Supplementary Information

Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma

Péneau C. et al.

Supplementary Materials and Methods

Supplementary Figures

Supplementary Tables

Supplementary table 1 - Clinical description of the series.
Supplementary table 2 - List of probe sets and primers.
Supplementary table 3 - List of HBV genotypes and human regions targeted in viral capture.
Supplementary table 4 - Characterization of 8809 HBV integration breakpoints identified by viral capture.
Supplementary table 5 - Survival analysis in patients treated with R0 curative resection.

SUPPLEMENTARY MATERIALS AND METHODS

Patients and tissue samples

A first series of 1128 hepatocellular carcinoma and 1063 non-tumor counterparts from 1128 patients was assembled and the study was approved by our institutional review board (IRB) committees (CCPRB Paris Saint-Louis, 1997 and 2004; Bordeaux 2010-A00498-31, Ile-de-France VII: projects C0-15-003 and PP 16–001). Patients were involved as advisers in hospital tumor collection boards and the definition of inform consent in IRB. They were not involved in the design of this study. Samples were collected in 12 academic hospitals in France (LICA-FR cohort) and frozen immediately at -80°C after resection or biopsy. DNA was systematically extracted, as RNA when the quantity of tissue was sufficient. The majority of tumor samples was from primary tumors except for 12 samples collected at relapse. Clinical features and serological data were gathered from each center, classified as previously described[1], and are summarized in **supplementary table 1**. A flowchart of the study inclusion at each step is provided in **supplementary figure 1**.

Viral DNA screening

Genomic DNA were screened to detect the presence of HBV DNA by quantitative RT-PCR (qRT-PCR) on Fluidigm 96.96 dynamic arrays using the BioMark Real-Time PCR system with TaqMan probe sets designed with Primer3Plus software, as previously described[2]. In total eleven probes were designed to detect specifically five viral regions for all the eight main HBV genotypes (**supplementary table 2**). Results were analyzed using the Fluidigm Real-Time PCR Analysis software (V.4.1.3) and reported to HMBS (Hydroxymethylbilane Synthase) as reference gene. The quantification was expressed in viral copy number/cell and based on the results obtained for a series of HBs and HBx plasmids with known concentration. The values obtained were tested for bimodal distribution using normalmixEM function of mixtools package in R[3].

HBV viral capture

Viral capture was performed for a selection of 177 frozen HCC and 170 matched non-tumor liver tissues from 177 patients with a known HBV etiology. 13 tumors and 8 non-tumor liver tissues from 13 HBV-negative patients were used as negative controls. Single-stranded biotinylated probes were designed with Roche NimbleGen (SeqCap EZ Designs, Roche NimbleGen Inc., Madison, WI, USA), to target 40 references representing the 8 main HBV genotypes and 4 human regions (*TERT* promoter, *CCNE1*, *CCNA2* and *KMT2B*). The reference of sequences used is shown in **supplementary table 3**.

Samples containing 1ug of DNA were used for DNA library preparation. DNA was sheared mechanically in fragments with an average length of 1kb, before following the SeqCap EZ HyperCap Workflow developed by Roche. We used KAPA Dual-Indexed Adapters for sample indexing and we multiplexed library samples by using one capture probe set for 12 to 42 samples. DNA samples were selected for multiplexing based on the HBV copy number per cell measured by qRT-PCR, in order to have similar viral loads between all library samples in one

capture pool. A sample not containing HBV DNA was added in each pool as negative control. We performed a double capture with two successive hybridization steps to enrich twice for DNA targets and increase the specificity of capture, following an adapted protocol developed by Roche. Final libraries were quantified with KAPA Library Quantification Kit and controlled with a DNA High Sensitivity assay on Caliper LabChip GX. Two final libraries were merged to be sequenced using an Illumina MiSeq instrument with paired-end reads of 2 × 250 nt.

Analysis of integration events and clonality definition

A flowchart described the overall pipeline used for the analysis of integration events is provided in **supplementary figure 2**.

Raw capture data obtained after sequencing were post-processed by trimming low-quality bases with Trimmomatic[4] and removing duplicates. Filtered reads were first aligned on the reference used (40 HBV sequences and 4 regions of HG19 genome) using Burrows-Wheeler Aligner (V.0.7.15)[5]. A normalization factor was assessed based on the mean coverage of the targeted human regions and was corrected to remove capture biases. The main HBV genotype of each sample was determined as the sequence with the higher number of mapped reads. Read pairs with at least one read aligned on a HBV sequence were extracted using samtools (V.1.3)[6], and realigned to a custom reference genome including the HG19 reference and the HBV sequence of the main genotype identified for each sample (one single reference was chosen for each of the 8 genotypes A-H) (**supplementary table 3**). For genotype A, the X02763 sequence was renumbered to use the EcoR1 restriction site as the +1. To compare HBV breakpoints from different genotypes, all the positions in HBV genomes were converted using this numbering.

HBV copy number per cell was assessed from the ratio (multiplied by 2) of the mean read coverage of the HBV genome and the normalization factor, and was compared with the measures obtained by qRT-PCR (**supplementary figure 3A**). HBV copy number per cell includes all HBV forms: both integrated HBV sequences and HBV episomal forms. We identified the HBV/human chimeric regions with at least one chimeric read supporting the junction and one read pair with a read mapping on HBV and its mate mapping on HG19. Coverage of reads were computed with bedtools multicov utility[7] and the characteristics of junction breakpoints (position and sequence) were extracted from hard clipped reads with htslib package. Clonality of integration events was assessed from the ratio (multiplied by 2) of the coverage of reads at the junction breakpoint and the normalization factor. All viral insertions with a clonality greater than 0.03 were validated by visual inspection on IGV (Integrative Genomics Viewer). Similar integration events (same position on HG19 and on HBV and same sequence) found in two independent samples within one single capture (and with at least a 10-fold difference in clonality) were considered as contaminations intra-capture and the event with the lowest clonality was removed from the analysis. Similar integration events found in the non-tumor and tumor tissues of the same patients were inspected specifically and considered as contaminations based on the very few number of reads detected in the non-tumor tissues compared to the tumors. Integration events located in repeated regions of the human genome with the same breakpoint and same orientation in HBV genome were considered as one unique event. The 8809 integration events identified were classified in 3 categories using the k-means method: (1)“clonal integrations” observed in more than 48% of cells, (2)“unique integrations”

observed in less than 3% of cells and (3)"subclonal integrations" with intermediate copy numbers (**supplementary figure 3B**).

Integration events identified from viral capture were compared with those identified from whole genome sequencing (WGS) for 20 tumors and 20 non-tumor liver tissues (7 pairs tumor/non-tumor had already been published) (**supplementary figure 3C-3D**). WGS was performed and analyzed as already described[8]. The bioinformatics analysis of HBV integrations was similar for WGS as for capture except for the normalization factor that was corresponding to the mean sequencing coverage for WGS (30x for non-tumor tissues and 60x or 90x for tumors).

We annotated the integration breakpoints with the following genomic features: replication timing in HepG2 cell line (ENCODE[9]), highly expressed (top 20%) genes in non-tumor liver from HBV-positive patients, common and rare fragile sites[10], repetitive sequences and CpG islands, and chromatin structure in adult liver (ROADMAP[11]). Insertions were named "classic" if the two breakpoints clustered within 25kb in opposite direction, "local inversion" if two breakpoints clustered within 25kb in the same direction, "large rearrangements" if only one breakpoint was identified and "cluster of insertions" if more than two breakpoints were identified.

Genomic DNA sequencing of driver genes

A selection of 265 tumor samples were sequenced with WGS (for 62 tumors with their adjacent liver tissues: 43 already published and 19 new cases) or Whole Exome Sequencing (WES, for 203 tumors with their adjacent liver tissues: 134 already published and 69 new cases. Protocols and methods of analysis have previously been described[1,8,12]. In particular, all tumor samples without WGS were re-sequenced to search for TERT promoter mutation with Sanger or MiSeq sequencing as previously described[13].

Detection of HBV episomal form

A specific DNAse/TaqMan-based assay was adapted from protocol by Werle-Lapostolle et al to detect HBV episomal form and performed as previously described for AAV2 episomal form detection[2]. All samples from the Capture series with enough DNA material were screened for the presence of HBV episomal form: 162 tumors and 155 non-tumor liver tissues from 172 patients. The HBV probes used are listed in **supplementary table 2**. The difference between HBV Ct values without and with PS-DNase digestion was analyzed to determine the presence of episomal form (viral DNA detected before and after digestion) or the absence of episomal form (viral DNA detected before but not after digestion). When the results did not enable to conclude directly due to close CT values, a second step of digestion was performed. The results obtained were only used as qualitative results.

Viral mRNA and specific genes mRNA screening

For viral mRNA, ten specific probe sets covering eight regions of the HBV genome were designed to detect HBV mRNA from the main eight HBV genotypes; they are listed in **supplementary table 2**. The presence of a complete transcript was defined as the positivity of all probes along the transcript. When not all probes were positive until the viral polyA, we

considered the transcript as partial or truncated. Probe sets also designed to detect HCV and HDV mRNA. For TERT expression analysis, we used human catalogue TaqMan probes (Hs00972656_m1). We performed qRT-PCR using BioMark Real-Time PCR system. Expression data were normalized with the $2^{-\Delta Ct}$ method relative to ribosomal 18S (Hs03928990_g1). Five normal tissues were used as reference.

HBV replicative forms analysis

We assessed the presence of HBV replicative form in a tissue as the presence of HBV episomal forms and the presence of pregenomic RNA (pgRNA). When episomal forms were detected in the absence of pgRNA, the tissue was considered as containing "Episomal not replicative HBV DNA".

RNAseq and transcriptomic analysis

A selection of 265 tumors (130 HBV-positive and 135 HBV-negative) and 24 HBV-positive non-tumor tissues were sequenced using Illumina TruSeq or Illumina TruSeq Stranded mRNA kit on HiSeq2000 sequencer by IntegraGen, Evry, France[14]. Among the 24 non-tumor samples, we used the Bioconductor limma package[15] to test for differential expression, between samples harboring a HBV clonal integration and other HBV-positive non-tumor tissues, of all genes expressed. We applied a q-value threshold of ≤ 0.05 to define differentially expressed genes. We used an in-house adaptation of the GSEA method[16] to identify gene sets from the MSigDB (v. 6.0) database overrepresented among up- and down-regulated genes.

In addition, qRT-PCR of 190 genes using BioMark Real-Time PCR system was performed on 133 HBV-positive tumors sequenced in viral capture, in order to classify them in the G1-G6 classification as previously described for the LICA-FR cohort[17,18].

HBV variants analysis

From viral capture, we extracted the number of each base per position of the HBV sequence. The positions on all HBV genotypes were modified to use the EcoR1 restriction site as the +1. We considered the positions 1762 and 1764 of the Basal Core Promoter region, 1896 of the PreCore region and the positions 646, 667, 670, 671, 739, 741, 836 and 877 of the Reverse transcriptase (RT) domain of HBV polymerase. For RT mutants, we considered the mutations inducing the following amino acids changes: V173L, L180M, A181T, A181V, M204V, M204I, N236T and M250V[19,20]. Only samples with a coverage greater than 20 reads at the selected position were used for analysis. A sample was considered as mutated if the number of reads harboring the mutation was greater than 20 or represented more than 10% of the total coverage.

Cell culture, transfection and dual luciferase assay

HuH7 cells were purchased from the American Type Culture Collection (ATCC) and cultured in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum and 100U/mL penicillin/streptomycin. Cells were co-transfected using Lipofectamine 3000 (Life Technologies) with a pGL3 plasmid containing the wild-type *TERT* promoter or promoter with the two hotspot mutations, or different HBV sequences (normal or scrambled) controlling a

firefly luciferase reporter gene and a plasmid encoding Renilla luciferase (Promega). Luminescence from firefly luciferase was normalized on the corresponding Renilla luciferase activity as previously described[21]. A fold change was then calculated relative to the values obtained for the construct containing the wild-type *TERT* promoter.

Long-read sequencing

Three HBV-positive tumors (#1151T, #1597T, #1994T) were selected to perform long-read sequencing using the technology SMRT (single molecule real time technology) with a PacBio-SMRTcell system (ICGex NGS platform, Curie Institute, Paris, France). Samples were selected based on the results of Fragment Analyzer to have an average length of DNA fragments of 10kb. Libraries were prepared based on the PacBio template Prep kit protocol. For analysis, error-prone raw subreads were firstly merged to obtain unique polymerase reads with the Consensus Circular Sequencing (CCS2) algorithm (Pacific Biosciences). The consensus reads obtained were aligned using minimap2 aligner (V.2.16)[22] on a custom reference genome including the HG19 reference and the HBV sequence of the main genotype previously identified for each sample.

Bionano whole-genome mapping

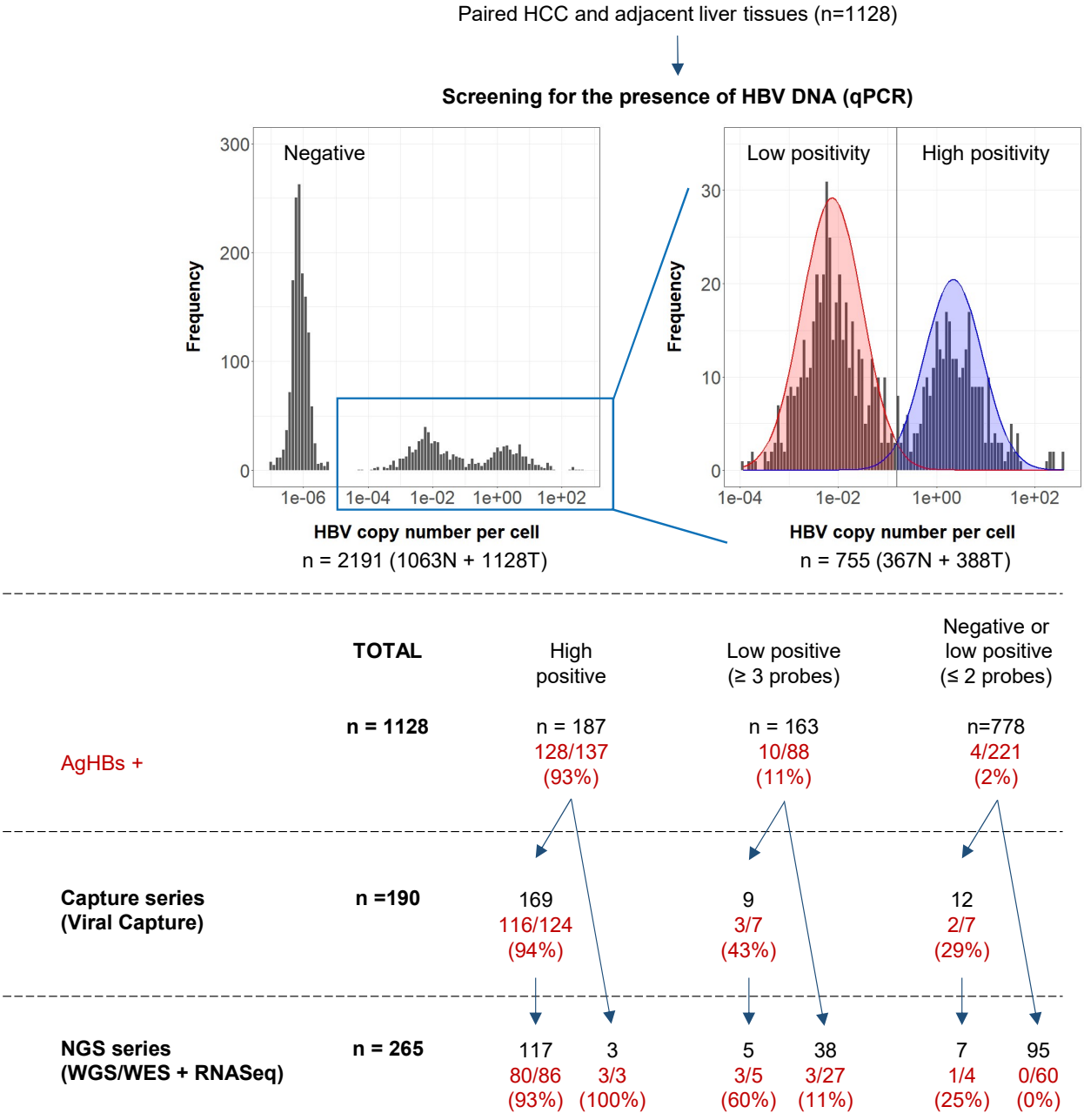
The three tumors analyzed with long-read sequencing (#1151T, #1597T, #1994T) were selected with their corresponding non-tumor liver tissues to perform whole-genome mapping using Bionano technology (Bionano Genomics, La Jolla, California, USA). 30mg of tissue were used to extract long molecules of DNA, labeled with Bionano reagents by incorporation of fluorophores at a specific sequence motif along the genome. The labeled genomic DNA was linearized in the SaphyrChip using NanoChannel arrays and single molecules were imaged and digitized. As molecules are uniquely identifiable by distinct distribution of sequence motif labels, they were then assembled by pairwise alignment into de novo genome maps. Genome-wide Structural Variant (SV) calling was performed in tumor samples by aligning these maps and molecules to a reference genome, and SV calls are annotated using the corresponding non-tumor tissue from the same patient.

Statistical analysis

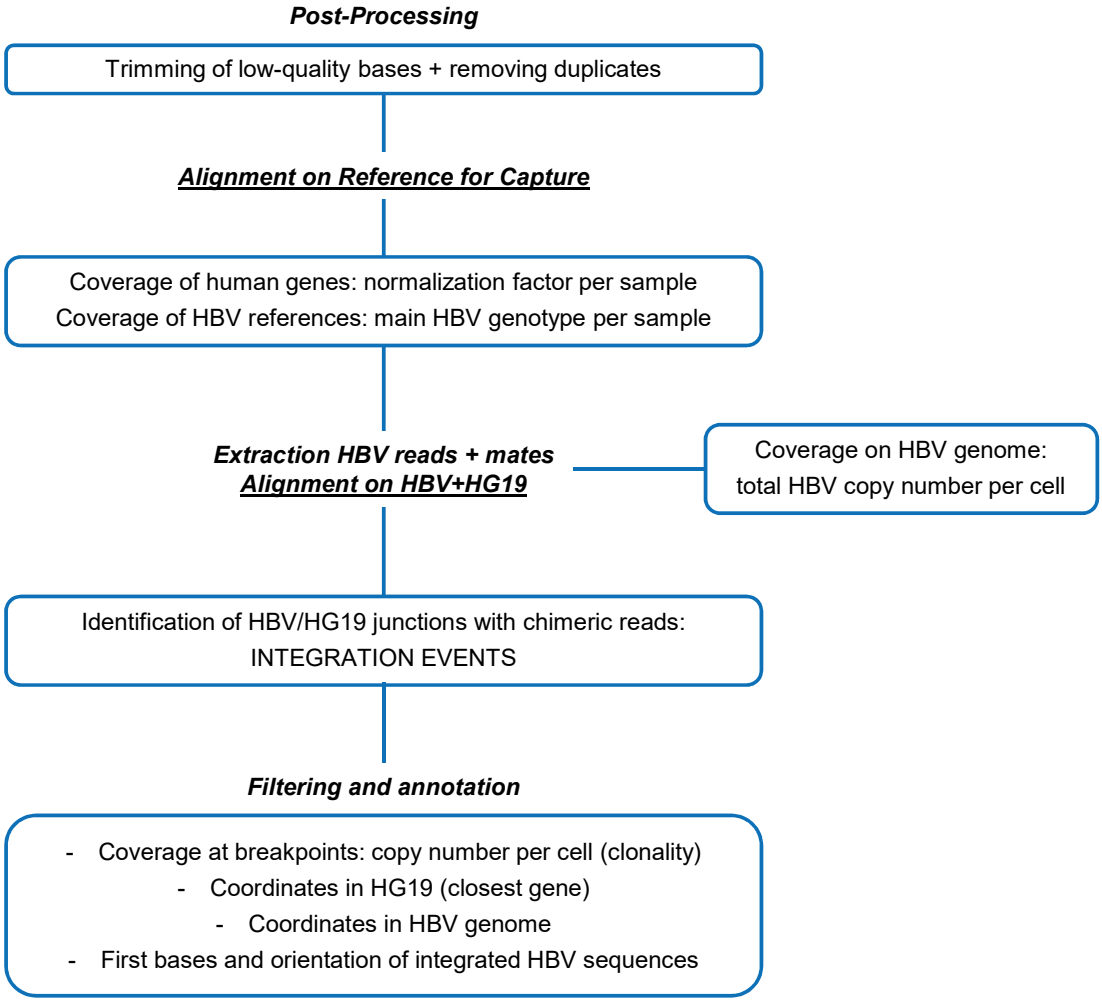
For statistical analysis, we used R version 3.6.0 (R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, (http://www.R-project.org). Wilcoxon, Fisher, Chi-square or Pearson's correlation statistical tests were applied with respect to the type of variable. Survival analysis was performed in patients treated for a primary HCC tumor by R0 liver resection as previously described[12]. We assessed overall survival defined by the interval between surgery and death. Survival curves were represented using the Kaplan-Meier method compared with the Log Rank test. Multivariate analysis was performed using Cox model. A p value <0.05 was considered as statistically significant.

References Supplementary Materials and Methods

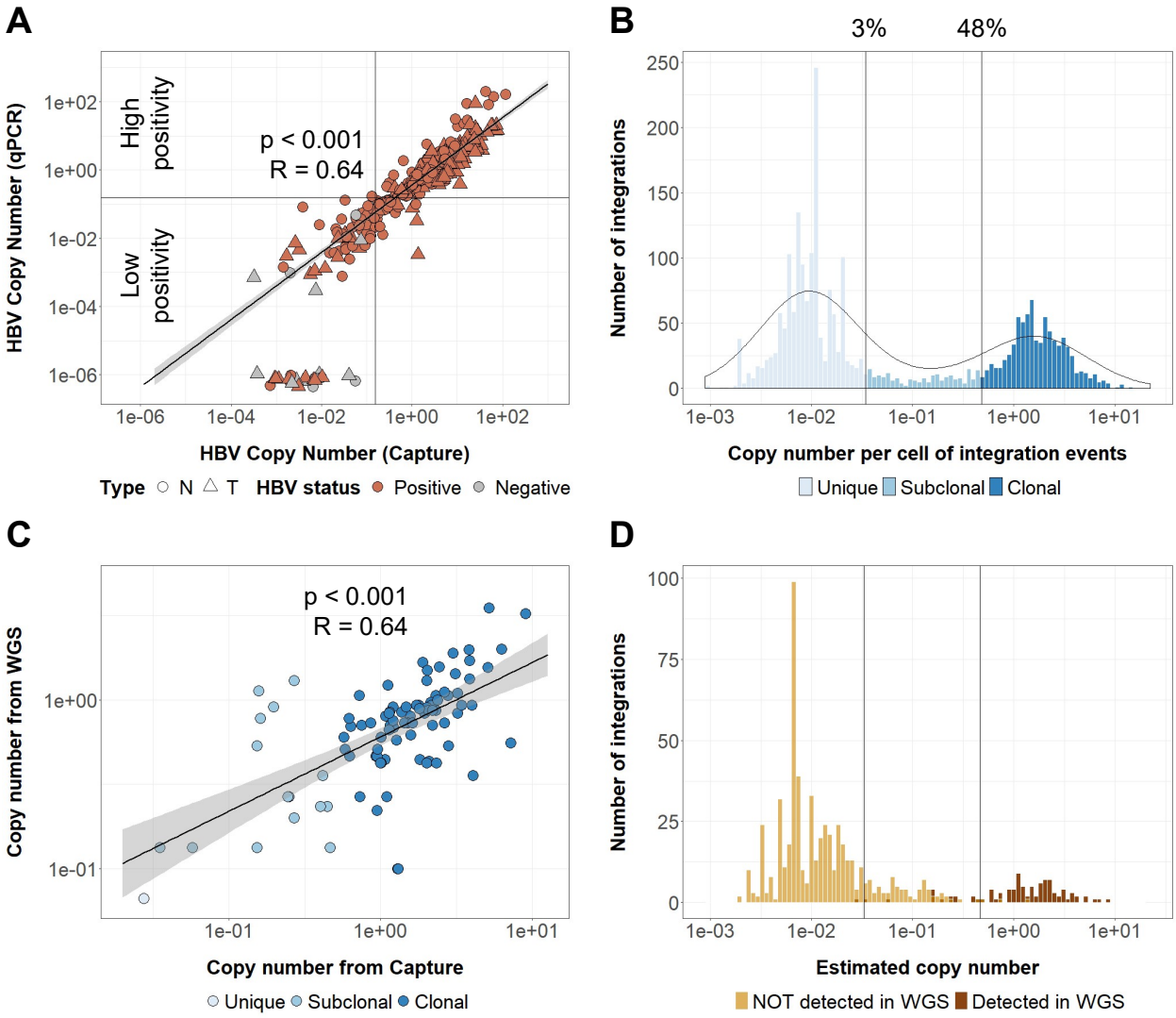
1. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
2. La Bella, T. *et al.* Adeno-associated virus in the liver: natural history and consequences in tumour development. *Gut* **69**, 737–747 (2020).
3. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. **mixtools** : An *R* Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32**, (2009).
4. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
5. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
6. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
7. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
8. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
9. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
10. Debacker, K. & Kooy, R. F. Fragile sites and human disease. *Hum. Mol. Genet.* **16**, R150–R158 (2007).
11. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
12. Nault, J. *et al.* Clinical Impact of Genomic Diversity From Early to Advanced Hepatocellular Carcinoma. *Hepatology* **71**, 164–182 (2020).
13. Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat. Commun.* **4**, 2218 (2013).
14. Bayard, Q. *et al.* Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. *Nat. Commun.* **9**, 5235 (2018).
15. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
16. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
17. Hirsch, T. Z. *et al.* BAP1 mutations define a homogeneous subgroup of hepatocellular carcinoma with fibrolamellar-like features and activated PKA. *J. Hepatol.* S0168827819307184 (2019) doi:10.1016/j.jhep.2019.12.006.
18. Boyault, S. *et al.* Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology* **45**, 42–52 (2007).
19. Lazarevic, I. Clinical implications of hepatitis B virus mutations: Recent advances. *World J. Gastroenterol.* **20**, 7653 (2014).
20. Tong, S. & Revill, P. Overview of hepatitis B viral replication and genetic variability. *J. Hepatol.* **64**, S4–S16 (2016).
21. Nault, J.-C. *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* **47**, 1187–1193 (2015).
22. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).



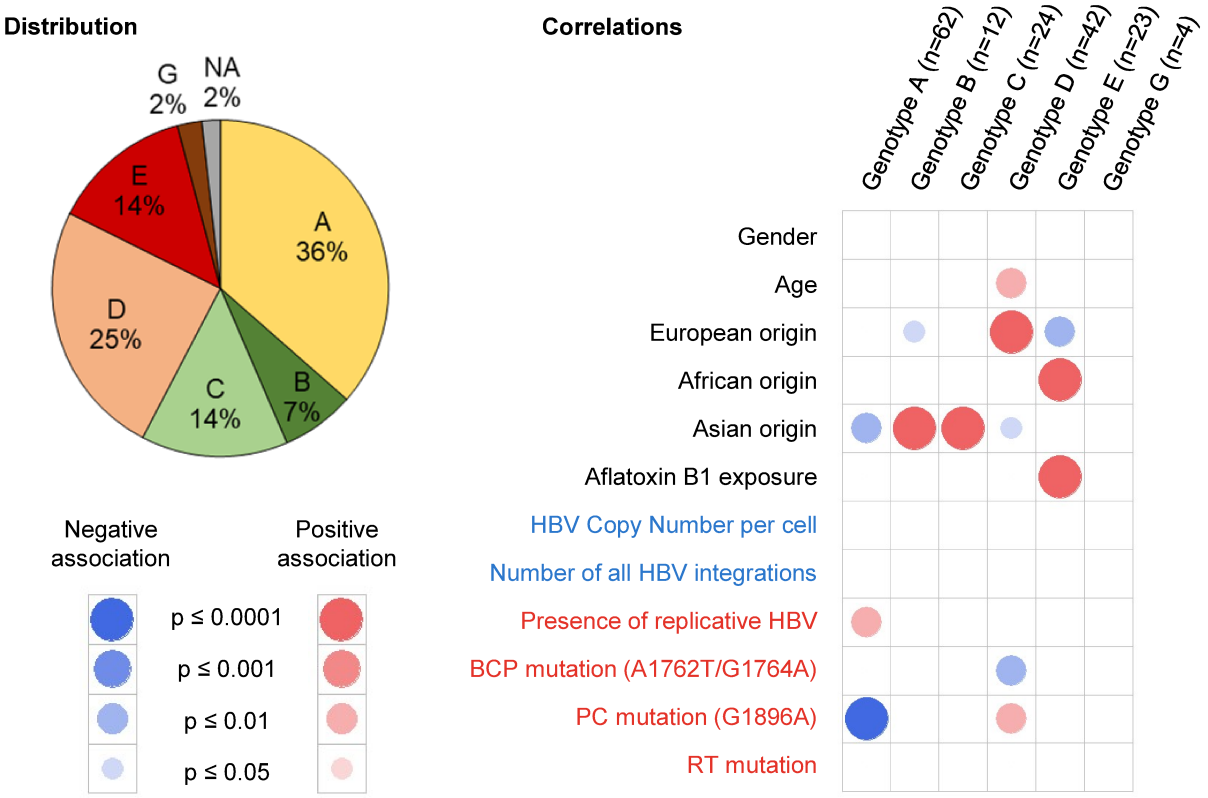
Supplementary figure 1 | Flow-chart of the study. Frozen paired HCC and adjacent liver tissues from 1128 patients were screened to detect the presence of HBV DNA by quantitative PCR with Taqman probes targeting 5 regions of the viral genome. Positive values were defined as low positive or high positive based on a bimodal distribution. Patients were then classified into 3 groups by considering for one patient the highest value between the corresponding two paired samples (tumor and adjacent liver tissue): high positive (at least 1 positive probe with a high positive value), low positive (at least 3 positive probes with low positive values), negative or low positive (0, 1 or 2 probes with low positive values). The positivity for AgHBs in patients' serum is indicated in red when available (See patient's characteristics in Supplementary Table 1). *HCC, hepatocellular carcinoma; N, non-tumor; T, tumor; NGS, next-generation sequencing; WGS, whole-genome sequencing; WES, whole-exome sequencing; RNASeq; RNA sequencing.*



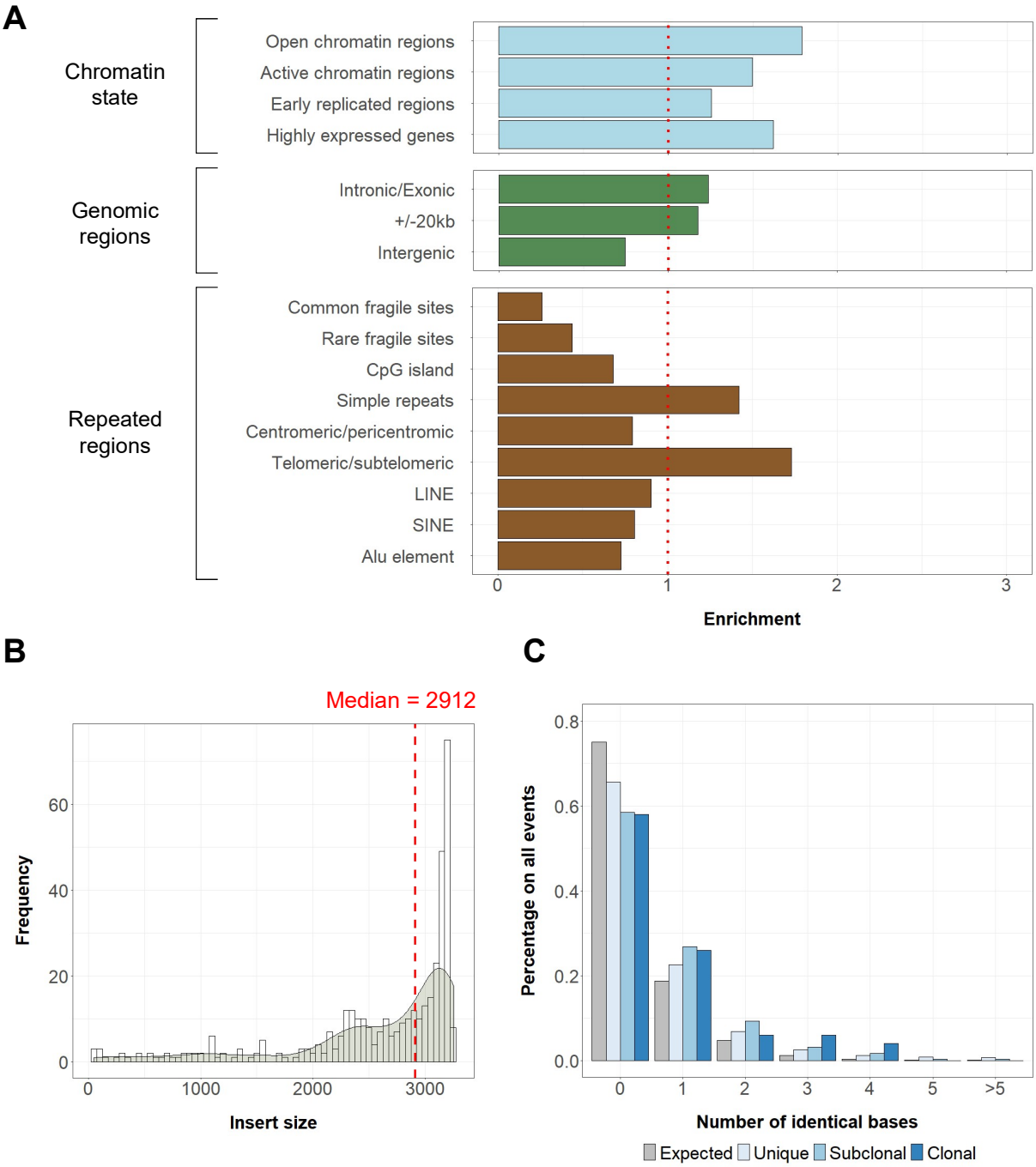
Supplementary figure 2 | Flow-chart of the HBV integrations analysis pipeline. The different steps are described in **Supplementary Materials and Methods**, with the references of the tools used. The list of the 8809 integration events obtained after filtering is detailed in **Supplementary table 4**.



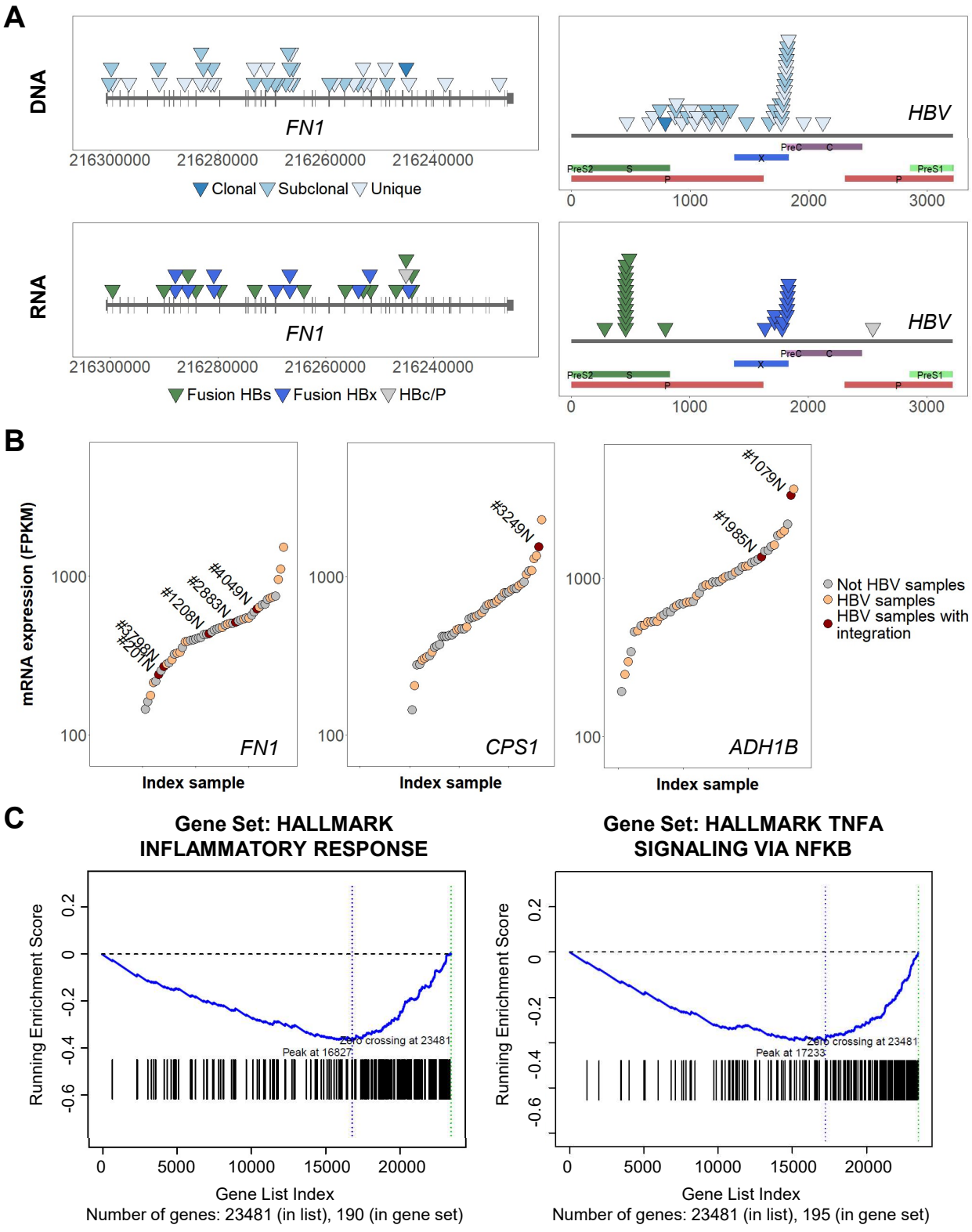
Supplementary figure 3 | Definition of clonality from viral capture. (A) Correlation between the values of HBV copy number per cell obtained from qPCR and viral capture in 347 samples from the Capture series of 190 patients (non-tumor liver, n=170, and tumor samples, n=177). The HBV status is annotated based on clinical data. The two techniques show a good correlation (Pearson correlation). (B) Definition of clonality with all 2199 integration events detected in tumors based on the k-means method. (C) Correlation between the values of copy number per cell for integration events obtained from WGS and from viral capture in 40 samples (non-tumor liver, n=20, and tumor samples, n=20). The two techniques show a good correlation (Pearson correlation). (D) Distribution of integration events detected in viral capture in 40 samples sequenced in WGS, based on their copy number per cell.



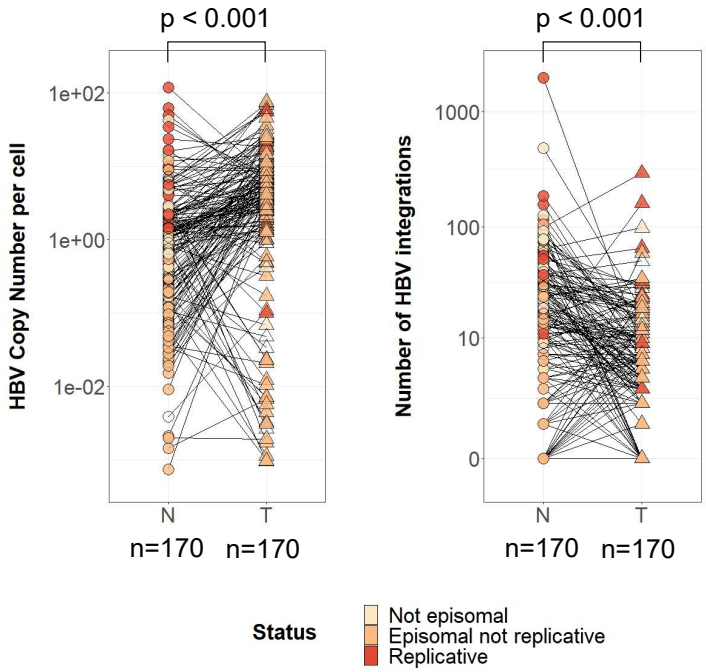
Supplementary figure 4 | HBV genotypes in non-tumor tissues. Correlation plot between the HBV genotypes identified from viral capture data and clinical or molecular features of non-tumor samples from the capture series (n=170). Blue circles indicate a negative association between features; red circles indicate positive associations. Color intensities represent different levels of statistical significance. Statistical analysis was performed using the chi-square test or Wilcoxon signed-rank test with respect to the type of variable. P values were adjusted for multiple testing using the Benjamini-Hochberg method (false discovery rate). A pie chart showing the distribution of HBV genotypes within the series is represented on the left side.



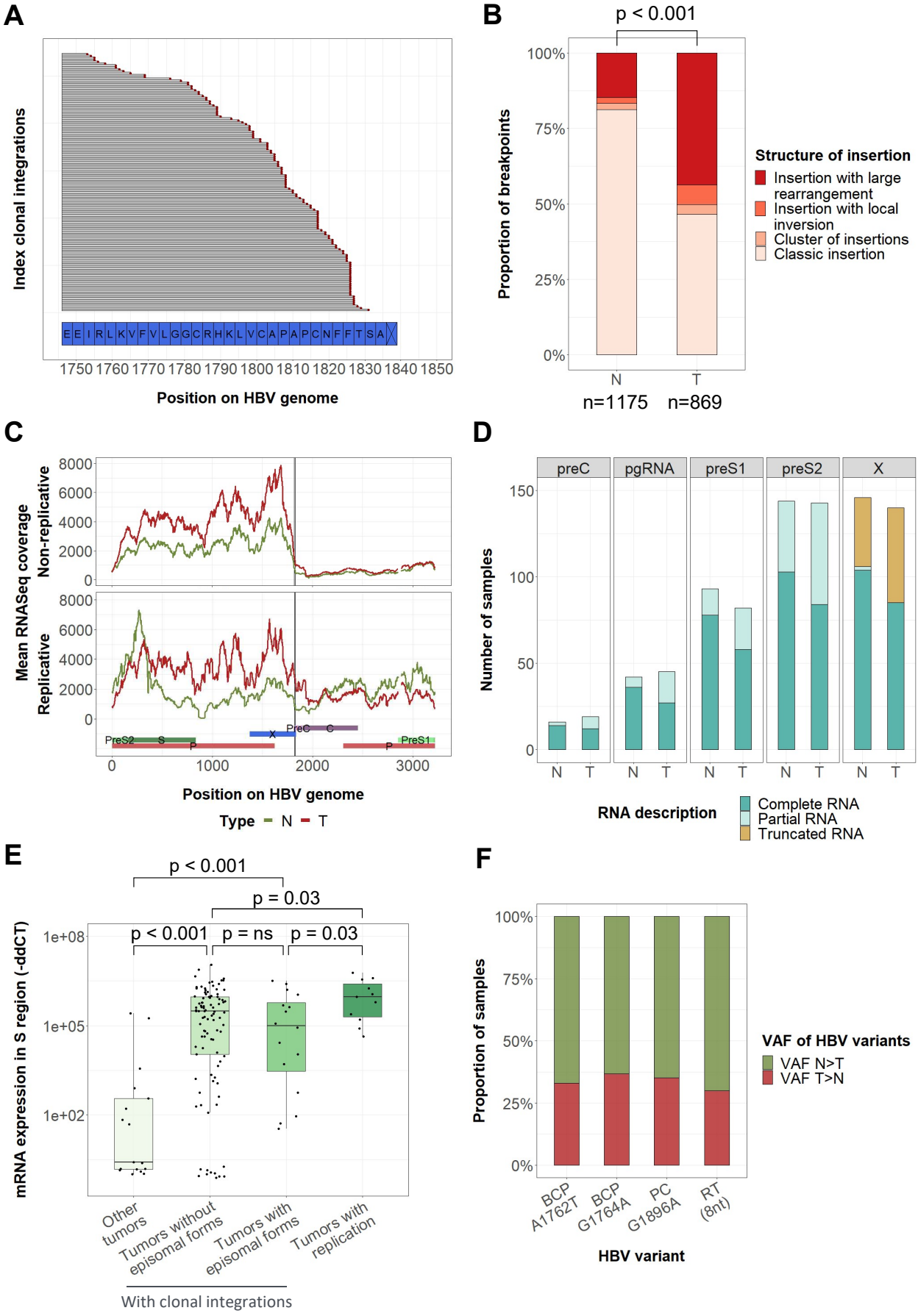
Supplementary figure 5 | Characterization of localization and sequences of HBV integrations in non-tumor samples. (A) Density of HBV integration breakpoints in a specific region compared to the complementary region. Regions were defined by chromatin state, genomic features and repeated motifs. (B) Size distribution of integrated sequences in simple subclonal or clonal integrations (n=394) identified in non-tumor tissues from the Capture series. (C) Homology analysis at HBV integration breakpoints between human genomic sequences and HBV integrated sequences, according to the clonality of the events.



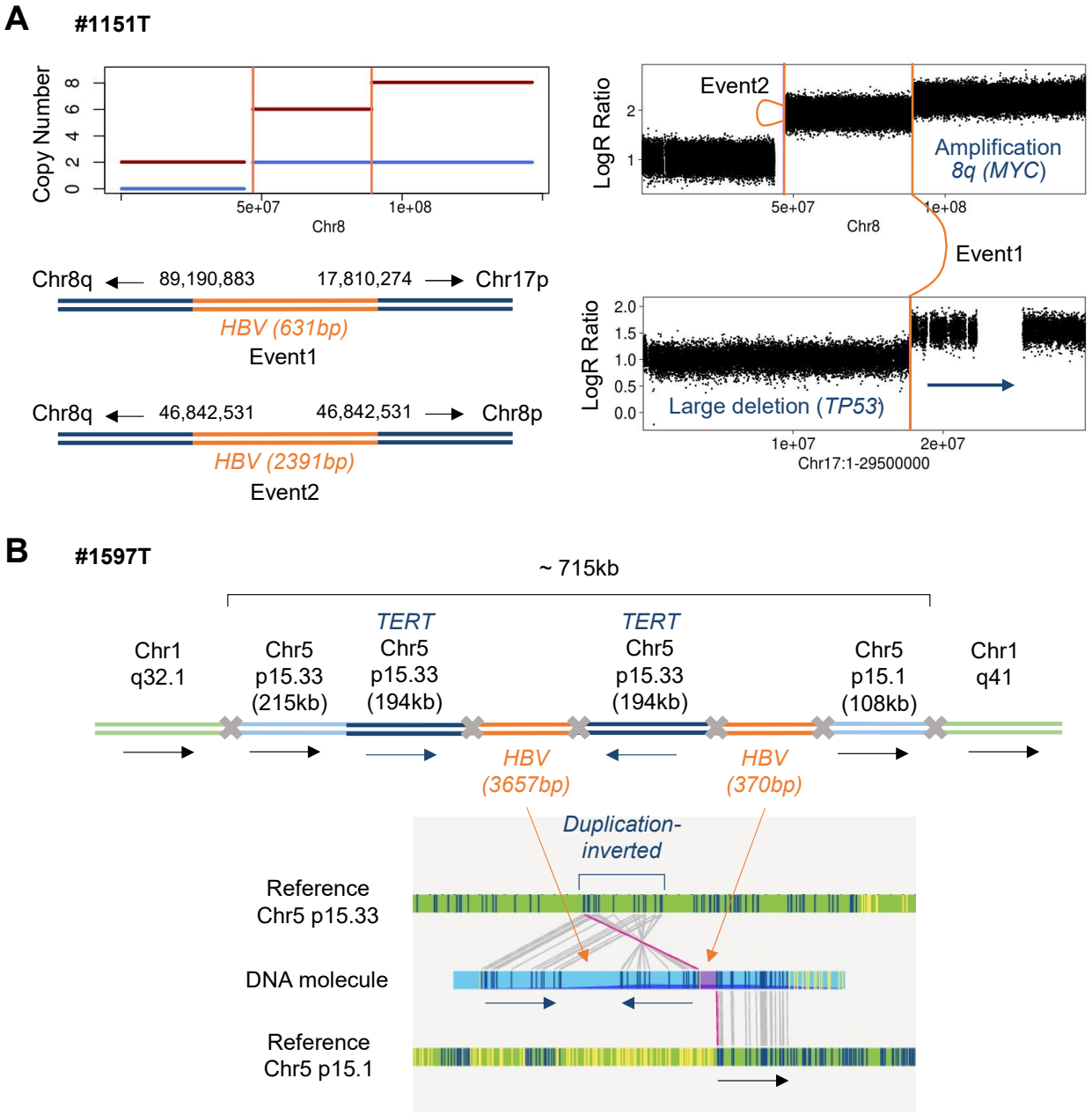
Supplementary figure 6 | Subclonal/clonal expansion in non-tumor tissues. (A) Breakpoints of HBV integrations in human genome at the *FN1* locus (left) and in HBV genome (right). Coordinates of breakpoints of HBV integrations in genomic DNA were identified from viral capture (up) and in transcripts from RNA-Seq (down). (B) mRNA expression for the 3 genes with integrations in more than two samples: *FN1*, *CPS1* and *ADH1B*, for 24 HBV-positive and 29 HBV-negative non-tumor liver samples. No RNAseq data was available for samples with integrations in *KCTN2*. (C) Gene-set enrichment analysis from RNA-Seq data, to compare HBV-positive non-tumor samples with a clonal HBV integration (n=5) or without (n=19). Two gene sets are shown to be downregulated in the first group: one gene set containing 190 genes defining inflammatory response (left) and one gene set containing 195 genes regulated by NF-κB in response to TNF (right).



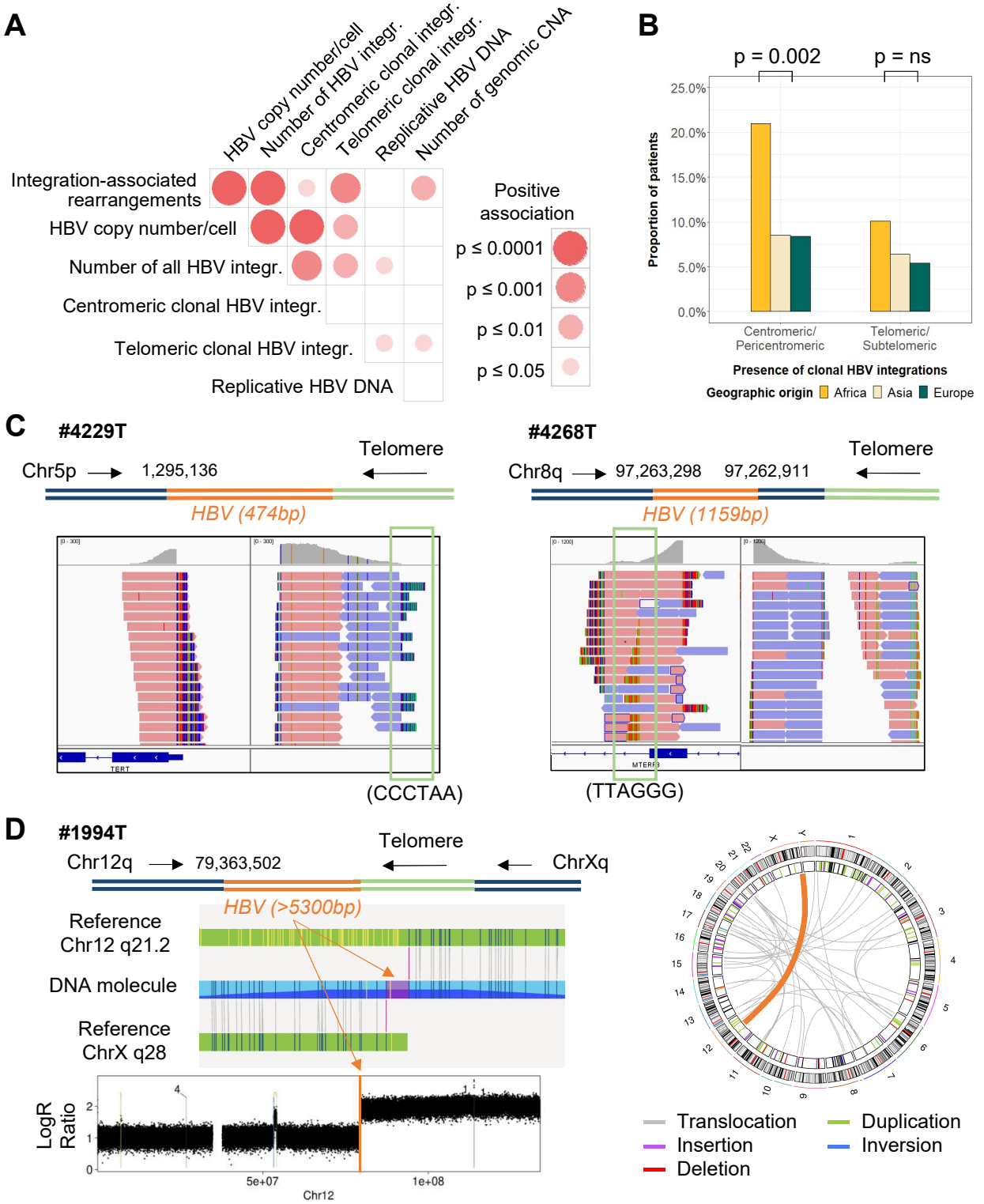
Supplementary figure 7 | Comparison of non-tumor and tumor samples. HBV copy number per cell (left) and number of HBV integration breakpoints (right) of paired tumor and non-tumor tissues of 170 HBV-positive patients from the Capture series (paired Wilcoxon signed-rank test).


A
Index clonal integrations
E E I R L K V F V L G G C R H K L V C A P A P C N F F T S A X
1750 1760 1770 1780 1790 1800 1810 1820 1830 1840 1850
Position on HBV genome
B
p < 0.001
Proportion of breakpoints
100%
75%
50%
25%
0%
N
T
n=1175 n=869
Structure of insertion
Insertion with large rearrangement
Insertion with local inversion
Cluster of insertions
Classic insertion
C
Mean RNASeq coverage
Non-replicative
Replicative
8000
6000
4000
2000
0
PreS2 S P X PreC C P PreS1
0 1000 2000 3000
Position on HBV genome
Type N T
D
preC pgRNA preS1 preS2 X
Number of samples
150
100
50
0
N T N T N T N T N T
RNA description
Complete RNA
Partial RNA
Truncated RNA
E
p < 0.001
p = 0.03
p < 0.001 p = ns p = 0.03
mRNA expression in S region (-ddCT)
1e+08
1e+05
1e+02
Other tumors
Tumors without episomal forms
Tumors with episomal forms
Tumors with replication
With clonal integrations
F
Proportion of samples
100%
75%
50%
25%
0%
BCP A1762T
BCP G1764A
PC G1896A
RT (8nt)
HBV variant
VAF of HBV variants
VAF N>T
VAF T>N

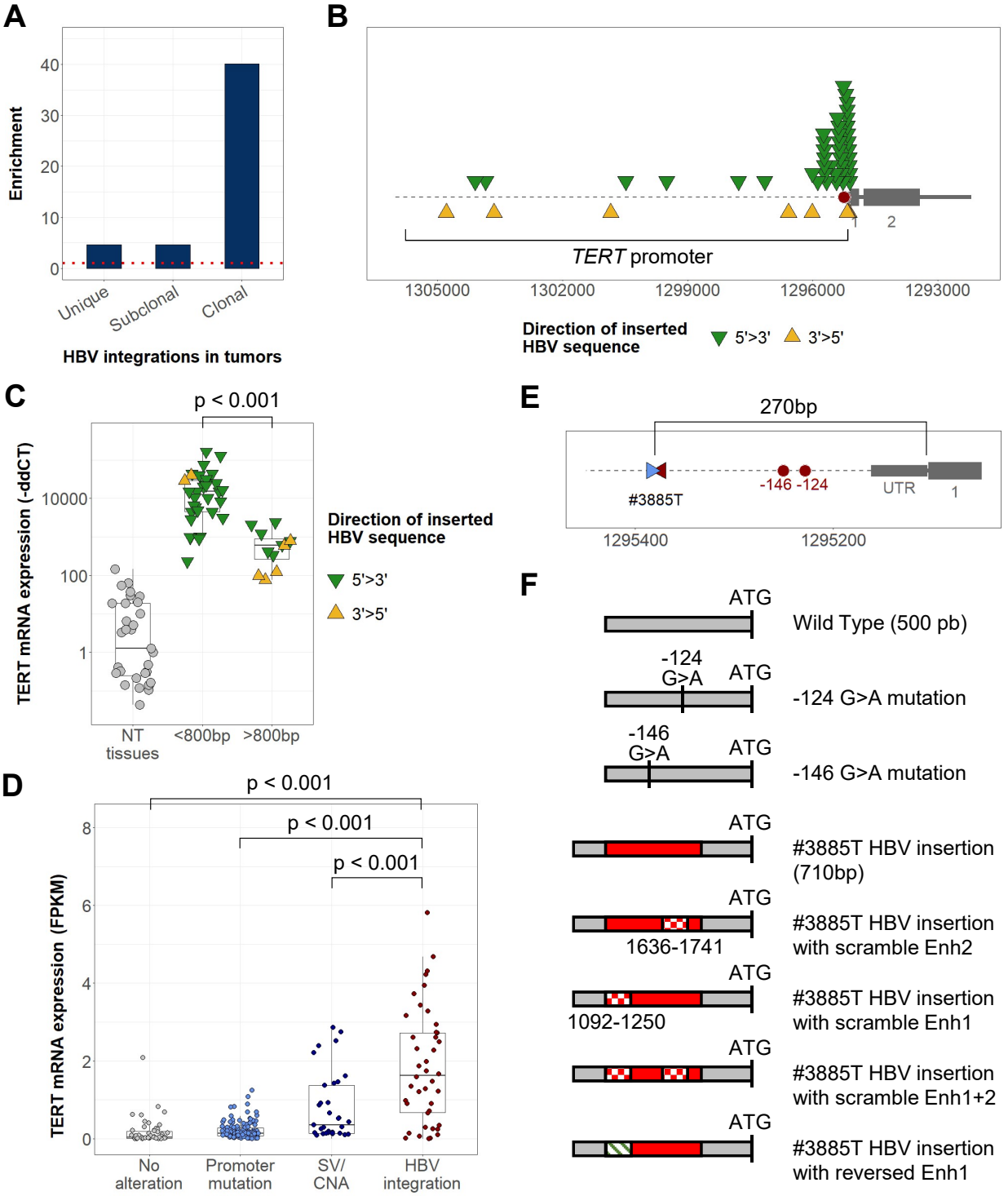
Supplementary figure 8 | Differences in HBV integrated sequences between non-tumor tissues and tumors. (A) HBV integrated sequences from 136 clonal events in tumors harboring a HBV/HG breakpoint around the 3' extremity of HBx (1750-1850) and with a 3'>5' orientation. Each line represents a different clonal integration and the breakpoints between viral and human genomes are represented with red dots. The last 30 amino acids of HBx protein are annotated below the graph. (B) Distribution of breakpoints in non-tumor and tumor tissues according to the structure of the clonal/subclonal integration (see **Materials and Methods** for definitions) (Chi-square test). (C) Mean HBV coverage from RNA-Seq of non-tumor samples (n=22) and tumors (n=73) according to the presence of replicative HBV DNA. (D) Number of samples positive for HBV mRNA according to the localization of the HBV probes. RNA are annotated as complete if a sample is positive for all probes until the polyA region on the HBV genome. (E) HBV mRNA expression (-ddCT) in the S region of HBV genome in tumors (Wilcoxon signed-rank test). (F) Proportion of non-tumor and tumor samples with HBV variants at different positions of the HBV genome. Variant Allele Frequency was determined in paired HCC and adjacent non-tumor tissue for 57 to 98 patients according to the position coverage on the HBV genome. *VAF, Variant Allele Frequency; BCP, Basal Core Promoter; PC, PreCore; RT, Reverse-Transcriptase; ns, not significant.*



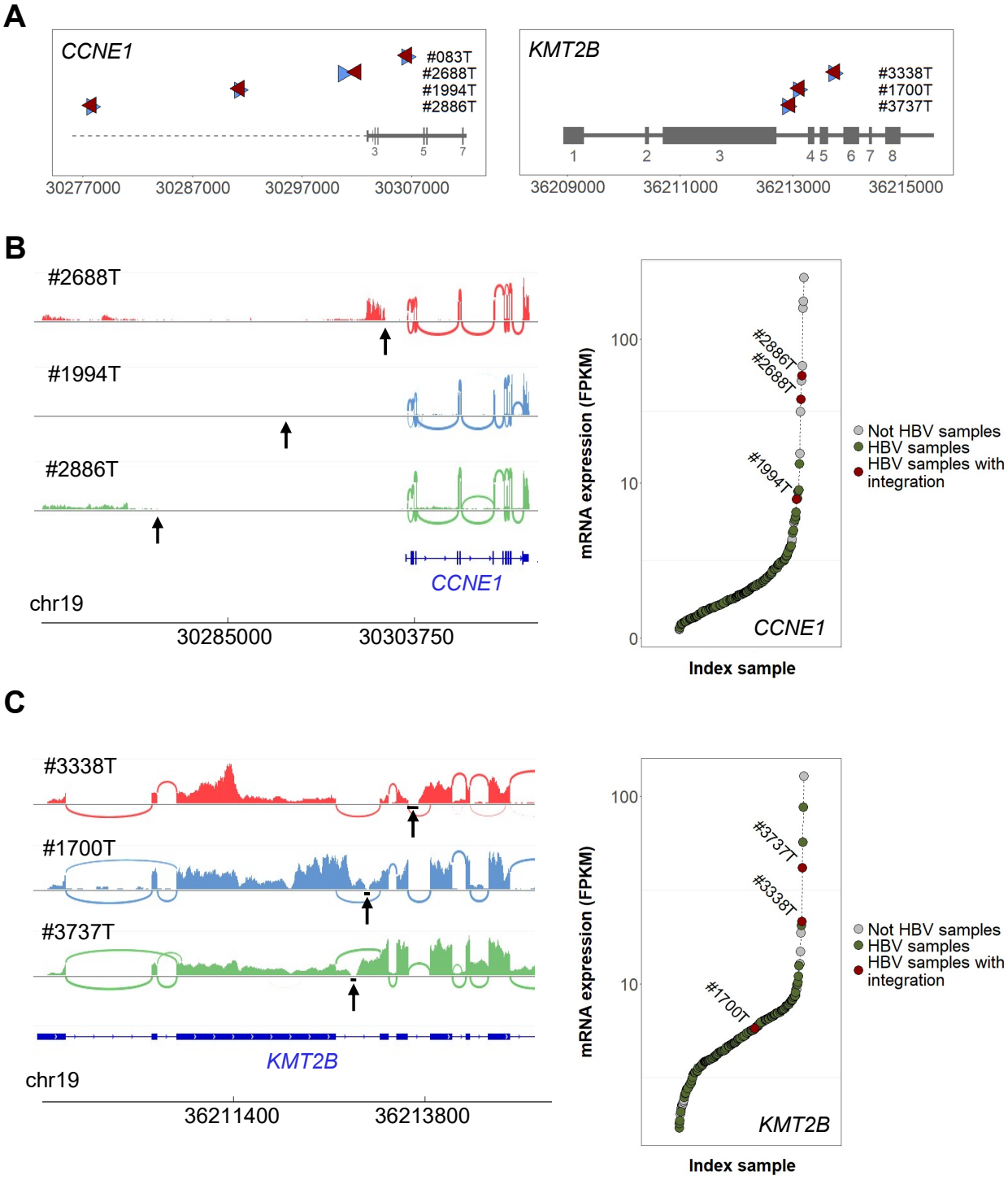
Supplementary figure 9 | Complex rearrangements involving integrations reconstructed with long-read and Bionano sequencing. (A) Reconstruction with long-read sequencing of two clonal integration events in tumor #1151T located in chr8q and inducing copy number alterations. Profiles of copy number (total in red, minor allele in blue) and LogR ratio are represented. Event 1 is a translocation-like event between chr8q and chr17p. Event 2 is a duplication-inverted-like event in the centromere of chr8. (B) Reconstruction with Bionano whole-genome sequencing of a complex rearrangement in tumor #1597T. The rearrangement is composed of two translocations between chr5p and chr1q, and two HBV clonal integrations inducing an amplification of *TERT*.



Supplementary figure 10 | HBV insertions in centromeric/telomeric regions. (A) Correlation plots between the presence of integration-associated rearrangements, viral features and the number of genomic CNA. All associations are positive and color intensities represent different levels of statistical significance. Statistical analysis was performed using the chi-square test, Wilcoxon signed-rank test or Pearson correlation with respect to the type of variable. P values were adjusted for multiple testing using the Benjamini-Hochberg method (false discovery rate). (B) Proportion of HCC according to the patients' geographic origin harboring HBV integrations in centromeric or telomeric regions (Fisher test). (C) Reconstruction of HBV clonal integration events involving telomeric regions in tumors #4229T and #4268T. Views from Integrative Genomics Viewer are shown to visualize telomeric sequences (green square). (D) Reconstruction with Bionano whole-genome sequencing of a translocation in tumor #1994T between chr12q and the telomeric region of chrXq. The logR Ration in chr12q is shown to see the CNA at HBV integration breakpoint. *CNA, Copy Number Alteration; ns, not significant.*



Supplementary figure 11 | HBV insertional mutagenesis: *TERT* activation. (A) Enrichment of HBV integration breakpoints in tumors around 72 HCC-associated genes (+/- 25kb). The HCC-associated genes were selected as previously described[45]. (B) Localization of HBV clonal integration breakpoints in the human genome in the promoter of the *TERT* gene. Red dots indicate the positions of classic mutations of the promoter (-124 and -146 from ATG) (C) mRNA expression (-ddCT from RT-qPCR) of *TERT* in HCC harboring a clonal HBV integration in the *TERT* promoter (n=48) and in the adjacent non-tumor liver samples (n=31) (Wilcoxon signed-rank test). (D) mRNA expression (log10(1+FPKM) from RNA-Seq data) of TERT in HCC harboring different alterations of the TERT promoter in the NGS Series (n=265) (Wilcoxon signed-rank test). (E) Localization of HBV clonal integration breakpoints in the promoter of the *TERT* gene for tumor #3885T. Red dots indicate the positions of mutations of the promoter (-124 and -146 from ATG). (F) Representation of plasmids constructs of TERT promoter (wild type, harboring -124 or -146 promoter mutations, harboring the HBV insertion identified in tumor #3885T wild type or containing scrambled/reversed Enhancer sequences). *SV, Structural Variant; CNA, Copy Number Alteration; Enh, Enhancer.*



Supplementary figure 12 | HBV insertional mutagenesis: *CCNE1* and *KMT2B*. (A) Localization of HBV clonal integration breakpoints in the human genome at the loci of *CCNE1* gene (left) and *KMT2B* gene (right). HBV integration in tumor #1994T have been previously described[21]. (B) Consequences of HBV integrations on *CCNE1* mRNA in 3 tumors. IGV-Sashimi plots show RNA-seq alignments (left) and sorted mRNA expression (FPKM) from RNA-Seq data of the NGS series (n=265) is represented (right). (C) Consequences of HBV integrations on *KMT2B* mRNA in 3 tumors. IGV-Sashimi plots show RNA-seq alignments (left) and sorted mRNA expression (FPKM) from RNA-Seq data of the NGS series (n=265) is represented (right). For IGV-Sashimi plots, alignments in exons are represented as read density, and alignments to splice junctions are shown as an arc connecting a pair of exons, where arc width is proportional to the number of reads aligning to the junction. The canonical transcripts for the genes are shown below. Black arrows indicate the position of HBV integration breakpoints. *FPKM, fragments per kilobase of exons per million reads.*