechanisms associated with the mobilization of active transposable elements (Kolomietz et al. 2002; Robberecht et al. 2013). Notably, somatic INSs and DELs were significantly shorter in length compared to somatic duplications (DUPs) and inversions (INVs), with the majority of INVs and DUPs exceeding 10,000 bp (Wilcoxon rank-sum test, p value $< 2.2 \times 10^{-16}$, Fig. 1D).

The mean variant allele frequency (VAF) of somatic SVs across all samples was 0.14. Notably, the VAF of somatic SVs in the adjacent nontumor samples from each patient was significantly lower than that in the tumor samples (Supplemental Fig. S3A), a phenomenon also observed in somatic SNVs in this project (Supplemental Fig. S3B) and prior studies (Huang et al. 2017; Strandgaard et al. 2020). This indicates that the fibrotic liver tissue adjacent to the tumor does not reflect a normal genomic state and has undergone accumulation of low-frequency somatic mutations.

Somatic SVs were prominent in adjacent nontumor tissues.

A recent study revealed that cirrhotic livers exhibit a higher mutation burden compared to normal livers, with the most significant divergence observed in SVs (Brunner et al. 2019). To explore the process of somatic mutation accumulation in HCC, we compared the burden of mutation between adjacent nontumor tissues and tumor tissues. The genome was segmented into 500-bp bins, and the number of bins harboring each mutation type was determined. We observed a significantly lower count of somatic SNVs in the adjacent nontumor liver samples compared to the tumor samples (Wilcoxon rank-sum test, p value $= 5.7 \times 10^{-9}$, Fig. 2A,). Besides, there were also fewer somatic CNV events in the adjacent nontumor liver samples, with 50% of these samples exhibiting no CNVs. However, somatic SVs were prominent in adjacent nontumor tissues, with no significant difference observed when compared to tumor tissues (Fig. 2A). These

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

results indicate that somatic SVs follow a distinct development process compared to somatic CNVs and somatic SNVs during the progression of HCC.

To compare the mutation profiles across different sampling sites, hierarchical clustering of all samples was performed on mutation matrices of 500 bp for both somatic SNVs and SVs across all samples. The clustering analysis of somatic SNVs revealed that tumor samples grouped according to their patient origin, whereas nontumor samples formed a distinct outgroup, separate from the tumor cohort (Fig. 2B). This separation likely arises from the significantly lower number of somatic SNVs in adjacent nontumor tissues compared to tumor tissues (an average of 851 versus 24,234), leading to their clustering based on their mutual dissimilarity to the tumors. In contrast, clustering analysis based on somatic SVs revealed a distinct pattern, with somatic SVs clustering according to patients, and both tumor and nontumor samples clustered together within individual (Fig. 2B). Due to the absence of somatic CNVs in half of the adjacent nontumor samples, we directly compared the mutational profiles of somatic CNVs across different sampling sites within the same patient (Supplemental Fig. S4A-E). Notably, there was a high degree of consistency in CNV changes among various tumor sampling sites. For instance, all 5 tumor sampling sites from HCC8 exhibited amplification of the entire chromosome 2 and 5, despite the absence of CNV alterations in the adjacent nontumor samples (Supplemental Fig. S4A). These findings underscore the similarity of nontumor tissues to tumors in terms of somatic SVs, while highlighting a marked distinction in terms of somatic CNVs and SNVs.

To further examine the similarities in mutations among samples in detail, we performed pairwise comparisons across different sampling sites (Supplemental Fig. S5A-E). The proportion of shared mutations between adjacent nontumor and tumor tissues exhibited significant differences among the categories of somatic SVs, SNVs, and CNVs. Specifically, in both nontumor and tumor samples, the proportion of shared somatic SVs between adjacent nontumor and tumor tissues was significantly higher than that of somatic SNVs (p value = 5.3×10^{-8} and

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

2.2 × 10−16, *t*-test, Supplemental Fig. S6A). Notably, approximately 12.98% of somatic SVs detected in the tumors were also present in the adjacent nontumor tissues, whereas somatic SNVs accounted for only 0.45%. There were no shared somatic CNVs between adjacent nontumor and tumor tissues. This observation aligns with the current understanding of tumor evolution, indicating that the acquisition of CNVs occurs late in clonal expansion (Li et al. 2020a), representing a consequence of the final malignant transformation of genomic stability.

To further understand tumor evolution, we constructed phylogenetic trees for each patient based on the mutation matrix of 500bp windows, and branch evolution patterns correlated with tumor heterogeneity were observed in all patients (Supplemental Fig. S7A). Notable branches shared between tumor and adjacent nontumor tissues were observed on the phylogenetic tree based on somatic SVs, whereas no such pattern was evident in the case of somatic SNVs (Supplemental Fig. S7B). These results suggest that somatic SVs might serve as the primary mutation type in the context of chronic liver disease during the early stages of HCC progression.

HCC was characterized by somatic INS and DEL at an early stage.

The adjacent nontumor tissue in HCC is not entirely normal (Carlessi et al. 2023; Chen et al. 2024b). A transcriptomic study examining HCC alongside seven other tumor types has demonstrated that adjacent nontumor tissues exhibit a unique intermediate state between health and tumor (Aran et al. 2017). To investigate the role of shared SVs between adjacent nontumor and tumor tissues in HCC pathogenesis, we categorized somatic SVs into three distinct categories within individual: shared (present in both adjacent nontumor and tumor tissues), T-specific (exclusive to tumor tissues), and N-specific (exclusive to adjacent nontumor tissues). For each patient, we found that the shared somatic SV categories exclusively comprised DELs and INSs, whereas translocations (TRAs) were unique to the T-specific category (Fig. 3A). Although DUPs and INVs were present in both T-specific and N-specific categories, the relative proportions varied significantly among patients. For instance, HCC9 exhibited a high proportion of DUPs,

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

whereas HCC13 had a higher proportion of INVs. Notably, while HCC9 lacked DUPs in the adjacent nontumor tissue, there was a high consistency among DUPs across different tumor sites, with approximately 68.46% being mutually shared among samples (Supplemental Fig. S8A). In contrast, although INVs were present in the adjacent nontumor tissue of HCC13, no shared INVs were identified between the tumor and adjacent tissues (Supplemental Fig. S8B). The accumulation of INVs in the tumor of HCC13 was localized to Chromosome 9, suggesting that this chromosome may have experienced a catastrophic event, such as chromothripsis (Supplemental Fig. S8C). Moreover, the percentage of DUPs, INVs, and TRAs within the T-specific category exhibited a significantly higher prevalence than those in the N-specific category (chi-square test, p value < 0.01). These specific types of SVs are frequently associated with chromosomal rearrangements in tumor evolution. These results indicate a significant difference in the types of SVs between adjacent nontumor tissues and tumors, suggesting that the tumor genome has undergone more complex genomic alterations.

We further investigated the differences in somatic INSs and DELs between the shared and N/T-specific categories. Notably, the lengths of somatic INSs and DELs shared between adjacent nontumor and tumor samples were significantly greater than those exclusive to either category (Fig. 3B, Wilcoxon rank-sum test, $< 2.2 \times 10^{-16}$). Additionally, the shared somatic INSs and DELs exhibited a higher prevalence of complex repetitive sequences and satellite elements, while non-repetitive and simple repeat sequences were less represented (Supplemental Fig. S9A,B).

Next, we analyzed the landscape of shared somatic SVs between tumor and nontumor samples within individual. Notably, we identified a subset of recurrent somatic SVs, present in two or more patients, affecting a total of 42 genes (Fig. 3C). These include tumor-related genes such as PGA5 and DPP6 (Shen et al. 2020; Choy et al. 2021; Munkhjargal et al. 2023), as well as newly identified candidate genes like *EVA1C*, which is present in four out of five patients. *EVA1C* is a membrane protein-encoding gene that plays a role in inflammatory and immunobiological

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

processes (Hu and Qu 2021). The SV observed on the *EVA1C* gene manifests as a deletion event spanning approximately 160 base pairs in length. Moreover, this deletion is consistently located in close proximity across different samples, with breakpoint variations confined within 10 bp. Short-read sequencing in the samples failed to detect this deletion, primarily due to the tandem-repeated *Alu* sequences present within the corresponding region of the reference genome (Supplemental Fig. S10A,B).

To assess the functional impact of these recurrent somatic SVs, we conducted long-read RNA sequencing on samples HCC10_T7 and HCC10_N1 (Supplemental Table S3). Transcript loss was observed for *EVA1C*, *GSTM1*, and *GSM2* in samples harboring DELs (Supplemental Fig. 11A,B), despite the DEL in *EVA1C* localized exclusively to an intronic region. Additionally, qPCR validation revealed that the expression level of the *EVA1C* gene in samples with DEL was significantly lower than in samples without DEL (Supplemental Fig. S11C). Furthermore, based on gene expression and clinical data obtained from the Cancer Genome Atlas (TCGA) project (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), survival analysis showed that the low expression of *EVA1C* was associated with poor survival of patients (Supplemental Fig. S11D, $P = 0.0055$). These findings indicate that somatic SVs shared between tumors and adjacent nontumor tissues have the potential to induce abnormal expression of genes associated with cancer.

HBV integrations induced the formation of SVs.

Viral DNA integration into the host genome can lead to the malignant transformation of tumors, resulting in an increased burden of mutations near the integration breakpoints(Zapatka et al. 2020). HBV infection is frequently linked to proliferative HCC, which is characterized by chromosomal instability (Villanueva 2019). To further explore the connection between somatic

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

SVs and HBV integrations, we developed a novel pipeline to comprehensively identify HBV integration events using long-read sequencing. We primarily identified two categories of long reads for detecting HBV integration events. First, chimeric reads that mapped simultaneously to the human genome and the HBV genome. Secondly, due to disparities in minimap2 alignment, certain HBV sequences were either mapped as insertions or remained unmapped due to clipping (Supplemental Fig. S12A). Basing on these two kinds of reads, we identified an average of 10 HBV integration breakpoints in the tumor samples (range from 0 to 37, median: 8) and 5 in the adjacent nontumor samples (range from 0 to 15, median: 4), which was significantly more than previous short-read datasets (Sung et al. 2012; Alvarez et al. 2021). Furthermore, 61.38% of HBV DNA integrated into the human genome exceeded 500 bp in length. (Supplemental Fig. S12B). In short-read sequencing, aligning these insert sequences larger than the DNA library fragment size is challenging, and complete insert sequences cannot be detected (Craven et al. 2022; Rajaby et al. 2023).

Within the 335 identified HBV integration breakpoints, 85.67% (287/335) were derived from HBV integration events fully captured by long reads, characterized by HBV integration sequences flanked by human sequences on both sides within the same read. These HBV integration events provided an opportunity for further investigation into the specific patterns of HBV integration into the HCC genome. Based on the alignment positions, orientations and order of human sequences flanking the HBV sequences, we inferred the integration patterns of HBV (Supplemental Fig. S12C-G). We observed that simple insertion of HBV accounted for only 20.20% of events, while the majority of HBV integrations resulting in chromosomal rearrangements within the HCC genome. Among these genomic alterations, translocations were the predominant type, constituting 44.80%, followed by inversions (22.20%), duplications (8.62%), deletions (4.12%), and complex SVs (1.78%). Notably, the complex SVs identified were all duplication-inversion events, originating from five distinct tumor sampling sites in HCC9. These rearrangements were

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

characterized by HBV integration resulting in a fold-back inversion, with human sequences on both sides of the integration site being inverted and duplicated (Fig. 4A). Moreover, a 76.23 Mb copy number amplification (Chr7:124,879,000-48,644,500) was observed 779 bp downstream of the HBV integration site, which is consistent with the breakage-fusion-bridge (BFB) inversion model.

Next, we further explored the relationship between HBV integration breakpoints and CNV boundaries. We found significant differences in the proximity of various HBV-induced SV types to CNV boundaries. INSs were located further from CNV boundaries, with all INS breakpoints situated more than 100 kb away from the CNV borders. In contrast, except for five cases of complex (duplication-inversion) events, TRAs and DUPs were found closer to the CNV boundaries, with 6.4% of TRA breakpoints located within 1 kb of the CNV boundary (Supplemental Fig. S13A). For instance, within the HBV integration event observed in HCC13_T3, approximately 2914 base pairs (bp) of HBV sequence integrated into the HCC genome, giving rise to a translocation event that linked Chr8:31,957,777 and Chr22:50,545,277. Notably, a CNV is located 214 bp downstream of the TRA breakpoint on Chromosome 8. This CNV corresponds to a 5.36 Mb gain on Chromosome 8p12 (Chr8: 31,957,991–37,314,348), and this CNV region harbors several tumor-associated genes, such as *NRG1*, *UNC5D*, and *KCNU1* (Repana et al. 2019). (Fig. 4B). These findings indicate that HBV-induced SVs may play a role in CNV formation and directly affect critical oncogenes involved in tumorigenesis.

In addition, in comparison to a random genomic background, we found a significant enrichment of somatic SVs in proximity to HBV integration sites (two-sided binomial test, p value < 0.01, Supplemental Fig. S13B). The above findings suggested that the HBV integration was closely related to the formation of SVs within the HCC genome.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Multisite shared HBV-induced SVs emerge as distinctive signatures of HCC.

Based on the HBV-induced SVs identified above, we explored their breakpoint location characteristics within the HCC genome. Among all samples, it was observed that 56.2% of breakpoints were shared across multiple samples (Supplemental Fig. S15A). It is noteworthy that these breakpoints were shared among tumor samples, with no shared breakpoints detected in the adjacent nontumor tissues (Supplemental Fig. S15B). Specifically, we observed a pronounced aggregation of HBV-induced SV breakpoints within tumors sampled from diverse regions of the same patient, while HBV-induced SV breakpoints in adjacent nontumor tissues exhibited a stochastic distribution (Fig. 5A). Furthermore, annotation of the HBV genome revealed that the HBV-induced SVs shared across the five samples consistently harbored similar HBV integration sequences, with an average sequence similarity of 96.78% (Fig. 5B). These integration sequences were derived from proximal loci within the HBV genome, with an average breakpoint position difference of 35 bp. (Supplemental Fig. S14A). For instance, all five tumor sampling sites in HCC9 revealed a consistent translocation linked Chr2:74,273,672 and Chr6:70,300,819, with the integrated HBV sequences derived from the HBV X gene (HBV:1544-1481). These findings suggest that HBV DNA integration into the liver occurs randomly, whereas integration at specific loci may lead to hepatocellular carcinogenesis.

Given the PCR-free nature of nanopore sequencing, each read is likely to represent a distinct single-cell origin. The probability of the same HBV DNA integration event occurring independently in multiple cells is minimal. Therefore, the shared HBV-induced SVs observed across different tumor sampling sites may indicate a common clonal origin for these tumor regions. To elucidate the clonal characteristics, we examined the mutational timing of the shared HBV-induced SVs within individual. Variant allele frequency (VAF) analysis revealed a significantly higher VAF of HBV-induced somatic SVs that shared within tumors compared to those appearing independently (Supplemental Fig. S15C, Wilcoxon rank-sum test, p value = 1.8 ×

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

321 10^{-24}). To further assess the cancer cell fraction of these shared HBV-induced SVs, we modified
322 the existing algorithm SVclone (Cmero et al. 2020) to suit long-read sequencing data. By
323 integrating established timing algorithms MutationTimeR (Gerstung et al. 2020), these shared
324 HBV-induced SVs were identified as early clonal events, preceding copy number gains (Fig. 5C).

325 **HBV-induced SVs influence genes associated with tumorigenesis.**

326 Next, we studied the potentially effects of HBV-induced SVs on genes, including those
327 genes directly intersecting with breakpoints or located closest to the breakpoints. Pathway
328 enrichment analysis of the relevant genes revealed their significant association with key HCC
329 pathways, such as Proteoglycans in cancer, PI3K/AKT signaling pathway, and Viral
330 carcinogenesis (Supplemental Fig. S16A, p value < 0.01). In addition, we found that shared HBV-
331 induced SVs within individual impact a total of 24 genes, including genes associated with
332 tumorigenesis, such as *SLC4A5*, *ABC13*, *NRG1*, and *SORCS1* (Fig. 6A). In addition, compared to
333 breakpoints occurring independently, breakpoints of shared HBV-induced SVs were significantly
334 concentrated in gene bodies and their flanking 5 kilobases (kb) regions (Supplemental Fig. S16B,
335 chi-square test, p value = 2.8×10^{-5}). This finding suggests that shared HBV-induced SVs were
336 more likely to have impact on gene regulation and expression. Notably, among these shared
337 HBV-induced SVs, TRAs account for 71.43%, highlighting the evolutionary advantage of this
338 type of SV.

339 We next investigated the functional consequences of HBV-induced TRAs on genes located
340 near the integration breakpoints. Among these genes, *NRG1*, *KLHDC7B*, and *KLHDC7B-DT*,
341 which have been implicated in various cancers (Maillet et al. 2018; Li et al. 2021; Yahiro et al.
342 2021), were directly affected by the translocations. In HCC13, a shared HBV-induced TRA was
343 detected across five sampling sites, with breakpoints located within the intronic region between
344 the first and second exons of NRG1, as well as upstream of *KLHDC7B* (620 bp) and *KLHDC7B-*
345 *DT* (2359 bp). While several previous studies have suggested their relevance to cancer (Martin-

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Pardillos and Cajal 2019; Rosas et al. 2021), mutations within these genes have not been identified in the context of HCC. Thus, we further investigated the impact of the HBV-induced TRA on the expression of genes situated on either side. Long-read RNA sequencing was performed on both the HCC13_T1 tumor sample and its corresponding adjacent nontumor tissue. To obtain high-quality sequences, we conducted local assembly of the TRA using reads supporting this integration events. By aligning RNA reads spanning the integration breakpoints to the assembled translocation sequence, we identified a fusion gene between the KLHDC7B-DT gene and the inserted HBV sequence in the tumor sample. However, no evidence of this fusion gene was detected in the adjacent nontumor sample (Fig. 6B). Subsequent PCR validation confirmed the presence of this fusion gene in the tumor sample and its absence in the adjacent nontumor sample (Fig. 6C). Additionally, quantitative PCR (qPCR) revealed that the expression levels of *KLHDC7B* in samples with HBV integration were significantly reduced, while *KLHDC7B-DT* expression was markedly increased (Fig. 6D,E). These findings suggest that HBV integration may disrupt the promoter of *KLHDC7B*, leading to decreased expression of this gene. Furthermore, qPCR demonstrated that HBV integration led to reduced expression of *NRG1* in the tumor (Supplemental Fig. S16C), potentially attributed to the integration disrupting the structural integrity of the gene. Overall, these findings indicate that HBV-induced TRA can directly impact genes adjacent to integration sites, including the formation of fusion genes and disruption of gene expression.

Discussion

Somatic SVs play a pivotal role in HCC, yet their accurate identification using short-read high-throughput sequencing poses challenges. To overcome limitations related to intratumor heterogeneity (ITH) and refine the estimation of the mutational landscape, we employed multisite sampling in tandem with long-read nanopore sequencing for the comprehensive genome-wide

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

analysis of somatic SVs. We depicted the characteristics of somatic SVs in both adjacent nontumor and tumor tissues, revealing somatic SVs as early genomic alterations that may contribute to the complexity of HCC development. Furthermore, we elucidated that HBV integration serves as a frequent driver of somatic SVs, particularly in the context of inter-chromosomal translocations. Our findings indicated that shared HBV-induced SVs represent early clonal events in the progression of HCC, with the potential to directly impact the transcription of genes relevant to tumorigenesis.

Building on these findings, it is important to note that while our study provides compelling insights into the role of somatic SVs in HCC, some of the results require additional validation to establish their broader significance. Specifically, we observed that the proportion of somatic DUPs, INVs, and TRAs was significantly higher in patients with a history of smoking compared to non-smokers. However, this observation warrants further validation through the inclusion of additional patient samples to improve statistical robustness. Utilizing multiple tumor sampling sites, we identified that HCC13 exhibits the highest number of somatic INVs in tumor sites T1, T2, and T5, which markedly differ from those at T3 and T4. The morphological characteristics of HCC13 indicate that T1, T2, and T5 are spatially distinct from T3 and T4, suggesting that the differing INV patterns may reflect their clonal origins. This highlights the potential of somatic SVs as biomarkers for pathological subtyping, offering insights into the clonal evolution of tumors and their underlying mechanisms. Further investigation into these patterns could enhance our understanding of tumor heterogeneity and its implications for personalized therapeutic strategies.

Recent studies have shown that adjacent nontumor tissues are not healthy and may harbor molecular features associated with tumors (Carlessi et al. 2023; Zhu et al. 2023; Chen et al. 2024b). A recent study has found that the complexity of HCC arises during the progression to chronic liver disease and subsequent malignant transformation (Brunner et al. 2019). However, to

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

our knowledge, the association of genomic mutation features between adjacent nontumor and tumor tissues remains to be further elucidated. In this study, we observed a comparable prevalence of somatic SVs in adjacent nontumor samples and tumor samples, contrasting with the distinct patterns observed for somatic SNVs and CNVs. This could be attributed to the instability of the host genome caused by HBV DNA integration (Zhao et al. 2016). The number of somatic SNVs and somatic CNVs in adjacent nontumor tissues was significantly lower than that observed in tumor tissues, suggesting that these alterations may share similar developmental processes in HCC (Huang et al. 2017). Notably, the differences in somatic CNVs observed between HCC and adjacent nontumor tissues are consistent with findings from a study on urothelial cell carcinoma (Li et al. 2020a), suggesting that this may represent a general feature of tumor evolution.

Furthermore, a considerable proportion of somatic SVs were found to be shared between the adjacent nontumor and tumor tissues. INSs and DELs have emerged as shared SV types between tumor and adjacent nontumor tissues, potentially exerting a direct influence on genes integral to normal hepatic function, including *TCP10L* and *EVA1C*. A possible explanation is that these somatic SVs may arise from pre-malignant intermediate state cells or tumor cells within the adjacent nontumor tissues (Aran et al. 2017; Carlessi et al. 2023). Additionally, a significant proportion of the insertions and deletions consisted of transposons and other repetitive sequences, indicating that these elements may play an important role in tumorigenesis. The challenge of accurately mapping such repetitive sequences using short-read sequencing likely accounts for the increased detection of somatic SVs in our study. Given that these repetitive sequences are often refractory to accurate mapping using short-read sequencing, this could explain the higher number of somatic SVs detected in our study. However, the functional implications of these shared somatic SVs between adjacent nontumor and tumor tissues require further investigation, particularly from the perspectives of transcriptomics and epigenetics.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

The integration of the HBV was considered a major factor driving the development of HCC. Studies had shown that HBV integration could lead to an increased burden of mutations in the genomic region within tens to hundreds of kilobases around the integration sites (Zapatka et al. 2020). However, the specific characteristics of HBV integration patterns were not yet clearly understood. Taking advantage of long-read sequencing, we could detect the complete sequence of HBV integration. Our findings indicated that HBV integration demonstrated a propensity towards generating somatic SVs linked to chromosomal rearrangements, particularly inter-chromosomal translocations, while simple insertions made up only about 20.20% of the occurrences. Notably, shared HBV-induced SVs emerged as distinctive signatures of HCC, exhibiting a clustering tendency proximal to various cancer-associated genes. However, we did not detect HBV integration in the promoter region of TERT, despite previous studies reporting its presence in approximately 30% of HCC cases (Chen et al. 2024a). This discrepancy may be stem from limitations in our sample size or sequencing depth. In comparison to adjacent nontumor tissues, HBV DNA integration in HCC leads to a higher incidence of gene fusion events (Zhuo et al. 2021). Using long-read RNA sequencing, we confirmed the formation of fusion transcripts between HBV and human genes. Previous studies have suggested that HBV DNA methylation could influence the expression patterns of viral genes (Fernandez et al. 2009; Mirabello et al. 2012). However, we did not observe significant differences between adjacent nontumor and tumor tissues in the methylation status of integrated HBV DNA sequences. This highlights the need for more advanced techniques, such as Cas9-targeted nanopore sequencing (Goldsmith et al. 2021), to enhance the detection sensitivity of HBV DNA and facilitate a deeper exploration of its potential epigenetic regulation.

In summary, our comprehensive analysis of somatic SVs in HCC, including those induced by HBV integration, shed light on the early genomic alterations and clonal events contributing to HCC development. The intricate relationship between viral integration and somatic SVs

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

emphasizes the importance of considering these factors in understanding the molecular landscape of HCC. This study extends our understanding of somatic SVs in HCC, providing valuable insights into their characteristics and significance in early tumorigenesis.

Methods

Long-read whole-genome sequencing

Tumor, nontumor, and blood samples from each patient were frozen in liquid nitrogen for DNA isolation. Genomic DNA from each sample was extracted with the MagAttract HMW DNA Kit (QIAGEN). DNA libraries were constructed using the SQK-LSK109 library preparation kit (ONT, UK) following the manufacturer's instructions. Then, the prepared libraries were sequenced on the PromethION sequencer (ONT, UK). Raw sequencing data were basecalled using Guppy 3.2.8 with the default parameters during sequencing.

Short-read whole-genome sequencing

Genomic DNA was isolated from fresh-frozen tumor, nontumor, and blood tissues using the DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer's protocol. The purity and integrity of DNA were assessed via agarose gel electrophoresis. The DNA concentration was measured using Qubit 2.0 (Invitrogen). The DNA was fragmented to approximately 350 bp using the Covaris ultrasonicator. The library was constructed using the established Illumina paired-end protocols. Following library quality control, sequencing was performed using the Illumina NovaSeq 6000 platform (Illumina).

Identification of somatic SVs

First, clean reads were obtained by excluding those with an average base quality < 7 and a length < 1000 bp. Then, minimap2 (v.2.17) (Li 2018) was used to align the clean reads to a custom reference genome (composed of hg38 and 21 HBV genomes (Yan et al. 2015; Zheng et al.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

2021)), with the following parameters: minimap2 -N 10 -p 0.3 -ax map-ont --MD. To calculate somatic SVs, blood samples were used as normal controls, and the BAM files from paired tumor or adjacent nontumor samples were merged with the BAM file from the blood sample after tagging. The merged files were then processed with Sniffles (v.2.0.6) (Smolka et al. 2024) to identify SV breakpoints and discern supporting reads, with the following parameters: Sniffles -i bam_merge -v vcf --tandem-repeats -t 36 --minsupport 1 --mapq 20 --min-alignment-length 500 --minsvlen 50 --output-rnames. All SVs with supporting reads >= 3 and a length >= 50 bp were identified. If all supporting reads originated from tumor or adjacent nontumor samples, and none from the normal control, the SV was categorized as a candidate somatic SV.

Next, we rigorously filtered the candidate somatic SVs. ONT sequencing data from the liver tissues of five healthy individuals (Pascarella et al. 2022), along with data from 63 healthy individuals from the HPRC cohort (Liao et al. 2023), were obtained (Supplemental Table S4). Together with the five blood samples from this study, SVs were identified using Sniffles with the following parameters: Sniffles -i bam_merge -v vcf --tandem-repeats -t 36 --minsupport 3 --mapq 20 --min-alignment-length 500 --minsvlen 50. Additionally, SVs derived from a cohort of 405 individuals from the Chinese population (Wu et al. 2021), also sequenced via nanopore technology, as well as SVs from the HGSVC database (Ebert et al. 2021), were incorporated. These SVs were used to construct a Panel of Normals (PON), which served to filter out false positives in somatic SV calls. Such false positives could arise due to insufficient sequencing depth in paired normal samples, biases associated with ONT technology, or tissue-specific SVs in normal liver tissue. Any candidate somatic SVs present in the PON were excluded. We also removed SVs occurring in unreliable alignment regions, including those where over 50% of reads in the SV region had mapping quality < 20, and SVs with variant allele frequency below 0.01. Furthermore, using genomview (Abeel et al. 2012) to visualize SVs, we manually inspected all candidate somatic SVs. Following the methodology employed in a previous study(Fujimoto et al.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

2021), candidate somatic SVs displaying matching breakpoints in corresponding normal samples were discarded from further analysis.

VCF merging of SVs and identification of shared somatic SVs

The Jasmine (v1.1.5) software was used to merge the VCF files of somatic SVs across different samples, incorporating PON filtering and identification of shared SVs. This tool is specifically designed for SV detection from long-read sequencing, considering factors such as the type, length, breakpoint positions, and sequence context of the SVs. Jasmine compares and merges SV calls across samples by representing variants as points in a multidimensional space and constructing a proximity graph based on their breakpoints and lengths. We employed Jasmine with the following parameters: jasmine file_list out_file --ignore_strand. The default parameters in Jasmine merge SVs of the same type if they are within 100 bp of each other and have length differences of less than 50%. SVs that could be merged are considered as shared SVs.

Detection of HBV integrations and HBV-induced SVs

Although the alignment results of long reads against the customized reference genome already contain partial information concerning HBV integration, notably the presence of chimeric reads aligning to both HBV and human genomes, our examination revealed that certain HBV DNA sequences present on some long reads failed to align to the 21 HBV reference genomes. These HBV sequences either received low alignment scores in minimap2, resulting in being labeled as INS, or were too short and directly clipped without alignment. Based on these characteristics of reads, we developed a workflow for detecting HBV integration using long-read sequencing, mainly consisting of the following steps: (1) extraction of chimeric reads, (2) extraction of INS sequences and clipped sequences from all long reads, (3) re-alignment of INS and clipped sequences to the custom reference genome using BLAST, and (4) determination of HBV DNA integration breakpoints.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Benefiting from the advantages of long-read sequencing, long reads can simultaneously contain HBV DNA sequences and their adjacent human sequences. We utilized these spanning reads to further deduce the structural consequences of HBV DNA integration on the human genome. Based on the alignment direction, position, and order of human sequences flanking HBV DNA, five types of HBV-induced structural variations were identified. Insertion (INS) denotes direct insertion of HBV DNA into the human genome without altering the flanking human sequences. Deletion (DEL) corresponds to the loss of human DNA segments caused by HBV DNA integration. Duplication (DUP) refers to the presence of duplicated human DNA segments flanking the HBV DNA integration breakpoint. Inversion (INV) occurs when the human DNA adjacent to one side of the HBV DNA integration breakpoint is reversed. Translocation (TRA) involves human sequences flanking HBV DNA originating from different chromosomes.

Identification of somatic SNVs and somatic CNVs

The clean reads were aligned to the human reference genome (hg38) using Burrows-Wheeler Aligner (BWA, v.0.7.17)(Li 2013), with the following parameters: BWA-MEM -M -Y -t 38 -R. SAMtools (v.1.9) (Li et al. 2009) was employed for sorting and indexing the resulting BAM files. PCR duplicate reads were marked using the MarkDuplicates module of the Genome Analysis Toolkit (GATK, v.4.1.2.0) (McKenna et al. 2010). Local realignment and base quality recalibration were conducted using the GATK BaseRecalibrator and ApplyBQSR modules.

Somatic SNVs were called from paired tumor and matched blood samples using GATK Mutect2, followed by variant filtering using FilterMutectCalls. Next, annotation of the filtered somatic SNVs was performed via ANNOVAR (Wang et al. 2010). To ensure high-quality somatic SNVs, the following quality control criteria were applied: (1) the variants supported by at least 7% of total reads in tumor sample and less than 2% of total reads in normal sample were retained. (2) the variants with sequencing depths less than 8 in tumor sample or less than 6 in normal sample were excluded. (3) Only variants with more than 3 mutation reads in the tumor

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

were considered. Additionally, potential germline variants were filtered out by excluding variants present in the 1000 Genomes Project database with a minor allele frequency >= 0.001.

Somatic CNVs were detected in paired tumor and normal samples using FACETS (v0.6.2) (Shen and Seshan 2016), alongside evaluation of tumor purity and ploidy. The analysis was performed with the following parameters: cnv_facets -t tumor_bam -n normal_bam -vcf snps --snp-nprocs 30 --depth 15 4000 --cval 25 400 --nbhd-snp 500. Copy numbers characterized by a total copy number of 2 and a lesser copy number of 1 were subsequently filtered out, as this is consistent with the normal diploid copy number state. To focus on large-scale CNV alterations, any CNV smaller than 100 kb in length was removed from the analysis.

Phylogenetic Tree Construction and Clustering Analysis

The genome was divided into 500-bp bins, and binary presence/absence matrices were constructed based on SV breakpoints and SNVs. A distance-based Neighbor-Joining (NJ) method was then applied to construct the phylogenetic trees. The resulting phylogenetic trees were subsequently visualized using the ggtree (Yu et al. 2018) package.

For clustering analysis, Jaccard indices were computed pairwise between samples. The resulting matrix of Jaccard indices was subjected to hierarchical clustering, using the pheatmap (v1.0.12) package in R (R Core Team 2023) with the “Pearson’s correlation” parameter.

Timing of somatic SVs

SVclone (v1.1.2) (Cmero et al. 2020) is a computational tool designed to infer the cancer cell fraction (CCF) of SV breakpoints based on short read sequencing data. To adapt SVclone for SVs detected from nanopore sequencing, several modifications were implemented: (1) SV supporting reads (including split and spanning reads), along with normal reads crossing the SV breakpoints, were utilized to calculate the variant allele fraction (VAF) of SV breakpoints. (2) As each long read in nanopore sequencing represents a single DNA molecule and does not overestimate the

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

number of normal reads in cases of DNA gain, the step of adjusting the number of normal reads was omitted. By integrating somatic SNVs, somatic CNVs, tumor purity, and ploidy, SVclone performed CCF estimation and clustering using a co-clustering mode. Next, clonality of somatic SVs was determined based on the somatic CNVs and SVclone clusters using the timing analysis algorithm MutationTimeR (Gerstung et al. 2020). This step assigned somatic SVs to four different timing categories，include early clonal, clonal not specified, late clonal and subclonal.

Long-read RNA sequencing (RNA-seq)

RNA from tumor and nontumor tissues was extracted using TRIzol Reagent (Invitrogen) according to the standard instructions. The RNA concentration was determined using Qubit RNA HS Assay Kit (Thermo Fisher Scientific2), and the purity and integrity were assessed by NanoDrop One, Qubit, and agarose gel electrophoresis. The cDNA library was prepared according to the SQK-DCS109 and EXP-NBD104 protocol (ONT). Then, the cDNA library was sequenced on a PromethION machine (ONT). Basecalling from raw data was performed in batches using Guppy software 3.2.8.

PCR validation of human-HBV fusion expression

Primers were designed to target the *KLHDC7B-DT* gene and HBV DNA sequences identified in the long-read RNA sequencing (Supplemental Table S5). A total of 35 ng SMART cDNA was mixed with 1 μL of forward primer, 1 μL of reverse primer, 12.5 μL of 2 × Taq Master Mix (Vazyme), and nuclease-free water to a total volume of 25 μL. The thermal cycling protocol was as follows: 3 min at 95 °C for initial denaturation, 35 cycles of 15 s at 95 °C, 15 s at 58 °C, and 4 min at 72 °C, followed by 5 min at 72 °C. Next, DNA was electrophoresed on a 1% agarose gel and sent to Youkang Bio for first-generation sequencing.

qPCR validation of gene expression

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

cDNA was prepared from 1 μg of RNA using the PrimeScript RT reagent Kit with gDNA Eraser (TAKARA) following the manufacturer's protocol. The qPCR primers were designed using Primer3Plus. Each pair of primers was designed to target exons to avoid amplification of genomic DNA. Oligos for qPCR are provided in Supplemental Table S5. For each sample, three replicates were performed using iTaq universal SYBR Green supermix (Bio-rad) following the manufacturer's instructions. The thermal cycling protocol on the CFX96™ Real-Time System (Bio-rad) was as follows: 3 min at 95 °C for initial denaturation, 41 cycles of (10 s at 95 °C, 30 s at 60 °C, and 15 s at 72 °C), followed by a melt curve from 65 °C to 95 °C in 0.5 °C increments. β-actin was employed as the internal control, and the Cq value was used for quantitative analysis. The experiment was repeated three times.

Data Access

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center(CNCB-NGDC Members Partners 2024), China National Center for Bioinformation/ Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA007177) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human. All code used is open source and available at https://github.com/tianfuzeng/ONT_SV, as well as in the Supplemental Code.

Competing interests

The authors declare no potential conflicts of interest.

Acknowledgments

We thank Ranlei Wei from the Laboratory of Omics Technology and Bioinformatics, West China Hospital of Sichuan University, Kefei Yuan from the West China Hospital of Sichuan University for their technical assistance. This work was supported by grants from the National Natural Science Foundation of China (82173383 to D. Xie, 32200508 to L. Xia, 82202260 to H.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Liao), the Sichuan Province Science and Technology Program (2022NSFSC1553 to L. Xia), the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC23024 to D. Xie), and China Postdoctoral Science Foundation (2022TQ0221 to H. Liao).

Author contributions

T.Z., H.L. and L.X. were responsible for the experimental design, execution, data analysis. T.Z. wrote the paper, with contributions from all other authors. Y.H. and J.Z. were responsible for RNA sequencing and validation. S.Y., Y.L. and X.L. were responsible for data downloading, organization, and processing of raw data. D.X. was responsible for supervision of research, data interpretation and manuscript preparation.

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y. 2012. GenomeView: a next-generation genome browser. *Nucleic Acids Res* **40**: e12.

Abyzov A, Li S, Kim DR, Mohiyuddin M, Stutz AM, Parrish NF, Mu XJ, Clark W, Chen K, Hurles M et al. 2015. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6**: 7256.

Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, Lee I, Kirsche M, Wappel R, Kramer M et al. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res* **30**: 1258-1273.

Alvarez EG, Demeulemeester J, Otero P, Jolly C, Garcia-Souto D, Pequeno-Valtierra A, Zamora J, Tojo M, Temes J, Baez-Ortega A et al. 2021. Aberrant integration of Hepatitis B virus DNA promotes major restructuring of human hepatocellular carcinoma genome architecture. *Nat Commun* **12**: 6910.

Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. 2017. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun* **8**: 1077.

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779-786.

Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* doi:10.3322/caac.21834.

Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, Sanders MA, Ellis P, Alder C, Hooks Y et al. 2019. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**: 538-542.

Carlessi R, Denisenko E, Boslem E, Kohn-Gaone J, Main N, Abu Bakar NDB, Shirolkar GD, Jones M, Beasley AB, Poppe D et al. 2023. Single-nucleus RNA sequencing of pre-

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

malignant liver reveals disease-associated hepatocyte state with HCC prognostic potential. *Cell Genom* **3**: 100301.

Chen L, Zhang C, Xue R, Liu M, Bai J, Bao J, Wang Y, Jiang N, Li Z, Wang W et al. 2024a. Deep whole-genome analysis of 494 hepatocellular carcinomas. *Nature* **627**: 586-593.

Chen Z, Shi Q, Zhao Y, Xu M, Liu Y, Li X, Liu L, Sun M, Wu X, Shao Z et al. 2024b. Long-read transcriptome landscapes of primary and metastatic liver cancers at transcript resolution. *Biomark Res* **12**: 4.

Choy TK, Wang CY, Phan NN, Khoa Ta HD, Anuraga G, Liu YH, Wu YF, Lee KH, Chuang JY, Kao TJ. 2021. Identification of Dipeptidyl Peptidase (DPP) Family Genes in Clinical Breast Cancer Patients via an Integrated Bioinformatics Approach. *Diagnostics (Basel)* **11**.

Cmero M, Yuan K, Ong CS, Schroder J, Evolution P, Heterogeneity Working G, Corcoran NM, Papenfuss T, Hovens CM, Markowetz F et al. 2020. Inferring structural variant cancer cell fraction. *Nat Commun* **11**: 730.

CNCB-NGDC Members Partners. 2024. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res* **52**: D18-D32.

Cosenza MR, Rodriguez-Martin B, Korbel JO. 2022. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annu Rev Genomics Hum Genet* **23**: 123-152.

Craven KE, Fischer CG, Jiang L, Pallavajjala A, Lin MT, Eshleman JR. 2022. Optimizing Insertion and Deletion Detection Using Next-Generation Sequencing in the Clinical Laboratory. *J Mol Diagn* **24**: 1217-1231.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**.

Fernandez AF, Rosales C, Lopez-Nieva P, Grana O, Ballestar E, Ropero S, Espada J, Melo SA, Lujambio A, Fraga MF et al. 2009. The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer. *Genome Res* **19**: 438-451.

Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M et al. 2016. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* **48**: 500-509.

Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, Shigemizu D, Nakagawa H, Mizokami M, Shimada M. 2021. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* **13**: 65.

George J, Lim JS, Jang SJ, Cun Y, Ozretic L, Kong G, Leenders F, Lu X, Fernandez-Cuesta L, Bosco G et al. 2015. Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**: 47-53.

Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K et al. 2020. The evolutionary history of 2,658 cancers. *Nature* **578**: 122-128.

Goldsmith C, Cohen D, Dubois A, Martinez MG, Petitjean K, Corlu A, Testoni B, Hernandez-Vargas H, Chemin I. 2021. Cas9-targeted nanopore sequencing reveals epigenetic heterogeneity after de novo assembly of native full-length hepatitis B virus genomes. *Microb Genom* **7**.

Ho XD, Nguyen HG, Trinh LH, Reimann E, Prans E, Koks G, Maasalu K, Le VQ, Nguyen VH, Le NTN et al. 2017. Analysis of the Expression of Repetitive DNA Elements in Osteosarcoma. *Front Genet* **8**: 193.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Hochhaus A, Larson RA, Guilhot F, Radich JP, Branford S, Hughes TP, Baccarani M, Deininger MW, Cervantes F, Fujihara S et al. 2017a. Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. *N Engl J Med* **376**: 917-927.

Hochhaus A, Masszi T, Giles FJ, Radich JP, Ross DM, Gomez Casares MT, Hellmann A, Stentoft J, Conneally E, Garcia-Gutierrez V et al. 2017b. Treatment-free remission following frontline nilotinib in patients with chronic myeloid leukemia in chronic phase: results from the ENESTfreedom study. *Leukemia* **31**: 1525-1531.

Hu Z, Qu S. 2021. EVA1C Is a Potential Prognostic Biomarker and Correlated With Immune Infiltration Levels in WHO Grade II/III Glioma. *Front Immunol* **12**: 683572.

Huang A, Zhao X, Yang XR, Li FQ, Zhou XL, Wu K, Zhang X, Sun QM, Cao Y, Zhu HM et al. 2017. Circumventing intratumoral heterogeneity to identify potential therapeutic targets in hepatocellular carcinoma. *J Hepatol* **67**: 293-301.

Khemlina G, Ikeda S, Kurzrock R. 2017. The biology of Hepatocellular carcinoma: implications for genomic and immune therapies. *Mol Cancer* **16**: 149.

Kolomietz E, Meyn MS, Pandita A, Squire JA. 2002. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer* **35**: 97-112.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *bioRxiv*: Preprint at https://arxiv.org/abs/1303.3997v1302.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li MX, Wang HY, Yuan CH, Ma ZL, Jiang B, Li L, Zhang L, Xiu DR. 2021. KLHDC7B-DT aggravates pancreatic ductal adenocarcinoma development via inducing cross-talk between cancer cells and macrophages. *Clin Sci (Lond)* **135**: 629-649.

Li R, Du Y, Chen Z, Xu D, Lin T, Jin S, Wang G, Liu Z, Lu M, Chen X et al. 2020a. Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* **370**: 82-89.

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE et al. 2020b. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**: 112-121.

Liao WW Asri M Ebler J Doerr D Haukness M Hickey G Lu S Lucas JK Monlong J Abel HJ et al. 2023. A draft human pangenome reference. *Nature* **617**: 312-324.

Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, Lencioni R, Koike K, Zucman-Rossi J, Finn RS. 2021. Hepatocellular carcinoma. *Nat Rev Dis Primers* **7**: 6.

Maillet V, Boussetta N, Leclerc J, Fauveau V, Foretz M, Viollet B, Couty JP, Celton-Morizur S, Perret C, Desdouets C. 2018. LKB1 as a Gatekeeper of Hepatocyte Proliferation and Genomic Integrity during Liver Regeneration. *Cell Rep* **22**: 1994-2005.

Martin-Pardillos A, Cajal SRY. 2019. Characterization of Kelch domain-containing protein 7B in breast tumours and breast cancer cell lines. *Oncol Lett* **18**: 2853-2860.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Mirabello L, Sun C, Ghosh A, Rodriguez AC, Schiffman M, Wentzensen N, Hildesheim A, Herrero R, Wacholder S, Lorincz A et al. 2012. Methylation of human papillomavirus type 16 genome and risk of cervical precancer in a Costa Rican population. *J Natl Cancer Inst* **104**: 556-565.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Muller M, Bird TG, Nault JC. 2020. The landscape of gene mutations in cirrhosis and hepatocellular carcinoma. *J Hepatol* **72**: 990-1002.

Munkhjargal B, Kondo K, Soejima S, Tegshee B, Takai C, Kawakita N, Toba H, Takizawa H. 2023. Aberrant methylation of dipeptidyl peptidase□like 6 as a potential prognostic biomarker for lung adenocarcinoma. *Oncol Lett* **25**: 206.

Pascarella G, Hon CC, Hashimoto K, Busch A, Luginbuhl J, Parr C, Hin Yip W, Abe K, Kratz A, Bonetti A et al. 2022. Recombination of repeat elements generates somatic complexity in human genomes. *Cell* **185**: 3025-3040 e3026.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rajaby R, Liu DX, Au CH, Cheung YT, Lau AYT, Yang QY, Sung WK. 2023. INSurVeyor: improving insertion calling from short read sequencing data. *Nat Commun* **14**: 3243.

Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T, Ciccarelli FD. 2019. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**: 1.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* **23**: 411-418.

Rosas D, Raez LE, Russo A, Rolfo C. 2021. Neuregulin 1 Gene (NRG1). A Potentially New Targetable Alteration for the Treatment of Lung Cancer. *Cancers (Basel)* **13**.

Sakamoto Y, Xu L, Seki M, Yokoyama TT, Kasahara M, Kashima Y, Ohashi A, Shimada Y, Motoi N, Tsuchihara K et al. 2020. Long-read sequencing for non-small-cell lung cancer genomes. *Genome Res* **30**: 1243-1257.

Schulze K, Nault JC, Villanueva A. 2016. Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *J Hepatol* **65**: 1031-1042.

Shen R, Seshan VE. 2016. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**: e131.

Shen S, Li H, Liu J, Sun L, Yuan Y. 2020. The panoramic picture of pepsinogen gene family with pan-cancer. *Cancer Med* **9**: 9064-9080.

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* doi:10.1038/s41587-023-02024-y.

Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H et al. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**: 561-566.

Strandgaard T, Nordentoft I, Lamy P, Christensen E, Thomsen MBH, Jensen JB, Dyrskjøt L. 2020. Mutational Analysis of Field Cancerization in Bladder Cancer. *Bladder Cancer* **6**: 253-264.

Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C et al. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* **44**: 765-769.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**: 82-93.

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S et al. 2011. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**: 593-596.

Villanueva A. 2019. Hepatocellular Carcinoma. *N Engl J Med* **380**: 1450-1462.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K et al. 2015. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**: 495-501.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Wu Z, Jiang Z, Li T, Xie C, Zhao L, Yang J, Ouyang S, Liu Y, Li T, Xie Z. 2021. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* **12**: 6501.

Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet* **19**: e1010514.

Yahiro K, Ogura K, Tsutsuki H, Iyoda S, Ohnishi M, Moss J. 2021. A novel endoplasmic stress mediator, Kelch domain containing 7B (KLHDC7B), increased Harakiri (HRK) in the SubAB-induced apoptosis signaling pathway. *Cell Death Discov* **7**: 360.

Yan H, Yang Y, Zhang L, Tang G, Wang Y, Xue G, Zhou W, Sun S. 2015. Characterization of the genotype and integration patterns of hepatitis B virus in early- and late-onset hepatocellular carcinoma. *Hepatology* **61**: 1821-1831.

Yu G, Lam TT, Zhu H, Guan Y. 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**: 3041-3043.

Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, Desai N, Sultmann H, Moch H, Pathogens P et al. 2020. The landscape of viral associations in human cancers. *Nat Genet* **52**: 320-330.

Zhao LH, Liu X, Yan HX, Li WY, Zeng X, Yang Y, Zhao J, Liu SP, Zhuang XH, Lin C et al. 2016. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat Commun* **7**: 12992.

Zheng B, Liu XL, Fan R, Bai J, Wen H, Du LT, Jiang GQ, Wang CY, Fan XT, Ye YN et al. 2021. The Landscape of Cell-Free HBV Integrations and Mutations in Cirrhosis and Hepatocellular Carcinoma Patients. *Clin Cancer Res* **27**: 3772-3783.

Zhu H, Lin Y, Lu D, Wang S, Liu Y, Dong L, Meng Q, Gao J, Wang Y, Song N et al. 2023. Proteomics of adjacent-to-tumor samples uncovers clinically relevant biological events in hepatocellular carcinoma. *Natl Sci Rev* **10**: nwad167.

Zhuo Z, Rong W, Li H, Li Y, Luo X, Liu Y, Tang X, Zhang L, Su F, Cui H et al. 2021. Long-read sequencing reveals the structural complexity of genomic integration of HBV DNA in hepatocellular carcinoma. *NPJ Genom Med* **6**: 84.

Zucman-Rossi J, Villanueva A, Nault J-C, Llovet JM. 2015. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. *Gastroenterology* **149**: 1226-1239.e1224.

Figure 1. Identification of somatic SVs through multi-region long-read sequencing.

A. Schematic diagram showing the sampling locations for five patients. Blue dots represent adjacent nontumor tissues. Red dots represent tumor tissues.

B. Bar plots depicting the number of somatic SVs across individual samples, including insertions (INS), deletions (DEL), duplications (DUP), inversions (INV), and translocations (TRA).

C. Stacked plots showing the distribution of somatic SV lengths categorized into two ranges:

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

(50bp, 1kb) and (1kb, 10kb), stratified based on SV types.

D. The proportion of different SV size ranges within each somatic SV type.

Figure 2. Somatic SVs were prominent in adjacent nontumors.

A. Bar plots showing the number of somatic mutations (classified into SVs, SNVs, and CNVs) in adjacent nontumor and tumor tissues from all patients.

B. Heatmap depicting the hierarchical clustering of 37 tumor and adjacent nontumor samples based on Jaccard indexes calculated from somatic SVs.

C. Heatmap depicting the hierarchical clustering of 37 tumor and adjacent nontumor samples based on Jaccard indexes calculated from somatic SNVs. The white dashed box outlines a cluster consisting of adjacent nontumor samples.

Figure 3. The characteristics of shared somatic SVs between adjacent nontumor and tumor tissues.

A. The frequency of each somatic SV type in different categories was shown: shared (in both adjacent nontumor and tumor tissues), T-specific (only in tumor tissues), and N-specific (only in adjacent nontumor tissues).

B. The length distribution of shared somatic SVs between adjacent nontumor and tumor tissues, as well as the SVs unique to either adjacent nontumor or tumor tissues.

C. An overview of genes overlapped by recurrently shared somatic SVs in adjacent nontumor and tumor tissues across patients. The left bar chart shows the mutation percentage across patients. The right bar chart displays the mutation percentage across samples.

Figure 4. Inferring SV types based on HBV DNA integration patterns.

A. In sample HCC9_T3, an HBV DNA insertion induces a fold-back inversion in the host DNA.

Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

The top panel illustrates the CNV increase occurring approximately 76.23 Mb downstream of the integration site. Sequencing data demonstrate that long reads encompass the sequences flanking the HBV integration, revealing identical breakpoints and reverse duplication of the adjacent sequences.

B. An example of HBV DNA integration resulting in a translocation (HCC13_T2). The IGV plot below illustrating the alignment of long reads to Chr22, HBV, and Chr8. The plot above depicting the concordance of its breakpoint on Chr8 (red dashed line) with a breakpoint of approximately 5.36 Mbp CNV (gray shade).

Figure 5. Integration breakpoints of HBV DNA in human and HBV genome.

A. Integration breakpoints of HBV DNA within the human genome. The same color scheme represents the same patient. Circular points represent tumor samples, while triangular points represent adjacent nontumor samples.

B. A schematic diagram of shared HBV-induced SVs at all tumor sampling sites in the patients. The horizontal line represents the relative position of the integrated HBV sequence within the HBV genome.

C. In HCC13, the clonal status of HBV integration sites was in the context of somatic SVs (top panel), with green indicating early clonal events. The red asterisk symbol represents the shared HBV integration sites. The middle panel represents the CNV status, while the lower panel displays the inferred mutation timeline.

Figure 6. Genes disrupted by HBV-induced SVs

A. Manhattan plot showing the genes affected by HBV-induced SVs. Tumor samples are represented by orange points, while adjacent nontumor samples are represented by blue points. Tumor genes annotated by COSMIC or NCG6 are marked in red.

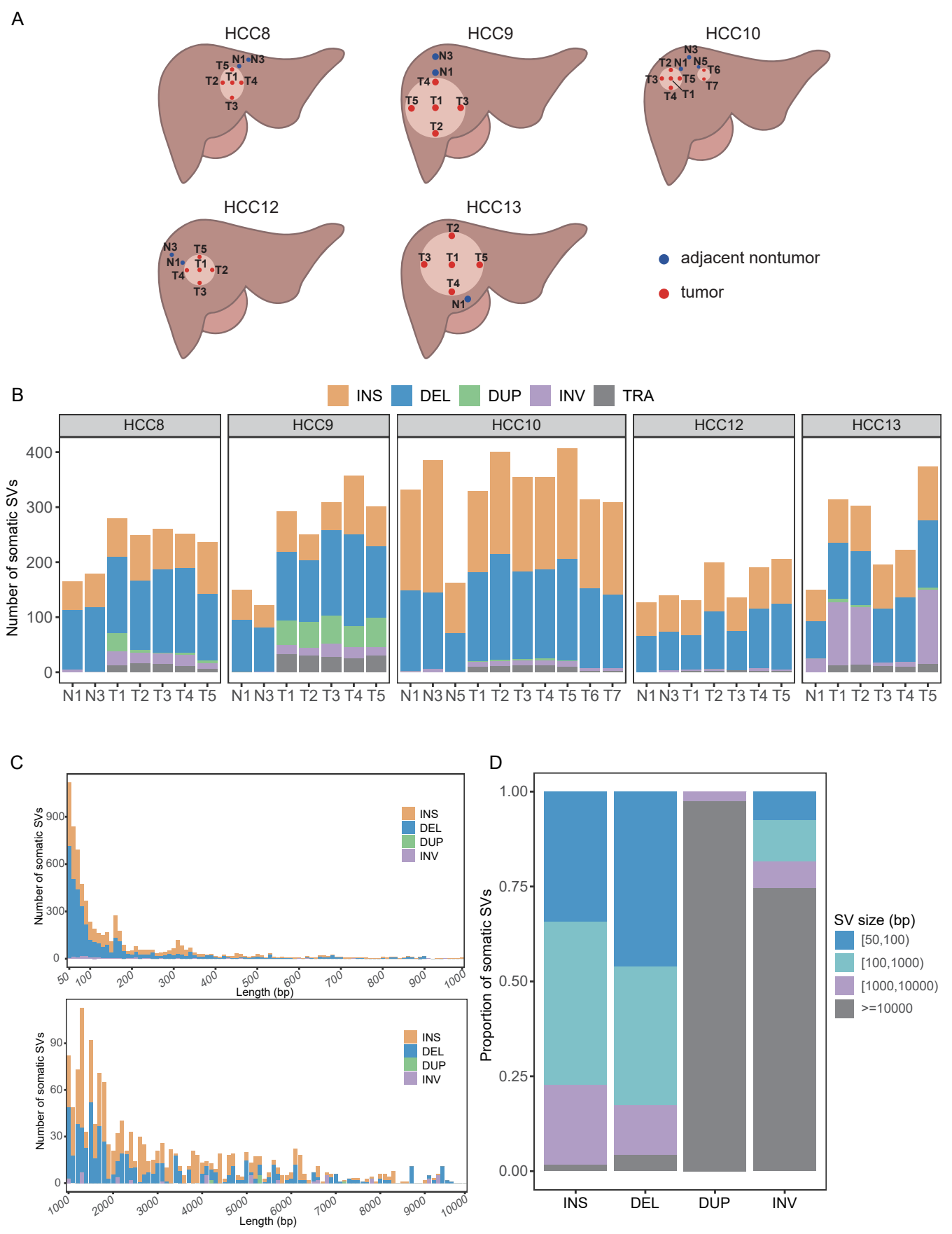
Downloaded from genome.cshlp.org on May 13, 2026 . Published by Cold Spring Harbor Laboratory Press

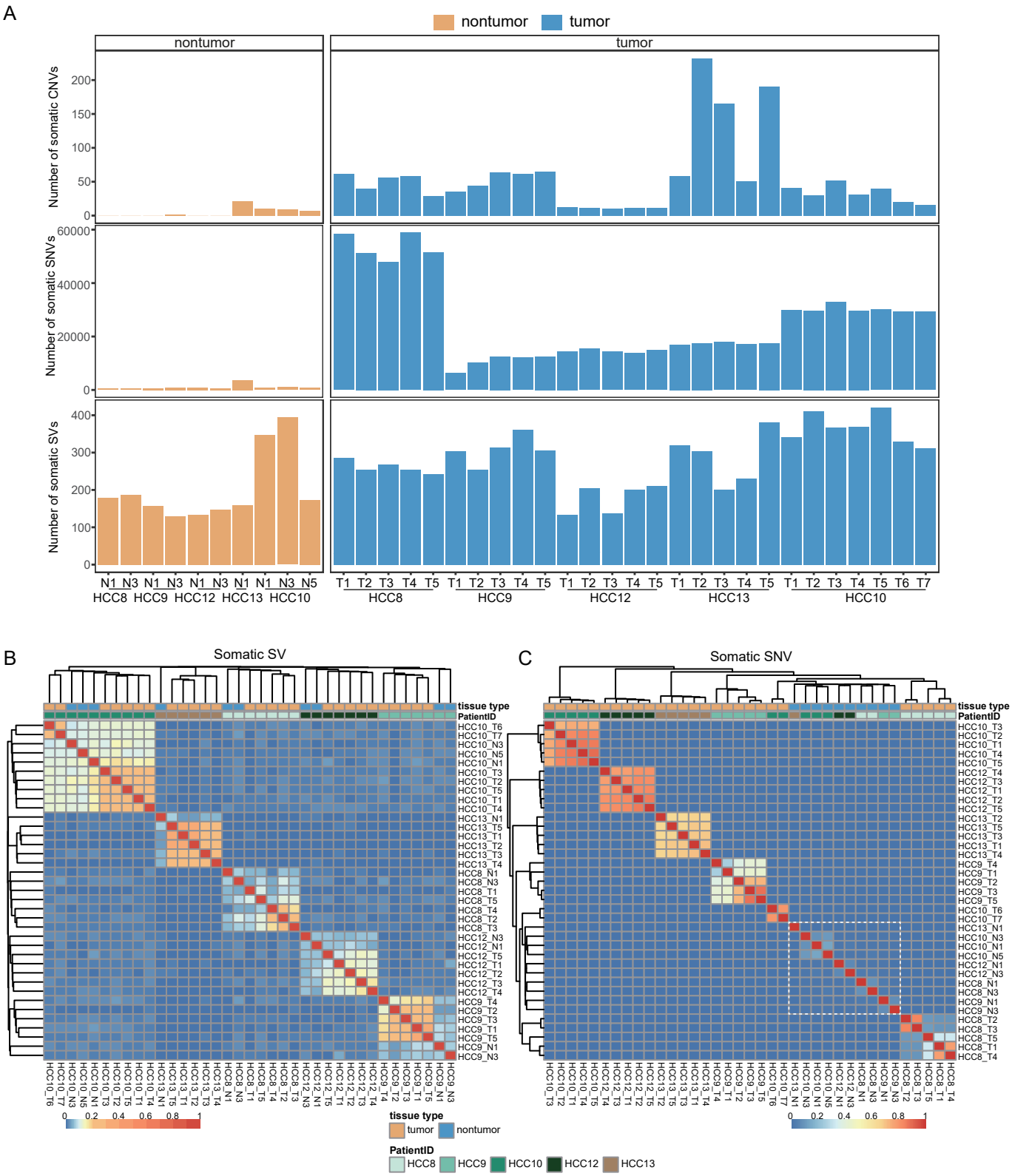
B. IGV plot depicting the fusion transcript formed between the *KLHDC7B-DT* gene and the HBV DNA sequence. The top track represents the assembled TRA sequence, while the bottom track illustrates gene annotations for the TRA sequence. The middle tracks include long-read RNA sequencing from both tumor and adjacent nontumor samples, and Sanger sequencing reads for the fusion transcript PCR product.

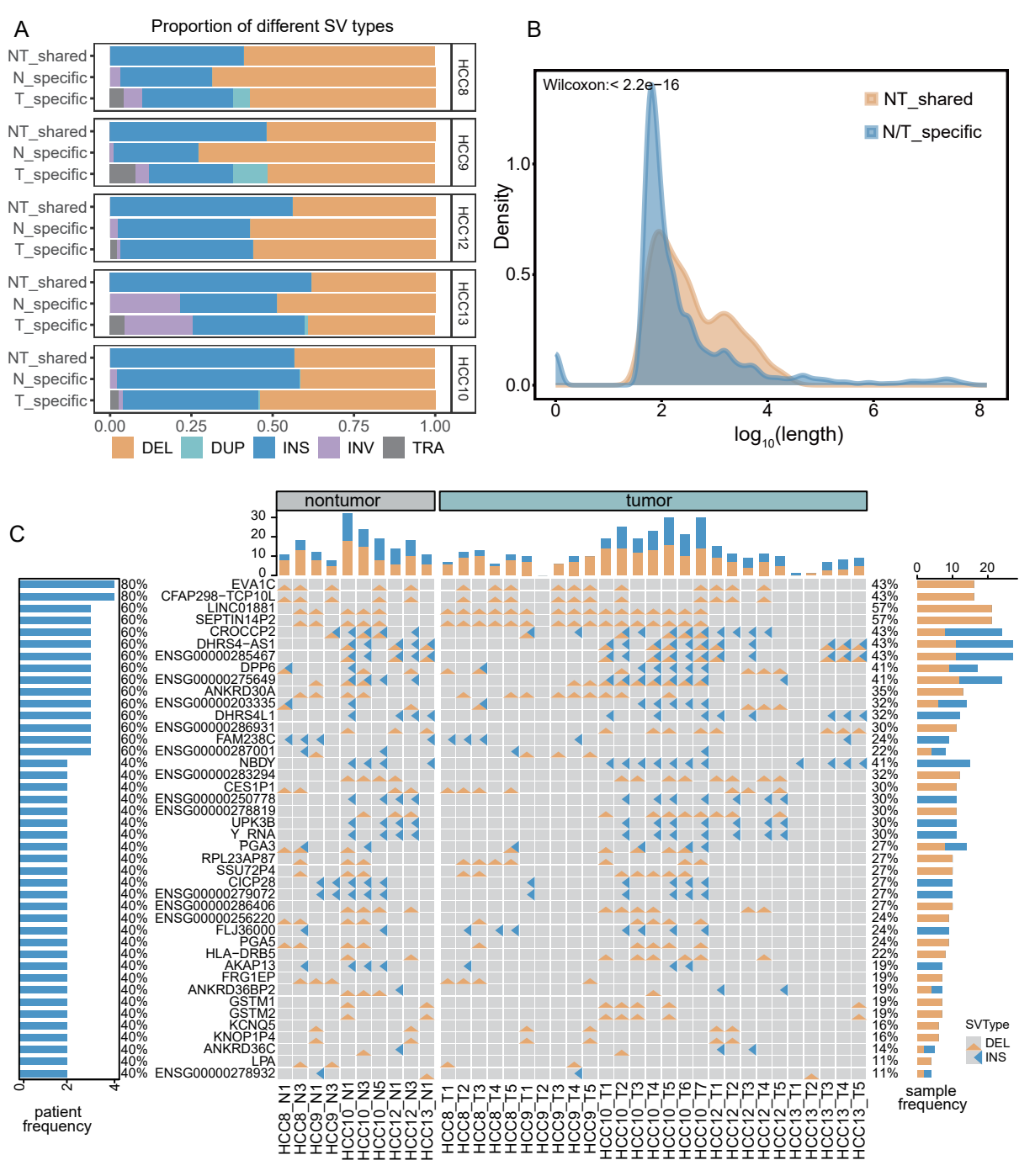
C. The PCR products of fusion transcript were detected by agarose gel electrophoresis.

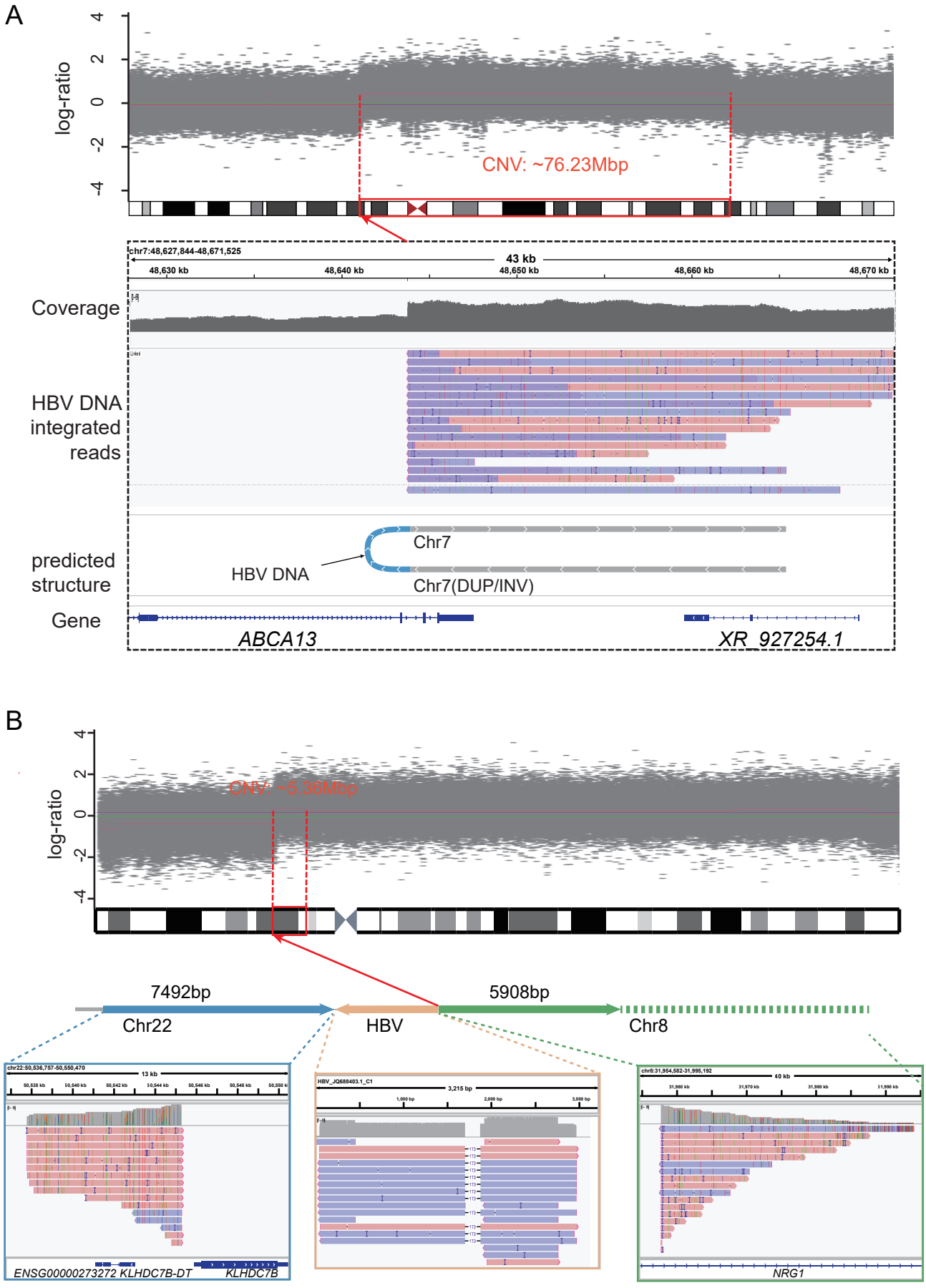
D. The relative expression levels of *KLHDC7B* determined using RT□PCR in samples with and without the HBV-induced TRA.

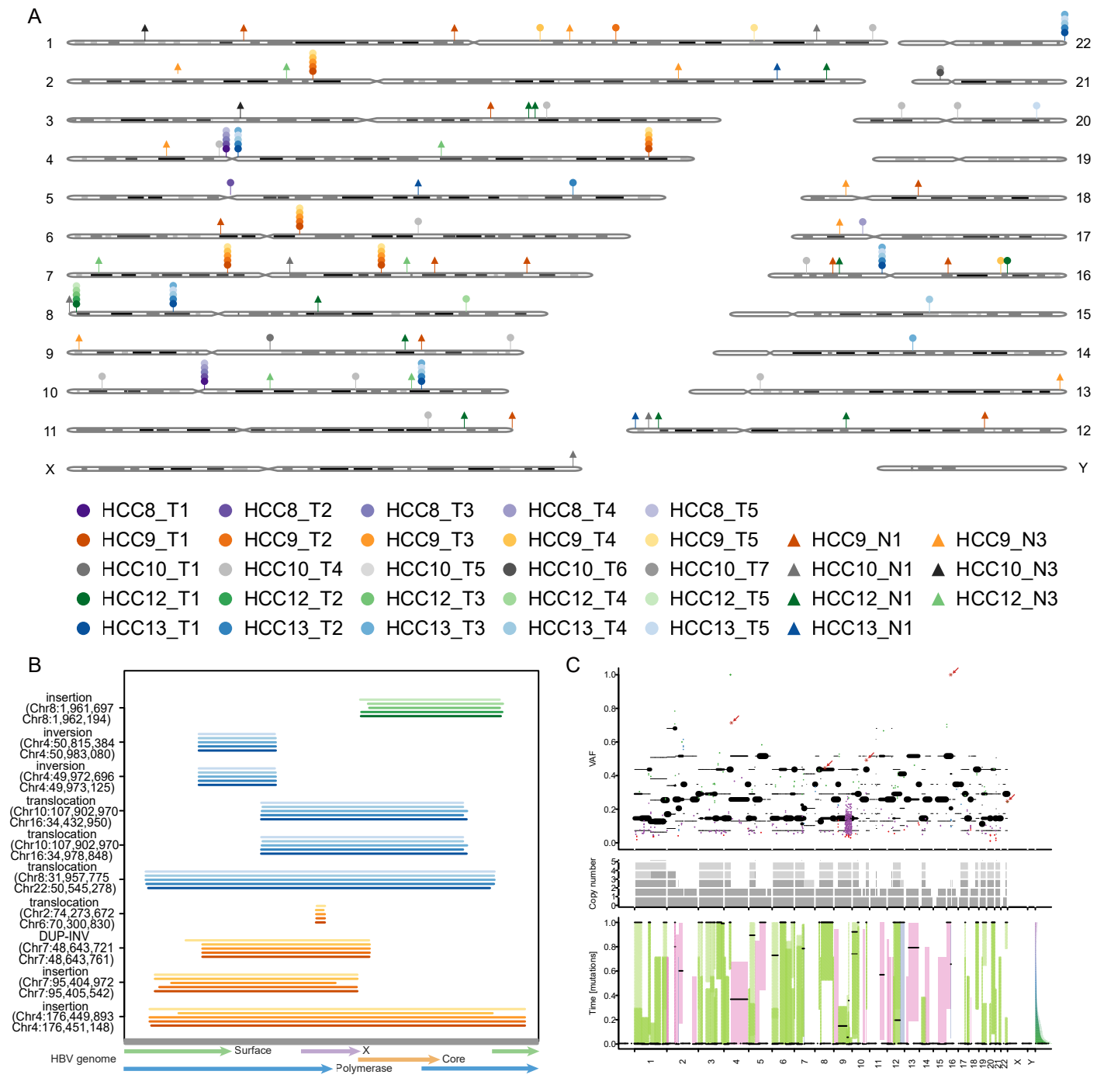
E. The relative expression levels of *KLHDC7B-DT* determined using RT□PCR in samples with and without the HBV-induced TRA.

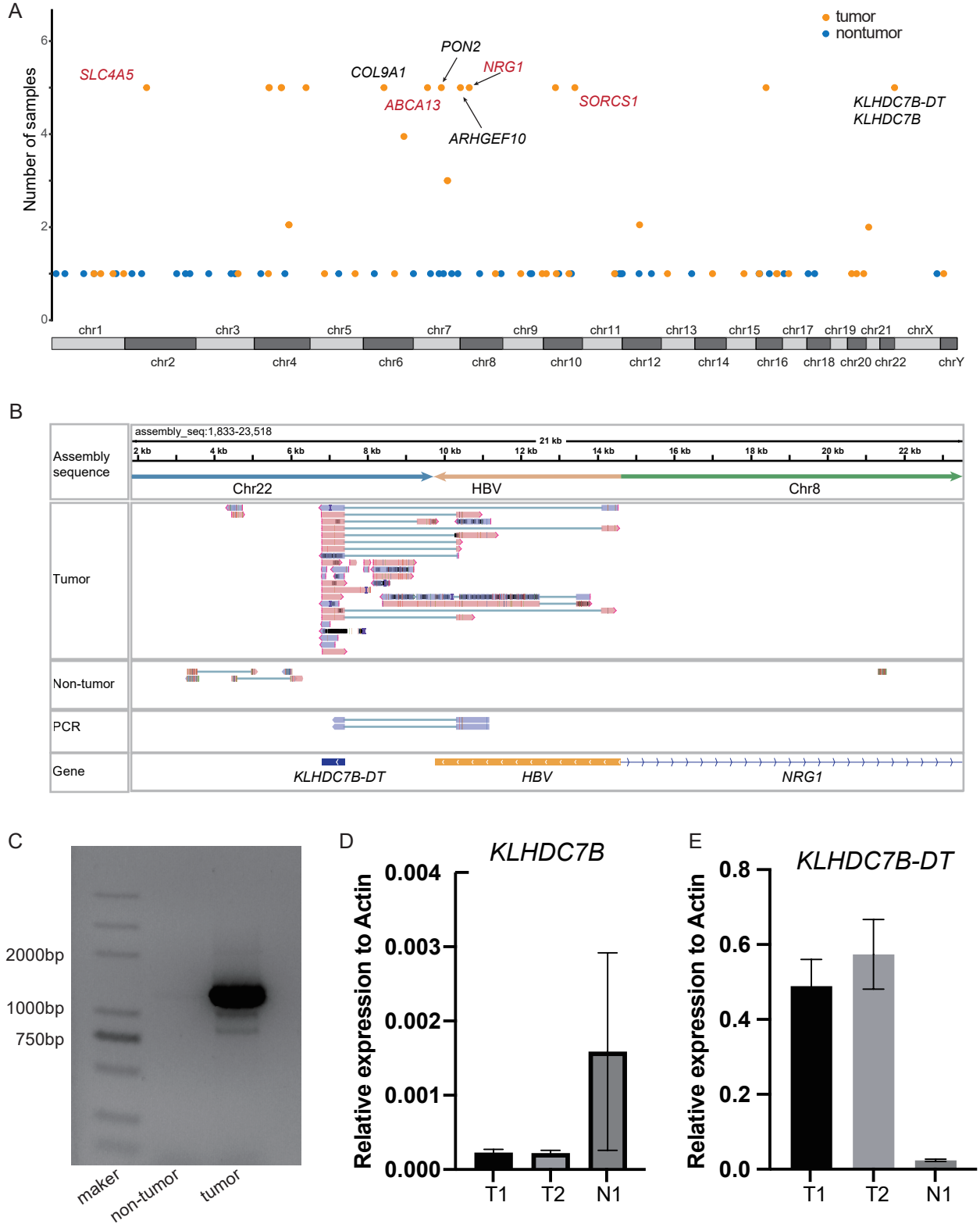

A
HCC8
HCC9
HCC10
HCC12
HCC13
adjacent nontumor
tumor
B
INS
DEL
DUP
INV
TRA
Number of somatic SVs
C
Length (bp)
D
Proportion of somatic SVs
SV size (bp)
[50,100)
[100,1000)
[1000,10000)
>=10000


A
nontumor
tumor
Number of somatic CNVs
Number of somatic SNVs
Number of somatic SVs
N1_N3 N1_N3 N1_N3 N1 N1_N3_N5
HCC8 HCC9 HCC12 HCC13 HCC10
T1_T2_T3_T4_T5 T1_T2_T3_T4_T5 T1_T2_T3_T4_T5 T1_T2_T3_T4_T5 T1_T2_T3_T4_T5_T6_T7
HCC8 HCC9 HCC12 HCC13 HCC10
B
Somatic SV
C
Somatic SNV
tissue type
PatientID
tumor
nontumor
HCC8
HCC9
HCC10
HCC12
HCC13


A
Proportion of different SV types
NT_shared
N_specific
T_specific
HCC8
HCC9
HCC12
HCC13
HCC10
0.00
0.25
0.50
0.75
1.00
DEL
DUP
INS
INV
TRA
B
Wilcoxon:< 2.2e−16
NT_shared
N/T_specific
Density
1.0
0.5
0.0
0
2
4
6
8
log10(length)
C
nontumor
tumor
30
20
10
0
80% EVA1C
80% CFAP298−TCP10L
60% LINC01881
60% SEPTIN14P2
60% CROCCP2
60% DHRS4−AS1
60% ENSG00000285467
60% DPP6
60% ENSG00000275649
60% ANKRD30A
60% ENSG00000203335
60% DHRS4L1
60% ENSG00000286931
60% FAM238C
60% ENSG00000287001
40% NBDY
40% ENSG00000283294
40% CES1P1
40% ENSG00000250778
40% ENSG00000278819
40% UPK3B
40% Y_RNA
40% PGA3
40% RPL23AP87
40% SSU72P4
40% CICP28
40% ENSG00000279072
40% ENSG00000286406
40% ENSG00000256220
40% FLJ36000
40% PGA5
40% HLA−DRB5
40% AKAP13
40% FRG1EP
40% ANKRD36BP2
40% GSTM1
40% GSTM2
40% KCNQ5
40% KNOP1P4
40% ANKRD36C
40% LPA
40% ENSG00000278932
0
2
4
patient frequency
43%
43%
57%
57%
43%
43%
43%
41%
41%
35%
32%
32%
30%
24%
22%
41%
32%
30%
30%
30%
30%
30%
27%
27%
27%
27%
27%
24%
24%
24%
22%
19%
19%
19%
19%
19%
16%
16%
14%
11%
11%
0
10
20
SVType
DEL
INS
sample frequency
HCC8_N1
HCC8_N3
HCC9_N1
HCC9_N3
HCC10_N1
HCC10_N3
HCC10_N5
HCC12_N1
HCC12_N3
HCC13_N1
HCC8_T1
HCC8_T2
HCC8_T3
HCC8_T4
HCC8_T5
HCC9_T1
HCC9_T2
HCC9_T3
HCC9_T4
HCC9_T5
HCC10_T1
HCC10_T2
HCC10_T3
HCC10_T4
HCC10_T5
HCC10_T6
HCC10_T7
HCC12_T1
HCC12_T2
HCC12_T3
HCC12_T4
HCC12_T5
HCC13_T1
HCC13_T2
HCC13_T3
HCC13_T4
HCC13_T5


A
log-ratio
4
2
0
-2
-4
CNV: ~76.23Mbp
chr7:48,627,844-48,671,525
43 kb
48,630 kb
48,640 kb
48,650 kb
48,660 kb
48,670 kb
Coverage
HBV DNA integrated reads
predicted structure
Chr7
HBV DNA
Chr7(DUP/INV)
Gene
ABCA13
XR_927254.1
B
log-ratio
CNV: ~5.36Mbp
7492bp
5908bp
Chr22
HBV
Chr8
chr22:50,536,757-50,550,470
13 kb
ENSG00000273272 KLHDC7B-DT
KLHDC7B
HBV_JQ688403.1_C1
3,215 bp
chr8:31,954,582-31,995,192
40 kb
NRG1


A
1
2
3
4
5
6
7
8
9
10
11
X
22
21
20
19
18
17
16
15
14
13
12
Y
HCC8_T1
HCC8_T2
HCC8_T3
HCC8_T4
HCC8_T5
HCC9_T1
HCC9_T2
HCC9_T3
HCC9_T4
HCC9_T5
HCC9_N1
HCC9_N3
HCC10_T1
HCC10_T4
HCC10_T5
HCC10_T6
HCC10_T7
HCC10_N1
HCC10_N3
HCC12_T1
HCC12_T2
HCC12_T3
HCC12_T4
HCC12_T5
HCC12_N1
HCC12_N3
HCC13_T1
HCC13_T2
HCC13_T3
HCC13_T4
HCC13_T5
HCC13_N1
B
insertion (Chr8:1,961,697 Chr8:1,962,194)
inversion (Chr4:50,815,384 Chr4:50,983,080)
inversion (Chr4:49,972,696 Chr4:49,973,125)
translocation (Chr10:107,902,970 Chr16:34,432,950)
translocation (Chr10:107,902,970 Chr16:34,978,848)
translocation (Chr8:31,957,775 Chr22:50,545,278)
translocation (Chr2:74,273,672 Chr6:70,300,830)
DUP-INV (Chr7:48,643,721 Chr7:48,643,761)
insertion (Chr7:95,404,972 Chr7:95,405,542)
insertion (Chr4:176,449,893 Chr4:176,451,148)
HBV genome
Surface
X
Core
Polymerase
C
VAF
Copy number
Time [mutations]
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


A
tumor
nontumor
Number of samples
SLC4A5
COL9A1
PON2
NRG1
ABCA13
ARHGEF10
SORCS1
KLHDC7B-DT
KLHDC7B
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
B
Assembly sequence
assembly_seq:1,833-23,518
Chr22
HBV
Chr8
Tumor
Non-tumor
PCR
Gene
KLHDC7B-DT
HBV
NRG1
C
2000bp
1000bp
750bp
maker
non-tumor
tumor
D
KLHDC7B
Relative expression to Actin
T1
T2
N1
E
KLHDC7B-DT
Relative expression to Actin
T1
T2
N1