JOURNAL OF VIROLOGY, Aug. 1984, p. 367–375
0022-538X/84/080367-09$02.00/0

Vol. 51, No. 2

Nucleotide Sequence of an Infectious Molecularly Cloned Genome of Ground Squirrel Hepatitis Virus

CHRISTOPH SEEGER,[1] DON GANEM,[1,2] AND HAROLD E. VARMUS[1]*

Department of Microbiology and Immunology[1] and Department of Medicine,[2] University of California, San Francisco, California 94143

Received 6 January 1984/Accepted 4 April 1984

We have determined the complete nucleotide sequence of an infectious cloned genome of ground squirrel hepatitis virus (GSHV), a nonpathogenic member of the hepadnavirus group. The genome is 3,311 base pairs long and contains the major open reading frames described for the related human and woodchuck hepatitis B viruses (HBV and WHV, respectively). These reading frames include genes for the major structural proteins (the surface and core antigens), unassigned open reading frames (*A* and *B*), the longer of which is presumed to encode the viral DNA polymerase, and an open reading frame preceding and continuous with the surface antigen gene. The arrangement of these open reading frames is similar to that encountered in the genomes of HBV and WHV: all of the reading frames are encoded on the same strand, they are positioned in the same fashion with respect to each other, and a large portion (at least 51%) of the genome can be translated in two reading frames. Comparisons of the predicted translational products of the three mammalian hepadnaviruses reveal 78% amino acid homology between the proteins of GSHV and WHV and 43% homology between those of GSHV and HBV. In addition, a perfect direct repeat of 10 to 11 base pairs, separated by ca. 46 to 223 base pairs, is present in the three mammalian viruses and in duck hepatitis B virus; the position of the repeats near the 5′ termini of the two strands of virion DNA suggests a role in viral replication.

Ground squirrel hepatitis virus (GSHV), a member of the hepadnavirus group, provides a useful experimental model for human hepatitis B virus (HBV) (17, 26). GSHV is similar in many respects to HBV and the other known mammalian hepadnavirus, woodchuck hepatitis virus (WHV) (28). The virus has a narrow host range and is highly hepatotropic (10, 16). Virus particles contain two major structural proteins, the surface and core antigens (sAg and cAg, respectively), both of which are cross-reactive with and biochemically related to the analogous proteins of the other mammalian hepadnaviruses (3, 5, 11, 17). Virion DNA is composed of two strands that overlap at their 5′ ends: a full-length, protein-linked strand (called the long or minus strand) of ca. 3.3 kilobases and a strand of variable length (called the short or plus strand) that can be extended at its 3′ end by a virion-associated DNA polymerase activity (9, 17, 24). The identified intracellular forms of viral DNA include both closed circular duplex DNA and heterogenous partially single-stranded species in which minus strands predominate over plus strands (33). These forms suggest that GSHV may follow the replication cycle proposed by Summers and Mason (27), with DNA synthesis proceeding from an RNA template. Despite these several similarities to other hepadnaviruses, GSHV is distinguished by its apparent lack of pathogenic effect. Persistent infection for up to 3 years is accompanied only by slight inflammatory changes in portal triads (10, 16; unpublished data); to date there are no reports of GSHV-associated hepatomas. HBV and WHV, in contrast, are agents of overt acute and chronic hepatitis and are probable factors in hepatic carcinogenesis (8, 26).

We have been studying GSHV as an experimentally accessible model for replication of hepadnaviruses and as a potentially useful reagent for identifying determinants of pathogenicity. We have recently cloned the genomes of two strains of GSHV distinguishable on the basis of restriction site polymorphisms (9). One of these cloned genomes, from GSHV strain 27, has proved to be infectious when injected directly into the livers of susceptible ground squirrels in the form of monomeric and oligomeric circles C. Seeger, D. Ganem, and H. E. Varmus, Proc. Natl. Acad. Sci. U.S.A., in press. We now report the complete nucleotide sequence of the infectious genome, as determined by the dideoxynucleotide method of Sanger and his colleagues (1, 22). The sequence reveals that the three mammalian hepadnaviruses contain the same number of coding domains arranged in a similar fashion, with all of the genes, some of them overlapping, encoded on the same strand of virion DNA. Furthermore, the deduced amino acid sequences are highly homologous to the corresponding products of HBV and WHV. A number of features of the sequences suggest signals that may be important in the replication and expression of the genome.

MATERIALS AND METHODS

DNA clones. GSHV DNA clone pBA131 was obtained by cloning an *Eco*RI digestion product of GSHV virion DNA, after repair of the gapped region, in the lambda bacteriophage vector gtWES.B, followed by subcloning the viral insert in pBR328 (9). The virion DNA was prepared from the serum of an animal infected with a virus stock arbitrarily named strain 27.

Nucleotide sequencing. The nucleotide sequencing reactions were done by the method of Sanger et al. (22) with modifications as described by Biggin et al. (1). For the construction of Bal 31 fragments, the following protocol was used. Replicative-form DNA from clone pRR12, consisting of an M13mp8 vector with a full-length insert of GSHV DNA from pBA131, was linearized at different sites within the GSHV sequence (*Xba*I, *Apa*I, *Hin*cII, and *Stu*I) and incubated with the exonuclease Bal 31 (0.1 U/μg of DNA; New England Biolabs) for various times (1 to 20 min). The reactions were stopped with a twofold molar excess of EDTA over Ca^{2+}, and the DNA was precipitated with ethanol. The Bal 31-treated molecules were digested with

* Corresponding author.

*Sma*I (which cuts within the mp8 polylinker sequence) and ligated under conditions favoring the formation of circles. The ligation mixture was then introduced into *Escherichia coli* by $CaCl_2$-mediated DNA transformation, and the resulting M13 phage was plaque purified. Deletion mutant DNAs were used in the dideoxy sequencing reactions as previously described (22).

RESULTS AND DISCUSSION

Generation of the substrates for sequencing. DNA from GSHV clone pBA131, derived from virus strain 27, was subcloned into the M13mp8 vector as (i) a full-length (3.3-kilobase) *Eco*RI fragment, (ii) a series of subgenomic fragments generated by cleavage with endonucleases known to recognize one or two sites in GSHV DNA, and (iii) relatively small fragments produced by complete digestion with either *Hae*III or *Sau*3A, each an enzyme that recognizes several hitherto unmapped sites in GSHV DNA. The regions sequenced from the useful subclones are diagrammed in Fig. 1. To obtain genomic segments that were underrepresented in the initial collection of subclones, we constructed deletion mutants that brought the region to be sequenced close to the cloning site in M13mp8. To do this, a recombinant (pRR12) that carries the complete GSHV genome in the *Eco*RI site of M13mp8 was opened at different restriction sites within the GSHV region, treated for various periods of time with the exonuclease Bal 31, cleaved with an enzyme specific for the M13mp8 polylinker, religated, and recloned (for details see above). This procedure produced a series of deletion mutants with deletion endpoints at different distances from the site of primary cleavage (Fig. 1). Over 70% of the genome was sequenced with strands of both polarities, and the rest was sequenced with at least two independent subclones.

Arrangement of the coding domains of GSHV DNA and similarities to other genomes. The complete nucleotide sequence of GSHV DNA, 3,311 base pairs (bp) long, is presented in Fig. 2. Several factors suggested that it would be possible to recognize the coding domains of GSHV by comparison of the GSHV DNA sequences with the published sequences of HBV and WHV DNAs (6, 7, 20, 30): (i) there is cross-hybridization among these genomes (9, 24), and there are immunological cross-reactivity and common peptides among their major antigens (sAg and cAg) (5, 11, 17); and (ii) previous DNA hybridization studies between GSHV and HBV indicated that the coding domains appear to be similarly disposed on their genomes (9, 24).

When all of the possible initiation and termination codons in GSHV DNA are displayed in the six potential reading frames (Fig. 3A), four major translatable domains are evident; each of these is more than 300 bp long and contains an initiation codon within 120 bp of the 5′ end of the open reading frame. As anticipated, the identity of these four regions is apparent from comparisons of GSHV nucleotide sequences with those of WHV and HBV; the positions and identities of the open reading frames on the circular map of GSHV DNA are diagrammed in Fig. 3B.

It is apparent that all of the four extended reading frames are present on the same strand. The translatable strand, the sequence of which is presented in Fig. 2, has the same chemical polarity as the strand of virion DNA; thus, the complementary template for GSHV mRNA is the minus strand. As in the genomes of HBV and WHV, the longest open reading frame (called gene *A*, since its product has not been identified) overlaps all or part of the other coding domains, each in a different frame from the *A* gene. As a consequence, at least 51% of the genome can be translated from more than one reading frame. The shortest open frame, called gene *B*, does not correspond to any known viral protein but is homologous to similarly positioned domains in the genomes of other hepadnaviruses (6, 7, 15, 20, 30). The sAg gene is preceded by an extended open frame called pre-S to conform with the nomenclature for other hepadnaviruses.

We have chosen to number our sequence beginning with the first nucleotide of the cAg gene, in the manner used for HBV by Valenzuela et al. (30), since the cAg gene is the only coding region preceded by a sequence (11 bp long) that does not appear to be translated. The sequence therefore ends, at position 3311, at the end of this short noncoding region. The relationship of this silent region to discontinuities in virion DNA will be discussed below.

Overall, the determined nucleotide sequence of GSHV DNA and the deduced amino acid sequences for the products of the major open reading frames display a striking degree of homology with the corresponding sequences of WHV (6): 82% of the nucleotides and 78% of the amino acids are identical (Table 1). GSHV appears to be less closely related to HBV (7, 20, 30), with 55% nucleotide and 46% amino acid identity. Only scattered homology is evident between GSHV DNA and protein and the recently determined sequences of duck hepatitis B virus (DHBV), and the homology pattern of GSHV and DHBV is not significantly different from that of WHV and DHBV (15). Moreover, the genetic organization of DHBV is slightly different: the reading frames of cAg and gene *B* are fused in DHBV DNA, and gene *A* overlaps only the 5′ end of the fused *B*-cAg domain (15).

The deduced amino acid sequences presented in Fig. 2 do not take into account the possibilities (i) that RNA splicing might occur to remove sequences within open reading frames or to join sequences from different frames, (ii) that initiation of protein synthesis might occur at sites other than the first AUG in each region, and (iii) that proteins may be proteolytically processed during or after synthesis. Howev-

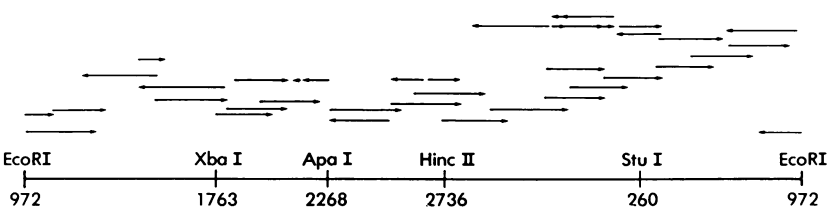

FIG. 1. Strategy for nucleotide sequencing. The sequenced regions of GSHV DNA are shown in relation to strategically important restriction sites, with the genome opened at the single *Eco*RI site at position 972. The illustrated restriction sites were used before Bal 31 treatment of pRR12 as described in the text. The arrows denote the direction of synthesis in the M13 sequencing reactions.

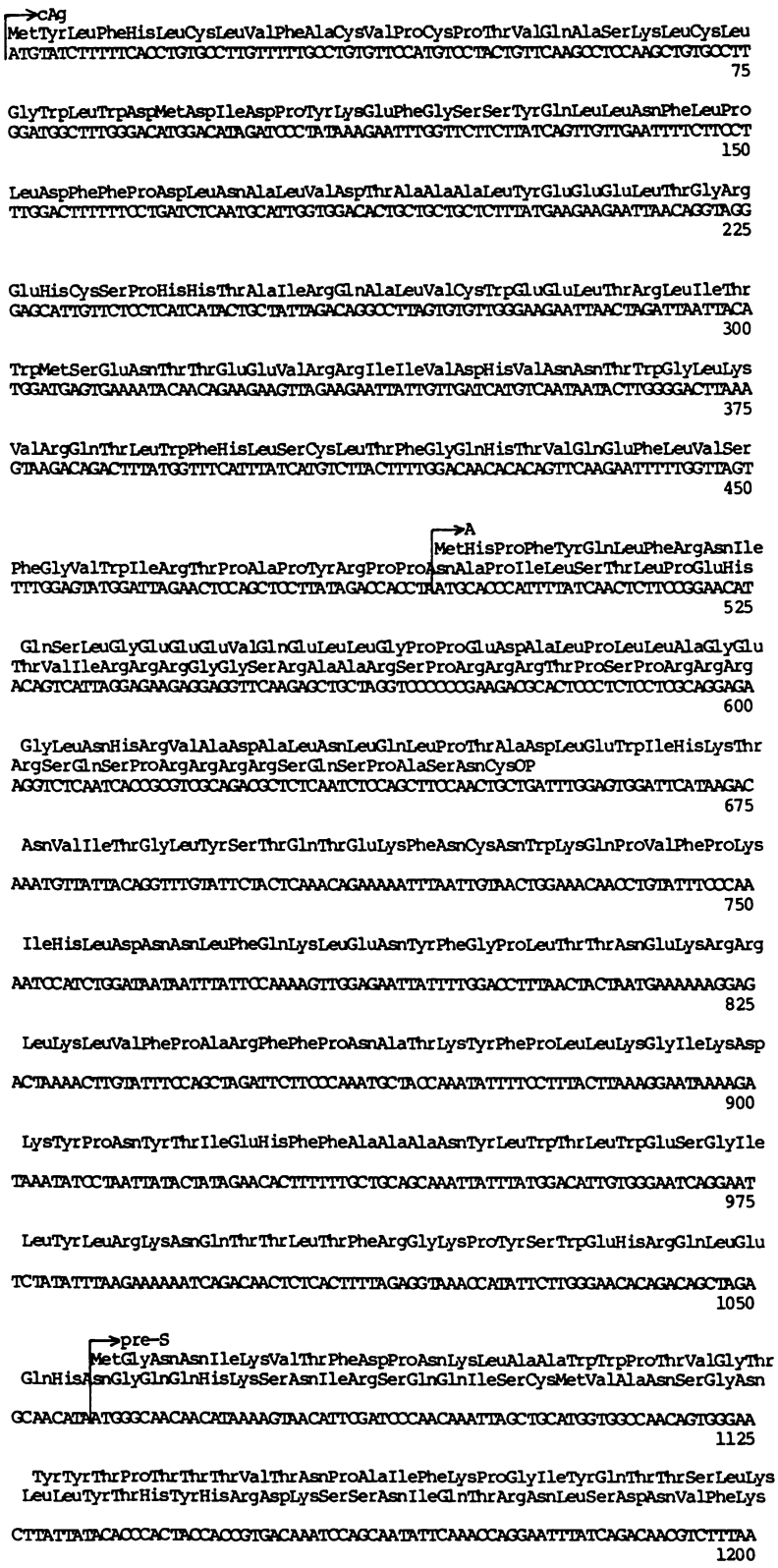

FIG. 2A.

FIG. 2. Nucleotide sequence of GSHV DNA. The sequence is arranged to begin with the first ATG of the open reading frame for cAg; it ends at nucleotide 3311 in the noncoding region. The four major open reading frames on the plus strand are bracketed, and the deduced amino acid sequences are presented at their corresponding positions.


AsnProLysAsnGlnGlnGluLeuAspAlaIleLeuMetThrArgTyrLysGluIleAspTrpAspAsnTrpGln
LysSerLysGluSerThrArgValArgCysTyrThrTyrAspLysIleGlnArgAsnArgLeuGlyGlnLeuAla
AAAATCCAAAGAATCAACAAGAGTTAGATGCTATACTTATGACAAGATACAAAGAAATAGATTGGGACAATTGGC
1275
GlyPheProValAsnGlnArgLeuProValSerAsnAsnAsnProProSerGlyGlnArgAlaGluThrPheGlu
ArgIleProCysGluSerLysAlaProSerGluGlnGlnSerSerLeuArgSerLysGlyArgAspPheArg
AAGGATTCCCTGTGAATCAAAGGCTCCCAGTGAGCAACAACAATCCTCCCTCAGGTCAAAGGGCAGAGACTTTCG
1350
IleLysSerArgProIleIleValProGlyIleArgAspIleProArgGlyIleValProProGlnThrProSer
AsnGlnIleGlnAlaTyrAsnSerSerArgAsnLysGlyTyrThrThrTrpHisSerThrThrSerAspSerIle
AAATCAAATCCAGGCCTATAATAGTTCCAGGAATAAGGGATATACCACGTGGCATAGTACCACCTCAGACTCCAT
1425
AsnArgAspGlnArgArgLysProThrProLeuThrProProLeuArgAspThrHisProHisLeuThrMetLys
GlnSerGlySerLysLysLysThrHisThrSerAsnSerSerPheGluArgHisThrProSerPheAspAsnGlu
CCAATCGGGATCAAAGAAGAAAACCCACACCTCTAACTCCTCCTTTGAGAGACACACACCCTCATTTGACAATGA
1500
AsnGlnThrGlyHisLeuGlnGlyPheAlaGluGlyLeuArgAlaLeuThrThrSerAspHisHisAsnSerAla
LysSerAspArgSerProAlaGlyIleCysArgGlyThrGluSerSerAsnHisLeuArgSerSerGlnLeuCys
AAAATCAGACAGGTCACCTGCAGGGATTTGCAGAGGGACTGAGAGCTCTAACCACCTCAGATCATCACAACTCTG
1575
TyrGlyAspProPheThrThrLeuSerProValValProThrValSerThrThrLeuSerProProLeuThrIle
LeuTrpArgSerPheTyrTyrThrLysProCysGlyThrTyrCysLeuHisHisIleValSerSerIleAspAsp
CCTATGGAGATCCTTTTACTACACTAAGCCCTGTGGTACCTACTGTCTCCACCACATTGTCTCCTCCATTGACGA
1650
sAg
GlyAspProValLeuSerThrGluMetSerProSerGlyLeuLeuGlyLeuLeuAlaGlyLeuGlnValValTyr
TrpGlyProCysThrPheAspGlyAspValThrIleArgSerProArgThrProArgArgIleThrGlyGlyIle
TTGGGGACCCTGTACTTTCGACGGAGATGTCACCATCAGGTCTCCTAGGACTCCTCGCAGGATTACAGGTGGTAT
1725
PheLeuTrpThrLysIleLeuThrIleAlaGlnSerLeuAspTrpTrpTrpThrSerLeuSerPheProGlyGly
PheLeuValAspLysAsnProTyrAsnSerSerGluSerArgLeuValValAspPheSerGlnPheSerArgGly
ATTTCTTGTGGACAAAAATCCTTACAATAGCTCAGAGTCTAGACTGGTGGTGGACTTCTCTCAGTTTTCCAGGGG
1800
IleProGluCysThrGlyGlnAsnLeuGlnPheGlnThrCysLysHisLeuProThrSerCysProProThrCys
HisSerArgValHisTrpProLysPheAlaValProAsnLeuGlnThrLeuAlaAsnLeuLeuSerThrAsnLeu
GCATTCCCGAGTGCACTGGCCAAAATTTGCAGTTCCAAACCTGCAAACACTTGCCAACCTCTTGTCCACCAACCT
1875
AsnGlyPheArgTrpMetTyrLeuArgArgPheIleIleTyrLeuLeuValLeuLeuLeuPheLeuThrPheLeu
GlnTrpLeuSerLeuAspValSerAlaAlaPheTyrHisIleProValSerProAlaAlaValProHisPheLeu
GCAATGGCTTTCGCTGGATGTATCTGCGGCGTTTTATCATATACCTGTTAGTCCTGCTGCTGTTCCTCACTTTCT
1950
LeuValLeuLeuAspTrpLysGlyLeuLeuProValCysProMetMetProAlaThrGluThrThrValAsnCys
ValGlySerProGlyLeuGluArgPheAlaSerCysMetSerHisAspAlaSerAsnArgAsnAsnSerLysLeu
TGTTGGTTCTCCTGGATTGGAAAGGTTTGCTTCCTGTATGTCCCATGATGCCAGCAACAGAAACAACAGTAAATT
2025
ArgGlnCysThrIleSerAlaGlnAspThrPheThrThrProTyrCysCysCysLeuLysProThrAlaGlyAsn
GlnThrMetHisHisIleCysSerArgHisLeuTyrAsnThrLeuLeuLeuLeuPheLysThrTyrGlyArgLys
GCAGACAATGCACCATATCTGCTCAAGACACCTTTACAACACCTTACTGCTGTTGTTTAAAACCTACGGCAGGAA
2100
CysThrCysTrpProIleProSerSerTrpAlaLeuGlySerTyrLeuTrpGluTrpAlaLeuAlaArgPheSer
LeuHisLeuLeuAlaHisProPheIleMetGlyPheArgLysLeuProMetGlyValGlyLeuSerProPheLeu
ATTGCACTTGTTGGCCCATCCCTTCATCATGGGCTTTAGGAAGCTACCTATGGGAGTGGGCCTTAGCCCGTTTCT
2175

FIG. 2B.

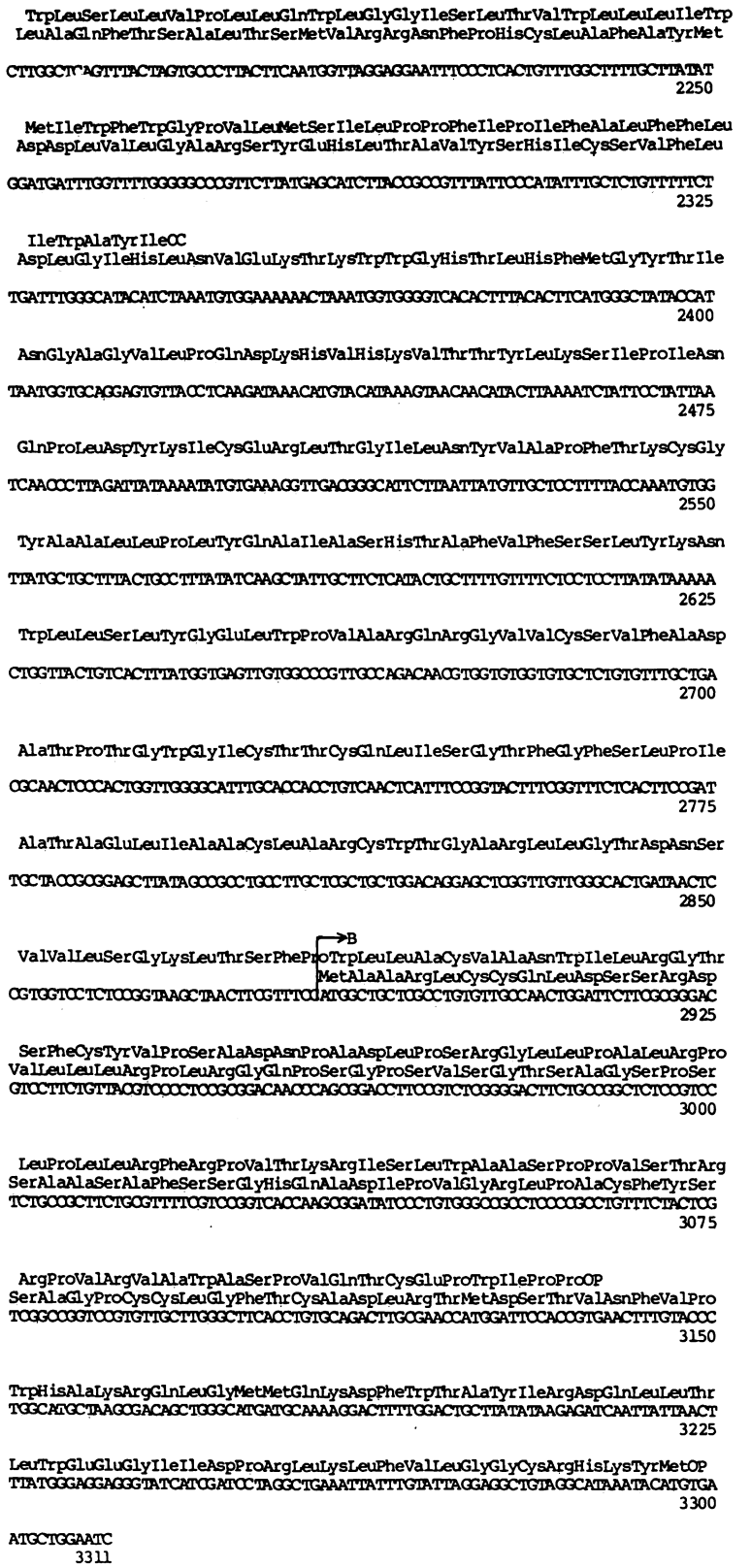
TrpLeuSerLeuLeuValProLeuLeuGlnTrpLeuGlyGlyIleSerLeuThrValTrpLeuLeuLeuIleTrp
LeuAlaGlnPheThrSerAlaLeuThrSerMetValArgArgAsnPheProHisCysLeuAlaPheAlaTyrMet
2250
MetIleTrpPheTrpGlyProValLeuMetSerIleLeuProProPheIleProIlePheAlaLeuPhePheLeu
AspAspLeuValLeuGlyAlaArgSerTyrGluHisLeuThrAlaValTyrSerHisIleCysSerValPheLeu
2325
IleTrpAlaTyrIleOC
AspLeuGlyIleHisLeuAsnValGluLysThrLysTrpTrpGlyHisThrLeuHisPheMetGlyTyrThrIle
2400
AsnGlyAlaGlyValLeuProGlnAspLysHisValHisLysValThrThrTyrLeuLysSerIleProIleAsn
2475
GlnProLeuAspTyrLysIleCysGluArgLeuThrGlyIleLeuAsnTyrValAlaProPheThrLysCysGly
2550
TyrAlaAlaLeuLeuProLeuTyrGlnAlaIleAlaSerHisThrAlaPheValPheSerSerLeuTyrLysAsn
2625
TrpLeuLeuSerLeuTyrGlyGluLeuTrpProValAlaArgGlnArgGlyValValCysSerValPheAlaAsp
2700
AlaThrProThrGlyTrpGlyIleCysThrThrCysGlnLeuIleSerGlyThrPheGlyPheSerLeuProIle
2775
AlaThrAlaGluLeuIleAlaAlaCysLeuAlaArgCysTrpThrGlyAlaArgLeuLeuGlyThrAspAsnSer
2850
B
ValValLeuSerGlyLysLeuThrSerPheProTrpLeuLeuAlaCysValAlaAsnTrpIleLeuArgGlyThr
MetAlaAlaArgLeuCysCysGlnLeuAspSerSerArgAsp
2925
SerPheCysTyrValProSerAlaAspAsnProAlaAspLeuProSerArgGlyLeuLeuProAlaLeuArgPro
ValLeuLeuLeuArgProLeuArgGlyGlnProSerGlyProSerValSerGlyThrSerAlaGlySerProSer
3000
LeuProLeuLeuArgPheArgProValThrLysArgIleSerLeuTrpAlaAlaSerProProValSerThrArg
SerAlaAlaSerAlaPheSerSerGlyHisGlnAlaAspIleProValGlyArgLeuProAlaCysPheTyrSer
3075
ArgProValArgValAlaTrpAlaSerProValGlnThrCysGluProTrpIleProProOP
SerAlaGlyProCysCysLeuGlyPheThrCysAlaAspLeuArgThrMetAspSerThrValAsnPheValPro
3150
TrpHisAlaLysArgGlnLeuGlyMetMetGlnLysAspPheTrpThrAlaTyrIleArgAspGlnLeuLeuThr
3225
LeuTrpGluGluGlyIleIleAspProArgLeuLysLeuPheValLeuGlyGlyCysArgHisLysTyrMetOP
3300
ATGCTGGAATC
3311

FIG. 2C.

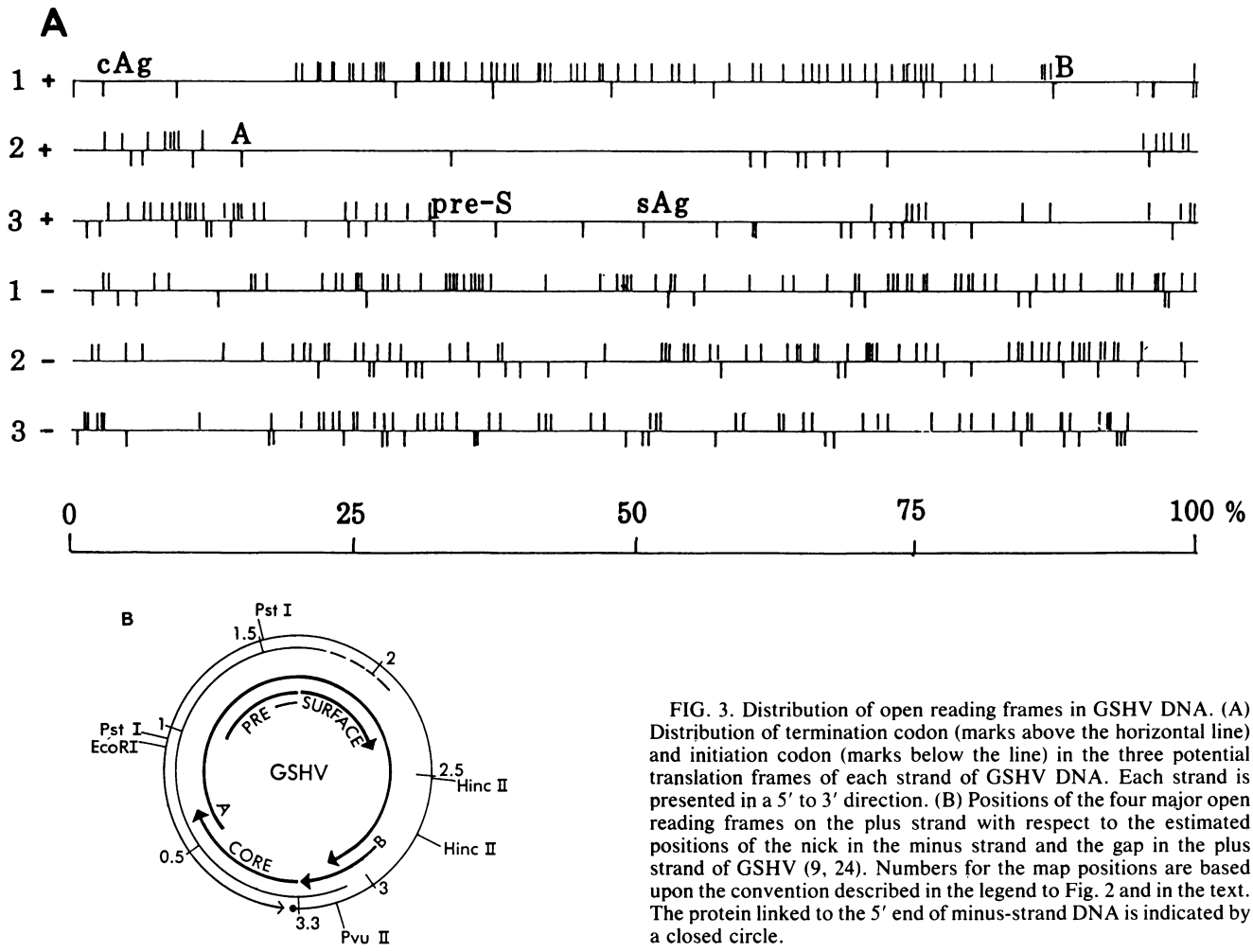

FIG. 3. Distribution of open reading frames in GSHV DNA. (A) Distribution of termination codon (marks above the horizontal line) and initiation codon (marks below the line) in the three potential translation frames of each strand of GSHV DNA. Each strand is presented in a 5′ to 3′ direction. (B) Positions of the four major open reading frames on the plus strand with respect to the estimated positions of the nick in the minus strand and the gap in the plus strand of GSHV (9, 24). Numbers for the map positions are based upon the convention described in the legend to Fig. 2 and in the text. The protein linked to the 5′ end of minus-strand DNA is indicated by a closed circle.

er, our knowledge of the transcriptional and translational programs of hepadnaviruses is primitive, and none of these possibilities has been excluded. Evidence for initiation of protein synthesis within the pre-S and sAg regions, for example, will be considered below.

sAg gene and pre-S region. We have assigned the beginning of the region coding for the structural ground squirrel hepatitis sAg (GSHsAg) protein to nucleotide 1677, although the amino terminus of GSHsAg has not been directly determined. This assignment would produce a protein of 222 amino acids, consistent with the apparent molecular weight of the major component of GSHsAg isolated from the sera of infected animals (23,000) (11). The predicted GSHsAg is highly homologous to woodchuck hepatitis sAg (WHsAg), with 90% amino acid identity. GSHsAg appears to be less closely related to hepatitis B sAg (HBsAg), with 61% amino acid identity; however, comparison with the directly determined amino terminus of HBsAg supports the assignment of the amino terminus of GSHsAg (21). A major portion of mature GSHsAg, like WHsAg and HBsAg, is believed to be glycosylated (11). Only one consensus site for glycosylation (Asn-Cys-Thr) is found in the predicted protein sequence for mature GSHsAg at position 2100 (Fig. 2). This sequence is conserved in WHsAg and HBsAg, though HBsAg contains two other potential glycosylation sites.

Although the coding domain for mature GSHsAg is likely to commence with nucleotide 1677, there are several reasons to suppose that translation of the sAg domain may be initiated, at least in part, at other sites. (i) The sAg region is preceded by the pre-S region, an open reading frame of 206 codons that includes initiation codons at positions 1059, 1239, and 1497. (ii) Pre-S domains of similar length are found in all four hepatitis B virus genomes, although the pre-S region is less highly conserved than most others (Table 1).

TABLE 1. Amino acid homologies among mammalian hepadnavirus proteins

Gene	Length (amino acids)	% Total coding capacity	% Amino acid identities between[a]:	
			GSHV and WHV	GSHV and HBV
cAg	218	13	92	68
A	881	53	76	45
Pre-S	206	12	71	17
sAg	222	13	90	61
B	138	8	71	33
Total	1,665	100	78	46

[a] The amino acid homology between the deduced protein sequences of GSHV and WHV (6) and GSHV and HBV (30) is summarized for each coding domain.

(The pre-S protein sequences of GSHV and HBV show only 17% identity, and they differ in length by 44 amino acids as a result of internal deletions or additions.) (iii) Proteins larger than mature sAg, with shared peptides and immunological reactivity, have been found in hepadnavirus particles (4, 14, 23, 25). The most prominent of these proteins has a molecular weight of ca. 31,000 (25) and has been proposed as a contributor to host cell tropism by its binding to polyalbumin (14). A protein of ca. 30,000 molecular weight could be initiated from the start codon positioned 60 codons from the 5′ end of the GSHsAg region, at nucleotide 1497. Moreover, synthesis of the major mRNA for HBsAg commences within the pre-S region near the final initiation codon in the pre-S region, the codon likely to be used to initiate the 31,000-molecular-weight protein (2; 24a).

cAg gene. The open reading frame for ground squirrel hepatitis cAg (GSHcAg) was identified on the basis of the homology between its deduced amino acid sequences and the sequences of hepatitis B cAg (HBcAg) and woodchuck hepatitis cAg (WHcAg). The gene for HBcAg was unambiguously identified by production of antigen in *E. coli* transformed by portions of HBV DNA (20). The cAg gene encodes the most highly conserved protein sequence (Table 1). However, the amino terminus of this protein cannot be unambiguously assigned since two initiation codons (at positions 1 and 91) are present near the 5′ end of the open reading frame; we have arbitrarily assumed that the first of these is used.

The predicted molecular weight of cAg is 24,000 (or 21,000 if the second ATG is the initiation site); both values are compatible with the molecular weight of 20,500 estimated from polyacrylamide electrophoresis (5). The deduced sequence of cAg is rich in basic amino acids at the carboxyl terminus (16 of 40), as is the case for WHcAg and HBcAg. The structure of cAg mRNA and the mechanism by which the expression of cAg is regulated are not understood. It is notable that the initiation site for cAg synthesis is preceded by the only nontranslated portion of the genomes of mammalian hepadnaviruses and that the cAg gene would be at the 5′ end of the linear molecule formed by denaturation of the cohesive 5′ termini in virion DNA. However, the significance of these features is uncertain, particularly since the predicted linear species has not been observed; moreover, the most likely template for RNA synthesis in infected cells appears to be covalently closed circular DNA (31).

Gene *B*. All three of the mammalian hepadnaviruses retain an open reading frame ca. 500 nucleotides long preceding the cAg genes. This region, arbitrarily called gene *B* (or *X*), begins with the ATG at position 2884 in the GSHV genome and ends at position 3300, 11 nucleotides before the start of the cAg gene. In the genome of DHBV, the cAg gene is continuous with the open frame for gene *B*, suggesting that the *B* gene product and cAg might be synthesized as a fusion protein (15). A polypeptide 138 amino acids long is deduced from the *B* domain of GSHV; the sequence of the putative product is somewhat less conserved than those of other viral proteins, with 71% identity between GSHV and WHV and 33% identity between GSHV and HBV B proteins. One possible function for the *B* gene might be to encode the protein found linked to mature (9, 12, 18) and nascent (18) minus-strand DNA; this protein may serve as a primer for minus-strand DNA synthesis (18, 27). However, such notions are at present purely speculative.

Gene *A*. The largest open reading frame in the genomes of all the hepadnaviruses, including GSHV, is also unassigned, but it seems likely to encode a protein containing the DNA polymerase activity described in virions (9, 13, 24) and in cytoplasmic cores (27). In the GSHV genome, this long reading frame, called gene *A*, begins at position 494, in the middle of the cAg gene, extends through all of the pre-S region and the sAg gene, and ends in the middle of gene *B* at position 3139. In all of these overlapping regions, gene *A* is read in a different frame than is the coincident coding domain. As a result, 51% of the genome is translatable in two reading frames, with gene *A* itself covering 80% of the genome and competent to encode a protein of 881 amino acids (molecular weight, 96,900).

We have examined the amino acid sequence similarities among the predicted gene *A* products of the three mammalian hepadnaviruses, expecting to find that overlapping reading frames promote conservation of sequence (Table 2). However, the regions of gene *A* interposed between two alternative coding regions were no less conserved than were the overlapping regions (e.g., the least identity was found in the region of gene *A* that overlaps the pre-S region). Furthermore, the components of the gene *A* product were not more conserved than were the alternative products, except in the region of overlap between gene *A* and gene *B*. Nor was gene *A* better conserved overall than were the others. For instance, the cAg gene was more highly conserved than was gene *A*, as judged, in particular, by a comparison between GSHV and HBV (Table 1). The degree of divergence of gene *A* may be influenced by the functional domains of its product, especially if the gene proves to encode a DNA polymerase responsive to an RNA or DNA template. Toh et al. have identified a region of gene *A*, overlapping the sAg genes of HBV and WHV, that encodes a polypeptide region homologous to portions of the reverse transcriptase of murine and avian retroviruses and to a portion of a putative RNA-directed DNA polymerase of cauliflower mosaic virus (29). This domain is somewhat more conserved than are the other domains of the gene *A* product when the GSHV protein sequence is compared with that of HBV (Table 2). However, the GSHV and WHV proteins are generally so similar, except in the region of gene *A* that overlaps with pre-S, that no significant statements about functional potential can be made.

At present, nothing is known about the mechanism of expression of genes *A* and *B*.

Short direct repeats in the vicinity of the 5′ ends of virion DNA. The life cycle of hepatitis B viruses is demonstrably complex—the genome passes through several forms of DNA and RNA, and overlapping coding domains are apparently expressed with differing efficiency—yet there is little infor-

TABLE 2. Amino acid homologies between gene *A* segments of mammalian hepadnaviruses

Gene *A* overlapping with:	Length (amino acids)	% Total coding capacity	% Amino acid identities between[a]:	
			GSHV and WHV	GSHV and HBV
cAg	53	6	92	36
cAg–pre-S interval	136	15	75	43
Pre-S	205	23	45	11
S	222	25	87	67
S-*B* interval	180	20	91	54
B	85	10	92	58
Total	881	100	76	45

[a] See Table 1, footnote *a*.

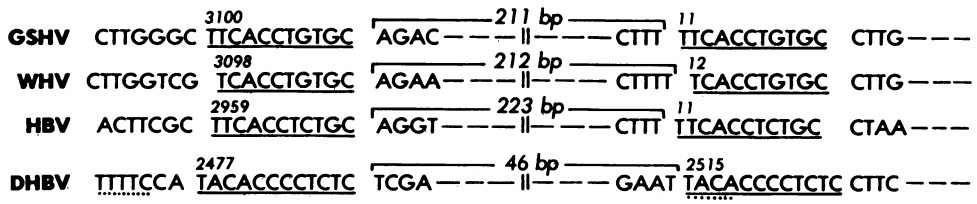

FIG. 4. Direct repeats in hepatitis B virus genomes. The 10- or 11-nucleotide direct repeats in homologous regions of the genomes of GSHV, WHV, and HBV are shown (underlined) with some adjacent sequences and the distances between the repeats. Also shown are related direct repeats from the genome of DHBV DNA (19); these have been shown to reside in the vicinity of the 5′ ends of plus and minus strands (dotted underlinings on the left and right sides, respectively). In all cases, the plus-strand sequence is shown (5′ to 3′ from left to right), and the numbering is based upon the assignment of position 1 to the first nucleotide of the first initiation codon in the cAg open reading frame of the mammalian viruses and in DHBV as described by Mandart et al. (15).

mation about the viral signals that regulate the cycle. With the help of a computer program, we have searched the sequence of GSHV DNA for dyad symmetries and direct repeats. Of the several short repeats observed, one set of 11 bp, beginning at positions 10 and 3099 (Fig. 4), is particularly intriguing for several reasons. (i) A similar set of 10- or 11-bp repeats is present in the corresponding regions of WHV and HBV DNAs (Fig. 4). (ii) The repeated sequences are separated by a distance of 211 to 223 bp, the approximate length of the cohesive 5′ ends of plus and minus strands in virion DNA, and the locations of the sequences are near the predicted ends of the two strands. (iii) In the DHBV genome, a similar 12-bp direct repeat is separated by only 46 bp, again the approximate length of the cohesive ends of virion DNA, and the 5′ termini of plus and minus strands have been mapped within or adjacent to the direct repeats (Fig. 4) (19).

One striking property of the direct repeats that emerges from the comparison in Fig. 4 is the retention of identity between the two sequences in each genome despite base differences between the repeat sequences in different genomes. This phenomenon suggests either that some crucial event requires exact identity or that the direct repeats are regenerated during each replication cycle from a single copy of the repeat. The latter explanation applies, for example, to the regeneration of exact long terminal repeats in retrovirus DNA from unique sequences in viral RNA during each round of replication (32).

Conclusions. The nucleotide sequences of the genomes of all four of the known hepatitis B viruses have now been determined (6, 7, 15, 20, 30). The results reveal a class of viruses with similar strategies for maximizing the coding potential of their small genomes and with an apparent requirement for elaborate mechanisms of gene expression. Despite the obvious medical importance of the hepadnaviruses, there is little understanding of these mechanisms; furthermore, only some of the protein products of the conserved coding domains have been identified.

A major finding in our study is the high degree of homology between the genomes and protein products of GSHV and WHV. Nevertheless, the two viruses differ markedly with respect to pathogenicity, whereas the two pathogenic viruses, WHV and HBV, are less closely related than are WHV and GSHV. Furthermore, all regions of amino acid homology between WHV and HBV are also shared with GSHV. Thus, subtle variations between WHV and GSHV genes or different host responses to infection seem likely to account for the apparent differences in virulence. We have recently demonstrated infection of ground squirrels with GSHV DNA from this sequenced clone (Seeger et al., in press). Thus, it may be possible to elucidate possible genetic differences between GSHV and WHV by the construction of recombinant clones and the subsequent infection of either ground squirrels or woodchucks with the recombinant molecules. In addition, it should be possible to assess the functional significance of the specific regions of the viral genome, including the unassigned reading frames *A* and *B* and the direct repeats, by examining the infectivity of viral DNA after site-directed mutagenesis in vitro.

ACKNOWLEDGMENTS

We thank K.-H. Klempnauer and D. A. Peattie for assisting in the initial nucleotide sequencing, Hugo Martinez for providing the computer programs, and Janine Marinos for excellent typing of the manuscript.

This work was supported by grants from the National Institutes of Health and from the American Cancer Society. C.S. was supported by a grant from the Swiss National Science Foundation. H.E.V. is an American Cancer Society Research Professor.

LITERATURE CITED

1. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. U.S.A. **80**:3963–3965.
2. **Cattaneo, R., H. Will, N. Hernandez, and H. Schaller.** 1983. Signals regulating hepatitis B surface antigen transcription. Nature (London) **305**:336–338.
3. **Cote, P. J., Jr., and J. L. Gerin.** 1983. Nonoverlapping antigenic sites of woodchuck hepatitis virus surface antigen and their cross-reactivity with ground squirrel hepatitis virus and hepatitis B virus surface antigens. J. Virol. **47**:15–23.
4. **Feitelson, M., P. L. Marion, and W. S. Robinson.** 1983. The nature of peptides larger in size than the major surface antigen components of hepatitis B and like viruses in ground squirrels, woodchucks and ducks. Virology **130**:76–90.
5. **Feitelson, M. A., P. L. Marion, and W. S. Robinson.** 1982. Core particles of hepatitis B virus and ground squirrel hepatitis virus. I. Relationship between hepatitis B core antigen- and ground squirrel hepatitis core antigen-associated polypeptides by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and tryptic peptide mapping. J. Virol. **43**:687–696.
6. **Galibert, F., T. N. Chen, and E. Mandart.** 1982. Nucleotide sequence of a cloned woodchuck hepatitis virus genome: comparison with the hepatitis B virus sequence. J. Virol. **41**:51–65.
7. **Galibert, F., E. Mandart, E. Fitoussi, P. Tiollais, and P. Charnay.** 1979. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature (London) **281**:646–650.
8. **Ganem, D.** 1982. Persistent infection of humans with hepatitis B virus: mechanisms and consequences. Rev. Infect. Dis. **4**:1026–1047.
9. **Ganem, D., L. Greenbaum, and H. E. Varmus.** 1982. Virion DNA of ground squirrel hepatitis virus: structural analysis and molecular cloning. J. Virol. **44**:374–383.
10. **Ganem, D., B. Weiser, A. Barchuk, R. J. Brown, and H. E. Varmus.** 1982. Biological characterization of acute infection with ground squirrel hepatitis virus. J. Virol. **44**:366–373.
11. **Gerlich, W. H., M. A. Feitelson, P. L. Marion, and W. S.**

Robinson. 1980. Structural relationships between the surface antigens of ground squirrel hepatitis virus and human hepatitis B virus. J. Virol. **36**:787–795.

12. **Gerlich, W. H., and W. S. Robinson.** 1980. Hepatitis B virus contains protein attached to the 5′ terminus of its complete DNA strand. Cell **21**:801–810.
13. **Kaplan, P. M., R. L. Greenman, J. L. Gerin, R. H. Purcell, and W. S. Robinson.** 1973. DNA polymerase associated with human hepatitis B antigen. J. Virol. **12**:995–1005.
14. **Machida, A., S. Kishimoto, H. Ohnuma, H. Miyamoto, K. Baba, K. Oda, T. Nakamura, Y. Miyakawa, and M. Mayumi.** 1983. A hepatitis B surface antigen polypeptide (P31) with the receptor for polymerized human as well as chimpanzee albumins. Gastroenterology **85**:268–274.
15. **Mandart, E., A. Kay, and F. Galibert.** 1984. Nucleotide sequence of a cloned duck hepatitis B virus genome: comparison with woodchuck and human hepatitis B virus sequences. J. Virol. **49**:782–792.
16. **Marion, P. L., S. S. Knight, F. H. Salazar, H. Popper, and W. S. Robinson.** 1983. Ground squirrel hepatitis virus infection. Hepatology **3**:519–527.
17. **Marion, P. L., L. S. Oshiro, D. C. Regnery, G. H. Scullard, and W. S. Robinson.** 1980. A virus in Beechey ground squirrels that is related to hepatitis B virus of humans. Proc. Natl. Acad. Sci. U.S.A. **77**:2941–2945.
18. **Molnar-Kimber, K. L., J. Summers, J. M. Taylor, and W. S. Mason.** 1983. Protein covalently bound to minus-strand DNA intermediates of duck hepatitis B virus. J. Virol. **45**:165–172.
19. **Molnar-Kimber, K. L., J. W. Summers, and W. S. Mason.** 1984. Mapping of the cohesive overlap of duck hepatitis B virus DNA and of the site of initiation of reverse transcription. J. Virol. **51**:181–191.
20. **Pasek, M., T. Goto, W. Gilbert, B. Zink, H. Schaller, P. MacKay, G. Leadbetter, and K. Murray.** 1979. Hepatitis B virus genes and their expression in *E. coli*. Nature (London) **282**:575–579.
21. **Peterson, D. L.** 1981. Isolation and characterization of the major protein and glycoprotein of hepatitis B surface antigen. J. Biol. Chem. **256**:6975–6983.
22. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**:5463–5467.
23. **Shih, J. W.-K., and J. L. Gerin.** 1977. Proteins of hepatitis B surface antigen. J. Virol. **21**:347–357.
24. **Siddiqui, A., P. L. Marion, and W. S. Robinson.** 1981. Ground squirrel hepatitis virus DNA: molecular cloning and comparison with hepatitis B virus DNA. J. Virol. **38**:393–397.

24a. **Standring, D. N., W. J. Rutter, H. E. Varmus, and D. Ganem.** 1984. Transcription of the hepatitis B surface antigen gene in cultured murine cells initiates within the presurface region. J. Virol. **50**:563–571.

25. **Stibbe, W., and W. H. Gerlich.** 1983. Structural relationships between minor and major proteins of hepatitis B surface antigen. J. Virol. **46**:626–628.
26. **Summers, J.** 1981. Three recently described animal virus models for human hepatitis B virus. Hepatology **1**:179–183.
27. **Summers, J., and W. S. Mason.** 1982. Replication of the genome of a hepatitis B-like virus by reverse transcription of an RNA intermediate. Cell **29**:403–415.
28. **Summers, J., J. M. Smolec, and R. Snyder.** 1978. A virus similar to human hepatitis B virus associated with hepatitis and hepatoma in woodchucks. Proc. Natl. Acad. Sci. U.S.A. **75**:4533–4537.
29. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305**:829-831.
30. **Valenzuela, P., M. Quiroga, J. Zaldivar, P. Gray, and W. J. Rutter.** 1981. The nucleotide sequence of the hepatitis B viral genome and the identification of the major viral genes. p. 57–70. *In* B. Fields, R. Jaenisch, and C. F. Fox (ed.), Animal virus genetics. Academic Press, Inc., New York.
31. **Varmus, H.** 1983. Reverse transcription in plants? Nature (London) **304**:116–117.
32. **Varmus, H. E.** 1982. Form and function of retroviral proviruses. Science **216**:812–820.
33. **Weiser, B., D. Ganem, C. Seeger, and H. E. Varmus.** 1983. Closed circular viral DNA and asymmetrical heterogeneous forms in livers from animals infected with ground squirrel hepatitis virus. J. Virol. **48**:1–9.