

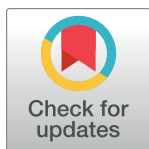
RESEARCH ARTICLE

Comprehensive analysis of full-length transcripts reveals novel splicing abnormalities and oncogenic transcripts in liver cancer

Hiroki Kiyose^{1,2}, Hidewaki Nakagawa³, Atsushi Ono⁴, Hiroshi Aikata⁴, Masaki Ueno⁵, Shinya Hayami⁵, Hiroki Yamaue⁵, Kazuaki Chayama^{6,7,8}, Mihoko Shimada¹, Jing Hao Wong¹, Akihiro Fujimoto^{1*}

1 Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, **2** Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan, **3** Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, **4** Department of Gastroenterology and Metabolism, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan, **5** Department of Surgery II, Wakayama Medical University, Wakayama, Japan, **6** Collaborative Research Laboratory of Medical Innovation, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan, **7** Research Center for Hepatology and Gastroenterology, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan, **8** RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

* afujimoto@m.u-tokyo.ac.jp



OPEN ACCESS

Citation: Kiyose H, Nakagawa H, Ono A, Aikata H, Ueno M, Hayami S, et al. (2022) Comprehensive analysis of full-length transcripts reveals novel splicing abnormalities and oncogenic transcripts in liver cancer. *PLoS Genet* 18(8): e1010342. <https://doi.org/10.1371/journal.pgen.1010342>

Editor: Peter McKinnon, St Jude Children's Research Hospital, UNITED STATES

Received: April 19, 2022

Accepted: July 14, 2022

Published: August 4, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pgen.1010342>

Copyright: © 2022 Kiyose et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The source code of SPLICE is freely available from Github (<https://github.com/hkiyose/SPLICE>). The sequencing data have been deposited in the National Bioscience

Abstract

Genes generate transcripts of various functions by alternative splicing. However, in most transcriptome studies, short-reads sequencing technologies (next-generation sequencers) have been used, leaving full-length transcripts unobserved directly. Although long-reads sequencing technologies would enable the sequencing of full-length transcripts, the data analysis is difficult. In this study, we developed an analysis pipeline named SPLICE and analyzed cDNA sequences from 42 pairs of hepatocellular carcinoma (HCC) and matched non-cancerous livers with an Oxford Nanopore sequencer. Our analysis detected 46,663 transcripts from the protein-coding genes in the HCCs and the matched non-cancerous livers, of which 5,366 (11.5%) were novel. A comparison of expression levels identified 9,933 differentially expressed transcripts (DETs) in 4,744 genes. Interestingly, 746 genes with DETs, including the *LINE1-MET* transcript, were not found by a gene-level analysis. We also found that fusion transcripts of transposable elements and hepatitis B virus (HBV) were overexpressed in HCCs. *In vitro* experiments on DETs showed that *LINE1-MET* and HBV-human transposable elements promoted cell growth. Furthermore, fusion gene detection showed novel recurrent fusion events that were not detected in the short-reads. These results suggest the efficiency of full-length transcriptome studies and the importance of splicing variants in carcinogenesis.

Author summary

Genes generate transcripts of various functions by alternative splicing. However, in most transcriptome studies, short-reads sequencing technologies (next-generation sequencers)

Database Center (NBDC) under accession JGAD000635 (<https://humandbs.biosciencedbc.jp/en/hum0182-v5>). The sequencing data is controlled access, and an application to the NBDC is needed.

Funding: This work was supported by grant 18H02680 from the Japan Society for the Promotion of Science (JSPS) to AF, and by grant JP21km0908001 from Japan Agency for Medical Research and Development (AMED) to AF, and by grant from Takeda Science Foundation to AF. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

have been used, leaving full-length transcripts unobserved directly. In this study, we developed an analysis pipeline named SPLICE for long-read transcriptome sequencing and analyzed cDNA sequences from 42 pairs of hepatocellular carcinoma (HCC), and matched non-cancerous livers with an Oxford Nanopore sequencer. Our analysis detected 5,366 novel transcripts and 9,933 differentially expressed transcripts in 4,744 genes between HCCs and non-cancerous livers. An analysis of hepatitis B virus (HBV) transcripts showed that fusion transcripts of the HBV gene and human transposable elements were overexpressed in HBV-infected HCCs. We also identified fusion genes that were not found in the short-reads. These results suggest that long-reads sequencing technologies provide a fuller understanding of cancer transcripts and that our method contributes to the analysis of transcriptome sequences by such technologies.

Introduction

Hepatocellular carcinoma (HCC) is the third-leading cause of death worldwide and the seventh most common form of cancer [1]. Common etiological factors in liver carcinogenesis include infection by hepatitis B virus (HBV) or hepatitis C virus (HCV), but other factors, such as alcohol intake, metabolic diseases, and exposure to specific carcinogens also play significant roles [2]. These factors cause liver inflammation, leading to cirrhosis and result in malignant transformations in hepatocytes [2]. To elucidate the molecular mechanisms underlying liver carcinogenesis, genetic and transcriptional aberrations have been investigated [3–7]. These previous studies identified numerous somatic mutations and differentially expressed genes (DEGs), and have led to the discovery of liver cancer-associated pathways, such as apoptosis, Wnt signaling, chromatin remodeling, and lengthening telomeres [3–7]. In HBV-infected liver cancers, HBV integrations and HBV-human fusion transcripts have been identified [8]. Fusion genes have been investigated, but recurrent fusions were rare [9]. Although these studies have expanded our knowledge of HCCs, our understanding of liver carcinogenesis is far from complete [10,11]. To obtain deeper insight into the mechanism responsible for HCCs, applications of novel technologies and analysis of new aspects of genetic and transcriptional aberrations are needed.

We aimed to analyze full-length transcripts for HCCs in the present study. The majority of previous transcriptome studies have used microarrays or short-reads sequencing technologies, and therefore lacked information on full-length transcripts that may assist in the detection of splicing variants expressed from each gene. Protein sequences are specified by the mRNA sequences of transcripts; thus, the direct observation of transcripts should provide essential information about carcinogenesis. Indeed, several recent studies have suggested that transcript-specific functions contribute to carcinogenesis in breast and ovarian cancers [12–14]. In HCCs, a recent study analyzed short-reads RNA-seq data and detected oncogenic splicing changes in the *AFMID* gene [15]. These splicing changes were estimated to occur at the early stage of HCC carcinogenesis and were associated with patient survival [15]. Overall, previous studies strongly suggest that splicing variants have important roles that are not recognized without an analysis of full-length transcripts.

One promising approach to observe transcripts is RNA-seq using long-reads sequencers [16], but their application in cancer research is still limited [17–21] for reasons such as high error rates in the long-reads sequencing [22] and the complexity or heterogeneity of the cancer genome. To overcome this problem and identify cancer-related transcriptional changes, we developed a method for analyzing RNA-seq data obtained by a long-reads sequencer, MinION

(Oxford Nanopore). In this study, we sequenced 42 pairs of cDNA from HCCs and matched non-cancerous livers that were previously used for whole-genome sequencing (WGS) and RNA-seq of short-reads [4,23,24]. We developed an analysis method named SPLICE and analyzed the long-reads data. Among the reads mapped to coding genes, 61.7% contained entire coding sequences, and a total of 46,663 transcripts were detected in the protein-coding genes in the HCCs and matched livers, of which 5,366 (11.5%) were novel. A comparison of expression levels found 9,933 differentially expressed transcripts (DETs) in 4,744 genes. Interestingly, 746 genes with DET were not detected by the gene-level analysis. Exonization of transposable elements (TEs) was also identified. In the analysis of HBV transcripts, fusion transcripts of the HBV HBx gene and human TEs were overexpressed in HBV-infected HCCs. These results suggest that long-reads sequencing technologies provide a fuller understanding of cancer transcripts and that the SPLICE method contributes to the analysis of transcriptome sequences by such technologies.

Results

cDNA sequencing with Oxford Nanopore sequencer

We sequenced cDNA from MCF-7 cells and previously published 42 cancer and their matched non-cancerous liver pairs (**S1 and S2 Tables and S1 and S2 Data**) [4]. cDNA was generated by reverse transcription with a SMARTer PCR cDNA synthesis kit (TakaraBio), and sequencing libraries were constructed using the SQK-LSK109 library construction kit (Oxford Nanopore) according to the manufacturer's instruction. We performed one run for each sample using flowcell R9.4 (Oxford Nanopore). The data yield ranged from 4.6 Gbp to 20.6 Gbp (average 12.7 Gbp), and the number of reads was 3,061,266 to 14,873,634 (average 9.1 M reads) (**S2 Data**).

Construction of analysis pipeline and evaluation of the SPLICE method

To analyze the long-reads data, we constructed an analysis pipeline named SPLICE (see **Methods**) (**Fig 1A**). Long-reads sequencers have the capability to sequence entire cDNA end-to-end [25], allowing us to capture full-length transcripts without assembly [26]. However, long-reads sequencers also have high error rates [22], which could lead to incorrect conclusions about splicing variants and their expression levels. To overcome these obstacles, we analyzed error patterns and developed an analysis pipeline.

Reads were mapped to the reference genome sequence (hg38), reference transcriptome sequence (GENCODE (version 28), and RefSeq (release 88)) using minimap2 software [27]. We found that false-positive splicing aberrations were caused by alignment errors at splicing junction sites, mapping errors due to highly homologous genomic regions, and artificial chimeric reads (**Fig 1B**). The following describes how the SPLICE pipeline removed possible error patterns from the mapping results (**S1 Fig**). (1) Reads were mapped to the reference genome and the reference transcriptome sequences, and reads were removed if the number of matched bases in a read to the reference genome sequence was smaller than that to the reference transcriptome sequence (removal of mapping errors). (2) For the identification of novel transcripts (novel splicing variants), we evaluated the error rate within ± 5 -bp regions from splicing junction sites and removed candidates for error rates $\geq 20\%$ (removal of alignment errors around splicing sites). (3) Because fusion gene candidates can be generated by mis-ligation of two transcripts during the library preparation, we removed fusion gene candidates that contained primer sequences for cDNA synthesis between two genes (removal of artificial chimeras). (4) Mapping errors due to highly homologous genomic regions also can cause fusion gene candidates; thus, we compared the results of mapping to the reference genome and the

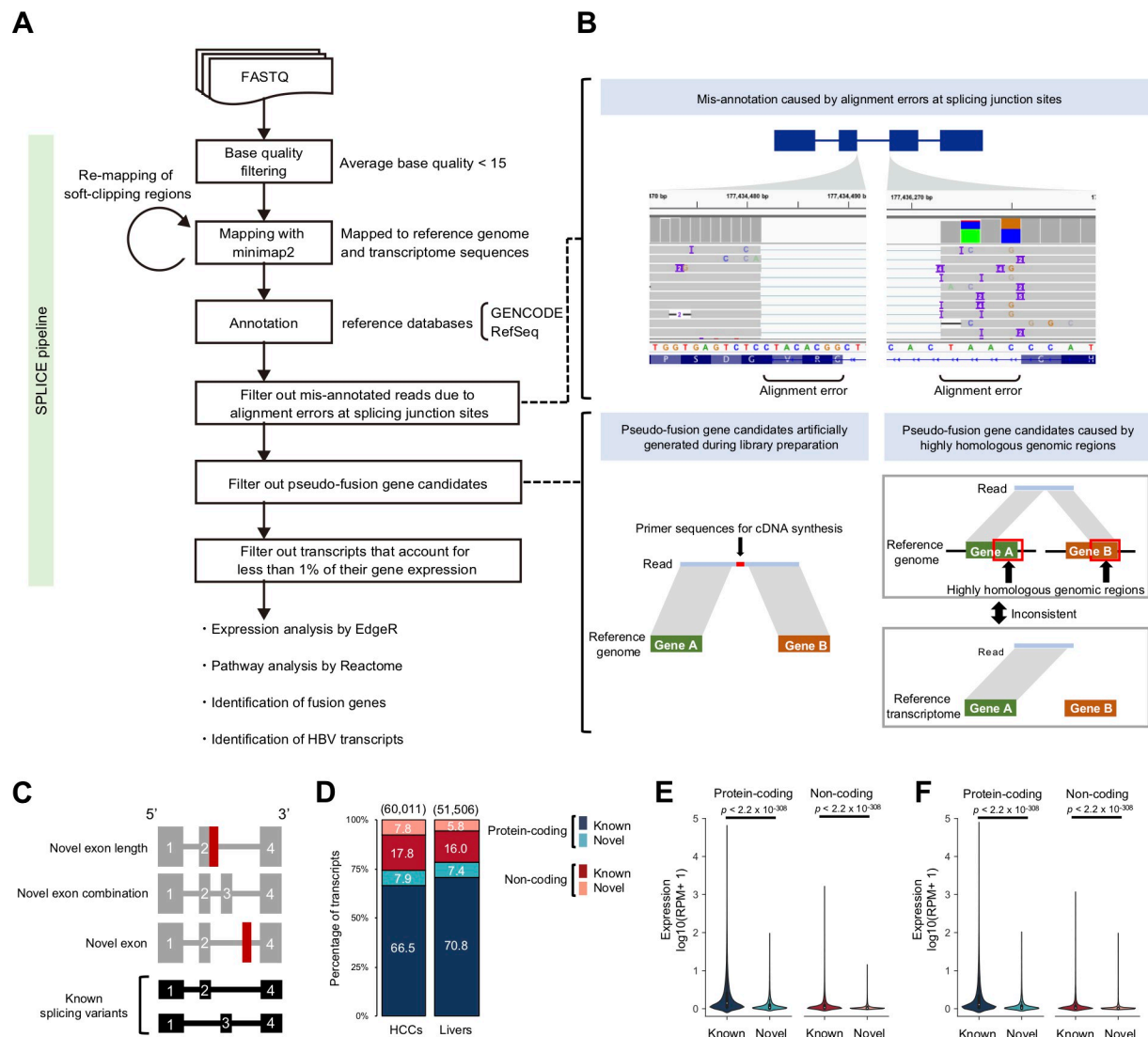


Fig 1. Overview of the SPLICE pipeline and HCC sequencing library. (A) Workflow of the SPLICE pipeline and data analysis of this study. Reads with an average base quality score < 15 were removed from the input FASTQ files. The read sequences were mapped to the reference genome, reference transcriptome sequences, and remapped soft-clipped regions (≥ 60 bp). Candidates corresponding to the error patterns or low-frequency transcripts were removed. The final data were used for the subsequent data analyses. (B) Error patterns in cDNA-seq using MinION. Alignment errors caused by sequencing errors, artificial chimeras, and mapping errors of highly homologous genes can result in errors. (C) Classification of patterns of novel transcripts. "Novel exon length" contains known exons with different lengths from the reference database. "Novel exon combination" has a different exon combination from the reference database. "Novel exon" contains exons expressed from regions that are not overlapped with known exons. (D) Composition of the transcripts. The stacked bar chart shows the distribution (%) of detected transcripts by transcript type. The total number of transcripts is shown above the bars. (E,F) Comparison of expression levels of known and novel transcripts. Transcript abundance was measured in reads per million reads (RPM), and $\log_{10}$ converted values for RPM+1 are shown in the violin plot. *P*-values were calculated by the Wilcoxon rank-sum test. (E) Comparison in HCC samples. (F) Comparison in non-cancerous liver samples.

<https://doi.org/10.1371/journal.pgen.1010342.g001>

reference transcriptome sequences and removed candidates if both results did not detect the same fusion genes (removal of mapping errors). (5) We removed a transcript if the expression level was less than 1% the total amount of the gene expression (removal of possible artificial chimeras). Using these filters, we estimated the expression level of transcripts and identified novel transcript candidates from long-reads, which we classified into transcripts with "novel exon length", "novel exon combination", and "novel exon" (Fig 1C).

To evaluate this method, we sequenced cDNA obtained from the MCF-7 cell line and analyzed it with the SPLICE pipeline. As a result, 13,424,802 reads (99.5%) were mapped ([S1 Table](#)). We calculated the number of reads mapped to each transcript per 1 million reads to estimate the expression levels. The numbers of reads of each transcript were then summed by each gene and compared to the expression level by short-reads sequencing technology [[28](#)]. The gene expression levels between long- and short-reads showed a strong correlation ($p < 2.2 \times 10^{-308}$, $r = 0.80$) ([S2 Fig](#)).

Comparison of SPLICE method with other tools

We evaluated the detection performance of splicing variants and fusion genes. Although several Nanopore long-reads RNA-seq analysis pipelines exist [[21,29–34](#)], to the best of our knowledge, no software can detect both splicing variants and fusion genes. In this study, we used sequence data from MCF-7 and an HCC sample (RK107C) and compared the performance of SPLICE and four other methods (TALON [[34](#)], FLAIR [[21](#)], StringTie [[31](#)], and Bambu [[29](#)]) for the splicing variant detection and SPLICE and two other methods (LongGF [[32](#)] and JAFFAL [[30](#)]) for the fusion gene detection. In the analysis of splicing variants, SPLICE identified the third highest number of transcripts after FLAIR and StringTie ([S3A Fig](#)). For MCF-7, SPLICE had the highest concordance rate with IsoSeq MCF-7 transcriptome data (see Data access) for known transcripts and second highest for novel transcripts ([S3B Fig](#)).

We next compared the fusion genes detected from MCF7 by SPLICE, LongGF, and JAFFAL. Previously known *BCAS3-BCAS4* [[35](#)] was detected only by SPLICE, and *BCAS3-ATXN7* was detected by SPLICE and LongGF ([S4A and S4B Fig](#)). In HCC, *AC138969.4-PDXDC1* was detected only by SPLICE, *C8B-PAH* was detected by SPLICE and JAFFAL, three fusions (*GCH1-SERPINA6*, *ABCD3-C1orf123*, and *CPS1-WNT10B*) were detected by SPLICE and LongGF, and two fusions (*TBC1D23-TF* and *NBEAL1-RPL12*) were detected only by LongGF. One of the fusions (*NBEAL1-RPL12*), detected only by LongGF, could not be validated by PCR ([S4B Fig](#)). These results indicate that SPLICE had the highest accuracy for detecting fusion genes. Based on these results, we considered our analysis pipeline to have sufficient accuracy for analyzing transcript aberrations in cancers.

Analysis of HCC and non-cancerous liver samples

We then analyzed the sequence data from HCCs and matched non-cancerous livers. After filtering low-quality reads (average base quality score < 15), we obtained an average of 9,043,803 reads (12.9 Gbp on average) in the HCCs and 9,013,602 reads (12.4 Gbp on average) in the livers ([S2 Data](#)). The average mapping rates were 98.9% for the HCCs and 97.3% for the livers ([S2 Data](#)). Using the SPLICE pipeline, we identified transcripts supported by ≥ 3 reads and obtained a total of 60,011 non-redundant transcripts in the HCCs and 51,506 in the livers ([Fig 1D](#)). Of these, 66.5% and 70.8% were known protein-coding transcripts, 7.9% and 7.4% were novel protein-coding transcripts, 17.8% and 16.0% were known non-coding transcripts, and 7.8% and 5.8% were novel non-coding transcripts in the HCCs and the livers, respectively ([Fig 1D](#)). We then compared the expression levels of known and novel transcripts. The average expression levels of novel transcripts were lower than those of known transcripts for both protein-coding and non-coding transcripts in the HCCs and the livers ($p < 2.2 \times 10^{-308}$ for protein-coding, $p < 2.2 \times 10^{-308}$ for non-coding) ([Fig 1E and 1F](#)).

Among reads mapped to known protein-coding transcripts, 61.7% had full-length coding sequences (CDSs) ([S5A and S5B Fig](#)). The percentage of reads mapped to the full-length CDS region was negatively correlated with the length of genes ([S5C Fig](#)). The percentages of reads

containing the full-length CDS were correlated with RIN (RNA Integrity Number) values of the RNA samples ($p = 9.8 \times 10^{-6}$, $r = 0.46$) (S6 Fig), indicating the importance of RNA quality for identifying full-length transcripts. We then compared the gene expression levels between previous short-reads and our long-reads data [4] and found high correlations ($p < 2.2 \times 10^{-308}$ in all samples, average $r = 0.79$) (S7 Fig).

Features of novel exons

We identified novel exons by comparing the exon locations in the GENCODE (version 28) and RefSeq (release 88) databases. Our analysis identified 767 and 531 novel exons in 4,769 and 3,786 transcripts of protein-coding genes in the HCCs and matched non-cancerous livers, respectively. Most (92.0%) were the first or last exon of the transcripts (Fig 2A). Since TEs are known to be exonized by tissue-specific or cancer-induced hypomethylation of the genome [36], we calculated the percentage of TE-derived novel exons. In the analysis, novel exons > 70% covered by TEs were considered TE-derived novel exons. The percentage of TE-derived novel exons was significantly higher than that of TE-derived known exons in the first, last, and middle exons ($p = 5.9 \times 10^{-153}$ for the first exon, $p = 2.2 \times 10^{-106}$ for the last exon, $p = 1.7 \times 10^{-79}$ for middle exon) (Fig 2B and S3 Table). A permutation test showed that the novel first exon was significantly enriched in L2 antisense, ERV1 antisense, and ERVL-MaLR sense (Fig 2C). The novel last exon was significantly enriched in *Alu* sense, and the middle exon was significantly enriched in *Alu* antisense, ERVL-MaLR antisense, and ERV1 antisense (Fig 2C). These results showed that exonized repeat families and strands differ by their position in the transcripts.

To estimate the biological importance of the novel exons, we estimated their evolutionary conservation using a conservation score obtained by 100 species. The average conservation score of the novel exon region was significantly lower than that of the known exon region ($p < 2.2 \times 10^{-308}$) (Fig 2D), suggesting that most of the novel exons had less importance in evolutionary aspects. However, a part of novel exons was highly conserved (Fig 2E and S3 Data), suggesting that they had important biological functionalities.

Identification of significantly differentiated genes and transcripts

One of the most important features of the long-reads transcriptome is detecting DETs. Thus, by comparing the expression level between HCCs and matched non-cancerous livers, we aimed to find DEGs and DETs. DETs are transcripts that show significant differentiation in expression level between HCCs and matched non-cancerous livers. DEGs are genes obtained by comparing the total amount of transcripts within a gene between HCCs and the livers. We estimated the expression level based on the number of reads and compared them using edgeR software [37]. Our analysis identified 4,744 DEGs and 9,933 DETs after adjusting for multiple testing (FDR < 0.01) (Fig 3A and S4 Data) [38]. A Reactome pathway analysis [39] for genes with up-regulated DETs was conducted, and genes with the terms “M Phase”, “Cell cycle, Mitotic”, and “Cell cycle” were significantly enriched (S5 Data). Among the genes with DETs, 746 genes (935 transcripts) were not detected as DEGs (DET-specific genes) (Fig 3B and S4 Data). To know the features of the DET-specific genes, we compared the numbers of transcripts identified in this study and transcription start sites (TSSs) of each gene in the database. DET-specific genes had a significantly larger number of transcripts than DEGs ($p = 6.2 \times 10^{-41}$) (Fig 3C) and also had a larger number of multiple TSSs (Fig 3D) ($p = 5.2 \times 10^{-13}$).

Among the 9,933 DETs, 200 transcripts (99 genes) were from previously reported driver genes (Fig 3E and S6 Data) [4,41,42], and 13 genes (16 transcripts) were DET-specific. Of

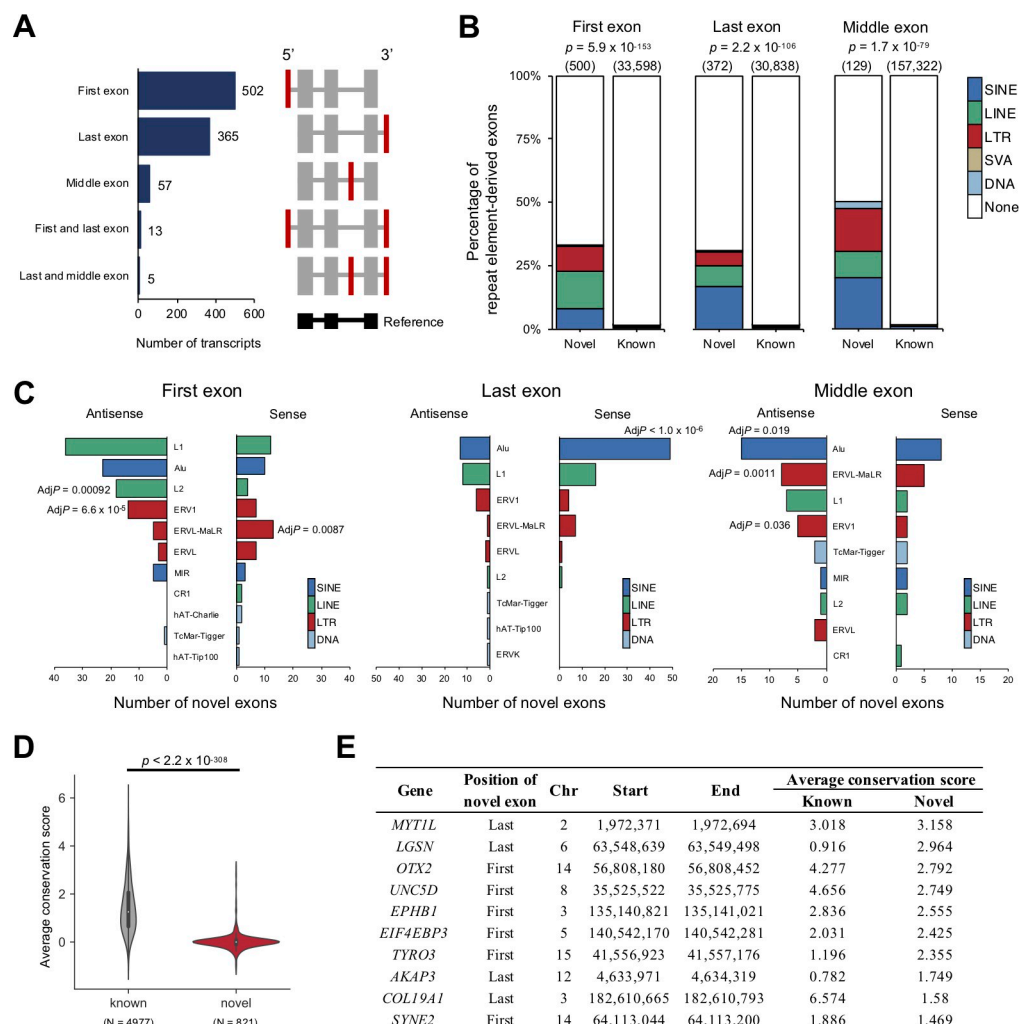


Fig 2. Features of novel exons. (A) Patterns of transcripts with novel exons. (Left panel) Bar graph showing the number of transcripts by the position of the novel exon. (Right panel) Schematic diagram of the location of novel exons. The novel exons are shown in red. (B) Percentage of TE-derived exons in novel and known exons by exon position. The stacked bar chart shows the distribution of TE-derived exons. The total number of exons is shown above the bars. Statistical significance of the total abundance of TE-derived exons between the known and novel exons was calculated by Fisher's exact test. (C) The number and enrichment of TE-derived novel exons. TEs were classified by repeat family, and the sense or antisense strand for the gene strand was distinguished. The statistical significance was examined by a permutation test. P -values were adjusted by Bonferroni correction (AdjP). (D) Average conservation scores between novel and known exon regions. For each gene, the average conservation scores were calculated based on the phastCons scores database (100 vertebrates). P -values were calculated by the Wilcoxon rank-sum test. (E) List of transcripts with novel exons with high average conservation scores. The 10 candidates are shown in order of average conservation score.

<https://doi.org/10.1371/journal.pgen.1010342.g002>

these, transcripts of the *MET* oncogene (XM_006715990.2, XM_011516223.1) were up-regulated in HCCs, and tumor suppressor genes, such as *HLA-B* (ENST00000359635.9, ENST00000450871.6) and *APC* (NM_000038.5), were down-regulated (S6 Data). Transcripts from 16 HCC driver genes (*ALB*, *APOB*, *ASH1L*, *CDKN2A*, *CPS1*, *CTNNB1*, *DHX9*, *IDH1*, *IL6ST*, *NCOR1*, *NFE2L2*, *NRAS*, *NSMCE2*, *NUP133*, *SETDB1*, and *XPO1*) were significantly differentiated (S6 Data) [4,41,42]. Of these, transcripts from *ASH1L*, *CDKN2A*, *CTNNB1*, *DHX9*, *NRAS*, *NSMCE2*, *NUP133*, *SETDB1*, and *XPO1* were up-regulated in the HCCs and other transcripts were down-regulated. *ALB* is known to be highly expressed in the liver, and 3

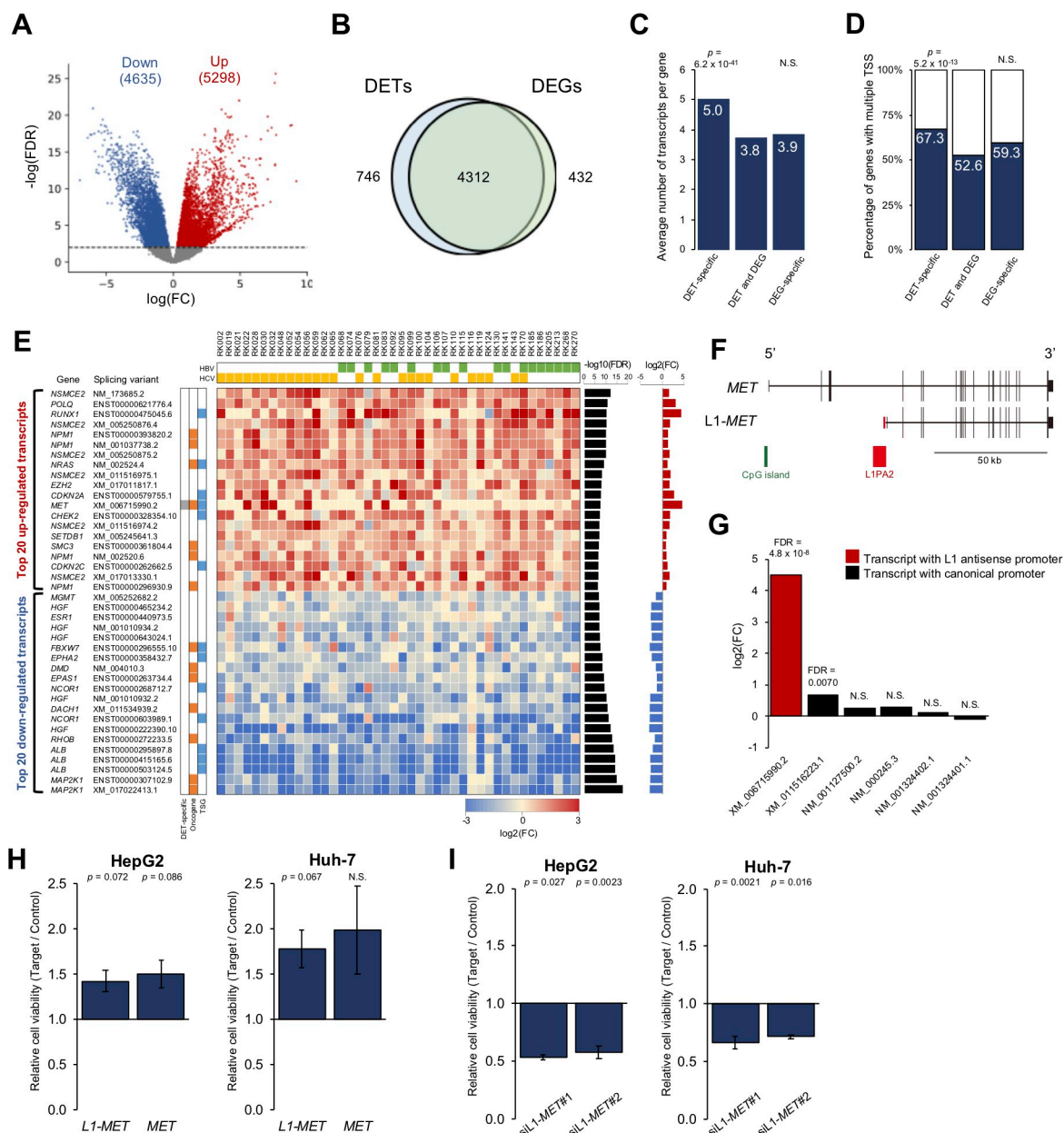


Fig 3. Landscape of significantly differentiated transcripts between HCCs and matched non-cancerous livers. (A) Volcano plot of DETs in cancers and non-cancerous livers. Red and blue dots represent transcripts that were significantly up- and down-regulated ($\text{FDR} < 0.01$), respectively. (B) Venn diagram of the number of genes with DETs and DEGs. (C) Comparison of the average number of transcripts per gene for each item in the Venn diagram in B. P-values for the abundance of transcripts per gene in DET-specific and DEG-specific genes were calculated by the Wilcoxon rank-sum test. The numbers of transcripts in each gene were compared for DET-specific genes and other genes and for DEG-specific genes and other genes. (D) Percentage of genes with multiple TSS for each item in the Venn diagram in B. P-values for the enrichment of genes with multiple TSS in DET-specific and DEG-specific genes were calculated by Fisher's exact test. The numbers of TSS in each gene were compared for DET-specific genes and other genes and for DEG-specific genes and other genes. (E) The landscape of DETs in driver genes. Twenty up-regulated and down-regulated DETs are shown in order of lower FDR. Heatmap showing the expression differences ($\log_2(\text{fold change})$) in cancer and matched non-cancerous livers. (F) Schematic diagram of the structural differences in L1-MET (XM_006715990.2) and other MET transcripts. L1-MET has the L1-derived first exon shown in red. (G) Bar chart showing expression differences between cancers and matched livers in MET transcripts. The red bar represents the MET transcript with the L1 antisense promoter, and black bars represent MET transcripts with the canonical promoter. P-values were obtained using quasi-likelihood methods in EdgeR software [40]. The Benjamini-Hochberg method was used for multiple testing correction (FDR). (H) Effect of L1-MET overexpression on the proliferation of HepG2 and Huh-7 cell lines assessed by CCK-8. P-values were calculated by the one-sample t-test ($n = 3$). (I) Effect of L1-MET knockdown on the proliferation of HepG2 and Huh-7 cell lines assessed by CCK-8. P-values were calculated by the one-sample t-test ($n = 3$). Error bars represent the standard error of the mean. N.S.: Not significant.

<https://doi.org/10.1371/journal.pgen.1010342.g003>

major transcripts (ENST00000503124.5, ENST00000415165.6, and ENST00000295897.8) were detected in HCCs and matched livers (S6 Data).

We then focused on genes that had both significantly up-regulated and down-regulated transcripts (referred to as bi-directionally significantly expressed (BiExp) genes). Our analysis identified 80 BiExp genes (270 transcripts) (S7 Data). Among them, *AFMID*, which is a previously reported oncogenic BiExp gene [15], was successfully detected in our analysis (S8 Fig). Among the 746 DET-specific genes, 42 were BiExp genes. As expected, DEGs had a smaller percentage of BiExp genes than did DET-specific genes ($p = 1.6 \times 10^{-15}$) (S4 Table), as BiExp genes had both up- and down-regulated transcripts whose expression levels canceled out each other.

Exonization of TEs in HCCs

In the HCCs, a transcript of *MET* (XM_6715990.2) was ranked first among the significantly up-regulated DET-specific driver genes ($\text{FDR} = 4.8 \times 10^{-8}$) (Fig 3E and S6 Data). A previous study reported that the first exon of XM_006715990.2 is L1-derived, and this transcript is expressed by the L1 antisense promoter (referred to as L1-*MET*) (Fig 3F) [43]. L1-*MET* was remarkably up-regulated in HCCs compared to other *MET* transcripts with the canonical promoter (Fig 3G). Since the role of L1-*MET* in HCCs is largely unknown, we performed overexpression and knockdown experiments of L1-*MET* in the HepG2 and Huh-7 cell lines. The overexpression of L1-*MET* and canonical *MET* (NM_001127500.2) increased cell proliferation (Figs 3H and S9), and the knockdown decreased cell proliferation of both cell lines (Figs 3I and S10), indicating that L1-*MET* contributes to the proliferation of HCC cells. Since the exonization of TEs was observed in HCCs and livers (Fig 2B), we next focused on cancer-specific TE-derived exons ($n = 829$) (S5 Table). We found that they were significantly enriched in the first and last exons ($p = 7.6 \times 10^{-43}$ for the first exon, $p = 3.9 \times 10^{-7}$ for the last exon) (S11 Fig). An analysis of repeat families showed that L1 sense and antisense, L2 antisense, ERVL sense and antisense, ERVL-MaLR sense, and LTR sense were significantly enriched in the first exon of cancer-specific transcripts (S5 Table). The liver-specific TE-derived exons were also significantly enriched in the first and last exons ($p = 2.9 \times 10^{-6}$ for the first exon, $p = 0.0012$ for the last exon) (S11 Fig), and by repeat families, the L2 antisense was significantly enriched in the first exon ($p = 8.6 \times 10^{-5}$) (S6 Table).

Expression from HBV

We then analyzed reads containing the HBV genome sequence in 18 HBV-positive patients. Of these samples, 13 HCCs and 10 matched non-cancerous livers expressed transcripts having the HBV genome sequence (HBV transcript) (S12 Fig). We classified transcripts having the HBV genome sequence into HBV-only and HBV-human fusion transcripts. We estimated transcriptional start and end sites and analyzed these locations on the HBV genome. Most of the TSSs of HBV transcripts were clustered in the upstream region of the PreS2 gene (around 1 bp–100 bp in the HBV reference genome) (Figs 4A and S12) as shown in a previous study using the CAGE method [44]. We then compared the expression levels of the transcripts with the HBV sequences. Expression levels were estimated by log10 converted support reads per million reads ($\log_{10}(\text{RPM})$) for each sample. The expression levels of each HBV transcript were significantly different between HCCs and the livers ($p = 0.046$) (Fig 4B); however, the comparison of the total amounts of HBV transcript was not (S13A Fig), suggesting that the transcript-level analysis with long-reads can detect significantly differentiated transcripts from HBV.

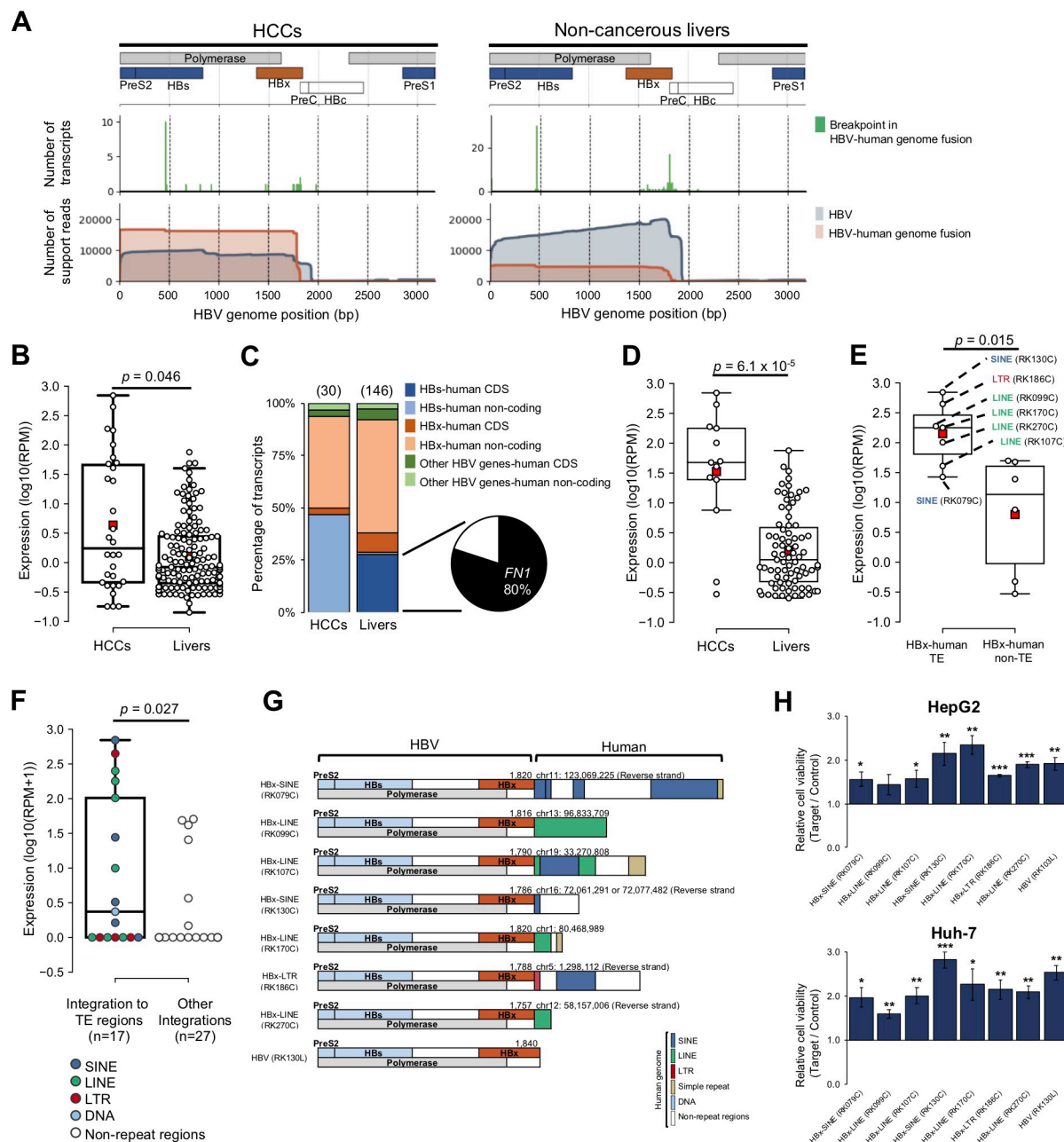


Fig 4. Identification of HBV-human genome fusion transcripts. (A) Visualization of the RNA-seq coverage in the HBV genome region in HCCs and non-cancerous livers. (Upper panels) Schematic diagram of the HBV genome structure. (Middle panels) Bar chart of the breakpoints in HBV-human fusion transcripts (HBV alone) and HBV-human fusion transcripts are shown in blue and red. The y-axis shows the total number of support reads (raw counts). (Lower panels) Coverage of HBV transcripts (HBV alone) and HBV-human fusion transcripts are shown in blue and red. The y-axis shows the total number of support reads (raw counts). (B) Expression levels of HBV-human fusion transcripts in HCCs and livers. Transcript abundance was measured in reads per million reads (RPM), and log₁₀ converted values for RPM+1 are shown in the boxplot. *P*-values were calculated by the Wilcoxon rank-sum test. (C) Percentage of HBV-human fusion transcripts in HCCs and livers. The stacked bar chart shows the distribution (%) of detected transcripts by type. The total number of transcripts is shown above the bars. (D) Expression levels of HBs-FN1 fusion transcripts in HBs-human CDS in the livers. (E) Expression levels of HBx-human non-coding fusion transcripts in HCCs and livers. *P*-values were calculated by the Wilcoxon rank-sum test. (F) Expression levels of HBx-human TE fusion and HBx human non-TE fusion transcripts in HCCs. *P*-values were calculated by the Wilcoxon rank-sum test. (G) Expression levels of HBV transcripts integrated into TE regions and those integrated into non-TE regions. *P*-values were calculated by the Wilcoxon rank-sum test. (H) Schematic diagram of HBV-human TE fusion transcripts. (I) Effect of HBx-TE fusion on the cell proliferation of HepG2 and Huh-7 cell lines assessed by CCK-8. Error bars represent the standard error of the mean. *P*-values were calculated by the one-sample t-test (**p* < 0.1, ***p* < 0.05, ****p* < 0.01) (*n* = 3).

<https://doi.org/10.1371/journal.pgen.1010342.g004>

Thirty and 146 HBV-human genome fusion transcripts were detected in the HCCs and matched livers (S8 Data), and all were selectively expressed in the respective tissues. Break-points of HBV-human fusion transcripts were clustered in the HBs gene region (450 bp– 470 bp) and downstream of the HBx gene region (1750 bp– 1840 bp) (middle panel of Fig 4A). Therefore, we focused on HBs and HBx genes and classified HBV-human fusion transcripts into “HBs-human CDS fusion”, “HBs-human non-coding fusion”, “HBx-human CDS fusion”, “HBx-human non-coding fusion”, “Other HBV genes-human CDS fusion”, and “Other HBV genes-human non-coding fusion” (Fig 4C). Forty HBs-human CDS fusion transcripts were detected in the livers but not in the HCCs (Fig 4C). Among them, 32 (80.0%) were HBs-*FN1* fusion transcripts (Fig 4C) [45]. Expression levels were then compared between the HCCs and matched livers in these categories (Figs 4D and S13A). HBx-human non-coding fusion showed higher expression levels in the HCCs than in the livers ($p = 6.1 \times 10^{-5}$) (Fig 4D); however, other categories did not show significant differences (S13B Fig). Previous studies reported that LINE1(L1), a retrotransposon, forms the HBx-L1 fusion transcript and induces carcinogenesis in HCCs [46,47]. We thus focused next on HBx-human TE fusion and identified 4 HBx-LINE fusion transcripts, 2 HBx-SINE fusion transcripts, and 1 HBx-LTR fusion transcript in HCCs (S13C Fig). The expression level of HBx-human TE fusion was significantly higher than that of HBx-human non-TE fusion in HCCs ($p = 0.015$) (Fig 4E).

The expression pattern of HBx-human TE fusions suggests their important roles in carcinogenesis. Therefore, we examined a possible mechanism for the promotion of expression and their functional effect on cell proliferation. Our samples in this study have been used in a previous WGS study, and genomic locations of HBV integrations have been determined [4]. The WGS identified 44 HBV integration sites, and 17 of them were integrated into TE regions. A comparison of their expression levels showed that integrations into the TE regions result in a higher expression than into the non-TE regions ($p = 0.027$) (Fig 4F). We then analyzed the role of HBx-human TE fusions on cell proliferation with an overexpression experiment of full-length HBx-TE fusions and canonical HBx in liver cancer cell lines (HepG2 and Huh-7) (Figs 4G and S9). We found that HBx-TE fusions significantly increased cell growth in both cells (Fig 4H).

Fusion genes

We detected 164 cancer-specific fusion transcripts (S9 Data). Although recurrent fusion events of the same gene pairs were not detected, six genes, *SENp6*, *TBRG1*, *SDC2*, *ABCD3*, *LGSN*, and *LDLR*, appeared recurrently but fused with different partners (Fig 5A). We experimentally validated all candidates (12/12) by RT-PCR (Fig 5B and S7 Table). Within the samples of the current study, 15 have been subjected to fusion gene detection by short-reads in the ICGC pan-cancer study [24] (S14A Fig). A comparison between the short- and long-reads revealed that 41 out of 50 fusions (82.0%) were not detected in the short-reads and only 9 fusions (8.0%) were commonly detected (Fig 5C). We considered three explanations for the low overlap. (1) Since most of the fusion genes missed by short-reads had very low expression levels (S14B Fig), many fusion genes with low expression levels were missed by the single approach. (2) In addition, 28.9% of transcripts in long-reads lacked the 5' region (S5 and S14C and S14D Figs). Therefore, fusion genes whose breakpoints are located in the 5' region are difficult to detect using long-reads. (3) Since artificially truncated cDNA can be produced by internal poly-A priming without priming to poly-A at the 3' end of the mRNA, fusion genes with breakpoints in the 3' region may also be missed. This comparison suggests that a combination of short- and long-reads can detect more fusion genes.

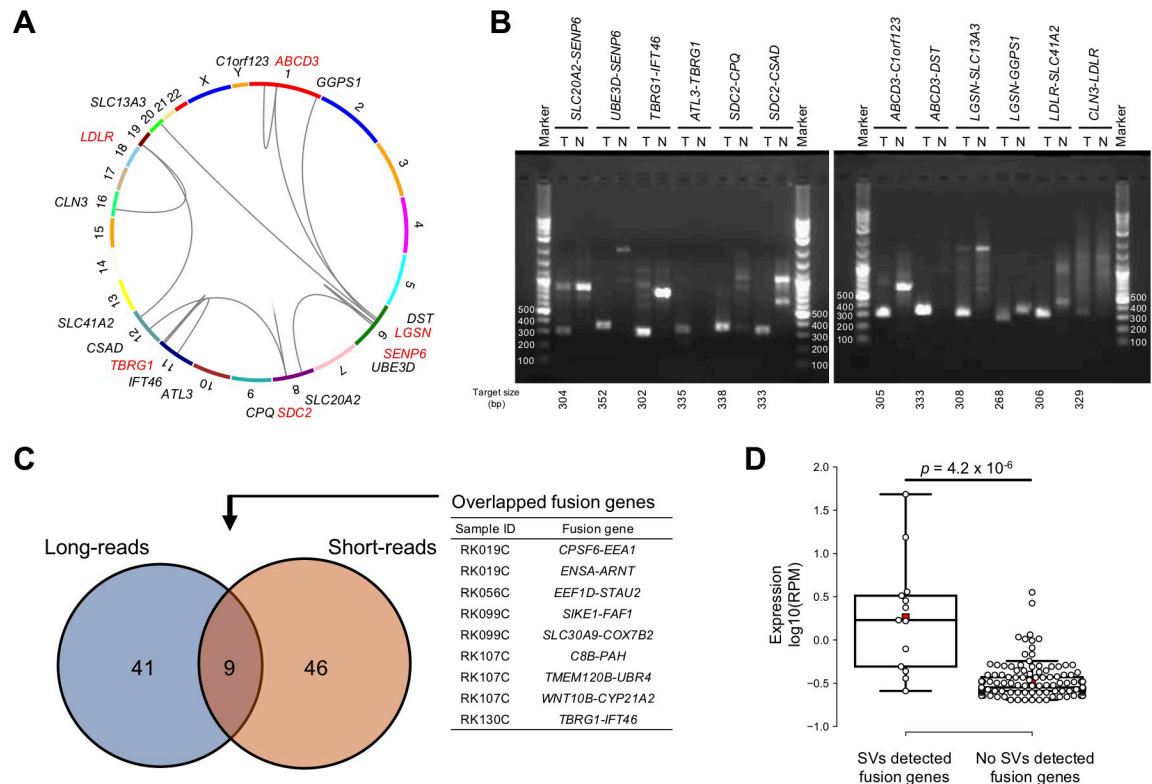


Fig 5. Identification of fusion genes. (A) Circos plot of recurrent fusion genes. The genes shown in red are recurrent fusion genes. (B) PCR validation of the recurrent fusion genes. The results of an electrophoresis of each fusion gene in cancer (T) and matched non-cancerous livers (N) are shown. (C) Venn diagram of fusion genes detected by long- and short-reads sequencers. We compared 15 HCC samples in which short-reads RNA-seq was analyzed in the ICGC study [24]. (D) Expression levels between fusion genes supported and not supported by SVs according to WGS. *P*-values were calculated by the Wilcoxon rank-sum test.

<https://doi.org/10.1371/journal.pgen.1010342.g005>

We also compared the fusion genes and structural variations (SVs) detected by previous WGS [4,23]. Among the 164 fusion transcripts, 21 had SVs that can cause the fusion transcripts, and their expression levels were significantly higher than other fusions ($p = 4.2 \times 10^{-6}$) (Fig 5D). This result suggests that there are many sub-clonal fusion genes, and their expression levels are lower than clonal fusion genes. Although the functional importance of sub-clonal fusion genes is currently unknown, deeper RNA-seq would detect more of them.

Expression profiles of HBV- and HCV-infected HCCs

Our non-cancerous liver samples have hepatitis by infection with HBV or HCV. To examine the influence of HBV and HCV on hepatitis and HCC, we performed a clustering analysis for 16 HBV-infected, 24 HCV-infected, and 2 dually infected samples in HCCs and non-cancerous livers, respectively (Fig 6A). In the non-cancerous livers, most HBV- and HCV-infected tissues were in separate clusters, whereas the HCCs were not separated (Fig 6A). Next, we analyzed genes with significantly differentiated transcripts between HCV and HBV samples. In the HCV- and HBV-infected livers, 284 and 68 transcripts were significantly up-regulated, respectively (Fig 6B and 6C). Highly expressed transcripts in the HCV-infected livers were significantly enriched in immune system-related pathways, especially in the interferon signaling pathway (Fig 6D and S8 Table). Transcripts up-regulated in the HBV-infected livers had no significantly enriched pathways. For the HCV- and HBV-infected HCCs, we performed the same analysis and found that immune system-related pathways were slightly enriched in DETs

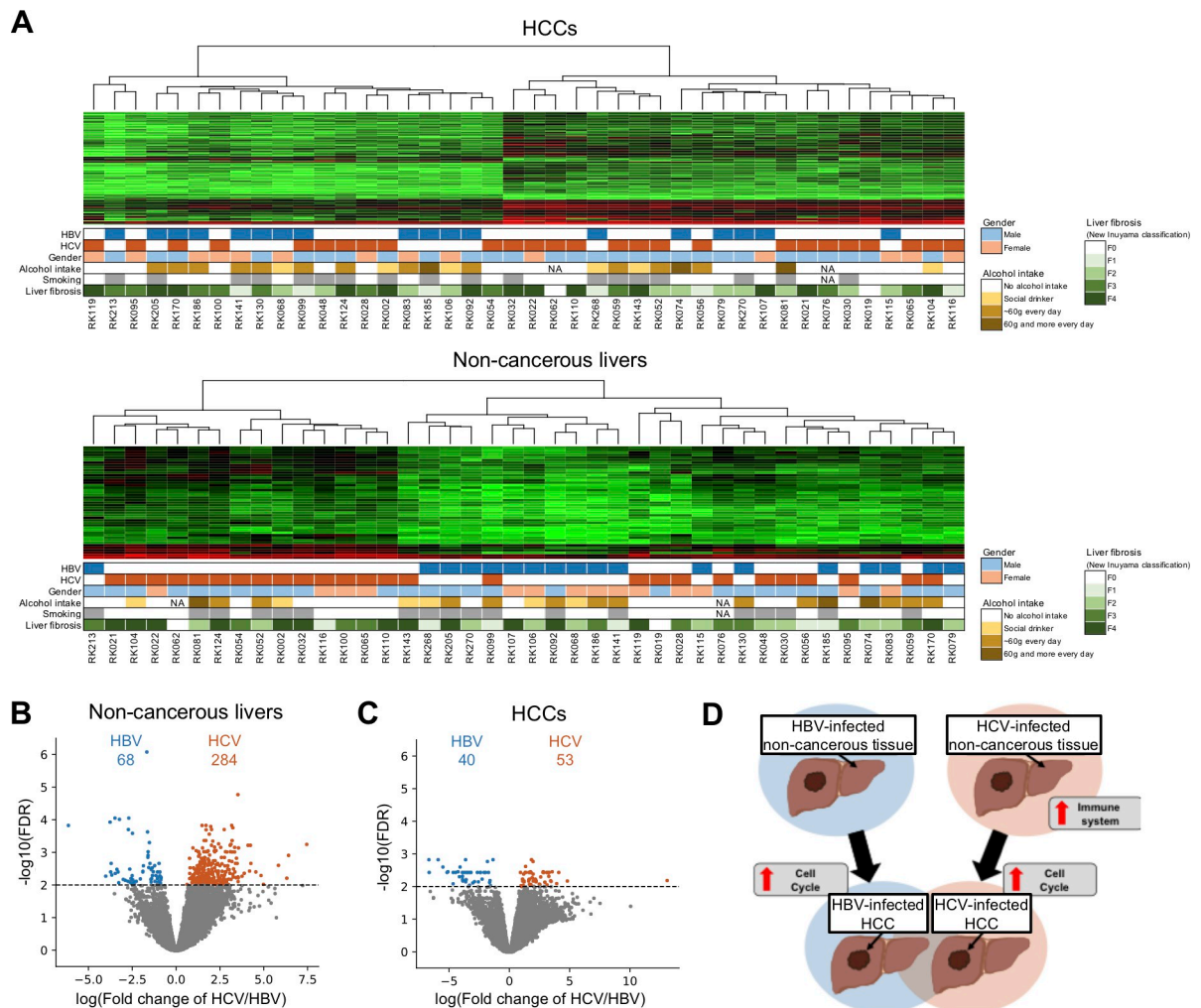


Fig 6. Comparison of expression profiles of HBV- and HCV-infected HCCs. (A) Clustering by the transcriptome of 16 HBV-infected tissues, 24 HCV-infected tissues, and 2 dually infected tissues in HCCs (upper) and in matched non-cancerous livers (lower). Clinical and pathological information of the samples is shown in the lower panels. (B, C) Volcano plots of DETs in HBV- and HCV-infected tissues. Light blue and orange dots represent transcripts that were significantly up-regulated in HBV- and HCV-infected tissues with FDR < 0.01. (B) Comparison in non-cancerous liver samples. (C) Comparison in HCC samples. (D) Schematic diagram of the expression profiles of HBV- and HCV-infected tissues.

<https://doi.org/10.1371/journal.pgen.1010342.g006>

only in the HCV-infected HCCs, but not as pronounced as in the non-cancerous livers (Fig 6C and S8 and S9 Tables). Genes of cell cycle-related pathways were significantly up-regulated in both HBV- and HCV-infected HCCs, as in all HCCs (Fig 6D and S5 and S10 and S11 Data).

Discussion

In this study, we used a long-reads sequencer (Oxford Nanopore) for a comprehensive analysis of transcriptional abnormalities in HCCs and non-cancerous livers. We first developed an analysis pipeline (SPlice) to analyze full-length transcripts (Fig 1A and 1B). SPlice removed possible errors around splicing junction sites and pseudo-fusion transcripts by considering alignment and mapping errors (Fig 1B). To our knowledge, SPlice is the only long-reads RNA-seq pipeline that can detect both splicing variants and fusion genes. While most long-reads RNA-seq analysis pipelines for detecting splicing variants use short-reads RNA-seq

data for the alignment correction of splice junctions [21,31,34], SPLICE considers error rates around splicing sites (S3C Fig) and can obtain accurate results using only long-reads RNA-seq data (S3 Fig). SPLICE also identified a larger number of fusion genes than existing software (S4 Fig). A comparison with short-reads showed high concordances in the estimation of gene expression levels, suggesting that our analysis was reliable (S7 Fig).

In the analysis of HCCs and matched livers, 15.7% and 13.2%, respectively, of transcripts were not found in the database and considered novel (Fig 1D). The expression levels of the novel transcripts were significantly lower than those of known transcripts, and deeper sequencing is expected to detect a larger number of novel transcripts (Fig 1E and 1F). Our analysis identified novel exons, the majority of which were first and last exons (Fig 2A). The novel first exons can be produced by TSSs from unknown alternative promoters and may contribute to tissue-specific or disease-associated mRNA regulation. Novel last exons can cause new UTRs and may affect the stability of mRNA. Since the average conservation scores of novel exons were lower than those of known exons, most novel exons are assumed to not have important functions (Fig 2D). However, some novel exons were highly conserved and could have essential functions for biological activities (Fig 2D and 2E). In the *MYT1L* gene, for example, a highly-conserved novel exon was detected (Fig 2E), and only a transcript with the novel exon was expressed. This transcript was significantly down-regulated in HCCs (S15 and S16 Figs). Previous studies reported that the expression of *MYT1L* is neuron-specific [48] and has been suggested to play a critical role in neuronal differentiation and maintenance [48,49]. Our analysis identified a new transcript with the novel exon of this gene in the liver. Several studies suggested that *MYT1L* suppresses the proliferation of glioblastoma, and this transcript could have tumor-suppressive effects in HCCs as well [50,51].

This study also identified the exonization of TEs. Interestingly, 30–50% of novel exons in protein-coding genes were derived from TEs (Fig 2B), suggesting that TEs can be a source of the functional diversity of genes [52]. Strands of the exonized TEs were not random, and patterns were different among the location of the exons (Fig 2C); first and middle exons had higher percentages of TEs encoded on opposite (antisense) strands, whereas the opposite pattern was observed in TEs of the last exons (Fig 2C). A previous study suggested that the antisense of LINE1 has a promoter function and can start transcription [43,53]. This observation can partly explain the strand bias of TEs of first exons. The strand bias of TEs of the middle exon can be caused by the asymmetric distribution of TEs in introns. TEs on sense strands generate harmful poly-A signals in introns, resulting in the reduction of TEs in sense strands in introns [54,55]. *Alu* has polyadenylation signals in the sense strand that can generate the last exons, which can explain the strand bias of *Alu* of the last exon (Fig 2C) [56].

A comparison of expression levels between HCCs and non-cancerous livers showed the differentiation of several transcripts with TEs. A transcript with a TE-derived exon from an important oncogene, *MET*, (L1-*MET*) was significantly overexpressed in HCCs (Fig 3F and 3G). L1-*MET* was reported to be involved in the progression of colorectal cancer and bladder cancer [43,57], and this transcript also promoted the proliferation of HCCs (Fig 3H and 3I). Since L1-*MET* is suggested to be regulated by the L1 antisense promoter (Fig 3F) [43,53,57], its expression is independent of other *MET* transcripts. Indeed, 6 transcripts were observed in *MET*; however, only L1-*MET* was highly overexpressed in HCCs (Fig 3G). Another oncogenic TE fusion event was found in the *HRH1* gene (L2-*HRH1*). Our study identified 3 transcripts from the *HRH1* gene; however, only L2-*HRH1* (NM_001098211.1) was significantly upregulated in HCCs, and the overexpression of L2-*HRH1* promoted cell proliferation (S17 Fig). The *HRH1* gene is known to promote the progression of HCCs [58], and our finding of the fusion events suggests an activation mechanism of the *HRH1* gene. Additionally, TE fusion transcripts were identified in HBV-infected HCCs (HBx-TE fusion) (Fig 4G). The overexpression

of HBx-TE fusion increased cell proliferation in the cell lines (Fig 4H), suggesting that HBx-TE has a similar function as canonical HBx [59]. Since the expression levels of the integrated HBV to the TE regions were significantly higher than those of the non-TE regions (Fig 4F), integrations into TE regions may cause the constitutive expression of HBV genes via TE promoters. These results strongly suggest the important role of TEs on gene expression. Most TEs are repressed by several inactivation mechanisms; however, epigenetic changes in cancer cells can activate TEs as promoters, resulting in the aberrant regulation of oncogenes [60].

Most RNA-seq studies identified DEGs but far fewer DETs. In this study, we detected 9,933 DETs (Fig 3A). Seven hundred forty-six genes with DETs were not detected by the gene-level analysis (DET-specific genes), indicating that the analysis of transcripts benefits the detection of novel driver genes (Fig 3B). Our analysis also showed possible mechanisms for regulating splicing variants. DET-specific genes had a higher number of transcripts per gene and a higher percentage of transcripts with multiple TSSs (Fig 3C and 3D). This observation suggests that multiple transcripts regulated by different promoters canceled each other out, and thus expression differences were not detected in the gene-level analysis. We also found 80 genes with both significantly up-regulated and down-regulated transcripts (BiExp genes) (S7 Data). In addition to the previously reported *AFMID* gene [15], genes related to aberrant cell proliferation (*CEACAM1*, *CASP8*, *LDHB*, *SPINT2*) [61–64] were detected as BiExp genes (S7 Data). Each transcript in BiExp genes is regulated differently and have different functions.

We then compared the expression level of transcripts in HCCs and non-cancerous livers upon HBV or HCV infection (Fig 6). A clustering analysis did not separate HBV- and HCV-infected HCCs, whereas the livers were clustered into three groups: HCV cluster, HBV cluster, and cluster with both types (Fig 6A). This observation suggests that the difference in virology mainly affects the phenotype of the livers (Fig 6D). A pathway analysis for genes differentiated between HCV- and HBV-infected livers showed an enrichment of immune-related genes (S8 Table). This result is consistent with previous studies [65,66] and suggests that HBV and HCV may induce different expression changes but that similar mechanisms work after cancer development (Fig 6A and S8 and S9 Table and S10 and S11 Data).

Although we obtained many interesting discoveries by analyzing full-length transcripts, our study has limitations that should be addressed in future studies. First, our analysis did not determine the structure of the entire CDS for 38.3% of the mapped reads. The majority of the incomplete transcripts lacked 5' regions (S5B and S5C Fig), which can be explained by the process of cDNA synthesis and degradation of RNA samples. cDNA synthesis starts from the 3' end and, therefore, the inactivation of reverse transcriptase or degradation of RNA samples results in cDNA lacking the 5' end (S6 Fig). This mechanism can also affect fusion-gene detection (Figs 5 and S14C and S14D). Improvements in cDNA synthesis and the use of high-quality RNA can increase the rate of full-length transcripts. Second, the accuracy of the analysis depends on the sequencing error rate. Although several filters were applied to currently available high-error reads (Figs 1B and S1), sequencing errors can still affect the accuracy of the result. With improvements in sequencing technologies and basecallers, the identification of larger numbers of splicing changes with high accuracy should be possible in the future (S10 Table). Third, our study identified a higher expression of HBx-human TE transcripts that promoted cell proliferation (Fig 4G and 4H). However, the detailed biological mechanisms of the transcripts are not clear. A previous study reported that HBx-LINE1 can change the expression levels of epithelial-mesenchymal transition markers [47]; however, our analysis did not detect significant changes in these genes (S18 Fig). Since the HBx-human transcripts may, at least in part, explain the carcinogenesis of liver cancer, further functional analysis is needed.

To conclude, here we developed the SPLICE pipeline for transcriptome analysis using a long-reads sequencer. Validation with the MCF-7 cell line and PCR showed the reliability of

our analysis. Using this method, we analyzed the full-length transcript of HCCs. We detected 746 genes with DETs, which were otherwise difficult to detect by short-reads sequencers. We also detected chimeras of TEs and cancer-related genes and found that they promote the proliferation of HCC cell lines. In the analysis of fusion genes, we found 6 recurrent genes that fused with other genes. Additionally, by comparing HBV- with HCV-infected livers, a significant difference in immune-related gene expression levels was detected. Our results strongly suggest that the direct observation of transcripts with long-reads sequencing contributes to understanding the true picture of transcript aberration in cancer.

Methods

Ethics statement

All subjects gave informed consent to participate in the study following International Cancer Genome Consortium (ICGC) guidelines. Institutional review boards at Institute of Physical and Chemical Research (RIKEN) and the University of Tokyo, and all groups participating in this study approved this work (RIKEN; H20-11(16), the University of Tokyo; 2020097G). All participants have provided written consent to participate in the study.

Samples

This study was conducted following the principles of the Declaration of Helsinki. RNA from HCCs and adjacent non-cancerous livers from 42 patients was used for this study. These samples have been used for short-reads sequencing and reported by several previous studies ([S1 Data](#)) [4,23,24], and a list of somatic mutations (point mutations, short indels and structural variations), gene expression levels, and fusion genes by short-reads analysis are available [4,24]. RNA sample qualities were evaluated with RIN (RNA Integrity Number) values using Bioanalyzer (Agilent), and samples with RIN values ≥ 8 were selected. RNA from 16 HBV-infected HCCs, 24 HCV-infected HCCs, 2 dually infected HCCs, and their matched non-cancerous livers were sequenced.

Library preparation and cDNA sequencing

Full-length cDNA was synthesized from total RNA (1 μ g) using the SMARTer PCR cDNA Synthesis kit (Clontech) according to the manufacturer's instructions. For primer digestion, cDNA samples were treated with exonuclease I (NEB) at 37°C for 30 min. After purification with Agencourt AMPure XP magnetic beads (Beckman Coulter), libraries were constructed using the Ligation Sequencing Kit (SQK-LSK109) (Oxford Nanopore) according to the manufacturer's protocol. Libraries were sequenced on a SpotON FlowCell MKI(R9.4) (Oxford Nanopore) using the MinION sequencer (Oxford Nanopore). Basecalling from raw data (FAST5 format) was performed using Guppy software (version 3.0.3) (Oxford Nanopore).

Analysis pipeline

We constructed an analysis pipeline named SPLICE ([Fig 1A](#)). Low-quality reads (average base quality < 15) were filtered out. Reads were mapped to the reference genome sequence (hg38), the reference transcriptome sequence (GENCODE (version 28), and RefSeq (release 88)) using minimap2 software (v2.17) [27]. For reads mapped to the reference genome sequence, a ≥ 60 -bp unmapped region (soft-clipping regions) was re-mapped to the reference genome sequence (hg38). Since sequencing errors near splicing junctions can lead to errors in predictions of the splicing sites, we compared the mapping results to the reference genome and the

reference transcriptome sequence and removed reads if the number of matches to the reference genome sequence was smaller than that to the reference transcriptome sequence.

In the present study, we analyzed sequence data from MinION, but our method is applicable to data from PromethION and flongle as well.

Identification of known transcripts

The genomic locations of known transcripts were obtained from GENCODE (version 28) and RefSeq (release 88), and transcripts that were commonly registered in both annotations were unified to GENCODE transcript IDs (Fig 1A). We assigned each exon in the transcripts to the exons in the database if the positions of the splicing junction sites at both ends of the transcript's exon and reference exon were perfectly matched. We then considered the combination of the exons and transcript ID in the database were assigned. Transcripts that were not found in the database were classified as novel. For the annotation, we did not consider UTR lengths for the classification, because the qualities of the ends of the reads were low. Reads that contained all coding exons of transcripts in GENCODE or RefSeq were classified as "full-length CDS". When reads had a part of the exons of a known transcript but did not cover all the exons, they were classified as "partial-length CDS". Among partial-length CDSs, if the reference transcripts had two or more exons and reads had only one exon, we removed the candidates as highly truncated reads. For partial-length CDSs reads annotated to multiple known transcripts, we annotated them to the longest known transcript among matched known transcripts. After the annotation, we removed transcripts if the percentage of the expression level of a transcript was less than 1% the total expression of the gene.

Identification of novel transcripts

Reads that were not annotated as known transcripts were classified as "novel exon length", "novel exon combination", or "novel exon" (Fig 1C). "Novel exon length" reads contain known exons with different lengths from the reference database. As described above, a change of splicing junction sites ≥ 5 -bp was considered a length change. If splicing junction sites changed < 20 bp for a known splicing junction site, we evaluated the mismatch rate within ± 5 bp regions from splicing junction sites and removed the candidates for error rates $\geq 20\%$. Novel exons were defined as expressed regions that were not overlapped with known exons. Novel exon candidates were removed if they were expressed only in a single novel exon. We removed candidates if the percentage of their transcript expression level was $< 1\%$ the total expression level of the gene.

Identification of fusion genes

Reads mapped to two protein-coding genes were considered fusion gene candidates. Possible errors in fusion gene candidates were filtered out as follow. (1) Because mapping errors due to highly homologous genomic regions can cause fusion gene candidates, we compared the results of mapping to the reference genome and the reference transcriptome sequences and removed candidates if both were inconsistent (Fig 1B). (2) Because fusion gene candidates can be generated by mis-ligation of two transcripts during the library preparation, we removed candidates that contained primer sequences for cDNA synthesis between two genes (Fig 1B). (3) Because artificial chimeric reads may be generated from highly expressed genes, there is a higher chance of mis-ligation in the library preparation; thus, we removed candidates if their percentage of transcript expression level was $< 1\%$ the average expression level of both genes. (4) Candidates with low expression levels (support reads < 3) in two or more samples in the matched non-cancerous livers were removed as artifacts. (5) Candidates were removed as

read-through transcripts if the two transcripts were the same strand and were in close proximity to each other in the genome ($< 200,000$ bp).

We further compared fusion genes with SVs detected by previous WGS [4,23]. If the break-points of an SV were located in intron or exon regions of the fusion gene, the SV was considered a possible cause of the fusion gene.

Benchmarking

To evaluate the performance of SPLICE, we compared it to four widely adopted long-reads RNA-seq analysis pipelines: TALON [34], FLAIR [21], StringTie [31], and Bambu [29]. For this comparison, we used MCF-7 and HCC (RK107C) sequence data. We randomly extracted 1 M reads from the MCF-7 and HCC (RK107C) sequence data using Seqtk (v1.3) (params: sample -s1 1000000). Reads were mapped to the reference genome sequence (hg38) with minimap2 (v2.17) (params: -ax splice—MD), and the output SAM files were converted to BAM files and sorted with samtools (v1.7) [67]. GENCODE v28 and RefSeq release 88 annotation databases were used in all data processes.

For the benchmarking of TALON (v5.0), we corrected aligned reads with TranscriptClean (v2.0.3) [68]. Next, we ran the *talon_label_reads* module to flagging reads for internal priming (params:—ar 20). The TALON database was initialized by running the *talon_initialize_database* module (params:—l o —5p 500 —3p 300). Then we ran the *talon* module to annotate the reads (params:—cov 0.8—identity 0.8). To output transcript abundance, we first obtained a whitelist using the *talon_filter_transcripts* module (params:—maxFracA 0.5—minCount 5) and then performed transcript quantification using the *talon_abundance* module based on the whitelist.

For the benchmarking of FLAIR (v1.5), the sorted BAM file was converted to BED12 using *bin/bam2Bed12.py*. Then we corrected misaligned splice sites with the *flair-correct* module. High-confidence isoforms were defined from the corrected reads using the *flair-collapse* module (params: -s 3—generate_map).

For the benchmarking of StringTie (v2.2.1), Stringtie was performed with input files consisting of long-reads alignments and reference annotations (params: -L -c 3).

For the benchmarking of Bambu (v2.0.0), Bambu was performed with input files consisting of long-reads alignments, reference annotations, and the reference genome (hg38) (params: min.readCount = 3).

Candidates with low expression levels (support reads < 3), antisense, and intergenic transcripts were excluded from the comparison.

Assessing the performance of fusion gene detection in SPLICE and other tools

To compare the fusion genes detected by SPLICE, LongGF [32] and JAFFAL [30], we used 1 M reads randomly extracted from the MCF-7 and HCC (RK107C) sequence data as described above.

For the benchmarking of LongGF (v0.1.2), reads were mapped to the reference genome sequence (hg38) with minimap2 (v2.17) (params: -ax splice—MD), and the output SAM files were converted to BAM files and sorted with samtools (v1.7). We then ran the *longgf* module and obtained the list of fusion genes (params: min-overlap-len 100 bin_size 50 min-map-len 200 pseudogene 0 secondary_alignment 0 min_sup_read 3).

For the benchmarking of JAFFAL (v2.2), we ran the *JAFFAL.groovy* module with zipped fastq files.

In this comparison, close gene pairs (< 200,000 bp) were not considered, because they cannot be distinguished from read-through transcripts [69] without information on SVs.

Permutation test

We performed a permutation test to examine the skew of the distribution of repeat types and strands. Random sequences corresponding to novel exon lengths were extracted from within the gene regions, and overlapping repeat types and their strands were obtained. This procedure was repeated 1,000,000 times, and the distributions of repeat types and their strands were used as a null distribution. The adjustment of the multiple testing was performed using Bonferroni's correction.

Comparison of gene expressions between HCCs and non-cancerous livers

To compare expression levels in HCCs and matched non-cancerous livers, we used edgeR software [37]. The number of mapped reads for each transcript was normalized using the trimmed mean of M value (TMM) normalization method [40]. Differential expression analysis was performed using the quasi-likelihood method [70]. The Benjamini-Hochberg method was used for multiple testing correction [38].

Clustering analysis

To compare the expression profiles in HBV- and HCV-infected tissues, we applied k-means clustering to the RNA-seq data. For the log10 converted value of TMM normalized RPM+1, we filtered out candidates with a variance < 0.1. We performed k-means clustering and then excluded candidates that were not involved in the clustering by the variance analysis (p-value < 1×10^{-5}).

Cell lines

HepG2 and Huh-7 cell lines were purchased from JCRB Cell Bank. Huh-7 was maintained in Dulbecco's modified Eagle's medium (DMEM; Nacalai tesque), and HepG2 was maintained in DMEM with high glucose (Nacalai tesque). All culture media contained 10% fetal bovine serum (FBS; Gibco) and 50 units/50 µg/mL penicillin-streptomycin (Nacalai tesque).

Gene overexpression

Cancer-related splicing variants were overexpressed in HepG2 and Huh-7 cell lines (S11 Table). The expression vectors were constructed from the pIRES2-AcGFP1-Nuc vector (Clontech). The target sequence was amplified with cloning primers from cDNA (S11 Table) and inserted into the EcoR1 site of the pIRES2-AcGFP1-Nuc vector using the In-Fusion HD Cloning Kit (Clontech) according to the manufacturer's protocol. The plasmid was amplified in *E. coli* DH5α Competent Cells (Takara Bio) and purified using the QIAGEN Plasmid Mini Kit (QIAGEN).

The cells were seeded in 24-well and 96-well plates at a concentration of 1×10^4 /mL from 60% confluent cells. After 48 h of incubation, the expression vectors were transfected into the cells using Lipofectamine 3000 (Invitrogen) according to the manufacturer's protocol. An empty vector of pIRES2-AcGFP1-Nuc was used as control. The medium was changed 2 h after the transfection. For the expression analysis, cells were cultured in 48-well plates, and RNA was extracted 48 h after the transfection. For the cell proliferation assay, cells were cultured in 96-well plates, and cell numbers were measured using the CCK-8 kit (DOJINDO) 96 h after the transfection according to the manufacturer's protocol.

Gene silencing

To investigate the effect of *L1-MET* on cell proliferation, *L1-MET* expression was suppressed in the HepG2 and Huh-7 cell lines. Cells were seeded in 24-well and 96-well plates at a concentration of 1×10^5 cells/mL. After 24 h of incubation, siRNA ([S12 Table](#)) was transfected into the cells using Lipofectamin RNAiMAX Transfection Reagent (ThermoFisher SCIENTIFIC) according to the manufacturer's protocol. MISSION siRNA Universal Negative Control #1 (Sigma-Aldrich) was used as a negative control for the siRNA transfection. RNA was extracted after 48 h incubation from the 24-well plates, and the cell proliferation assay was performed for cells in the 96-well plates using CCK-8 (DOJINDO) after 96 h of incubation according to the manufacturer's protocol.

Supporting information

S1 Fig. Error patterns in the Nanopore RNA-seq analysis and filtering of SPLICE. (PDF)

S2 Fig. Correlation of expression levels between long-reads and short-reads RNA-seq in MCF-7. Transcript abundance was measured in reads per million mapped reads (RPM) for long-read RNA-seq data and fragments per kb of exon per million fragments mapped (FPKM) for short-reads RNA-seq data. log10 converted values of RPM+1 and FPKM+1 are shown. *r*, Pearson's correlation coefficient; *p*, *p*-value. (PDF)

S3 Fig. Performance of SPLICE pipeline in splicing variant detection. (A) Number of transcripts detected by SPLICE and four other methods (TALON, FLAIR, StringTie and Bambu). (B) Percentage of transcripts supported in PacBio data with SPLICE and the other methods. (C) Average alignment error rate of SPLICE and the other methods for novel splice sites. The gray dotted line indicates the location of the splice site. (PDF)

S4 Fig. Performance of SPLICE pipeline in fusion gene detection. (A) Venn diagram of the fusion genes detected by SPLICE and two other methods (LongGF and JAFFAL). The neighboring gene pairs (< 200,000 bp) detected by LongGF (MCF-7: VMP1-RPS6KB1, GTF2IRD1-GFT2I and PITPNC1-BPTF, HCC (RK107C): C11orf52-DIXDC1) were excluded from the list because they cannot be distinguished from read-through transcripts without information on structural variations. (B) RT-PCR validation of fusion genes detected by SPLICE, LongGF, and JAFFAL. (PDF)

S5 Fig. Percentage of reads mapped to full-length CDS regions. (A) Schematic diagram of the read length types. "Entire CDS" reads contain all coding exons of known transcripts. "Lacking 3' region of CDS" reads lack coding exons in the 5' region of known transcripts. "Lacking 5' region of CDS" reads lack the coding exons in the 3' region of known transcripts. "Lacking both ends of CDS" reads lack both the coding exon in the 5' region and the 3' region of known transcripts. (B) Pie chart of the detected transcripts by read length type. (C) Relationship between the reference gene length and percentage of read length type in the mapped reads. (Upper panel) Distribution of mapped reads to the reference gene length. (lower panel) Percentage of the mapped read length type to the reference gene length. (PDF)

S6 Fig. Correlation between RNA integrity number (RIN) and percentage of "Entire CDS" reads in mapped reads. *r*, Pearson's correlation coefficient; *p*, *p*-value.

(PDF)

S7 Fig. Expression levels of long-reads and short-reads RNA-seq in HCC clinical samples. Transcript abundance was measured in RPM and FPKM for long-reads and short-reads RNA-seq data. log10 converted values for RPM+1 and FPKM+1 are shown. *r*, Pearson's correlation coefficient. $p < 2.2 \times 10^{-308}$ for all samples.

(PDF)

S8 Fig. Differential expression levels and structure of AFMID transcripts. (A) Significantly up- and down-regulated AFMID transcripts in HCCs are shown in red and blue, respectively. (B) Structure of AFMID transcripts.

(PDF)

S9 Fig. RT-PCR validation of the expression of target transcripts in HepG2 and Huh-7 cell lines. Target transcript size bands were confirmed by agarose gel electrophoresis. Cell lines transfected with empty vectors were used as control (Ctrl).

(PDF)

S10 Fig. Effect of L1-MET knockdown on proliferation of HepG2 and Huh-7 cell lines. (A) Schematic diagram of the target sequence for L1-MET-specific knockdown. (B) Expression levels of L1-MET in L1-MET knockdown in both cell lines.

(PDF)

S11 Fig. Enrichment analysis of TE-derived exons in HCC-specific transcripts and non-cancerous liver-specific transcripts. (A) Venn diagram of cancer-specific transcripts and non-cancerous liver specific transcripts. (B) Percentage of TE-derived exons in each item of the Venn diagram in A. *P*-values for the enrichment of TE-derived exons in cancer-specific transcripts and non-cancerous liver-specific transcripts were calculated by Fisher's exact test. Comparisons were done for TE-derived exons in cancer-specific transcripts and other transcripts and for TE-derived exons in liver-specific transcripts and other transcripts. N.S.: Not significant.

(PDF)

S12 Fig. Visualization of sequencing coverage in the HBV genome. HBV transcripts were classified as "HBV (HBV alone)" and "HBV-human genome fusion". (Upper panels) Schematic diagrams of the HBV genome structure. (Lower panels) RNA-seq coverage of HBV transcripts (HBV alone) and HBV-human genome fusion transcripts are shown in blue and red. N.D.: Not detected.

(PDF)

S13 Fig. Expression analysis of HBV transcripts. (A) Comparison of HBV transcript expression levels in cancer and matched non-cancerous livers. HBV transcripts were classified as "HBV (HBV alone)" and "HBV-human genome fusion". Total expression levels of transcripts were plotted for each sample. Statistical significances were calculated by the Wilcoxon signed-rank test (B) Expression levels of HBV-human genome fusion transcripts by each transcript type in HCCs and non-cancerous livers. Statistical significances was calculated by the Wilcoxon rank-sum test. (C) Expression levels of HBx-human TE fusion transcripts. HBx-human non-coding fusion transcripts were classified according to the repeat types of the human genome. Their expression levels are shown for each sample. (D) Expression levels of HBx-human TE fusion transcripts in HCCs and livers. Statistical significances was calculated by the

Wilcoxon rank-sum test. (E) Expression levels of HBx-human TE fusion transcripts and non-HBx human TE fusion transcripts in livers. Statistical significances was calculated by the Wilcoxon rank-sum test.

(PDF)

S14 Fig. Features of fusion transcripts. (A) Number of mapped bases between long-reads and short-reads RNA-seq. (B) Expression levels of fusion genes detected by both long-reads and short-reads RNA-seq and those detected by long-reads RNA-seq only. P-values were calculated by the Wilcoxon rank-sum test. (C) Average number of exons at the 5' and 3' sides of the fusion gene. (D) Average transcript length (bp) of the 5' and 3' sides of the fusion gene.

(PDF)

S15 Fig. Expression levels of a novel splicing variant of MYT1L in HCCs and non-cancerous livers. P-values were calculated by quasi-likelihood methods. The Benjamini-Hochberg method was used for multiple testing correction (FDR).

(PDF)

S16 Fig. A genome browser view of a novel transcript region of MYT1L identified in this study. From top, a novel transcript, reference transcripts, and two types of multiple sequence comparisons are shown.

(PDF)

S17 Fig. Structural and functional analysis of L2-HRH1. (A) Schematic diagram of the structural differences in L2-HRH1 (NM_001098211.1) and other HRH1 transcripts. (B) Bar chart showing the expression differences between HCCs and non-cancerous livers in HRH1 transcripts. P-values were calculated by quasi-likelihood methods. The Benjamini-Hochberg method was used for multiple testing correction (FDR). (C) Effect of L2-HRH1 overexpression on the proliferation of HepG2 and Huh-7 cell lines as assessed by CCK-8 expression. P-values were calculated by the one-sample t-test ($n = 3$).

(PDF)

S18 Fig. Expression levels of epithelial-mesenchymal transition markers (CDH1, JUP and FN1) in HBx-TE fusion overexpressed cell lines. P-values were calculated by the one-sample t-test (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

(PDF)

S1 Table. Summary of RNA-seq of MCF-7.

(PDF)

S2 Table. Primer sequences for validation of fusion genes, related to S3B Fig.

(PDF)

S3 Table. Ratios of TE-derived exons among novel exons and known exons.

(PDF)

S4 Table. Percentage of BiExp DET-specific and DET and DEG genes.

(PDF)

S5 Table. Ratios of TE-derived exons in cancer-specific transcripts.

(PDF)

S6 Table. Ratios of TE-derived exons in non-cancerous liver-specific transcripts.

(PDF)

S7 Table. Primer sequences for validation of fusion genes, related to Fig 5B.
(PDF)

S8 Table. Significantly enriched reactome pathways of genes with up-regulated DETs in HCV-related non-cancerous livers vs. HBV-related non-cancerous livers.
(PDF)

S9 Table. Significantly enriched reactome pathways of genes with up-regulated DETs in HCV-related HCCs vs. HBV-related HCCs.
(PDF)

S10 Table. Analysis results by SPLICE of sequences basecalled by Guppy v3.0.3 and Guppy v6.0.6.
(PDF)

S11 Table. Cloning and RT-PCR primers for overexpression experiments.
(PDF)

S12 Table. siRNA design for knockdown experiments and qPCR primers.
(PDF)

S1 Data. Clinical and pathological information of the 42 HCC patients.
(XLSX)

S2 Data. Summary of RNA-seq of HCC clinical samples.
(XLSX)

S3 Data. List of transcripts with novel exons.
(XLSX)

S4 Data. List of significantly differentially expressed transcripts.
(XLSX)

S5 Data. Significantly enriched reactome pathways of genes with up- or down-regulated DETs.
(XLSX)

S6 Data. List of significantly expressed driver gene's transcripts.
(XLSX)

S7 Data. List of bi-directionally significantly expressed genes.
(XLSX)

S8 Data. List of HBV-human genome fusion transcripts.
(XLSX)

S9 Data. List of fusion genes.
(XLSX)

S10 Data. Significantly enriched reactome pathways of genes with up-regulated DETs in HBV-infected HCCs vs HBV-infected non-cancerous livers.
(XLSX)

S11 Data. Significantly enriched reactome pathways of genes with up-regulated DETs in HCV-infected HCCs vs. HCV-infected non-cancerous livers.
(XLSX)

Acknowledgments

The super-computing resource was provided by Human Genome Center, Institute of Medical Science, the University of Tokyo. We acknowledge the staff at Laboratory for Cancer Genomics, RIKEN, for their technical assistances.

Author Contributions

Conceptualization: Akihiro Fujimoto.

Data curation: Hiroki Kiyose.

Formal analysis: Akihiro Fujimoto.

Funding acquisition: Akihiro Fujimoto.

Methodology: Hiroki Kiyose.

Project administration: Akihiro Fujimoto.

Resources: Hidewaki Nakagawa, Atsushi Ono, Hiroshi Aikata, Masaki Ueno, Shinya Hayami, Hiroki Yamaue, Kazuaki Chayama, Akihiro Fujimoto.

Software: Hiroki Kiyose, Akihiro Fujimoto.

Supervision: Akihiro Fujimoto.

Validation: Hiroki Kiyose.

Writing – original draft: Hiroki Kiyose, Akihiro Fujimoto.

Writing – review & editing: Mihoko Shimada, Jing Hao Wong, Akihiro Fujimoto.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018; 68: 394–424. <https://doi.org/10.3322/caac.21492> PMID: 30207593
2. El-Serag HB, Rudolph KL. Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. *Gastroenterology*. 2007; 132: 2557–2576. <https://doi.org/10.1053/j.gastro.2007.04.061> PMID: 17570226
3. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*. 2017; 169: 1327–1341.e23. <https://doi.org/10.1016/j.cell.2017.05.046> PMID: 28622513
4. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet*. 2016; 48: 500–509. <https://doi.org/10.1038/ng.3547> PMID: 27064257
5. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad I Ben, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet*. 2012; 44: 694–8. <https://doi.org/10.1038/ng.2256> PMID: 22561517
6. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet*. 2015; 47: 505–511. <https://doi.org/10.1038/ng.3252> PMID: 25822088
7. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, et al. High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet*. 2011; 43: 464–9. <https://doi.org/10.1038/ng.804> PMID: 21499249
8. Shiraishi Y, Fujimoto A, Furuta M, Tanaka H, Chiba K, Boroevich KA, et al. Integrated Analysis of Whole Genome and Transcriptome Sequencing Reveals Diverse Transcriptomic Aberrations Driven by Somatic Genomic Changes in Liver Cancers. Creighton C, editor. *PLoS One*. 2014; 9: e114263. <https://doi.org/10.1371/journal.pone.0114263> PMID: 25526364

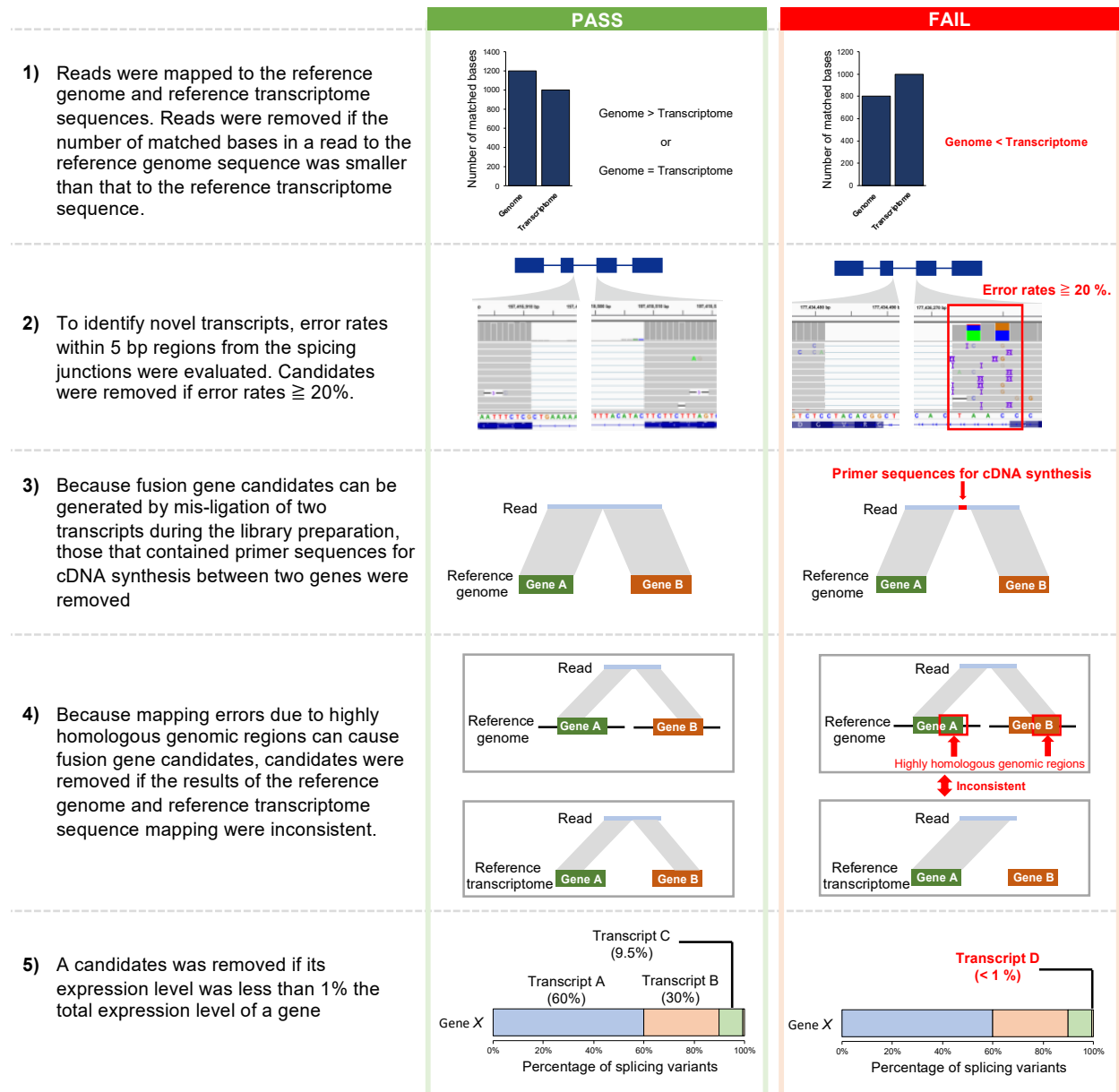
9. Vellichirammal NN, Albahrani A, Banwait JK, Mishra NK, Li Y, Roychoudhury S, et al. Pan-Cancer Analysis Reveals the Diverse Landscape of Novel Sense and Antisense Fusion Transcripts. *Mol Ther—Nucleic Acids*. 2020; 19: 1379–1398. <https://doi.org/10.1016/j.omtn.2020.01.023> PMID: 32160708
10. Alqahtani A, Khan Z, Alloghbi A, Said Ahmed TS, Ashraf M, Hammouda DM. Hepatocellular Carcinoma: Molecular Mechanisms and Targeted Therapies. *Medicina (Kaunas)*. 2019; 55. <https://doi.org/10.3390/medicina55090526> PMID: 31450841
11. Liu C-Y, Chen K-F, Chen P-J. Treatment of Liver Cancer. *Cold Spring Harb Perspect Med*. 2015; 5: a021535. <https://doi.org/10.1101/cshperspect.a021535> PMID: 26187874
12. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev*. 2010; 24: 2343–64. <https://doi.org/10.1101/gad.1973010> PMID: 21041405
13. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, et al. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol*. 2009; 16: 670–6. <https://doi.org/10.1038/nsmb.1608> PMID: 19448617
14. Venables JP. Aberrant and alternative splicing in cancer. *Cancer Res*. 2004; 64: 7647–54. <https://doi.org/10.1158/0008-5472.CAN-04-1910> PMID: 15520162
15. Lin K-T, Ma WK, Schamer J, Liu Y-R, Krainer AR. A human-specific switch of alternatively spliced AFMID isoforms contributes to TP53 mutations and tumor recurrence in hepatocellular carcinoma. *Genome Res*. 2018; 28: 275–284. <https://doi.org/10.1101/gr.227181.117> PMID: 29449409
16. Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. *J Hum Genet*. 2020; 65: 3–10. <https://doi.org/10.1038/s10038-019-0658-5> PMID: 31474751
17. Bueno R, Stawiski EW, Goldstein LD, Durinck S, De Rienzo A, Modrusan Z, et al. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nat Genet*. 2016; 48: 407–16. <https://doi.org/10.1038/ng.3520> PMID: 26928227
18. Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, et al. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. *Genome Biol*. 2021; 22: 44. <https://doi.org/10.1186/s13059-021-02261-x> PMID: 33482911
19. Lian B, Hu X, Shao Z-M. Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. *Sci Rep*. 2019; 9: 6032. <https://doi.org/10.1038/s41598-019-42184-z> PMID: 30988345
20. Oka M, Xu L, Suzuki T, Yoshikawa T, Sakamoto H, Uemura H, et al. Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer. *Genome Biol*. 2021; 22: 9. <https://doi.org/10.1186/s13059-020-02240-8> PMID: 33397462
21. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun*. 2020; 11: 1438. <https://doi.org/10.1038/s41467-020-15171-6> PMID: 32188845
22. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*. 2018; 19: 90. <https://doi.org/10.1186/s13059-018-1462-9> PMID: 30005597
23. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet*. 2012; 44: 760–4. <https://doi.org/10.1038/ng.2291> PMID: 22634756
24. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*. 2020; 578: 82–93. <https://doi.org/10.1038/s41586-020-1969-6> PMID: 32025007
25. Bolisetty MT, Rajadinakaran G, Graveley BR. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol*. 2015; 16: 204. <https://doi.org/10.1186/s13059-015-0777-z> PMID: 26420219
26. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010; 28: 511–5. <https://doi.org/10.1038/nbt.1621> PMID: 20436464
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; 34: 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> PMID: 29750242
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489: 57–74. <https://doi.org/10.1038/nature11247> PMID: 22955616
29. Chen Y, Sim A, Wan Y, Goeke J. 2022. bambu: Reference-guided isoform reconstruction and quantification for long read RNA-Seq data. R package version 2.0.6, <https://github.com/GoekeLab/bambu>.
30. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, et al. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol*. 2022; 23: 10. <https://doi.org/10.1186/s13059-021-02588-5> PMID: 34991664

31. Kovaka S, Zimin A V., Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 2019; 20: 278. <https://doi.org/10.1186/s13059-019-1910-1> PMID: 31842956
32. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics.* 2020; 21: 793. <https://doi.org/10.1186/s12864-020-07207-4> PMID: 33372596
33. Stangl C, de Blank S, Renkens I, Westera L, Verbeek T, Valle-Inclan JE, et al. Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nat Commun.* 2020; 11: 2861. <https://doi.org/10.1038/s41467-020-16641-7> PMID: 32504042
34. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv.* 2020. <https://doi.org/10.1101/672931>
35. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* 2011; 12: R6. <https://doi.org/10.1186/gb-2011-12-1-r6> PMID: 21247443
36. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. *Science* (80-). 2016; 351: aac7247–aac7247. <https://doi.org/10.1126/science.aac7247> PMID: 26912865
37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; 26: 139–140. <https://doi.org/10.1093/bioinformatics/btp616> PMID: 19910308
38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B.* 1995; 57: 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
39. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2020; 48: D498–D503. <https://doi.org/10.1093/nar/gkz1031> PMID: 31691815
40. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010; 11: R25. <https://doi.org/10.1186/gb-2010-11-3-r25> PMID: 20196867
41. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell.* 2018; 173: 371–385.e18. <https://doi.org/10.1016/j.cell.2018.02.060> PMID: 29625053
42. Ding X-X, Zhu Q-G, Zhang S-M, Guan L, Li T, Zhang L, et al. Precision medicine for hepatocellular carcinoma: driver mutations and targeted therapy. *Oncotarget.* 2017; 8: 55715–55730. <https://doi.org/10.18632/oncotarget.18382> PMID: 28903454
43. Hur K, Cejas P, Feliu J, Moreno-Rubio J, Burgos E, Boland CR, et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut.* 2014; 63: 635–46. <https://doi.org/10.1136/gutjnl-2012-304219> PMID: 23704319
44. Altinel K, Hashimoto K, Wei Y, Neuveut C, Gupta I, Suzuki AM, et al. Single-Nucleotide Resolution Mapping of Hepatitis B Virus Promoters in Infected Human Livers and Hepatocellular Carcinoma. *J Virol.* 2016; 90: 10811–10822. <https://doi.org/10.1128/JVI.01625-16> PMID: 27681123
45. Furuta M, Tanaka H, Shiraishi Y, Uchida T, Imamura M, Fujimoto A, et al. Characterization of HBV integration patterns and timing in liver cancer and HBV-infected livers. *Oncotarget.* 2018; 9: 25075–25088. <https://doi.org/10.18632/oncotarget.25308> PMID: 29861854
46. Lau C-C, Sun T, Ching AKK, He M, Li J-W, Wong AM, et al. Viral-Human Chimeric Transcript Predisposes Risk to Liver Cancer Development and Progression. *Cancer Cell.* 2014; 25: 335–349. <https://doi.org/10.1016/j.ccr.2014.01.030> PMID: 24582836
47. Liang H-W, Wang N, Wang Y, Wang F, Fu Z, Yan X, et al. Hepatitis B virus-human chimeric transcript HBx-LINE1 promotes hepatic injury via sequestering cellular microRNA-122. *J Hepatol.* 2016; 64: 278–291. <https://doi.org/10.1016/j.jhep.2015.09.013> PMID: 26409216
48. Mall M, Kareta MS, Chanda S, Ahlenius H, Perotti N, Zhou B, et al. Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. *Nature.* 2017; 544: 245–249. <https://doi.org/10.1038/nature21722> PMID: 28379941
49. Matsushita F, Kameyama T, Kadokawa Y, Marunouchi T. Spatiotemporal expression pattern of Myt/NZF family zinc finger transcription factors during mouse nervous system development. *Dev Dyn.* 2014; 243: 588–600. <https://doi.org/10.1002/dvdy.24091> PMID: 24214099
50. Hu J, Ho AL, Yuan L, Hu B, Hua S, Hwang SS, et al. Neutralization of terminal differentiation in gliomagenesis. *Proc Natl Acad Sci.* 2013; 110: 14520–14527. <https://doi.org/10.1073/pnas.1308610110> PMID: 23918370

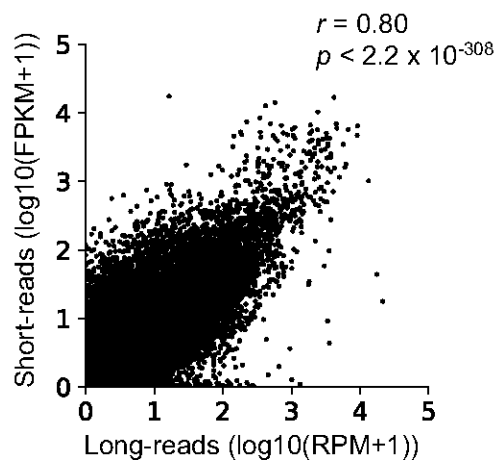
51. Melhuish TA, Kowalczyk I, Manukyan A, Zhang Y, Shah A, Abounader R, et al. Myt1 and Myt1l transcription factors limit proliferation in GBM cells by repressing YAP1 expression. *Biochim Biophys Acta Gene Regul Mech*. 2018; 1861: 983–995. <https://doi.org/10.1016/j.bbagr.2018.10.005> PMID: 30312684
52. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nat Rev Genet*. 2009; 10: 691–703. <https://doi.org/10.1038/nrg2640> PMID: 19763152
53. Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol*. 2001; 21: 1973–85. <https://doi.org/10.1128/MCB.21.6.1973-1985.2001> PMID: 11238933
54. Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, et al. Human retroelements may introduce intragenic polyadenylation signals. *Cytogenet Genome Res*. 2005; 110: 365–371. <https://doi.org/10.1159/000084968> PMID: 16093688
55. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev*. 1999; 9: 657–63. [https://doi.org/10.1016/s0959-437x\(99\)00031-3](https://doi.org/10.1016/s0959-437x(99)00031-3) PMID: 10607616
56. Chen C, Ara T, Gautheret D. Using Alu Elements as Polyadenylation Sites: A Case of Retroposon Exaptation. *Mol Biol Evol*. 2009; 26: 327–334. <https://doi.org/10.1093/molbev/msn249> PMID: 18984903
57. Wolff EM, Byun H-M, Han HF, Sharma S, Nichols PW, Siegmund KD, et al. Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer. *Ren B, editor. PLoS Genet*. 2010; 6: e1000917. <https://doi.org/10.1371/journal.pgen.1000917> PMID: 20421991
58. Zhao J, Hou Y, Yin C, Hu J, Gao T, Huang X, et al. Upregulation of histamine receptor H1 promotes tumor progression and contributes to poor prognosis in hepatocellular carcinoma. *Oncogene*. 2020; 39: 1724–1738. <https://doi.org/10.1038/s41388-019-1093-y> PMID: 31740780
59. Koike K, Moriya K, Iino S, Yotsuyanagi H, Endo Y, Miyamura T, et al. High-level expression of hepatitis B virus HBx gene and hepatocarcinogenesis in transgenic mice. *Hepatology*. 1994; 19: 810–9. Available: <http://www.ncbi.nlm.nih.gov/pubmed/8138251> PMID: 8138251
60. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, et al. Transposable elements drive widespread expression of oncogenes in human cancers. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, et al. 2019. Transposable elements drive widespread expression of oncogenes in human c. *Nat Genet*. 2019; 51: 611–617. <https://doi.org/10.1038/s41588-019-0373-3> PMID: 30926969
61. Abou-Rjaily GA, Lee SJ, May D, Al-Share QY, Deangelis AM, Ruch RJ, et al. CEACAM1 modulates epidermal growth factor receptor—mediated cell proliferation. *J Clin Invest*. 2004; 114: 944–52. <https://doi.org/10.1172/JCI21786> PMID: 15467833
62. Boege Y, Malehmir M, Healy ME, Bettermann K, Lorentzen A, Vucur M, et al. A Dual Role of Caspase-8 in Triggering and Sensing Proliferation-Associated DNA Damage, a Key Determinant of Liver Cancer Development. *Cancer Cell*. 2017; 32: 342–359.e10. <https://doi.org/10.1016/j.ccell.2017.08.010> PMID: 28898696
63. Brisson L, Bański P, Sboarina M, Dethier C, Danhier P, Fontenille M-J, et al. Lactate Dehydrogenase B Controls Lysosome Activity and Autophagy in Cancer. *Cancer Cell*. 2016; 30: 418–431. <https://doi.org/10.1016/j.ccell.2016.08.005> PMID: 27622334
64. Morris MR, Gentle D, Abdulrahman M, Maina EN, Gupta K, Banks RE, et al. Tumor Suppressor Activity and Epigenetic Inactivation of Hepatocyte Growth Factor Activator Inhibitor Type 2/SPINT2 in Papillary and Clear Cell Renal Cell Carcinoma. *Cancer Res*. 2005; 65: 4598–4606. <https://doi.org/10.1158/0008-5472.CAN-04-3371> PMID: 15930277
65. Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, et al. Comparison of gene expression profiles between hepatitis B virus- and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data on the basis of a supervised learning method. *Cancer Res*. 2002; 62: 3939–44. Available: <http://www.ncbi.nlm.nih.gov/pubmed/12124323> PMID: 12124323
66. Sun S, Li Y, Han S, Jia H, Li X, Li X. A comprehensive genome-wide profiling comparison between HBV and HCV infected hepatocellular carcinoma. *BMC Med Genomics*. 2019; 12: 147. <https://doi.org/10.1186/s12920-019-0580-x> PMID: 31660973
67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25: 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
68. Wyman D, Mortazavi A. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Berger B, editor. Bioinformatics*. 2019; 35: 340–342. <https://doi.org/10.1093/bioinformatics/bty483> PMID: 29912287

69. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, et al. Transcription-mediated gene fusion in the human genome. *Genome Res.* 2006; 16: 30–6. <https://doi.org/10.1101/gr.4137606> PMID: [16344562](https://pubmed.ncbi.nlm.nih.gov/16344562/)
70. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika.* 1974; 61: 439–447. <https://doi.org/10.1093/biomet/61.3.439>

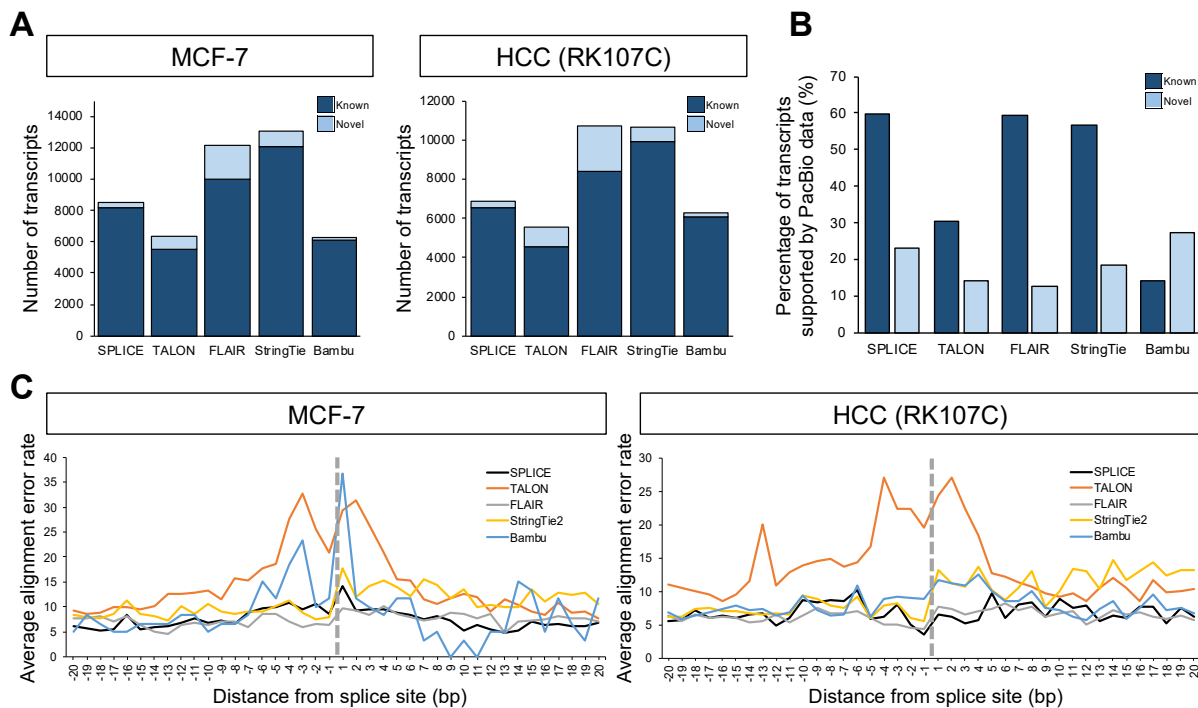
S1 Fig.



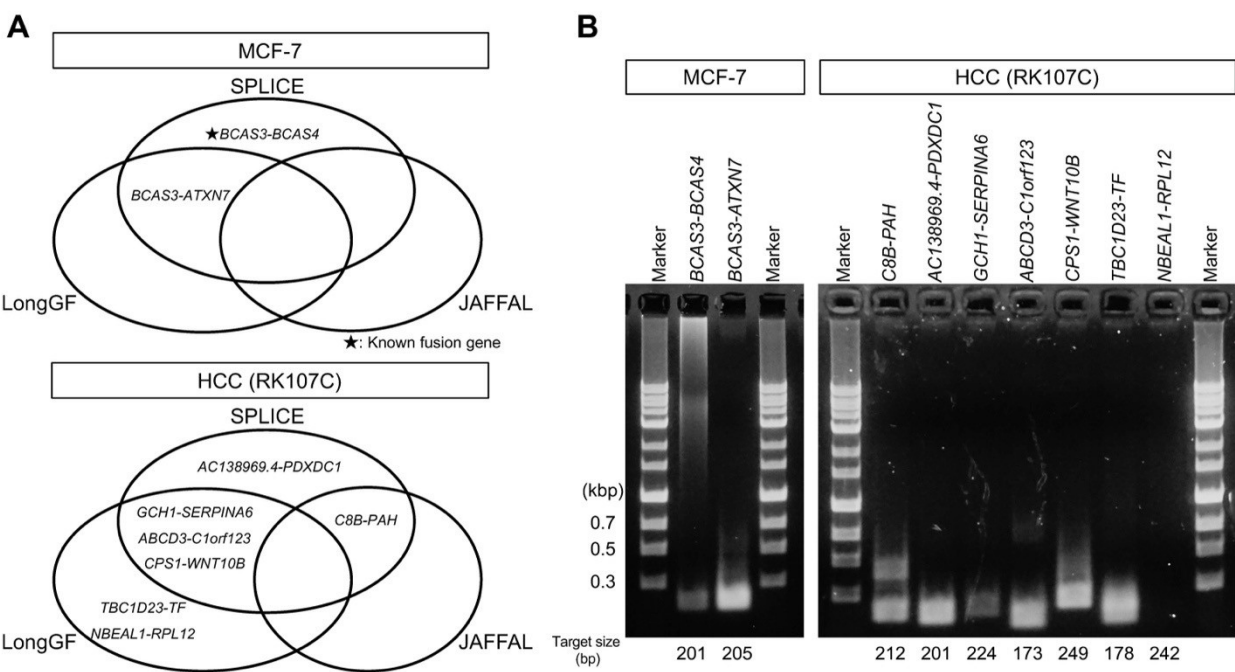
S2 Fig.



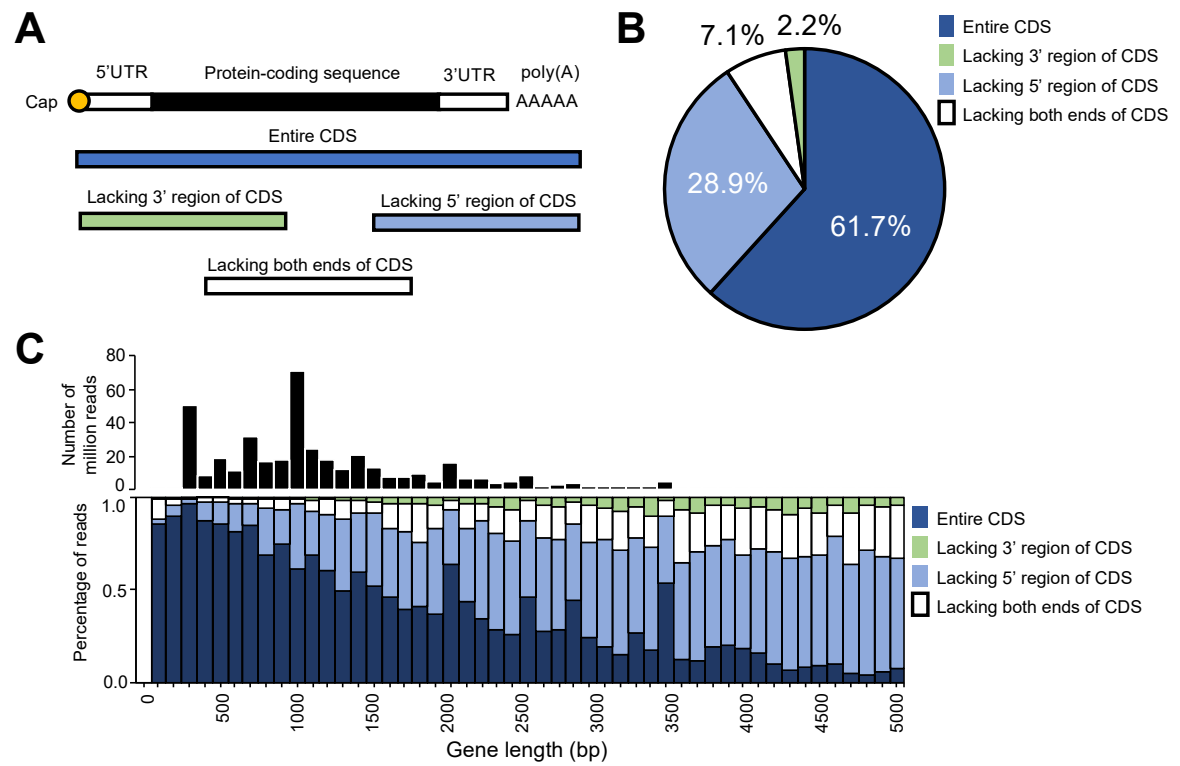
S3 Fig.



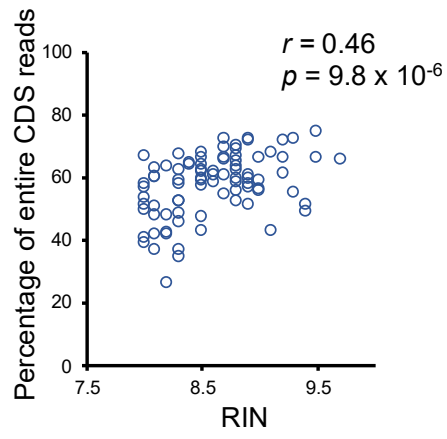
S4 Fig.



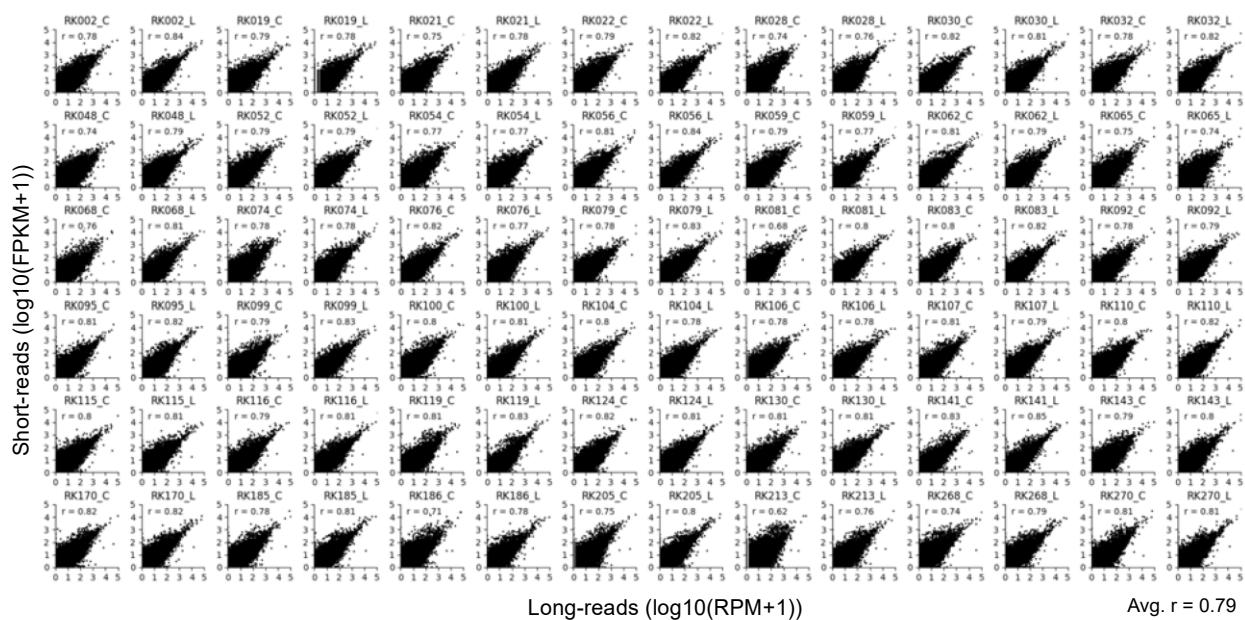
S5 Fig.



S6 Fig.

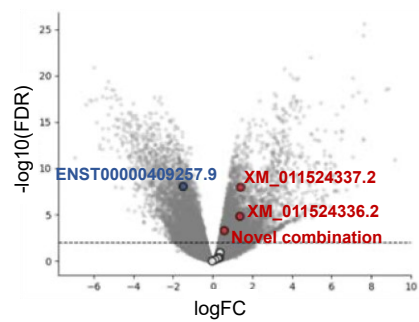


S7 Fig.

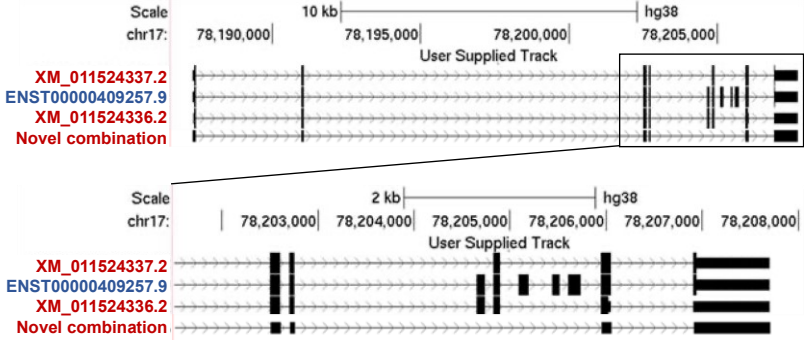


S8 Fig.

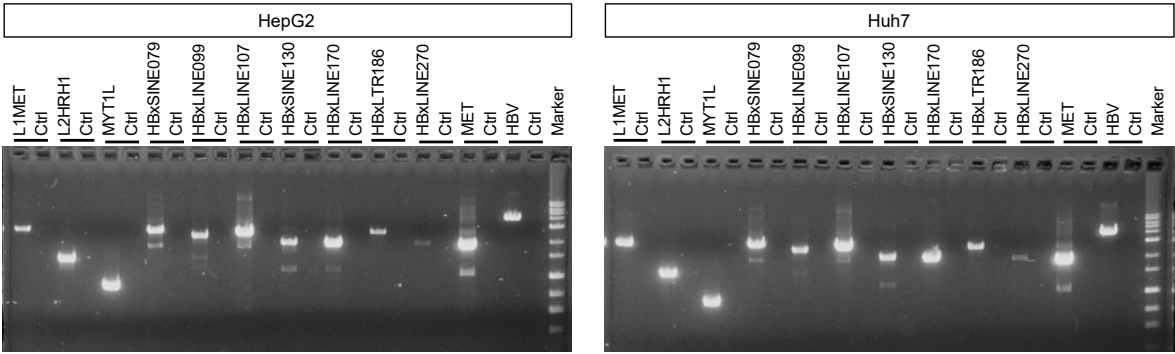
A



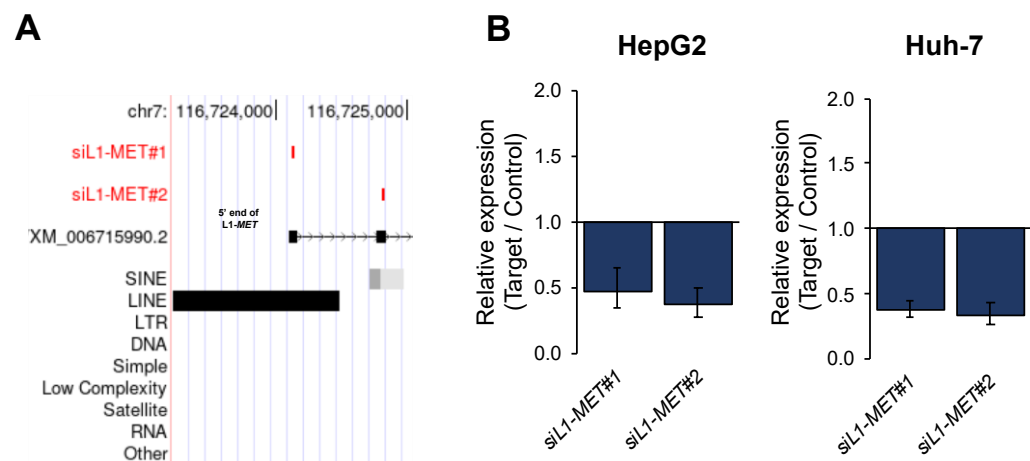
B



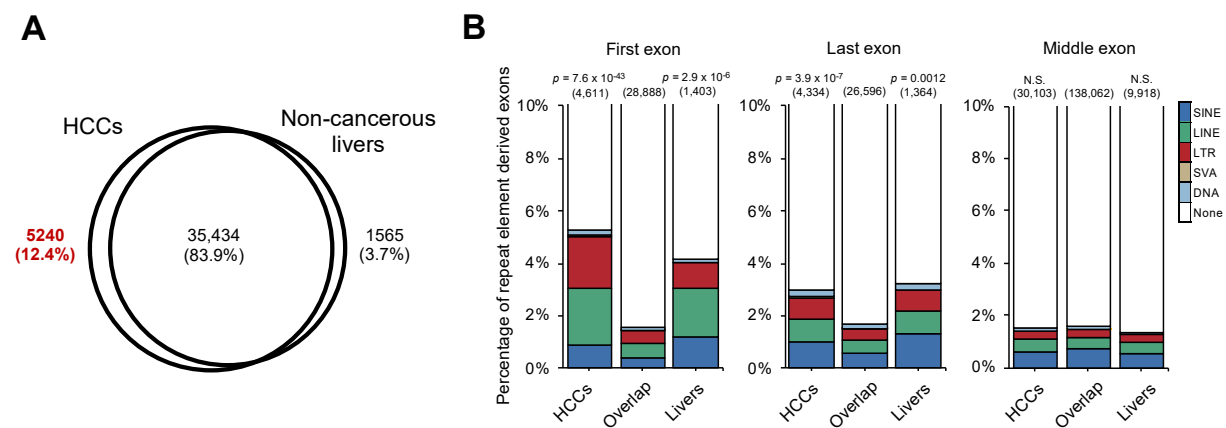
S9 Fig.



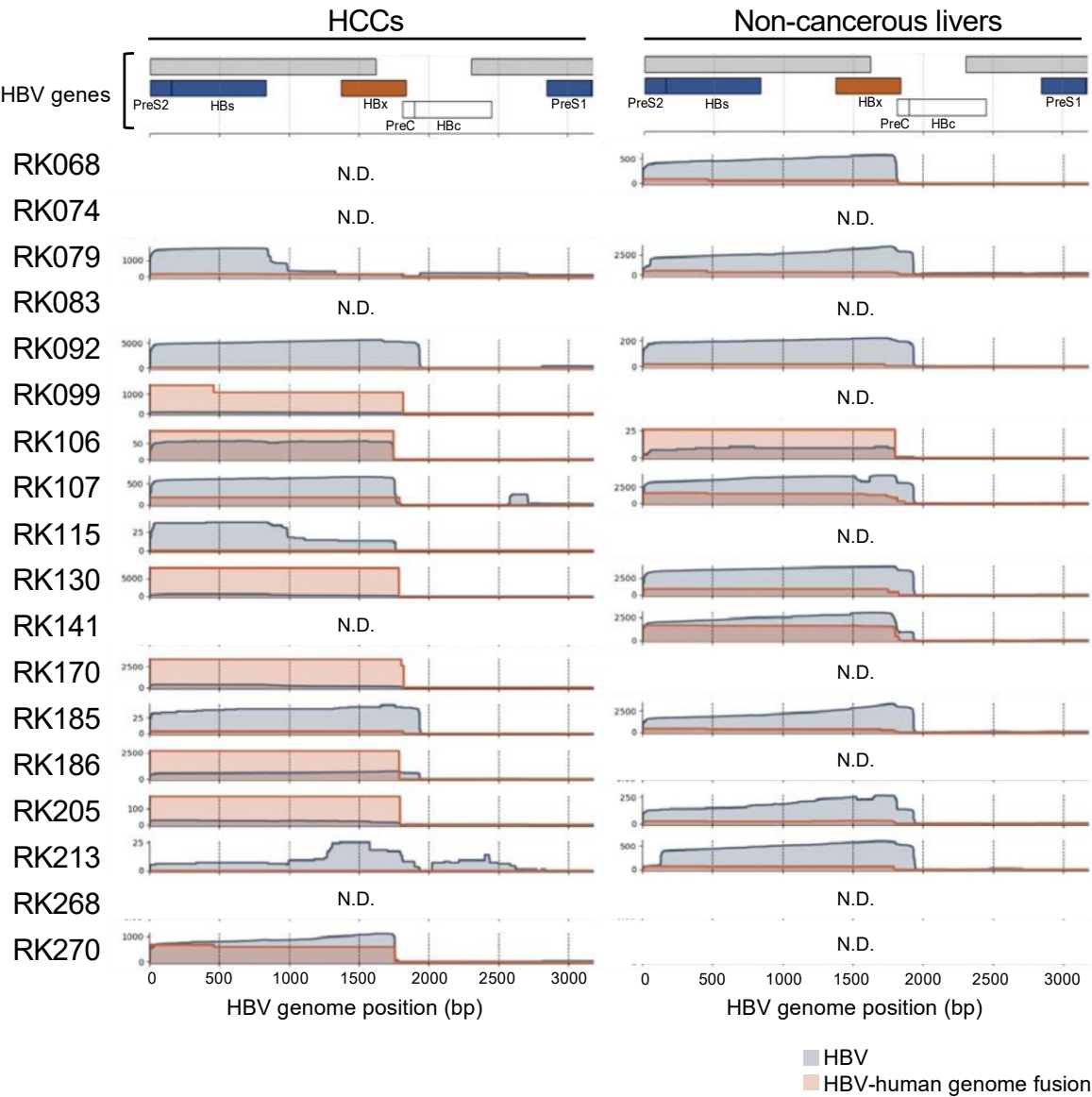
S10 Fig.



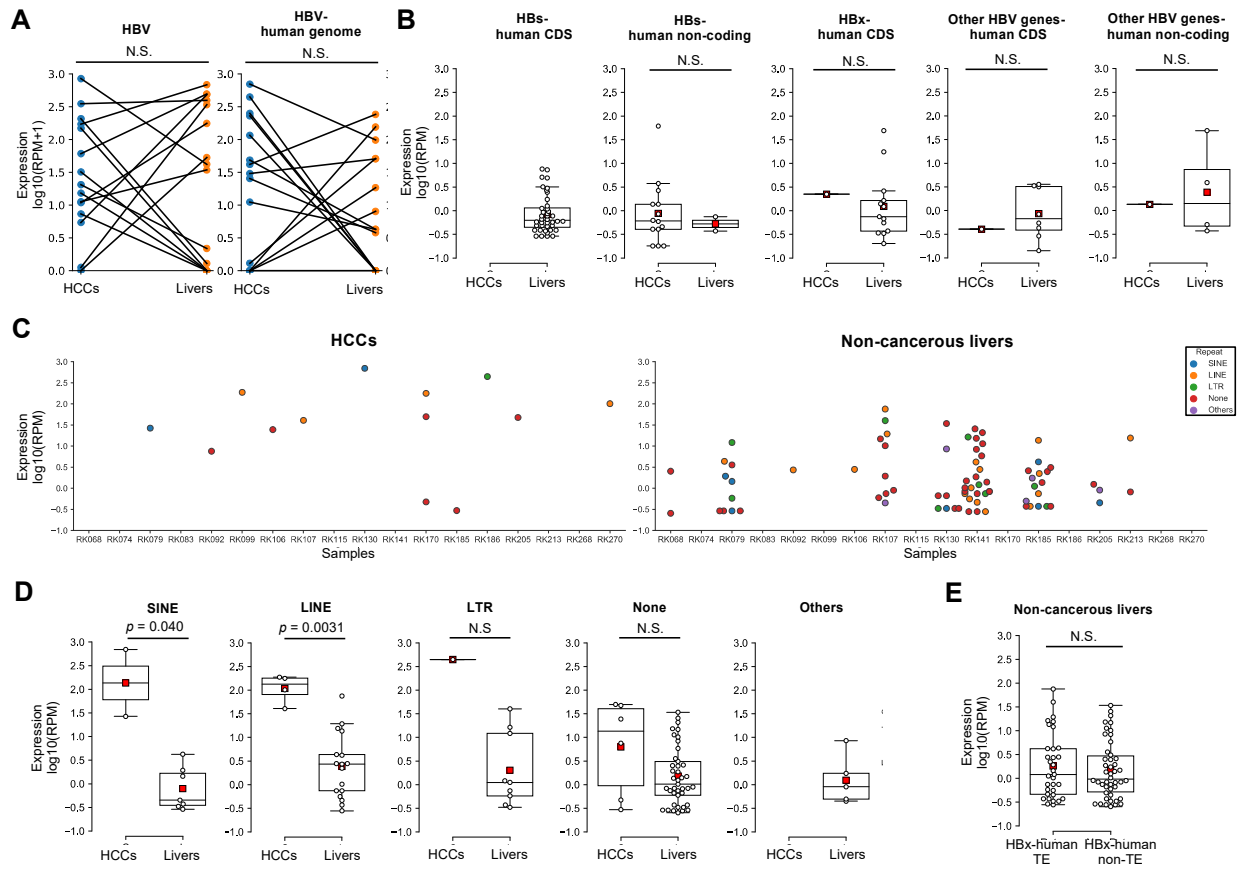
S11 Fig.



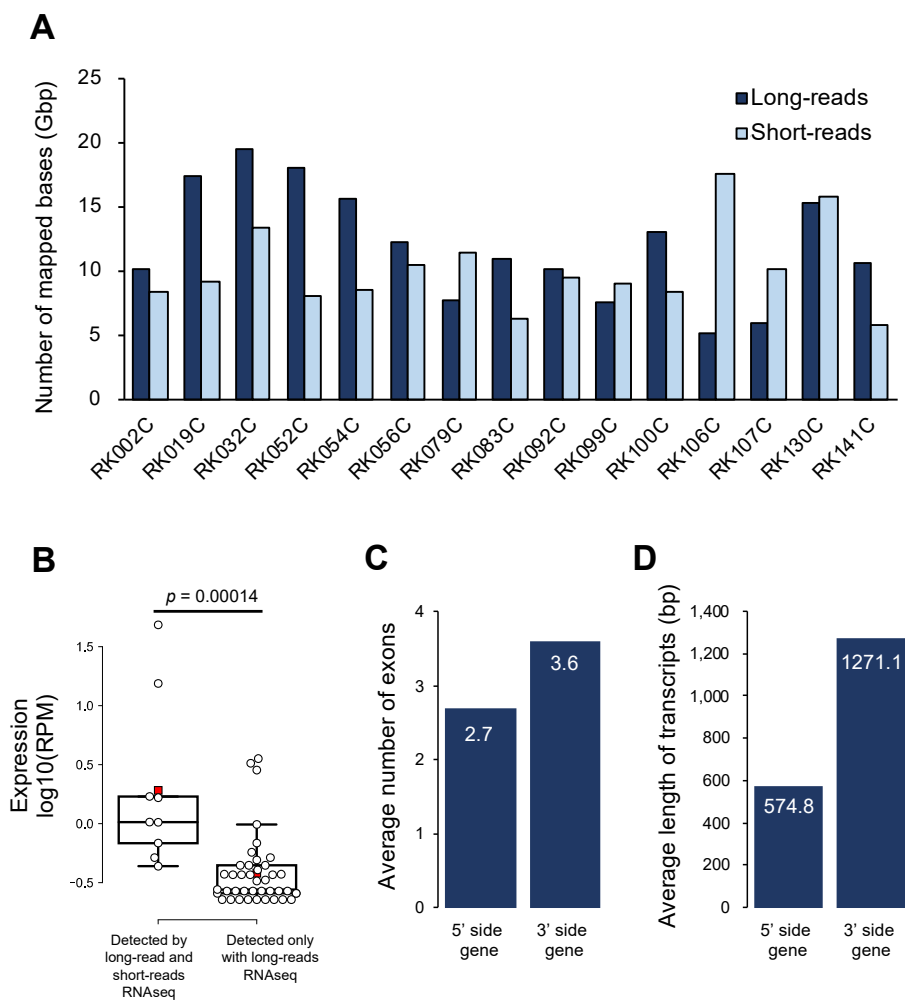
S12 Fig.



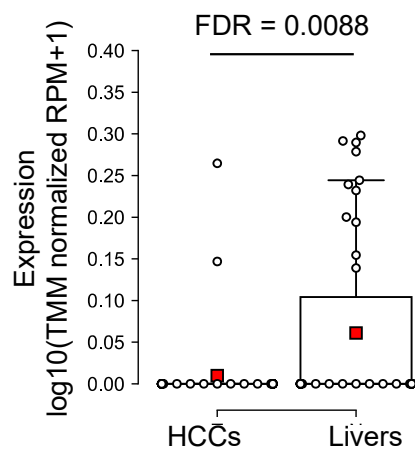
S13 Fig.



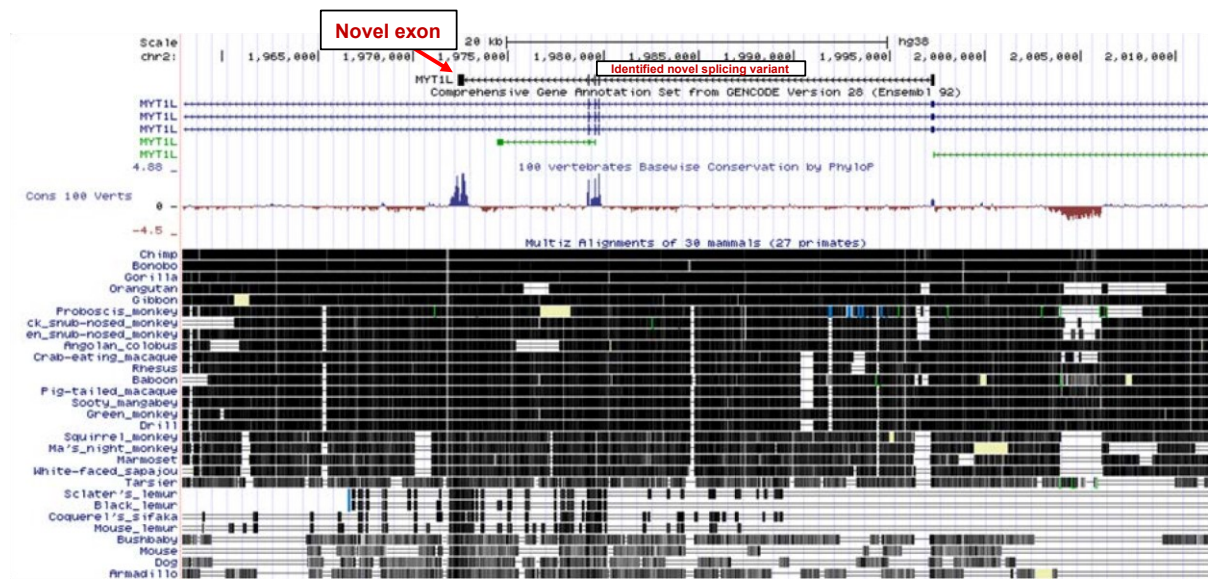
S14 Fig.



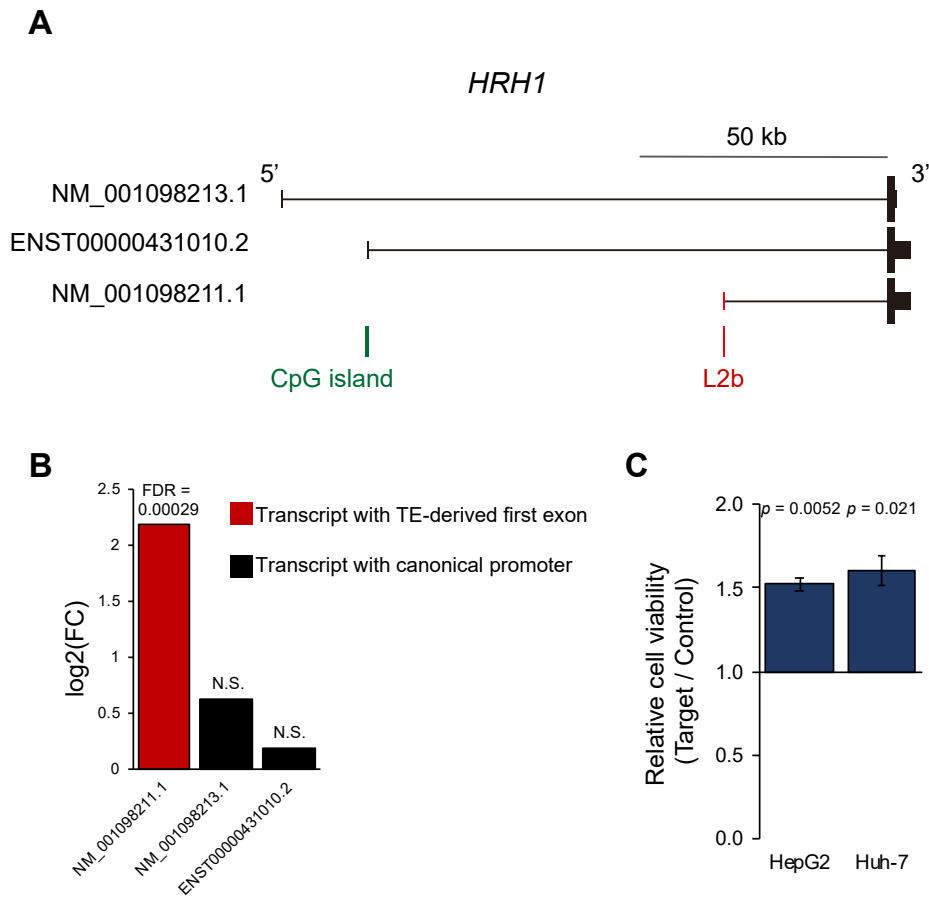
S15 Fig.



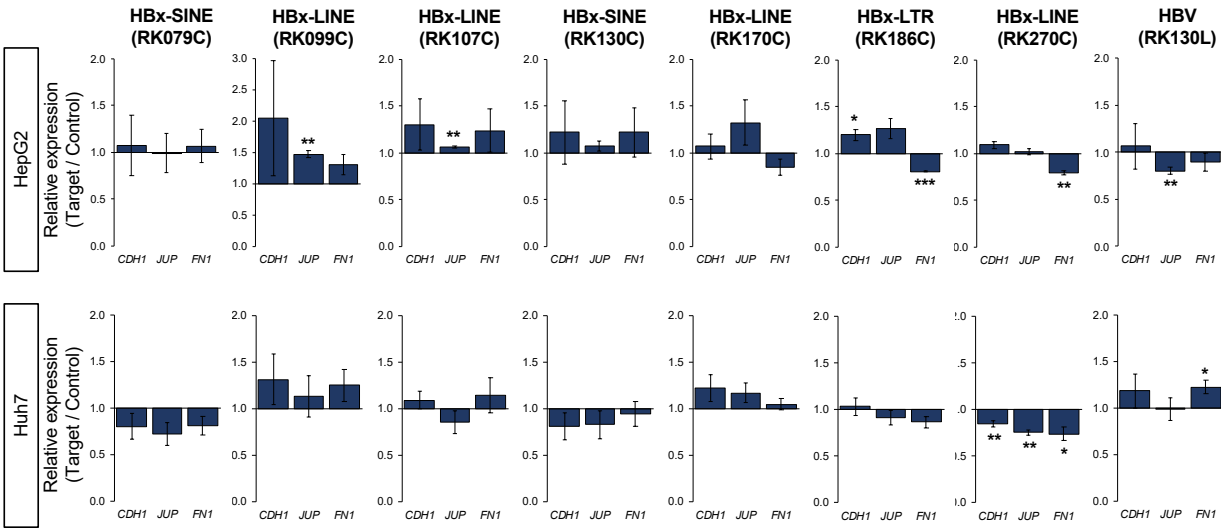
S16 Fig.



S17 Fig.



S18 Fig.



S1 Table

Items	Values
Total sequenced reads	16,538,905
Filtered reads	13,494,736
Filtered read bases (bp)	15,054,986,136
Longest read length	25,575
Mapped reads	13,424,802
Mapping rate (%)	99.48
Mapped bases	15,037,722,889
Average read length (bp)	1,120.15
Mismatch rate (%)	2
Insertion rate (%)	2
Deletion rate (%)	3

S2 Table

Fusion gene	Sample	Forward primer	Reverse primer
<i>BCAS3-BCAS4</i>	MCF-7	CCTCCTGATGCTGCTCGT	CTCCTCCGTGTGCTCCAT
<i>BCAS3-ATXN7</i>	MCF-7	CAGTGGTGAAGCACAAGAGC	ATGTAGAGACGGTGGCTGCT
<i>C8B-PAH</i>	RK107C	GCCAATGCACCAACTTCTTC	CTCTGTGGCATCTCCTGTCA
<i>CPS1-WNT10B</i>	RK107C	GACAGAAAGGAGCCTGATGC	TCTGACAAGGGGACAGAACC
<i>GCH1-SERPINA6</i>	RK107C	TTGCTGGGAAACAACAAAGA	GGACAAGGGGAAGATGAACA
<i>ABCD3-Clorf123</i>	RK107C	AGACTCCTGGGCCTTTTCAT	CCAGACCAGCTGCTTACACA
<i>AC138969.4-PDXDC1</i>	RK107C	ACTTGTGCAGAGGGAGATGG	GCGGGAGGAGGAAGTAGAG
<i>TBC1D23-TF</i>	RK107C	GCTCTTTGTCTGGCCATCTC	AGAAGGGAGATGTGGCCTTT
<i>NBEAL1-RPL12</i>	RK107C	ATCTCGTTGGGGTCGAACCT	GGCAATGTCATCACCAACCT

S3 Table

First exon (n = 500 for novel exons, n = 33,598 for known exons)

Repeat	Novel	Known	Odds ratio	<i>p</i>-value	Bonferroni adjusted <i>p</i>-value
SINE	41	138	21.66	9.60E-37	4.80E-36
LINE	72	205	27.40	1.66E-68	8.29E-68
LTR	49	184	19.73	8.17E-42	4.09E-41
SVA	0	3	0.00	1	1
DNA	5	41	8.27	0.000555002	0.0028

Last exon (n = 372 for novel exons, n = 30,838 for known exons)

Repeat	Novel	Known	Odds ratio	<i>p</i>-value	Bonferroni adjusted <i>p</i>-value
SINE	62	152	40.38	3.02E-68	1.51E-67
LINE	30	148	18.19	1.12E-25	5.59E-25
LTR	22	132	14.62	1.50E-17	7.52E-17
SVA	0	4	0.00	1	1
DNA	2	63	2.64	0.181584153	0.91

Middle exon (n = 129 for novel exons, n = 157,322 for known exons)

Repeat	Novel	Known	Odds ratio	<i>p</i>-value	Bonferroni adjusted <i>p</i>-value
SINE	26	1130	34.89	1.49E-29	7.47E-29
LINE	13	778	22.55	1.63E-13	8.13E-13
LTR	22	432	74.67	2.13E-32	1.07E-31
SVA	0	3	0.00	1	1
DNA	4	210	23.94	3.20E-05	0.00016

S4 Table

DET-specific genes			DET and DEG genes			P-value	Odds ratio
BiExp genes	Non-Biexp gene	Percentage of BiExp genes	BiExp genes	Non-Biexp gene	Percentage of BiExp genes		
42	704	5.630	38	4274	0.881	1.591×10^{-15}	6.706

S5 Table

※The total number of TE-derived exons in cancer-specific transcripts was 829

First exon

Position of the exon	Repeat	Repeat family	First exons of cancer-specific transcripts		First exons of transcripts that are not cancer-specific transcripts		Enrichment analysis of TE-derived exons in the first exon of cancer-specific transcripts and in the first exon of other transcripts					
			Sense	Antisense	Sense	Antisense	Sense			Antisense		
							Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
first	DNA	TcMar-Mariner	0	1	0	0	nan	1.000	1.000	inf	0.133	1.000
first	DNA	TcMar-Tigger	3	3	11	4	1.782	0.418	1.000	4.901	0.054	1.000
first	DNA	hAT-Charlie	2	0	7	8	1.867	0.340	1.000	0.000	0.608	1.000
first	LINE	CR1	5	0	7	1	4.669	0.015	0.406	0.000	1.000	1.000
first	LINE	L1	17	49	36	78	3.092	3.333E-04	0.00933	4.136	8.120E-13	2.274E-11
first	LINE	L2	4	23	12	47	2.178	0.256	1.000	3.207	2.078E-05	5.819E-04
first	LTR	ERV1	12	11	34	26	2.309	0.016	0.453	2.768	0.007	0.194
first	LTR	ERVL	27	8	27	10	6.564	9.792E-11	2.742E-09	5.233	0.001	0.034
first	LTR	ERVL-MaLR	22	7	23	10	6.273	9.074E-09	2.541E-07	4.578	0.004	0.116
first	LTR	Gypsy	1	0	2	3	3.266	0.348	1.000	0.000	1.000	1.000
first	LTR	LTR	5	0	0	0	inf	4.119E-05	0.00115	nan	1.000	1.000
first	Retroposon	SVA	1	0	1	1	6.533	0.248	1.000	0.000	1.000	1.000
first	SINE	Alu	8	21	22	84	2.378	0.052	1.000	1.636	0.059	1.000
first	SINE	MIR	6	7	6	26	6.539	0.002	0.069	1.760	0.194	1.000

Last exon

Position of the exon	Repeat	Repeat family	Last exons of cancer-specific transcripts		Last exons of transcripts that are not cancer-specific transcripts		Enrichment analysis of TE-derived exons in the last exon of cancer-specific transcripts and in the last exon of other transcripts					
			Sense	Antisense	Sense	Antisense	Sense			Antisense		
							Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
last	DNA	TcMar-Tigger	3	2	10	5	1.918	0.404	1.000	2.557	0.243	1.000
last	DNA	hAT-Blackjack	1	0	0	0	inf	0.135	1.000	nan	1.000	1.000
last	DNA	hAT-Charlie	2	1	22	6	0.581	0.763	1.000	1.065	1.000	1.000
last	DNA	hAT-Tip100	0	1	0	1	nan	1.000	1.000	6.391	0.252	1.000
last	LINE	L1	19	18	61	51	1.995	0.013	0.360	2.260	0.004	0.126
last	LINE	L2	1	1	15	10	0.426	0.713	1.000	0.639	1.000	1.000
last	LTR	ERV1	12	2	40	10	1.919	0.064	1.000	1.278	0.672	1.000
last	LTR	ERVK	5	1	14	2	2.284	0.167	1.000	3.195	0.354	1.000
last	LTR	ERVL	3	1	10	3	1.918	0.404	1.000	2.130	0.441	1.000
last	LTR	ERVL-MaLR	8	3	32	8	1.599	0.244	1.000	2.397	0.178	1.000
last	Retroposon	SVA	1	1	2	0	3.195	0.354	1.000	inf	0.135	1.000
last	SINE	5S-Deu-L2	2	0	1	0	12.785	0.050	1.000	nan	1.000	1.000
last	SINE	Alu	26	14	93	70	1.791	0.015	0.410	1.279	0.423	1.000
last	SINE	MIR	0	1	6	4	0.000	1.000	1.000	1.598	0.517	1.000

Middle exon

Position of the exon	Repeat	Repeat family	Middle exons of cancer-specific transcripts		Middle exons of transcripts that are not cancer-specific transcripts		Enrichment analysis of TE-derived exons in the middle exon of cancer-specific transcripts and in the middle exon of other transcripts					
			Sense	Antisense	Sense	Antisense	Sense			Antisense		
							Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
middle	DNA	MULE-MuDR	1	0	1	0	4.747	0.318	1.000	nan	1.000	1.000
middle	DNA	TcMar-Mariner	0	1	3	4	0.000	1.000	1.000	1.187	1.000	1.000
middle	DNA	TcMar-Tigger	4	6	40	29	0.475	0.167	1.000	0.982	1.000	1.000
middle	DNA	hAT-Ac	1	0	7	6	0.678	1.000	1.000	0.000	0.599	1.000
middle	DNA	hAT-Charlie	10	6	46	40	1.032	0.861	1.000	0.712	0.560	1.000
middle	LINE	CR1	7	9	25	30	1.329	0.485	1.000	1.424	0.395	1.000
middle	LINE	L1	21	54	120	253	0.831	0.505	1.000	1.013	0.940	1.000
middle	LINE	L2	22	26	94	139	1.111	0.625	1.000	0.888	0.681	1.000
middle	LINE	RTE-BovB	4	0	12	6	1.582	0.504	1.000	0.000	0.599	1.000
middle	LINE	RTE-X	0	2	3	2	0.000	1.000	1.000	4.748	0.142	1.000
middle	LTR	ERV1	17	12	50	55	1.614	0.105	1.000	1.036	0.872	1.000
middle	LTR	ERVK	2	2	13	4	0.730	1.000	1.000	2.374	0.281	1.000
middle	LTR	ERVL	11	9	40	39	1.306	0.458	1.000	1.096	0.849	1.000
middle	LTR	ERVL-MaLR	26	26	77	90	1.603	0.050	1.000	1.372	0.176	1.000
middle	LTR	Gypsy	0	1	1	2	0.000	1.000	1.000	2.374	0.436	1.000
middle	SINE	5S-Deu-L2	1	0	2	2	2.374	0.436	1.000	0.000	1.000	1.000
middle	SINE	Alu	20	98	98	650	0.969	1.000	1.000	0.715	0.002	0.066
middle	SINE	MIR	16	37	75	176	1.013	1.000	1.000	0.998	1.000	1.000
middle	SINE	tRNA	1	3	2	14	2.374	0.436	1.000	1.017	1.000	1.000
middle	SINE	tRNA-Deu	0	1	0	0	nan	1.000	1.000	inf	0.174	1.000

S6 Table

※The total number of TE-derived exons in non-cancerous liver-specific transcripts was 235

Position of the exon	Repeat	Repeat family	First exons of matced liver-specific transcripts		First exons of splicing variants that are not matched liver-specific transcripts		Enrichment analysis of TE-derived exons in the first exon of matched liver-specific transcripts and in the first exon of other transcripts					
							Sense			Antisense		
			Sense	Antisense	Sense	Antisense	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
first	DNA	hAT-Tip100	0	1	1	1	0.000	1.000	1.000	23.474	0.080	1.000
first	LINE	L1	6	6	46	120	3.069	0.019	0.303	1.174	0.649	1.000
first	LINE	L2	1	13	15	58	1.564	0.487	1.000	5.298	5.379E-06	8.607E-05
first	LTR	ERV1	2	5	44	31	1.066	0.713	1.000	3.794	0.015	0.239
first	LTR	ERVL	4	1	51	16	1.842	0.287	1.000	1.466	0.508	1.000
first	LTR	ERV1-MaLR	0	2	44	15	0.000	0.263	1.000	3.131	0.152	1.000
first	SINE	Alu	3	10	28	96	2.517	0.132	1.000	2.454	0.012	0.186
first	SINE	MIR	2	2	10	30	4.697	0.084	1.000	1.565	0.378	1.000

Position of the exon	Repeat	Repeat family	Last exons of matced liver-specific transcripts		Last exons of splicing variants that are not matched liver-specific transcripts		Enrichment analysis of TE-derived exons in the last exon of matched liver-specific transcripts and in the last exon of other transcripts					
							Sense			Antisense		
			Sense	Antisense	Sense	Antisense	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
last	DNA	TcMar-Tigger	1	0	12	7	1.844	0.437	1.000	0.000	1.000	1.000
last	DNA	hAT-Charlie	1	1	23	6	0.962	1.000	1.000	3.688	0.266	1.000
last	LINE	CR1	0	1	2	0	0.000	1.000	1.000	inf	0.043	0.865
last	LINE	L1	4	5	75	63	1.180	0.586	1.000	1.758	0.220	1.000
last	LINE	L2	0	2	16	9	0.000	1.000	1.000	4.920	0.079	1.000
last	LTR	ERV1	1	2	50	10	0.442	0.727	1.000	4.428	0.093	1.000
last	LTR	ERV1	1	0	18	3	1.229	0.569	1.000	0.000	1.000	1.000
last	LTR	ERVL	0	1	13	3	0.000	1.000	1.000	7.377	0.162	1.000
last	LTR	ERV1-MaLR	4	2	36	9	2.462	0.093	1.000	4.920	0.079	1.000
last	SINE	Alu	13	5	106	78	2.729	0.002	0.039	1.419	0.411	1.000

Position of the exon	Repeat	Repeat family	Middle exons of matced liver-specific transcripts		Middle exons of transcripts that are not matched liver-specific transcripts		Enrichment analysis of TE-derived exons in the middle exon of matched liver-specific transcripts and in the middle exon of other transcripts					
							Sense			Antisense		
			Sense	Antisense	Sense	Antisense	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -	Odds ratio	<i>p</i> -value	Bonferroni adjusted <i>p</i> -
middle	DNA	TcMar-Tigger	1	1	43	34	0.359	0.522	1.000	0.455	0.722	1.000
middle	DNA	hAT-Charlie	2	3	52	42	0.595	0.772	1.000	1.104	0.753	1.000
middle	LINE	CR1	0	4	29	33	0.000	0.419	1.000	1.874	0.284	1.000
middle	LINE	L1	6	16	133	283	0.697	0.479	1.000	0.874	0.716	1.000
middle	LINE	L2	3	12	110	149	0.421	0.165	1.000	1.245	0.411	1.000
middle	LTR	ERV1	4	6	57	59	1.085	0.787	1.000	1.573	0.290	1.000
middle	LTR	ERVL	0	2	50	46	0.000	0.074	1.000	0.672	1.000	1.000
middle	LTR	ERV1-MaLR	10	7	91	103	1.700	0.138	1.000	1.051	0.841	1.000
middle	SINE	Alu	7	34	109	704	0.993	1.000	1.000	0.746	0.104	1.000
middle	SINE	MIR	6	8	80	201	1.160	0.651	1.000	0.615	0.194	1.000
middle	SINE	tRNA	0	1	3	17	0.000	1.000	1.000	0.909	1.000	1.000

S7 Table

Fusion genes	Sample	Forward primer	Reverse primer
<i>SLC20A2-SEN6</i>	RK059C	GCTCCGGGGCTCTTTACC	TATCTACGTGCCGGGTCCAT
<i>UBE3D-SEN6</i>	RK119C	ACCTCTCCTGGTCCTGTTGA	TGGTCCAGTAGACTCATTCAAAGT
<i>TBRG1-IFT46</i>	RK130C	GATCATCACCGACCGACCTG	GCATCAGCGTGTCAATGTCG
<i>ATL3-TBRG1</i>	RK116C	CGTTTTTCGGGTGAACTTGACA	AGGCAGTGTTTTTCAAGGTGT
<i>SDC2-CPQ</i>	RK107C	AAACTTCTGCCGTAGCTCCC	GCTTTAGCAACATCTCCACAGC
<i>SDC2-CSAD</i>	RK116C	TCCAGGCTAAGGACTTTTGTT	AACATTGCGCAGGTCACTGT
<i>ABCD3-Clorff123</i>	RK107C	GCTAATCCAGACCAGCTGCT	AAAGGAACGAGCCTCTGTGG
<i>ABCD3-DST</i>	RK076C	GCTGTGGTGGACAAGGTGTT	TCTGGCAGCTGCAGTTGTTA
<i>LGSN-SLC13A3</i>	RK059C	ACAGTCACCCACCTAGGAGG	TGATGAGCAGCACAGCCAG
<i>LGSN-GGPS1</i>	RK052C	ACAGTCACCCACCTAGGAGG	GTCCTCTGGAACTTTCAGCCA
<i>LDLR-SLC41A2</i>	RK124C	GGCTGAGCTGGATCACTTGA	CGGGCTTCTTTAAAAGAGCAA
<i>CLN3-LDLR</i>	RK170C	GGAGAGCAGTGGCACAATCT	ACAGTGTTTTGTCCCTGGGG

S8 Table

Pathway name	Entities found	Entities total	Entities ratio	Entities <i>p</i> - value	Entities FDR
Interferon alpha/beta signaling	65	191	0.0130	0.0000	1.10E-14
Antigen Presentation: Folding, assembly and peptide loading of class I MHC	38	102	0.0069	0.0000	1.10E-14
ER-Phagosome pathway	42	165	0.0112	0.0000	1.10E-14
Endosomal/Vacuolar pathway	36	82	0.0056	1.11E-16	1.10E-14
Interferon Signaling	66	401	0.0272	1.11E-16	1.10E-14
Antigen processing-Cross presentation	43	187	0.0127	1.11E-16	1.10E-14
Interferon gamma signaling	45	255	0.0173	1.11E-16	1.10E-14
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	39	317	0.0215	1.11E-16	1.10E-14
Cytokine Signaling in Immune system	82	1108	0.0753	1.11E-16	1.10E-14
Class I MHC mediated antigen processing & presentation	44	465	0.0316	5.00E-15	4.45E-13
Immune System	122	2713	0.1843	2.94E-14	2.38E-12
Adaptive Immune System	56	1003	0.0681	1.08E-09	8.00E-08

S9 Table

Pathway name	Entities found	Entities total	Entities ratio	Entities <i>p</i> -value	Entities FDR
Interferon alpha/beta signaling	6	191	0.0130	1.6E-04	0.0330
Iron uptake and transport	4	83	0.0056	4.3E-04	0.0330
The citric acid (TCA) cycle and respiratory electron transport	6	233	0.0158	4.5E-04	0.0330
RUNX3 Regulates Immune Response and Cell Migration	2	10	6.79E-04	8.5E-04	0.0451
Nef mediated downregulation of MHC class I complex cell surface expression	2	11	7.47E-04	0.00102	0.0451

S10 Table

Summary of RNA-seq and mapping of MCF-7

Items	v3.0.3	v6.0.6
Total sequenced reads	1,000,000	1,000,000
Filtered reads	926,530	989,031
Filtered read bases (bp)	1,035,718,911	1,105,567,703
Longest read length	16,997	17,151
Mapped reads	921,744	983,477
Average length	1,122.33	1,122.49
Mismatch_rate (%)	2.00	1.45
Insertion_rate (%)	2.03	1.24
Deletion_rate (%)	2.78	1.87

Result of splicing variant analysis

Items	v3.0.3	v6.0.6
Number of known transcripts	8,919	9,228
Number of novel transcripts	410	446

Result of fusion analysis

Items	v3.0.3	v6.0.6
Fusion genes (Number of support reads)	<i>BCAS3-BCAS4</i> (3)	<i>BCAS3-BCAS4</i> (4), <i>BCAS3-ATXN7</i> (4)

※ In v3.0.3, the number of support reads for *BCAS3-ATXN7* was 2.

S11 Table

Target transcript	Clone type	Sample	Forward Sequence	Reverse sequence	Target size (bp)
L1-MET (XM_006715990.2)	CDS	RK030C	ATGGGTCAATTCAGCGAAGT	CTATGATGTCTCCAGAAGG	2883
MET (NM_001127500.2)	CDS	pFN21AB5617	ATGAAGGCCCGCTGTGCTT	CTATGATGTCTCCAGAAGG	4227
L2-HRH1 (NM_001098211.1)	CDS	RK048C	ATGAGCCTCCCAATTCCTC	TTAGGAGCGAATATGCAGAA	1464
HBx-SINE_079C	Full length	RK079C	TCCACAACATTCACCGAGC	TTGTGTTGTTTCCTTTGCAG	2786
HBx-LINE_099C	Full length	RK099C	TCCACAACCTTCCACCAAAC	GTCTCACTTTAATACAGGTTCA	2394
HBx-LINE_107C	Full length	RK107C	CAGGCCATGCAGTGGAATC	TGGAACACAGCCACACCCAT	2552
HBx-SINE_130C	Full length	RK130C	TCCACCACATTCACCAAGC	TTGTGTTCACTCTTGAGTCT	2011
HBx-LINE_170C	Full length	RK170C	TCCACAACATTCACCAAAC	TTAGAAAACCTGAATTAATTTA	1973
HBx-LTR_186C	Full length	RK186C	TCCACAACATTCACCAAGC	TAGGAGGGATGGACTTTAAT	2600
HBx-LINE_270C	Full length	RK270C	TCCACAACATTCACCAAGC	TTTGAATCTTTTCAGATCTTCAC	1946
HBV_130L	Full length	RK130L	TCCACCACATTCACCAAGC	GGTCAATGTCCATGCCCAA	1998

S12 Table

siRNA design

siRNA ID	Target Name	Sense/Antisense	siRNA design	Start on Target	Target sequence
siL1-MET#1	<i>MET</i> XM_006715990.2	s	GCGGGACUCCGUGGGCGUAdTdT	31	GCGGGACTCCGTGGGCGTA
siL1-MET#1	<i>MET</i> XM_006715990.2	a	UACGCCACGGAGUCCCGCdTdT	31	TACGCCACGGAGTCCCGC
siL1-MET#2	<i>MET</i> XM_006715990.2	s	GAGCCUAGGCUUAGUCCUAdTdT	109	GAGCCTAGGCTTAGTCCTA
siL1-MET#2	<i>MET</i> XM_006715990.2	a	UAGGACUAAGCCUAGGCUCdTdT	109	TAGGACTAAGCCTAGGCTC

qPCR primers

Target	Forward/Reverse	length (bp)	Sequence
L1-MET	F	20	GAGCCAGGCAGAAAATGTGC
L1-MET	R	20	CTTCAGTGCAGGGCTAGGAC

S1 Data. Clinical and pathological information of the 42 HCC patients.

<https://doi.org/10.1371/journal.pgen.1010342.s031>

S2 Data. Summary of RNA-seq of HCC clinical samples.

<https://doi.org/10.1371/journal.pgen.1010342.s032>

S3 Data. List of transcripts with novel exons.

<https://doi.org/10.1371/journal.pgen.1010342.s033>

S4 Data. List of significantly differentially expressed transcripts.

<https://doi.org/10.1371/journal.pgen.1010342.s034>

S5 Data. Significantly enriched reactome pathways of genes with up- or down-regulated DETs.

<https://doi.org/10.1371/journal.pgen.1010342.s035>

S6 Data. List of significantly expressed driver gene's transcripts.

<https://doi.org/10.1371/journal.pgen.1010342.s036>

S7 Data. List of bi-directionally significantly expressed genes.

<https://doi.org/10.1371/journal.pgen.1010342.s037>

S8 Data. List of HBV-human genome fusion transcripts.

<https://doi.org/10.1371/journal.pgen.1010342.s038>

S9 Data. List of fusion genes.

<https://doi.org/10.1371/journal.pgen.1010342.s039>

S10 Data. Significantly enriched reactome pathways of genes with up-regulated DETs in HBV-infected HCCs vs HBV-infected non-cancerous livers.

<https://doi.org/10.1371/journal.pgen.1010342.s040>

S11 Data. Significantly enriched reactome pathways of genes with up-regulated DETs in HCV-infected HCCs vs. HCV-infected non-cancerous livers.

<https://doi.org/10.1371/journal.pgen.1010342.s041>