Hepatitis B virus preS2Δ38–55 variants: A newly identified risk factor for hepatocellular carcinoma

Damien Cohen, Sumantra Ghosh, Yusuke Shimakawa, Njie Ramou, Pierre Simon Garcia, Anaëlle Dubois, Clément Guillot, Nora K Deluce, Valentin Tilloy, Geoffroy Durand, Catherine Voegele, Gibril Ndow, Céline Brochier-Armanet, Sophie Alain, Florence Le Calvez-Kelm, Janet Hall, Fabien Zoulim, Maimuna Mendy, Mark Thursz, Maud Lemoine, Isabelle Chemin

Table of Contents

Fig. S1. 2

Fig. S2. 3

Fig. S3. 6

Fig. S4. 7

Supplementary information for alignments.......... 10

References.......... 13

Fig. S1. Flow diagram of study participants

The individuals in this study were participants in the PROLIFICA project conducted in The Gambia. HBsAg screening was carried out using a rapid point-of-care test (Determine, Alere, USA), and confirmed by ELISA (AxSYM HBsAg V2, Abbott, USA).

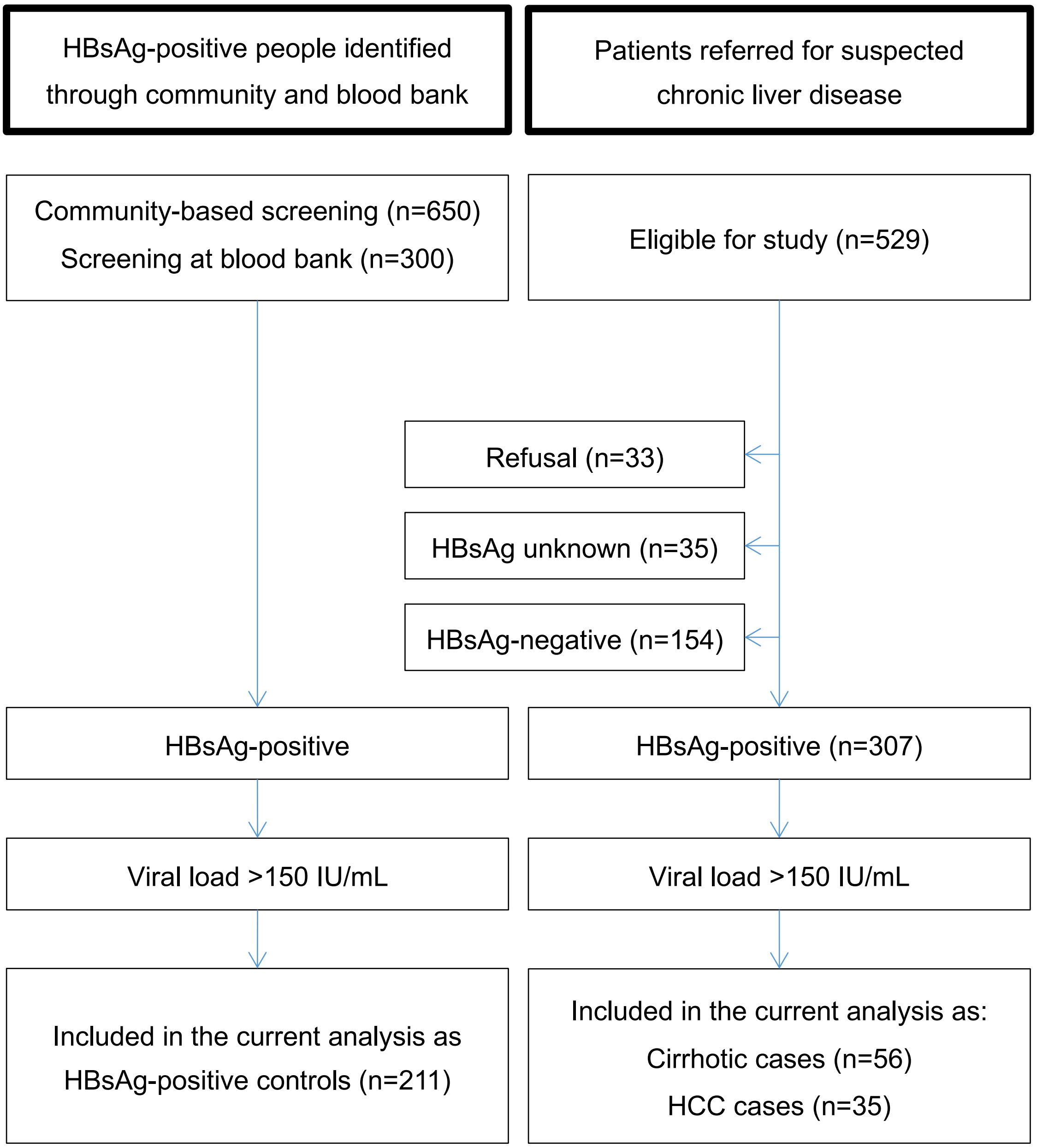
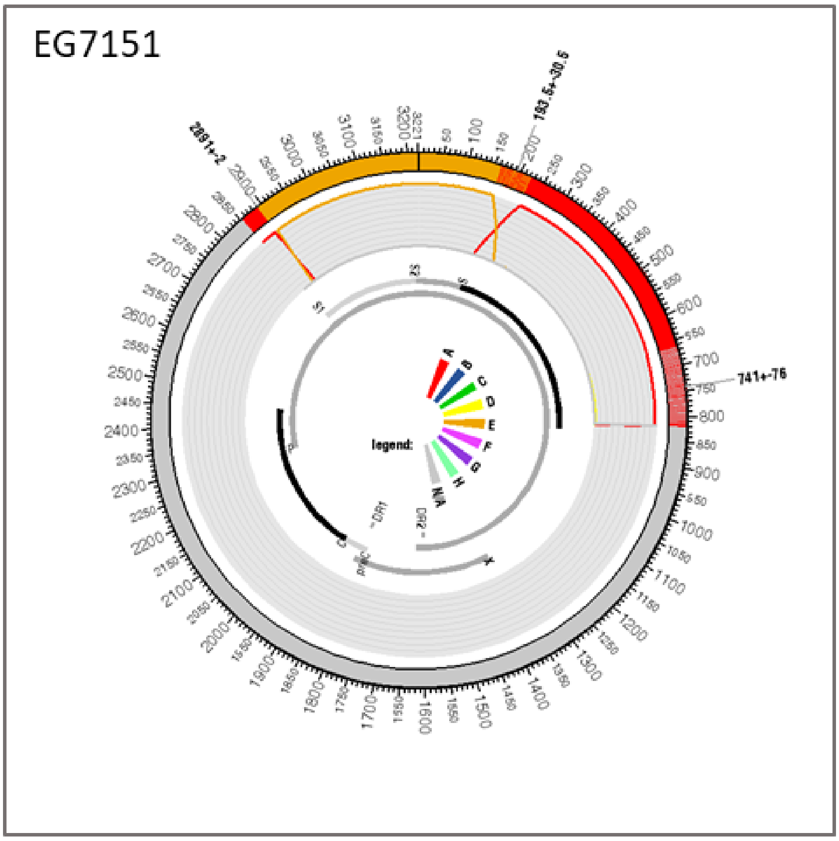

HBsAg-positive people identified through community and blood bank
Patients referred for suspected chronic liver disease
Community-based screening (n=650)
Screening at blood bank (n=300)
Eligible for study (n=529)
Refusal (n=33)
HBsAg unknown (n=35)
HBsAg-negative (n=154)
HBsAg-positive
HBsAg-positive (n=307)
Viral load >150 IU/mL
Viral load >150 IU/mL
Included in the current analysis as HBsAg-positive controls (n=211)
Included in the current analysis as:
Cirrhotic cases (n=56)
HCC cases (n=35)

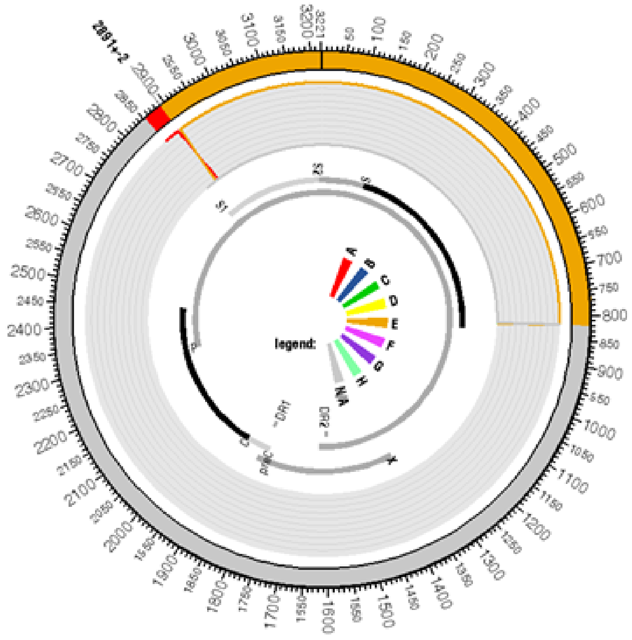
Fig. S2. Genotyping profile of A/E recombinants

Using the jpHMM on Sanger sequenced S genes and based on reference genome AM282986 numbering (number 1 is the historical annotation of the EcoRI cleavage site), the recombination sites were identified for sequences EG7151, EG7168, EG7190 and EG7194. For phylogenetic analysis, sequences were cut at the breakpoint. Colours indicate the genotypes: red = A genotype, orange = E genotype.

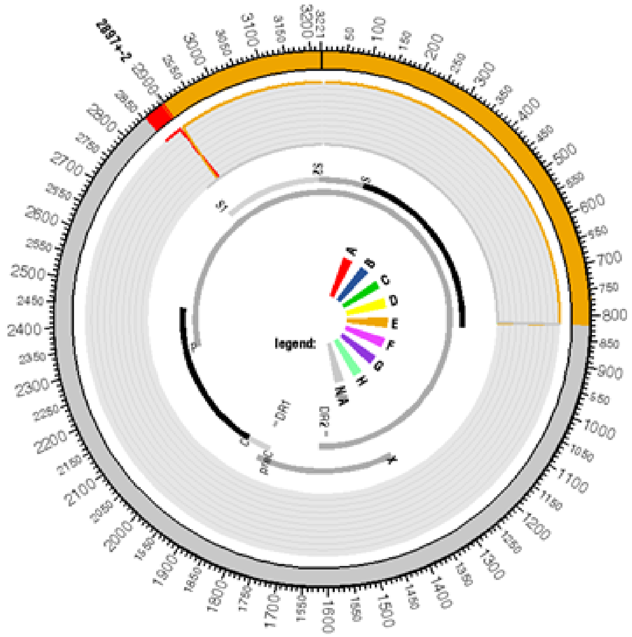
Sequence	Breakpoint interval	Breakpoint
EG7190	2479 – 2893	2891
EG7168	3023 – 3047	3035
EG7194	2895 – 2989	2987
EG7151	2889 – 2893	2891
	163 – 224	193.5



EG7190



EG7194



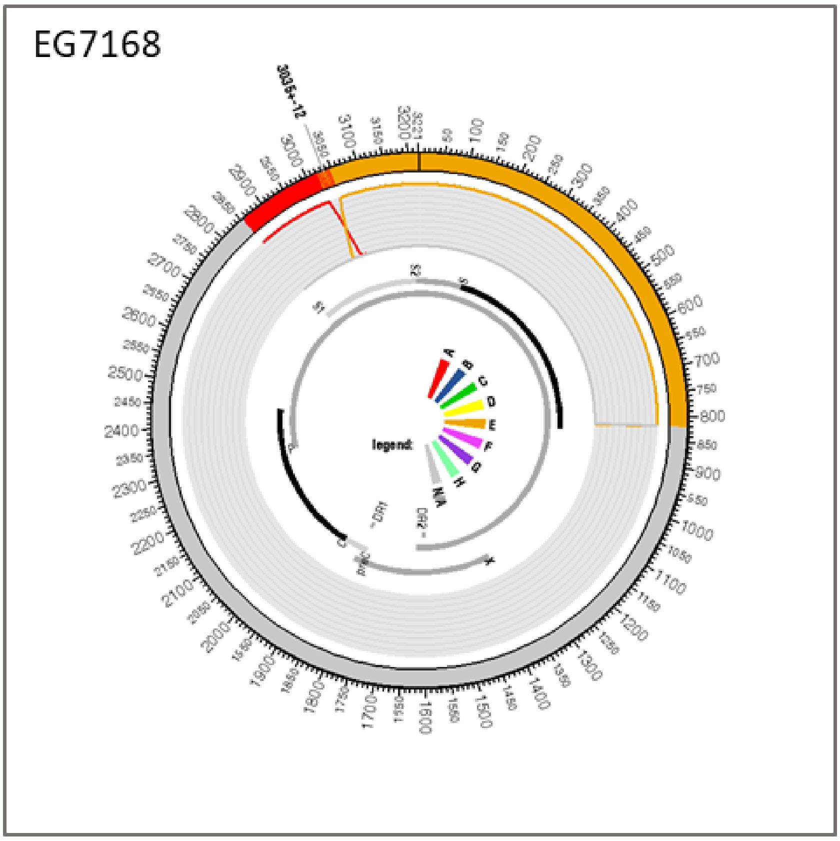

EG7168
3035+12
legend:
A
B
C
D
E
F
G
H
N/A
S1
S2
S
P
preC
C
DR1
DR2
X

Fig. S3. Representative NGS PreS2 domain sequencing profiles

Each panel shows an NGS profile for a representative sample containing either the 12 base-pair, 18 base-pair deletion or the wild-type (WT)-HBV virus (y-axis: number of reads, x-axis: EcoRI-nomenclature position in the PreS2-domain). In the "deleted" samples, the number of reads is reduced reflecting the presence of a gap, but some reads remain showing the systematic coexistence of mutant and WT populations.

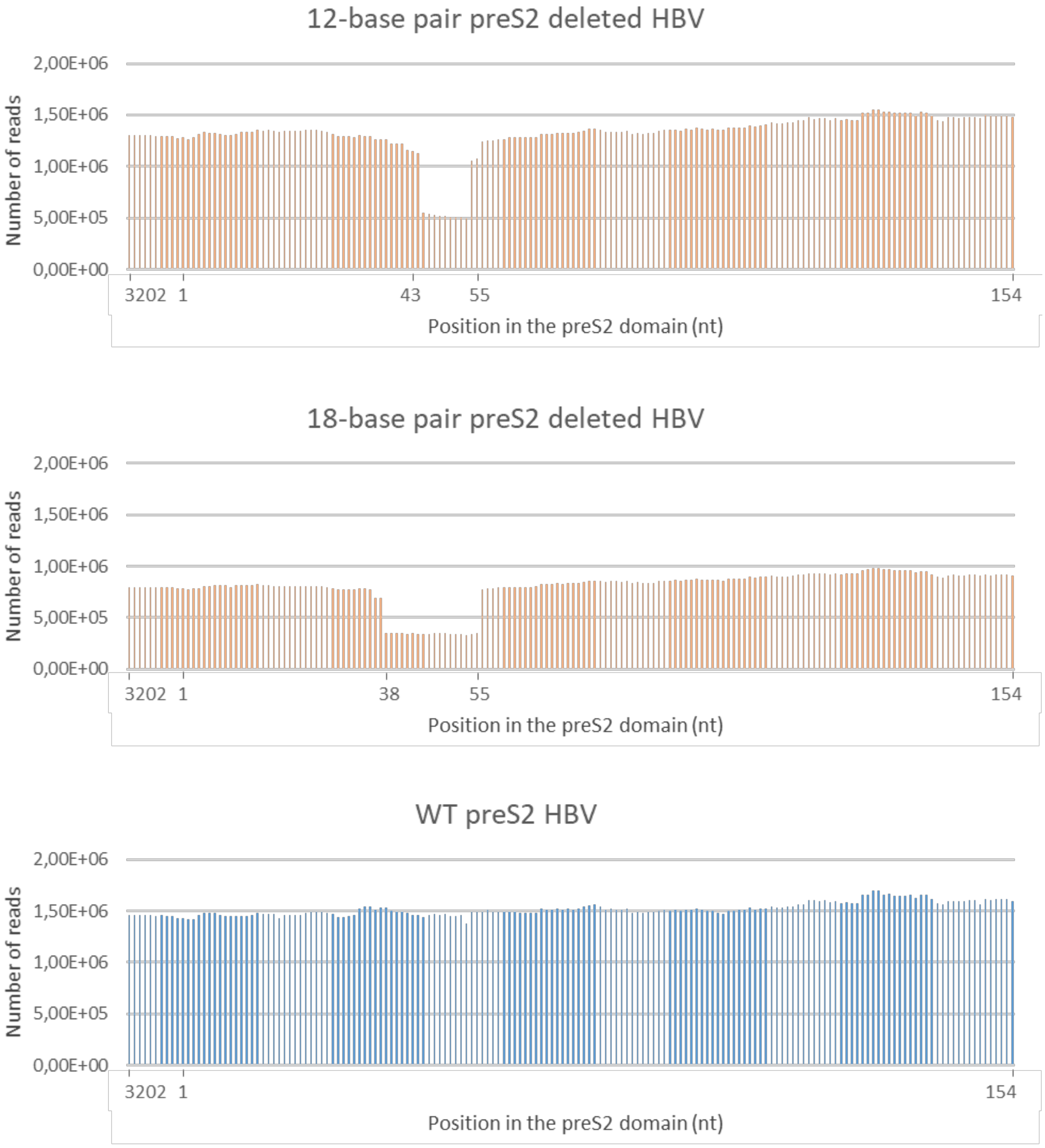
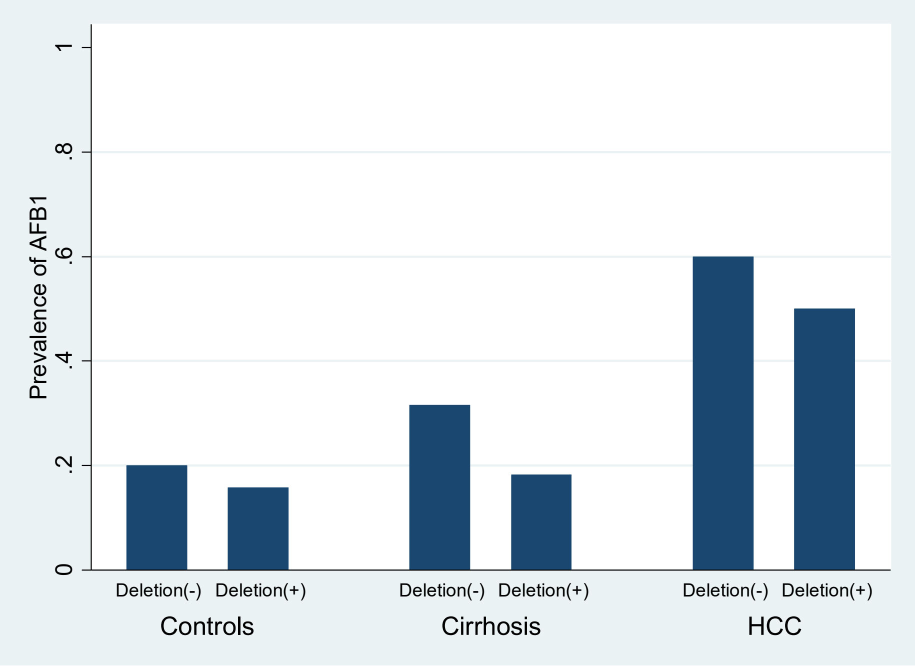


Fig. S4. Box and whisker plots of serum HBV DNA levels, HBsAg and HBe levels, and AFB1 exposure by the disease group and the presence of PreS2Δ38-55

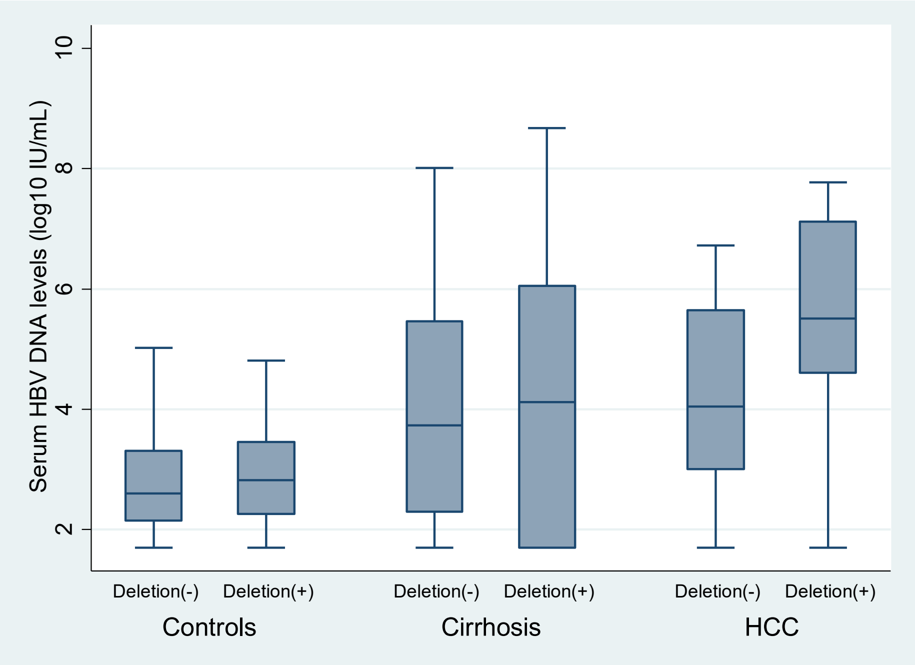
4-A. Prevalence of AFB1 exposure

No significant difference in prevalence of AFB1 exposure between those with and without deletions 38-55bp, in each disease category



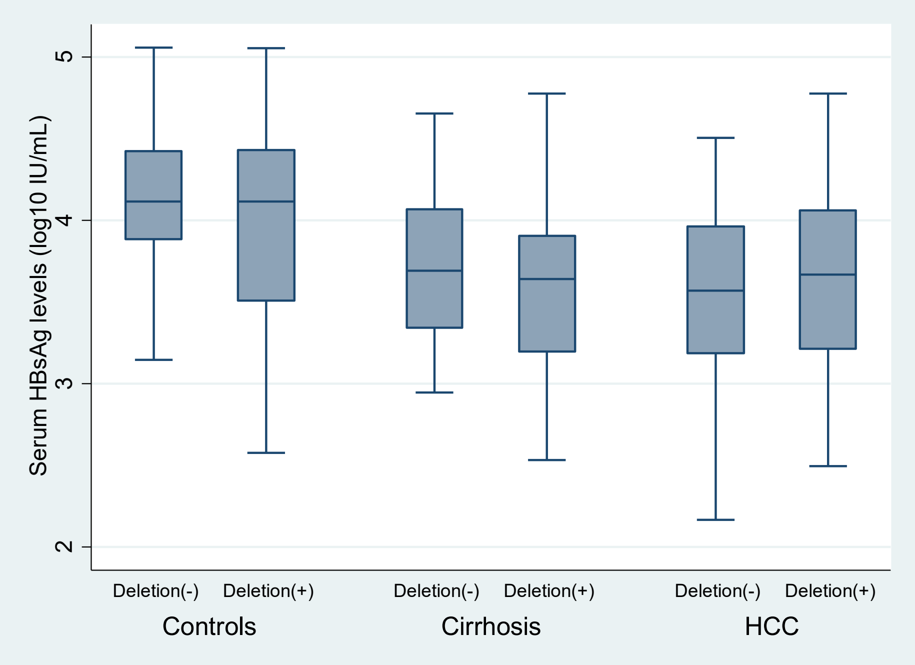
4-B. Serum HBV DNA levels

No significant difference in median HBV DNA levels between those with and without deletions 38-55bp, in each disease category



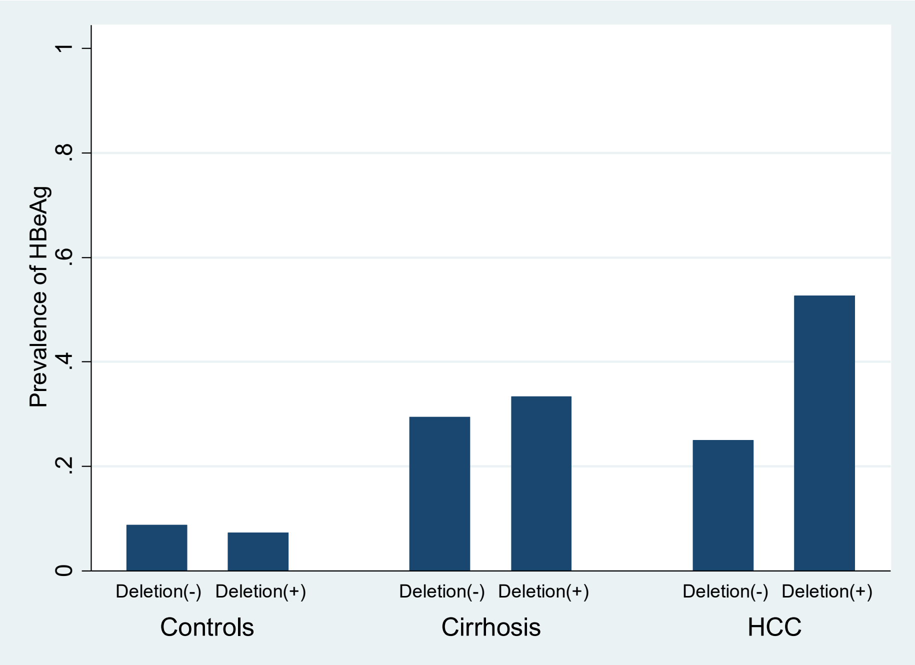
4-C. Serum HBsAg levels

No significant difference in median HBsAg levels between those with and without deletions 38-55bp, in each disease category



4-D. Prevalence of HBeAg-seropositivity

No significant difference in HBeAg-seropositivity *prevalence between those with and without deletions 38-55bp, in each disease category*



Supplementary information for individuals' alignments

patient_alignment.fasta

Dataset of 308 sequences of S region. 304 sequences come from 302 individuals. Three sequences have been split in two regions, 1 sequence in 3 regions, regarding recombinations. 3 regions have been removed because of their short length (EG7151_Cter : 37 nt, EG7194_Cter : 44 nt, EG7190_Cter : 37 nt). Reference sequences of genotype B, C, D and F were used (AB010289, AB014360, AB048701 and F AB036905 for B, C, D and F, respectively). The alignment includes 1,233 positions.

patient_trimmed_alignment.phy

Dataset of 308 sequences of S region from 302 patients after trimming. The alignment includes 1,176 positions.

patient_phyml_tree.txt

PhyML tree of 308 sequences of S region from 302 patients (GTR+G4). Supports correspond to Bootstrap values.

Supplementary materials and methods

Isolation of Circulating cell-free DNA for p53R249S mutation analysis

Circulating cell-free DNA (cfDNA) was isolated from plasma (600 µl) for 117 subjects, using the QIAamp Circulating Nucleic Acid Kit (Qiagen), and quantified using the Quant-iT™ PicoGreenR dsDNA Assay (Molecular Probes, Invitrogen) PicoGreen®).

An 83bp fragment of *TP53* exon 7 (codons 237 to 261, hg19: ch17: 7,577,489 – ch17: 7,577,571, forward primer 5'-TGTGTAACAGTTCCTGCAT-3', reverse 5'-GGCTCCTGACCTGGAGTCTT-3') was PCR amplified using 5ng of cfDNA, 5X AccuStart Buffer, 200 nM forward and reverse primers and 0.04 U/mL of AccuStart HiFi Taq Polymerase (Quanta BioSciences) with the following conditions. After quantification using the Qubit^{TM} dsDNA HS Assay Kit and (Invitrogen) and Qubit$^{®}$ 2.0 fluorometer, 20 ng of the PCR product were purified with Serapure magnetic beads at a final concentration of 2.5X and 28% isopropanol.

NGS analysis by Ion Torrent technology.

For the mutation analysis of *TP53*, library preparation was done using the NEB Next® Fast DNA Library Prep Set for Ion Torrent™ kit (New England Biolabs) as previously described [8] with the size selection of pooled libraries (~180 bp) being performed in a 2% agarose gel. Briefly, 12.5µl of the 20µl purified products were end-repaired in 15µl, and added to 8.6µl of ligation reaction mix, 0.7µl of the Ion P1 Adapter and 0.7 µl of each Ion Barcode for the ligation step. The barcoded products were purified using Serapure magnetic beads at final concentration of 1.8X, amplified in 25µl and quantified using Qubit quantification system. 40 ng of amplified barcoded products were pooled into a single tube and the cleanup and size selection of pooled libraries (~180 bp) was performed in a 2% agarose gel and MinElute Gel Extraction Kit (Qiagen). The pool of purified barcoded libraries was quantified using the Qubit quantification system and the assessment of the library quality (molarity and size analysis) was done using the Agilent$^{®}$ High Sensitivity DNA Kit and the Agilent Technologies 2100 Bioanalyzer^{TM} (Agilent Technologies). Emulsion amplification was performed on the Ion OneTouch 2 system (Thermoscience Fisher, Waltham, MA) using 7 µL of 100 pM library and the Ion PI Hi-Q OT2 200 Kit (Thermoscience Fisher), according to the manufacturer's protocol. The purified 1.4 Kb fragment from the full HBV S region was used to prepare libraries using the same protocol, but was automatized on a Library Builder instrument (Life technologies) with size selection using E-Gel (Invitrogen).

The purified barcoded libraries were sequenced on an Ion Proton System (Thermoscience Fisher, Waltham, MA). The accuracy and detection threshold and the read error rate for the p53R249S mutation was determined as previously reported [9] using genomic DNA from the cell-line PLCPRF5 harboring a homozygous p53R249S mutation serially diluted into wildtype *TP53* genomic DNA. For the analysis of HBV sequences, the detection limits for variant sequences was estimated to be 2% for a coverage of 10,000 reads from repeats of a 2500bp cytomegalovirus DNA PCR product with the same GC content [10].

NGS bioinformatics and statistical analyses

For HBV, signal processing and base-calling were performed with Torrent Suite Software version 5.0.2. For samples EG7132, EG7141, EG7156, EG7181, EG7030, EG7129, EG7116, EG7166 and EG7201 variant calling was obtained using Torrent Variant Caller 5.8 and applying Somatic variant frequency (including down sample to coverage of 50000) and HBV EG0270 genome (genotype E) as reference. PreS mutations were filtered and highlighted using either IGV_2.3.55 or a python local script.

For TP53, the bioinformatics analysis was performed using a combination of the Torrent Suite Software version 5.8 and the Needlestack algorithm in extra-robust regression mode [8] (https://github.com/IARCbioinfo/needlestack). A threshold of Phred scale q-values QVAL>30 was used to call variants (QVAL= -10 log10 (q-value)). Reads were mapped to the human whole genome and BAM files were generated by the Ion Torrent Proton server using default parameters. The variant calling was perfomed using Needlestack in extra-robust regression mode and using only reads with a base quality above 13 at the position TP53c747G. As the p53R249S could be recurrently identified in our sample set and Needlestack aims at calling rare variants, samples with an allelic frequency >5% were considered as true variants and the regression was fitted on the remaining set of samples (i.e extra robust regression). We calculated for each sample the p-value for being a variant (outlier from the regression) that we further transformed into q-values to account for multiple testing. *q*-values are reported in Phred scale Q=-10 $\log_{10}(q\text{-value})$, and we used a threshold of Q>30 to call variants.

REFERENCES

1 Ghosh S, Banerjee P, RoyChoudhury A, *et al.* Unique Hepatitis B Virus Subgenotype in a Primitive Tribal Community in Eastern India. *Journal of Clinical Microbiology* 2010; **48**: 4063–71.

2 Ghosh S, Banerjee P, Deny P, *et al.* New HBV subgenotype D9, a novel D/C recombinant, identified in patients with chronic HBeAg-negative infection in Eastern India. *Journal of Viral Hepatitis* 2013; **20**: 209–18.

3 Crooks GE. WebLogo: A Sequence Logo Generator. *Genome Research* 2004; **14**: 1188–90.

4 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010; **59**: 307–21.

5 Schultz A-K, Bulla I, Abdou-Chekaraou M, *et al.* jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic Acids Res* 2012; **40**: W193–8.

6 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–9.

7 Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016; **44**: W242-245.

8 Le Calvez-Kelm F, Foll M, Wozniak MB, *et al.* KRAS mutations in blood circulating cell-free DNA: a pancreatic cancer case-control. *Oncotarget* 2016; **7**: 78827–40.

9 Fernandez-Cuesta L, Perdomo S, Avogbe PH, *et al.* Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. *EBioMedicine* 2016; **10**: 117–23.

10 Garrigue I, Moulinas R, Recordon-Pinson P, *et al.* Contribution of next generation sequencing to early detection of cytomegalovirus UL97 emerging mutants and viral subpopulations analysis in kidney transplant recipients. *J Clin Virol* 2016; **80**: 74–81.