Relative Abundance of Integrant-Derived Viral RNAs in Infected Tissues Harvested from Chronic Hepatitis B Virus Carriers

Natalia Freitas,^a Tetyana Lukash,^a Sumedha Gunewardena,^b Benjamin Chappell,^a Betty L. Slagle,^c Severin O. Gudima^a

^aDepartment of Microbiology, Molecular Genetics and Immunology, University of Kansas Medical Center, Kansas City, Kansas, USA

^bDepartment of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, Kansas, USA

^cDepartment of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA

ABSTRACT Five matching sets of nonmalignant liver tissues and hepatocellular carcinoma (HCC) samples from individuals chronically infected with hepatitis B virus (HBV) were examined. The HBV genomic sequences were determined by using overlapping PCR amplicons covering the entire viral genome. Four pairs of tissues were infected with HBV genotype C, while one pair was infected with HBV genotype B. HBV replication markers were found in all tissues. In the majority of HCC samples, the levels of pre-genomic/precore RNA (pgRNA) and covalently closed circular DNA (cccDNA) were lower than those in liver tissue counterparts. Regardless of the presence of HBV replication markers, (i) integrant-derived HBV RNAs (id-RNAs) were found in all tissues by reverse transcription-PCR (RT-PCR) analysis and were considerably abundant or predominant in 6/10 tissue samples (2 liver and 4 HCC samples), (ii) RNAs that were polyadenylated using the cryptic HBV polyadenylation signal and therefore could be produced by HBV replication or derived from integrated HBV DNA were found in 5/10 samples (3 liver and 2 HCC samples) and were considerably abundant species in 3/10 tissues (2 livers and 1 HCC), and (iii) cccDNA-transcribed RNAs polyadenylated near position 1931 were not abundant in 7/10 tissues (2 liver and 5 HCC samples) and were predominant in only two liver samples. Subsequent RNA sequencing analysis of selected liver/HCC samples also showed relative abundance of id-RNAs in most of the examined tissues. Our findings suggesting that id-RNAs could represent a significant source of HBV envelope proteins, which is independent of viral replication, are discussed in the context of the possible contribution of id-RNAs to the HBV life cycle.

IMPORTANCE The relative abundance of integrant-derived HBV RNAs (id-RNAs) in chronically infected tissues suggest that id-RNAs coding for the envelope proteins may facilitate the production of a considerable fraction of surface antigens (HBsAg) in infected cells bearing HBV integrants. If the same cells support HBV replication, then a significant fraction of assembled HBV virions could bear id-RNA-derived HBsAg as a major component of their envelopes. Therefore, the infectivity of these HBV virions and their ability to facilitate virus cell-to-cell spread could be determined mainly by the properties of id-RNA-derived envelope proteins and not by the properties of replication-derived HBsAg. These interpretations suggest that id-RNAs may play a role in the maintenance of chronic HBV infection and therefore contribute to the HBV life cycle. Furthermore, the production of HBsAg from id-RNAs independently of viral replication may explain at least in part why treatment with interferon or nucleos(t)ides in most cases fails to achieve a loss of serum HBsAg.

KEYWORDS HBV life cycle, hepatitis B virus, chronic HBV infection, integrant-derived viral RNAs

Received 20 December 2017 Accepted 17 February 2018

Accepted manuscript posted online 28 February 2018

Citation Freitas N, Lukash T, Gunewardena S, Chappell B, Slagle BL, Gudima SO. 2018. Relative abundance of integrant-derived viral RNAs in infected tissues harvested from chronic hepatitis B virus carriers. *J Virol* 92:e02221-17. <https://doi.org/10.1128/JVI.02221-17>.

Editor Rozanne M. Sandri-Goldin, University of California, Irvine

Copyright © 2018 American Society for Microbiology. All Rights Reserved.

Address correspondence to Severin O. Gudima, sgudima@kumc.edu.

Human hepatitis B virus (HBV) is a prototype hepadnavirus. HBV is a significant human pathogen that causes transient and chronic liver infections. About 240 million people worldwide are chronic carriers of HBV. More than 50% of all known hepatocellular carcinoma (HCC) cases are associated with chronic HBV infection. Approximately 1 million chronic HBV carriers per year die because of advanced liver disease or HCC (1–8).

HBV is a small DNA virus that replicates via reverse transcription (RT). It infects hepatocytes using the sodium taurocholate-cotransporting polypeptide (NTCP) as the receptor (1, 9, 10). HBV encodes the small (S), middle (M), and large (L) envelope proteins, which all share the C-terminal S domain. Compared to S, M has an additional N-terminal pre-S2 domain. L has an extra N-terminal domain, pre-S1, compared to M. The N terminus of pre-S1 bears most of the infectivity determinants and interacts with NTCP. The coexpression of S and L is sufficient for the assembly and release of infectious virions (1, 9–12). The relaxed circular DNA (rcDNA) is the major endpoint product of viral reverse transcription. However, if strand transfer does not occur at the beginning of plus-strand DNA synthesis, *in situ* priming can take place. In this case, the final product of reverse transcription is a double-stranded linear DNA genome (DSL). The DSL is the main precursor of HBV DNA integration into the host chromosomal DNA. This integration is facilitated by the host DNA repair machinery and is considered a by-product of HBV replication. It is commonly assumed that the HBV life cycle does not depend on integrated HBV DNA. Integration is random and takes place during all phases of HBV infection. After years of chronic HBV infection, a considerable fraction of hepatocytes may bear integrated HBV DNA. Not all HBV integrants remain intact. Alterations such as mutations, deletions, insertions, and rearrangements were found in integrated HBV DNA (1, 9, 13–19). Progeny virus cannot be produced from integrants. While integrated HBV sequences can be transcribed, the roles of RNAs derived from integrated HBV sequences in relation to the HBV life cycle and pathogenesis are far from being well understood (14, 20–22).

Some integrants can serve as an alternative to the HBV replication source of the virus envelope proteins (i.e., surface antigens [HBsAg]). In this case, integrant-derived RNAs (id-RNAs) encoding the envelope proteins would contain the 5'-CAP, the 5' untranslated region (UTR), an open reading frame (ORF) for the HBV envelope protein (L or M/S), most of the 3' UTR of HBV, and a host sequence insert of an unknown length immediately next to the poly(A) tail (15, 16, 20, 21, 23, 24). Previously, we demonstrated that integrant-derived envelope proteins of HBV (id-HBsAg) were able to facilitate the assembly of infectious human hepatitis delta virus (HDV) (24), which is a subviral agent of HBV that uses HBsAg to form its own virions and to infect susceptible hepatocytes via the HBV receptor (10, 25, 26). Therefore, fully functional HBV envelope proteins can be translated from id-RNAs regardless of HBV replication.

In the present study, we examined the accumulation of HBV RNAs in matching liver and HCC tissues harvested from five untreated chronic HBV carriers. By using a conventional RT-PCR approach, id-RNAs were found in all tissues. Interestingly, unlike covalently closed circular DNA (cccDNA)-transcribed RNAs (i.e., RNAs produced by HBV replication), id-RNAs were considerably abundant in 6 out of 10 tissues (2 liver and 4 HCC samples), regardless of the fact that an HBV replication marker, such as cccDNA, was found in all analyzed tissues. In addition, we subsequently analyzed selected liver/HCC samples using RNA sequencing (RNA-seq), which showed that id-RNAs were predominant in most of the examined tissues, although a comparison of the results of RT-PCR and RNA-seq suggested that different areas of the same tissue can have different levels of cccDNA-derived RNAs and id-RNAs. Overall, our findings have several implications. (i) In cells that replicate HBV and efficiently produce unaltered id-RNAs, the virus envelope proteins produced from id-RNAs can considerably impact the properties of the assembled HBV virions. In other words, they can influence the virions' infectivity that in turn will affect the virus cell-to-cell spread and superinfection, which, according to our recent findings made with the woodchuck model (woodchucks infected with woodchuck hepatitis virus [WHV], which is closely related to HBV and is often used as an HBV

surrogate to study the mechanisms of HBV replication and infection [27]), could represent important determinants of the maintenance of chronic hepadnavirus infection (28). (ii) It also became apparent that id-HBsAg (especially if it is abundant) can play an important role in HBV-inflicted liver pathogenesis independently of HBV replication. Previous reports linked HBsAg to liver inflammation, HBV-induced liver disease, and carcinogenesis (12, 29–33). (iii) Guidelines for the treatment of chronic HBV carriers consider the loss of HBsAg to be the major goal of antiviral treatments, which current anti-HBV drugs (interferon and nucleos[t]ide analogs) have failed to achieve in the majority of cases (8). Our results suggest that the failure to achieve a loss of HBsAg could be due at least in part to the production of id-HBsAg, which occurs independently of viral replication and therefore cannot be suppressed by any of the FDA-approved anti-HBV drugs.

The above-mentioned considerations call for the inclusion of the potential contributions of id-RNAs and id-HBsAg into the model of the HBV life cycle. Follow-up studies that focus on further analysis of the properties of id-RNAs and their impact on the HBV life cycle and on HBV-related pathogenesis are therefore warranted. In addition, it became apparent that id-RNAs should now be considered important targets for novel anti-HBV interventions.

RESULTS

Sequences of HBV DNA genomes. We analyzed five pairs of nonmalignant liver tissue samples and matching HCC samples that were collected from untreated individuals chronically infected with HBV. For each pair, the odd number corresponds to a liver tissue, while the even number corresponds to an HCC sample. Thus, the analyzed pairs of tissues were 3/4, 7/8, 13/14, 25/26, and 39/40. The strategy for sequencing HBV genomes from the tissue samples via PCR amplification of three overlapping fragments covering the entire HBV genome is presented in Fig. 1A. Figure 1B shows the results of PCR amplifications of fragments B, F1, and F2, using DNAs isolated from liver/HCC tissues as the templates. PCR products of approximately the expected sizes were gel purified and cloned into the pCRII-TOPO vector, and the resulting plasmids were then subjected to sequencing. For tissue 13, during the amplification of fragment F1, a PCR product of an unexpectedly larger size was efficiently produced along with fragment F1 of approximately the expected size. The larger product was also gel purified, cloned, and sequenced. It appeared to be a recombinant in which (i) the 5' end contained 51 nucleotides (nt) of the expected correct HBV sequence (positions 1673 to 1723) downstream of the forward primer and (ii) the 3' end bore 21 nt of the expected sequence (positions 2276 to 2296) upstream of the reverse primer, but the middle part of the PCR product harbored an HBV fragment spanning positions 501 to 1555. The origin of such a recombinant was not immediately apparent. In addition, the F1 fragment for tissue 40 needed to be analyzed differently. Because of the uncovered mutations in the overlapping B/F1 region and an internal deletion, the F1 sequence was divided in two fragments, M and N, which were amplified and then subjected to cloning and sequencing as described above (Table 1). We used at least 7 sequences for each of the fragments for further analysis. Using these data and employing the sequences of the overlapping regions between the fragments, the consensus sequences of HBV genomes were assembled for each tissue. Figures 2 and 3 demonstrate the HBV genome consensus sequences for tissue 3 (nonmalignant liver tissue) and tissue 4 (matching HCC tissue) as examples. The consensus HBV genome sequences for the other tissues are not shown.

Information about the consensus sequences is summarized in Table 2. All consensus HBV genomes were 3,215 nt long. All tissues with the exception of tissues 25 and 26 were infected with HBV genotype C. Sequences of genotype B were found in tissues 25 and 26. The closest sequence matches among the reported HBV sequences were found for all tissues, but there was not a single complete match detected. For tissue 25, two different HBV sequences that were equally similar to the consensus sequence were found in the NCBI database. The degree of HBV genome identity between liver tissues

FIG 1 PCR amplification of three overlapping DNA fragments covering the entire HBV genome. (A) Positions of the three PCR amplicons. The positions of the three overlapping PCR fragments (excluding the primers) are indicated. The rcDNA genome of HBV is shown as a reference. The HBV genome shown reflects the numbering of HBV genotypes B and C and has a length of 3,215 nt (79). The strands of the rcDNA are shown in blue. The minus strand is complete, and its 5' end is covalently attached to the viral polymerase (shown as a green circle). The plus strand is incomplete, which is indicated with a dashed line. The positions of direct repeat 1 (DR1), DR2, as well as the RNA primer at the 5' end of the plus strand are indicated. The PCR fragments are shown in the following colors: (i) gray for fragment F1, (ii) yellow for fragment F2, and (iii) red for fragment B. The positions of the primers are given in parentheses. For all tissues, fragment B (excluding the primers) covers the region between positions 267 and 1768 (inclusive). For all tissues with the exception of tissue 39, fragment F1 (excluding the primers) spans the region between positions 1673 and 2296. In the case of tissue 39, a version of fragment F1 covers the region between positions 1673 and 2377 (shown as F1' in Table 1). Fragment F2 (for all tissues with the exception of tissues 39 and 40) covers the region between positions 2287 and 375 (through positions 3215/1). For tissues 39 and 40, F2 spans the region from position 2318 through 3215/1 to position 375 (shown as F2' in Table 1). (B) Results of PCR assays amplifying fragments B, F1, and F2. Each panel represents the results for a single PCR assay. The amplified fragment is indicated at the right-hand side of each gel image. The tissues analyzed are indicated at the top. "no T" stands for the negative control that did not have the template for amplification. M indicates double-stranded DNA size markers (a 1-kb plus DNA ladder [Invitrogen] was used). The sizes of certain DNA markers are indicated. The main PCR products are marked with arrows. The expected size of fragment B is 1,545 nt. The expected size of fragment F1 for all tissues with the exception of tissue 39 is 665 nt. For tissue 39, the size of fragment F1 was expected to be 743 nt. The size of fragment F2 (for all tissues with the exception of tissues 39 and 40) is 1,344 nt. For tissues 39 and 40, the size of fragment F2 was 1,314 nt. All major PCR products that migrated in the areas of the predicted sizes were gel purified and cloned into the pCRII-TOPO vector for subsequent sequencing. In the middle gel image, the black asterisk indicates the PCR product of an unexpectedly larger size that was observed for tissue 13. It was also gel purified and analyzed. The red asterisk points to fragment F1 of tissue 40, which had to be divided into two fragments, M and N, for further analysis, because it appeared to bear mutations in the overlapping B/F1 region as well as an internal deletion. The overlap between fragments B and F2 (for all tissues) spanned the region between

(Continued on next page)

TABLE 1 Primers and amplicons for nested PCR assays used to amplify HBV genomic sequences that were present in collected tissue samples

Fragment, positions ^a	Round of nested PCR, amplicon positions ^b	Forward primer sequence (genomic orientation)	Reverse primer sequence (antigenomic orientation)
F1, 1673–2296	1st, 1593–2395 2nd, 1653–2317	1593-ACCTCTGCACGTGTCATGG-1610 1653-CATAAGAGGACTCTTGACT-1672	2395-TCTGCGAGGCGAGGGAGT-2378 2317-GATAGGGGCATTGGTGGTCT-2297
F2, 2287–3215/1–375	1st, 1653–3215/1–449 2nd, 2267–3215/1–395	1653-CATAAGAGGACTCTTGACT-1672 2267-GGAGTGTGGATTCGACTCC-2286	449-CCAGAAGAACCAACAAGAAGATGA-426 395-TAAACGCCGAGACACATC-376
B, 267–1768	1st, 58–1820 2nd, 246–1790	58-TGCTGGTGGCTCCAGTT-74 246-GTCTAGACTCGTGGTGACTT-266	1820-GTTGCATGGTGTGGTGAACA-1800 1790-ATGCCTACAGCCTCTAGTACA-1769
F1', 1673–2377	1st, 1593–3215/1–449 2nd, 1653–2395	1593-ACCTCTGCACGTGTCATGG-1610 1653-CATAAGAGGACTCTTGACT-1672	449-CCAGAAGAACCAACAAGAAGATGA-426 2395-TCTGCGAGGCGAGGGAGT-2378
F2', 2318–3215/1–375	1st, 1593–3215/1–449 2nd, 2297–3215/1–395	1593-ACCTCTGCACGTGTCATGG-1610 2297-AGACCACCAATGCCCTATC-2317	449-CCAGAAGAACCAACAAGAAGATGA-426 395-TAAACGCCGAGACACATC-376
M, 1679–2114	1st, 1593–2168 2nd, 1653–2141	1593-ACCTCTGCACGTGTCATGG-1610 1653-CATAAGAGGACTCTTGACTCTCAGC-1678	2168-TGACATAGCTGACTACTAATTCCTGG-2142 2141-ATGCTGGATCTTCCAAATTACTTCCCA-2115
N, 2033–2377	1st, 1593–3215/1–449 2nd, 2012–2395	1593-ACCTCTGCACGTGTCATGG-1610 2012-TACGGGAGGCCTTAGAGTCT-2032	449-CCAGAAGAACCAACAAGAAGATGA-426 2395-TCTGCGAGGCGAGGGAGT-2378

^aThe boundaries of the fragment excluding the sequences of the primers are shown.

^bThe boundaries of the PCR amplicons, including the sequences of the primers, are indicated. Amplicon F1' was used in the case of tissue 39. Amplicon F2' was used for tissues 39 and 40. In the case of tissue 40, there were (i) mutations found in the overlapping fragment B/F1 region as well as (ii) an internal deletion. Therefore, amplicon F1 for tissue 40 had to be divided into two smaller fragments, M and N. Additional details are provided in Materials and Methods and in the legend to Fig. 1.

and matching HCC samples was 97.7 to 99.8% (Table 2). Interestingly, during the comparison of HBV sequences from tissues 26 and 25, we noted that mismatches heavily clustered within the region spanning positions 1726 to 2266. For tissue 26, this region of the consensus sequence had a 100% match to the corresponding fragment of the HBV genotype A sequence (GenBank accession number [KJ854710.1](#)) (34). Thus, the majority of HBV genomes from tissue 26 were recombinants between genotypes B and A. The region of the HBV genome that includes the area between positions 1726 and 2266 was previously reported to be a frequent intergenotype recombination site (35, 36).

The numbers of nucleotide mismatches between HCC and the corresponding liver tissue were 6 for HBV genomes recovered from tissues 3 and 4, 8 for tissues 7 and 8, 15 for tissues 13 and 14, 74 for tissues 25 and 26, and 44 for tissues 39 and 40. The nucleotide changes were found in ORFs for the polymerase, envelope proteins, pre-core/core, and X protein. The majority of these changes (92/147; 62.6%) did not result in amino acid substitutions. There were a number of changes that were previously reported to be associated with HCC risk (37–50). Here we mention several of them. Thus, for example, the C(2875)A substitution that causes the amino acid changes Thr(190)Lys in polymerase and Gln(10)Lys in the pre-S1 domain of the L envelope protein (37, 38) was found in tissue 4 (6/8 sequences; 75%) but not in tissue 3. The A(1846)T change, which does not cause an amino acid mutation (38, 40), was found in tissue 13 (12.5%), tissue 14 (100%), and tissue 26 (88.9%) but not in tissue 25. The G(1896)A change transforming the Trp-28 codon into a stop codon in precore (38–41) was observed in tissue 13 (12.5%), tissue 14 (100%), and tissue 39 (100%) but not in

FIG 1 Legend (Continued)

positions 267 and 375 (109 nt). The overlap between fragments B and F1 was 96 nt and spanned the region between positions 1673 and 1768. For the F1/F2 junction, the overlap was 10 nt (positions 2287 to 2296). For tissue 39, the F1/F2 junction was 60 nt (positions 2318 to 2377). For tissue 40, the junctions were 90 nt for the B/M junction (positions 1679 to 1768), 82 nt for the M/N junction (positions 2033 to 2114), and 60 nt for the N/F2 junction (positions 2318 to 2377). The amplification of fragments M and N is not shown.

Tissue 3 HBV genome consensus sequence

	CTCCACAACATTCCACCAAGCTCTGCTAGACCCAGAGTGAGGGCCTATACCTTCTGCTGGTGGCTCCAGTTC	75
4	.c.....	
6, 7	g.....	
	CGGAACAGTAAACCCGTGTTCCGACTACTGCCCTCACCATAATCTTCTCGAGGACTGGGGACCTGCACC	150
3	g.....	
5	c.....	
	GAACATGGAGAACACAACATCAGGATTCCTAGGACCCTGCTCTGTGTACAGCGGGGTTTTCTTGTGTGACAAG	225
1, 6, 7	c.....	
	AATCCTCACAATACCAAGAGTCTAGACTCGTGGTGGACTTCTCTCAATTTTCTAGGGGGAGCACCCACGTGTCC	300
5	Y.....	
8	R.....	
	TGGCCAAAATTGCGAGTCCCCAACCTCCAATCACTCACCAACTCTTGTCTCTCAATTGTCTGGCTATCGCTG	375
	GATGTGTCTGCGGCGTTTTATCATATTCCTCTTCATCTGCTGCTATGCCTCATCTTCTGTGTGTTCTTCGGA	450
4	c.....	
6	t.....	
8	g.....	
	CTACCAGGTATGTGCGCGTTTGCTCTACTTCCAGGAACATCAACTACCAGCACGGGACCTGCAAGACTTG	525
	CACGATTCTGCTCAAGGAACCTCTATGTTTCCCTCTGTTGCTGTACAAAACCTTCGGAAGGAACTGCACITG	600
	TATTCCTATCCCATCATCTCTGGGCTTTTCGAAGATTCTATGGGAGTGGGCTCAGTCCGTTTCTCTGGCTCAG	675
3	t.....	
	TTTACTAGTGCATTTGTTGAGTGTTCGTAGGGCTTTCCCCACIGTTTGGCTTTCAGTTATATGGATGATGTG	750
	GTATTGGGGGCCAAGTCGTACAAATCTTGAGTCCCTTTTACCTCTATTACCAATTTCTTTGTCTTTGGGT	825
	ATACATTTGAACCTTAACAAAACAAACGTTGGGGCTACTCCCTTAACCTCATGGGATATGTAAATGGATGTTGG	900
8	a.....a.....	
	GGTACTTTACCGCAAGAACATATTGTACTAAAAATCAAGCAATGTTTTCGAAACGCTGTAAATAGACCTATT	975
8	a.....	
	GATTGGAAAGTATGTCAGAGAAITGTGGGTCTTTTGGGCTTTGCTGCCCCCTTTACACAATGTGGCTATCTGCC	1050
7	c.....	
	TTAATGCTTTTATATGATGTATACAATCTAAGCAGGCTTTTCACTTTCTGCGCAACTTACAAGGCTTTTCTGTGT	1125
	AAACAATATCTGAACCTTTACCCCGTTGCCCGGCAACGGTCAGGTCTCTGCCAAGTGTTTGCTGACGCAACCCCC	1200
6, 8	g.....	
	ACTGGATGGGGCTTGGCTATTGGCCATCGCCGCATGCGTGGAACCTTTGTGGCTCCCTGCGGATCCATACTGCG	1275
6	ag.....	
	GAACTCCTGGCAGCTTGTTTTGCTCGCAGCCGGTCTGGAGCGAACTTATCGGAACGACAACTCTGTTGTCTC	1350
7	t.....	
8	c.....c.....	
	TCTCGGAAATACACCTCTTTCATGGCTGCTAGGGTGTGCTGCCAACTGGATCCGCGCGGACGTCCTTTGTCT	1425
1, 3, 5, 7	c.....	
2, 4, 6, 8	t.....	
	TACGTCCTGTCGGCGCTGAATCCGCGGACGACCCGCTCGGGGCGGTTTGGGACTCTACGTCCTCTTCTCAT	1500
5, 6	c.....g.....	
8	g.....	
	CTGCCGTTCCGGCCGACACGGGGCGCACCCTCTCTTACGCGTCTCCCGTCTGTGCTTCTCATCTGCCGGAC	1575
6	g.....	

FIG 2 Consensus sequence of the HBV DNA genome recovered from tissue 3 (nonmalignant liver). The consensus sequence was generated based on the eight complete HBV genomic sequences (eight sequences for each PCR fragment were used), which were assembled based on the degree of sequence identity within the overlapping areas of the PCR fragments. The consensus sequence is shown at the top. The positions in the HBV genomic sequence are shown on the right-hand side of the alignment. The identification number for a particular assembled genome sequence that has a nucleotide(s) that differs from the nucleotide(s) in the same position(s) in the consensus sequence is shown on the left side of the corresponding line of the alignment. The dots represent identical nucleotides, while nonidentical nucleotides are shown in small letters. "R" stands for purine (A or G), "Y" stands for pyrimidine (C or T), "K" stands for G or T, and "W" stands for A or T. Tissue 3 was infected with an HBV genome of genotype C, which was 3,215 nucleotides long.

	CGTGTGCACTTCGCTT CACCTCTG CACGTG CATGGAAACCAC CGTGAACGCCA CCAGGTCTTGCCCAAGGCTC	1650
4, 8	TACATAAGAGGACTCT TGGACTCT CAGCAATGTCAA CGACCGA CCTTGAGGCATACTTCAAAGACTGTTTGTITA	1725
6	..t..... ..t.....K.....	
1, 2, 7, 8	AGGACTGGGAGGAGTT GGGGAGGAGATTAGGTTAAAGATTATGTACTAGGAGGCTGTAGGCATAAATGGTCT	1800
3	W.....	
5	W.....a.....K.....W.....	
6	GTTCACCAGCACCATGCAACTTTTTCACCTCTGCTTAATCATCTCTTGTTTCATGTCCTACTGTTCAAGCCTCCAA	1875
	a.....	
8	GCTGTGCTTGGGTGGCITTGGGGCATGGACATTGACCCGTATAAGAATTGGAGCTTCTGTGGAGTTACTCTC	1950
	C.....	
	TTTTTGCCTTCGACTTCCTTCTCTATTCGAGATCTCCTCGACACCGCTCAGCTCTGTATCGGGAGGCTT	2025
	AGAGTCTCCGGAACATGTTTCACTCACCATACGCACTCAGGCAAGCTATTCTGTGTTGGGGTGAGTTGATGAA	2100
7	TCTGGCCACCTGGGTGGGAAGTAATTTGGAAGACCCAGCATCCAGGGAATTAGTAGTCGGCTATGTCAATGTTAA	2175
	c.....t.....	
1	TATGGGCTAAAACCTAGACAACCTACGTGGTTTCACATTCTCTGCTTACTTTTGAAGAGAAACTGTTCTTGA	2250
	t.....	
1-5, 7, 8	GTATTTGGTGTCTTTTGGAGTGTGATTGCACTCTCTCTGCTTACAGACCACCAATGCCCTATCTTATCAAC	2325
	Y...Y.....	
	ACTTCCGAAACTACTGTTGTTAGACGACGAGGCAGGTCCCTAGAAGAAGAACTCCCTCGCCTCGCAGACGAAG	2400
	GTCTCAATCGCCGCGTGCAGAGAATCTCAATCTCGGGAATCTCAATGTTAGTATCCCTTGGACTCATAAGGTGG	2475
2	GAAACTTTACTGGGCTTATTCTTCTACTGTACCTGCTTTAATCTGAGTGGCAAACCTCCTCTTCTCTCACA	2550
6, 7	g.....c.....	
2	TTCATTTACAGGAGGCATTATTAATAGATGTCAACAATATGTGGGCCCTCTTACAGTTAATGAAAAAGGAGAT	2625
3	a.....a.....a..a.....	
2	TAAATTAATTATGCTAGGTCTATCTTAACCTTACCAAATATTTGCCCTTGGACAAAGGCATTAAACCAT	2700
	c.....	
1	ATTATCCIGAACATGCAGTTAATCACTTCAAACTAGGCATTATTTACATACTCTGTGGAGGCGGGCATTCT	2775
4	a.....c.....	
2	TATATAAGAGAGAACTACACGCAAGCCTCATTTTGTGGGTACCATATTCTTGGGAACAAGGCTACAGCATG	2850
4	g.....a.....	
	GGAGGTGGTCTTCCAACCTCGACAAGGCATGGGACGAATCTTCTGTTCCCAATCCTCTGGGATTCCTTCCC	2925
	GATCACCAGTTGGACCTGCGTTCGGAGCCAATCAACAATCAGATTGGGACTTCAACCCCAACAAGGATCAT	3000
	TGGCCAGAGGCAATCAGGTAGGAGCGGGAGCAATCGGGCCAGGGTTCACCCACACACGGCGGTCTTTGGGG	3075
8	TGGAGCCTCAGGCTCAGGGCGTATTGACAACAGTGCAGTAGCACCTCCTCTGCTCCACCAATCGGCAGTCA	3150
	c.....	
	GGAAGACAGCCTACTCCATCTCTCACCTCTAAGAGACAGTCATCTCAGGCCATGCAAGTGAG	3215

FIG 2 (Continued)

tissue 40. One previous study, however, did not find that this change was significantly associated with an increased HCC risk (42). The paired A(1762)T and G(1764)A mutations, which were considered highly predictive of HCC (38–43), were found in tissue 25 (42.9%), tissue 26 (16.7%), tissue 39 (100%), and tissue 40 (37.5%). The G(1613)A change (39, 42, 44) was detected in tissue 39 (100%) and tissue 40 (12.5%). Another change, C(1653)T (38, 39, 41–44), was also found in tissue 39 (87.5%) and tissue 40 (12.5%). Another HCC-linked Met5 mutant in the X protein (47) was seen in tissue 13 (50%), tissue 14 (62.5%), and tissue 39 (50%) but not in tissue 40. The T(2170)C substitution (48) was observed in tissue 25 (100%), tissue 26 (22.2%), and tissue 40 (12.5%) but not

FIG 3 Consensus sequence of the HBV DNA genome recovered from tissue 4 (HCC). The consensus HBV genome sequence was generated as described in the legend to Fig. 2. The consensus is based on the eight assembled HBV sequences (eight sequences for each PCR fragment were used). The consensus sequence is shown at the top. The positions in the HBV genome are shown on the right side of the alignment. The identification number for a particular assembled genome sequence that has a nucleotide(s) that differs from the nucleotide(s) in the same position(s) in the consensus sequence is shown on the left side of the corresponding line of the alignment. The dots represent identical nucleotides, while nonidentical nucleotides are shown in small letters. "R" stands for purine (A or G), "Y" stands for pyrimidine (C or T), and "W" stands for A or T. Sequences 7 and 8 have the same 33-nt deletion spanning positions 24 to 56 (inclusive). The deletion is shown as a dashed line. Tissue 4 was infected with an HBV genome of genotype C (the total length is 3,215 nucleotides).

4, 6	GAACCTCTGGCAGCTTGTGTTTGTCTCGCAGCCGGTCTGGAGCGAACTTATCGGAACCGACAACCTCTGTTGTCCTC	1350
	C.....a.....	
4, 6	TCTCGGAATAACCTTCCTTCCATGGCTGCTAGGGTGTGCTGC CAACTGGATCCTGCGCGGGACGTCTTTGTC	1425
7	..C.....C.....	
3	TACGTCCCGTCGGCGCTGAATCCCGCGGACGACCCGTCTCGGGCCGTTTGGGACTCTACCGTCCCTTCTTCAT	1500
4, 6	C.....g.....	
5	C.....	
2	CTGCCGTTCGGCCGACACGGGGCGCACCTCTCTTTACGCGGTCTCCCCGTCTGTGCCTTCTCATCTGCCGGAC	1575
7	C.....g.....	
1	CGTGTGCATTTCGTTTCACTCTGCACTGCGCATGGAACCACTGTAACGCCACAGGTCTTGCCCAAGGTCT	1650
4	C.....g.....g.....	
7	g.....	
1	TACATAAGAGGACTCTTGGACTCTCAGCAATGTCAACGACCGACCTTGAGGCATACTTCAAGACTGTTTGTTTA	1725
3	R.....	
6	..t.....Y.....	
1, 3	AGGACTGGGAGGAGTTGGGGGAGGAGATTAGGTTAAAGATTATGTACTAGGAGGCTGTAGGCATAAATTGGTCT	1800
2	W.....W.....	
5	W.....W.....a.....	
7	.R.....W.....	
8	a.....	
5, 8	GTTCAACGACCATGCAACTTTTTCACTCTGCCTAATCATCTCTTGTTCAATGCTCTACTGTTCAGCCCTCCAA	1875
7	C.....a.....a.....	
1	GCTGTGCCCTGGGTGGCTTTGGGGCATGGACATTGACCCGTATAAAGAATTGGAGCTTCTGTGGAGTTACTCTC	1950
3	...g.....g.....	
5, 8	a.c.....	
2	TTTTTGTCTTCTGACTCTTTCTTCTATTCTGAGATCTCTCGACACCGCCTCAGCTCTGTATCGGGAGGCCTT	2025
5, 8	C.....g.....	
6	C.....	
5, 8	AGAGTCTCCGGAACATTGTTCACTCACCATACAGCACTCAGGCAGCTATTCTGTGTTGGGTGAGTTGATGAA	2100
	a.....	
	TCTGGCCACCTGGGTGGGAAGTAATTGGAGACCCAGCATCCAGGAATTAGTAGTCGGCTATGTCAATGTAA	2175
	TATGGGCCTAAAACTCAGCACTACTGTGTTTCACTTTCTGTCTTACTTTTGAAGAGAACTGTTCTTGA	2250
1, 2, 3, 4	GTATTGTGCTCTTTGGAGTGTGGATTGCACTCTCTCTGCTTACAGACCACCAATGCCCTATCTTATCAAC	2325
6	Y...Y.....	
7	t...c.....c.....	
4	ACTTCCGGAACTACTGTTGTTAGACGACGAGGCAGGTCCTTGAAGAAGAACTCCTCGCTCGCAGACGAAG	2400
8	t.....c.....	
2	GTCTCAATCGCCGCTCGCAGAAGATCTCAATCTCGGAATCTCAATGTTAGTATCCTTGGACTCATAAGGTGG	2475
4	...g.....g.....a.....	

FIG 3 (Continued)

in tissue 39. Furthermore, a deletion of nt 24 to 56 (pre-S2 domain) was found in 25% of sequences from tissue 4, and a deletion of nucleotides 3053 to 3115 (pre-S1 domain) was found in tissue 39 (50%). Previously, deletions in pre-S1/pre-S2 domains were reported in connection with the risk of HCC (50). In addition, there were also changes

	GAAACTTTACTGGGCTTATTCTTCTACTGTACCTGTCCTTAATCCTGAGTGGCAACTCCCTCCTTTCCTCACA	2550
1, 3, 8	g.....	
7	t.....a.....	
	TTCATTACAGGAGGACATTATTAATAGATGTCAACAATATGTGGCCCTCTTACAGTAATGAAAAAGGAGAT	2625
5	t.....	
6	a.....	
8	t.....	
	TAAATTAATTATGCCTGCTAGGTCTATCCTAAOCTTACCAATATTGCCCCTTGACAAAGGCATTAAACCAT	2700
5	c.....	
	ATTATCCTGAACATGCAGTAATCATTACTTCAAAA CYAGGCAT TATTACATACTCTGTGGAAGGCGGGCATT	2775
1	g.....c.....c.....	
2, 3, 8	c.....	
4, 6, 7	t.....	
5	t.....t.....	
	TATATAAGAGAGAACTACACGCAGCGCCTCATTITGTGGGTACCATATTCTTGGGAACAAGAGCTACAGCATG	2850
4	g.....	
	GGAGGTGGTCTTCCAACTCGAAAGGCATGGGACGAATCTTCTGTTCCTCAATCCTCTGGGATTCTTTCCC	2925
5	cc.g.....cg.....	
6	c.....	
	GATCACCAGTTGGACCCGCGTTCCGGAGCCAATCAACAATCCAGATTGGGACITCAACCCCAACAAGGATCAT	3000
	TGGCCAGAGGCAATCAGTAGGAGCGGAGCATTCTGGCCAGGTTTACCCCAACACGGCGGTCTTTTGGGG	3075
8	t.....	
	TGGAGCCCTCAGGCTCAGGGCGTATTGACAACAGTCCAGTAGCACTCCTCCTGCCTCCACCANTCGGCAGTCA	3150
1, 8	a.....a.....	
5	a.....g.....	
7	g.....	
	GGAAGACAGCCTACTCCATCTCTCCACCCTAAGAGACAGTCACTCTCAGGCCATGCAGTGGAA	3215

FIG 3 (Continued)

associated with a reduced risk of HCC. For example, the C(1799)G change (38) was detected in tissue 25 (100%) and tissue 26 (11.1%). It is worth noting, however, that a recent study that analyzed precore/core sequences found no significant difference in the total mutation ratios for either nucleotide mutations or amino acid changes between HCCs and adjacent nonmalignant liver tissues (46). Collectively, these data suggest that a number of changes, which were previously linked to HCC risk (37–50), could be valuable in terms of their appearance in serum HBV genomes, but they are not exclusively restricted to malignant or nonmalignant liver tissues.

In addition, there were other mismatches found in more than one liver/HCC pair, which were not previously linked to HCC risk. Among them were mutations of amino acid 88 in X (pairs 7/8 and 39/40); mutations of amino acid 60 in core (pairs 13/14 and 39/40); mutations of amino acids 31, 59, and 91 in core (pairs 25/26 and 39/40); nucleotide changes in the Leu-95 codon of core that did not change the amino acid (pairs 13/14 and 25/26); silent mutations in codons 45 and 111 of core and in codon 620 of polymerase (i.e., codon 284 in the reverse transcriptase domain) (pairs 25/26 and 39/40); and nucleotide substitutions that change or do not change the amino acid at position 27 in core (pairs 39/40 and 25/26, respectively). The functional significance of these mutations could be a subject for follow-up studies and is outside the scope of this study.

Markers of intrahepatic HBV replication. Quantification of HBV replication markers in liver/HCC tissues is summarized in Table 3. Levels of cccDNA and pregenomic/precore RNA (pgRNA) were measurable in all tissues with the exception of tissue 26, in which pgRNA levels were below the cutoff value for quantitative real-time PCR (qPCR). In all paired tissues, with the exception of pair 13/14, the levels of cccDNA and pgRNA were considerably higher in liver tissues than in matching HCC tissues. The difference

TABLE 2 Characteristics of the HBV genome consensus sequences identified in HBV-infected tissues

Tissue (liver or HCC)	HBV genotype ^a	GenBank accession no. of closest match, % identity ^b	Degree of HBV sequence identity in liver/HCC (%) ^c
3 (liver)	C	GQ205441.1, 99.4	99.8
4 (HCC)	C	GQ205441.1, 99.4	
7 (liver)	C	JQ688404.1, 99.2	99.8
8 (HCC)	C	JQ688404.1, 98.9	
13 (liver)	C	KJ173439.1, 98.7	99.5
14 (HCC)	C	KJ173439.1, 98.6	
25 (liver)	B	JQ688405.1 (GQ205440.1), 99.4 ^d	97.7
26 (HCC)	B	JQ688405.1, 97.4 ^e	
39 (liver)	C	GQ205441.1, 98.7	98.5
40 (HCC)	C	GQ205441.1, 98.7	

^aThe genotype of HBV was determined based on the closest sequence match of the consensus sequence for each infected tissue to the reported HBV genome sequence of a known genotype.

^bThe closest match to a particular HBV genome consensus sequence was determined by comparison between the consensus HBV sequence recovered from the HBV-infected tissue and the HBV genome sequence from the NCBI database.

^cThe degree of sequence identity was determined by comparison of the consensus HBV genome sequences from nonmalignant liver tissue and matching HCC tissue.

^dFor the HBV genome consensus sequence found in tissue 25, two matching previously reported HBV sequences were found. The sequence identity in each case was 99.4%.

^eThe HBV genome consensus sequence for tissue 26 represents a recombinant between genotypes B and A (see the text), and in the majority of recovered HBV genomes, the area between nucleotides 1726 and 2266 belongs to genotype A. However, the complete genome consensus sequence still had the closest match to HBV genotype B.

in cccDNA levels was within the range of 22.7- to 35.8-fold, while the difference in pgRNA levels was between 9.9-fold and 379.4-fold. It needs to be considered that since the level of pgRNA in tissue 26 could not be measured, the difference in pgRNA levels between tissues 25 and 26 could be higher than the above-mentioned value of 379.4-fold (which was found for pair 7/8) (Table 3). These differences were consistent with the previously described hypothesis that immune-mediated selection (via killing hepatocytes efficiently replicating HBV, which is mediated by cytotoxic T lymphocytes [CTLs]) drives the clonal outgrowth of cells that poorly replicate HBV and that, even-

TABLE 3 Quantities of cccDNA and pgRNA in tissues harvested from untreated individuals chronically infected with HBV^a

Tissue (liver or HCC)	cccDNA level (mean HBV GE/ μ g of total DNA $\pm$ SD)	pgRNA level (mean GE/ μ g of total RNA $\pm$ SD)
3 (liver)	$(2.34 \pm 1.46) \times 10^3$	$(8.22 \pm 6.59) \times 10^3$
4 (HCC)	$(1.03 \pm 0.46) \times 10^2$	$(8.27 \pm 4.21) \times 10^2$
7 (liver)	$(3.97 \pm 3.43) \times 10^2$	$(1.51 \pm 0.74) \times 10^4$
8 (HCC)	$(1.11 \pm 0.49) \times 10^1$	$(3.98 \pm 1.93) \times 10^1$
13 (liver)	$(1.69 \pm 0.48) \times 10^2$	$(1.77 \pm 2.02) \times 10^4$
14 (HCC)	$(7.20 \pm 4.81) \times 10^1$	$(8.78 \pm 2.11) \times 10^3$
25 (liver)	$(1.46 \pm 1.19) \times 10^3$	$(1.88 \pm 1.51) \times 10^3$
26 (HCC)	$(6.39 \pm 3.62) \times 10^1$	ND
39 (liver)	$(1.34 \pm 0.86) \times 10^3$	$(1.45 \pm 0.13) \times 10^4$
40 (HCC)	$(5.43 \pm 0.25) \times 10^1$	$(1.27 \pm 0.72) \times 10^2$

^aThe results of measurements of DNA concentrations in total DNA preparations and of RNA concentrations in total RNA preparations were used for calculations. ND, not determined (pgRNA levels in this tissue were below the detection limits of qPCR).

tually, one such cell can acquire an additional mutation(s) and give rise to clonal HCC. As the result, nonmalignant liver tissues usually represent a diverse population of hepatocytes that replicate HBV to different levels, while HCC cells usually have considerably lower levels of HBV replication markers (13, 15, 51–53). For pair 13/14, the levels of both HBV replication markers were not significantly different between liver and HCC tissues (between 2.0- and 2.4-fold) (Table 3). One possible explanation is that HCC-14 originated from a hepatocyte that did not have greatly reduced HBV replication levels, and the mechanism of HCC induction in this case could have been different from the mechanism that mediated the onset of the other tested HCCs.

PCR amplification of the 5'-end and 3'-end sequences of HBV RNAs that accumulated in liver/HCC tissues. To facilitate the subsequent analysis of the origins of HBV RNAs that accumulated in the tissue samples, we performed PCR amplification of HBV RNA sequences. Our primary focus was HBV RNAs coding for the viral envelope proteins. We anticipated that these mRNAs could either (i) originate from HBV genome replication or (ii) be produced independently of viral replication from integrated HBV DNA sequences, which could contain unaltered viral envelope promoters and coding sequences for HBsAg (1, 9, 16, 20, 21, 24). Total RNA or mRNA fractions were extracted from liver/HCC tissues and used for PCR analysis. The PCR strategy was designed to amplify RNA sequences coding for the L envelope protein and included cDNA synthesis using a VNN-anchored primer(s) (see Materials and Methods) for reverse transcription (54) and subsequent amplification of HBV mRNA sequences using two overlapping PCR amplicons. This analysis was expected to (i) determine what kinds of HBV RNA accumulated in liver/HCC tissues during chronic HBV infection and (ii) help us to understand if id-RNAs can serve as the templates for the translation of the L, M, and S envelope proteins. The PCR strategy is summarized in Fig. 4. For the analysis of the 5'-end HBV sequences, a nested PCR was employed, while the 3'-end region sequences were amplified via seminested PCR. The two amplicons overlapped by 158, 338, or 609 nt depending on the forward primer used for the amplification of the 3'-end region sequences (see Materials and Methods) (Fig. 4).

The above-mentioned 5'-end region sequences could be present on the cccDNA-transcribed pregenomic RNA, precore RNA, and L mRNA and also on id-RNAs for L. As shown in Fig. 5, the main precursor for HBV integration is a double-stranded linear HBV genome (DSL). The DSL has the left-hand side at or around position 1820, while the right-hand side is located at or around position 1831. Some ORFs are disrupted in the DSL (4, 15). However, if the integrated DSL did not undergo alterations, the complete ORF for L (which includes the sequences of M and S as well) should be not disrupted, and the promoters for the transcription of the L mRNA and M/S mRNA should be functional. Therefore, the transcription of these two mRNAs can start within the integrant; copy the HBV sequences, including the very-right-hand-side HBV nucleotide of the integrant; and then pass into the host sequence. As shown in Fig. 5, the conventional HBV poly(A) signal/site is located in the "wrong" position upstream of the ORFs for HBsAg. Thus, id-RNAs, for which transcription began from the envelope promoter(s), were expected to (i) be hybrids bearing HBV and host sequences and (ii) be polyadenylated using the poly(A) signal/site of the host, located downstream of the HBV-host junction at the right-hand side of the DSL (Fig. 5) (4, 15, 20, 21, 24). This would be the case if the viral cryptic poly(A) signal (55, 56) is not used. If the cryptic poly(A) signal is used, then RNAs transcribed from HBV promoters and originating from either cccDNA or integrated HBV DNA would be indistinguishable (4, 15). Since HBV integrations are random, the location of the above-mentioned host poly(A) signal/site is unknown, and the length of the host sequence insert incorporated into id-RNA immediately prior to the poly(A) tail also remains unknown (Fig. 5) (4, 15, 20, 21). In addition, another kind of integrant-derived HBV RNA may be transcribed from the host promoter that could be located upstream of the left-hand side of the DSL (Fig. 5). In this case, the resulting RNA will bear a 5'-end UTR of an unknown length, because the location of the host promoter relative to the DSL is unknown. However, it remains to be determined which signal(s) these RNAs use for polyadenylation, which regions of HBV they can

FIG 4 PCR strategy to amplify the entire lengths of HBV mRNAs coding for the large envelope protein (L). HBV mRNAs coding for the L envelope protein can be transcribed from cccDNA (i.e., produced by HBV replication) or derived from integrated HBV DNA. The strategy is expected to amplify mRNAs coding for L regardless of their origin. The cccDNA-derived HBV RNAs, id-RNAs, and positions of the primers are shown above the dashed green line. Only one set of the forward primers used to amplify the 3'-end region RNA sequences is shown, while all three sets of the forward primers used are described in Materials and Methods. Also indicated are the location of the ORF for the L envelope protein and the boundaries of the pre-S1, pre-S2, and S domains. The pregenomic/precursor RNA (pgRNA) (~3.5 kb), the mRNA for the L envelope protein (~2.4 kb), mRNA for the M/S (middle/small) envelope proteins (~2.1 kb), and mRNA for the X protein (~0.8 kb) are shown as the major cccDNA-derived HBV RNAs. They all have the same 3' end and are expected to be polyadenylated near position ~1931 using the conventional viral poly(A) signal UAUAAA located at positions 1916 to 1921 (shown as a green rectangle on pgRNA). The id-RNAs coding for L are also shown. The main substrate for HBV integration is the double-stranded linear HBV genome (DSL). The DSL is presented in detail in Fig. 5. The L mRNA produced from such HBV integrant (and transcribed from the pre-S1 promoter) is expected to have 5'-end sequences identical to those of cccDNA-derived L mRNA. However, the 3'-end part is different. The HBV sequence is expected to end at or near position ~1831 (corresponding to the right end of the DSL) (Fig. 5) (4, 15), and it will be joined to a host sequence (that is located next to the HBV-host junction at the integration site) of an unknown size (indicated in red). The id-RNAs are virus-host hybrids (20, 21, 24). They acquire their poly(A) tails using host poly(A) signals/sites that are located at an unknown distance away from the HBV-host junction (downstream of the right end of the integrant) (Fig. 5). cDNA synthesis and PCRs are described in detail in Materials and Methods. The entire L mRNA sequence was amplified by using two overlapping PCR amplicons (one for the 5'-end part and another for the 3'-end part). The 5'-end region was amplified by using nested PCR, while the 3'-end part was amplified by seminested PCR. For the 5'-end region, the outer amplicon employed primers spanning nt 2807 to 2836 and 1253 to 1274, while the inner amplicon used primers covering nt 2816 to 2843 and 1232 to 1251. The product of nested PCR spanning positions 2816 to 1251 (i.e., the 5'-end region product) is approximately 1,651 bp long (shown as a purple line below the dashed green line). This 5'-end region product can be produced only from replication-derived pgRNA, L mRNA, as well as integrant-derived L mRNA (that is transcribed from the pre-S1 promoter) (Fig. 5). As can be deduced from Fig. 5, there is also some possibility that HBV RNA sequences could be transcribed from the integrated DSL by using the host promoter that could be located upstream of the left-hand side of the integrated DSL. In this case, the resulting RNA will bear a 5' UTR of an unknown length, because the location of the upstream cellular promoter relative to the integrated DSL is unknown. In addition, the site(s) at which such RNAs could be polyadenylated, and whether these RNAs could facilitate the translation of L or other envelope proteins, remains to be determined. The 3'-end region amplicon utilized a VN LNA oligo(dT) primer (see Materials and Methods) and a primer that spans positions 754 to 774 in the S domain for the first round of PCR, while for the second round of PCR, the same VN LNA oligo(dT) primer and a primer covering nt 865 to 893 (in the 3' UTR of HBV) were used. The expected two types of final 3'-end region PCR products are shown below the dashed green line. The final 3'-end region PCR product related to HBV replication can be produced from pgRNA, L mRNA, and M/S mRNA. It will span positions ~865 to 1931 and is expected to be about 1,102 nt long. The size of the final PCR product generated from id-RNAs is unknown, because the size of the host sequence at the 3' end of id-RNAs remains unknown. It is expected to bear HBV sequences covering positions ~865 to 1831 and, unlike replication-related PCR products, is not expected to bear the HBV region between nt ~1831 and 1931 (as reported previously, the exact positions of the left end and right end of the integrated HBV DNA can vary [4, 15]). The cryptic HBV poly(A) signal CAUAAA (positions 1788 to 1793) (55, 56), which is located upstream of the above-mentioned conventional poly(A) signal, is also shown as a white rectangle on pgRNA.

FIG 5 The double-stranded linear HBV DNA genome is the main substrate for HBV DNA integration into host chromosomal DNA. (A) Features of the DSL. The schematic representation of the DSL is based on diagrams from two previously published studies that analyzed HBV-host junctions at the left- and right-hand sides of HBV DNA integrants (4, 15). The DSL is shown as a solid black line. Numbering is displayed for HBV serotype AYW. The DSL left-hand side is expected to be located at or near position ~1820, while the right-hand side is expected to be located at or near position ~1831 (4, 15). The exact positions of the left-hand side and the right-hand side of the integrated HBV DNA can vary (4, 15). At the left and right ends, the integrated DSL is joined to the host DNA sequences. Some ORFs are disrupted in the DSL. The disrupted ORFs for the precore/core and X proteins are indicated. The positions of direct repeat 1 (DR1) and the posttranscriptional regulatory element (PRE) that is needed for mRNA export to the cytoplasm for subsequent translation (1, 9) are indicated. In the unaltered DSL, the L ORF is not disrupted, and promoters for both the L mRNA and the M/S mRNA are in the correct positions and are expected to be functional. Thus, transcription of the L mRNA or M/S mRNA can be initiated from the HBV promoters, and the synthesis of these mRNAs can be completed. However, the conventional viral poly(A) signal/site in the DSL is located upstream of the L ORF. Therefore, transcription (initiated from the L or M/S promoter) is expected to pass the virus-host junction (at the right end of the DSL) and continue into the host sequences until the RNA acquires the poly(A) tail using the host poly(A) addition signal/site [if the viral cryptic poly(A) signal (Fig. 4) is not used]. (B) mRNA for the L envelope protein that can be transcribed from the integrated DSL independently of HBV replication. This RNA (transcribed from the viral pre-S1 promoter) is a hybrid RNA that bears 5'-CAP, the 5' UTR of HBV, the ORF for L, and the 3' UTR of HBV, which is expected to end at or around position ~1831. Unlike replication-derived RNA, this id-RNA also contains a host insert of an unknown length [immediately next to the poly(A) tail] that is expected to bear the host poly(A) signal and part of the poly(A) addition site that remains after cleavage during the processing of the mRNA [if the conventional host poly(A) signal/site is used] and the poly(A) tail. Pol, polymerase.

bear, and whether HBsAg could be translated from such RNAs. Therefore, these RNAs were not considered during the analysis conducted by using conventional RT-PCR, but RNAs bearing 5'-host-HBV-3' junctions and representing this category of transcripts were analyzed later by using RNA sequencing (see "Analysis of accumulated HBV RNAs in selected tissues using RNA sequencing," below). Regardless of the origin of HBV RNA, it was anticipated that the final PCR product for the 5'-end region would be about 1,651 nt long (Fig. 4).

The final PCR product for the 3'-end region can be derived (i) from pregenomic RNA, precore RNA, L mRNA, and M/S mRNA produced by HBV replication or (ii) from id-RNAs (Fig. 4 and 5). There was expected to be a difference between 3'-end region PCR products depending on the origin of HBV RNA. The products generated from the cccDNA-transcribed RNAs were expected (i) to be about 925, 1,102, and 1,368 nt long (because three different sets of forward primers were used, as described in Materials and Methods) and (ii) to contain only HBV sequences preceding the poly(A) tail. The PCR products generated from id-RNAs transcribed from one of the HBV envelope promoters were expected to bear HBV and host sequences and to be of various lengths for the reasons explained above (Fig. 4). Considering the boundaries of the DSL (4, 15), id-RNA-derived PCR products were expected to have variable 3' ends of HBV sequences at or around position 1831. Thus, the region between approximately nt 1831 and 1931 was expected to be absent on these products, but it was expected to be present on the PCR products generated from cccDNA-transcribed RNAs (Fig. 4 and 5).

The outcomes of PCR are shown in Fig. 6 and 7. The products of amplification of the 5'-end regions of HBV RNAs are shown in Fig. 6. We observed PCR products of the expected size of ~1,651 nt (Fig. 4) for all tissues, and we gel purified and analyzed

FIG 6 PCR amplification of the 5'-end region sequences of HBV RNAs that accumulated in matching liver/HCC tissues. (Top) Results of the RT-PCR procedure conducted in the presence of reverse transcriptase (RT+). (Bottom) Results obtained in control experiments in the absence of reverse transcriptase (RT-). "no T" stands for the negative control that did not have the template for amplification. The lanes marked M contained the size markers of double-stranded DNA (a 1-kb plus DNA ladder [Invitrogen] was used). The tissue numbers are shown at the top. The products marked with asterisks were gel purified, cloned, and sequenced.

them. However, for tissues 25 and 26, these products were not the major ones. The major product was of a lower molecular weight (about ~1,200 bp). Therefore, for tissues 25 and 26, we also analyzed this lower-molecular-weight product. Products of roughly similar sizes (they appeared to be slightly smaller than those of tissues 25 and 26) were also observed for tissues 3, 7, 13, 39, and 40. All the products marked with asterisks in Fig. 6 were gel purified and cloned, and the resulting constructs were sequenced.

During the analysis of 3'-end region sequences, we focused on the major and distinct PCR products and their sizes relative to the expected sizes of the PCR products produced from the cccDNA-transcribed RNAs (i.e., either 925, 1,102, or 1,368 nt, depending on the forward primers used [see Materials and Methods] [Fig. 4]). In the case of the PCR analysis presented in Fig. 7, the expected size of the PCR product generated from cccDNA-transcribed RNAs was 1,102 nt. For all tissues, no highly intense PCR products were observed above the DNA marker of 1,650 bp. All gel-purified, cloned, and sequenced PCR products are marked by asterisks in Fig. 7. For tissues 3, 7, 8, 13, 14, 39, and 40, we gel purified and analyzed the PCR products of the highest intensity, which migrated between the markers of 850 bp and 1,650 bp. In the case of tissue 4, we analyzed the only major PCR product that migrated between 500 and 650 bp. For this tissue, no major distinct PCR product of a higher molecular weight was observed. For tissues 25 and 26, we analyzed relatively weak bands that migrated in the area between approximately 1,000 and 1,100 bp. We have not analyzed the other relatively weak PCR products. The PCR products with sizes above 2,000 bp or below ~500 bp were not considered for further analysis as well. High-molecular-weight PCR bands (>2,000 bp) were observed for tissues 3, 7, 13, and 39 and appeared rather weak. During the analysis of RNA isolated from tissue 39, when the reaction mixture after the completion of the PCR process was boiled prior to loading onto an agarose gel, the above-mentioned high-molecular-weight band was no longer observed (not shown),

FIG 7 PCR amplification of the 3'-end region sequences of HBV RNAs that accumulated in matching liver/HCC tissues. In the case of the PCR analysis presented here, the expected size of PCR products amplified from cccDNA-derived RNAs was 1,102 nt (Fig. 4). (Top) Results of RT-PCR conducted in the presence of reverse transcriptase (RT+). (Bottom) Results obtained in the absence of reverse transcriptase (controls) (RT-). "no T" stands for the negative control that did not have the template for amplification. The lanes marked M contain the size markers of double-stranded DNA (a 1-kb plus DNA ladder [Invitrogen] was used). The numbers of the tissues analyzed are shown at the top. The products marked with asterisks were gel purified, cloned, and sequenced.

which suggested that the bands of >2,000 bp initially observed for tissues 3, 7, 13, and 39 were aggregates of DNA species produced by PCR. The other lower-molecular-weight bands (~500 bp or lower) could represent RNA species that lost a significant part of the 3' UTR of HBV (Fig. 5) and may not serve as functional templates for HBsAg translation.

Analysis of 5'-end region HBV RNA sequences amplified from liver/HCC tissue samples. HBV-specific sequences of 5'-end region PCR products (described above) were obtained from all tissues. The complete sequences of this PCR amplicon (excluding the primers) were expected to include 4 nt of the 5' UTR; sequences coding for the pre-S1, pre-S2, and S domains (i.e., the entire ORF for the L protein, which incorporates ORFs for M and S); and a part of the 3' UTR (Fig. 4). As mentioned above, these 5'-end sequences do not reflect the origins of the HBV RNAs (i.e., cccDNA transcribed or integrant derived). At least nine sequences were obtained per tissue, and the corresponding alignments were generated (not shown). The results of the sequence analysis are summarized in Table 4, which shows the numbers of sequences with deletions/insertions and a number of mRNAs that potentially could serve as the templates for the translation of functional L, M, or S envelope proteins. Insertions were found only in the sequences from tissues 7 and 25 (Table 4). Only sequences from tissue 40 did not contain any deletions/insertions, which suggested that each sequence in this case can represent an mRNA that could be used for the translation of functional L, M, or S. Our study did not analyze functional properties of accumulated mRNA sequences, and therefore, every observed deletion/insertion represented a potentially nonfunctional mRNA sequence. Thus, for example, for tissue 3, 6/10 sequences had the same 18-nt deletion in the 5'-UTR/pre-S1 domain, and another sequence had a 1-nt deletion in the pre-S1-coding area. However, in other regions (coding for pre-S2, S, and the 3' UTR), no deletions/insertions were found (Table 4). Based on the locations of the deletions and

TABLE 4 Summary of data from analyses of 5'-end sequences of HBV RNAs that accumulated in liver/HCC tissues

Tissue (liver or HCC)	No. sequences with deletions or insertions ^a in different domains of HBV RNAs/ total no. of recovered RNA sequences					No. of potentially functional mRNAs for L, M, and S proteins/total no. of RNAs ^b
	5' UTR	Pre-S1	Pre-S2	S	3' UTR	
3 (liver)	6/10	7/10	0/10	0/10	0/10	L, 3/10; M, 10/10; S, 10/10
4 (HCC)	8/12	8/12	2/12	0/12	1/12	L, 2/12; M, 9/12; S, 11/12
7 (liver)	0/22	13/22	12/22	12/22 ^d	1/22	L, 7/22; M, 7/22; S, 7/22
8 (HCC)	0/20	9/20	9/20	9/20	1/20	L, 10/20; M, 10/20; S, 10/20
13 (liver)	0/10	0/10	0/10	2/10	1/10	L, 7/10; M, 7/10; S, 7/10
14 (HCC)	0/9	0/9	1/9	1/9	0/9	L, 7/9; M, 7/9; S, 8/9
25 (liver)	0/18	13/18 ^c	12/18	12/18 ^e	1/18 ^f	L, 4/18; M, 4/18; S, 4/18
26 (HCC)	0/15	10/15 ^c	9/15	9/15	1/15	L, 5/15; M, 5/15; S, 5/15
39 (liver)	0/11	1/11	0/11	2/11	0/11	L, 9/11; M, 9/11; S, 9/11
40 (HCC)	0/12	0/12	0/12	0/12	0/12	L, 12/12; M, 12/12; S, 12/12

^aThe numbers of sequences with deletions per total number of recovered RNA sequences are shown. The insertions were found only for the material recovered from tissues 7 and 25, and they are described in footnotes below. Not all deletions/insertions in the same area were identical.

^bAs explained in the text, any deletion/insertion was considered an alteration that potentially could render the mRNA defective in terms of the production of functional HBsAg. Therefore, the potentially functional mRNAs for the envelope proteins bear no deletions and/or insertions.

^cFor tissue 25, 12/18 sequences contained a deletion of 478 nt spanning the parts coding for the pre-S1, pre-S2, and S domains. In the case of tissue 26, 9/15 sequences bear the same 478-nt-long deletion. The finding of a 478-nt-long deletion was consistent with the lower-molecular-weight major PCR products observed for tissues 25 and 26 (Fig. 6). We conducted an additional alternative PCR assay and confirmed the presence of the 478-nt-long deletion for HBV RNAs from tissues 25 and 26 (not shown).

^dIn this case, 12/22 sequences had the same deletion. In addition, 8 out of these 12 sequences had the same 1-nt insertion. There was also another sequence that had the same 1-nt insertion and no deletion. In addition, another sequence contained a different 1-nt insertion.

^eIn this case, 1/18 sequences (with no deletions in the S region) had a 1-nt insertion.

^fIn this case, 2/18 sequences (that contained the same deletion in the S region) had a 1-nt insertion.

a number of deletions per individual sequence, at least 3/10 sequences could represent potentially functional L mRNAs, while 10/10 sequences could reflect functional mRNAs for M or S. For tissue 4, 8/12 sequences had the same deletion in the 5'-UTR/pre-S1 domain (as for tissue 3), while 2 sequences had an identical deletion of 33 nt in the pre-S2 domain, and one of the sequences bearing an 18-nt deletion in the 5'-UTR/pre-S1 domain (as for tissue 3) also had an additional 2-nt deletion in the 3' UTR (Table 4). These deletions suggested that 2/12 mRNAs could produce functional L protein, while 9/12 mRNAs could produce functional M, and 11/12 could produce functional S. For tissue 7, 12/22 sequences had a 685-nt deletion spanning the areas coding for pre-S1, pre-S2, and S, and one of these 12 sequences had an additional 1-nt deletion in the pre-S1 area. Another sequence had a unique 3-nt deletion in the pre-S1 area. In addition, in the S-coding area, 8/22 sequences bore an identical 1-nt insertion along with the above-mentioned 685-nt deletion, while another sequence contained the same 1-nt insertion and no deletion. Furthermore, another sequence had a 1-nt deletion in the 3' UTR along with the above-mentioned 685-nt deletion and 1-nt insertion. In addition, another sequence contained a different 1-nt insertion in the S-coding area and no deletion. Overall, only 7/22 sequences did not have any deletions/insertions, and thus, 7/22 mRNAs from this tissue potentially could produce functional L, M, and S proteins (Table 4). The potential impacts of the observed deletions and insertions on the RNAs recovered from other liver/HCC tissues are also presented in Table 4. Overall, potentially functional mRNAs for all three envelope proteins (as judged by analysis of the 5'-end regions of HBV RNAs) could have been produced in all tissues but with different frequencies (Table 4).

Analysis of 3'-end region HBV RNA sequences amplified from liver/HCC tissue samples. The 3'-end region sequences were expected to indicate the origin of the HBV RNA species. Accordingly, several types of 3'-end sequences were found in liver/HCC tissues. The first type was the RNAs produced by HBV genome replication (i.e., transcribed from cccDNA). An example of the 3'-end region of the cccDNA-transcribed RNA is presented in Fig. 8, which shows the RNA sequence recovered from tissue 3. The 3'

TCCCTTAACCTTCATGGGATATGTAATTGG (forward primer, 865 to 893)–
 (894) ATGTTGGGGTACTTTACCGCAAGAACATATTGTACTAAAAATCAAGCAATGTTTTTCGAAACTGCCTG
 TAAATAGACCTATCGATTGGAAAGTATGTGAGAGAATTGTGGGTCTTTTGGGCTTTGCTGCCCCCTTTTACACA
 ATGTGGCTATCCTGCCTTAATGCCTTTATATGCATGTATACAATCTAAGCAGGCTTTTACTTTCTCGCCAACT
 TACAAGGCCTTTCTGTGTAACAATATCTGAACCTTTACCCCGTTGCCCGGCAACGGTCCGGTCTCTGCCAAG
 TGTCTGCTGACGCAACCCCACTGGATGGGGCTTGGCTATTGGCCATCAGCGCATGCGTGGAACCTTTGTGGC
 TCCTCTGCCGATCCATACTGCGGAACCTCCTGGCAGCTTGTCTGCTCGCAGCCGGTCTGGAGCGAACCTTATC
 GGAACCGACAACCTCTGTTGTCTCTCTCGGAAATACACCTCCTTTCCATGGCTGCTAGGGTGTGCTGCCAACT
 GGATCTGCGCGGACGTCCTTGTCTACGTCCCGTCGGCGCTGAATCCCGCGACGGCCCTCTCGGGGCCG
 TTTGGGACTCTACCGTCCCCCTTCTTCGTCTGCCGTTCCGGCCGACCACGGGGCGCACCTCTCTTACGCGGTC
 TCCCGTCTGTGCCCTTCTCATCTGCCGGCCGTGCGCACTTCGCTTCACCTCTGCACGTGCGATGGAACCCAC
 CGTGAACGCCCCACAGGTCTTGCCCAAGGTCTTATATAAGAGGACTCTTGGACTCTCAGCAATGTCAACGACC
 GACCTTGAGGCATACTTCAAGACTGTTTGTTTAAGGACTGGGAGGAGCTGGGGAGGAGATTAGGTTAAAGA
 TTTATGTACTAGGAGGCTGTAGG**CATAAA**TTGGTCTGTTCACCAGCACCATGCAACTTTTTTACCTCTGCCTA
 ATCATCTCATGTTTCATGTCTACTGTTCAAGCCTCCAAGCTGTGCCTTGGGTGGCTTTGGGCGATGGACATTG
 ACCCG**TATAAA**GAATTTGGAG (1931)
 – **AT**–AAAAAAAAAAAAAAAAACCTATAGTGAGTCG (reverse primer)

FIG 8 Example of a 3'-end region sequence of cccDNA-derived HBV RNA. The sequence (T3_1) recovered from tissue 3 is shown. The PCR product contains an HBV sequence spanning positions 894 to 1931. Polyadenylation occurred via the conventional HBV poly(A) signal TATAAA (shown in the DNA sequence in boldface type and with underlining). Also shown are the cryptic HBV poly(A) signal CATAAA (shown in the DNA sequence in boldface type and with underlining) (55, 56) and the sequences of the 3'-end and 5'-end boundaries of the obtained PCR product, which match the sequences of the forward and reverse primers. The AT dinucleotide immediately downstream of position 1931 likely reflects the use of a VN or VNN sequence that was placed at the 3' end of the reverse primer or the primer used for cDNA synthesis, respectively.

UTR of this RNA contains an HBV sequence that ends at position 1931. Polyadenylation occurred by using the conventional HBV poly(A) signal TATAAA (shown as in the DNA sequence). The second type was the RNAs polyadenylated by using the cryptic HBV polyadenylation signal CATAAA (Fig. 9), which is located upstream of the conventional HBV polyadenylation signal (Fig. 4) (55, 56). Unlike the conventional poly(A) signal that is located close to the left-hand side of the DSL in the “wrong” position (Fig. 5), the cryptic signal is present in the correct position in the integrants (downstream of the ORFs for HBsAg) and can facilitate the HBV-mediated polyadenylation of the integrant-derived RNAs. Therefore, RNAs that are polyadenylated via the cryptic poly(A) signal can be produced by HBV replication or be transcribed from integrated HBV DNA, and thus, their origin cannot be established by analysis of the 3'-end regions of HBV RNAs. The third type was the id-RNAs (Fig. 10). The 3'-end region of id-RNA bears an HBV sequence that usually ends at position ~1831 (15), which is joined to a human sequence of an unknown length (Fig. 5), followed by the poly(A) tail. These RNAs always bear the host sequence insert next to the poly(A) tail. For the id-RNAs from tissues 3, 7, 8, 13, 14, 25, 26, 39, and 40, most of the time, the 3' ends of the HBV sequence were

TCCCTTAACCTTCATGGGATATGTAATTGG (forward primer, 865 to 893)–
 (894) ATGTTGGGGTACTTTACCGCAAGAACATATTGTACTAAAAATCAAGCAATGTTTTTCGAAACTGCCTG
 TAAATAGACCTATGATTGGAAAGTATGTGAGAGAATTGTGGGTCTTTTGGGCTTTGCTGCCCCCTTTTACACA
 ATGTGGCTATCCTGCCTTAGTGCCCTTTATATGCATGTATACAATCTAAGCAGGCTTTTACTTTCTCGCCAACT
 TACAAGGCCTTTCTGTGTAACAATATCTGAACCTTTACCCCGTTGCCCGGCAACGGTCCGGTCTCTGCCAAG
 TGTCTGCTGACGCAACCCCACTGGATGGGGCTTGGCTATTGGCCATCGCCGATGCGTGGAACCTTTGTGGC
 TCCTCTACCGATCCATACTGCGGAACCTCCTGGCAGCTTGTCTGCTCGCAGCCGGTCTGGAGCGAACTTATC
 GGAACCGACAACCTCTGTTGTCTCTCTCGGAAATACACCTCCTTTCCATGGCTGCTAGGGTGTGCTGCCAACT
 GGATCTGCGCGGGACGTCCTTGTCTACGTCCCGTCGGCGCTGAATCCCGCGGACGACCCGCTCTCGGGGCCG
 TTTGGGACTCTACCGTCCCCCTTCTTCATCTGCCGTTCCGGCCGACCACGGGGCGCACCTCTCTTACGCGGTC
 TCCCGTCTGTGCCCTTCTCATCTGCCGACCGTGTGCACTTCGCTTCACCTCTGCACGTGCGATGGAACCCAC
 CGTGAACGCCCCACAGGTCTTGCCCAAGGTCTTACATAAGAGGACTCTTGGACTCTCAGCAATGTCAACGACC
 GACCTTGAGGCATACTTCAAGACTGTTTGTTTAAGGACTGGGAGGAGTTGGGGAGGAGATTAGGTTAATGA
 TTTTGTACTAGGAGGCTGTAGG**CATAAA**TTGGTCTGCTCACCAGCACCATCC (1817)
 – **AAAA**AAAAAAAAAAAAAAAAACCTATAGTGAGTCT (reverse primer)

FIG 9 Example of a 3'-end region sequence of HBV RNA that was polyadenylated using the cryptic HBV polyadenylation signal. The sequence (T3_7) recovered from tissue 3 is shown. The PCR product contains an HBV sequence spanning positions 894 to 1817. Polyadenylation occurred using the cryptic HBV poly(A) signal CATAAA (shown in the DNA sequence in boldface type and with underlining) (55, 56). Also shown are the sequences of the 3'-end and 5'-end boundaries of the PCR product that match the sequences of the forward and reverse primers.

TCCCTTAACCTTCATGGGATATGTAATTGG (forward primer, 865 to 893)–
 (894) AAGTTGGGGTACCTTACCGCAGGAACATATTGTATAAAAATCAAGCAATGTTTTCGAAAACCTGCCTG
 TAAATAGACCTATTGATTGGAAGTATGTCAAAGAATTGTGGGTCTTTTGGGCTTGCCGCCCTTTTACACA
 ATGTGGCTATCCTGCTTAAATGCTTATGTCATGTATACAATCTAAGCAGGCTTTCACCTTTCGCGCAACT
 TATAAGGCCTTTCTGTGTAACAATATCTGAACCTTTACCCCGTTGCCCGGCAACGGTCAGGTCTCTGCCAAG
 TGTGTGCTGACGCAACCCCACTGGATGGGGCTTGGCCATAGGCCATCGGCGCATGCGTGAACCTTTGTGGC
 TCCTCTGCCGATCCATACTGCGGAACCTCTAGCAGCTTGTGTTGCTCGCAGCAGGTCTGGAGCGAACTTATC
 GGAACCGACAACCTCTGTGCTCTCTCGGAAATACATCTCCTTTCCATGGCTGCTAGGGTGTGCTGCCAACT
 GGATCCTGCGCGGGACGTCCTTTGTCTACGTCCCGTCGGCGCTGAATCCCGCGGACGCCGCTCTCGGGGCCG
 TTTGGGACTCTACCGTCCCTTCTTCATCTGCCGTTCCGGCCAACCGGGCGCACCTCTCTTTACGCGGTC
 TCCCCGTCTGTGCTTCTCATCTGCCGACCGTGTGCACCTTCGCTTACCTCTGCACGTGCGATGGAACCCAC
 CGTGAACGCCCCACCGGTCTTGCCCAAGGTCTTACATAAGAGGACTCTTGACTCTCAGCAATGTCAACGACC
 GACCTTGAGGCATACTTCAAAGACTGTTTGTGTTAAGGACTGGGAGGAGTTGGGGAGGAGATTAGGTTAAAGG
 TCTTTGTACTAGGAGGCTGTAGG**CATAAA**ATTGGTCTGCGCACCAGCACCATGCAACT**TT** (1822)
 – GAAAAATGCTGAAT**TTTAA**AGTCCCTGTTTCATGGGGAATATGAACTTTTCATTCTAAAAAGACTGTACTGT
 ATGAAAG**CATAAA**TTTATGTTATATTGTGGTC (Homo sapiens chromosome 5 clone CTC–
 529J12, complete sequence, ID: gb|AC113167.2| Range 1: 100333/35 to
 100232)– **TAAAAA**AAAAAACCCTATAGTGAGTCG (reverse primer)

FIG 10 Example of a 3'-end region sequence of the HBV id-RNA. The sequence (T7_3) recovered from tissue 7 is shown. The PCR product contains an HBV sequence spanning positions 894 to 1822, which is joined to a fragment of a human sequence. The identity of the human sequence in this id-RNA is shown in parentheses. The TT dinucleotide (boldface type and underlined) near position 1822 could belong to either the HBV sequence or the human sequence. Polyadenylation occurred by using the host poly(A) signal. At least two potential poly(A) signals were identified in the human sequence insert (shown in the DNA sequence in boldface type and with underlining). The signal located closer to the 3' end of the id-RNA was likely used for polyadenylation. Also shown are the cryptic HBV poly(A) signal CATAAA (boldface type and underlined) (55, 56) and the sequences of the PCR product 3'-end and 5'-end boundaries matching the sequences of the forward and reverse primers. The T nucleotide (boldface type and underlined) immediately downstream of the host insert likely reflects the 3'-end version of the VN sequence that was placed at the 3' end of the reverse primer.

mapped within the region of positions 1733 to 1827, while a few times, the HBV nucleotide at the HBV-host junction was mapped to position 1899. The id-RNAs from tissue 4 all had the 3' end of the HBV sequence at position 1306 (thus reflecting a loss of a considerable part of the HBV 3' UTR [Fig. 4 and 5]). Our observations were consistent with data from a previous report on mapping the positions of the right-hand side of integrated DSL sequences in human livers chronically infected with HBV (15).

During our analysis, we also uncovered 3'-end region sequences that (i) contain only HBV sequences but lost a rather substantial portion of the 3' UTR of HBV, (ii) bear several different HBV fragments joined together (but do not bear host sequences at the very 3' ends) and thus likely represent the products of recombination, or (iii) bear HBV fragments recombined together and human sequences at the very 3' ends. These RNAs likely cannot produce functional HBsAg, and thus, they were not considered further in our study. Therefore, only the above-described three types of HBV RNAs that potentially could produce HBsAg were considered for further analysis.

It was reasonable to expect that the frequency of recovery of a particular type of HBV RNA sequence by PCR would likely correlate with the abundance of this particular RNA template in the tissue analyzed. The relative abundances of the three types of HBV RNAs in liver/HCC tissues harvested from chronic HBV carriers are summarized in Table 5. Table 5 also shows the fractions of the RNAs that bear deletions or insertions in HBV sequences and thus could have been defective. For tissue 8, all cccDNA-transcribed RNAs contained deletions. In all other cases, the fraction of RNAs with deletions/insertions did not exceed 50%, and in a number of tissues, no deletions/insertions were found for id-RNAs and RNAs polyadenylated via the cryptic viral poly(A) signal (Table 5).

Interestingly, id-RNAs were the only species found in all tissues. Furthermore, in the majority of the tissues (6/10; 2 nonmalignant liver tissues [tissues 7 and 25] and 4 HCC tissues [tissues 4, 14, 26, and 40]), id-RNAs were considerably abundant species, and in 5/6 of the above-mentioned tissues (tissues 4, 14, 25, 26, and 40), they were the most abundant species (Table 5). These results were somewhat unexpected, given that for all tissues, we were able to measure HBV replication markers (Table 3). Surprisingly, cccDNA-transcribed RNAs were predominant species in only two nonmalignant liver tissues, tissues 3 and 13 (i.e., in 2/10 tissues). In tissues 7, 8, and 25, cccDNA-transcribed

TABLE 5 Relative abundances of different types of HBV RNAs that accumulated in liver/HCC tissues determined by using a conventional RT-PCR approach

Tissue (liver or HCC)	No. (%) of sequences of different types of HBV RNAs accumulated in liver/HCC tissues; no. of insertions or deletions/total no. of sequences ^a			Total no. of HBV RNA sequences recovered from each tissue
	cccDNA-derived RNAs polyadenylated at or near position ~1931 ^b	RNAs polyadenylated using the cryptic poly(A) signal of HBV ^c	Integrand-derived RNAs ^d	
3 (liver)	15 (53.6); 1/15	10 (35.7); 2/10 ^e	3 (10.7); 1/3	28
4 (HCC)	0 (0.0)	0 (0.0)	<u>16 (100.0)</u> ; 0/16	16
7 (liver)	2 (10.0); 1/2	9 (45.0); 0/9	<u>9 (45.0)</u> ; 1/9	20
8 (HCC)	2 (7.7); 2/2	22 (84.6); 1/22	2 (7.7); 1/2	26
13 (liver)	4 (80.0); 2/4	0 (0.0)	1 (20.0); 0/1	5
14 (HCC)	0 (0.0)	0 (0.0)	<u>13 (100.0)</u> ; 1/13	13
25 (liver)	4 (23.5); 2/4	0 (0.0)	<u>13 (76.5)</u> ; 0/13	17
26 (HCC)	0 (0.0)	0 (0.0)	<u>22 (100.0)</u> ; 2/22	22
39 (liver)	8 (40.0); 2/8	9 (45.0); 3/9	3 (15.0); 0/3	20
40 (HCC)	0 (0.0)	2 (13.3); 0/2	<u>13 (86.7)</u> ; 1/13	15

^aShown are the numbers of sequences of each type of RNA and the fraction (percentage) of each type of HBV RNAs out of total number of HBV RNA sequences found in each tissue. Also shown are the fractions of sequences with deletions or insertions within HBV sequences. Insertions were found only for RNAs from tissue 3 that were polyadenylated using the cryptic viral poly(A) signal (see below). Our study did not analyze the ability of HBV RNAs to serve as the templates for the production of HBsAg. The host sequences incorporated into the 3' ends of id-RNAs could influence their functional potential (21). Thus, there is a possibility that host inserts in id-RNAs may compensate for a considerable loss of the 3' UTR of HBV, as in case of id-RNAs from tissue 4.

^bHBV RNAs that bear only HBV sequences at the 3' end and are polyadenylated at or near position ~1931 using the conventional HBV poly(A) signal UAUAAA located at positions 1916 to 1921 of the HBV genome (which is not present on the integrated DSL downstream of the ORFs for HBsAg [4, 15]) represent RNAs transcribed from cccDNA.

^cHBV RNAs that bear only HBV sequences at the 3' end and are polyadenylated using the cryptic HBV poly(A) signal CAUAAA located at positions 1788 to 1793 could represent either RNAs transcribed from cccDNA or integrant-derived RNAs, because the cryptic poly(A) signal is present in the correct position on the integrated DSL (downstream of the ORFs for HBsAg) and thus can facilitate the HBV-mediated polyadenylation of integrant-derived RNAs (4, 15, 55, 56).

^dHBV RNAs that bear the host sequences at the 3' end downstream of HBV sequences and immediately prior to the poly(A) tail and are polyadenylated using the host poly(A) signal represent id-RNAs. The cases when id-RNAs were relatively abundant/predominant HBV RNA species are underlined.

^eIn this case, 2 sequences (2/10) with insertions are indicated.

RNAs were the least abundant species, and in tissues 4, 14, 26, and 40, this type of HBV RNA was not detected by our PCR assay (Table 5). In addition, HBV RNAs that were polyadenylated via the viral cryptic poly(A) signal were detected in 5/10 tissues and were relatively abundant in tissues 3, 7, 8, and 39. Moreover, in tissues 7, 8, and 39, this type of HBV RNA was highly abundant or predominant (Table 5). Others also detected RNAs polyadenylated via the HBV cryptic poly(A) signal as frequent species in liver/HCC tissues (55). Given that this type of HBV RNA could have been produced either via HBV replication or from integrated HBV DNA (55), it is possible that we underestimated the levels of accumulation of the RNAs derived from HBV DNA integrants. In summary, the observed relative abundance of id-RNAs in a significant percentage of chronically infected tissues raise questions regarding the contribution of these RNAs to the HBV life cycle and HBV pathogenesis.

The fact that cccDNA-transcribed RNAs were predominant in only 2/10 tissues and were not found in 4/10 tissues (Table 5) suggested that the RNAs of the other two types [i.e., id-RNAs and/or the ones that acquired the poly(A) tail using the viral cryptic poly(A) signal] accumulated in most tissues to higher numbers. Therefore, the use of an assay that amplifies only cccDNA-transcribed RNAs could detect this type of HBV RNA in all tissues, because in all tissues, markers of HBV replication were quantified by qPCRs (Table 3). We thus designed a nested PCR assay that captures only cccDNA-transcribed RNAs. The PCR amplicon included a large part of an area of ~100 nt (positions ~1831 to 1931) (Fig. 4) that is expected to be present only on cccDNA-transcribed RNAs. In fact, primers 1045 and 1046 were placed within this 100-nt area, which, along with the design of the PCR amplicon (see Materials and Methods), ensured the exclusive amplification of cccDNA-transcribed RNAs. As expected, using this selective PCR, we obtained specific products of the expected sizes for every tissue (Fig. 11). The presence

FIG 11 Selective PCR amplification of the 3'-end regions of cccDNA-derived HBV RNAs present in liver/HCC tissues. This PCR assay exclusively amplifies cccDNA-derived RNAs. (Top) Results of RT-PCR amplification performed in the presence of reverse transcriptase (RT+). (Bottom) RT-PCR results obtained in the absence of reverse transcriptase (RT-) (controls). "no T" stands for the negative control that did not have the template for amplification. The lanes marked M contain double-stranded DNA size markers (a 1-kb plus DNA ladder [Fermentas] was used). The tissues assayed are shown at the top of the gel image. The arrow marks products of the expected size. For tissues 3, 4, 8, and 13, the PCR products of the expected size were gel purified, cloned, and sequenced, which confirmed the presence of the expected HBV sequences. Other PCR products were not analyzed further.

of the expected HBV sequences was confirmed by sequencing of the selected PCR products (not shown).

We then further analyzed the features of id-RNAs (Table 6). We recovered id-RNAs derived from 26 different integrations, of which 8/26 (30.8%) were within the ORFs of the host genes. The other two integrants were mapped to RNA genes that do not encode any proteins. All 10 of these above-mentioned id-RNA-producing integrants were mapped to intronic regions, while the rest of the integrants were located in intergenic regions. The id-RNAs were transcribed from integrants that mapped to chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 17, and X. The integrations in liver and HCC tissues were different. In a number of the id-RNAs, we found a relatively short sequence between the HBV sequence and the host insert, which could not be identified as being of either HBV or host origin. We speculate that such a sequence could result from DNA repair that occurred during the nonhomologous integration of HBV DNA into host DNA. In addition, the sequences of the id-RNAs derived from the same HBV integrant frequently demonstrated microheterogeneity in the 3'-end area (Table 6). In other words, the length of the host sequence insert could vary for id-RNAs derived from the same integrant, and/or there also could be up to several nonhost nucleotides located at the very 3' end (immediately downstream of the host insert). Possible explanations for the observed heterogeneity include (i) the utilization of different cleavage sites or different poly(A) signals during RNA processing leading to polyadenylation (57–59) and (ii) the use of primers that contained VN or VNN sequences at their very 3' ends. The former scenario could be a likely explanation for the heterogeneity in the lengths of host sequences in the 3'-end regions of id-RNAs. The latter scenario likely explains the appearance of the nonhost nucleotides at the very 3' end (right after the host insert). The sizes of host inserts at the 3' ends of id-RNAs ranged between 18 and

TABLE 6 Features of integrant-derived HBV RNAs found in matching liver/HCC tissues

Tissue (liver or HCC)	No. of different integrants ^a	Group ^b	No. of sequences_length of host insert at the 3' end (nt) ^c	No. of integrations into host ORFs ^d /total no. of integrants	No. of integrations into known recurrent sites ^e
3 (liver)	2	1 2	1_50 (+6) 2_281 (+14)	0/2	0
4 (HCC)	1	1 1 1 1	9_145 (± 2) 1_144 (± 2) 3_143 (± 2) 3_141 (± 2)	0/1	0
7 (liver)	7	1 2 3 4 5 6 7	1_35 (± 4) 1_40 (+1) 1_85 1_85 (+23) ^f 2_102 (± 2) 2_252 (+7) 1_354 (± 2)	2/7	0
8 (HCC)	2	1 2	1_21 (+2) 1_27 (+6) ^f	1/2	0
13 (liver)	1	1	1_29 (+3)	0/1	0
14 (HCC)	1	1 1 1 1	3_222 (± 1) 6_221 (± 1) 3_220 (± 1) 1_219 (± 1)	0/1	0
25 (liver)	7	1 2 2 2 3 4 5 6 7	1_58 1_64 (± 2) 4_63 (± 2) 1_18 (± 2) 1_96 (± 1) 2_101 1_215 1_237 (± 1) 1_291 (± 3) ^g	3/7	1/7 (FN1)
26 (HCC)	1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1_85 (± 2) 1_75 (± 2) 1_72 (± 2) 2_69 (± 2) 1_68 (± 2) 3_65 (± 2) 1_64 (± 2) 3_63 (± 2) 1_62 (± 2) 1_61 (± 2) 2_60 (± 2) 4_59 (± 2) 1_57 (± 2)	1/1	0
39 (liver)	2	1 2	2_31 (+2) 1_105 (± 4)	1/2	0
40 (HCC)	2	1 1 1 1 1 1 1	1_376 (± 1) 1_373 (± 1) 1_281 (± 1) 1_81 (± 1) 3_53 (± 1) 1_50 (± 1) 1_25 (± 1)	0/2	0

(Continued on next page)

TABLE 6 (Continued)

Tissue (liver or HCC)	No. of different integrants ^a	Group ^b	No. of sequences_length of host insert at the 3' end (nt) ^c	No. of integrations into host ORFs ^d /total no. of integrants	No. of integrations into known recurrent sites ^e
		1	1_23 (± 1)		
		2	2_83 (± 1) + 77 (+3) ^h		
		2	1_83 (± 1) + 15 (+3) ⁱ		

^aThe number of HBV integrants was determined by analysis of the host sequences at the 3' ends of the id-RNAs and by analysis of sequences at HBV-host junctions of the id-RNAs.

^bThe group number indicates that all RNAs that belong to the same group were derived from the same integrant. If, for example, the tissue has 7 HBV integrants that produced id-RNAs, then for this tissue, 7 groups of the id-RNAs are indicated (from 1 to 7) in order to distinguish the id-RNAs derived from different HBV integrants.

^cFor each group, the subsets of RNA sequences that have the same lengths of the host (i.e., human) sequence are shown. The reasons for the occurrence of the id-RNAs that are derived from the same HBV integrant but bear different lengths of host sequence (regardless of the same virus-host junction site) are discussed in the text. $\pm$ represents the number of nucleotides at the virus-host junction site that can be either of HBV or host origin, while + represents the number of nucleotides at the junction site that have no homology to either HBV or host sequences. The origin of the non-HBV/nonhuman sequences is discussed in the text. Thus, for example, for tissue 14, the information on the first line of the id-RNA characteristics, "1_3_222 (± 1)," means that within this subset of sequences of group 1, three id-RNAs that bear a human sequence of 222 nt were derived from the same integrant, and 1 nt at the junction site could have been of HBV or host origin. For cloned PCR products, we often encountered 1 to 3 nt at the very 3' end (immediately after the host insert) that were not of host origin. Such microheterogeneity likely reflects our RT-PCR strategy that employed primers bearing VN or VNN sequences at their 3' ends (see Materials and Methods). Cases of such microheterogeneity are not reflected here.

^dNumber of integrations into known human ORFs that occurred.

^eThere are approximately 20 human genes that currently are considered recurrent integration sites for HBV (see the text), which have been found in a number of HBV-infected individuals (1, 9, 14, 19, 60–63).

^fThis RNA contained a deletion in the HBV sequence.

^gThe host sequence found in this id-RNA had apparently 10 nt missing (as a series of small deletions) compared to the sequence retrieved from the NCBI database.

^hThese id-RNAs had a combination of two different host sequences of 83 and 77 nucleotides at the 3' end downstream of the virus-host junction (the second host sequence belonged to the known ORF). In addition, "2_83 (± 1) + 77 (+3)" also means that 1 nt at the virus-host junction site could belong to HBV or the host, while 3 nt of unknown origin were found between the two host sequences. The origin of such a combination of two host sequences is not immediately apparent and may represent complex recombination.

ⁱThis id-RNA had a combination of two different host sequences of 83 and 15 nt at the 3' end downstream of the virus-host junction. In addition, "1_83 (± 1) + 15 (+3)" also means that 1 nt at the virus-host junction site could belong to HBV or the host, while 3 nt of unknown origin were found between the two host sequences. Of note, the host insert 83 (± 1) + 15 (+3) represents a shortened version of the 83 (± 1) + 77 (+3) insert mentioned above, and it thus has shorter part of the second host sequence fragment.

376 nt. Interestingly, we found one integration (into fibronectin 1 [FN1] in tissue 25) that was reported previously to be a recurrent HBV integration site (60, 61). Currently, about 20 human genes have been recognized as recurrent targets for HBV integration, including TERT1 (telomerase reverse transcriptase), FN1, SMAD5 (SMAD family member 5), MLL4 (myeloid/lymphoid or mixed leukemia 4), ESRG (estrogen-related receptor gamma), RAI1 (retinoid acid induced 1), and MYOM1 (myosin 1, 185 kDa). It was suggested that recurrent integrations may be linked to HCC induction via insertional mutagenesis (14, 19, 60–63). However, the mechanism behind apparent recurrent HBV integrations remains unknown. One possible explanation is that HBV integration into the recurrent target gene is linked to a survival advantage of the cell bearing the integrant and/or may stimulate the proliferative activity of such a cell, which in turn could promote the clonal expansion of such an integrant-bearing cell and thus lead to frequent observations of integration into a particular recurrent target.

Analysis of the overlap area amplified by both 3'-end and 5'-end region PCR assays. We next analyzed the area of the overlap between the 3'-end and 5'-end region sequences obtained by using two different but overlapping PCR amplicons (see Materials and Methods) (Fig. 4). We found a number of completely matching 3'-end and 5'-end region sequences in the overlap area. Figure 12 shows an example of completely matching overlap sequences for id-RNAs from tissue 14. Table 7 summarizes closely matching overlap sequences for HBV RNAs found in all 5 sets of liver/HCC tissues. Completely matching sequences were found for all types of HBV RNAs. For the id-RNAs, 100% matching sequences in the overlapping area were found for tissues 4, 7, 8, 14, 25, 26, 39, and 40. For the rest of the tissues, we found id-RNAs that differ in the overlap area by one to several nucleotides (Table 7). The finding of completely matching sequences in the overlap area for the id-RNAs suggests that these RNAs could serve as functional mRNAs coding for HBsAg. Similar findings for the cccDNA-transcribed RNAs and RNAs polyadenylated using the HBV cryptic poly(A) signal

FIG 12 Example of a complete sequence match in the overlap area between the 3'-end and 5'-end regions of HBV RNA sequences amplified by two different PCR strategies. The alignment represents an example of a complete sequence match in the overlap area (in this case, positions 1074 to 1231 of the HBV sequence) and includes three 3'-end region id-RNA sequences (14_2, 14_4, and 14_8) and six 5'-end region RNA sequences (5_9_4, 5_9_5, 5_9_6, 5_9_11, 5_9_13, and 5_9_18) recovered from tissue 14. All three 3'-end region id-RNA sequences were identical in the overlap area. In addition, each of the 3'-end region id-RNA sequences had a 100% match with the above-mentioned six 5'-end region sequences (identical nucleotides are shown as dots). The identification numbers of the sequences are shown on the left-hand side of the alignment. The finding of a complete sequence match between 3'-end and 5'-end region sequences in the overlap area for the id-RNAs is consistent with the notion that functional HBsAg could be translated from these id-RNAs.

suggest that these two types of RNAs could also facilitate the translation of functional HBsAg.

Analysis of accumulated HBV RNAs in selected tissues using RNA sequencing (RNA-seq). We next used RNA-seq to further analyze the accumulation of HBV RNAs in tissues 3, 4, 7, and 8. In each case, another piece of tissue was cut, and total RNA was isolated, treated with DNase, and subjected to RNA-seq. Thus, different areas of the above-mentioned four tissues were analyzed in the present study by either conventional RT-PCR or RNA-seq. The results of RNA-seq are summarized in Table 8. Due to the nature of the RNA-seq procedure, which uses RNA fragmentation, it was not an option to readily identify and separate all the sequences located immediately next to the poly(A) tails and analyze just them. Therefore, RNA-seq could not be used for analysis of RNAs polyadenylated via the HBV cryptic poly(A) signal. Using specific sequence features, we quantified three types of HBV RNA. The first type was the RNAs, the transcription of which was initiated from the host promoter located upstream of the left-hand side of the HBV integrant (Fig. 5). Such RNAs contained a human sequence at the 5' end that was joined downstream to the HBV sequence at the 3' end. Therefore, these RNAs were identified as RNA-seq reads that bore the 5'-human-HBV-3' type of host-virus junction. These hybrid RNAs were found only in tissue 3 and represented 10.1% of HBV RNAs (Table 8). These RNAs were derived from integrants that mapped to one intronic region and one exonic region of the CFHR1 (complement factor H related 1) gene that is located on chromosome 1 and to one exonic region of DYSF (Dysferlin) gene that is located on chromosome 2. The number of RNA reads produced from each location ranged between 18 and 84. The HBV nucleotides at the host-virus junctions were mapped to positions 1818, 1985, and 402. The latter case is consistent with the significant loss of HBV sequences from the DSL (Fig. 5) (if in this case the DSL

TABLE 7 Closely matching sequences in the area of overlap between the 3'-end and 5'-end regions amplified from HBV RNAs that accumulated in liver/HCC tissues^a

Tissue (liver or HCC)	Type of HBV RNA	Overlap region (nt) ^b	3'-end region sequence identification no. ^c	No. of sequence matches (% identity), no. of nt mismatches ^d
3 (liver)	Integrand derived	338	3_T3_2	2 (99.7), 1
		338	3_T3_6	2 (99.7), 1
	Cryptic poly(A)	338	3_T3_4	2 (99.7), 1
		338	3_T3_7	2 (99.7), 1
		338	3_T3_8	2 (99.7), 1
		338	3_T3_9	2 (99.7), 1
		338	3_T3_12	2 (99.7), 1
	Cryptic poly(A)	158	3_2	2 (100)
	cccDNA transcribed	338	3_T3_10	2 (99.7), 1
	cccDNA transcribed	158	3_7	2 (100)
		158	3_12	2 (100)
		158	3_13	2 (100)
		158	3_15	2 (100)
		158	3_17	2 (100)
		158	3_18	2 (100)
4 (HCC)	Integrand derived	609	4RNA_1	1 (99.5), 3
		609	4RNA_3	1 (99.5), 3
		609	4RNA_4	1 (99.3), 4
		609	4RNA_5	1 (99.5), 3
		609	4RNA_6	1 (99.5), 3
		609	4RNA_7	1 (99.0), 6
		609	4RNA_11	1 (99.3), 4
		609	4RNA_12	1 (99.3), 4
		609	4RNA_14	1 (99.2), 5
	Integrand derived	338	T4_2	1 (100)
		338	T4_4	1 (100)
		338	T4_6	1 (100)
		338	T4_7	1 (100)
7 (liver)	Integrand derived	609	7RNA_5	1 (100)
		609	7RNA_1	1 (99.8), 1
		609	7RNA_7	1 (99.7), 2
		609	7RNA_8	1 (99.8), 1
	Integrand derived	338	T7_9	9 (100)
		158	7_9	10 (99.4), 1
	Cryptic poly(A)	609	7RNA_4	1 (100)
	Cryptic poly(A)	338	T7_2	9 (100)
		338	T7_11	9 (100)
		338	T7_12	9 (100)
		338	T7_13	9 (100)
		338	T7_14	9 (100)
	Cryptic poly(A)	158	7_11	10 (100)
		158	7up4	10 (100)
	cccDNA transcribed	158	T7_8	9 (100)
8 (HCC)	Integrand derived	338	T8_4	8 (100)
		338	3_8_1	8 (100)
	Cryptic poly(A)	338	3_8_3	8 (100)
		338	3_8_6	8 (100)
		338	3_8_7	8 (100)
		338	3_8_8	8 (100)
		338	3_8_11	8 (100)
		338	3_8_12	8 (100)
		338	3_8_15	8 (100)
		338	3_8_16	8 (100)
		338	T8_2	8 (100)
		338	T8_8	8 (100)
		158	8_4	10 (100)
	Cryptic poly(A)	158	8_7	10 (100)
		338	3_8_13	8 (99.7), 1

(Continued on next page)

TABLE 7 (Continued)

Tissue (liver or HCC)	Type of HBV RNA	Overlap region (nt) ^b	3'-end region sequence identification no. ^c	No. of sequence matches (% identity), no. of nt mismatches ^d
13 (liver)	Integrand derived cccDNA transcribed	338	3_T13_4	4 (99.7), 1
		338	3_T13_2	4 (100)
14 (HCC)	Integrand derived	158	14_2	6 (100)
		158	14_4	6 (100)
		158	14_8	6 (100)
25 (liver)	Integrand derived	609	25RNA9	1 (99.5), 3
		609	25RNA10	1 (99.5), 3
		609	25RNA15	1 (99.7), 2
	Integrand derived	338	T25_4RNA	2 (100)
		338	T25_12RNA	2 (100)
	Integrand derived	158	25_3	5 (100)
		158	25_5	5 (100)
		158	T25_3	5 (100)
		158	T25_4	5 (100)
	cccDNA transcribed	338	T25_13RNA	1 (100)
	cccDNA transcribed	338	25_13	5 (100)
26 (HCC)	Integrand derived	158	26_1	6 (100)
		158	26_5	6 (100)
		158	26_6	6 (100)
		158	26_10	6 (100)
		158	26_11	6 (100)
		158	T26_1	6 (100)
		158	T26_2	6 (100)
		158	T26_3	6 (100)
		158	T26_5	6 (100)
		158	T26_6	6 (100)
		158	T26_7	6 (100)
		158	T26_8	6 (100)
		158	T26_9	6 (100)
		158	T26_10	6 (100)
	Integrand derived	609	26RNA1	1 (100)
		609	26RNA2	1 (100)
		609	26RNA5	1 (100)
		609	26RNA7	1 (100)
		609	26RNA8	1 (100)
		609	26RNA9	1 (100)
39 (liver)	Integrand derived	338	3_39_01	2 (100)
		338	3_39_17	2 (100)
	Cryptic poly(A)	609	T39_6	1 (100)
		609	T39_7	1 (100)
		609	T39_10	1 (100)
		609	T39_12	1 (100)
	Cryptic poly(A)	338	3_39_13	2 (100)
		338	3_39_15	2 (100)
		338	3_39_16	2 (100)
	cccDNA transcribed	609	T39_1	1 (100)
		609	T39_8	1 (100)
	cccDNA transcribed	609	39_11	1 (100)
		338	3_39_18	2 (99.7), 1
40 (HCC)	Integrand derived	609	40_15	6 (100)
		609	40_14	6 (100)
		609	40_16	6 (100%)
	Integrand derived	338	3_40_13	7 (99.7), 1
		338	3_40_01	7 (99.7), 1
		338	3_40_05	7 (99.7), 1
		338	3_40_12	7 (99.7), 1
		338	3_40_62	7 (99.7), 1

(Continued on next page)

TABLE 7 (Continued)

Tissue (liver or HCC)	Type of HBV RNA	Overlap region (nt) ^b	3'-end region sequence identification no. ^c	No. of sequence matches (% identity), no. of nt mismatches ^d
	Cryptic poly(A)	609	40_21	6 (100)
		609	40_24	6 (100)

^aOnly data for closely matching sequences that have a degree of sequence identity of 99 to 100% in the overlap area are shown.

^bThe size of the overlap area depends on the position of the forward primer that was used to amplify the 3'-end region of HBV RNAs (in the second PCR round).

Three different forward primers were used in the second round of seminested PCR analyzing the 3'-end region of HBV RNAs. Details are described in Materials and Methods.

^cEach individual sequence was assigned a unique identification number.

^dThe data show how many 5'-end region sequences were close matches to a single 3'-end region sequence. The degree of sequence identity within the overlap area is shown in parentheses. The number of mismatched nucleotides is indicated for very similar sequences (<100% match).

was the substrate for integration). Further follow-up studies would be required to understand (i) the portion of the HBV sequence that these transcripts can bear, (ii) how these RNAs were polyadenylated, (iii) whether these RNAs can be relatively infrequent, and (iv) the kind of functions related to the course of HBV infection that these RNAs can mediate. The second type of HBV RNAs was cccDNA-derived RNAs. These RNAs were identified as reads that bore the region from positions ~1831 to ~1931 of the HBV genome, which was not expected to be present on the classic id-RNAs that originated from viral promoters and thus should be present almost exclusively on cccDNA-derived RNAs (4, 15) (Fig. 5). Two aspects, however, needed to be considered. First, the region covering positions ~1831 to 1931 would be present twice on pregenomic RNA and precore RNA [once in their 5'-end region and in addition next to the poly(A) tail], while other HBV RNAs produced by replication bear this region only once at their 3' ends (1, 9). RNA-seq, which produces short sequence reads, does not identify the two RNAs bearing the above-mentioned region twice. Therefore, it was not possible to account for two such kinds of RNA transcripts, and as a result, our counts of cccDNA-derived RNAs using RNA-seq could be somewhat overestimated. This was a limitation of the RNA-seq approach. Second, the above-described transcripts that initiated from the cellular promoters located upstream of the left-hand side of HBV integrants could also

TABLE 8 Relative abundances of different types of HBV RNAs that accumulated in selected liver/HCC tissues determined by using the RNA-seq approach^a

Tissue (liver or HCC)	No. of sequence reads (% [normalized no. of reads])			Total no. of reads corresponding to all three types of HBV RNAs recovered from each tissue (normalized no. of reads) ^e
	RNAs transcribed from integrants via an upstream host promoter ^b	cccDNA-derived RNAs ^c	id-RNAs ^d	
3 (liver)	142 ^f (10.1 [150.47])	144 (10.3 [152.59])	<u>1,115</u> (79.6 [1,181.49])	1,401 (1,484.55)
4 (HCC)	0 (0.0 [0])	3 (100.0 [1.12])	0 (0.0 [0])	3 (1.12)
7 (liver)	0 (0.0 [0])	7 (2.6 [15.05])	<u>260</u> (97.4 [559.15])	267 (574.20)
8 (HCC)	0 (0.0 [0])	0 (0.0 [0])	<u>142</u> (100.0 [151.06])	142 (151.06)

^aThree types of HBV RNAs are presented (see below). Shown are the numbers of sequence reads for each type of RNA and the fraction (percentage) of each type of RNA out of the total number of HBV RNA sequences (see below). The normalized numbers of reads (see Materials and Methods) are shown in square brackets. This value allows not only comparisons of numbers corresponding to the same tissue but also comparisons of numbers between different tissues.

^bRNA-seq allows the identification of 5'-human-HBV-3' junctions on the RNAs that were transcribed from the host promoter located upstream of the left end of the integrated DSL (Fig. 5). These hybrid RNAs were not targeted by our RT-PCR approach. This type of HBV RNA was found only in tissue 3.

^cRNAs that were produced by HBV genome replication (i.e., cccDNA-derived RNAs). The numbers of these RNAs were calculated by counting HBV reads that do not bear any host inserts and contain the sequences of the region spanning positions 1831 to 1931 in the HBV genome (see the text).

^dThe number of id-RNAs was calculated by counting the number of reads that contain sequences of 5'-HBV-human-3' junctions corresponding to the right-hand side of the integrated DSL (Fig. 5). Cases where id-RNAs were relatively abundant/predominant HBV RNA species are underlined.

^eAs mentioned above, RNA-seq cannot identify RNAs that were polyadenylated using the HBV cryptic poly(A) signal. Therefore, the total number of HBV RNA transcripts was calculated as a sum of cccDNA-derived RNAs, id-RNAs (Fig. 5), and integrant-derived HBV RNAs that were transcribed via the host promoter located upstream of the left end of the HBV integrant. Since RNA-seq produces only ~100-nt-long reads, it was impossible to conclude whether the RNAs of the latter type most of the time were terminated/polyadenylated within the HBV sequence of the DSL or whether they incorporated the entire sequence of the corresponding HBV integrant, crossed the HBV-host junction site at the right-hand side of the integrant, and then acquired the poly(A) tail using the host poly(A) signal. Therefore, for the purposes of the calculations presented here, we did not consider the above-described scenario.

bear the region of positions 1831 to 1931 (by copying the HBV sequences located close to the left end of the integrated DSL [Fig. 5]), which could potentially increase the total number of RNAs bearing the above-mentioned ~ 100 -nt region. In addition, the id-RNAs (Fig. 5) on occasion could bear a portion of the region spanning positions 1831 to 1931 (thus, using RT-PCR, we mapped a few 3'-end HBV nucleotides at virus-host junctions to position 1899). However, using bioinformatics tools, we excluded the above-described 5'-host-HBV-3' hybrid RNAs and the reads corresponding to id-RNAs prior to counting of the reads bearing the region spanning positions 1831 to 1931. The third type of HBV RNAs was the id-RNAs (Fig. 5). The id-RNAs were identified as reads that bear 5'-HBV-human-3' junctions corresponding to the right end of HBV integrant (Fig. 5). The id-RNAs identified by RNA-seq originated from a number of different random integration sites located at different chromosomes (i.e., chromosomes 2, 4, 6, 7, 8, 9, 11, 12, 13, and 15). On four occasions, the id-RNA-producing integrants were located within the intronic regions of the genes. These genes were MMADHC (methylmalonic aciduria and homocystinuria, CblD type), DST (Dystonin), ATP6V0D2 (ATPase H⁺-transporting V0 subunit D2), and RASAL1 (RAS protein activator like 1). All other id-RNA-producing integrants were mapped to intergenic regions. Therefore, as in the case of id-RNAs found by conventional RT-PCR, all id-RNAs identified by RNA-seq were also transcribed from the integrants located in the noncoding regions. The significance of this observation is not immediately apparent. The numbers of id-RNA-producing integrants were 4 for tissue 3, 0 for tissue 4, 7 for tissue 7, and 1 for tissue 8. The numbers of reads produced from the same integrant ranged between 12 and 609, which may suggest differences in transcription efficiencies. The HBV nucleotides at the virus-host junctions on id-RNAs were mapped between positions 1630 and 1827. Since RNA-seq output is sequence reads of about 100 nt, it was not possible to evaluate whether the full-length RNAs of the above-mentioned first type bore the sequences that overlap the sequences of id-RNAs. Therefore, such a possibility was not considered when the number of id-RNAs was quantified.

The id-RNAs were predominant (79.6%) in the area of tissue 3 analyzed by RNA-seq. In the same sample, cccDNA-derived RNAs represented 10.3% of HBV transcripts (Table 8). Similar to tissue 3, the id-RNAs were predominant species in RNA samples isolated from tissues 7 and 8. In tissue 7, id-RNAs represented 97.4% of HBV RNAs, while in tissue 8, id-RNAs represented 100%. Tissue 4 contained a very small amount of HBV RNAs. Only 3 reads of cccDNA-derived RNAs were found in this tissue. Overall, id-RNAs were the predominant HBV RNA species in 3 out of 4 analyzed tissues (Table 8). The relative abundance of id-RNAs was also observed in another area of tissue 7 by using RT-PCR (Table 5). In a different area of tissue 3, cccDNA-derived RNAs were more abundant than id-RNAs (Table 5). Equal abundance of cccDNA-derived RNAs and id-RNAs was found in another area of tissue 8, and only id-RNAs were found in a different area of tissue 4 (Table 5). Therefore, we observed certain differences in terms of the abundances of id-RNAs and cccDNA-derived RNAs between different areas of the same tissues that were analyzed by two different approaches (Tables 5 and 8), which, however, was not surprising because HBV RNAs and DNAs are not uniformly distributed among different areas of the infected tissues (4, 64). Importantly, the overall outcome of the second independent approach, RNA-seq, was consistent with that of the conventional RT-PCR approach and showed again that id-RNAs were relatively abundant in most of the examined tissues.

DISCUSSION

Progeny virus cannot be produced from HBV integrants, and it is commonly assumed that the HBV life cycle does not depend on integrated HBV DNA (1, 9, 13). In the present study, by using an RT-PCR approach, HBV integrant-derived RNAs (Fig. 5) were found in all tissues tested. In 7/10 tissues (4 HCC tissues and 3 nonmalignant liver tissues), id-RNAs represented a considerable fraction (between 20% and 100%) of the accumulated HBV RNAs (Table 5), which included (i) cccDNA-transcribed RNAs that were polyadenylated using the conventional viral poly(A) signal, (ii) RNAs that were

polyadenylated using the viral cryptic poly(A) signal and that can be generated either via HBV replication or by transcription from HBV integrants, and (iii) classic id-RNAs bearing an insertion of the host sequence at the 3' end immediately next to the poly(A) tail (Fig. 4 and 5). Surprisingly, in 6/10 tissues (2 liver tissues and 4 HCC tissues), id-RNAs were either very abundant or predominant HBV RNA species (between 45% and 100%) (Table 5) regardless of the presence of the markers of HBV replication in the same tissues (Table 3). In addition, different areas of tissues 3, 4, 7, and 8 were analyzed via RNA-seq that showed the predominance of id-RNAs in 3/4 analyzed tissues (Table 8), which was consistent with the overall findings of the conventional RT-PCR approach (although the comparison of RNA-seq and RT-PCR data suggested that different areas of the same tissue can have different levels of accumulation of cccDNA-derived RNAs and id-RNAs [Tables 5 and 8]).

During the comparison of the outcomes of RT-PCR and those of RNA-seq, we considered several important aspects.

(i) RT-PCR allows the analysis of sequences much longer than the ~100-nt reads produced by RNA-seq. This allows the analysis of sequence alterations within a considerable area of a single-molecule sequence. Using this feature of RT-PCR, we were able to discard sequences that likely represented defective HBV RNAs.

(ii) We focused on major and distinct RT-PCR products and thus could have missed products that were relatively not abundant, while RNA-seq analyzed all the products of the amplification reaction.

(iii) RNA-seq produces reads of a much larger number of sequences (while RT-PCR is expected to generate a somewhat limited number of clones for sequencing) and may demonstrate higher sensitivity than RT-PCR.

(iv) The RNA fragmentation used by RNA-seq and the short length of final reads make it not a simple task to identify the origins of HBV RNAs, since all types of HBV RNAs have significant regions of overlapping sequences. Conventional RT-PCR has some advantages in this respect, which depend on the location and size of the amplified regions.

(v) RT-PCR could be biased toward shorter PCR products, which likely can result in the poor amplification of id-RNAs that bear a relatively long host insert at the 3' end. No such bias is noted for RNA-seq.

(vi) It was not an option to simply identify and extract all the sequences adjacent to the poly(A) tails using RNA-seq. Thus, RNA-seq could not identify RNAs that were polyadenylated via the cryptic HBV poly(A) signal and polyadenylated HBV RNAs that were truncated in the 3'-end area and contained only HBV sequences. In contrast, RT-PCR easily identified both of these RNA categories.

(vii) Only RT-PCR could produce full-length host inserts located at the 3' ends of id-RNAs (but, as mentioned above, the long host inserts are less likely to be amplified by RT-PCR), while RNA-seq was unable to provide information on how long the host sequence inserts were in id-RNAs.

(viii) RNA-seq allowed the detection of hybrid transcripts that initiated from cellular promoters located upstream of the left end of the integrants and contained human sequences at the 5' end that were joined to HBV sequences at the 3' end (Table 8), while our RT-PCR approach was not designed to capture this type of RNA.

Overall, either of the two approaches has its own limitations, and the use of a particular approach or a combination of approaches should depend on the task of the study. In our study, these two approaches complemented each other. Furthermore, the observed relative abundance of HBV id-RNAs (using two independent approaches) in chronically infected tissues have several interesting implications.

Given that (i) we found completely matched sequences within the overlap area between the 3'-end and 5'-end region sequences (via RT-PCR) for id-RNAs (for tissues 4, 7, 8, 14, 25, 26, 39, and 40) (Table 7), (ii) id-RNAs were relatively abundant (Tables 5 and 8), and (iii) HBV envelope promoters are in the correct positions and could remain fully functional in HBV integrants (Fig. 5) (4, 15, 20, 21, 24), it became apparent that id-RNAs could produce quite a significant fraction of viral envelope proteins. This can

at least in part explain the observations that a loss of HBsAg was not achieved in at least 93% of cases for all anti-HBV drugs analyzed, which included pegylated interferon, lamivudine, telbivudine, entecavir, adefovir, and tenofovir (8). Clearly, anti-HBV drugs can reduce only the levels of HBsAg produced by HBV replication and have no effect on HBsAg translated from id-RNAs independently of HBV replication. Thus, our data (i) suggest at least one reason why serum HBsAg could not reflect cccDNA levels (i.e., the extent of HBV replication [65]) and further support the significance of monitoring intrahepatic cccDNA levels and (ii) indicate that guidelines on the management of HBV infection could be revised in terms of the expectation of a loss of serum HBsAg as a result of anti-HBV treatments.

Our data are also consistent with the fact that HBV-infected individuals can support HDV infection regardless of the presence of HBV replication markers (66–68). Furthermore, our results confirmed and extended our previous work that suggested that HDV can likely persist in the absence of HBV replication *in vivo*, if an alternative source of functional HBsAg is available (24). The relative abundance of id-RNAs in tissues from chronic HBV carriers once again draws our attention to this novel mechanism of the maintenance of HDV persistence that does not depend on HBV replication and thus further support the development and implementation of antivirals that target HDV directly. Currently, there are no drugs in clinical practice that target HDV directly (68, 70).

Considering the observed relative abundance of id-RNAs, it became apparent that id-RNAs can contribute to the life cycle of HBV during the chronic stage of infection. Figure 13 depicts two scenarios. First, when an infected cell supports HBV replication but either it does not contain integrated HBV DNA or the integrant is altered (i.e., bears a deletion[s], mutation[s], or rearrangement[s], etc.) and does not generate id-RNA that could produce HBsAg, all HBV virions produced would be enveloped with replication-derived HBsAg (rd-HBsAg) (Fig. 13A). Second, the cell supports HBV replication and contains an HBV integrant(s) that produces considerable amounts of id-RNAs coding for HBsAg, compared to the amounts of the envelope mRNAs produced by viral replication. In this case, a sizeable fraction of assembled and released HBV virions can bear id-HBsAg (and not rd-HBsAg) as a major component of their envelopes (Fig. 13B). Furthermore, our data (Tables 5 and 8) suggest that during chronic HBV infection, a considerable fraction of hepatocytes and HCC cells could produce id-RNAs. This is consistent with previous reports that most HCCs bear integrated HBV DNA, and high levels of HBV integration in infected hepatocytes were observed relatively early, in the first (immunotolerant) phase of chronic HBV infection (14, 19, 61, 71, 72). Thus, in the second scenario (Fig. 13B), the properties of HBV virions would be defined largely by the properties of the id-HBsAg. Therefore, id-HBsAg could considerably influence the infectivity of HBV virions and their ability to support virus spread and superinfection, which indicates that id-RNAs (i) can contribute to the diversity and variability of HBV virions, (ii) may influence HBV survival, and (iii) may have a so-far-unrecognized role in the maintenance of the chronic state of HBV infection (because we recently obtained data that suggested that hepadnavirus spread and superinfection may represent determinants of the maintenance of chronic hepadnavirus infection [28]). The contribution of id-HBsAg would have a greater impact in a situation where cccDNA-transcribed RNAs (i) do not produce high levels of HBsAg or (ii) produce defective HBsAg, while the envelope proteins translated from id-RNAs are relatively abundant/fully functional. This case would represent a potential positive impact of id-RNAs on the HBV life cycle. In the opposite situation, where the majority of HBsAg comes from id-RNAs and bears defects, the contribution of such id-RNAs could suppress the infectivity and spread of HBV, which would represent a negative impact on the HBV life cycle. In another situation, where HBV envelope proteins that came from replication and from id-RNAs have similar functional properties, the impact of id-HBsAg may not be significant. However, if the majority of HBsAg came from id-RNAs and is fully functional, then, in theory, we may expect increased levels of HBV assembly. However, the feasibility of the latter scenario needs to be addressed in follow-up experiments, which would examine how different ratios of id-HBsAg/

FIG 13 Potential contribution of id-RNAs to the life cycle of HBV. The model of the HBV life cycle is based on the one that was reported previously by Mason et al. (13). Briefly, about 90% of HBV virions contain rcDNA inside the nucleocapsids, while the other ~10% bear the DSL. The virions enter susceptible cells via a specific receptor, sodium taurocholate-cotransporting polypeptide (NTCP) (10). Inside the cell, uncoating takes place, and virions lose the envelope that consists of three HBV envelope proteins, L, M, and S (shown as bars), and some lipid. DNA synthesis resumes within the nucleocapsid, and the plus DNA strand is completed. Viral polymerase is shown as a green circle. After nucleocapsid entry into the nucleus and subsequent disassembly, the rcDNA loses the polymerase and is converted into cccDNA, which is the template for the synthesis of all HBV RNAs produced by viral replication. The newly synthesized pgRNA forms a complex with the polymerase and is assembled into the nucleocapsid, after which reverse transcription starts inside the capsid. In 80 to 95% of cases, the end product of reverse transcription is rcDNA, while in 5 to 20% of cases (if, during plus DNA strand synthesis, the RNA primer is not translocated and *in situ* priming occurs), the final product of reverse transcription is the DSL. Both encapsidated rcDNA and the DSL can be enveloped and released as HBV virions or can be recycled into the nucleus to replenish the cccDNA pool. In addition, the DSL is the main substrate for HBV integration, which occurs randomly and is facilitated by cellular DNA repair enzymes. As shown in Fig. 5, the integrated DSL, independently of HBV replication, can serve as the template for the synthesis of id-RNAs coding for the viral envelope proteins. For simplicity, the mRNAs for HBsAg are shown as a single horizontal line. The red line represents replication-derived mRNAs for HBsAg, while the blue line indicates the id-RNAs coding for HBsAg. (A) Shown is an HBV-infected cell that supports HBV but does not produce id-RNAs. In this case, either the cell does not contain an HBV integrant or the integrant is altered (i.e., can bear a mutation[s], deletion[s], or rearrangement[s], etc.), and therefore id-RNAs are not produced, which is indicated by a purple cross. In this case, only cccDNA-derived mRNAs for HBsAg are produced, and HBV replication is the only source of HBsAg. (B) Case where mRNAs for HBsAg are produced by virus replication and are also derived from the integrated HBV DNA. In this situation, the envelope proteins are produced from cccDNA-derived RNAs and from id-RNAs. As our data suggest, when id-RNAs represent a relatively abundant/predominant fraction of mRNAs for HBsAg, the newly made HBV virions will contain id-RNA-derived HBsAg as a major component of their outer envelopes, which further suggests that the properties of these virions will be determined largely by the properties of the id-RNA-derived envelope proteins and not by the properties of HBsAg produced by HBV replication. Thus, id-RNA-derived HBsAg can determine the ability of HBV virions to support virus spread and superinfection, which in turn may affect the ability of the virus to maintain a state of chronic infection. Recently, we showed that limited hepadnavirus spread and superinfection continue during chronic infection *in vivo* and thus may represent the determinants of the maintenance of the chronicity state (28). In addition, id-RNA-derived HBsAg, if abundant, could make a significant contribution to HBsAg-associated liver pathogenesis independently of HBV replication. The above-described conclusions suggest that id-RNAs may have a considerable contribution to the HBV life cycle and HBV-related pathogenesis.

rd-HBsAg may influence the rate of HBV assembly in the context of ongoing HBV genome replication.

Therefore, further studies that would compare the properties of id-HBsAg and rd-HBsAg in terms of their abilities to facilitate the assembly and infectivity of HBV are

warranted in order to better define the impact of id-RNAs on the HBV life cycle. In addition, considering that a significant fraction of HBsAg during chronic infection could be produced from id-RNAs (Tables 5 and 8) independently of HBV replication, follow-up studies should also investigate the impact of id-HBsAg on liver pathogenesis. This would help to determine whether id-RNAs and/or HBV integrants should be considered as targets for novel antiviral interventions. The significance of HBsAg-mediated pathogenesis was recognized previously (12, 29–33). It needs to be noted that when HBV replication is inhibited by a drug, the role of id-HBsAg in HBV-related pathogenesis could increase, because the production of id-RNAs and id-HBsAg is independent of HBV replication and is unaffected by anti-HBV treatments.

Several caveats, however, need to be mentioned. First, although the use of five sets of matching liver/HCC samples allowed us to recognize the phenomenon of the relative abundance of id-RNAs during chronic HBV infection, examination of a considerably larger number of tissues collected at different clinical stages of chronic HBV infection would now be warranted in order to better evaluate the extent of the observed phenomenon and to determine its significance for the understanding of the mechanisms regulating chronic HBV infection/pathogenesis and possibly for the development of novel treatments targeting id-HBsAg production that occurs independently of HBV replication.

Second, the data in Table 5 are based on the amplification of the 3'-end regions of HBV RNAs and not the complete sequences of the mRNAs coding for the L envelope protein that were amplified within a single PCR amplicon. In reconstruction experiments using transfected Huh7 cells replicating HBV, we were able to amplify the entire L mRNA sequences within a single amplicon (not shown), while this appears not to be feasible for total RNA isolated from infected tissues. This result suggests that lower copy numbers of RNA were present in HBV-infected tissues than in transfected Huh7 cells, which precluded the amplification of a rather long PCR product (for cccDNA-derived L mRNA, the PCR product was expected to be about 2.4 kb long). Of note is the fact that by using the same cDNA, we were able to amplify the 5'-end sequences of L mRNAs, which overlapped with the above-mentioned 3'-end region sequences (completely matched sequences in the overlap area are summarized in Table 7). This is consistent with the expression of full-length L mRNAs from the integrants. Furthermore, the main precursor for HBV integration is the DSL, and as shown in Fig. 5, the HBV promoters/ORFs for the L and M/S mRNAs are in the correct positions and likely to be functional (if the integrant is unaltered). In addition, the interpretation of data regarding the transcription of the full-size L mRNAs from the integrants is also supported by the fact that by using RT-PCR for analysis of 3'-end HBV RNA sequences, we found that (i) in 7/10 tissues, the fraction of id-RNAs was quite considerable, i.e., between 20% and 100%, and in 6/10 tissues, id-RNAs represented between 45% and 100% of all viral RNAs (Table 5) and (ii) RNA-seq also found a predominance of id-RNAs in the majority of analyzed tissues (Table 8). There is also an alternative possibility that upstream of the left-hand side of the DSL integrant, there could be a cellular promoter that could drive the expression of RNA from the HBV insert. RNA-seq found transcripts that contained HBV sequences and were initiated from the upstream (relative to the HBV integrant) cellular promoters but only in tissue 3 (Table 8). However, whether such transcripts (i) can be abundant and (ii) may function as efficient templates for HBsAg translation needs to be addressed in follow-up studies.

Third, we have not directly quantified HBV RNA levels using our RT-PCR approach. Our interpretations regarding the abundances of different HBV RNA species are based on the analysis of PCR-amplified and cloned sequences. The assumption here was that the sequences from more abundant RNAs will also be more abundant among the generated PCR products. As also mentioned above, our RT-PCR approach used for the analysis of the 3'-end regions of HBV RNAs could favor the amplification of shorter PCR products, such as id-RNAs bearing relatively short host inserts. However, our PCR strategy amplified HBV RNAs of different origins (Table 5), and the lengths of host

sequences preceding the poly(A) tail in id-RNAs varied between 18 and 376 nucleotides (Table 6), although no id-RNAs with longer host inserts were amplified. On the other hand, we showed that in tissues 4, 14, 26, and 40, for which the RT-PCR approach did not reveal the presence of cccDNA-transcribed RNAs (Table 5), cccDNA-derived RNAs could be detected when a selective PCR assay amplifying exclusively replication-produced RNAs (and not id-RNAs) was used (Fig. 11). These data are not inconsistent with the interpretation that in a number of tissues, accumulated cccDNA-transcribed RNAs were not abundant. Finally, it needs to be considered that another independent approach, RNA-seq, that has no bias toward the lengths of the host sequences at the 3' ends of id-RNAs and produces a number of sequence reads that is proportional to the number of RNA templates showed a predominance of id-RNAs in 3/4 analyzed tissues (Table 8).

Fourth, our interpretations regarding the abundance of the integrant-derived RNAs could represent an underestimation, because the species that are polyadenylated using the viral cryptic poly(A) signal represented a considerable fraction of HBV RNAs in 4/10 tissues (between 35% and 85%) (Table 5), and they could be derived from integrated HBV DNA (4, 15, 55, 56). Furthermore, a previous study suggested that the cryptic HBV poly(A) signal is used efficiently if the conventional poly(A) signal is somewhat inactivated or removed (56), thus indicating why, during HBV replication, the conventional poly(A) signal is used at least as the preferable (if not exclusive) means of polyadenylation. That work (56) and the fact that we did not observe alterations of either conventional or cryptic poly(A) signals in most of the HBV genomes [i.e., we found only a single nucleotide substitution in the cryptic poly(A) signal in 2 sequences from tissue 7 and in 1 sequence from tissue 39, and we also found a single nucleotide substitution in the conventional poly(A) signal in 1 sequence from tissue 39] further increase the likelihood that RNAs that were polyadenylated via the cryptic poly(A) signal were derived from the integrants, because in the DSL, the conventional poly(A) signal is located upstream of the envelope promoters and ORFs and in a sense can be considered as removed (Fig. 5), while the cryptic poly(A) signal is in the correct position downstream of the posttranscriptional regulatory element (PRE) and can facilitate the polyadenylation of integrant-derived RNAs. Thus, if RNAs that used the cryptic HBV poly(A) signal were transcribed mostly from the integrants, then integrant-derived RNAs [id-RNAs and RNAs that used the cryptic poly(A) signal] could have been the predominant HBV RNA species in 8/10 tissues analyzed (Table 5).

Fifth, we have not conducted analyses at the single-cell level. The interpretation that a considerable number of HBV virions may bear a significant fraction of id-HBsAg per virion is based on our data (Tables 5 and 8) that suggest that in many tissues, id-RNAs were abundant/predominant. Therefore, it was reasonable to assume that a considerable portion of the cells replicating HBV could also bear HBV integrants that produce id-RNAs. The impact of id-RNAs would depend, of course, on the ratio of id-RNAs/cccDNA-transcribed RNAs within the cell. Our interpretation was further supported by the findings of apparently not altered HBV sequences in the context of id-RNAs and by our previous observations that at least a portion of naturally integrated HBV DNA inserts is able to produce functional HBV envelope proteins that support the infectivity of assembled virions (24).

In addition, during the preparation of the manuscript, we became aware of a recent paper by Wooddell et al. (73). Using RNA from livers of two chimpanzees chronically infected with HBV that were treated with entecavir for several weeks, the authors of that study performed single-molecule real-time (SMRT) sequencing and showed that in one animal, (i) id-RNAs represented 66.4% of all HBV transcripts and (ii) RNAs that were polyadenylated using the HBV cryptic poly(A) signal accumulated to considerable levels among the rest of the HBV transcripts. These data are consistent with our findings and conclusions.

MATERIALS AND METHODS

Tissue samples. Five sets of matching nonmalignant liver tissue samples and HCC samples were obtained from a collection of tissues harvested from untreated individuals chronically infected with HBV. For each pair, the odd number corresponds to a nonmalignant liver tissue, while the even number represents an HCC sample. Thus, in pair 3/4, the number 3 corresponds to the liver tissue, while the number 4 corresponds to the matching HCC tissue. Pair 3/4 was obtained from a 40-year-old female patient, pair 7/8 was obtained from a 31-year-old female patient, pair 13/14 was obtained from a 26-year-old male patient, pair 25/26 was obtained from a 55-year-old male patient, and pair 39/40 was obtained from a 50-year-old male patient. The corresponding serum samples were not available. Certain features of this collection of tissues were reported previously; however, extensive sequencing of HBV genomes, characterization of HBV replication markers, and analysis of the accumulated HBV RNA species in liver and HCC tissues have not been conducted (69, 74–77).

Sequencing of HBV genomes. Liver (or HCC) tissues (approximately 50 mg) were homogenized in approximately 1 ml of a 10:10 TE buffer (10 mM Tris-HCl [pH 7.6], 10 mM EDTA) by using a glass grinder (28). The 105 μ l of the resulting tissue homogenate was then mixed with 195 μ l of a 10:10 TE buffer. The diluted homogenate was mixed with 300 μ l of a buffer containing 10 mM Tris-HCl (pH 7.6), 10 mM EDTA, 0.2% SDS, and 4 mg/ml pronase and incubated at 37°C for 2 h. After incubation, total DNA was extracted twice with an equal volume of a 1:1 phenol-chloroform mixture and precipitated by the addition of 2 volumes of 100% ethanol and 0.1 volume of 3.0 M sodium acetate (pH 5.5). DNA pellets were resuspended in 25 μ l of DNase-free/RNase-free water (78). The complete HBV genome sequences from all analyzed liver/HCC tissues, with the exception of tissues 39 and 40, were amplified by using nested PCR assays that spanned three different overlapping fragments, covering the entire HBV genome (Fig. 1). Fragment B spans positions 246 to 1790, fragment F1 covers positions 1653 to 2317, and fragment F2 spans positions 2267 to 395 (Table 1). The positions of the fragments, primers, and real-time PCR (qPCR) probes throughout the paper are shown based on the sequences of HBV genotypes B and C (79). HBV genomic sequences from tissues 39 and 40 were first amplified by using 2 overlapping fragments. Thus, the initial PCR product spanning positions 58 to 1820 (i.e., fragment 1) was then used as the template in a second PCR to generate fragment B, whereas another PCR product covering the region spanning positions 1593 to 3215/1 to 449 (i.e., fragment 2) was subsequently used as the template in 2 different PCRs in order to produce fragments F1' (spanning the region of positions 1653 to 2395) and F2' (positions 2297 to 3215/1 to 395). Fragment F1' amplified from tissue 40 had to be subdivided into 2 smaller fragments, fragments M (positions 1653 to 2141) and N (positions 2012 to 2395). The nucleotide positions of the primers and primer sequences that were used for the above-described PCR amplifications are listed in Table 1. All PCRs were performed with a 25- μ l reaction mixture containing 1 $\times$ KlenTaq buffer, 200 μ M deoxynucleoside triphosphate (dNTP) mix (each dNTP at final concentration of 200 μ M), 0.5 μ M forward and reverse primers, 5 U of KlenTaq DNA polymerase (AB Bioscience), and 2 μ l of either total DNA isolated from the tissue sample or the first PCR product (i.e., the product generated by the first round of nested PCR). The PCR products were analyzed by agarose gel electrophoresis and then gel purified and cloned into the pCRII-TOPO vector. The resulting plasmids were analyzed by sequencing. The sequences were then assembled in order to produce complete HBV genomes. The consensus HBV sequences for each tissue sample were subsequently generated by using MacVector 11.1.1 software.

Quantification of cccDNA of HBV. Specificity for the quantification of cccDNA was achieved by combining an isolation procedure that enriches for cccDNA and a PCR assay that is highly selective for cccDNA. cccDNA isolation was performed as previously described (80). Briefly, for each sample, 210 μ l of the above-mentioned tissue homogenate was first diluted with 930 μ l of a 10:10 TE buffer. The diluted homogenate was then mixed with 60 μ l of a 10% SDS solution and 300 μ l of a 2.5 M KCl solution. The resulting mixture was vortexed and incubated for 30 min at room temperature. After incubation, the mixture was centrifuged for 30 min at 4°C, and the resulting supernatant was first extracted with 1 volume of phenol, followed by a second extraction with an equal volume of a 1:1 phenol-chloroform mixture. The DNA was then precipitated by the addition of 2 volumes of ethanol. The resulting DNA pellet was resuspended in 50 μ l DNase-free/RNase-free H₂O, and 40 μ l of the DNA sample was subjected to overnight digestion at 37°C with 50 to 100 U of plasmid-safe ATP-dependent DNase (Epicentre) in a 500- μ l reaction mixture containing 1 $\times$ NEB buffer 4, 1 mM ATP, and 40 μ g/ml of RNase A. Quantification of cccDNA levels was performed by qPCR using the following primers: forward primer 943 (1685-AAC GACCGACCTTGAGGC-1702), reverse primer 945 (1946-GTAACTCCACAGAAGCTCCAAATTCT-1921), and TaqMan probe 944 (1859–6-carboxyfluorescein [FAM]–ACTGTTCAAGCCTCAAGCTGTGCCTT–3' black hole quencher 1 [3BHQ_1]–1885). A 10-fold dilution series of plasmid pCB102 linearized with HindIII was used to generate a calibration curve (range from 20 to 2.0 $\times$ 10⁷ genome equivalents [GE] of HBV). Plasmid pCB102 is a modified version of plasmid pODW282, which contains a 1.3 $\times$ HBV genome sequence (GenBank accession number [D00330](#)) (genotype B) cloned at the BamHI site of vector pSP65 (81). pODW282 was digested with the EcoRI and EcoRV restriction enzymes, and the larger DNA fragment was gel purified and religated. The resulting plasmid contains a slightly-shorter-than-full-length HBV monomer spanning nucleotide positions 1065 to 3215/1 to 1043. Furthermore, a single nucleotide substitution, A(2465)T, was introduced into the HBV sequence by oligonucleotide-directed mutagenesis. It reflected the nucleotide that was different between plasmid pODW282 and the actual HBV genomes recovered from infected tissues. Thus, the resulting construct, pCB102, served as the HBV monomer standard for qPCR. Two types of DNA controls were used in order to evaluate the efficiency of the cccDNA isolation procedure. Plasmid pCB102 was used as the surrogate for cccDNA, and a gel-purified greater-than-HBV-genome-length PCR product (positions 1820 to 3215/1 to 1831) was used as the surrogate for the double-stranded HBV linear genome (DSL) (4, 15). The use of these two surrogates

allowed us to monitor the DNase treatment that was designed to eliminate HBV DNA intermediates other than cccDNA (i.e., rcDNA, single-stranded DNA [ssDNA], DSL DNA, and other incomplete HDV DNA intermediates) and to verify that, after treatment, PCRs for total HBV DNA (see the details of the corresponding qPCR assay below) and for cccDNA produce very similar numbers, thus indicating that the measured resulting HBV DNA species practically represent only cccDNA.

The total DNA concentrations in DNA preparations isolated from the collected liver/HCC tissues were measured by using Hoechst 33258 and used for the normalization of cccDNA levels.

A qPCR assay that measured total intrahepatic HBV DNA levels in the DNA preparations obtained from tissues 25 and 26 employed primers 946 (2340-CTGTTGTTAGACGAAGATGCAGGT-2363) and 948 (2475-CCACCTTATGTGTCCAAGGAATACTAAC-2448) and TaqMan probe 947 (2407-FAM-ATCGCCGCTCG CAGAAGATCTCAATC-3BHQ_1-2433). A 10-fold dilution series of plasmid pCB101 linearized with Sall was used to generate the calibration curve for quantification, which spanned a range from 20 to 2.0×10^7 GE of HBV. Plasmid pCB101 was constructed by the insertion of the HBV genome sequence reported under GenBank accession number [AB205119](#) (genotype B) into the XbaI and SacI restriction sites of vector pSVL (Pharmacia). The inserted HBV sequence spanned nucleotide positions 1820 to 3215/1 to 1825 and thus exceeds the length of the HBV monomer by 5 nucleotides, but regardless, plasmid pCB101 served as the HBV monomer standard for qPCR.

For the DNA preparations derived from the remainder of the liver/HCC tissue samples, qPCR to measure total HBV DNA levels was performed by using primers 997 (2307-ATGCCCTATCTTATCAACAC TTCCG-2332) and 1038 (2487-CAGTAAAGTTTCCACCTTATGAGTCCA-2460) and TaqMan probe 1039 (2366-FAM-CCTAGAAGAAGAACTCCCTCGCTCGC-3BHQ_1-2392). A 10-fold dilution series of plasmid pCB102 linearized with HindIII was used to generate the calibration curve (range from 20 to 2.0×10^7 GE of HBV).

Quantification of pregenomic/precore RNA of HBV. Total RNA from liver/HCC tissue samples was extracted by using Tri reagent (Molecular Research Center) according to the manufacturer's instructions. The isolated total RNA (5 μ g) was incubated with 10 U of Turbo DNase (Fisher) for 2 h at 37°C. After this incubation, 10 more units of Turbo DNase were added to the reaction mixture, and incubation proceeded further for two additional hours at 37°C. DNase-treated RNA (dtrRNA) samples were reextracted with Tri reagent and then used for cDNA synthesis. Up to 2 μ g of the reextracted dtrRNA was reverse transcribed by using a high-capacity cDNA reverse transcription kit (Fisher) and 10 pmol of the HBV-specific primers. Primer 950 (2523-TAGGATTAAGCAAGGTACCGTAGAAG-2497) was used to reverse transcribe dtrRNA samples isolated from tissues 25 and 26. pgRNA levels were quantified by qPCR using primers 946 and 948 and TaqMan probe 947, as described above for total DNA quantification. cDNA synthesis for the remaining dtrRNA samples was carried out with primer 1038, and qPCR was performed by using primers 997 and 1038 and TaqMan probe 1039. The locations of the primers and probes used in these two qPCRs ensure the amplification of the RNA sequences that are present only on the pregenomic/precore RNAs of HBV (1, 9, 82, 83).

PCR amplification of 5'-end and 3'-end sequences of HBV RNAs. Using the NucleoTrap mRNA kit (Macherey-Nagel), poly(A) RNA was isolated from total RNA that was extracted from the tissue samples using Tri reagent. Up to 200 μ g of total RNA was used for each isolation cycle. Following elution, the poly(A) RNA was treated with Turbo DNase and reextracted with Tri reagent. The final poly(A) RNA pellet was resuspended in 12 μ l of DNase-free/RNase-free water. Up to 4 μ l of DNase-treated poly(A) RNA was used for reverse transcription with ProtoScript II reverse transcriptase (New England BioLabs), which was conducted for 2 h at 42°C. cDNA synthesis was carried out by using 100 pmol of a 3-base-anchored oligo(dT) primer. Two versions of the anchored oligo(dT) primer were used: (i) 980, which contained a stretch of 12 dT residues (5'-GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTTTTTTNN-3'), and 982, which harbored a stretch of 18 dT residues (5'-AGCACAGAATTAATACGACTCACTATAGGTTTTTTTTTT TTTTTTTTTTNN-3' [V represents either dA, dC, or dG but not dT, while N could be any of the four nucleotides dA, dC, dG, and dT]). Having the VNN sequences at the 3' ends is supposed (i) to avoid amplifications from internal A-rich regions of the RNA and (ii) to ensure that the copied species are true polyadenylated RNAs (84). The 5'-end sequences of both primers 980 and 982 played the role of adapter sequences during subsequent conventional PCRs. Using the cDNA generated, the sequences of the 5'-end and 3'-end regions of the HBV poly(A) RNAs were amplified by nested and seminested PCRs, respectively. KlenTaq DNA polymerase (AB Bioscience) was used for all PCRs.

The 5'-end HBV sequences, including the 5' UTR, the entire ORF for the large (L) envelope protein, as well as part of the 3' UTR, were amplified by using either oligonucleotide 820 (2807-ATTTTGCGGT CACCATATCTTGGGAACA-2836) or oligonucleotide 815 (2807-ATTTTGCGGT CACCATATCTTGGGAACA-2836) as a forward primer (genomic sense) and reverse primer 1013 (1274-GCAGTATGGATCGCAGAG GAG-1253) in the first round of PCR amplification. The second round of PCR was performed by using the product of the first PCR as the template and either primer 821 (2816-GTACCATATCTTGGGAACAAGA GCTA-2843) or primer 816 (2816-GTACCATATCTTGGGAACAAGATCTAC-2844) as the forward primer and reverse primer 1012 (1251-CACAAAGGTTCCACGCATGC-1232). Forward primers 815 and 816 were used exclusively to amplify the HBV sequences from tissue samples 25 and 26. This PCR strategy amplified the sequences of both HBV cccDNA-transcribed RNAs and HBV DNA integrant-derived RNAs (id-RNAs).

The 3'-end HBV sequences of the poly(A) RNAs, which were produced exclusively by means of virus replication, were amplified by using primers 1017 (754-TTGGGGGCAAGTCTGTACAA-774) and 1045 (1930-TCCAAATCTTTATACGGGTCAATGTCCAT-1901) in the first round of PCR and primers 1015 (865-T CCCCTAACTTCATGGGATATGTAATTGG-893) and 1046 (1905-TCCATGCCCTAAAGCCACC-1887) in the second PCR round. This nested PCR selectively amplified only cccDNA-transcribed RNAs polyadenylated at

or around position 1931 and did not amplify RNAs that were produced from the integrated HBV DNA regardless of HBV replication and bore various-length inserts of the host sequences at their 3' ends.

Another seminested PCR strategy was used to amplify the 3'-end regions of either id-RNAs or cccDNA-transcribed RNAs. The PCR employed either primer 1017 or 1015 (on some occasions) as the forward primer for the first PCR round and the VN-anchored LNA (locked nucleic acid) oligo(dT) primers (54) as the reverse primers (for both PCR rounds). Reverse primer 910 (5'-ATACGACTCACTATAGGT+TTT+TTT+TTT+TTVN-3') was used when cDNA synthesis was done by using primer 980, whereas primer 908 (5'-CGACTCACTATAGGTTTTT+TTTTT+TTTTT+TVN-3') was employed when reverse transcription was carried out with primer 982. "+" represents locked dT nucleotides. The second round of PCR was carried out by using forward primer 1015 (when 1017 was used in the first round of PCR) or 1016 (1042-TATCCTGCTTAAATGCCCTTATATGCATGTAT-1073) (when 1015 was used in the first round of PCR). The longer 3'-end fragments of the isolated poly(A) RNAs were amplified by using the oligonucleotide 882 (541-AGGAACCTCTATGTTCCCTCATGTTG-567) or primer 884 (541-AGGAACCTCTATGTTCCCTCTTG-567) as the forward primer in the first PCR round and primer 883 (599-TGTATCCCATCCCATCATC-TTGG-622) or primer 885 (599-TGTATCCCATCCCATCATCCTGG-622) in the second round of PCR. Forward primers 882 and 883 were used to amplify HBV sequences from tissues 25 and 26 only.

On some occasions, cDNA synthesis and PCR were conducted by using the Qiagen LongRange 2Step RT-PCR kit. In these cases, reverse transcription reactions were performed by using up to 4 µg of total RNA and primer 835 (5'-GCTGTCAACGATACGCTACGTAACGGCATGACAGTGTTTTTTTTTTTTTTT-3'). Primers 882 (or 884) and 836 (5'-GCTGTCAACGATACGCTACGTAACG-3') were used in the first round of PCR, and primers 883 (or 885) and 837 (5'-CGCTACGTAACGGCATGACAGTG-3') were used in the second round of PCR amplification.

The generated PCR products were separated by agarose gel electrophoresis, gel purified, and cloned into the pCRII-TOPO vector. The resulting constructs were analyzed by sequencing.

RNA sequencing. RNA sequencing was performed on the individual libraries generated by using DNase-treated RNAs isolated from tissues 3, 4, 7, and 8 on an Illumina HiSeq2500 sequencing system (Illumina, San Diego, CA). Sequencing libraries were prepared by using the Illumina TruSeq stranded mRNA sample preparation kit (A, catalog number FC-122-2101; B, catalog number FC122-2102). RNA sequencing was performed at a 100-cycle paired-end resolution. Viral transcripts and virus-host junctions on hybrid (5'-host-HBV-3' or 5'-HBV-host-3') transcripts were detected with the primary aid of three software tools, Virus-Clip (85), STAR (86), and IGV (87). Following the steps of Virus-Clip, reads were first aligned to the HBV genome by using BWA software (88). Soft-clipped reads that resulted from the alignment were extracted by using SAMTools software (89). These soft-clipped segments were then mapped to the human genome (GRCh38) by using BLASTN software (90). This provided mapping of the exact location, to a single-base resolution, of the virus-host junction sites on the human and viral genomes. The consensus HBV genome sequences generated for each tissue were used as the corresponding references. The strand-specific nature of the RNA sequencing protocol was used to determine the precise strand orientation and direction of transcription of the hybrid (or chimeric) transcripts (i.e., 5'-host-HBV-3' or 5'-HBV-host-3') or viral (HBV) transcripts. The virus-host junctions were further validated by identifying chimeric HBV-human reads after alignment of the sequences to the combined human (GRCh38) and viral HBV genomes, where the viral genome was added as an additional chromosome, using STAR software as part of STAR-Fusion. By using this information, the 5'-host-HBV-3' or 5'-HBV-host-3' transcripts were manually assembled (joined at their respective junction sites), and sequenced reads were mapped to them by using BWA. The mapped reads were visually inspected by using IGV software. Only those transcripts with mapped reads that spanned at least 5 bases on both sides of the junction site were retained. The cccDNA-derived RNAs (i.e., the RNAs that were produced by HBV genome replication) were identified (after the reads corresponding to chimeric 5'-host-HBV-3' and 5'-HBV-host-3' transcripts were identified and removed) as the reads that bore the sequences of the region between positions 1831 and 1931 on the HBV genome that was expected to be uniquely present (after the exclusion of the above-mentioned reads) only on RNAs produced by virus replication and polyadenylated at or around position 1931 using the HBV conventional poly(A) signal (1, 4, 15). The above-mentioned region was not expected to be present on hybrid 5'-HBV-host-3' transcripts that were initiated from HBV promoters. Further details about the approach used for the identification of the origins of the transcripts of interest are described in Results. The levels of cccDNA-derived RNAs in each sample were estimated from the average read count per base in the above-mentioned region of positions 1831 to 1931 after the removal of already identified virus-host hybrid RNAs. The abundances of the HBV-human transcripts (5'-host-HBV-3' and 5'-HBV-host-3') were estimated by counting the number of reads spanning the virus-host junction sites using MATLAB software (MathWorks Inc.). For normalization, the number of reads mapped to each gene in the human genome was estimated by using HTSeq software (91). The raw read counts of the cccDNA-derived RNAs and virus-host hybrid transcripts along with the read counts of individual genes were normalized by using the trimmed mean of M values (TMM) normalization method (92).

ACKNOWLEDGMENTS

S.O.G. was supported by NIH grants R21AI099696 and R01CA166213 and by a School of Medicine of the University of Kansas bridging grant. The Genomics Core is supported by the Kansas Intellectual and Developmental Disability Research Center at the University of Kansas Medical Center (KUMC) (which is supported by NIH grant U54

HD090216) and by COBRE grant Molecular Regulation of Cell Development and Differentiation (5P20GM104936-10).

We are very grateful to Jianming Hu for constructive discussions and valuable comments. We thank the KUMC Genomics Core for generating the sequence data sets. We are also very grateful to Clark Bloomer, Rosanne Skinner, and Yafen Niu for the guidance and excellent assistance with RNA-seq experiments.

REFERENCES

- Seeger C, Mason WS. 2015. Molecular biology of hepatitis B virus infection. *Virology* 479–480:672–686. <https://doi.org/10.1016/j.virol.2015.02.031>.
- Lupberger J, Hildt E. 2007. Hepatitis B virus-induced oncogenesis. *World J Gastroenterol* 13:74–81. <https://doi.org/10.3748/wjg.v13.i1.74>.
- Zoulim F, Mason WS. 2012. Reasons to consider earlier treatment of chronic HBV infection. *Gut* 61:333–336. <https://doi.org/10.1136/gutjnl-2011-300937>.
- Mason WS, Low HC, Xu C, Aldrich CE, Scougall CA, Grosse A, Clouston A, Chavez D, Litwin S, Peri S, Jilbert AR, Lanford RE. 2009. Detection of clonally expanded hepatocytes in chimpanzees with chronic hepatitis B virus infection. *J Virol* 83:8396–8408. <https://doi.org/10.1128/JVI.00700-09>.
- Arbuthnot P, Kew M. 2001. Hepatitis B virus and hepatocellular carcinoma. *Int J Exp Pathol* 82:77–100. <https://doi.org/10.1111/j.1365-2613.2001.iep178.x>.
- Shariff MI, Cox IJ, Gomaa AI, Khan SA, Gedroyc W, Taylor-Robinson SD. 2009. Hepatocellular carcinoma: current trends in worldwide epidemiology, risk factors, diagnosis and therapeutics. *Expert Rev Gastroenterol Hepatol* 3:353–367. <https://doi.org/10.1586/egh.09.35>.
- Lun-Gen L. 2014. Antiviral therapy of liver cirrhosis related to hepatitis B virus infection. *J Clin Transl Hepatol* 2:197–201. <https://doi.org/10.14218/JCTH.2014.00022>.
- Lampertico P, Agarwal K, Berg T, Buti M, Janssen HLA, Papatheodoris G, Zoulim F, Tacke F. 18 April 2017. EASL clinical practice guidelines on the management of hepatitis B virus infection. *J Hepatol* <https://doi.org/10.1016/j.jhep.2017.03.021>.
- Seeger C, Mason WS. 2000. Hepatitis B virus biology. *Microbiol Mol Biol Rev* 64:51–68. <https://doi.org/10.1128/MMBR.64.1.51-68.2000>.
- Yan H, Zhong G, Xu G, He W, Jing Z, Gao Z, Huang Y, Qi Y, Peng B, Wang H, Fu L, Song M, Chen P, Gao W, Ren B, Sun Y, Cai T, Feng X, Sui J, Li W. 2012. Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. *Elife* 1:e00049. <https://doi.org/10.7554/eLife.00049>.
- Blanchet M, Sureau C. 2007. Infectivity determinants of the hepatitis B virus pre-S domain are confined to the N-terminal 75 amino acid residues. *J Virol* 81:5841–5849. <https://doi.org/10.1128/JVI.00096-07>.
- Pollicino T, Cacciola I, Saffiotti F, Raimondo G. 2014. Hepatitis B virus PreS/S gene variants: pathobiology and clinical implications. *J Hepatol* 61:408–417. <https://doi.org/10.1016/j.jhep.2014.04.041>.
- Mason WS, Litwin S, Xu C, Jilbert AR. 2007. Hepatocytes turnover in transient and chronic hepatitis B virus infection. *J Viral Hepat* 14(Suppl 1):22–28. <https://doi.org/10.1111/j.1365-2893.2007.00911.x>.
- Jiang Z, Jhunhunjwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S, Diao J, Yeung S, Jubb A, Ye W, Wu TD, Kapadia SB, de Sauvage FJ, Gentleman RC, Stern HM, Seshagiri S, Pant KP, Modrusan Z, Ballinger DG, Zhang Z. 2012. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res* 22:593–601. <https://doi.org/10.1101/gr.133926.111>.
- 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* 84:8308–8315. <https://doi.org/10.1128/JVI.00833-10>.
- Shaul Y, Ziemer M, Garcia PD, Crawford R, Hsu H, Valenzuela P, Rutter WJ. 1984. Cloning and analysis of integrated hepatitis virus sequences from a human hepatoma cell line. *J Virol* 51:776–787.
- Su TS, Hwang WL, Yauk YK. 1998. Characterization of hepatitis B virus integrant that results in chromosomal rearrangement. *DNA Cell Biol* 17:415–425. <https://doi.org/10.1089/dna.1998.17.415>.
- Takada S, Gotoh Y, Hayashi S, Yoshida M, Koike K. 1990. Structural rearrangement of integrated hepatitis B virus DNA as well as cellular flanking DNA is present in chronically infected hepatic tissues. *J Virol* 64:822–828.
- Hai H, Tamori A, Kawada N. 2014. Role of hepatitis B virus DNA integration in human carcinogenesis. *World J Gastroenterol* 20:6236–6243. <https://doi.org/10.3748/wjg.v20.i20.6236>.
- Ou J, Rutter WJ. 1985. Hybrid hepatitis B virus-host transcripts in human hepatoma cell. *Proc Natl Acad Sci U S A* 82:83–87.
- von Loringhoven AF, Koch S, Hofschneider PH, Koshy R. 1985. Co-transcribed 3' host sequences augment expression of integrated hepatitis B virus DNA. *EMBO J* 4:249–255.
- Niknafs YS, Chinnaiyan AM. 2014. RNA identity crisis: hepatitis B walks the LINE. *Cancer Cell* 25:259–260. <https://doi.org/10.1016/j.ccr.2014.02.011>.
- Chakraborty PR, Ruiz-Opazo N, Shouval D, Shafritz DA. 1980. Identification of integrated hepatitis B virus DNA and expression of viral RNA in HBsAg-producing human hepatocellular carcinoma cell line. *Nature* 286:531–533. <https://doi.org/10.1038/286531a0>.
- Freitas N, Cunha C, Menne S, Gudima SO. 2014. Envelope proteins derived from naturally integrated hepatitis B virus DNA support assembly and release of infectious hepatitis delta virus particles. *J Virol* 88:5742–5754. <https://doi.org/10.1128/JVI.00430-14>.
- Taylor J. 2000. Hepatitis delta virus. *Virology* 344:71–76. <https://doi.org/10.1016/j.virol.2005.09.033>.
- Smedile A, Rizzetto M. 2011. HDV: thirty years later. *Dig Liver Dis* 43(Suppl 1):S15–S18. [https://doi.org/10.1016/S1590-8658\(10\)60687-1](https://doi.org/10.1016/S1590-8658(10)60687-1).
- Menne S, Cote P. 2007. The woodchuck model as an animal model for pathogenesis and therapy of chronic hepatitis B virus infection. *World J Gastroenterol* 13:104–124. <https://doi.org/10.3748/wjg.v13.i1.104>.
- Rodrigues L, Freitas N, Kallakury BV, Menne S, Gudima SO. 2015. Superinfection with woodchuck hepatitis virus strain WHVNY of livers chronically infected with strain WHV7. *J Virol* 89:384–405. <https://doi.org/10.1128/JVI.02361-14>.
- Block TM, Mehta A, Fimmel CJ, Jordan R. 2003. Molecular viral oncology of hepatocellular carcinoma. *Oncogene* 22:5093–5107. <https://doi.org/10.1038/sj.onc.1206557>.
- Xu HZ, Liu YP, Guleng B, Ren JL. 2014. Hepatitis B virus-related hepatocellular carcinoma: pathogenic mechanism and novel therapeutic interventions. *Gastrointest Tumors* 1:135–145. <https://doi.org/10.1159/000365307>.
- Churin Y, Roderfeld M, Roeb E. 2015. Hepatitis B virus large surface protein: function and fame. *Hepatobiliary Surg Nutr* 4:1–10. <https://doi.org/10.3978/j.issn.2304-3881.2014.12.08>.
- Block TM, Guo H, Guo JT. 2007. Molecular virology of hepatitis B virus for clinicians. *Clin Liver Dis* 11:685–706. <https://doi.org/10.1016/j.cld.2007.08.002>.
- Li YW, Yang FC, Lu HQ, Zhang JS. 2016. Hepatocellular carcinoma and hepatitis B virus surface protein. *World J Gastroenterol* 22:1943–1952. <https://doi.org/10.3748/wjg.v22.i6.1943>.
- Lago BV, Mello FC, Kramvis A, Niel C, Gomes SA. 2014. Hepatitis B virus subgenotype A1: evolutionary relationships between Brazilian, African and Asian isolates. *PLoS One* 9:e105317. <https://doi.org/10.1371/journal.pone.0105317>.
- Yang J, Xing K, Deng R, Wang J, Wang X. 2006. Identification of hepatitis B virus putative intergenotype recombinants by using fragment typing. *J Gen Virol* 87(Part 8):2203–2215. <https://doi.org/10.1099/vir.0.81752-0>.
- Sugachi F, Orito E, Ichida T, Kato H, Sakugawa H, Kakumu S, Ishida T, Chutaputti A, Lai CL, Ueda R, Miyakawa Y, Mizokami M. 2002. Hepatitis B virus genotype B with or without recombination with genotype C over the pre-core region plus core gene. *J Virol* 76:5985–5992. <https://doi.org/10.1128/JVI.76.12.5985-5992.2002>.
- Xie JX, Zhao J, Yin JH, Zhang Q, Pu R, Lu WY, Zhang HW, Wang HY, Cao GW. 2010. Association of novel mutations and haplotypes in the preS

- region of hepatitis B virus with hepatocellular carcinoma. *Front Med China* 4:419–429. <https://doi.org/10.1007/s11684-010-0160-0>.
38. Xie J, Zhang Y, Zhang Q, Han Y, Yin J, Pu R, Shen Q, Lu W, Du Y, Zhao J, Han X, Zhang H, Cao G. 2013. Interaction of signal transducer and activator of transcription 3 polymorphisms with hepatitis B virus mutations in hepatocellular carcinoma. *Hepatology* 57:2369–2377. <https://doi.org/10.1002/hep.26303>.
 39. Zhu Y, Jin Y, Guo X, Bai X, Chen T, Wang J, Qian G, Groopman JD, Gu J, Li J, Tu H. 2010. Comparison study on the complete sequence of hepatitis B virus identifies new mutations in core gene associated with hepatocellular carcinoma. *Cancer Epidemiol Biomarkers Prev* 19:2623–2630. <https://doi.org/10.1158/1055-9965.EPI-10-0469>.
 40. Caligiuri P, Cerruti R, Icardi G, Bruzzone B. 2016. Overview of hepatitis B virus mutations and their implications in the management of infection. *World J Gastroenterol* 22:145–154. <https://doi.org/10.3748/wjg.v22.i1.145>.
 41. Liao Y, Hu X, Chen J, Cai B, Tang J, Ying B, Wang H, Wang L. 2012. Precore mutation of hepatitis B virus may contribute to hepatocellular carcinoma risk: evidence from an updated meta-analysis. *PLoS One* 7:e38394. <https://doi.org/10.1371/journal.pone.0038394>.
 42. Liu S, Zhang H, Gu C, Yin J, He Y, Xie J, Cao G. 2009. Association between hepatitis B virus mutations and the risk of hepatocellular carcinoma: a meta-analysis. *J Natl Cancer Inst* 101:1066–1082. <https://doi.org/10.1093/jnci/djp180>.
 43. Yano Y, Azuma T, Hayashi Y. 2015. Variations and mutations in the hepatitis B virus genome and their associations with clinical characteristics. *World J Hepatol* 7:583–592. <https://doi.org/10.4254/wjh.v7.i3.583>.
 44. Tatsukawa M, Takaki A, Shiraha H, Koike K, Iwasaki K, Kobashi H, Fujioka S, Sakaguchi K, Yamamoto K. 2011. Hepatitis B virus core promoter mutations G1613A and C1653T are significantly associated with hepatocellular carcinoma in genotype C HBV-infected patients. *BMC Cancer* 11:458. <https://doi.org/10.1186/1471-2407-11-458>.
 45. Sung FY, Lan CY, Huang CJ, Lin CL, Liu CJ, Chen PJ, Lin SM, Yu MW. 2016. Progressive accumulation of mutations in the hepatitis B virus genome and its impact on time to diagnosis of hepatocellular carcinoma. *Hepatology* 64:720–731. <https://doi.org/10.1002/hep.28654>.
 46. Jia J, Li H, Wang H, Chen S, Wang M, Feng H, Gao Y, Wang Y, Fang M, Gao C. 2017. Hepatitis B virus core antigen mutations predict post-operative prognosis of patients with primary hepatocellular carcinoma. *J Gen Virol* 98:1399–1409. <https://doi.org/10.1099/jgv.0.000790>.
 47. Lee JH, Han KH, Lee JM, Park JH, Kim HS. 2011. Impact of hepatitis B virus (HBV) x gene mutations on hepatocellular carcinoma development in chronic HBV infection. *Clin Vaccine Immunol* 18:914–921. <https://doi.org/10.1128/CVI.00474-10>.
 48. Sung JJ, Tsui SK, Tse CH, Ng EY, Leung KS, Lee KH, Mok TS, Bartholomewsz A, Au TC, Tsoi KK, Locarnini S, Chang HL. 2008. Genotype-specific genomic markers associated with primary hepatomas, based on complete genomic sequencing of hepatitis B virus. *J Virol* 82:3604–3611. <https://doi.org/10.1128/JVI.01197-07>.
 49. Dong H, Qian Z, Zhang L, Chen Y, Ren Z, Ji Q. 2015. Genomic and transcriptome profiling identified both human and HBV genetic variations and their interactions in Chinese hepatocellular carcinoma. *Genom Data* 6:1–3. <https://doi.org/10.1016/j.gdata.2015.07.018>.
 50. Yeung P, Wong DK, Lai CL, Fung J, Seto WK, Yuen MF. 2011. Association of hepatitis B virus pre-S deletions with the development of hepatocellular carcinoma in chronic hepatitis B. *J Infect Dis* 203:646–654. <https://doi.org/10.1093/infdis/jiq096>.
 51. Mason WS, Litwin S, Jilbert AR. 2008. Immune selection during chronic hepatitis B virus infection. *Hepatology* 47:23–36. <https://doi.org/10.1007/s12072-007-9024-3>.
 52. Xu C, Yamamoto T, Zhou T, Aldrich CE, Frank K, Cullen JM, Jilbert AR, Mason WS. 2007. The liver of woodchucks chronically infected with the woodchuck hepatitis virus contains foci of virus core antigen-negative hepatocytes with both altered and normal morphology. *Virology* 359:283–294. <https://doi.org/10.1016/j.virol.2006.09.034>.
 53. Freitas N, Salisse J, Cunha C, Toshkov I, Menne S, Gudima SO. 2012. Hepatitis delta virus infects the cells of hepatitis B virus-induced hepatocellular carcinoma in woodchucks. *Hepatology* 56:76–85. <https://doi.org/10.1002/hep.25663>.
 54. Jacobsen N, Eriksen J, Nielsen PS. 2011. Efficient poly(A)⁺ RNA selection using LNA oligo(dT) capture. *Methods Mol Biol* 703:43–51. https://doi.org/10.1007/978-1-59745-248-9_4.
 55. Kairat A, Beerheide W, Zhou G, Tang ZY, Edler L, Schroder CH. 1999. Truncated hepatitis B virus RNA in human hepatocellular carcinoma: its representation in patients with advancing age. *Intervirology* 42:228–237. <https://doi.org/10.1159/000024982>.
 56. Schutz T, Kairat A, Schroder CH. 1996. DNA sequence requirements for the activation of a CATAAA polyadenylation signal within the hepatitis B virus X reading frame: rapid detection of truncated transcripts. *Virology* 223:401–405. <https://doi.org/10.1006/viro.1996.0495>.
 57. Pauws E, van Kampen AH, van de Graaf SA, de Vijlder JJ, Ris-Stalpers C. 2001. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res* 29:1690–1694. <https://doi.org/10.1093/nar/29.8.1690>.
 58. Beaudoin E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* 10:1001–1010. <https://doi.org/10.1101/gr.10.7.1001>.
 59. Neilson JR, Sandberg R. 2010. Heterogeneity in mammalian RNA 3'-end formation. *Exp Cell Res* 316:1357–1364. <https://doi.org/10.1016/j.yexcr.2010.02.040>.
 60. Ding D, Lou X, Hua D, Yu W, Li L, Wang J, Gao F, Zhao N, Ren G, Li L, Lin B. 2012. Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next generation sequencing-based approach. *PLoS Genet* 8:e1003065. <https://doi.org/10.1371/journal.pgen.1003065>.
 61. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, Mulawadi FH, Wong KF, Liu AM, Poon RT, Fan ST, Chan KL, Hu Y, Lin Z, Wang G, Zhang Q, Barber TD, Chou WC, Aggarwal A, Hao K, Zhou W, Zhang C, Hardwick J, Buser C, Xu J, Kan Z, Dai H, Mao M, Reinhard C, Wang J, Luk JM. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* 44:765–769. <https://doi.org/10.1038/ng.2295>.
 62. Toh ST, Jin Y, Liu L, Wang J, Babrzadeh F, Gharizadeh B, Ronaghi M, Toh HC, Chow PK, Chung AY, Ooi LL, Lee CG. 2013. Deep sequencing of the hepatitis B virus in hepatocellular carcinoma patients reveals enriched integration events, structural alterations and sequence variations. *Carcinogenesis* 34:787–798. <https://doi.org/10.1093/carcin/bgs406>.
 63. 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* 29:703–708. <https://doi.org/10.1002/humu.20701>.
 64. Zhang X, Lu W, Zheng Y, Wang W, Bai L, Chen L, Feng Y, Zhang Z, Yuan Z. 2016. In situ analysis of intrahepatic virological events in chronic hepatitis B virus infection. *J Clin Invest* 126:1079–1092. <https://doi.org/10.1172/JCI83339>.
 65. Lin LY, Wong VW, Zhou HJ, Chan HY, Gui HL, Guo SM, Wang H, Huang L, Bao SS, Xie Q, Chan HL. 2010. Relationship between serum hepatitis B virus DNA and surface antigen with covalently closed circular DNA in HBeAg-negative patients. *J Med Virol* 82:1494–1500. <https://doi.org/10.1002/jmv.21863>.
 66. Govindarajan S, Fields HA, Humphrey CD, Margolis HS. 1986. Pathologic and ultrastructural changes of acute and chronic delta hepatitis in an experimentally infected chimpanzee. *Am J Pathol* 122:315–322.
 67. Tavanez JP, Cunha C, Silva CM, David E, Monjardino J, Carmo-Fonseca M. 2002. Hepatitis delta virus ribonucleoproteins shuttle between the nucleus and the cytoplasm. *RNA* 8:637–646. <https://doi.org/10.1017/S1355838202026432>.
 68. Pollicino T, Raffa G, Santantonio T, Gaeta GB, Iannello G, Alibrandi A, Squadrito G, Cacciola I, Calvi C, Colucci G, Levrero M, Raimondo G. 2011. Replicative and transcriptional activities of hepatitis B virus in patients coinfecting hepatitis B and hepatitis delta viruses. *J Virol* 85:432–439. <https://doi.org/10.1128/JVI.01609-10>.
 69. Connolly E, Melegari M, Landgraf P, Tchaikovskaya T, Tennant BC, Slagle BL, Rogler LF, Zavolan M, Tuschl T, Rogler CE. 2008. Elevated expression of the miR-17-92 polycistron and miR-21 in hepatitis B virus-associated hepatocellular carcinoma contributes to the malignant phenotype. *Am J Pathol* 173:856–864. <https://doi.org/10.2353/ajpath.2008.080096>.
 70. Koytak ES, Yurdaydin C, Glenn JS. 2007. Hepatitis D. *Curr Treat Options Gastroenterol* 10:456–463. <https://doi.org/10.1007/s11938-007-0045-8>.
 71. Kennedy PT, Litwin S, Dolman GE, Bertoletti A, Mason WS. 2017. Immune tolerant chronic hepatitis B: the unrecognized risks. *Viruses* 9:E96. <https://doi.org/10.3390/v9050096>.
 72. Mason WS, Gill US, Litwin S, Zhou Y, Peri S, Pop O, Hong ML, Naik S, Quaglia A, Bertoletti A, Kennedy PT. 2016. HBV DNA integration and clonal hepatocyte expansion in chronic hepatitis B patients considered immune tolerant. *Gastroenterology* 151:986–998. <https://doi.org/10.1053/j.gastro.2016.07.012>.
 73. Wooddell CI, Yuen MF, Chan HL, Gish RG, Locarnini SA, Chavez D, Ferrari

- C, Given BD, Hamilton J, Kanner SB, Lai CL, Lau JYN, Schluep T, Xu Z, Lanford RE, Lewis DL. 2017. RNAi-based treatment of chronically infected patients and chimpanzees reveals that integrated hepatitis B virus DNA is a source of HBsAg. *Sci Transl Med* 9:eaan0241. <https://doi.org/10.1126/scitranslmed.aan0241>.
74. Scorsone KA, Zhou YZ, Butel JS, Slagle BL. 1992. p53 mutations cluster at codon 249 in hepatitis B virus-positive hepatocellular carcinomas from China. *Cancer Res* 52:1635–1638.
 75. Slagle BL, Zhou YZ, Butel JS. 1991. Hepatitis B virus integration event in human chromosome 17p near the p53 gene identifies the region of the chromosome commonly deleted in virus-positive hepatocellular carcinomas. *Cancer Res* 51:49–54.
 76. Slagle BL, Zhou YZ, Birchmeier W, Scorsone KA. 1993. Deletion of the E-cadherin gene in hepatitis B virus-positive Chinese hepatocellular carcinomas. *Hepatology* 18:757–762. <https://doi.org/10.1002/hep.1840180402>.
 77. Becker SA, Zhou YZ, Slagle BL. 1996. Frequent loss of chromosome 8p in hepatitis B virus-positive hepatocellular carcinomas from China. *Cancer Res* 56:5092–5097.
 78. Mason WS, Xu C, Low HC, Saputelli J, Aldrich CE, Scougall C, Grosse A, Colonna R, Litwin S, Jilbert AR. 2009. The amount of hepatocytes turnover that occurred during resolution of transient hepadnavirus infection was lower when virus replication was inhibited with entecavir. *J Virol* 83:1778–1789. <https://doi.org/10.1128/JVI.01587-08>.
 79. Kramvis A, Kew M, Francois G. 2005. Hepatitis B virus genotypes. *Vaccine* 23:2409–2423. <https://doi.org/10.1016/j.vaccine.2004.10.045>.
 80. Freitas N, Lukash T, Rodrigues L, Litwin S, Kallakury BV, Menne S, Gudima SO. 2015. Infection patterns induced in naive adult woodchucks by virions of woodchuck hepatitis virus collected during either the acute or chronic phase of infection. *J Virol* 89:8749–8763. <https://doi.org/10.1128/JVI.00984-15>.
 81. Okamoto H, Tsuda F, Sakugawa H, Sastrosowignjo RI, Imai M, Miyakawa Y, Mayumi M. 1988. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *J Gen Virol* 69: 2575–2583. <https://doi.org/10.1099/0022-1317-69-10-2575>.
 82. Jeong JK, Yoon GS, Ryu WS. 2000. Evidence that the 5'-end cap structure is essential for encapsidation of hepatitis B virus pregenomic RNA. *J Virol* 74:5502–5508. <https://doi.org/10.1128/JVI.74.12.5502-5508.2000>.
 83. Laras A, Koskinas J, Hadziyannis SJ. 2000. In vivo suppression of pre-core mRNA synthesis is associated with mutations in hepatitis B virus core promoter. *Virology* 295:86–96. <https://doi.org/10.1006/viro.2001.1352>.
 84. Boehm V, Gerbracht JV, Marx MC, Gehring NH. 2016. Interrogating the degradation pathways of unstable mRNAs with XRN1-resistant sequences. *Nat Commun* 7:13691. <https://doi.org/10.1038/ncomms13691>.
 85. Ho DW, Sze KM, Ng IO. 2015. Virus-clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* 6:20959–20963. <https://doi.org/10.18632/oncotarget.4187>.
 86. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gindera TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
 87. Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative genomic viewer (IGV): high performance genomic data visualization and exploration. *Brief Bioinform* 14:178–192. <https://doi.org/10.1093/bib/bbs017>.
 88. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
 89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
 90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
 91. Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169. <https://doi.org/10.1093/bioinformatics/btu638>.
 92. Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11:R106. <https://doi.org/10.1186/gb-2010-11-10-r106>.