The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients

Zhaoshi Jiang, Suchit Jhunjhunwala, Jinfeng Liu, et al.

Genome Res. published online January 20, 2012
Access the most recent version at doi:10.1101/gr.133926.111

The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients

Zhaoshi Jiang[1*], Suchit Jhunjhunwala[1*], Jinfeng Liu[1], Peter M. Haverty[1], Michael I. Kennemer[2], Yinghui Guan[3], William Lee[1], Paolo Carnevali[2], Jeremy Stinson[3], Stephanie Johnson[4], Jingyu Diao[5], Stacy Yeung[3], Adrian Jubb[4], Weilan Ye[3], Thomas D. Wu[1], Sharookh B. Kapadia[5], Frederic J. de Sauvage[3], Robert C. Gentleman[1], Howard M. Stern[4], Somasekar Seshagiri[3], Krishna P. Pant[2], Zora Modrusan[3], Dennis G. Ballinger[2] and Zemin Zhang[1§]

Keywords:
Cancer, next generation sequencing, hepatocellular carcinoma, hepatitis B virus.

1.Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA 94080, USA
2.Complete Genomics Inc., Mountain View, CA 94043, USA
3.Department of Molecular Biology, Genentech Inc., South San Francisco, CA 94080, USA
4.Department of Pathology, Genentech Inc., South San Francisco, CA 94080, USA
5.Department of Microbial Pathogenesis, Genentech Inc., South San Francisco, CA 94080, USA

* These authors contributed equally to this study

§ Corresponding author:
Zemin Zhang, Ph.D
Principal Scientist
Department of Bioinformatics and Computational Biology,
Genentech Inc., South San Francisco, California 94080,
USA
Phone: +(01) 650-225-4293
Fax: +(01) 650-225-5389
Email: zhang.zemin@gene.com

Hepatitis B virus (HBV) infection is a leading risk factor for hepatocellular carcinoma (HCC). HBV integration into the host genome has been reported but its scale, impact and contribution to HCC development is not clear. Here, we sequenced the tumor and non-tumor genomes (>80X coverage) and transcriptomes of four HCC patients and identified 255 HBV integration sites. Increased sequencing to 240X coverage revealed a proportionally higher number of integration sites. Clonal expansion of HBV-integrated hepatocytes was found specifically in tumor samples. We observe a diverse collection of genomic perturbations near viral integration sites, including direct gene disruption, viral promoter-driven human transcription, viral-human transcript fusion and DNA copy number alteration. Thus, we report the most comprehensive characterization of HBV integration in hepatocellular carcinoma patients. Such widespread random viral integration will likely increase carcinogenic opportunities in HBV-infected individuals.

More than 350 million people are infected by hepatitis B virus (HBV) worldwide (Lavanchy 2004). HBV is a leading risk factor for hepatocellular carcinoma (HCC), with over eighty percent of HCC cases occurring in the regions where HBV is endemic (Michielsen and Ho 2011). Approximately 30-50% of the estimated 320,000 annual HBV-related deaths are due to hepatocellular carcinoma (HCC) (Farazi and DePinho 2006). Despite clear evidence supporting the involvement of HBV in HCC (Farazi and DePinho 2006; Chemin and Zoulim 2009; Bouchard and Navas-Martin 2011) the underlying nature of viral-host interaction remains elusive (Block et al. 2003). HBV integration into the host genome has been reported both in tumors (Gozuacik et al. 2001; Murakami et al. 2005; Saigo et al. 2008) and in non-tumor liver tissue from HBV-infected individuals (Mason et al. 2010), although such integration is not essential for HBV replication. The relative extent, mutation model and the functional impact of HBV integration in host genomes is not clear due to lack of an unbiased approach to identify and quantify genome-wide HBV integration sites. Recent advances in sequencing technologies (Meyerson et al. 2010) provide an opportunity to investigate the global extent, mutation model and functional impact of viral integration in the host genome. Recently, a primary hepatitis C virus-infected HCC patient has been subjected to whole genome sequencing and many somatic mutations were reported (Totoki et al. 2011). However, as a RNA virus, HCV never integrates into the host genome during its life cycle, therefore liver cancer with HCV infection is not an optimal model to study viral-human genomic interactions. To that end, sequencing the genome and transcriptome of HBV

positive HCC patient provides a great opportunity to reveal functional impact of viral integration on the host genome.

Results

Detection of HBV integration based on whole genome sequencing

We performed whole genome deep sequencing (>80X coverage) and transcriptome sequencing on primary HCC tumors and matched adjacent non-neoplastic liver tissues from four patients (Supplemental Table 1; Supplemental Fig. 1; Supplemental Material Sections 1-3). Three of these patients are HBV positive and one is HBV negative. For comparison, we also sequenced blood samples from one HBV positive and one HBV negative patient. Deep coverage sequencing enabled detection of rare integration events, quantification of abundance of each event and investigation of the genomic impact of HBV integration on the human genome. Besides, transcriptome sequencing (RNA-seq) enabled us to evaluate the transcriptional impact of the integration events.

To detect HBV sequences in our samples, we aligned all short reads from whole genome sequencing against a comprehensive list (n=73) of HBV reference genomes (genotype A-I and a strain of Woolly Monkey HBV, Supplemental Table 2) (Zöllner et al. 2006; Mulyanto et al. 2011). HBV sequences were not detected in samples from the HBV negative individual, but were clearly present in both tumor and non-tumor liver samples from the HBV positive patients (Fig.1AB). We found that these patients were infected by three different strains of HBV (B, C and D genotypes, Supplemental Material Section 4) based on the clear majority representation, in each patient, of only one out of the constellation of HBV variants.

The total number of viral reads is substantially higher in the HCC tumors than in their matching non-tumor liver tissue (Fig. 1A), at similar overall coverage (Supplemental Table 1). Based on total number of viral and human reads, we estimate that on average, the tumor samples contain at least 2 copies of the viral genome per diploid human genome. We identified HBV integration sites by searching for human-virus chimeric paired-end reads, where one end mapped to the human genome and the other mapped to the viral genome. We found such chimeric reads in both tumor and non-tumor liver tissue, but not in blood. The tumor samples again harbor a much higher number of chimeric reads (Fig. 1B) than the non-tumor samples. Chimeric reads supporting the same viral integration event were then clustered, yielding 255 unique HBV insertion sites in the three HBV positive patients (Supplemental Table 3A and Supplemental

Material Section 4), 48 of which are supported by multiple chimeric reads (Supplemental Table 3A).

Next we questioned whether the number of detected viral integration events is dependent on the sequencing depth. We selected patient 31656 for additional sequencing, bringing the total coverage to 234X for the tumor and 243X for the non-tumor sample. In total, we detected 142 integration sites in the tumor sample, and 136 sites in the non-tumor tissue (Supplemental Fig. 2A; Supplemental Table 4). After simulating lower-coverage sequencing from this high-coverage data, we found that the number of unique viral integrations detected was proportional to the sequencing depth for both the normal and tumor samples (Supplemental Fig. 2B; Supplemental Material Section 9). This supports many stochastic viral integration events without clonal expansion, which are likely to be underestimated by previous PCR-based approaches (Bréchot et al. 2000; Gozuacik et al. 2001; Saigo et al. 2008).

Both high-depth coverage (~80X) and ultra high-depth (~240X) sequencing reveals a heterogeneous, widespread viral integration landscape in tumor as well as in non-tumor liver tissue from HCC patients. However, HCC tumor samples and their adjacent non-tumor liver tissues exhibit strikingly distinct patterns of viral insertion (Fig. 1C). Based on high-depth coverage (~80X) data, we found non-tumor tissues contained 107 viral integration sites among the three HBV positive patients, each with 11 or fewer supporting chimeric reads. This suggests that viral DNA integration occurs commonly and at many sites in non-tumor liver tissue with HBV infection, resulting in a heterogeneous collection of insertion-carrying hepatocytes, each representing a small proportion of the population. In contrast, the tumor samples contained 148 insertion sites, with a small subset (9 sites) at much higher frequencies (supported by 23 or more chimeric reads, Fig.1C; Supplemental Table 3A), designated as Major Integration Sites (MIS, Supplemental Fig. 3; Supplemental Material Section 5). The tumor samples also contain large numbers of low-frequency viral insertion sites, likely due to contamination of non-tumor tissue or late viral integration in the expanding tumor. Occurrence of MIS in each of the three HBV positive HCC tumors we examined is most likely the result of clonal expansion of hepatocytes carrying these insertions, suggesting that the events leading to MIS occur fairly early during tumorigenesis. Therefore, any functional genomic impact of viral integration would be restricted to these few clonal sites.

Transcriptional and genomic impact of HBV integration

Detailed expression analysis, using RNA sequencing, of the genomic region flanking the most abundant MIS in each patient revealed a distinct transcriptional impact of viral integration. In patient 31107, we observed a MIS within the Mixed-Lineage Leukemia 4 (*MLL4*) gene, which has been previously reported as a recurrent HBV integration target among HCC patients (Saigo et al. 2008). *MLL4*, a histone-lysine N-methyltransferase, is a part of the ASC-2 complex implicated in the p53 tumor suppressor pathway (Lee et al. 2009). Other members of the MLL family are frequently mutated in solid tumors (Lee et al. 2008; Natarajan et al. 2010). This viral integration within the *MLL4* gene is accompanied by >20-fold increase in the *MLL4* transcript level (Fig. 2A; Supplemental Fig. 4A). Detailed sequence analysis reveals that the inserted HBV sequence contains two partial copies of the HBV genome, driving adjacent increased expression on both strands (Supplemental Fig. 4A). In patient H442, the most abundant viral insertion occurs ~10kb upstream of *ANGPT1* (Supplemental Table 3A), leading to >8-fold increase in expression than that of the matched liver sample (Fig. 2B). Overexpression of *ANGPT1* in HCC has been previously reported (Zeng et al. 2008). Interestingly, in patient 31656, the most abundant MIS is in a non-genic region. However, we observed novel transcription precisely next to the viral insertion site that was not observed in the samples without this viral insertion (Fig. 2C; Supplemental Fig. 5).

Consistent with the MIS described above, adjacent transcriptional activation appears to be a common feature of viral integration. We examined the relative RNA-seq read abundance between paired samples with and without viral insertion, regardless of tumor and non-tumor status. In half of the 48 viral integration sites investigated, there is obvious local transcriptional activation. This effect is usually directional (Fig. 2D; Supplemental Material Section 7), suggesting that some of the observed transcriptional activity changes can be attributed to the “run-through” of viral transcripts into human sequences. Indeed, paired-end RNA-seq data reveal the strong presence of chimeric transcripts between the viral HBx gene and the human *MLL4* gene in patient 31107 and with the non-genic region in patient 31656.

We found that besides direct expression alteration, viral insertion can also lead to genomic instability and introduce copy number changes, indirectly affecting gene expression. As detailed above, the most abundant MIS in the tumor sample of patient 31656 shows local increased human transcript level of an unannotated genomic locus in the immediate vicinity (Fig. 2C; Fig. 3AB). DNA copy number analysis in this region revealed that this viral integration colocalized precisely with the junction of a large DNA copy number loss (chr11q22.3, Fig. 3A). Interestingly, this

deletion leads to the heterozygous loss of a cluster of Caspases (*CASP12*, *CASP4*, *CASP5*, *CASP1*) and Caspase recruitment domain family genes (*CARD16* and *CARD17*), proteases that play a central role in the execution phase of cell apoptosis. RNA-seq read coverage shows that a string of these genes are down regulated in the tumor sample with this viral insertion (Fig. 3B). We reason that a random viral insertion at this site led to genomic instability, resulting in the loss of a large adjacent genomic region, likely due to non-allelic recombination between two copies of integrated viral sequences.

HBV integration at the DR1 site favors fusion transcripts

Viral-human fusion transcripts are prevalent, based on the number of viral-human chimeric RNA-seq reads. Many such events were detected in both tumor and non-tumor tissues (Fig. 4A). Based on fusion transcripts supported by 2 or more chimeric reads, we found that the viral arms of chimeric reads map preferentially to a region between 1500-2000 base pairs on the viral genome (Fig. 4A). Specifically, the fusion junctions obtained from RNA-seq reads, predominantly map near the direct repeat 1 (DR1) region located towards the end of the HBx gene (Fig. 4B). In contrast, the majority of viral integration junctions obtained from DNA-seq are close to either DR1 or DR2. We also observe a drop in viral transcription downstream of the DR1 region, (Supplemental Fig. 6; Supplemental Material Section 3). The DR1 and DR2 regions have been previously found to be involved in multiple insertion events (Dejean et al. 1984; Mason et al. 2010). We again examined the transcriptional change between paired samples with and without viral-human transcript chimeras, and confirmed the trend of unidirectional human transcriptional activation at the sites of RNA fusion (Fig. 4C). In contrast to the strong preference for the viral DR1 site, the human sequences in these chimeric RNA-seq reads map to many distinct locations in the human genome (Fig. 4A). No bias in terms of genomic location or local sequence preference for HBV integration was observed in the human genome, suggesting a stochastic viral DNA integration model (Supplemental Fig. 7; Supplemental Material Section 6).

It is worth noting that the preferred DR1 site (1824-1834 bp) of viral-human fusion transcripts is located near the 3’ end of the HBx gene (1374-1838 bp), just before its stop codon. Therefore many fusion transcripts may extend the open reading frame of the HBx gene. We examined a number of individual cases by local *de novo* assembly of chimeric RNA-seq reads (Supplemental Material Section 8), and identified 76 precise fusion junctions (Supplemental Table 5). Although the biological consequence of most of these potentially elongated proteins is not clear, it is intriguing that the viral insertion within the *MLL4* gene leads to the formation of an in-frame

fusion between HBx and truncated *MLL4* (Supplemental Fig. 4). Although the overall *MLL4* transcription output is much higher in the affected genome (Fig. 2A), the resulting fusion transcript lacks the AT-hook DNA-binding domain of *MLL4* (Supplemental Fig. 4B). We speculate that this over-expressed fusion product acts as a dominant negative allele, perhaps replacing the normal *MLL4* protein in the ASC-2 tumor suppressor complex without conferring its normal DNA binding activity.

Mutation spectrum of HCC revealed by whole genome sequencing

In addition to identifying viral integrations, whole genome sequencing also provides us the opportunity to identify other somatic alterations that may not be directly related to viral integration in these tumor genomes, and to compare the somatic changes between HBV-infected and non-infected HCC patients. To obtain a collection of high-confidence somatic mutations, we systematically identified somatic single-base mutations, and attempted to validate (Supplemental Fig. 8) a large number (1319) of these mutations (Supplemental Material Section 10). The three HBV positive tumors had 3180 to 5862 somatic point substitutions, while the HBV negative patient had 6362 such substitutions. The number of non-synonymous mutations in these 4 HCC samples ranges from 22 to 54 (Supplemental Table 6). The only gene mutated in all 4 tumors is *TP53* (Supplemental Fig. 9), with all four patients carrying predicted protein-altering mutations in *TP53* (Supplemental Table 7). The nucleotide substitution pattern (designated as mutation signature) that we observed in HCC (Fig. 5A) is distinct from that associated with tobacco smoking (Lee et al. 2010; Pleasance et al. 2010b) or UV damage (Pleasance et al. 2010a)(Fig. 5A). The most prevalent substitutions are A>G and C>T transition events, a pattern similar to the one recently reported in a hepatitis C virus infected HCC patient (Totoki et al. 2011). The mutation signature in HCC patients was the same, irrespective of HBV infection status.

We computationally predicted (Fig. 5B) a large number of structural variations and then experimentally validated a subset of these structural variations (Supplemental Table 8; Supplemental Material Section 11) in the four HCC patients. The number of intra-chromosomal somatic structural variations in the uninfected individual was at least 10-fold higher than the HBV-infected individuals. Similarly, the uninfected individual carried at least a 3-fold higher number of inter-chromosomal structural variations compared with the infected individuals (Fig. 5B). Noteworthy experimentally confirmed structural variations included a fusion between *AXIN1* and *LUC7L* genes in patient H442, resulting in a truncated *AXIN1* and up-regulation of *LUC7L* (Supplemental Fig. 10). Point mutation and/or epigenetic silencing of *AXIN1*, a Wnt

antagonist, were reported in various types of human solid tumors (Baeza et al. 2003; Segditsas and Tomlinson 2006; Zucman-Rossi et al. 2007). In this case, the truncated *AXIN1* presumably disrupts the normal function of APC-depend destruction complex and consequently activates the Wnt signaling pathway. It is not clear whether *LUC7L* gene, the fusion partner, plays any functional role in liver cancer development except that it was reported as a survival predictor of breast cancer (Crawford et al. 2008). DNA copy number variation (CNV) and allele-imbalance (AIB/LOH) (Supplemental Material Section 12) indicated a copy number loss of *TP53* (chr17p13 region) in all four patients, and amplification of *CCND1* (chr11q13 region) in the HBV negative patient H384 (Supplemental Fig. 11).

Interestingly, in HBV-infected patients, the MIS sites tend to coincide with boundaries of copy number alterations (Fig. 6A-C, Supplemental Table 9), suggesting an underlying mechanistic connection between HBV integration and genomic instability. The association of MIS to copy number boundaries is statistically significant ($p < 10^{-5}$, Supplemental Material Section 13). We note that the single HBV negative patient we sequenced has a higher rate of mutation and larger number of structural variations (Supplemental Table 6; Fig. 5B). Incidentally, genes involved in telomere maintenance (*PARP1*, *BLM* and *MLH3*) were mutated in this patient, but not in the HBV positive patients (Supplemental Fig. 9). In addition, the same patient showed a pattern of structural catastrophe typical of chromothripsis (Stephens et al. 2011) (chromosome 11, Fig. 6D), while the HBV-infected patients did not show such an event. We believe the higher mutation rate that we observed in the HBV negative patient is likely due to mutations of telomere maintenance genes rather than the HBV status.

Discussion

HCC is a consequence of multiple complex mutation processes (Farazi and DePinho 2006). To our best knowledge, this study provides the first comprehensive analysis of multiple dimensions of genomic alterations in HBV-infected and uninfected HCC patients, including viral integration, single nucleotide changes and large genomic alterations (Fig. 6). While the conventional PCR based methods can be used to detect the presence of viral integration, only a small subset of insertions can be detected (Saigo et al. 2008), or only insertions close to targeted human (Murakami et al. 2005) or viral sequences (Mason et al. 2010) can be found. Whole genome and transcriptome sequencing provides an unbiased and sensitive method for comprehensively identifying viral insertion events and quantifying their frequencies, thus providing the first opportunity to interrogate the global extent of viral impact on the human genome and

transcriptome. We found that HBV integration occurs frequently in both tumor and non-tumor hepatocytes, but they show distinct patterns of integration. Clonal expansion of MIS-carrying hepatocytes was found specifically in the tumor samples but not in the matched liver samples. However, a heterogeneous background population of cells harboring low-frequency viral integrations was detected both in the tumor and matched liver samples. This finding is consistent with a random integration model followed by a positive selection of MIS-carrying hepatocytes during hepatocarcinogenesis, resulting more virus-integrated hepatocytes (clonally expanded sub population) in the tumor samples when compared to their matched non-tumor counterparts. We argue that the impact of HBV integration is multi-faceted. Given the observed stochastic nature of viral integration, it appears that HBV integration effectively surveys the human genome, exerting insertional mutation pressure, and thus may expand the oncogenic opportunities for patients infected by HBV. In these samples, the most dominant HBV integration sites occur within the *MLL4* gene, a frequently observed target for HBV insertion among HCC patients (Saigo et al. 2008); near the *ANGPT1* gene, a key player in angiogenesis; and next to the cluster of Caspase genes on chromosome 11, causing a copy number loss at this locus. Recurrence of integration sites in the *MLL4* gene argues for a causative role of HBV integration in HCC. Other recurrent integration sites such as one within the *h-TERT* gene, *PDGFRB* and *MAPK1* have also been reported (Paterlini-Bréchot et al. 2003; Murakami et al. 2005). We also examined the gene expression profile of *ANGPT1* in an independent large collection of tumor samples across thirty-five different types of tissues. Significant over-expression of *ANGPT1* was found specifically in the liver cancer samples, which argues *ANGPT1* might play important functional role during tumorigenesis in a tissue-specific manner (Supplemental Fig. 12). In order to check if the deletion at the chromosome 11 Caspase locus is an isolated event, we examined copy number alteration in two independent liver cancer data sets (GSE34957 and GSE9829) and found about ~10% of liver cancer samples in these two data sets show copy number loss on the Caspase locus (Supplemental Fig. 13).

The whole genome sequencing data show that viral integration frequently occurs near the DR1 or DR2 sites (Fig. 4B). A hot spot of integration on the viral genome might suggest a sequence specific integration mechanism. However, on the human side, we did not observe any specific sequence features near the fusion breakpoints (Supplemental Fig. 7). Since DRs are 5' ends of the minus and plus DNA strands of the linearized HBV genome, it is reasonable to argue that the frequent use of the DRs as integration sites is likely due to the preferred use of the free ends of replication intermediates of HBV. Although fusions can occur close to both DR1 and DR2, the

viral-human fusion transcripts were strongly biased to regions near DR1 (Fig. 4B). Previous studies (Guo et al. 1993; Raney and McLachlan 1997; Yu and Mertz 2001) reported several important cis-elements near the DR1 regions, such as the enhancer II, the preC promoter and the hormone response element. Members of the nuclear receptors, a superfamily of transcription factors, can bind to DR1 hormone response element and regulate the transcription and replication of HBV. A basepair-level view of fusion transcripts breakpoints (Supplemental Fig. 14) shows that the majority of fusion breakpoints mapped close to DR1, resulting in intact DR1 cis-elements. In contrast to integration using DR2 as breakpoint, a fusion at DR1 region will also juxtapose the DR1 cis-elements close to the flanking human genomic sequences. We therefore speculate that intact DR1 cis-elements and a short physical distance between DR1 elements and the human fusion partner are two critical factors for a successful viral-human transcript formation. In addition, the linearization of viral genome after integration also explains the significant down-regulation of Polymerase (Supplemental Fig. 6 right panel), a gene otherwise located downstream of the DR1 region in a circular viral genome. In the linearized viral genome, the Polymerase gene would be at the 5' end, disconnected from the cis-elements upstream of the DR1 region, which would now be at the 3' end of the linearized genome.

In summary, it is evident that viral integration affects the human genome via insertional mutagenesis, viral promoter-driven transcriptional upregulation and induction of genomic instability. Despite the diversity of insertion sites and their varied effects, it is conceivable that virus-mediated mutagenesis functions, in conjunction with other genomic alterations such as *TP53* mutation, to drive hepatocarcinogenesis. Our observations support a model wherein the frequent assault of the human genome by widespread viral integration significantly widens "oncogenic opportunities" in patients with chronic HBV infection.

Methods

Sample preparation

All specimens were obtained from patients with appropriate consent. Tissue samples were examined by pathologists. All tumor samples contained greater than 80% of tumor content. The HBV infection status of patient samples was confirmed by a polymer chain reaction (PCR) assay. The DNA and RNA were extracted by using standard DNA/RNA extraction kit (Supplemental Material section 1).

Whole genome and transcriptome sequencing

Whole genome DNA paired-end sequencing was performed by "unchained combinatorial probe anchor ligation sequencing", as described previously (Drmanac et al. 2010). The average coverage was greater than 80X. The single nucleotide variation (SNV) calls for each sample with respect to the human reference genome (NCBI Build37) were made as described previously (Drmanac et al. 2010). Detection of structural variation (SV), copy number variation (CNV) and loss of heterozygosity (LOH) are described in Supplemental Material section 11 and 12.

Transcriptome sequencing of both tumor and matched liver samples was performed on the Illumina HiSeq Platform using standard paired-end protocol. On average, 25-35 million 75 base pairs reads were obtained per sample. The reads were mapped to both the human genome and a collection of HBV reference genomes by GSNAP (Wu and Nacu 2010). The number of reads mapped to the exons of each RefSeq gene was calculated and the corresponding RPKM (reads mapping to the genome per kilobase of transcript per million reads obtained from sequencing) value was derived (Supplemental Material section 3).
Statistical analysis of differentially expressed genes was performed using the DEseq package from Bioconductor (Anders and Huber 2010) (Supplemental Material section 14)

HBV integration detection

We first selected a HBV reference genome from a collection (n=73) HBV reference sequences by finding the best match. We then utilized the paired-end nature of reads to search for human-virus chimeric reads, an indication of HBV integration in the human genome. Adjacent chimeric reads were clustered to obtain non-redundant integration events (Supplemental Material section 4).

Experimental validation

Validation of candidate single nucleotide variants was performed by Sequenom MassARRAy platform (Supplemental Material section 10). Validation of HBV insertions and a subset of structural variations was performed by polymer chain reaction (PCR) followed by Sanger sequencing (Supplemental Material section 4 and 11).

Figure Legends

Figure 1. Tumor-specific clonal expansion of virus-integrated hepatocytes in HCC.

For each sequenced human genome, viral integration is quantified as (*A*) the total number of paired-end reads where at least one arm maps to the HBV genome, (*B*) the number of human-

viral chimeric reads, and (*C*) the number of chimeric reads as a function of the genomic location of HBV integration sites in the human genome. In contrast to the matched non-tumor samples, tumor samples carry a few loci with substantially larger number of chimeric reads. T: Hepatocellular carcinoma tumor samples, N: Matched non-tumor liver samples. The legend in panel (*C*) indicates internal identifiers for the three HBV positive patients in this study.

Figure 2. Transcriptional effect of HBV integration on the human genome.
Local transcriptional effect of HBV integration on the human genome is shown, for the most abundant integration site for each patient (*A*-*C*), and for all integration sites (*D*), based on RNA-seq data. (*A*) *MLL4* is highly over-expressed in the tumor sample with the HBV integration event (31107). (*B*) Substantial over-expression of *ANGPT1* in the tumor from patient H442 with HBV integration upstream (~10kb) of *ANGPT1*. (*C*) A novel human-viral fusion transcript in the tumor sample with HBV integration in a non-genic region (patient: 31656). Asterisks indicate samples with viral integration. (*D*) Transcriptional effect at human-viral junctions defined by DNA-seq. The human-viral junctions supported by multiple DNA-seq chimeric reads (n=48) are represented as a dotted line at the center. We then used the RNA-seq data to infer the transcriptional changes on each side of these junctions. The color in the heatmap represents the fold-change for each interval, measured as the difference in the generalized log of the RPKM of the altered genome (i.e. the genome containing the viral insertion) versus the unaltered genome. Samples carrying the insertion are indicated as either N (Non-tumor) or T (Tumor). The rows were grouped by hierarchical clustering.

Figure 3. Genomic instability at the viral integration site near the Caspase locus.
(*A*) Copy number and transcription around a chr11q22 HBV integration site in patient 31656. The top panel shows normalized, GC-corrected copy number values in 50kb windows with the horizontal red line representing the resulting copy number segments. There is a copy number breakpoint right at the HBV integration site, with a copy number loss of the chromosomal region 3' of the integration site. Red and blue bar plots show transcription (RNA-seq read coverage) in this locus in the tumor and matched normal, respectively. (*B*) Genes closest to the integration site within this deletion were significantly downregulated, including *CASP12*, *CASP4*, *CASP5*, *CASP1*, *CARD16* and *CARD17*.

Figure 4. Viral-human fusion transcripts are common in both HCC and non-tumor samples.

(*A*) The genomic coordinate on the HBV genome of each chimeric RNA-seq read is plotted against its genomic coordinate on the human genome (linearized after concatenating all chromosomes). Only locations supported by two or more chimeric reads are shown. (*B*) Viral junctions determined from clusters of two or more chimeric reads are shown as vertical bars, with part of the remaining integrated viral sequence (50 bp) indicated as horizontal lines. Reads from both RNA- and DNA-seq are shown. A large majority of the RNA-seq junctions are in close vicinity (10 bp) of the DR1 (Direct repeat 1) region on the HBV genome. Clusters from non-tumor liver samples are indicated in blue whereas clusters from tumor samples are in red. (*C*) A global view of the transcriptional consequence of viral integration on the flanking human genome. The data was organized in the same manner as Figure. 2D, except that the human junctions in this panel are based on RNA-seq data instead of DNA-seq data. Most of the sites show strong unidirectional transcriptional upregulation, starting at the integration site, while relatively fewer sites correlate with non-directional transcriptional downregulation or upregulation.

Figure 5. Mutation signature and structural variations in HCC patients.
(*A*) High confidence, somatic single-base substitutions were classified into all six categories of base substitutions. The fraction of mutations belonging to each category is shown for the four HCC patients, and compared to signatures previously found in NSCLC (Lee et al. 2010), SCLC (Pleasance et al. 2010b), melanoma (Pleasance et al. 2010a) and in germline variations. (*B*) Number of predicted structural variations detected in both HBV positive and negative HCC patients. Intra: Intra-chromosomal SVs and Inter: Inter-chromosomal SVs.

Figure 6. Summary of somatic genomic alterations in HCC patients.
Various types of somatic alterations in the four HCC patient genomes using circos plots (Krzywinski et al. 2009)(*A-D***)**. High confidence somatic structural variations (SVs) are shown as lines, with red lines representing interchromosomal SVs and blue lines indicating intrachromosomal SVs. Regions of loss of heterozygosity and allelic imbalance are illustrated as green bars. Somatic copy number alterations are shown as bar plots with copy number gain shown in red and copy number loss in blue (the scale ranges from -2 to 4). Each surrounding red dot represents number of high-confidence somatic SNVs within a 1-million basepair window. Major HBV integration sites are illustrated as triangles. Patient's identifier and HBV status is shown at the center of each circular view.

Data Access

Sequencing data can be accessed at dbGAP (Accession ID phs000384.v1.p1). Copy number data for HCC patients are accessible at GEO (GEO Series Accession GSE34957).

Acknowledgments

We thank Eric Brown for valuable discussion; Peter Dijkgraaf, Julie Rae, Carlo Santos and May Wittke for sample handling; Robert Soriano for generation of microarray data; Florian Gnad, Kiran K. Mukhyala, Colin Watanabe, Jim Fitzgerald, Meg Green and Albion Baucom for computational assistance.

Author Contributions

ZJ: study design, project coordination, overall data analysis, preparation of manuscript; SJ: overall data analysis, preparation of manuscript; JL: transcriptome sequencing data analysis, preparation of manuscript; PMH: CNV and LOH data analysis, preparation of manuscript; WL: viral integration breakpoint analysis, preparation of manuscript; MIK, KPP and PC: whole genome sequencing and HBV integration site analysis; YG and ZM: PCR validation of structural variations and viral integration sites; JS and SS: mutation experimental validation; JD and SBK: HBV status validation sample preparation; HMS and SJ: sample handling and histopathological evaluation; SY, AJ and WY: Experimentation studies on ANGPT1; DGB, RCG, FJD: project coordination and manuscript critiques; ZZ: study design, data interpretation and manuscript preparation.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**(10): R106.

Baeza N, Masuoka J, Kleihues P, Ohgaki H. 2003. AXIN1 mutations but not deletions in cerebellar medulloblastomas. *Oncogene* **22**(4): 632-636.

Block TM, Mehta AS, Fimmel CJ, Jordan R. 2003. Molecular viral oncology of hepatocellular carcinoma. *Oncogene* **22**(33): 5093-5107.

Bouchard MJ, Navas-Martin S. 2011. Hepatitis B and C virus hepatocarcinogenesis: Lessons learned and future challenges. *Cancer Lett* **305**(2): 123-143.

Bréchot C, Gozuacik D, Murakami Y, Paterlini-Bréchot P. 2000. Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). *Semin Cancer Biol* **10**(3): 211-231.

Chemin I, Zoulim F. 2009. Hepatitis B virus induced hepatocellular carcinoma. *Cancer Lett* **286**(1): 52-59.

Crawford NP, Walker RC, Lukes L, Officewala JS, Williams RW, Hunter KW. 2008. The Diasporin Pathway: a tumor progression-related transcriptional network that predicts breast cancer survival. *Clin Exp Metastasis* **25**(4): 357-369.

Dejean A, Sonigo P, Wain-Hobson S, Tiollais P. 1984. Specific hepatitis B virus integration in hepatocellular carcinoma DNA through a viral 11-base-pair direct repeat. *Proc Natl Acad Sci U S A* **81**(17): 5350-5354.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**(5961): 78-81.

Farazi PA, DePinho RA. 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nat Rev Cancer* **6**(9): 674-687.

Gozuacik D, Murakami Y, Saigo K, Chami M, Mugnier C, Lagorce D, Okanoue T, Urashima T, Bréchot C, Paterlini-Bréchot P. 2001. Identification of human cancer-related genes by naturally occurring Hepatitis B Virus DNA tagging. *Oncogene* **20**(43): 6233-6240.

Guo W, Chen M, Yen TS, Ou JH. 1993. Hepatocyte-specific expression of the hepatitis B virus core promoter depends on both positive and negative regulation. *Mol Cell Biol* **13**(1): 443-448.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9): 1639-1645.

Lavanchy D. 2004. Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *J Viral Hepat* **11**(2): 97-107.

Lee J, Kim D-H, Lee S, Yang Q-H, Lee DK, Lee S-K, Roeder RG, Lee JW. 2009. A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc Natl Acad Sci U S A* **106**(21): 8513-8518.

Lee S, Lee J, Lee S-K, Lee JW. 2008. Activating signal cointegrator-2 is an essential adaptor to recruit histone H3 lysine 4 methyltransferases MLL3 and MLL4 to the liver X receptors. *Mol Endocrinol* **22**(6): 1312-1319.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D et al. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**(7297): 473-477.

Mason WS, Liu C, Aldrich CE, Litwin S, Yeh MM. 2010. Clonal expansion of normal-appearing human hepatocytes during chronic hepatitis B virus infection. *J Virol* **84**(16): 8308-8315.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**(10): 685-696.

Michielsen P, Ho E. 2011. Viral hepatitis B and hepatocellular carcinoma. *Acta Gastroenterol Belg* **74**(1): 4-8.

Mulyanto, Depamede SN, Wahyono A, Jirintai, Nagashima S, Takahashi M, Okamoto H. 2011. Analysis of the full-length genomes of novel hepatitis B virus subgenotypes C11 and C12 in Papua, Indonesia. *J Med Virol* **83**(1): 54-64.

Murakami Y, Saigo K, Takashima H, Minami M, Okanoue T, Bréchot C, Paterlini-Bréchot P. 2005. Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. *Gut* **54**(8): 1162-1168.

Natarajan TG, Kallakury BV, Sheehan CE, Bartlett MB, Ganesan N, Preet A, Ross JS, Fitzgerald KT. 2010. Epigenetic regulator MLL2 shows altered expression in cancer cell lines and tumors from human breast and colon. *Cancer Cell Int* **10**: 13.

Paterlini-Bréchot P, Saigo K, Murakami Y, Chami M, Gozuacik D, Mugnier C, Lagorce D, Bréchot C. 2003. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. *Oncogene* **22**(25): 3911-3916.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR et al. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**(7278): 191-196.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C et al. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**(7278): 184-190.

Raney AK, McLachlan A. 1997. Characterization of the hepatitis B virus major surface antigen promoter hepatocyte nuclear factor 3 binding site. *J Gen Virol* **78 (Pt 11)**: 3029-3038.

Saigo K, Yoshida K, Ikeda R, Sakamoto Y, Murakami Y, Urashima T, Asano T, Kenmochi T, Inoue I. 2008. Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. *Hum Mutat* **29**(5): 703-708.

Segditsas S, Tomlinson I. 2006. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**(57): 7531-7537.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T et al. 2011. High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* **43**(5): 464-469.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7): 873-881.

Yu X, Mertz JE. 2001. Critical roles of nuclear receptor response elements in replication of hepatitis B virus. *Journal of virology* **75**(23): 11354-11364.

Zeng W, Gouw ASH, van den Heuvel MC, Zwiers PJ, Zondervan PE, Poppema S, Zhang N, Platteel I, de Jong KP, Molema G. 2008. The angiogenic makeup of human hepatocellular carcinoma does not favor vascular endothelial growth factor/angiopoietin-driven sprouting neovascularization. *Hepatology* **48**(5): 1517-1527.

Zöllner B, Feucht H-H, Sterneck M, Schäfer H, Rogiers X, Fischer L. 2006. Clinical reactivation after liver transplantation with an unusual minor strain of hepatitis B virus in an occult carrier. *Liver Transpl* **12**(8): 1283-1289.

Zucman-Rossi J, Benhamouche S, Godard C, Boyault S, Grimber G, Balabaud C, Cunha AS, Bioulac-Sage P, Perret C. 2007. Differential effects of inactivated Axin1 and activated beta-catenin mutations in human hepatocellular carcinomas. *Oncogene* **26**(5): 774-780.

Figure 1

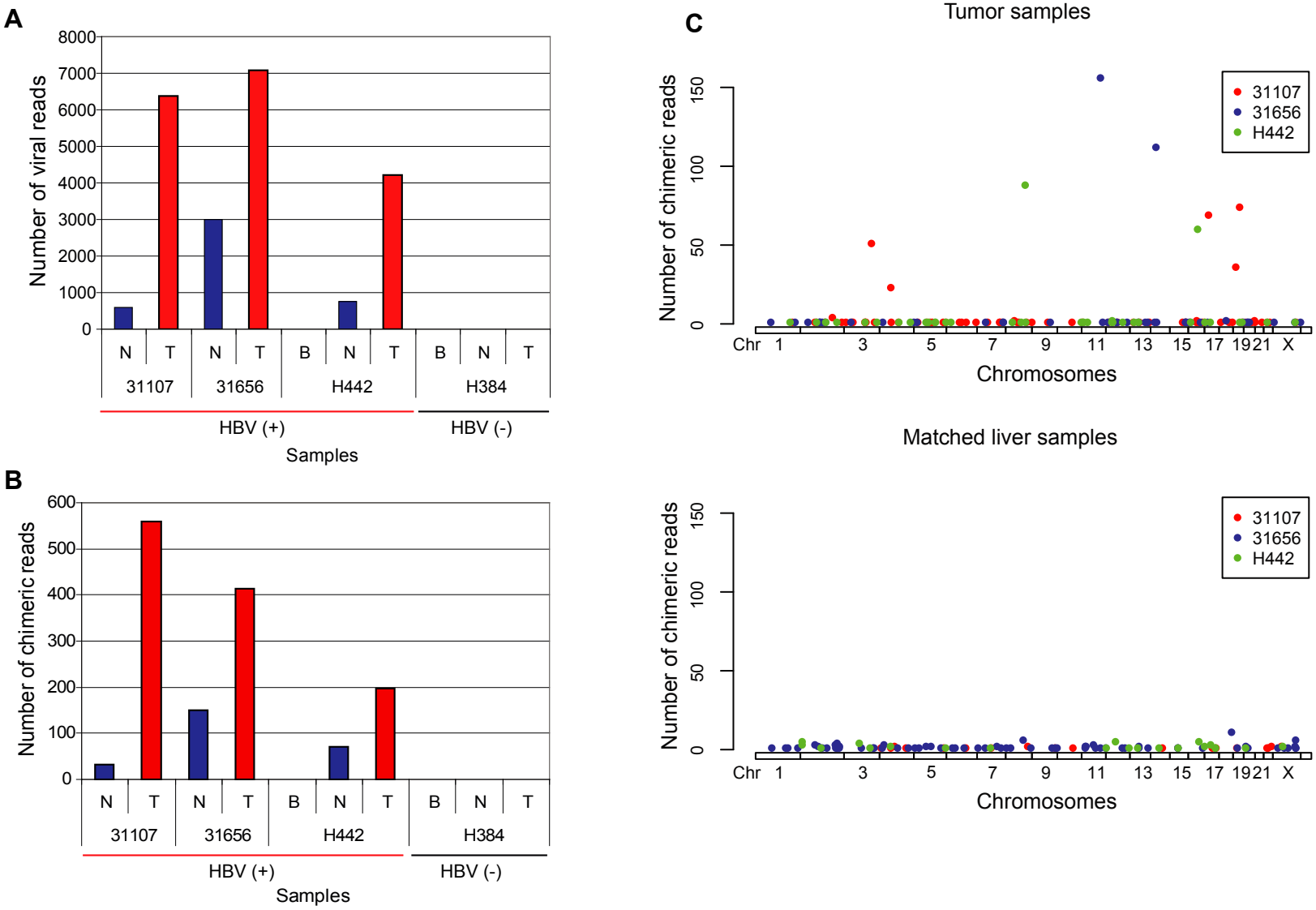

A
Number of viral reads
8000
7000
6000
5000
4000
3000
2000
1000
0
N T N T B N T B N T
31107 31656 H442 H384
HBV (+)
HBV (-)
Samples
B
Number of chimeric reads
600
500
400
300
200
100
0
N T N T B N T B N T
31107 31656 H442 H384
HBV (+)
HBV (-)
Samples
C
Tumor samples
Number of chimeric reads
150
100
50
0
31107
31656
H442
Chr 1 3 5 7 9 11 13 15 17 19 21 X
Chromosomes
Matched liver samples
Number of chimeric reads
150
100
50
0
31107
31656
H442
Chr 1 3 5 7 9 11 13 15 17 19 21 X
Chromosomes

Figure 2

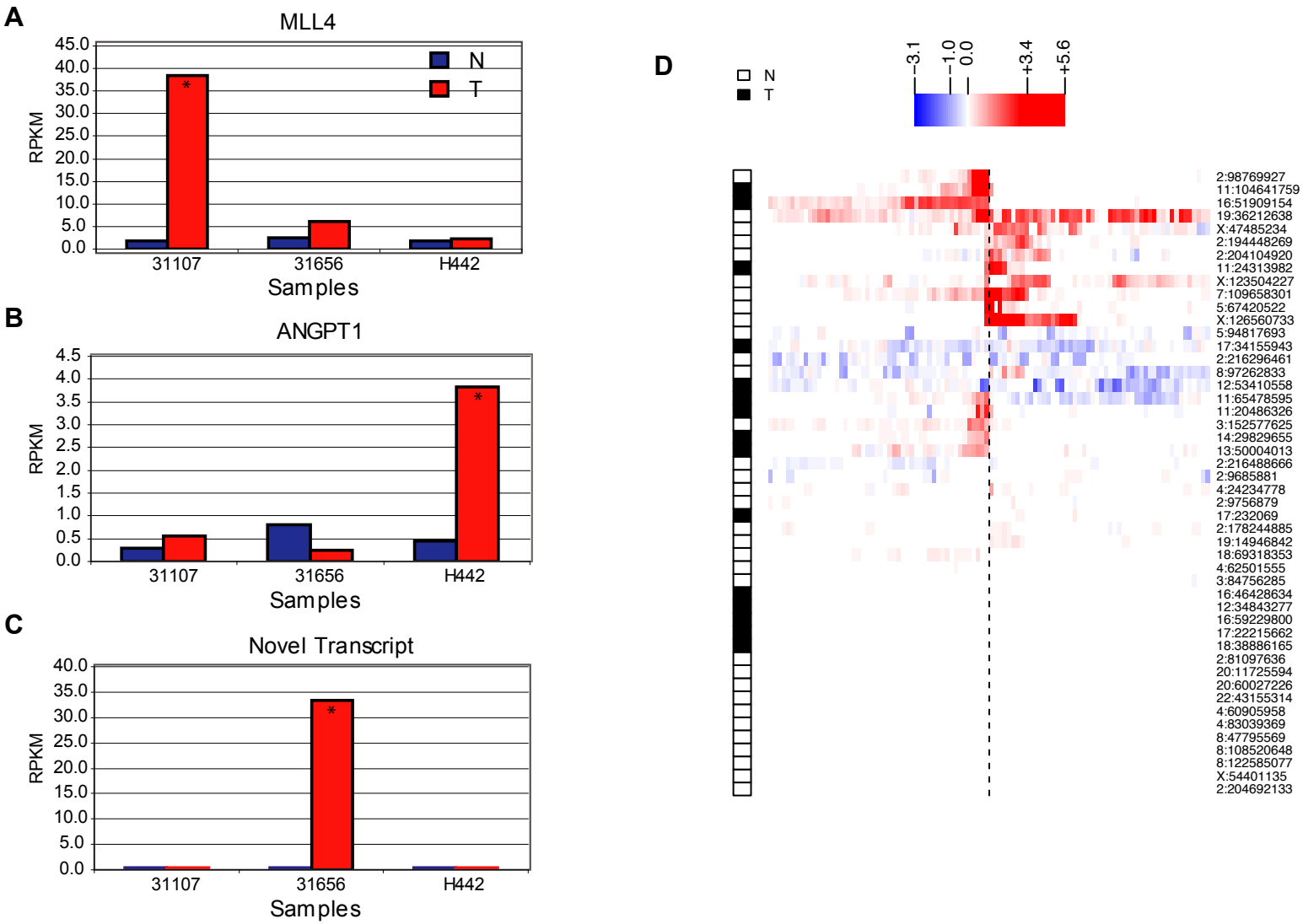
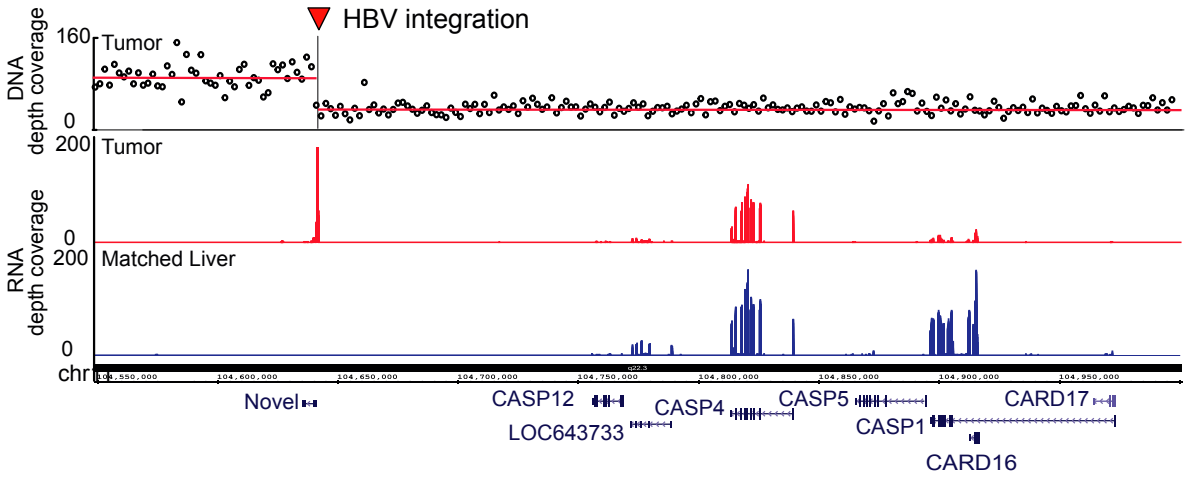

A
MLL4
RPKM
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
N
T
*
31107
31656
H442
Samples
B
ANGPT1
RPKM
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
*
31107
31656
H442
Samples
C
Novel Transcript
RPKM
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
*
31107
31656
H442
Samples
D
N
T
-3.1
-1.0
0.0
+3.4
+5.6
2:98769927
11:104641759
16:51909154
19:36212638
X:47485234
2:194448269
2:204104920
11:24313982
X:123504227
7:109658301
5:67420522
X:126560733
5:94817693
17:34155943
2:216296461
8:97262833
12:53410558
11:65478595
11:20486326
3:152577625
14:29829655
13:50004013
2:216488666
2:9685881
4:24234778
2:9756879
17:232069
2:178244885
19:14946842
18:69318353
4:62501555
3:84756285
16:46428634
12:34843277
16:59229800
17:22215662
18:38886165
2:81097636
20:11725594
20:60027226
22:43155314
4:60905958
4:83039369
8:47795569
8:108520648
8:122585077
X:54401135
2:204692133

Figure 3

A



B

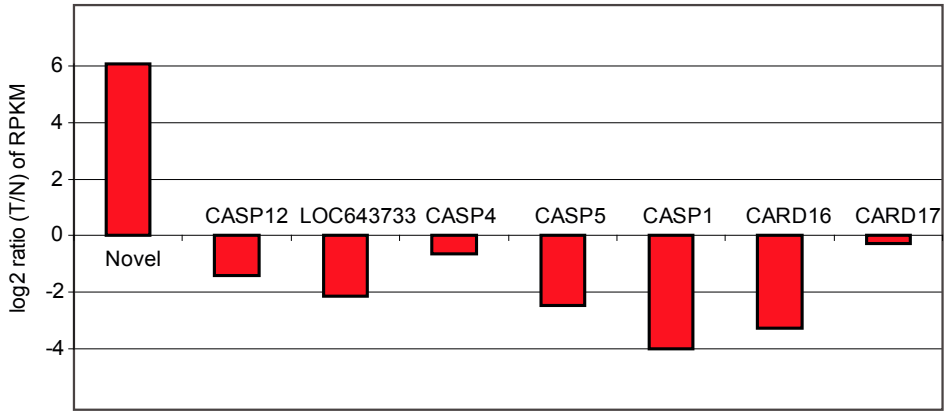


Figure 4

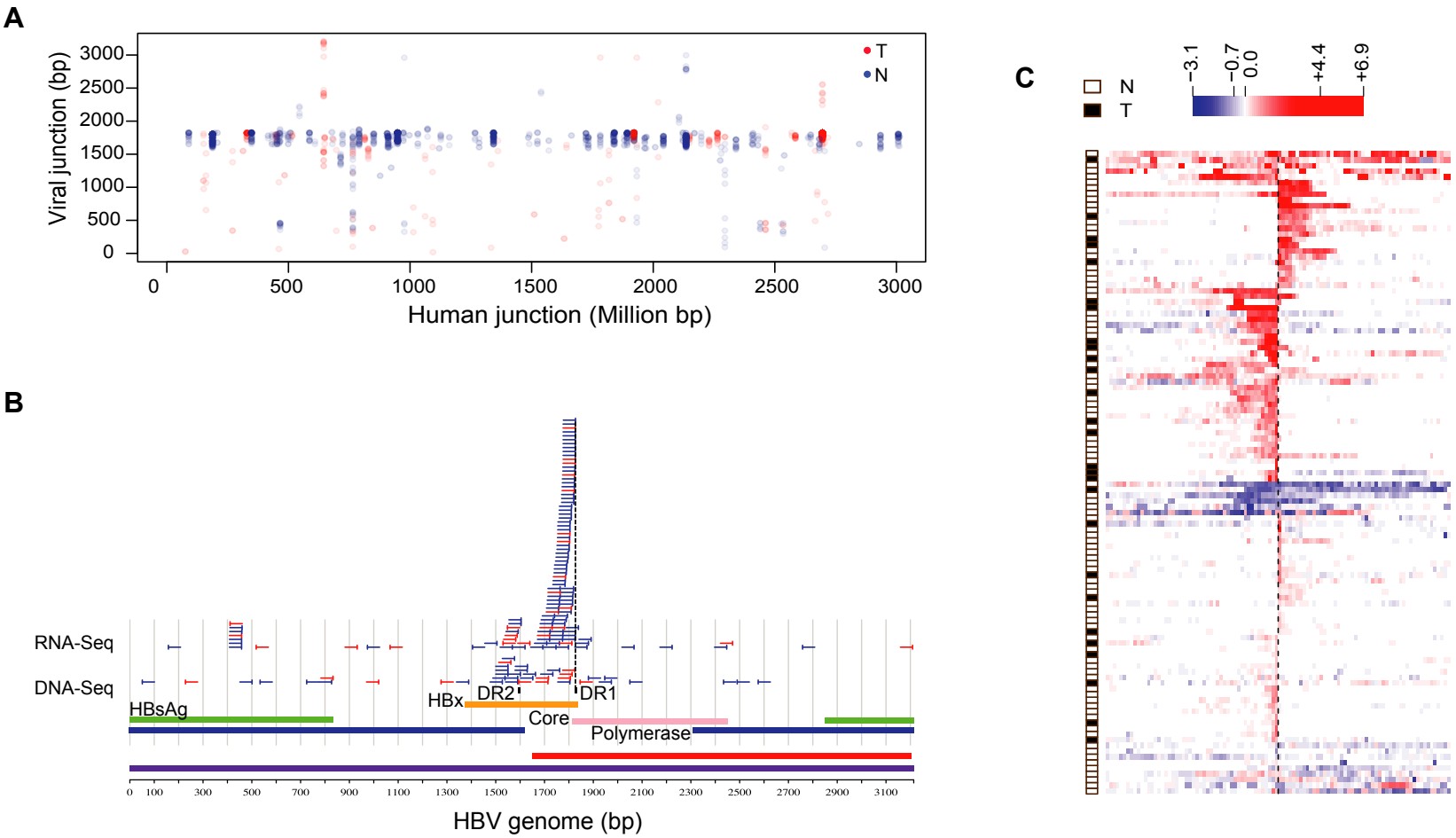

A
Viral junction (bp)
3000
2500
2000
1500
1000
500
0
T
N
0
500
1000
1500
2000
2500
3000
Human junction (Million bp)
B
RNA-Seq
DNA-Seq
HBsAg
HBx
DR2
DR1
Core
Polymerase
0
100
300
500
700
900
1100
1300
1500
1700
1900
2100
2300
2500
2700
2900
3100
HBV genome (bp)
C
N
T
−3.1
−0.7
0.0
+4.4
+6.9

Figure 5

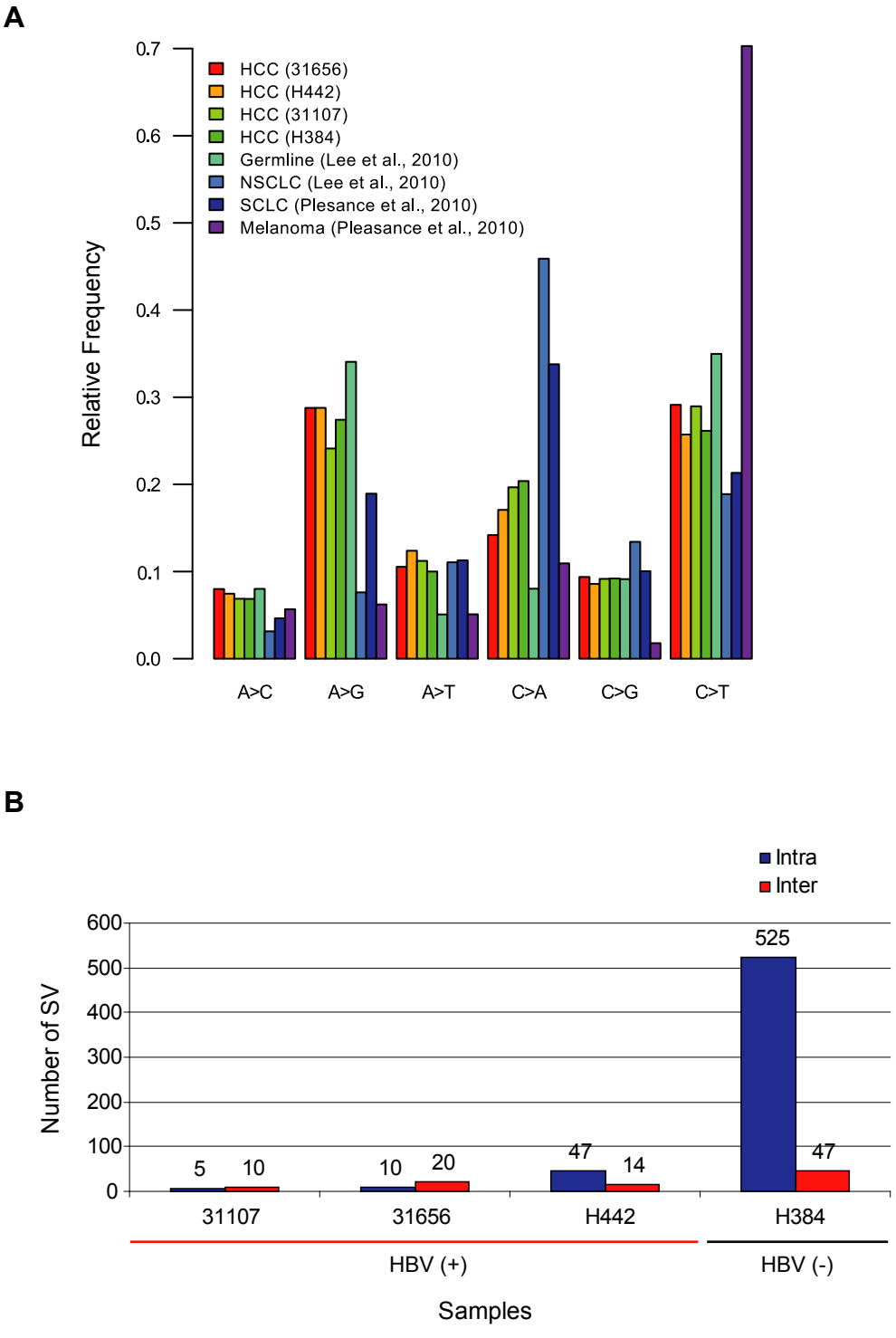


Figure 6