SUPPLEMENTARY MATERIALS AND METHODS

Supplementary Table 1 Reference sequences used for genotyping

NCBI GenBank accession numbers of the 86 reference sequences used in HBV genotyping:

HBV genotype (N sequences)	GenBank accession numbers
A (11)	GQ331047, GQ331048, GQ477501, AF090839, AY233278, AB241115, AB194952, AB194951, AY934764, FJ692609, AM180623
B (17)	AB073835, AP011084, GQ924653, AB073858, D00329, AB287316, DQ463787, AP011093, AP011094, GQ358148, GQ358152, M54923, AP011085, EF473977, AP011091, AB219427, AP011086
C (22)	AP011099, AB241109, AB048704, AB540583, AB112066, AB031265, EU670263, AB554025, AB644287, AB644281, AB644286, X75665, AP011108, AP011107, AP011104, AB554019, AP011102, AP011103, X52939, AB644284, X75656, AB033553
D (13)	DQ315779, AB033558, AB048701, GQ922005, AJ132335, EU939680, AB493846, GU456647, AF280817, AB210820, Z35716, AM494716, FJ904405
E (5)	X75664, AM494694, DQ060828, FJ349237, FJ349226
F (8)	AY090459, HE981184, AB166850, DQ823090, AY090455, X69798, X75663, AB036911
G (5)	EF464098, GU565217, AB064311, HE981176, AP007264
H (5)	AY090454, AP007261, AB275308, AB516393, AY090460

Supplementary Protocol 1 Amplification of the preS1 region and ultra-deep pyrosequencing data treatment

HBV DNA was extracted from 200 µL of serum with the QIAamp DNA MiniKit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions. A 585-bp fragment was amplified by nested PCR in all HBV genotypes, except for genotypes D (552-bp) and E (582-bp). The amplification procedure was carried out as follows: the first PCR round was performed in a final volume of 10 µL, including 0.2 µM of the forward and reverse primers (preS_M13_Fw, ***GTTGTAAAACGACGGCCAGT***-GTCACCATATTCTTGGGAACAA-3' and preS_M13_Rv, 5'- ***CACAGGAAACAGCTATGACC***-GAACTGGAGCCACCAGCAG-3'), 1x Pfu Ultra II Polymerase Buffer, 0.25 mM of each dNTP, 0.2 µL of Pfu Ultra II DNA Polymerase (Stratagene, Agilent Technologies), and 3 µL of DNA. The primers included an M13 universal adaptor sequence in the first 20 nts of their 5' ends (in bold and italics) followed by template-specific sequences that hybridize between nt positions 2822 and 2843 (genotype A) 2816 and 2837 (genotypes B, C, D, E and H), and 2815 and 2836 (genotype F). After a single denaturation step of 2 min at 95ºC, samples underwent 30 cycles of 20 s at 95ºC, 20 s at 57ºC, and 15 s at 72ºC, and a single final 3 min step at 72ºC. The second PCR round was performed using a specific pair of primers for each sample: forward 5'-CGTATCGCCTCCCTCGCGCCATCAG-**MID**-***GTTGTAAAACGACGGCCAGT***-3' and reverse 5'-CTATGCGCCTTGCCAGCCCGCTCAG-**MID**-***CACAGGAAACAGCTATGACC***-3'. These primers contain the 5' sequences A and B, which are adaptors for the elements of the ultra-deep pyrosequencing (UDPS) system, followed by a unique identifier that enables grouping of the sequences derived from each sample (multiplex identifier sequences [MID], in bold), and the same M13 universal adaptor sequences as the primers used in the first PCR (in bold and italics). Briefly, PCR was performed using 1x Pfu Ultra II polymerase buffer, 0.25 mM of each dNTP, 0.2 µM of forward and reverse primers, and 1 µL of Pfu Ultra II DNA polymerase (Stratagene, Agilent Technologies) in a final volume of 50 µL. After a single denaturation step of 2

min at 95°C, samples underwent 20 cycles of 20 s at 95°C, 20 s at 60°C, and 15s at 72°C, and a single final 3-min step at 72°C.

The final PCR products (amplicons) were purified with Agencourt AMPure XP magnetic beads (Beckman Coulter, Beverly, USA). The quality of the purified products was verified with the Agilent 2200 TapeStation System using the D1000 ScreenTape kit (Agilent Technologies, Waldbronn, Germany), and these products were quantified using the Quant-iT Picogreen dsDNA Assay kit (Thermo Fisher Scientific – Life Technologies, Austin, USA). Before UDPS sequencing, each amplicon was diluted, separately, to $1x10^9$molecules/μL in 1xTE Buffer, and mixed equally (10 μL of each of these diluted amplicons) to prepare the amplicon pool, which was diluted to $1x10^7$ molecules/μL, and finally, to $2x10^6$ molecules/μL using 1xTE Buffer. Molecules in this working solution were denatured and mixed with the capture beads needed for emulsion PCR and UDPS at a ratio of 0.8 molecules per bead for both the forward and reverse single strains. UDPS was performed with the GS-Junior platform (454 Life sciences-Roche, Branford, USA), using titanium chemistry (GS-Junior Titanium Sequencing Kit).

After UDPS, the sequencing data underwent an in-house bioinformatics filtering procedure, based on scripts developed in R[1], which consisted of the following steps, as previously described[2]:

(1) Raw data: The fasta files provided by the 454 GS-Junior system were used as raw data.

(2) Demultiplexing: Identification of the MID sequence, M13 primer, and amplicon-specific primers was used to split the reads (equivalent to sequences) into different fasta files. One mismatch was tolerated in the MID identification and three mismatches in the specific primer identification.

(3) General quality filter: Reads that did not cover the full amplicon, and those showing more than 2 positions with an indeterminate nt or more than 3 gaps with respect the dominant haplotype were excluded. The remaining positions with indeterminate nts and gaps were repaired as per the dominant haplotype

(unique sequences covering the full amplicon observed on the clean set of sequences[3]) in each sample.

(4) Forward and reverse intersection: Only haplotypes common to the forward and reverse strands present in abundances of at least 0.1% were accepted. The final frequency of each haplotype was calculated as the sum of the reads (equivalent to sequences) observed in each strand.

(5) Final abundance filter: haplotypes with abundances below 0.25% were excluded.

Supplementary Protocol 2 Detection of the NTCP rs2296651 SNP (S267F)

Genomic DNA was isolated from 200 µL of serum using the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions.

Primers were designed to flank a 111 bp region of the SLC10A1 gene, located on chromosome 14 (GenBank accession number, NC_000014.9): rs2296651_F: 5′-CGGCACAGGTGCAGACG-3′ (nucleotides [nt] 69778537-69778521) and rs2296651_R: 5′-GAAAAGTGGTCCAATGACTTCAG-3′ (nt 69778427-69778449). The fluorescence resonance energy transfer (FRET) probes used, which hybridize very close together, were Anchor: 5′-ACATTTTGGCATCCAGTCTCCA-3′ (nt 69778467-69778485), labeled with fluorescein at the 3′ end, and Sensor: 5′-AGGATGGTGAAACAGAGTT-3′ (nt 69778488-69778502), labeled with red fluorophore LC640 at the 5′ end. This latter probe hybridizes over the mutation position and perfectly matches the mutated sequence (g.69778476G>A). A schematic representation of the oligonucleotide primers, FRET probes, and sequence used for detecting rs2296651 is shown in the Figure.

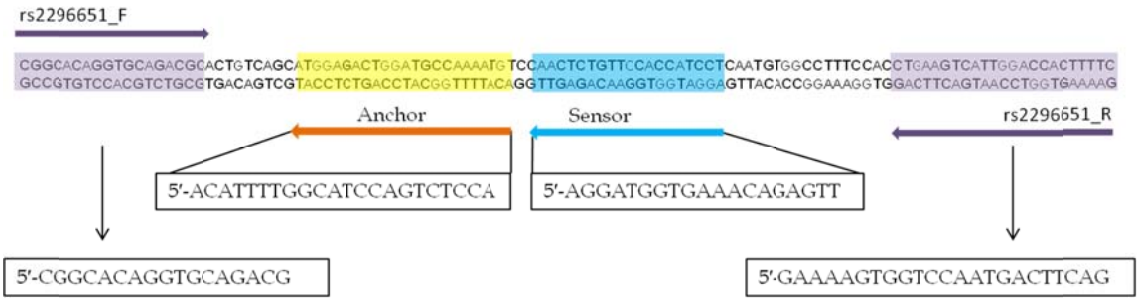


Figure Schematic representation of oligonucleotide primers, FRET probes, and the sequence amplified for detection of rs2296651 by real-time PCR.

The reaction was conducted at a final volume of 20 µL, containing 2 mM $MgCl_2$, 0.5 µM of each primer, 0.2 µM of each FRET probe, 2 µL of LightCycler Fast Start DNA Master Hybridization probe mix (Roche Diagnostics, Mannheim, Germany), and 5 µL of genomic DNA. The PCR program was run as follows: an initial denaturation step of 95°C for 10 min, and thereafter, 45 cycles including denaturation at 95 °C for 5 s, annealing at 50 °C for 10 s, and extension at 72 °C

for 15 s. After amplification, melting curves were generated by denaturation at 95°C for 20 s, maintaining the samples at 40°C for 20 s and slowly heating them to 85°C. Fluorescent measurements were performed at a wavelength of 640 nm and recorded during each annealing step and continuously during heating in the melting step. LightCycler software converted melting curves into melting peaks. To recognize the presence or absence of the rs2296651 SNP in the samples, we designed two 159-bp oligonucleotides as positive (nt 69778476 A) and negative (nt 69778476 G) controls (nts 69778403-69778561; GenBank accession number NC_000014.9). These controls were processed by the real-time PCR protocol described above at the same time as the samples.

Sanger sequencing was used to validate the ability of real-time PCR to detect the presence/absence of this SNP in genomic DNA extracted from blood. As a 111-bp amplicon was too short to sequence by Sanger, we designed external primers, NTCP_ext_Fw: 5'-CTAGATTCCATCTGCTGCGA-3' (nts 69778631-69778612) and NTCP_ext_Rv: 5'-CCTTGGGAGTCTTGAATTTC-3' (nts 69778333- 69778352) to obtain a 259-bp fragment for a single PCR. Briefly, PCR was performed in a final volume of 25 µL, using 1x FastStart High Fidelity reaction Buffer (with 18mM $MgCl_2$), 0.2 mM of each dNTP, 0.4 µM of the forward and reverse primers, and 1.25 U of FastStart High Fidelity Enzyme Blend (Roche Diagnostics, Mannheim, Germany), and 5 µL of genomic DNA. After a single denaturation step of 2 min at 95ºC, samples underwent 35 cycles of 30 s at 95ºC, 30 s at 55ºC, and 15 s at 72ºC, and a single final 3-min step at 72ºC. PCR products were identified on 1.5% agarose gel and were cleaned with ExoSap-IT® (Affymetrix Inc., Santa Clara, US), according to the manufacturer's instructions. The sequencing reaction used the standard Big Dye 3.1. protocol (Thermo Fisher Scientific – Life Technologies, Austin, USA).

SUPPLEMENTARY RESULTS

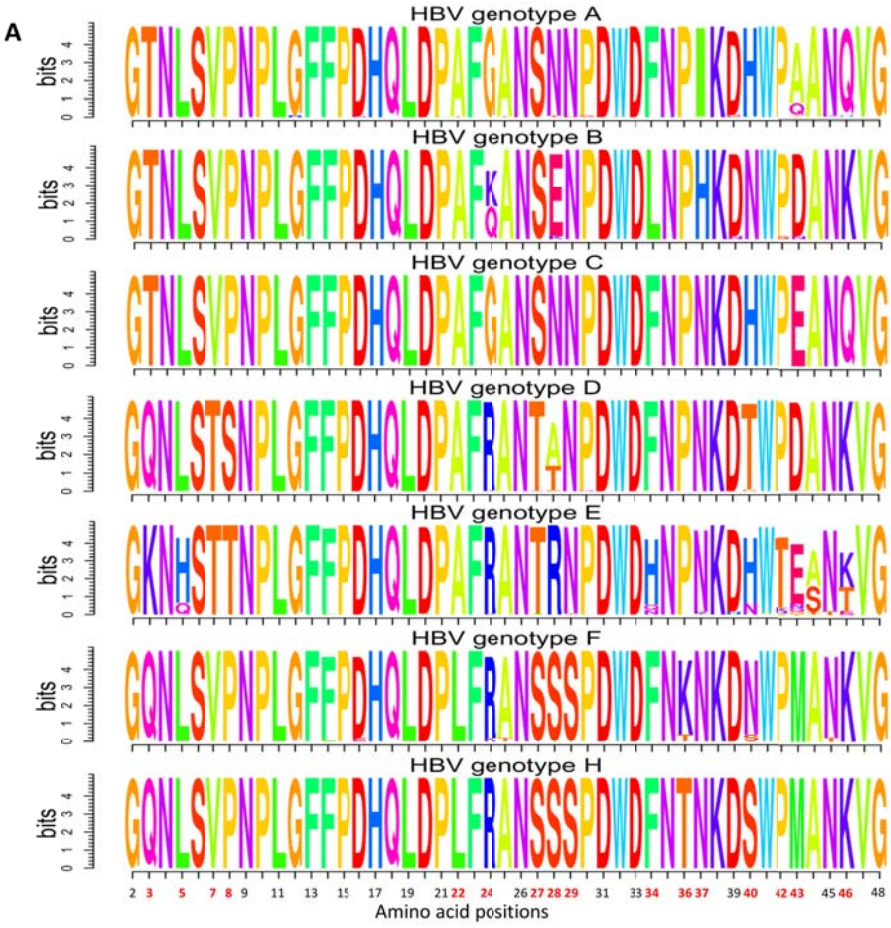
Supplementary Table 2 Percentages of amino acid changes in the sodium-taurocholate cotransporting polypeptide (NTCP) interacting domain and virion morphogenesis domain

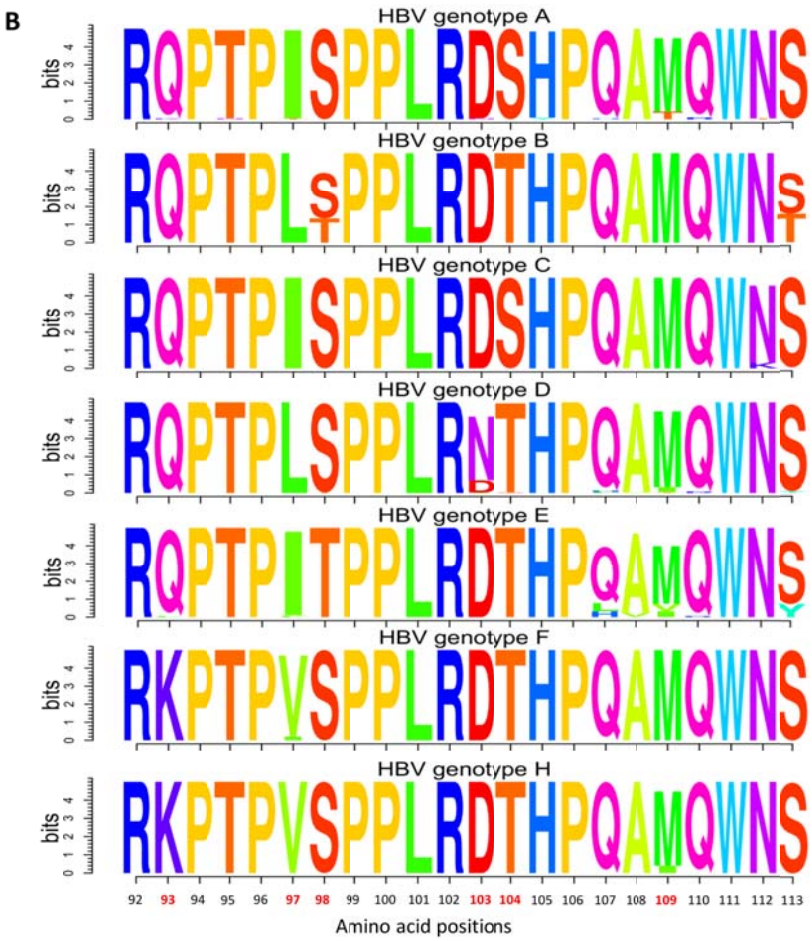
aa Position in genotype D	aa Position (Genotype)	aa Wildtype>aa substitution (Genotype; % changes)	Genotypes (% changes)						
			A (*n* = 4)*	B (*n* = 3)*	C (*n* = 4)*	D (*n* = 5)	E (*n* = 2)	F (*n* = 2)	H (*n* = 1)
5	15 (E)	H > Q (E; 5.89)					5.89		
24	35 (B)	K > Q (B; 22.95)		22.95					
28	39 (B)	D > E (B; 99.33)		100		40.29			
		D > K (B; 0.67)							
	28 (D)	A > T (D;39.91)							
		A > S (D; 0.17)							
		A > E (D; 0.15)							
		A > V (D; 0.06)							
34	44 (E)	H > Q (E; 2)					10.13		
		H > N (E; 8.13)							
36	47 (F)	K > T (F;50)						50	
39	49 (E)	D > H (E; 1.23)					1.23		
40	50 (E)	H > N (E; 13.79)					13.79	50	100
	51 (F,H)	N > S (F; 50)							
		N > S (H; 100)							
42	52 (E)	T > N (E; 4.43)					6.31		
		T > K (E; 1.31)							
		T > P (E; 0.35)							
		T > I (E; 0.23)							
43	54 (A)	A > Q (A; 25)	25.32				4.91		
		A > T (A; 0.22)							
		A > V (A; 0.10)							
	53 (E)	E > N (E; 2.11)							

		E > Q (E;1.29)								
		E > G (E; 0.94)								
		E > T (E; 0.57)								
aa Position in genotype D	**aa Position (Genotype)**	**aa Wildtype>aa substitution (Genotype; % changes)**	**Genotypes (% changes)**							
			A (n=4)*	**B (n=3)***	**C (n=4)***	**D (n=5)**	**E (n=2)**	**F (n=2)**	**H (n=1)**
		A > S (E; 25.46)							
44	54 (E)	A > T (E; 1.41)					27.43		
		A > P (E; 0.57)							
45	55 (E)	N > T (E; 6.75)					6.75		
		K > T (E; 25.72)							
46	56 (E)	K > E (E; 2.69)					29.48		
		K > Q (E; 1.08)							
98	109 (B)	S > T (B; 9.60)		9.60					
103	103 (D)	N > D (D; 13.45)				13.94			
		N > I (D; 0.49)							
107	117 (E)	Q > L (E; 8.99)					14.76		
		Q > H (E; 5.78)							
108	118 (E)	A > V (E; 2.69)					2.69		
109	120 (A,H)	M > T (A; 4.47)	4.47			4.37	12.94		100
		V > M (H; 92.51)							
	109 (D)	M > I (D; 3.80)							
		M > V (D; 0.49)							
		M > T (D; 0.08)							
	119 (E)	M > V (E; 9.13)							
		M > I (E; 3.82)							
		V > I (H; 7.49)							
110	110 (D)	Q > R (D; 0.90)				1.00			
		Q > K (D; 0.10)							

113	124 (B) 123 (E)	S > T (B; 22.95) S > Y (E; 14.36)	22.95	14.36

The aa positions included from the domains studied are those in which aa changes in any of the HBV genotypes are present in >1% (columns 4 to 10). For each position indicated in the second column, the first column shows its numeration according to genotype D, to facilitate comparison of positions with those of Figure 2 and 3 of the main text. Since the haplotypes obtained by next-generation sequencing were grouped according to genotype, and two patients showed haplotypes belonging to more than one genotype, the number of patients per genotype (n in columns 4 to 10) sometimes differed from the number of patients described in Materials and Methods because each group of haplotypes with the same genotype within a single patient was considered as a different patient. For example, a patient with 3 haplotypes belonging to 3 different genotypes would be represented as 3 different patients. Genotypes where this occurred are marked with an asterisk.


A
HBV genotype A
HBV genotype B
HBV genotype C
HBV genotype D
HBV genotype E
HBV genotype F
HBV genotype H
bits
0 1 2 3 4
2 3 5 7 8 9 11 13 15 17 19 21 22 24 26 27 28 29 31 33 34 36 37 39 40 42 43 45 46 48
Amino acid positions



Supplementary Figure 1 Sequence logos from all haplotypes obtained by next generation sequencing grouped according to HBV genotype. The sequence logos represent calculation of the information content (IC) of each amino acid (aa) position. To simplify variations due to HBV genotype, the numeration of aa positions from the two domains is reported according to genotype D: **(A)** sodium-taurocholate cotransporting polypeptide (NTCP) interacting domain from residue 2 to 48 of the N-terminal end of preS1 and **(B)** virion morphogenesis domain, from residue 92 to 108 of the C-terminal end of preS1 and first 5 residues of the N-terminal end of preS2. Positions where the wild-type aa varies according to HBV genotype are highlighted in bold and red. Haplotypes obtained by next-generation sequencing were grouped according to their HBV genotype. In this way, sequence logos from both domains assessed in the preS1 region were created from the calculated IC from haplotypes of 4 patients with genotype A, 3 patients with genotype B, 4 patients with genotype C, 5 patients with genotype D, 2 patients with genotype E, 2 patients with genotype F, and only 1 patient with genotype H.

REFERENCES

1. **R Development Core Team**. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2016, Vienna, Austria. Available from: URL: https://www.r-project.org/
2. **Ramírez C**, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. *Antiviral Res* 2013; **98**: 273–283 [PMID: 23523552 DOI: 10.1016/j.antiviral.2013.03.007]
3. **Gregori J**, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. *Virology* 2016; **493**: 227–237 [PMID: 27060566 DOI: 10.1016/j.virol.2016.03.017]