Adeno-associated virus in the liver: natural history and consequences in tumor development

Tiziana La Bella, Sandrine Imbeaud, Camille Péneau, Iadh Mami, Shalini Datta, Quentin Bayard, Stefano Caruso, Theo Hirsch, Julien Calderaro, Guillaume Morcrette, et al.

Adeno-associated virus in the liver: natural history and consequences in tumor development

Tiziana La Bella[1,2*], Sandrine Imbeaud[1,2*], Camille Peneau[1,2], Iadh Mami[1,2], Shalini Datta[1,2], Quentin Bayard[1,2], Stefano Caruso[1,2], Theo Z Hirsch[1,2], Julien Calderaro[1-3], Guillaume Morcrette[1,2,4], Catherine Guettier[4], Valérie Paradis[5], Giuliana Amaddeo[6], Alexis Laurent[7], Laurent Possenti[8], Laurence Chiche[9], Paulette Bioulac-Sage[10,11], Jean-Frédéric Blanc[10,12], Eric Letouzé[1,2], Jean-Charles Nault[1,2,13], Jessica Zucman-Rossi[1,2,14.]
* These authors contributed equally to this work.

1. Centre de Recherche des Cordeliers, Sorbonne Universités, Inserm, UMRS-1138, F-75006 Paris, France.
2. Functional Genomics of Solid Tumors, Université de Paris, Université Paris 13, Labex Immuno-Oncology, équipe labellisée Ligue Contre le Cancer, Paris, France.
3. APHP, Department of Pathology, CHU Henri Mondor, Créteil, France
4. Service d'anatomopathologie, Hôpitaux Paul Brousse et Bicêtre, Le Kremlin Bicêtre, InsemU1193 Université Paris-Sud, Orsay, France.
5. Service d'Anatomopathologie, Hôpital Beaujon, Clichy, France.
6. APHP, Service d'Hépatologie, Groupe Hospitalier Henri Mondor,Inserm U955, Université Paris-Est Créteil, France.
7. APHP, Service de Chirurgie Digestive et Hépatobiliaire, Groupe Hospitalier Henri Mondor, Créteil, France.
8. CHU Bordeaux, Department of Hepato-Gastroenterology and Digestive Oncology, Hôpital Haut-Lévêque,Bordeaux, France.
9. CHU Bordeaux, Department of Digestive Surgery, Centre Médico Chirurgical Magellan, Haut-Lévêque Hospital, Pessac, France.
10. Université Bordeaux, Bordeaux Research in Translational Oncology, Bordeaux, France
11. Service de Pathologie, Hôpital Pellegrin, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France.
12. CHU Bordeaux, Service Hépato-Gastroentérologie et Oncologie Digestive, Hôpital Haut-Lévêque, Bordeaux, France.
13. APHP, Hôpitaux Universitaires Paris–Seine Saint-Denis, Site Jean Verdier, Pôle d'Activité Cancérologique Spécialisée, Service d'Hépatologie, Bondy, France.
14. APHP, Hopital Europeen Georges Pompidou, Department of Oncology, Paris, France.

Correspondance:
Jessica Zucman-Rossi; MD, PhD
INSERM, Centre de Recherche des Cordeliers
Génomique fonctionnelle des tumeurs solides
15 rue de l'école de médecine, 75006 Paris, France
Tel : 33153725166
Email: Jessica.zucman@gmail.com

Word Count : 3978

Abbreviation : AAV, Adeno-associated virus; AdV, adenoviruse; AAP, assembly-activating protein; BWA, Burrows-Wheeler Aligner; CCK, cholangiocarcinoma; CNA, copy number alteration; FLC, fibrolamellar carcinoma; FNH, focal nodular hyperplasia; HBV, hepatitis B virus; HCV, hepatitis C virus; HB/TLCT, hepatoblastoma or transitional tumor; HCA, hepatocellular adenoma; HCC, hepatocellular carcinoma; HHV1 or HSV1, herpes simplex virus type 1; HHV2 or HSV2, herpes simplex virus type 2; HHV4 or EBV, Epstein-Barr virus; HHV5 or CMV, Cytomegalovirus; HHV8 or KSHV, Kaposi's sarcoma-associated herpesvirus; HHV6, human herpesvirus type 6; HHV7, Human betaherpesvirus 7; HHV, human herpesviruses; HPV16, human papillomavirus type 16; HVR, hypervariable region; IGV, Integrative Genomics Viewer ; IRB, institutional review board; ITR, inverse tandem repeat; NT, non-tumor; rAAV, recombinant AAV; shHCA, qRT-PCR, quantitative Real-Time Polymerase Chain Reaction; sonic hedgehog molecular adenoma; TFBS, transcription factor binding site; TSS, transcription start site; T, tumor.

Key words: Adeno Associated Virus, AAV2, AAV13, hepatocellular carcinoma, hepatocellular adenoma, cyclin, telomerase, insertional mutagenesis.

Abstract

Objective: Adeno-associated virus (AAV) is a defective mono-stranded DNA virus, endemic in human population (35-80%). Recurrent clonal AAV2 insertions are associated with the pathogenesis of rare human hepatocellular carcinoma (HCC) developed on normal liver. This study aimed to characterize the natural history of AAV infection in the liver and its consequence in tumor development.

Design: Viral DNA was quantified in tumor and non-tumor liver tissues of 1461 patients. Presence of episomal form and viral mRNA expression were analyzed using a DNAse/TaqMan based assay and quantitative RT-PCR. *In silico* analyses using viral capture data explored viral variants and new clonal insertions.

Results: AAV DNA was detected in 21% of the patients, including 8% of the tumor tissues, equally distributed in 2 major viral subtypes: one similar to AAV2, the other hybrid between AAV2 and AAV13 sequences. Episomal viral forms were found in 4% of the non-tumor tissues, frequently associated with viral RNA expression and human herpesvirus type 6 (HHV6), the candidate natural AAV helper virus. In 30 HCC, clonal AAV insertions were recurrently identified in *CCNA2*, *CCNE1*, *TERT*, *TNFSF10*, *KMT2B* and *GLI1*/*INHBE*. AAV insertion triggered oncogenic overexpression through multiple mechanisms that differ according to the localization of the integration site.

Conclusion: We provided an integrated analysis of the wild type AAV infection in the liver with the identification of viral genotypes, molecular forms, helper virus relationship and viral integrations. Clonal AAV insertions were positive selected during HCC development on non-cirrhotic liver challenging the notion of AAV as a non-pathogenic virus.

Significance of this study

What is already known about this subject?

- The seroprevalence of AAV in general population is 40-80% and AAV2 is the most frequent serotype in human.
- AAV has a biphasic life cycle characterized by latent and lytic phases. The presence of a helper virus is required for the AAV replication.
- It is commonly believed that Adenovirus is the natural AAV helper virus.
- Although AAV is considered a non-pathogenic virus, recurrent clonal AAV2 insertions were associated with HCC development.

What are the new findings?

- Two viral subtypes are present in 21% of the liver tissues: AAV2 and hybrid AAV2/13 sequences.
- Episomal AAV forms are found in 4% of non-tumor liver tissues, mainly in young, female patients without liver fibrosis.
- Human herpesvirus type 6 (HHV6) is the most frequent AAV helper virus in the liver.
- The 2% of HCC patients displayed clonal AAV integration in cancer driver genes.
- AAV clonal insertion in HCC activates oncogenes using various mechanisms.

How might it impact on clinical practice in the foreseeable future?

- These findings are important to understand wild type AAV biology and its association with hepatocarcinogenesis. Our data are particularly relevant considering the large usage of AAV vector in liver-targeted gene therapy.
- Even if rare, AAV insertional mutagenesis is a new risk factor of HCC development, therefore the notion of AAV as non-pathogenic virus should be reviewed.

INTRODUCTION

Adeno-associated virus (AAV) is a small non-enveloped DNA virus with an icosahedral capsid that contains a 4.7 kb linear single-stranded genome.[1, 2] AAV genome codes for non-structural proteins (Rep78, 68, 52 and 40), capsid proteins (VP1, VP2, VP3) and the assembly-activating protein (AAP).[3, 4] At the extremities, inverse tandem repeats (ITR) are important for the integration in host genome.[5, 6] AAV is a defective virus that requires a helper virus for an active infection, otherwise it can establish a latent infection through integration into host genome or maintenance as circular episomal form.[7-9] AAV seroprevalence showed that the infection is endemic in human populations (30-80%) starting during childhood.[10-12] Twelve distinct serotypes and more than 100 natural variants have been identified, among which AAV2 is the most frequent type in human.[13-16]

This small virus is attractive for gene therapy because of the lack of identifiable associated disease and the remarkable ability of recombinant AAV (rAAV) vectors to transduce dividing and non-dividing cells with high efficiency, long-term transgene expression, low immunogenicity and specific tissue tropism.[17] Although AAV was discovered in 1965, many questions regarding the natural history of AAV infection in human remain unanswered.[2] It is well known that the vector predominantly persists in the nucleus as episomal form with sustained RNA expression raising question on putative episomal AAV form in wild type infection.[8] Several helper viruses have been identified but their precise association with wild type liver AAV infection remain unclear. The frequency of the different AAV genotypes in the human population and AAV persistence in tissues after first infection remains to be determined.[18] Moreover, AAV link with tumor development is controversial, with some studies reporting an oncogenic effect of AAV infection in animal model and others suggesting a tumor suppressive role.[19-24]

Recently, we reported the involvement of AAV2 in the pathogenesis of human hepatocellular carcinoma (HCC) developed on normal liver in the absence of classical HCC risk factors such as infection with hepatitis B or C viruses (HBV and HCV), high alcohol intake, hemochromatosis or aflatoxin B1 exposure.[25] Similarly to HBV, recurrent AAV2 clonal insertions were described in *TERT*, *CCNE1* and *CCNA2* cancer driver genes, leading to their overexpression.[25-28] The AAV insertions can activate oncogenes located nearby in the human genome by a liver promoter recently identified within the minimal common AAV inserted sequence adjacent to the 3'ITR of the virus.[29]

In this work, we investigated the natural history of wild type AAV infection in the liver and its consequences in tumor development in a large cohort of 1461 patients with benign or malignant liver tumors.

MATERIALS AND METHODS

Patients and tissue samples

A series of 1461 patients was included in the study approved by our local 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). Liver tissues were frozen immediately after surgery in French hospitals. Tumor and non-tumor counterparts were analyzed in 1269 patients, only the tumor or non-tumor tissues were investigated for 138 and 54 patients respectively. The present series included HCC (n=936), hepatocellular adenomas (HCA, n=225), focal nodular hyperplasia (FNH, n=97), hepatoblastoma or transitional tumors (HB/TLCT, n=87), cholangiocarcinoma (CCK, n=46), fibrolamellar carcinoma (FLC, n=36) and other tumors (n=34, Supplementary table 1).

Viral DNA screening

Genomic DNA were analyzed for the presence of viral DNA by quantitative PCR on Fluidigm 96.96 dynamic arrays using the BioMark Real-Time PCR system with TaqMan probe sets designed with Primer3Plus software (Supplementary figure 1A and supplementary table 2). Results were analyzed using the Fluidigm Real-Time PCR Analysis software (4.1.3) and reported to a reference gene, HMBS. The quantification was expressed in viral copy number/cell. Copy number/cell values were tested for unimodal and bimodal distribution using normalmixEM function of *mixtools* package in R.[30]

Isolation of human AAV using viral capture sequencing

Viral capture of genomic DNA was performed for tumor and matched normal sample, sequence as previously described using 120-mer primers recognizing all AAV genotypes 1 to 13 already described with around 305 probes/genotype.[25] Viral reads were mapped to all AAV1- to 13 reference sequences using BWA (Burrows-Wheeler Aligner, version 0.7.15).[31] The number of AAV reads correlates with the number of viral copies/cell (Supplementary figure 1B). Read pairs with at least one read aligned on the virus were extracted using samtools (v1.3),[32] and aligned to a custom reference genome including human chromosomes and virus sequences. We calculated the number of reads mapping the AAV/human chimeric and mate regions in each samples by generating a 20k-bin size bed for hg19 genome which was used for computations with bedtools multicov utility.[33] For each bin, we calculated the mean of coverage in the samples displayed in a pan genomic plot. We used chimeric reads to identify insertion breakpoints at base resolution by mapping sequences on both sides of the junctions. Clonal events were considered when more than 25 reads overlapped the same locus, putative subclonal insertions when 4 to 24 overlapping reads were identified. All viral insertions were validated by visual inspection on IGV (Integrative Genomics Viewer). Sequences have been deposited in the Genbank database MK231253 to MK231264 and KT258720 to KT258730.

The analysis of full-length human-AAV sequences is detailed in Supplementary Material and Methods. Sequences have been deposited in the Genbank database MK139243 to MK139299 and MK163929 to MK163942.

RNAseq

Samples enriched in poly(A)+ RNA were sequenced using Illumina® TruSeq Stranded mRNA kit on HiSeq2000 sequencer, yielding approximately 45 million 100-base-pair (bp) paired-end reads (IntegraGen, Evry).[34] Reads were aligned and chimeric sequences reconstructed with TopHat2[35] and Cufflinks v2.2.1.[36] We used ElemeNT[37] to predict transcription start sites (TSS), Alamut Visual software (Interactive Biosoftware) to identify splicing signals on the chimeric DNA sequence, ATGpr[38] to identify translation initiation sites and Poly(A) Signal Miner to identify PolyA sites.[39] Sequences were deposited in EGA database (EGAS00001002879, EGAS00001001284 and EGAS00001003310).

Detection of viral episomal form

A specific DNAse/TaqMan based assay was adapted from Werle-Lapostolle's protocol[40] to detect AAV episomal form (detailed procedures in Supplementary Material and Methods). Junctions of the circular AAV were amplified using 2 couples of primers surrounding the ITRs (Supplementary table 2) in 2.5% glycerol and 5% DMSO. PCR products were sequenced by Sanger after ExoSAP-IT (Applied Biosystem) purification.[41]

Quantitative RT-PCR

AAV mRNA and inserted target genes expressions were analyzed using quantitative Real-Time Polymerase Chain Reaction (qRT-PCR). Specifically, we used 7 AAV custom made and human catalog TaqMan probes (Supplementary table 2) with AB7900HT PCR System (Applied Biosystem) and 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.

Site Directed Mutagenesis

The role of the viral poly-A signal in AAV-induced gene over-expression was investigated in 2 plasmids containing AAV insertions in the 3'UTR of *TNFSF10*.[25] QuikChange Lightning site-directed mutagenesis kit (Agilent) was used to introduce 4 point mutations in the viral poly-A signal (NC_001401: 4424 A>C, 4426 T>G, 4427 A>C, 4429 A>C). All mutations were verified using Sanger sequencing.

Cell Culture, transfection and dual luciferase assay

HuH7, HepG2 and HuH6 cells were purchased from ATCC and cultured in DMEM supplemented with 10% FBS and 100 U/mL penicillin/streptomycin. Cells were tested for mycoplasma contamination. Identity was verified by exome sequencing. Cells were transfected using Lipofectamine 3000 (Life Technologies) with pmirGLO plasmid (Promega) containing wild-type *TNFSF10* 3'UTR, the 3'UTR with AAV2 insertions or scrambled AAV2 sequence downstream a luciferase reporter gene. Luminescence from firefly luciferase was normalized on the corresponding renilla luciferase activity. Fold change was calculated relative to the wild type *TNFSF10* 3'UTR construct.

Statistical analysis

Statistical analyses were performed using RStudio (v1.0.136) and GraphPad Prism (v6.0a). Relationship between AAV and clinical, histological features of the patients was investigated using Chi-square test. *P*-values adjustment was computed for a Monte Carlo test with 2,000 permutations. Statistical significance of quantitative variable was determined by Wilcoxon test. Association among variables was modelled by a multinomial logistic regression. Luciferase activity of transfected vs control cells was compared using Student's t-test. All tests were 2-tailed and a *P*-value < 0.05 was considered as significant.

RESULTS

Identification of two major AAV genotypes in the liver

Screening of frozen liver tissues from 1,319 patients with 6 Taqman probes distributed along the genome that collectively recognize all AAV genotypes 1 to 13 identified AAV DNA in 18% (n=233) of non-tumor liver tissues (Supplementary figure 1). For viral AAV DNA capture of all known genotypes 1 to 13, we selected 80 non-tumor liver samples including 68 positive samples ranging from $2x10^{-4}$ to 0.18 copy number/ cell. After sequencing, a full-length AAV sequence was reconstructed in 57 samples and two major AAV subtypes were identified (Figure 1A-B). The first subtype (n=25) is highly similar to AAV2 reference sequence (NC_001401) and to VP1 Clade B genotype isolated in human[14, 42] (Supplementary figure 2). The second subtype (n=32), showed hybrid sequences including various parts of the AAV13 capsid (similar to Clade C[14, 42]) and c-ter in the context of an AAV2 5'part, it was named AAV2/13 (Figure 1B and Supplementary figure 2). We identified along the viral genome 42 silent variants shared by both AAV subtypes, but different from the AAV2 reference NC_001401 (Figure 1C). In contrast, several nucleotide variants leading to amino acid substitutions in AAV2/13 sequences were located in the hypervariable regions (HVRs) 5, 6 7 and 10 and originated from AAV13 sequence (Figure 1B-C). Screening the overall series of 1,319 samples with two probes specific of AAV2/13 subtype and located in the CAP2 region (Supplementary figure 1), identified 47.6% AAV2 and 52.4% AAV2/13 genotypes among 143 samples positive for the variable region.

AAV infection and episomal form

In the 233 AAV positive liver samples, quantification of the viral DNA showed a bimodal distribution: 97% of the tissues exhibited a low number of copy/cell (ranging from $4.6x10^{-5}$ to 0.04) and only 8 patients showed a higher quantity of AAV ranging from 0.07 to 0.18 copy/cell (Figure 2A). AAV was significantly enriched in female ($p<0.001$), young patients ($p=0.016$) and occurred more frequently in a background of non-fibrotic liver ($p<0.001$; Figure 2B).
In 64/233 (27.5%) of the tissues positive for AAV, all the genomic AAV regions were amplified suggesting the presence of the entire viral genome. We designed a DNAse/TaqMan based assay (Supplementary figure 3A), which allowed to detect episomal AAV in 60 patients, corresponding to 26% of AAV positive samples and 4.6% of all patients. Using *in silico* analyses of the AAV capture sequencing, among the 57 cases with a complete reconstructed AAV genomic sequence, we identified 14 cases with 3'ITR-5'ITR junctions. Circularized concatemeric structures may escape from our experimental method to identify episomal form,[43] however, we did not identified insertion of concatemer *in silico*. The 3'ITR-5'ITR junctions showed various sequences presenting a double-D ITR structure, in flip or flop configuration, with a 125bp deletion confirmed by Sanger sequencing (Supplementary figure 3C-D and 4).

AAV transcription is associated with episomal form

Then, we screened for AAV RNA expression in 101 non-tumor liver tissues positive for AAV by qRT-PCR. AAV transcript was identified in 64% of the tested liver tissues. Either AAV

REP or CAP expression were enriched in liver tissues with episomal form ($p<0.001$) and both transcripts were more frequently associated in presence of episomal than not-episomal AAV form ($p=0.022$), defining a population of patients with an "episomal-expressed AAV" (Figure 2C). A higher AAV copies per cell was identified in liver tissues with episomal-expressed AAV, supporting the hypothesis of a viral active infection in these liver samples (Figure 2D). Episomal AAV were also more frequent in female ($p<0.001$) and non-cirrhotic patients ($p<0.001$; Supplementary figure 5A). Analysis of AAV positivity in function of age showed a peak of frequency at 25% in the 30-40 years class. AAV episomal form was more frequent in young patients (<40 years old) reaching the highest frequency level in the twenties (Figure 2E and Supplementary figure 5B). These results suggest that AAV active infection is more frequent in the second and third decade during life, while inactive not-episomal forms subsist after the primary infection.

Co-infection with AAV helper viruses

As AAV is a defective virus, we searched for the presence of potential AAV helper viruses by screening the entire cohort of 1,319 liver tissues for human adenoviruses (AdV types A to F), human herpesviruses (HHV type 1, 2, 4, 5, 6, 7 and 8) and human papillomavirus type 16 (HPV16) by quantitative PCR. At least one of these viruses was detected in 43% of the patients (n=570), and only one per patient in 39% (n=520). HHV6 was the most frequent (39%), then HHV4 (Epstein-Barr virus, EBV, 6%), while HHV7 and adenovirus were only rarely detected (2% and 0.5% respectively, Figure 3A). No HPV16 and HHV type 1 (HSV1), 2 (HSV2), 5 (CMV) and 8 (KSHV) were found in our cohort of liver tissues. HHV6 was the only helper virus enriched in AAV positive patients (37.3% versus 44.8%, $p=0.039$), in particular in patients with episomal or expressed-episomal forms (52.5% and 67.9% respectively, $p<0.001$; Figure 3B).

To identify independent features associated with AAV infection in the overall cohort of patients, we performed a multivariate analysis (Figure 3C). Female gender (odds ratio, OR=1.83, $p<0.001$), the age (OR=1.42, $p=0.044$), non-cirrhotic liver (OR=1.96, $p<0.001$) and co-infection with HHV6 (OR=1.15, $p=0.031$) were independently associated with AAV positivity. Three factors were also significantly associated to the presence of episomal and expressed AAV: female gender (OR=4.71, $p=0.013$), non-fibrotic liver (OR=12.13, $p=0.018$) and co-infection with HHV6 (OR=1.61, $p=0.01$).

AAV in tumor tissues

AAV DNA positivity was less frequently identified in the tumor tissues (n=109, 8%) compared to non-tumor liver tissues (n=233, 18%) with only 4.7% of patients presenting AAV in both tumor and non-tumor compartments (Figure 4A). Twenty out of the 109 positive tumors showed a high number of AAV copies/cell ranging from 0.07 to 6.08. This value might be underestimated considering both potential contamination by normal cells and ploidy of tumor hepatocytes. The vast majority (n=83, 76%) had only 1 or 2 amplified viral regions with an enrichment for the 3'ITR region of the virus (Supplementary figure 6A-B). AAV was detected with a similar frequency in malignant and benign tumors, but with a higher number of copies/cell in malignant tumors corresponding to the clonal AAV insertion events (Figure 4B-

C; Supplementary table 3). Conversely, in all patients with benign tumors except one with focal nodular hyperplasia (FNH), AAV was more highly positive in the non-tumor counterpart than in the corresponding tumor (Figure 4C). Finally, viral episomal forms were rarely identified in tumors (n=8, 0.6%), mostly in benign tumors (4 HCA and 2 FNH) and only 2 HCC (Supplementary table 3).

AAV insertion in liver tissues

We identified 7 novel clonal insertions in 6 HCCs, in *GLI1*/*INHBE*, *TERT* and *CCNA2*. Only one clonal insertion was identified in a benign focal nodular hyperplasia, it occurred in an intergenic region of chromosome 10 without consequences on the expression of the nearest genes (Figure 4D and 5). Combining with AAV insertions identified in TCGA and ICGC sequenced HCC[44, 45] and previously described cases in our cohort,[25, 34] we re-analyzed a total of 30 independent AAV insertions in liver tumors (Supplementary table 4). Viral insertions occurred in both directions, AAV2 and AAV2/13 subtypes were equally represented (55% versus 45% of the interpretable cases, respectively) and the minimal AAV region commonly inserted (nucleotide 4390-4570) was identified in 25 out of the 30 insertions.

Six oncogenes were recurrently activated by AAV (Supplementary figure 7). Insertions in *GLI1*/*INHBE* (4 adenomas transformed into HCC), *TERT* (2 HCC), *CCNE1* (7 HCC), *TNFSF10* (2 HCC) and *KMT2B* (2 HCC) led in almost all the cases to an overexpression of full-length coding region of these oncogenes by a promoter and/or a enhancer cis mechanism (Figure 5). *CCNA2* was inserted in 9 HCC; all insertions but one clustered in *CCNA2* intron 2, they resulted in an abnormal AAV-CCNA2 transcript leading to a stable oncogenic truncated protein lacking the N-terminal regulatory domain (Figure 5D).[34] The 3'UTR of *TNFSF10* showed AAV insertions in two HCC inducing TNFSF10 overexpression with transcripts that prematurely ended at the viral polyadenylation (Figure 5F). Here, using site-directed mutagenesis of both insertions, we demonstrated that the viral poly-A signal is required to ensure a strong luciferase overexpression in 3 different tested cell lines (Supplementary figure 8).

In the non-tumor liver tissues, no clonal AAV insertions were identified; non-clonal insertions were significantly associated to the presence of episomal AAV ($p<0.001$), in contrast to the tumor samples. In both non-tumor and tumor tissues, non-clonal AAV insertions were randomly distributed along the genome (Figure 4D and Supplementary figure 9). No specific enrichment was found in major target of AAV previously described in cell lines.[46, 47]

AAV features and tumor heterogeneity

We explored inter-tumor heterogeneity by analysing multi-nodules (n=475) from 186 patients for the presence of viral DNA, clonal insertions and episomal form. Of those, AAV DNA was detected in 25 patients (Supplementary figure 6C), including 4 patients with clonal AAV insertion in at least one nodule. Two HCC patients displayed clonal AAV integrations in all nodules. Thanks to the NGS data, we were able to predict the evolution of these tumors by looking at the common and private somatic mutations and copy number alterations (CNA) in each nodule. Interestingly, the 2 tumors from patient #2557 showed the same viral insertion in

TNFSF10, similar gene mutations and CNA profiles, demonstrating that AAV insertion is a truck alteration occurring before intra-hepatic metastasis (Figure 6A). Conversely, the three tumors from patients #1919, resulting from a malignant transformation of adenoma in carcinoma, harboured 3 different clonal insertions all targeting *GLI1*, with different gene mutation profile and no CNA suggesting that the three nodules have an independent origin (Figure 6B).

DISCUSSION

In this study, we provided a comprehensive description at large scale of the different AAV viral forms in the liver and of its oncogenic consequences, contributing to better understand the natural history of AAV infection in human.

The prevalence of AAV was observed in 21% of patients in non-tumor and/or tumor liver in agreement with the seroprevalence of antibody against AAV identified in 30 to 80% of the general population.[10-12, 48] Our result showed that one out of 5 patients demonstrates persistent AAV DNA in the liver during life, mainly in the population of young and female patients without liver fibrosis (Figure 7). However, since most of our liver tissues were sampled from patients with liver diseases, the exact prevalence of AAV DNA in the liver of healthy individuals remains to be evaluated.

Only two AAV genotypes, AAV2 and hybrids AAV2/13, were identified in our cohort equally distributed among the patients. AAV2/13 sequences were hybrids between AAV2 in the 5'part and AAV13 in the 3' corresponding to the previous clade C of the VP1 classification.[14] Since only one full-length AAV sequence from clade C was publicly available[42], our work significantly increased the number human AAV full-length sequences enlightening the genomic variants associated with an efficient natural AAV infection in the liver. In contrast to previous serological analysis,[10, 11, 49] we did not identified other AAV genotypes in the liver, even if AAV5 and 8 were frequent in circulating monocytes.[48]

AAV episomal form was identified in the non-tumor tissues of 4.6% of the patients, representing 26% of all AAV positive liver samples, whereas episomal AAV has only be described in human tonsil and adenoid previously.[9] It was frequently associated to viral mRNA expression suggesting that the episomal AAV are also transcriptionally active in a significant proportion of the population in the liver (Figure 7). Several viruses[50, 51] are able to support AAV replication *in vitro*, and it was commonly admitted that adenovirus is the natural AAV helper. Here, we identified HHV6 as the virus most frequently associated with episomal and transcribed AAV in the liver. This co-occurrence was previously described in healthy blood donors[48] and HHV6 is able to infect hepatocytes.[52-54] The increased frequency of HHV6 in patients with episomal-expressed AAV form could indicate an ongoing active infection in the liver of 2.1% of the patients. In contrast, only very rare patients showed an association with adenovirus or other candidate helper viruses even in livers with episomal and expressed AAV (Figure 7). All these results may suggest the role of HHV6 as the natural helper virus of AAV in the liver. However, co-infection with other helper viruses could occur at the initial acute AAV infection, followed by its clearance. Replication-competent infectious AAV has been rescued from human tonsil and adenoid tissue and lymphocytes, it remain to be searched in fresh liver tissues.[48, 55] Viral clones were isolated and their infectivity was tested *in vitro* in Hela cells showing that only AAV clones with a complete double-D ITR structure were able to replicate and gave rise to infectious virus.[55] Interestingly, the analysis of the ITRs junctions of the episomal form in our series highlighted the presence of the same double-D structure supporting its role in an active AAV infection. Moreover, here a peculiar link between episomal-expressed AAV in the liver and age suggested that AAV active infection occurs during the first 3 decades of life and then remains latent.

Analyses of the tumor tissues confirmed the selection of clonal AAV insertion in HCC development in non-cirrhotic liver. Recurrent somatic viral integrations were identified in 2% of our HCC cohort, targeting *CCNA2* (33.3%), *CCNE1* (27.8%), *GLI1*/*INHBE* (11.1%), *TERT* (11.1%), *TNFSF10* (11.1%) and *KMT2B* (5.6%). AAV insertion induced the overexpression of the target genes through multiple mechanisms that differ according to the target and the localization of the integration. Clonal insertions upstream the TSS or within the 5' region of the gene lead to the gain of a positive regulatory mechanism such as the usage of viral enhancers and transcription factor binding sites (TFBS). Interestingly, a recent work by Logan and collaborators has described a liver specific enhancer-promoter element in wild-type AAV genome within the common inserted region in HCC tumors.[29] It consists of 124 nucleotides sequence that contains TFBSs for HNF1-α, HNF6 and GATA6. Noteworthy, this region is absent in many AAV vectors currently in use and should raise a biosafety flag or be deleted in the remaining. In line with our finding, this result strongly supports the mechanism of AAV induced over-expression of the target gene. In addition, viral insertions in *CCNA2* and *TNFSF10* genes led the expression of a truncated protein or the premature ending of the transcript within the viral poly-A, respectively.

AAV oncogenic integrations were not only identified in our cohort of European HCC patients. They were also observed in the ICGC-Japan cohort in 3 HCC cases out of 268 HCC (1.1%),[45] in 4 out of 334 HCC (1.2%) of the TCGA cohort[34] and in 2 out of 289 HCC (0.7%) from Korea.[56] Interestingly the most frequent AAV integrated oncogenes are similar to HBV, i.e. *CCNA2*, *CCNE1, TERT* and *KMT2B*. The lower prevalence of AAV could be due to the lack of chronic liver disease associated to active AAV replication in contrast to chronic HBV infection. In the present series, we reinforced the link between AAV oncogenic insertion and the occurrence of HCC in normal liver, including recurrent AAV insertions in the malignant transformation of hepatocellular adenoma in carcinoma targeting *GLI1* that defines the activated sonic hedgehog molecular subgroup of adenoma, shHCA.[57] In the same line, AAV insertions in cyclin A2 or E1 in HCC are associated with unique chromosomal rearrangement signature and poor prognosis mainly occurring in HCC developed in normal liver.[34] These results underline the role of AAV insertion in the development of a specific subgroup of HCC without other etiologies.

In conclusion, we provided a portrait of AAV infection in the liver with a description of viral genotypes, molecular forms and helper virus paving the way for a renovated interest in wild type AAV biology. New highlights on the understanding of the oncogenic consequences of AAV integration in HCC tumors emerged from this work. However, further studies are necessary to clarify the impact of AAV infection in additional cohort of patients and the frequency of insertional mutagenesis across different countries.

REFERENCES

1. Rose JA, Berns KI, Hoggan MD, et al. Evidence for a single-stranded adenovirus-associated virus genome: formation of a DNA density hybrid on release of viral DNA. Proc Natl Acad Sci U S A 1969;64:863-9.
2. Atchison RW, Casto BC, Hammon WM. Adenovirus-Associated Defective Virus Particles. Science 1965;149:754-6.
3. Balakrishnan B, Jayandharan GR. Basic biology of adeno-associated virus (AAV) vectors used in gene therapy. Curr Gene Ther 2014;14:86-100.
4. Sonntag F, Schmidt K, Kleinschmidt JA. A viral assembly factor promotes AAV2 capsid formation in the nucleolus. Proc Natl Acad Sci U S A 2010;107:10220-5.
5. Yang CC, Xiao X, Zhu X, et al. Cellular recombination pathways and viral terminal repeat hairpin structures are sufficient for adeno-associated virus integration in vivo and in vitro. J Virol 1997;71:9231-47.
6. Nakai H, Wu X, Fuess S, et al. Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. J Virol 2005;79:3606-14.
7. Huser D, Gogol-Doring A, Chen W, et al. Adeno-associated virus type 2 wild-type and vector-mediated genomic integration profiles of human diploid fibroblasts analyzed by third-generation PacBio DNA sequencing. J Virol 2014;88:11253-63.
8. Schultz BR, Chamberlain JS. Recombinant adeno-associated virus transduction and integration. Mol Ther 2008;16:1189-99.
9. Schnepp BC, Jensen RL, Chen CL, et al. Characterization of adeno-associated virus genomes isolated from human tissues. J Virol 2005;79:14793-803.
10. Boutin S, Monteilhet V, Veron P, et al. Prevalence of serum IgG and neutralizing factors against adeno-associated virus (AAV) types 1, 2, 5, 6, 8, and 9 in the healthy population: implications for gene therapy using AAV vectors. Hum Gene Ther 2010;21:704-12.
11. Calcedo R, Vandenberghe LH, Gao G, et al. Worldwide epidemiology of neutralizing antibodies to adeno-associated viruses. J Infect Dis 2009;199:381-90.
12. Erles K, Sebokova P, Schlehofer JR. Update on the prevalence of serum antibodies (IgG and IgM) to adeno-associated virus (AAV). J Med Virol 1999;59:406-11.
13. Gao GP, Alvira MR, Wang L, et al. Novel adeno-associated viruses from rhesus monkeys as vectors for human gene therapy. Proc Natl Acad Sci U S A 2002;99:11854-9.
14. Gao G, Vandenberghe LH, Alvira MR, et al. Clades of Adeno-associated viruses are widely disseminated in human tissues. J Virol 2004;78:6381-8.
15. Mori S, Wang L, Takeuchi T, et al. Two novel adeno-associated viruses from cynomolgus monkey: pseudotyping characterization of capsid protein. Virology 2004;330:375-83.
16. Schmidt M, Voutetakis A, Afione S, et al. Adeno-associated virus type 12 (AAV12): a novel AAV serotype with sialic acid- and heparan sulfate proteoglycan-independent transduction activity. J Virol 2008;82:1399-406.
17. Chandler RJ, Sands MS, Venditti CP. Recombinant Adeno-Associated Viral Integration and Genotoxicity: Insights from Animal Models. Hum Gene Ther 2017;28:314-322.
18. Berns KI, Muzyczka N. AAV: An Overview of Unanswered Questions. Hum Gene Ther 2017;28:308-313.
19. Donsante A, Miller DG, Li Y, et al. AAV vector integration sites in mouse hepatocellular carcinoma. Science 2007;317:477.
20. Chandler RJ, LaFave MC, Varshney GK, et al. Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. J Clin Invest 2015;125:870-80.
21. Wang PR, Xu M, Toffanin S, et al. Induction of hepatocellular carcinoma by in vivo gene targeting. Proc Natl Acad Sci U S A 2012;109:11264-9.
22. Hermonat PL, Plott RT, Santin AD, et al. Adeno-associated virus Rep78 inhibits oncogenic transformation of primary human keratinocytes by a human papillomavirus type 16-ras chimeric. Gynecol Oncol 1997;66:487-94.

23. Liu T, Cong M, Wang P, et al. Adeno-associated virus Rep78 protein inhibits Hepatitis B virus replication through regulation of the HBV core promoter. Biochem Biophys Res Commun 2009;385:106-11.
24. Kokorina NA, Santin AD, Li C, et al. Involvement of protein-DNA interaction in adeno-associated virus Rep78-mediated inhibition of HIV-1. J Hum Virol 1998;1:441-50.
25. Nault JC, Datta S, Imbeaud S, et al. Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. Nat Genet 2015;47:1187-93.
26. Zhao LH, Liu X, Yan HX, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. Nat Commun 2016;7:12992.
27. Sung WK, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 2012;44:765-9.
28. Llovet JM, Zucman-Rossi J, Pikarsky E, et al. Hepatocellular carcinoma. Nat Rev Dis Primers 2016;2:16018.
29. Logan GJ, Dane AP, Hallwirth CV, et al. Identification of liver-specific enhancer-promoter activity in the 3' untranslated region of the wild-type AAV2 genome. Nat Genet 2017;49:1267-1273.
30. Benaglia T, Chauveau D, Hunter DR, et al. mixtools: An R Package for Analyzing Mixture Models. Journal of Statistical Software 2009;32:1-29.
31. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.
32. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.
33. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841-2.
34. Bayard Q, Meunier L, Peneau C, et al. Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. Nat Commun 2018;9:5235.
35. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14:R36.
36. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562-78.
37. Sloutskin A, Danino YM, Orenstein Y, et al. ElemeNT: a computational tool for detecting core promoter elements. Transcription 2015;6:41-50.
38. Nishikawa T, Ota T, Isogai T. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. Bioinformatics 2000;16:960-7.
39. Liu H, Han H, Li J, et al. An in-silico method for prediction of polyadenylation signals in human sequences. Genome Inform 2003;14:84-93.
40. Werle-Lapostolle B, Bowden S, Locarnini S, et al. Persistence of cccDNA during the natural history of chronic hepatitis B and decline during adefovir dipivoxil therapy. Gastroenterology 2004;126:1750-8.
41. Mroske C, Rivera H, Ul-Hasan T, et al. A capillary electrophoresis sequencing method for the identification of mutations in the inverted terminal repeats of adeno-associated virus. Hum Gene Ther Methods 2012;23:128-36.
42. Chen CL, Jensen RL, Schnepp BC, et al. Molecular characterization of adeno-associated viruses infecting children. J Virol 2005;79:14781-92.
43. Penaud-Budloo M, Le Guiner C, Nowrouzi A, et al. Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. J Virol 2008;82:7875-85.
44. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017;169:1327-1341 e23.
45. Fujimoto A, Furuta M, Totoki Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 2016;48:500-9.

46. Kotin RM, Siniscalco M, Samulski RJ, et al. Site-specific integration by adeno-associated virus. Proc Natl Acad Sci U S A 1990;87:2211-5.
47. Samulski RJ, Zhu X, Xiao X, et al. Targeted integration of adeno-associated virus (AAV) into human chromosome 19. EMBO J 1991;10:3941-50.
48. Huser D, Khalid D, Lutter T, et al. High Prevalence of Infectious Adeno-associated Virus (AAV) in Human Peripheral Blood Mononuclear Cells Indicative of T Lymphocytes as Sites of AAV Persistence. J Virol 2017;91.
49. Li C, Narkbunnam N, Samulski RJ, et al. Neutralizing antibodies against adeno-associated virus examined prospectively in pediatric patients with hemophilia. Gene Ther 2012;19:288-94.
50. Mast TC, Kierstead L, Gupta SB, et al. International epidemiology of human pre-existing adenovirus (Ad) type-5, type-6, type-26 and type-36 neutralizing antibodies: correlates of high Ad5 titers and implications for potential HIV vaccine trials. Vaccine 2010;28:950-7.
51. Steininger C, Rassenti LZ, Vanura K, et al. Relative seroprevalence of human herpes viruses in patients with chronic lymphocytic leukaemia. Eur J Clin Invest 2009;39:497-506.
52. Ozaki Y, Tajiri H, Tanaka-Taya K, et al. Frequent detection of the human herpesvirus 6-specific genomes in the livers of children with various liver diseases. J Clin Microbiol 2001;39:2173-7.
53. Ishikawa K, Hasegawa K, Naritomi T, et al. Prevalence of herpesviridae and hepatitis virus sequences in the livers of patients with fulminant hepatitis of unknown etiology in Japan. J Gastroenterol 2002;37:523-30.
54. Cermelli C, Concari M, Carubbi F, et al. Growth of human herpesvirus 6 in HEPG2 cells. Virus Res 1996;45:75-85.
55. Schnepp BC, Jensen RL, Clark KR, et al. Infectious molecular clones of adeno-associated virus isolated directly from human tissues. J Virol 2009;83:1456-64.
56. Park KJ, Lee J, Park JH, et al. Adeno-Associated Virus 2-Mediated Hepatocellular Carcinoma is Very Rare in Korean Patients. Ann Lab Med 2016;36:469-74.
57. Nault JC, Couchy G, Balabaud C, et al. Molecular Classification of Hepatocellular Adenoma Associates With Risk Factors, Bleeding, and Malignant Transformation. Gastroenterology 2017;152:880-894 e6.

Acknowledgments: We thank surgeons, pathologists and all the clinicians that collected samples and clinical data. Sophie Prevost, Service d'anatomie pathologique, AP-HP, Hôpital Antoine-Béclère, Clamart, France. Anne de Muret, Service d'anatomopathologie, Centre Hospitalier Régional Universitaire de Tours, Tours, France. Eric Viber, Centre Hépato-Biliaire, INSERM U785, Hôpital Paul Brousse, Villejuif, France. Philippe Merle, Department of Hepatology, Hospices Civils de Lyon, Croix-Rousse University Hospital, Lyon, France. Monique Fabre, Service Anatomie, HU-Necker Enfants Malades AP-HP, Paris. Nathalie Sturm, Department of Anatomie et Cytologie Pathologiques, CHU de Grenoble, Grenoble, France. Thomas Decaens, Service Hépato-gastroentérologie et Tumeurs du foie, CHU de Grenoble, Grenoble, France. Sophie Michalak, département de Pathologie cellulaire et tissulaire, CHU ANGERS. Georges-Philippe Pageaux, service d'hépato-gastro-entérologie Hôpital St Eloi CHU Montpellier. Jean-Michel Fabre, service de chirurgie digestive Hôpital St Eloi CHU Montpellier. Emmanuel Boleslawski, service de chirurgie digestive et transplantation. Hôpital Huriez. Chru de lille. 59037 lille cedex. Marie Christine Saint Paul, service d'Anatomie Pathologique, CHU de Nice, Nice. Dominique Wendum, Department of Pathology, Saint-Antoine Hospital, AP-HP, Paris, France. Olivier Rosmorduc, Department of Gastroenterology and Hepatology, Hôpital de la Pitié-Salpêtrière, AP-HP, Université Pierre et Marie Curie UPMC, Paris. Jean Christophe Vaillant, service de chirurgie hépato-bilio-pancréatique, CHU Pitié-Salpetriere, Université Pierre et Marie Curie UPMC, Paris. Marianne Ziol, Service d'Anatomopathologie, Hôpital Jean Verdier, Hôpitaux universitaires Paris-Seine-Saint-Denis, AP-HP, Bondy, France. Nathalie Ganne, Department of Hepatogastroenterology, Hôpital Jean Verdier, AP-HP, Bondy, France. Luigi Terraciano, Basel University Hospital, Department of Pathology, Basel, Switzerland. Vincenzo Mazzaferro, university of Milan at the Istituto Nazionale Tumori IRCCS (National Cancer Institute). Celine Bazille, Service d'Anatomie Pathologie, CHU de Caen, Caen, France.

Author contributions

Study design: JCN, JZR, TLB, SI
Generation of experimental data: TLB, SI, CP, IM, SD
Analysis and interpretation of data: TLB, SI, CP, IM, SD, QB, SC, TZH, EL, JCN, JZR
Collection of samples and related histological and clinical data: JCN, JZR, JC, GM, CG, VP, GA, AL, LP, LC, PBS, JFB and investigators.
Drafting of the manuscript: JZR, TL, SI
Revision of the manuscript and approval of the final version of the manuscript: JZR, JCN, TLB, SI, IM, SD, QB, SC, TZH, EL, JC, GM, CG, VP, GA, AL, LP, LC, PBS, JFB.

Grants supports: This work was supported by Inserm, by INCa within the ICGC project, France Génomique, Cancéropole Ile de France (ExhauTrans project), ITMO Cancer AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé, National Alliance for Life Sciences & Health) within the framework of the Cancer Plan ("HTE program-HetColi network" and "Cancer et environnement program"), the Réseau national CRB Foie, Ligue Nationale contre le cancer: project équipe Labellisée, Fondation Schueller Bettencourt "coup d'élan", Prix Ligue contre le Cancer comité de Paris René et Andrée Duquesne 2018, the SIRIC CARPEM and Fondation Mérieux, Labex OncoImmunology (investissement d'avenir) ANRS

and the French Liver Biobanks network – INCa, BB-0033-00085, Hepatobio bank. QB is supported by a fellowship from the HOB doctoral school and the ministry of Education and Research, TLB is supported by an “Attractivité IDEX" fellowship from IUH and CP is supported by a doctoral fellowship funded by ANRS.

Conflict of interests for all authors: the authors have no conflicts of interests related to the manuscript. A European patent application EP19305521.7 was filed on April 23, 2019 entitled “New adeno-associated virus (aav) variants and uses thereof for gene therapy”.

Data availability: The sequencing data reported in this paper have been deposited to Genbank (**accessions:** KT258720-KT258730, MK139243-MK139299, MK163929-MK163942 and MK231253-MK231264) and EGA (European Genome-phenome Archive) database (RNA-seq accessions: EGAS00001002879, EGAS00001001284 and EGAS00001003310). All supplemental data and sequences are available at http://zucmanlab.com/wp-content/uploads/2019/05/Zucman-CaptureVirusProbes.xlsx

Figure 1. AAV Full-length sequences in 57 human liver tissues. A) Schematic representation of AAV genome (reference NC_001401) with location of the two open reading frames encoding replication proteins (Rep78, Rep68, Rep52 and Rep40), structural proteins (VP1, VP2 and VP3) and AAP protein. Inverted terminal repeats (ITR) are represented on the 5' and 3' ends. Promoters (p5, p19 and p40) are indicated with arrows. B) Nucleotides sequences (4679 bp) from 57 full-length AAV isolated from human liver tissues (ID number indicated with #) multi-aligned with the ClustalW algorithm compared to reference sequences on the top, AAV2 (NC_001401, in white), AAV3 (NC_001729.1) and AAV13 (EU285562.1). Two distinct viral genotypes, AAV2 and AAV2/13 were identified. Color bars indicated nucleotide divergence with the AAV2 reference genome similar to AAV3 and/or AAV13 genomes (green) or not (grey), similarities with NC_001401 are in white. Variations due to flip-flop ITR configurations compared to AAV2 reference are labeled in light grey. The liver-specific enhancer–promoter element (LSP) described by Logan et al. is indicated.[29] C) Amino acid variations compared to the AAV2 reference are indicated. The triangles indicate genome location of specific AAV2/13 (top) or AAV2 (middle) variants in the series of 57 human liver AAV isolates. Common variants shared by both genotypes are shown (bottom). Grey and black colors refer to silent and missense AAV variants, respectively; numbers correspond to wild-type AAV2 nucleotide sequence coordinates (NC_001401).

Figure 2. AAV DNA in non-tumor tissues and viral episomal form. A) Copy number/cell distribution in 233 samples. The density line defines the low and high positivity groups in blue and red, respectively. B) Contingency analysis of AAV positive and negative patients according to gender, age and Metavir fibrosis score. Frequency of AAV positive patients is displayed (χ^2 test with Monte Carlo simulation and χ^2 test for trend in proportions for Metavir score). C) Frequency of RNA expression according to REP and CAP viral transcripts in patients with episomal and not-episomal AAV (χ^2 test with Monte Carlo simulation). D) Viral copy number/cell ($\log_{10}$) in AAV positive samples according to the episomal status and the transcriptional activity of the episome (Wilcoxon rank-sum test). E) Distribution of the different viral molecular forms according to the age of the patients. ***P< 0.001, **P< 0.01, *P<0.05.

Figure 3. Helper viruses according to AAV status. A) Frequency of helper viruses' infections and co-infection in non-tumor tissues (N=1319). B) Global frequency of HHV6, EBV, HHV7 and AdV infection according to AAV presence and form (χ^2 test for trend in proportions). C) Multivariate analysis for global AAV positivity (left) including the variables closely related to AAV presence in the univariate analysis (logistic regression). The same analysis was performed for the presence of episomal AAV (middle) and episomal and expressed form (right). ***P< 0.001, **P< 0.01, *P<0.05.

Figure 4. AAV in tumor tissues and non-tumor liver counterparts. A) Copy number/cell (log_{10}) of paired tumor (T) and non-tumor (NT) tissues of each patient (n=1269). Solid and dashed line define respectively the threshold of positivity and the boundaries between high and low number of viral copies per cell. The frequency of patients with AAV in both tumor and non-tumor counterparts or only in one of them is indicated. B) Frequency of AAV in tumor and non-tumor tissues of patients with malignant and benign tumors (χ^2 test with Monte Carlo simulation and Cochran–Mantel–Haenszel for gender adjustment). C) AAV copy number/cell of paired tumor and non-tumor tissues of 270 AAV positive patients grouped in malignant and benign tumor patients. Triangles represent the tumors with clonal AAV insertions (Wilcoxon rank-sum test). D) Pan-genomic views of genomic location of the human/virus matching chimeric and mate reads in tumor (top) and non-tumor (bottom) samples. A line corresponds to a 20k-bin region, color refers to the average number of reads counted per bin. The height of the lines corresponds to the frequency of presence of reads in the series of samples, considering 94 tumors and 82 non-tumors investigated with viral capture deepseq. ***P< 0.001, **P< 0.01, *P<0.05.

Figure 5. AAV clonal integration sites and transcripts consequences in tumors. Genes structure are schematized with boxes referring to exons and lines to introns regions. TSS location is shown on 5' of the gene. Arrows indicate viral insertion sites in our series, in red, and in TCGA and ICGC tumors, in green. Asterisks refers to new inserted cases. Top lines refer to inserted AAV viral regions and arrows to 5'>3' sequence orientation. Flip or flop 3'ITR are indicated. Observed transcripts are represented at the bottom of the gene structure with fusion viral sequences in red.

Figure 6. Tumor development in patients with multiple nodules and clonal AAV insertions. The relation between the tumors is determined according to gene mutation profile and copy number alteration (CNA) of each nodule. The number of shared and private alterations is indicated above each branch. The major alterations with amino acid consequences are listed; mutations in driver genes and main CNAs are in bold. The AAV status, diagnosis

and sources of genomic information (WGS, WES) are specified for each nodule. The thickness of the branch indicates the number of alterations. The position of the nodules for each patient is represented on the right. A) The two HCC nodules of patients #2557 display the same AAV insertion in *TNFSF10* and they share 199 somatic mutations and several CNAs. This profile suggests that the nodules originate from the same primary tumor. B) The three nodules of patient #1919 are heterogeneous for mutation profile and AAV insertions suggesting an independent origin of the tumors.

Figure 7. AAV and helper viruses in the general population and in human liver. Frequency of different AAV genotypes and seroprevalence of AAV[10-12] and helper viruses[50, 51] in the general population are showed in the upper panel. The error bar in the histogram represents the range of helper viruses seroprevalence according to the literature. The bottom panel summarizes the results found in this study, with estimated frequencies in the general population. For men and women, the global AAV frequency, the presence of episomal transcribed AAV and the prevalence of oncogenic clonal AAV insertions are indicated. *This prevalence is normalized according to the frequency of clonal AAV insertion in HCC (2%) and the prevalence of HCC in France (0.013%).

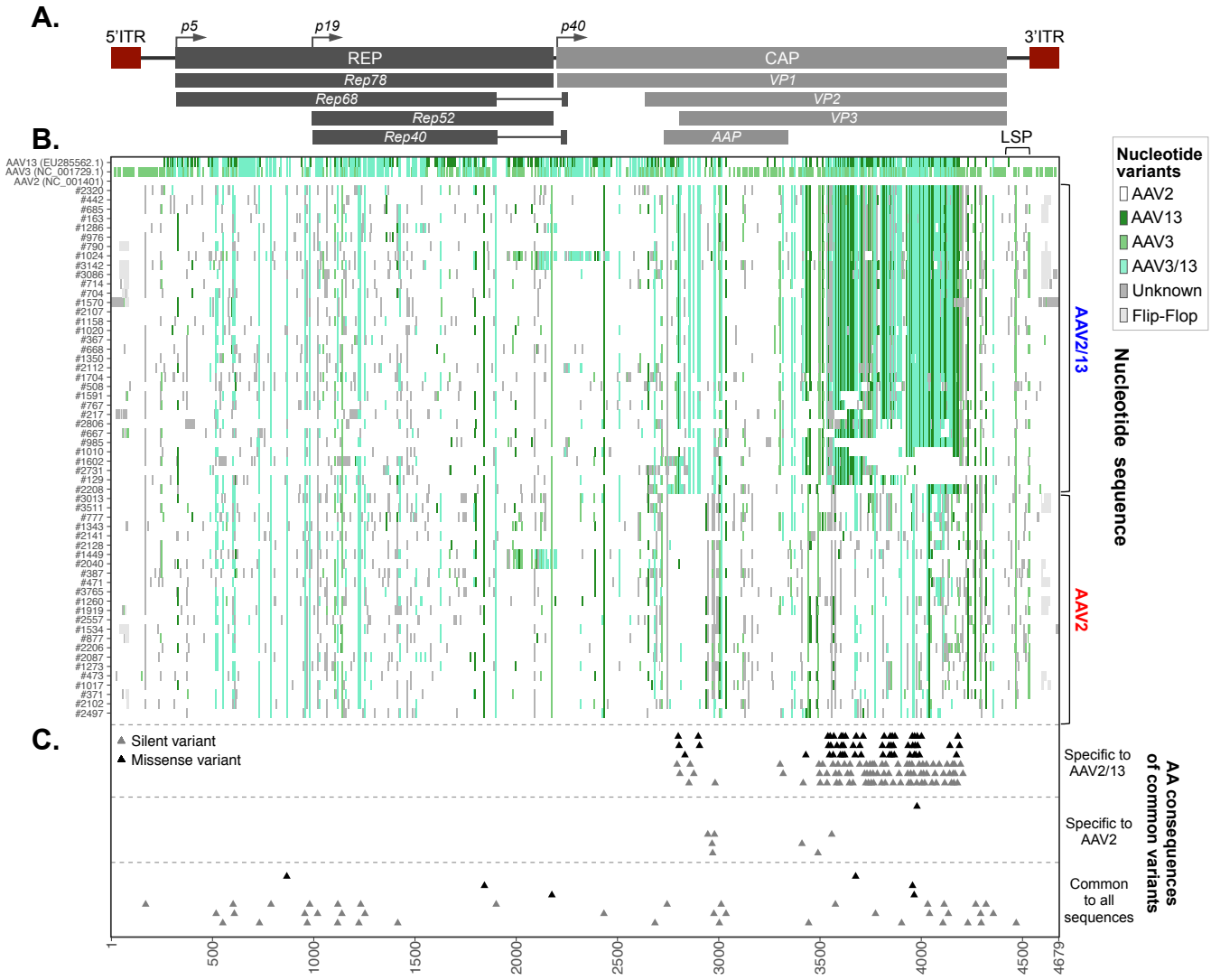

A.
5'ITR
p5
p19
p40
3'ITR
REP
CAP
Rep78
VP1
Rep68
VP2
Rep52
VP3
Rep40
AAP
B.
LSP
AAV13 (EU285562.1)
AAV3 (NC_001729.1)
AAV2 (NC_001401)
Nucleotide variants
AAV2
AAV13
AAV3
AAV3/13
Unknown
Flip-Flop
AAV2/13
AAV2
Nucleotide sequence
C.
Silent variant
Missense variant
Specific to AAV2/13
Specific to AAV2
Common to all sequences
AA consequences of common variants
1
500
1000
1500
2000
2500
3000
3500
4000
4500
4679

Figure 1

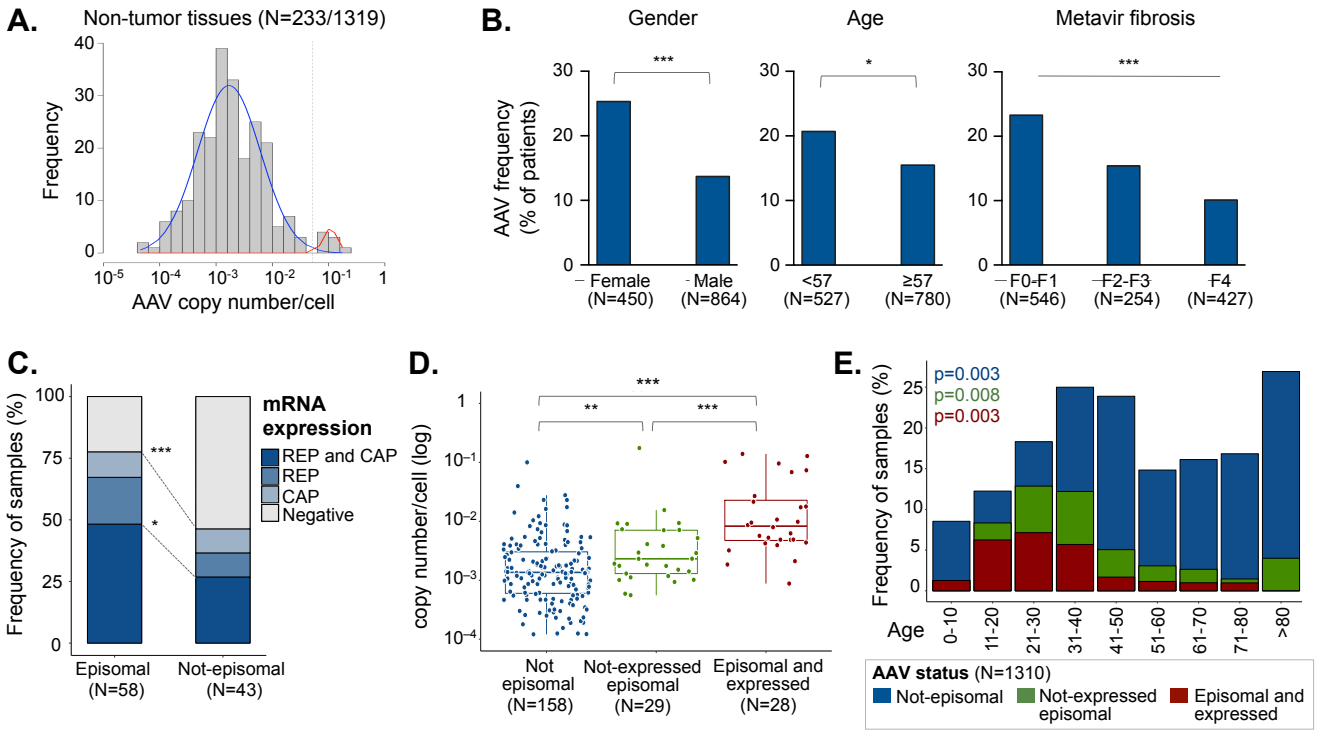

A.
Non-tumor tissues (N=233/1319)
Frequency
AAV copy number/cell
B.
Gender
Age
Metavir fibrosis
AAV frequency (% of patients)
Female (N=450)
Male (N=864)
<57 (N=527)
≥57 (N=780)
F0-F1 (N=546)
F2-F3 (N=254)
F4 (N=427)
C.
Frequency of samples (%)
mRNA expression
REP and CAP
REP
CAP
Negative
Episomal (N=58)
Not-episomal (N=43)
D.
copy number/cell (log)
Not episomal (N=158)
Not-expressed episomal (N=29)
Episomal and expressed (N=28)
E.
Frequency of samples (%)
p=0.003
p=0.008
p=0.003
Age
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
>80
AAV status (N=1310)
Not-episomal
Not-expressed episomal
Episomal and expressed

Figure 2

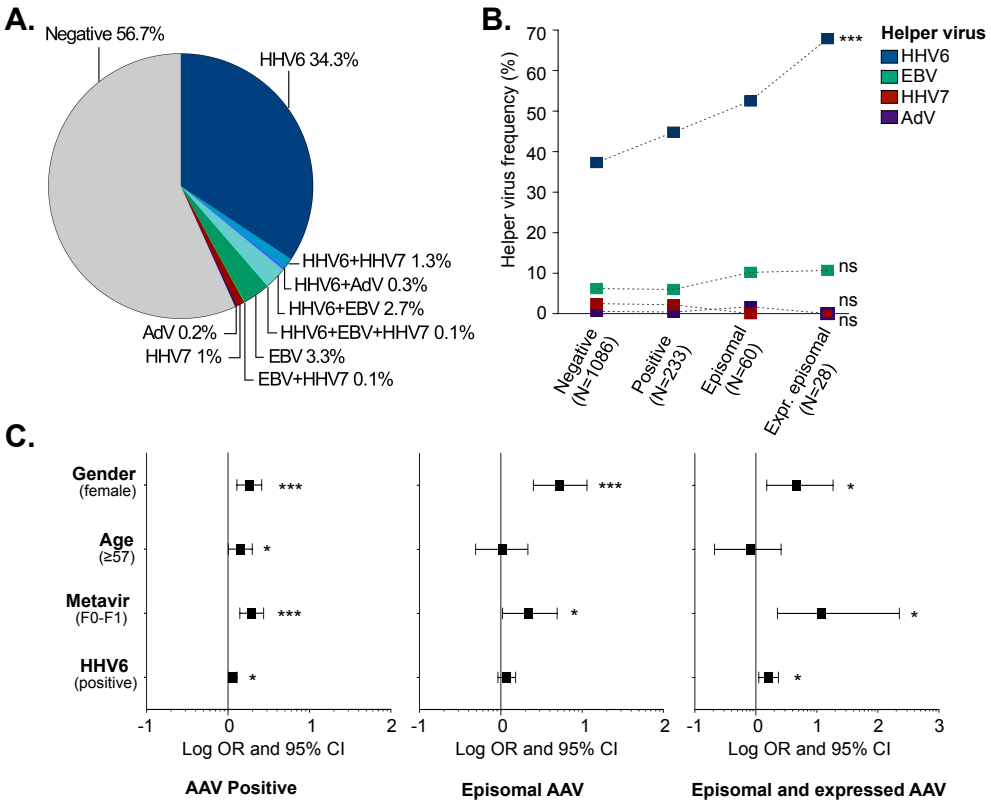

A.
Negative 56.7%
HHV6 34.3%
HHV6+HHV7 1.3%
HHV6+AdV 0.3%
HHV6+EBV 2.7%
HHV6+EBV+HHV7 0.1%
AdV 0.2%
HHV7 1%
EBV 3.3%
EBV+HHV7 0.1%
B.
Helper virus frequency (%)
Helper virus
HHV6
EBV
HHV7
AdV
***
ns
ns
ns
Negative (N=1086)
Positive (N=233)
Episomal (N=60)
Expr. episomal (N=28)
C.
Gender (female)
Age (≥57)
Metavir (F0-F1)
HHV6 (positive)
Log OR and 95% CI
AAV Positive
Episomal AAV
Episomal and expressed AAV

Figure 3

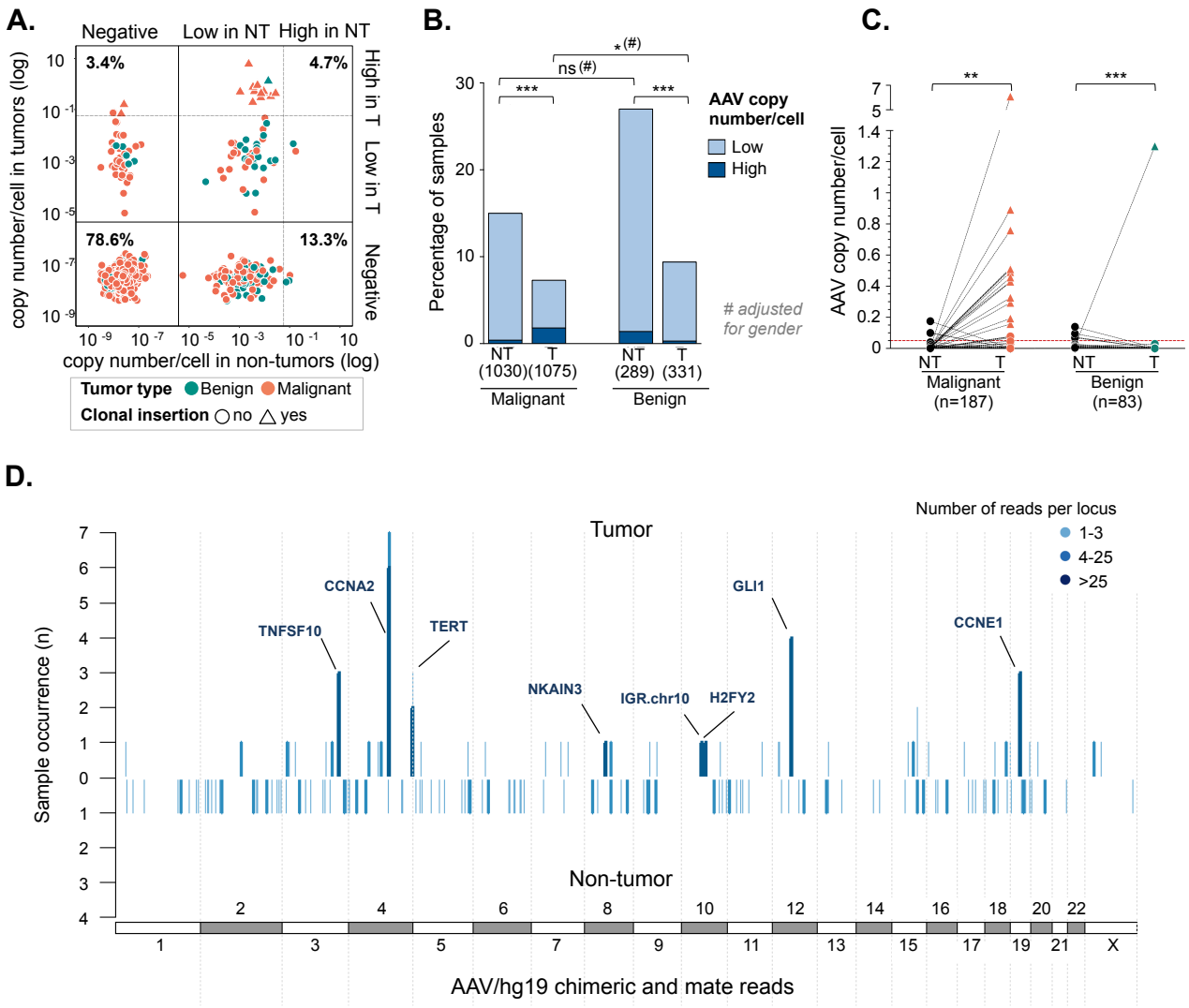

A.
Negative
Low in NT
High in NT
3.4%
4.7%
78.6%
13.3%
High in T
Low in T
Negative
copy number/cell in tumors (log)
copy number/cell in non-tumors (log)
Tumor type Benign Malignant
Clonal insertion no yes
B.
*(#)
ns(#)
***
***
Percentage of samples
AAV copy number/cell
Low
High
adjusted for gender
NT T NT T
(1030)(1075) (289) (331)
Malignant
Benign
C.
**
***
AAV copy number/cell
NT T NT T
Malignant (n=187)
Benign (n=83)
D.
Number of reads per locus
1-3
4-25
>25
Tumor
CCNA2
TNFSF10
TERT
NKAIN3
IGR.chr10
H2FY2
GLI1
CCNE1
Sample occurrence (n)
Non-tumor
AAV/hg19 chimeric and mate reads

Figure 4

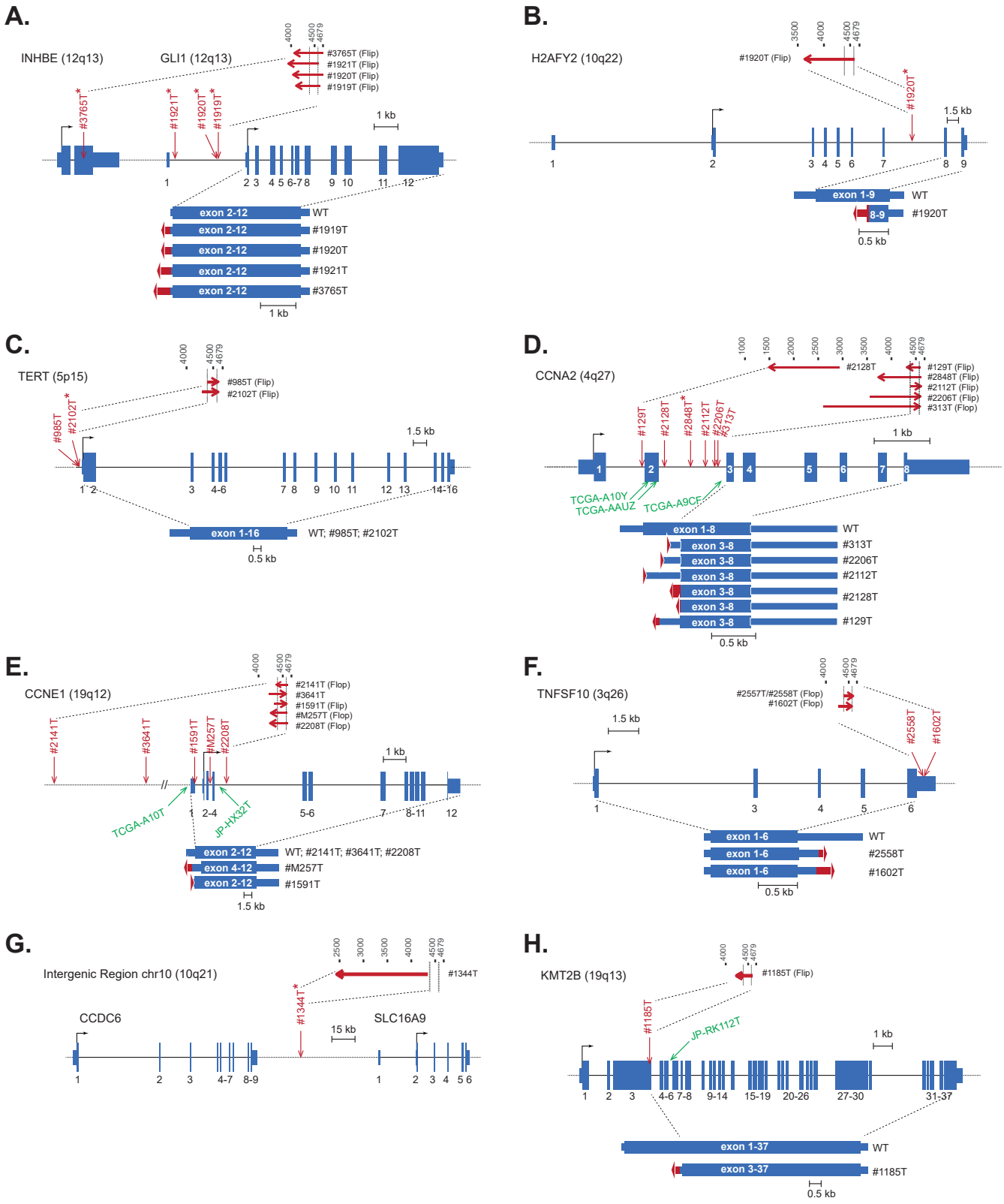

A.
INHBE (12q13)
GLI1 (12q13)
#3765T (Flip)
#1921T (Flip)
#1920T (Flip)
#1919T (Flip)
#3765T*
#1921T*
#1920T*
#1919T*
1 kb
exon 2-12
WT
#1919T
#1920T
#1921T
#3765T
B.
H2AFY2 (10q22)
#1920T (Flip)
#1920T*
1.5 kb
exon 1-9
WT
8-9
#1920T
0.5 kb
C.
TERT (5p15)
#985T (Flip)
#2102T (Flip)
#985T
#2102T*
1.5 kb
exon 1-16
WT; #985T; #2102T
0.5 kb
D.
CCNA2 (4q27)
#2128T
#129T (Flip)
#2848T (Flip)
#2112T (Flip)
#2206T (Flip)
#313T (Flop)
#129T
#2128T
#2848T*
#2112T
#2206T
#313T
1 kb
TCGA-A10Y
TCGA-AAUZ
TCGA-A9CF
exon 1-8
WT
exon 3-8
#313T
#2206T
#2112T
#2128T
#129T
0.5 kb
E.
CCNE1 (19q12)
#2141T (Flop)
#3641T
#1591T (Flip)
#M257T (Flop)
#2208T (Flop)
#2141T
#3641T
#1591T
#M257T
#2208T
1 kb
TCGA-A10T
JP-HX32T
exon 2-12
WT; #2141T; #3641T; #2208T
exon 4-12
#M257T
exon 2-12
#1591T
1.5 kb
F.
TNFSF10 (3q26)
#2557T/#2558T (Flop)
#1602T (Flop)
#2558T
#1602T
1.5 kb
exon 1-6
WT
#2558T
#1602T
0.5 kb
G.
Intergenic Region chr10 (10q21)
#1344T
#1344T*
CCDC6
SLC16A9
15 kb
H.
KMT2B (19q13)
#1185T (Flip)
#1185T
JP-RK112T
1 kb
exon 1-37
WT
exon 3-37
#1185T
0.5 kb

Figure 5

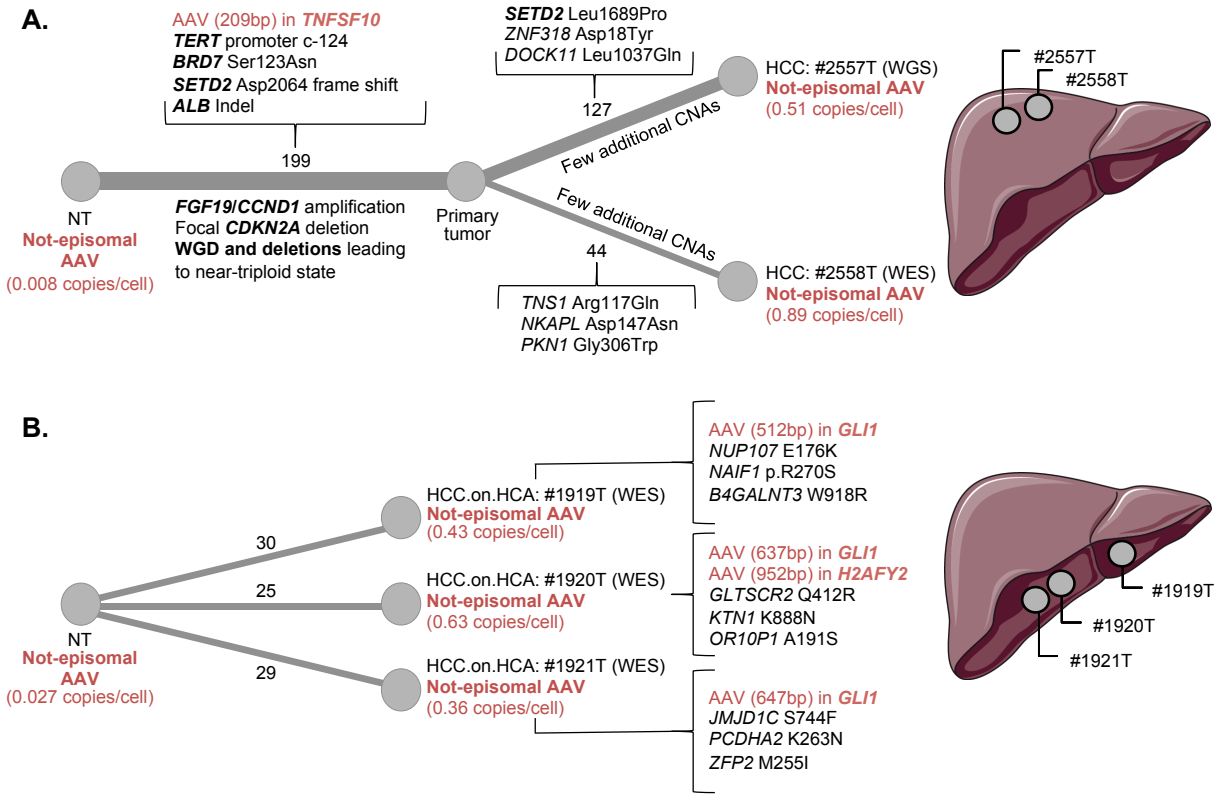

A.
AAV (209bp) in TNFSF10
TERT promoter c-124
BRD7 Ser123Asn
SETD2 Asp2064 frame shift
ALB Indel
199
SETD2 Leu1689Pro
ZNF318 Asp18Tyr
DOCK11 Leu1037Gln
127
HCC: #2557T (WGS)
Not-episomal AAV
(0.51 copies/cell)
Few additional CNAs
#2557T
#2558T
NT
Not-episomal AAV
(0.008 copies/cell)
FGF19/CCND1 amplification
Focal CDKN2A deletion
WGD and deletions leading to near-triploid state
Primary tumor
Few additional CNAs
44
TNS1 Arg117Gln
NKAPL Asp147Asn
PKN1 Gly306Trp
HCC: #2558T (WES)
Not-episomal AAV
(0.89 copies/cell)
B.
AAV (512bp) in GLI1
NUP107 E176K
NAIF1 p.R270S
B4GALNT3 W918R
HCC.on.HCA: #1919T (WES)
Not-episomal AAV
(0.43 copies/cell)
30
AAV (637bp) in GLI1
AAV (952bp) in H2AFY2
GLTSCR2 Q412R
KTN1 K888N
OR10P1 A191S
HCC.on.HCA: #1920T (WES)
Not-episomal AAV
(0.63 copies/cell)
25
NT
Not-episomal AAV
(0.027 copies/cell)
29
HCC.on.HCA: #1921T (WES)
Not-episomal AAV
(0.36 copies/cell)
AAV (647bp) in GLI1
JMJD1C S744F
PCDHA2 K263N
ZFP2 M255I
#1919T
#1920T
#1921T

Figure 6

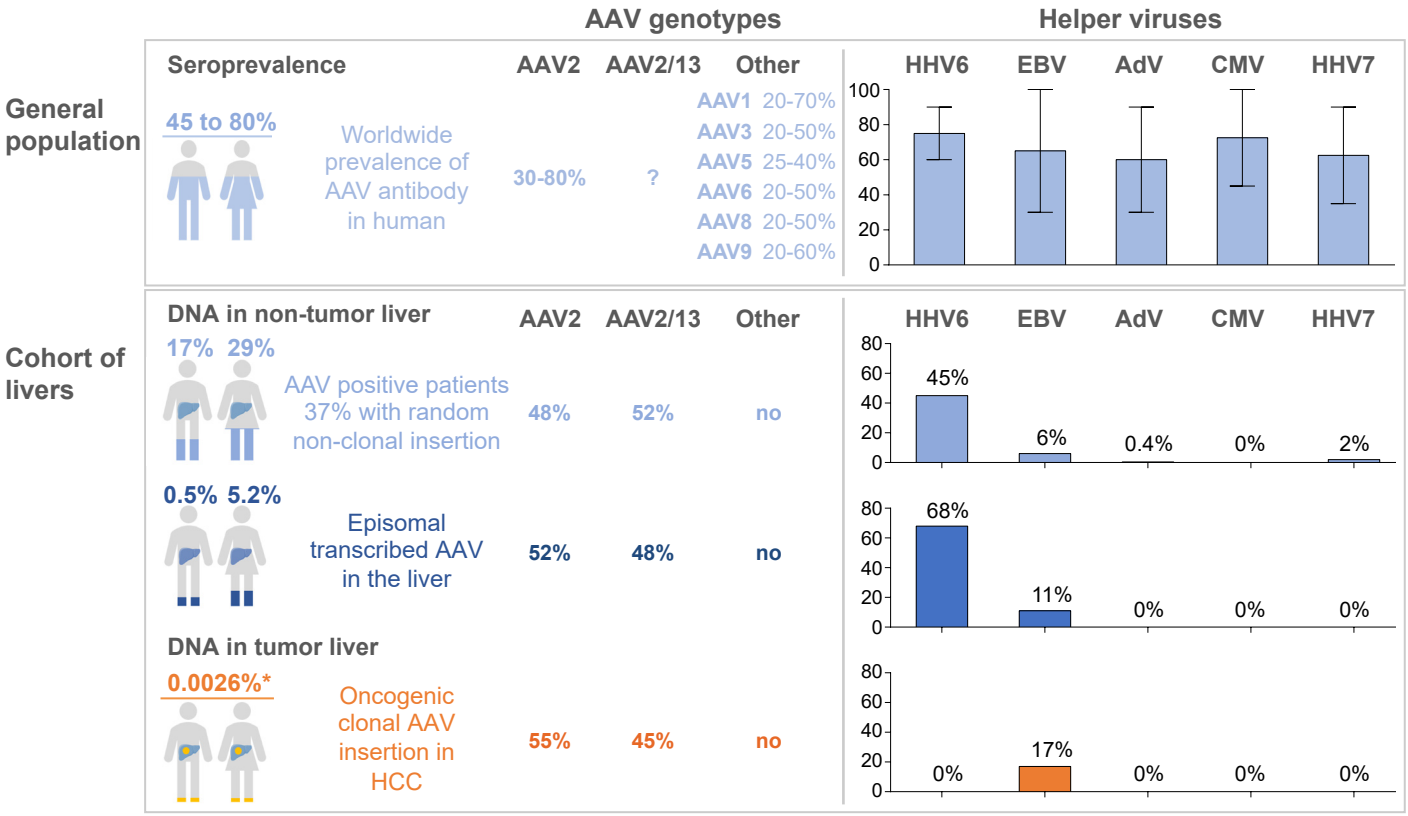
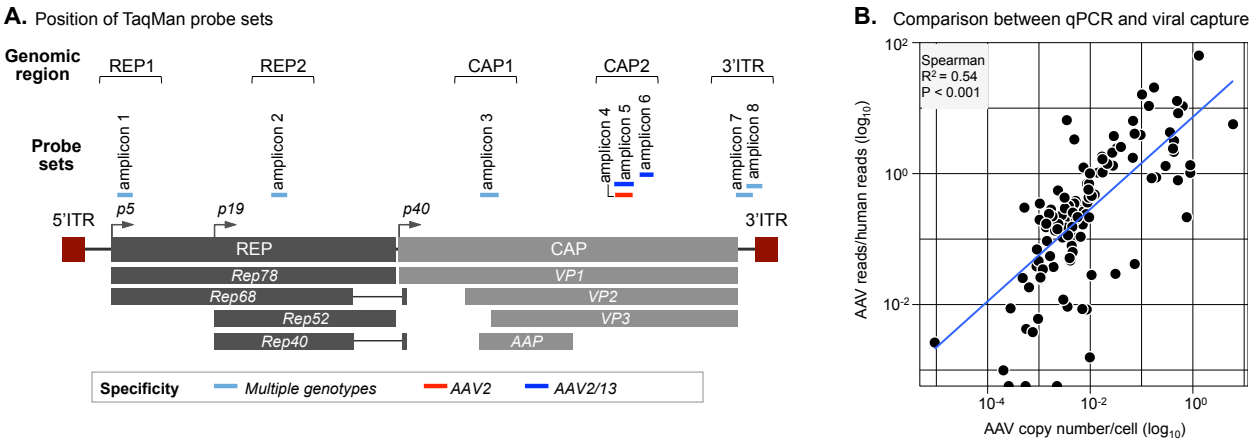
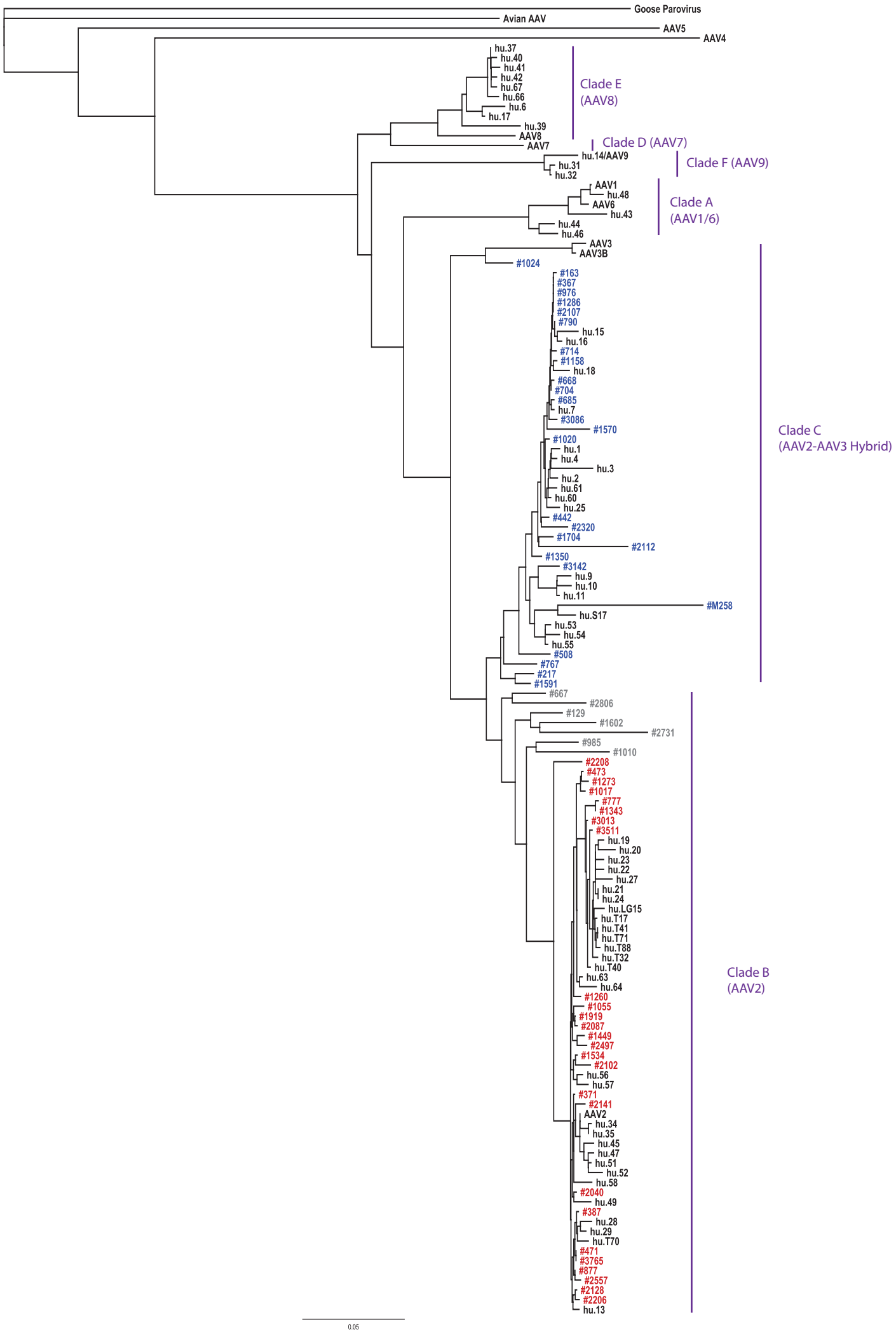
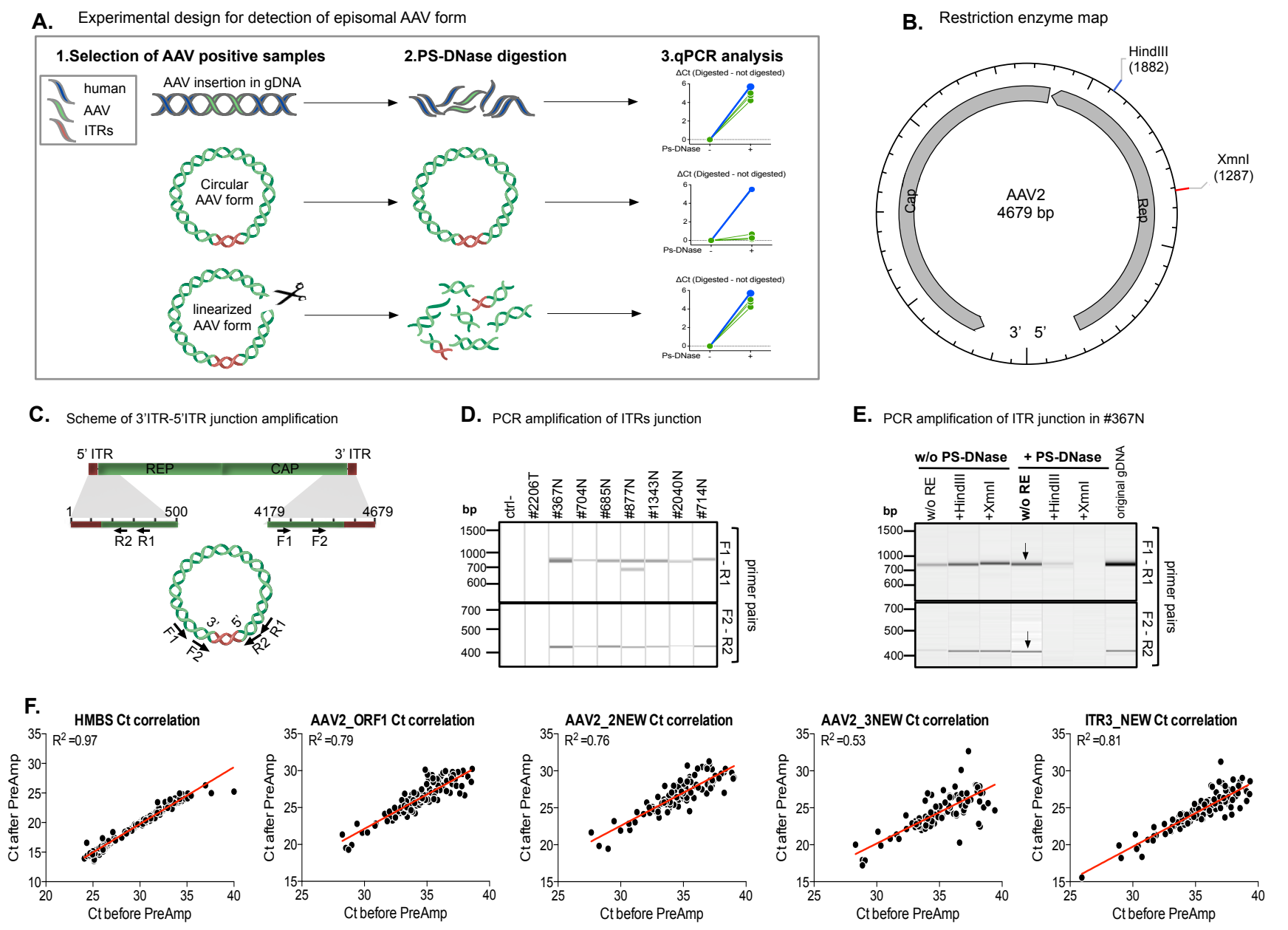
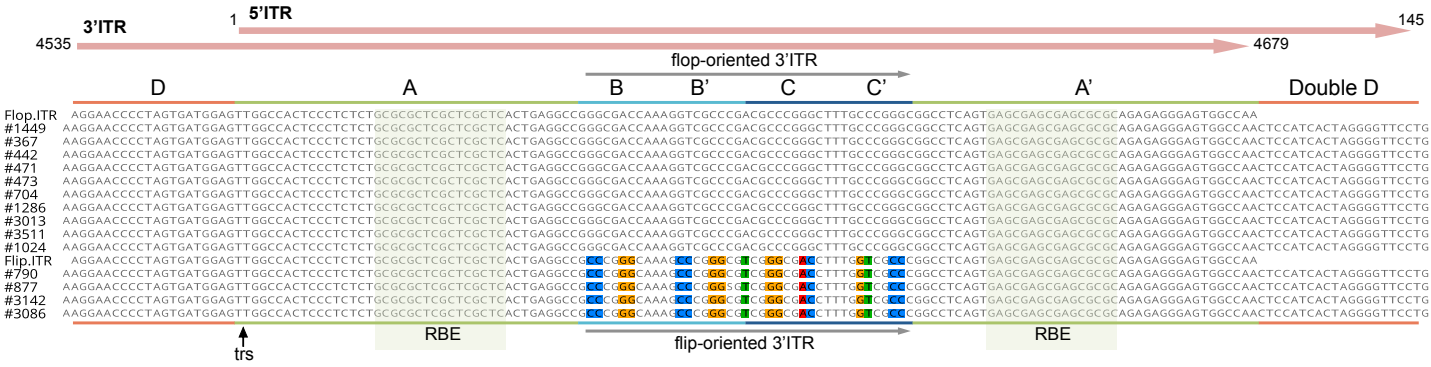

AAV genotypes
Helper viruses
General population
Seroprevalence
AAV2
AAV2/13
Other
45 to 80%
Worldwide prevalence of AAV antibody in human
30-80%
?
AAV1 20-70%
AAV3 20-50%
AAV5 25-40%
AAV6 20-50%
AAV8 20-50%
AAV9 20-60%
HHV6
EBV
AdV
CMV
HHV7
100
80
60
40
20
0
Cohort of livers
DNA in non-tumor liver
17% 29%
AAV positive patients 37% with random non-clonal insertion
48%
52%
no
0.5% 5.2%
Episomal transcribed AAV in the liver
52%
48%
no
DNA in tumor liver
0.0026%*
Oncogenic clonal AAV insertion in HCC
55%
45%
no
45%
6%
0.4%
0%
2%
68%
11%
0%
0%
0%
0%
17%
0%
0%
0%

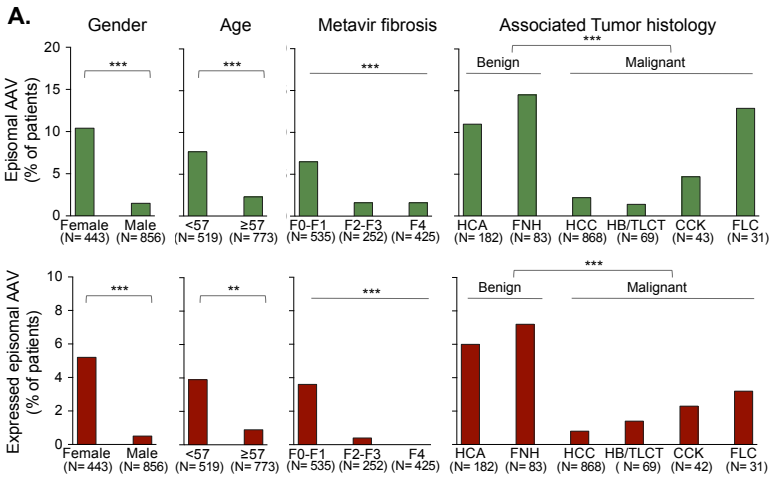
Figure 7

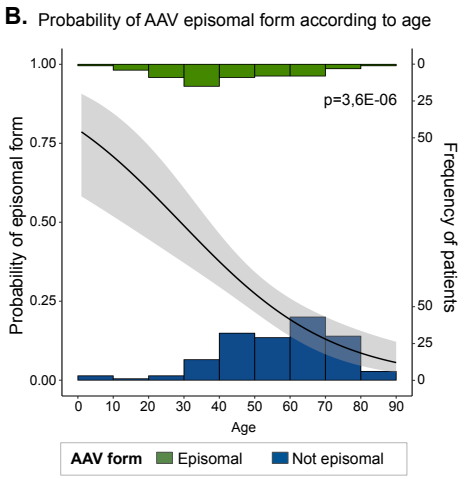

A. Position of TaqMan probe sets
Genomic region
REP1
REP2
CAP1
CAP2
3'ITR
Probe sets
amplicon 1
amplicon 2
amplicon 3
amplicon 4
amplicon 5
amplicon 6
amplicon 7
amplicon 8
5'ITR
p5
p19
p40
3'ITR
REP
CAP
Rep78
VP1
Rep68
VP2
Rep52
VP3
Rep40
AAP
Specificity
Multiple genotypes
AAV2
AAV2/13
B. Comparison between qPCR and viral capture
Spearman
R² = 0.54
P < 0.001
AAV reads/human reads (log10)
AAV copy number/cell (log10)
10²
10⁰
10⁻²
10⁻⁴
10⁻²
10⁰

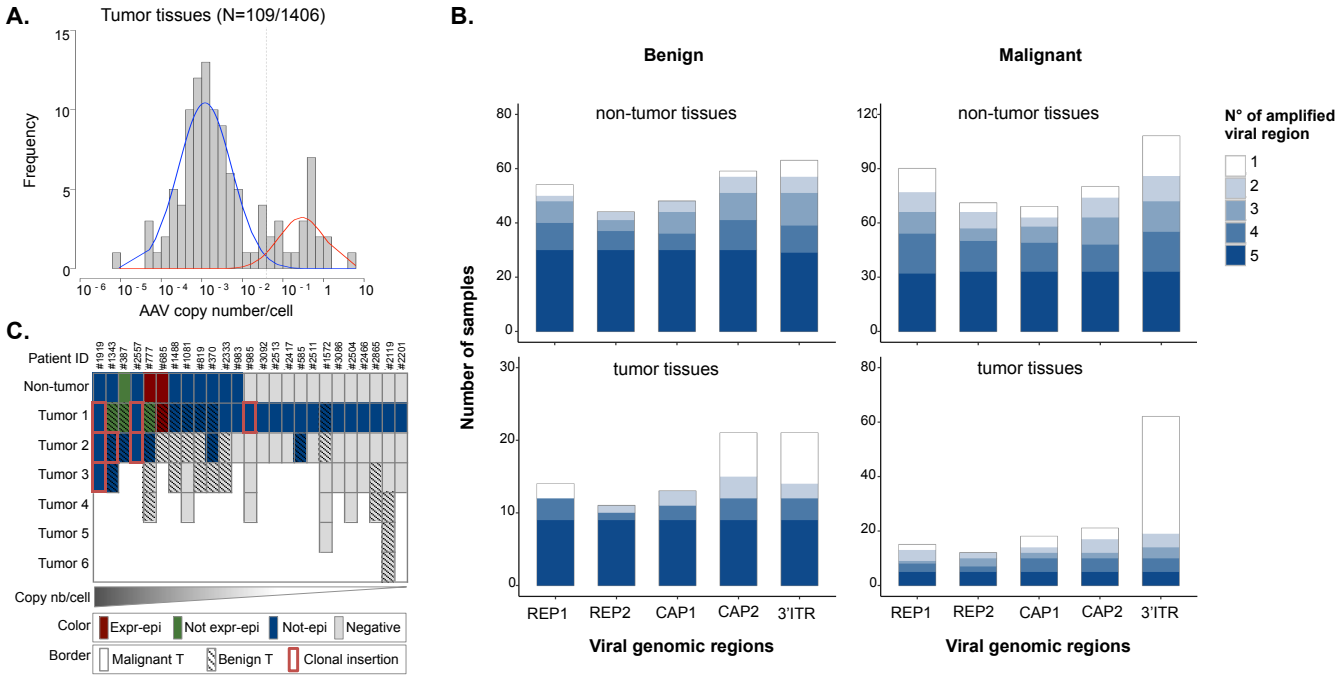

Goose Parovirus
Avian AAV
AAV5
AAV4
hu.37
hu.40
hu.41
hu.42
hu.67
hu.66
hu.6
hu.17
hu.39
AAV8
AAV7
Clade E
(AAV8)
Clade D (AAV7)
hu.14/AAV9
hu.31
hu.32
Clade F (AAV9)
AAV1
hu.48
AAV6
hu.43
hu.44
hu.46
Clade A
(AAV1/6)
AAV3
AAV3B
#1024
#163
#367
#976
#1286
#2107
#790
hu.15
hu.16
#714
#1158
hu.18
#668
#704
#685
hu.7
#3086
#1570
#1020
hu.1
hu.4
hu.3
hu.2
hu.61
hu.60
hu.25
#442
#2320
#1704
#2112
#1350
#3142
hu.9
hu.10
hu.11
#M258
hu.S17
hu.53
hu.54
hu.55
#508
#767
#217
#1591
Clade C
(AAV2-AAV3 Hybrid)
#667
#2806
#129
#1602
#2731
#985
#1010
#2208
#473
#1273
#1017
#777
#1343
#3013
#3511
hu.19
hu.20
hu.23
hu.22
hu.27
hu.21
hu.24
hu.LG15
hu.T17
hu.T41
hu.T71
hu.T88
hu.T32
hu.T40
hu.63
hu.64
#1260
#1055
#1919
#2087
#1449
#2497
#1534
#2102
hu.56
hu.57
#371
#2141
AAV2
hu.34
hu.35
hu.45
hu.47
hu.51
hu.52
hu.58
#2040
hu.49
#387
hu.28
hu.29
hu.T70
#471
#3765
#877
#2557
#2128
#2206
hu.13
Clade B
(AAV2)
0.05

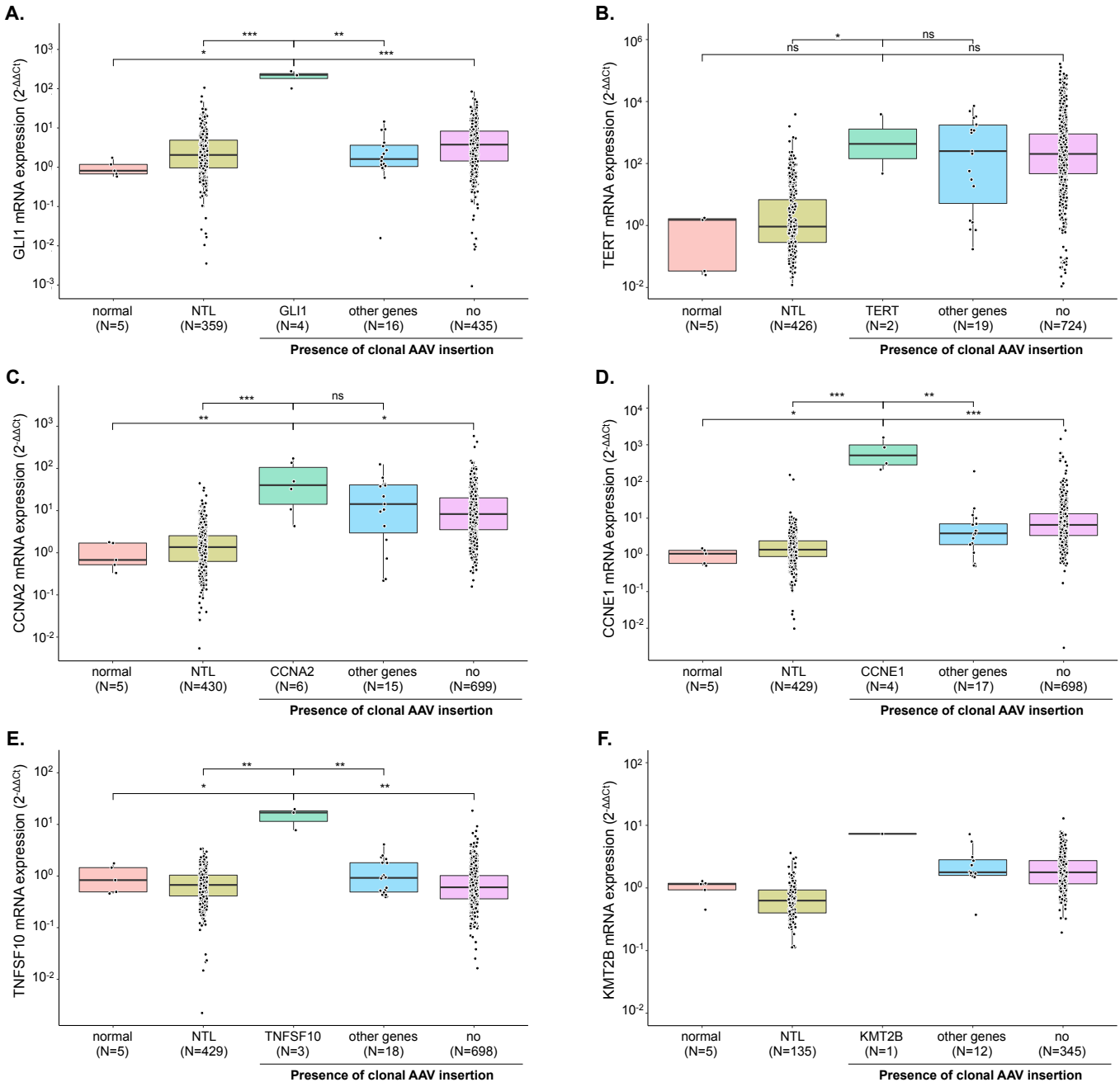

A.
Experimental design for detection of episomal AAV form
1.Selection of AAV positive samples
2.PS-DNase digestion
3.qPCR analysis
human
AAV
ITRs
AAV insertion in gDNA
Circular AAV form
linearized AAV form
ΔCt (Digested - not digested)
Ps-DNase
B.
Restriction enzyme map
HindIII (1882)
XmnI (1287)
AAV2 4679 bp
Cap
Rep
3' 5'
C.
Scheme of 3'ITR-5'ITR junction amplification
5' ITR
3' ITR
REP
CAP
1
500
4179
4679
R2 R1
F1 F2
D.
PCR amplification of ITRs junction
bp
ctrl-
#2206T
#367N
#704N
#685N
#877N
#1343N
#2040N
#714N
1500
1000
700
600
500
400
F1 - R1
F2 - R2
primer pairs
E.
PCR amplification of ITR junction in #367N
w/o PS-DNase
+ PS-DNase
w/o RE
+HindIII
+XmnI
original gDNA
F.
HMBS Ct correlation
AAV2_ORF1 Ct correlation
AAV2_2NEW Ct correlation
AAV2_3NEW Ct correlation
ITR3_NEW Ct correlation
R2 =0.97
R2 =0.79
R2 =0.76
R2 =0.53
R2 =0.81
Ct after PreAmp
Ct before PreAmp

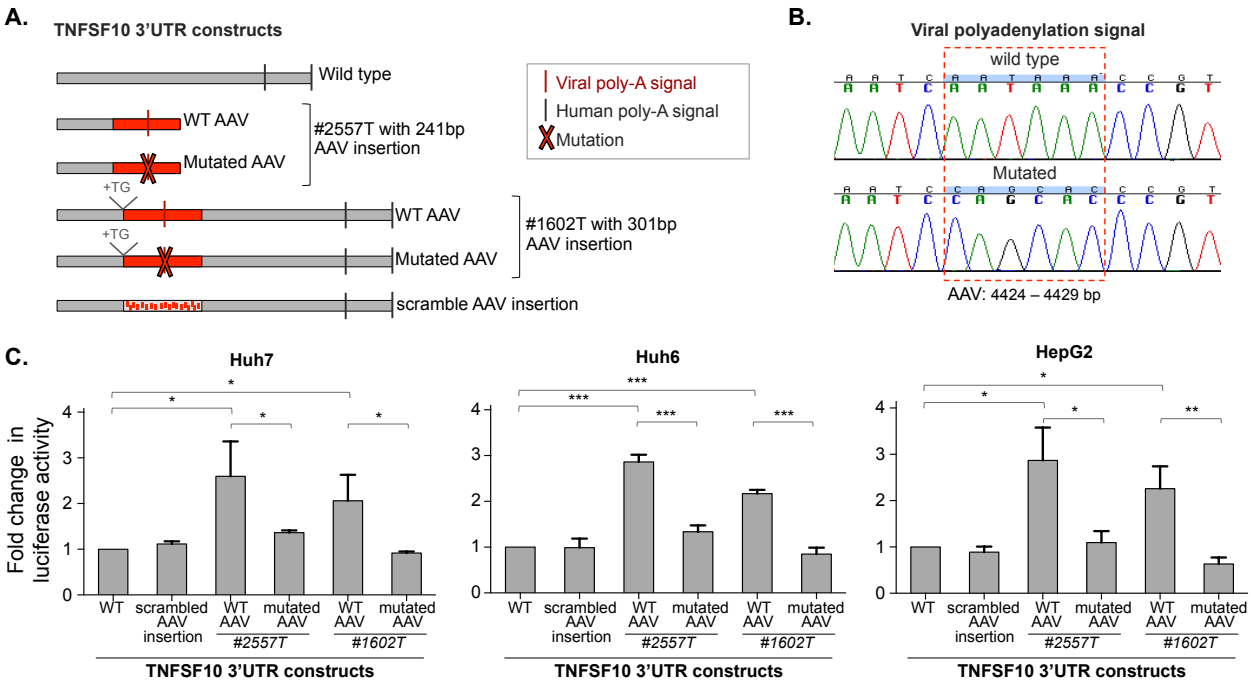

3'ITR
4535
1
5'ITR
145
4679
flop-oriented 3'ITR
D
A
B
B'
C
C'
A'
Double D
Flop.ITR
#1449
#367
#442
#471
#473
#704
#1286
#3013
#3511
#1024
Flip.ITR
#790
#877
#3142
#3086
trs
RBE
flip-oriented 3'ITR
RBE


A.
Gender
Age
Metavir fibrosis
Associated Tumor histology
Episomal AAV (% of patients)
Expressed episomal AAV (% of patients)
***
**
Benign
Malignant
Female (N= 443)
Male (N= 856)
<57 (N= 519)
≥57 (N= 773)
F0-F1 (N= 535)
F2-F3 (N= 252)
F4 (N= 425)
HCA (N= 182)
FNH (N= 83)
HCC (N= 868)
HB/TLCT (N= 69)
CCK (N= 43)
FLC (N= 31)
CCK (N= 42)
HB/TLCT (N= 69)

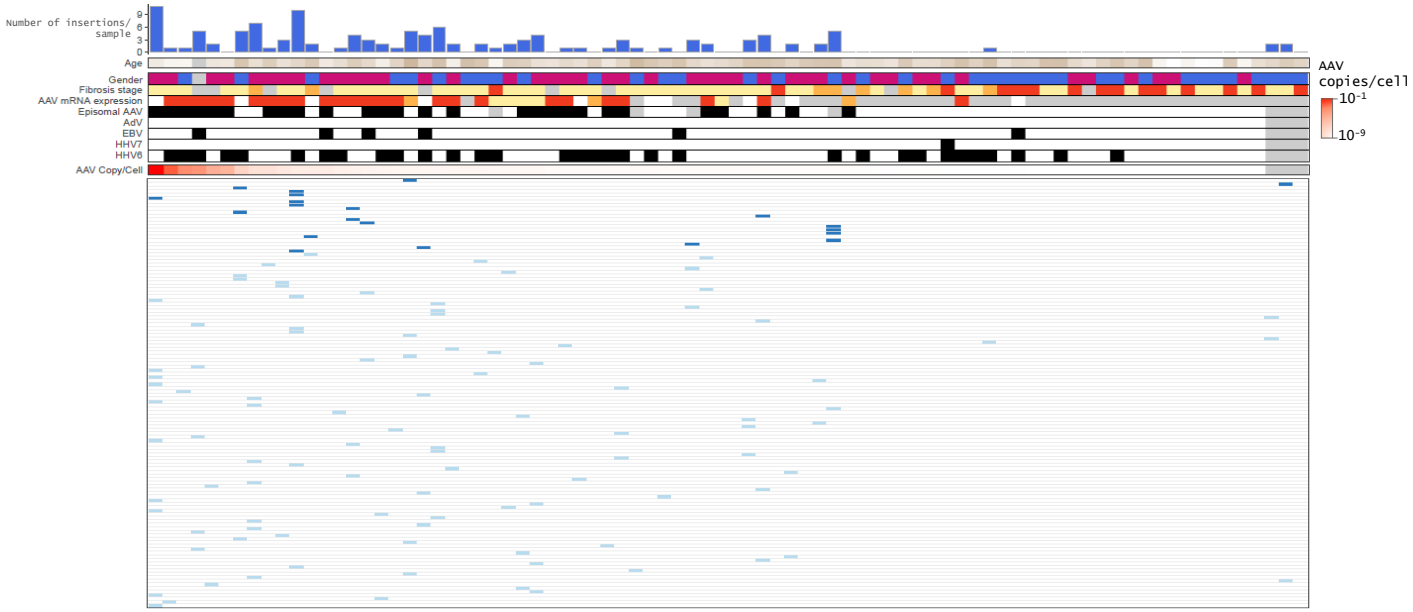

B. Probability of AAV episomal form according to age
p=3,6E-06
Probability of episomal form
Frequency of patients
Age
AAV form
Episomal
Not episomal


A.
Tumor tissues (N=109/1406)
Frequency
AAV copy number/cell
B.
Benign
Malignant
non-tumor tissues
tumor tissues
Number of samples
N° of amplified viral region
REP1
REP2
CAP1
CAP2
3'ITR
Viral genomic regions
C.
Patient ID
Non-tumor
Tumor 1
Tumor 2
Tumor 3
Tumor 4
Tumor 5
Tumor 6
Copy nb/cell
Color
Expr-epi
Not expr-epi
Not-epi
Negative
Border
Malignant T
Benign T
Clonal insertion

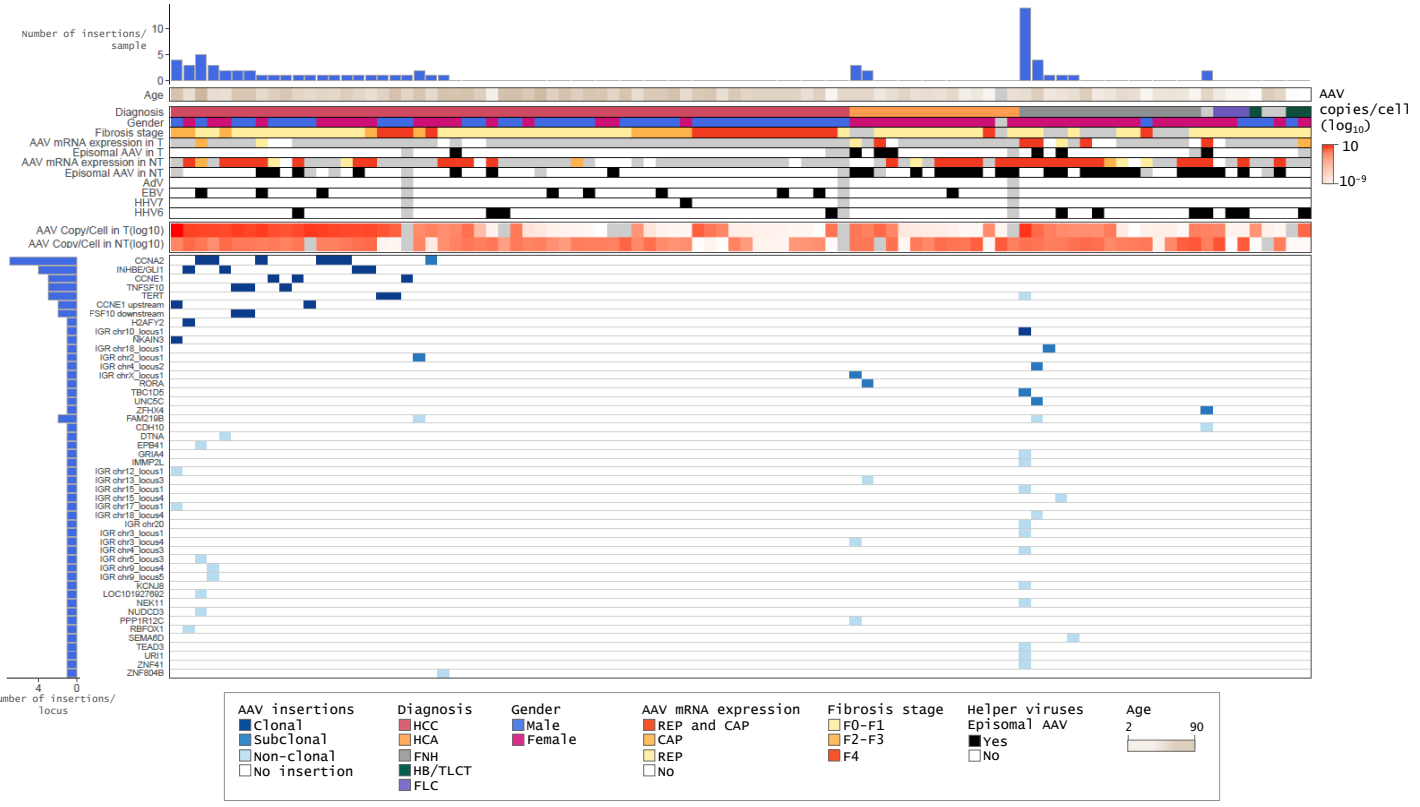

A.
GLI1 mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=359)
GLI1 (N=4)
other genes (N=16)
no (N=435)
Presence of clonal AAV insertion
B.
TERT mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=426)
TERT (N=2)
other genes (N=19)
no (N=724)
Presence of clonal AAV insertion
C.
CCNA2 mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=430)
CCNA2 (N=6)
other genes (N=15)
no (N=699)
Presence of clonal AAV insertion
D.
CCNE1 mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=429)
CCNE1 (N=4)
other genes (N=17)
no (N=698)
Presence of clonal AAV insertion
E.
TNFSF10 mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=429)
TNFSF10 (N=3)
other genes (N=18)
no (N=698)
Presence of clonal AAV insertion
F.
KMT2B mRNA expression (2-ΔΔCt)
normal (N=5)
NTL (N=135)
KMT2B (N=1)
other genes (N=12)
no (N=345)
Presence of clonal AAV insertion


A.
TNFSF10 3'UTR constructs
Wild type
WT AAV
Mutated AAV
#2557T with 241bp AAV insertion
+TG
WT AAV
+TG
Mutated AAV
#1602T with 301bp AAV insertion
scramble AAV insertion
Viral poly-A signal
Human poly-A signal
Mutation
B.
Viral polyadenylation signal
wild type
Mutated
AAV: 4424 – 4429 bp
C.
Huh7
Huh6
HepG2
Fold change in luciferase activity
WT
scrambled AAV insertion
WT AAV
mutated AAV
WT AAV
mutated AAV
#2557T
#1602T
TNFSF10 3'UTR constructs

A. AAV insertion in non-tumor liver tissues (N=82)



B. AAV insertion in tumor liver tissues (N=94)



Supplementary Information

Adeno-associated virus in the liver: natural history and consequences in tumor development

La Bella T, Imbeaud S. et al.

Supplementary Tables

Supplementary Table 1 : Description of the series (n=1461 patients)
Supplementary Table 2: List of probe setes and primers
Supplementary Table 3: Results of the viral DNA screening
Supplementary Table 4: List of AAV clonal inserted sites in liver tumors

Supplementary Figures

Supplementary figure 1. AAV detection using quantitative PCR and viral capture.
Supplementary figure 2. Phylogenetic tree on VP1 coding protein region.
Supplementary figure 3. Viral episomal form investigation.
Supplementary figure 4. Viral 3'ITR-5'ITR episomal junction determined by in silico analysis.
Supplementary figure 5. AAV presence in tumor tissues and comparison with non-tumor counterparts.
Supplementary figure 6. AAV presence in tumor tissues and comparison with non-tumor counterparts.
Supplementary figure 7. Impact of clonal AAV integration on the expression of the target oncogenes.
Supplementary figure 8. Viral polyadenylation signal usage in TNFSF10.
Supplementary figure 9. Description of AAV insertions in non-tumor and tumor liver tissues.

Supplementary Materials and Methods

Computational analysis of human-AAV sequences

Experimental procedure to search for episomal forms of AAV

Supplementary table 1. Description of the series (n=1461 patients)

			Available data	n (%) or median (range)
HCC patients (n=936)	**Gender** (female)		934	163 (17.5%)
	Age		930	63 (0-90)
	Transformed HCA (HCC on HCA)		936	17 (1.8%)
	Largest nodule diameter (>5cm)		909	425 (46.8%)
	Number Nodules (≥2)		871	244 (28%)
	Preoperative serum AFP (>20ng/ml)		727	293 (40.3%)
	Risk factors	Alcohol	895	393 (43.9%)
		HBV	907	164 (18.1%)
		HCV	896	224 (25%)
		Metabolic syndrome	843	201 (23.8%)
		Without etiology	912	128 (14%)
	Metavir score	F0-F1	929	264 (28.4%)
		F2-F3	929	236 (25.4%)
		F4	929	439 (46.2%)
	G1-G6 classification	G1	596	44 (7.4%)
		G2	596	47 (7.9%)
		G3	596	105 (17.6%)
		G4	596	210 (35.2%)
		G5	596	129 (21.6%)
		G6	596	61 (10.2%)
	Tumor differentiation	Edmonson I-II	869	381 (43.8%)
		Edmonson III-IV	869	488 (56.2%)
	Vascular invasion	Macrovascular	827	127 (15.4%)
		Microvascular	827	287 (34.7%)
HCA and FNH patients (n=322)	**Gender** (female)		320	270 (84.4%)
	Age		312	37 (2-68)
	Borderline tumors (HCA/HCC)		322	14 (4.3%)
	Largest nodule diameter (>5cm)		221	136 (61.5%)
	Number Nodules (≥2)		216	100 (46.3%)
	BMI (>25kg/m^2)		182	69 (37.9%)
	Children (women only)		132	100 (75.8%)
	Oral contraception (women only)		147	128 (87.1%)
	Diabetes		174	19 (10.9%)
	Glycogenesis		160	7 (4.4%)
	Metavir score	F0-F1	229	210 (91.7%)
		F2-F3	229	11 (4.8%)
		F4	229	8 (3.5%)
	Molecular subgroups	HHCA	314	62 (19.7%)
		b$^{\varepsilon\alpha\gamma,\upsilon}$-HCA	314	7 (2.2%)
		b$^{\varepsilon\alpha\gamma,\upsilon}$-IHCA	314	14 (4.5%)
		IHCA	314	66 (0,21)
		b$^{\varepsilon\alpha\upsilon}$-IHCA	314	24 (7.6%)
		b$^{\varepsilon\alpha\upsilon}$-HCA	314	19 (6.1%)
		shHCA	314	10 (3.2%)
		UHCA	314	15 (4.8%)
		FNH	314	97 (30.9%)
Other tumors	**Histological tumor type**	HB/TLCT	203	87 (42.9%)
		CCK	203	46 (22.7%)
		FLC	203	36 (17.7%)
		others*	203	34 (16.7%)
	*DMN, LGDN, HGDN, angiomyolipoma, liver carcinosarcoma, liver rhabdoid tumor, solitary fibrous tumor, mesenchymal Hamartoma, embryonal sarcoma, leiomyoma, Yolk-salk-tumor, neuroendocrine carcinoma, neuroblastoma, pecoma, nephoblastoma, malignant peripheral nerve sheath tumor			

Supplementary table 2. List of probe setes and primers

Probe Name	Forward sequence 5'-3'	Reverse sequence 5'-3'	Probe sequence 5'-3'	Target	used for
AAV2 ORF1	CGGCATTTCTGACAGCTTTG	GGGTGCCTGCTCAATCA	TGGCCGAGAAGGAATGGGAGTTG	AAV	qPCR, DNase/TaqMan assay
AAV2 2 NEW	CTTCTACGGGTGCGTAAACT	GACCTTGGCGGTCATCTT	AGATGGTGATCTGGTGGGAGGAGG	AAV	qPCR, DNase/TaqMan assay
AAV2 3 NEW	AGACGCAGACTCAGTACCT	CCCTCGTTATTGTCTGCCAT	AATACGATGGCTACAGGCAGTGGC	AAV	qPCR, DNase/TaqMan assay
AAV2 4 NEW	TGACATTCGGGACCAGTCTA	TAGCTCCAGTCCACGAGTATT	TTACCGCCAGCAGCGAGTATCAAA	AAV2	qPCR, DNase/TaqMan assay
AAV2 ITR3 NEW	CCTGACTCGTAATCTGTAATTGC	ACGTAGCCATGGAAACTAGATAAG	TCGTTTCAGTTGAACTTTGGTCTCTGC	AAV	qPCR, DNase/TaqMan assay
AAV2 ITR3	GGTCTCTGCGTATTTCTTTCTTATC	GTGGCCAACTCCATCACTA	ACGTAGATAAGTAGCATGGCGGGT	AAV	qPCR, DNase/TaqMan assay
AY6 2	CTATGGCCAGCCACAAAGA	GTCGTTGGCATTTGTTCCTT	CCCCATGCATGGAACCCTGATA	AAV2/13	qPCR, DNase/TaqMan assay
AY6 1	CCACCAGCATGTCTCTTCAA	ATACTTTGTAGCCGCAGTCCA	AACTGGCTGCCTGGACCTTGCTA	AAV2/13	qPCR, DNase/TaqMan assay
AAV HidIII	CCCCAGTGACGCAGATATAAGT	CCACGTGACGAGAACATTTG	CAGACGCGGAAGCTTCGATCA	AAV	DNase/TaqMan assay
AAV XmnI	TGGAACTGAACGGGTACGAT	ACAGCCAGATGGTGTTCCTC	TCTGGGATGGGCCACGAAAA	AAV	DNase/TaqMan assay
AdV hexon1C	GCTACCCCTTCGATGATGC	ACCGTGGGGTTTCTAAACTTG	AGTGGTCTTACATGCACATCTCGGG	AdV	qPCR
AdV hexon2C	CACGCGCTAGACATGACTTTT	ACCACGTCAAAGACTTCAAACA	CATGGACGAGCCCACCCTTCTTT	AdV	qPCR
AdV hexonACF	CGCAGTGGTCTTACATGCAC	ACCGTGGGGTTTCTAAACTTG	CAGGACGCCTCGGAGTACCTGA	AdV	qPCR
AdV hexon1B	CCGTGCAACAGACACCTACTT	TGTAAGAGTATGTATTGTCCTCCCG	ACCCACGATGTGACCACCGAC	AdV	qPCR
AdV hexon2B	GGACATGACTTTTGAGGTGGA	GTGGCTGGTGCACTCTGA	CCCACCCTGCTTTATCTTCTTTTCGA	AdV	qPCR
AdV hexon1F	AACACGGAGCTGTCTTACCAGT	GTCTGGGTCATAGCTGTCCAC	TCGATACTTCTCCATGTGGAACCAGG	AdV	qPCR
AdV hexon2F	TCCTTTGGACAAGCTCCCTA	GCTCCGGTTGGTACGTTTTAT	CAAGTGGGCTCAGACTCCAACAATC	AdV	qPCR
AdV hexon1A	CATACCAGTTAATGCTGGATGCT	CATCCTCTACCCCGTGATTCT	ACGACCCTGACGTTCGCGTTATT	AdV	qPCR
AdV hexon2A	CTCCCACCCATGATGTTACC	GCGTAAAGCGAGCCTTGTAG	TGCGTTTTGTGCCCGTGGAT	AdV	qPCR
AdV hexon1D	GGTCTGGTGCAGTTTGCC	ACCGTGGGGTTCCTAAACTT	CCACCGACACGTACTTCAGCCTG	AdV	qPCR
AdV hexon2D	GAGTTCCCTCGGAAACGAC	TTGAAGGACTGGTCGTTGGT	AACCTCTACGCCACATTCTTCCCCA	AdV	qPCR
AdV hexon1E	AGTGCAACATGACCAAGGACT	AGTTGCGGAAGAAGGAGTACAT	CTGGCCCACTACAACATCGGCT	AdV	qPCR
AdV hexon2E	CGCTTCGGAGTACCTGAGTC	TGGGGTTCCTAAACTTGTTCC	CGCGCCACAGACACCTACTTCAGT	AdV	qPCR
AdV polBE	CCCATCCAGGTGTAGGTTTC	CACATCAACAGCCATTCCTC	AAGAACTGGATCTCCTGCCACCAGT	AdV	qPCR
AdV polD	CCCATCCAGGTGTAGGTTTC	TTCATCACATCAACAGCCACT	AAGAACTGGATTTCCTGCCACCAGT	AdV	qPCR
AdV polA	CCCATCCAGGTGTAGGTTTC	GATCCACTTTTTCCCAATCG	AAGGCGCTCAGTGCGAGGATG	AdV	qPCR
AdV polF	CCAGCTGTTCGGGTGAGTAT	ATCAAATCCACCTCGTTTGTG	CGCGTTGAAAGGTGGGCATAAC	AdV	qPCR
AdV polC	GCGGTTGAGGACAAACTCTT	TTCTACATGCTAGGCTCTTACCG	TCCAGTACTCTTGGATCGGAAACCC	AdV	qPCR
HSV 1.1	TCCTGGCTCTGCGAGTAGTT	GTTCGTGTCGGTCAAGGAGT	CGTTGGCCGTGAGCCACTTT	HSV1/2	qPCR
HSV 1.2	TCTGGGAGTAGTTGGTGATGC	CGTTTGTGTCGGTCAAAGAGT	GCGTTGGCCGTGATCCACTT	HSV1/2	qPCR
HSV 2.1	TTTGACTACGACCAGAAGTTGC	CTCCGTGACATACAGGGTCAT	CCCCTACGTCGACCATAGCCAATC	HSV1/2	qPCR
HSV 2.2	ACTGCTGATCGACCTGTGTG	TTATATGCGTGGACGAGAAGG	CCCCCTATGTTGATCACAGCCAA	HSV1/2	qPCR
HSV 3	AGGCAGACGTTTGACTTTGG	TGGGTTGTAGCGTGCTGTC	GCGTGCGTGCTGATGGTGAA	HSV1/2	qPCR
HSV 4	GAGGGACATCCAGGACTTTGT	CGGGCCATGAGCTTGTAA	ACCGCCGAACTGAGCAGACAC	HSV1/2	qPCR
HHV4 1	TGGAGCGAAGGTTAGTGGTC	GGGGAAAATTGTCACATTGG	TGCAGTGCCTGGTGCTGCTTTA	EBV/HH	qPCR
HHV4 2	CTGAGGCCCTTCTTCCTTTT	CCTGAACAGCCTTGGATAGC	TTTGAGGGGTGGGGGAATATGG	EBV/HH	qPCR
HHV4 3	CAGCAGTGAGCGGTAAAACA	CCTACAAGGACCTGGTCAAGAG	CGCAAAGGGGTTGCACGAGT	EBV/HH	qPCR
HHV4 4	TAGTGTCCGGGAATAGGGTCT	CCACTACCAGGAGGGAGAACTA	TCATGTAGGGCCCGTTGATGATG	EBV/HH	qPCR
HHV5 1	AGGCTTTCTACCTCATCATCCA	TGGGAATCCGTAACCACAAC	GGAAACGCGTCGATGCTTTTATCC	CMV/HH	qPCR
HHV5 2	GTTAGGTGACACGGCAACACT	ATTTTATCCTGCCCAACCACT	TTTCGGGCACCGTCGTACCTTT	CMV/HH	qPCR
HHV5 3	TGGCCAAACGTTCGAGTT	ACGTGTCCGTCTTTGAAACC	TGCCAGAACACTACCAAACCACCA	CMV/HH	qPCR
HHV5 4	TCAAACGTGCGCTTATCTTG	AAGACGTACAGCAGCGGATT	CGAGTCGCTCGCCTTTTGTCACT	CMV/HH	qPCR
HHV5 5	CATCTCCGTACTTGAGGGTAGTG	CTCTCAACGCGTAACTTCTTCC	CGTTAACACCATGGCTGACCGTTT	CMV/HH	qPCR
HHV6 1	GATAGTATCCCTCTTCTCCCCATC	ACGGTTAACTTTTGAGGACGAA	CCAGAATCGAGAAGTGCGCCAG	HHV6	qPCR
HHV6 2	CATCTCCTCCGTTATGTCTTTCA	GGAGTTATTGGACCACGACATT	AGCGAGAGATAGGGATGGTTGGGA	HHV6	qPCR
HHV6 3	TCTAATAGAGCTTGGTTGGATGAA	GCGAAATACAATACTCGTGCTCT	GGCAGAGCGTTTTGTAAGAACTGGC	HHV6	qPCR
HHV6 4	CCGCACTATGTCGTGGATTAG	ACTGGATCTGTACTGTAGGAATCGTT	TATAAGTTCGGGGGCCGTATGGGTG	HHV6	qPCR
HHV6 5	CCGCACTATGTCGTGGATTAG	GGATCTGTACCGTAGGAATCGTT	TAAGTTCGGGGCCGTAGGGGT	HHV6	qPCR
HHV7 1	CATCCAGAGCGTAGACAGCA	GCCGTGGTTATCGGAAAGT	AAGACCACGACCGAGGCATCTTC	HHV7	qPCR
HHV7 2	GTCCGTTAGAACCTCCATCAAC	TTGGCTGTACCACGAATCAC	TCAAGGCCTGAAAAAGCGAACGT	HHV7	qPCR
HHV7 3	GGATTTAGGTAGAGTTGTGGCG	CATTTAGTGGCTGTTCACTTTCC	TGCGGATGTCATTCCGTATAGAGAGG	HHV7	qPCR
HHV7 4	AGTTGGCCGATACGACACA	TCACCGGTTCTTCTGTTCCTAC	CCGGAAGGCTGGATGGTAACTCTAGG	HHV7	qPCR
HHV8 1	AGAGTCCGCCATCACCAA	ACCCCGTTGACATTTACCTTC	ACCCGTGCCACTCTATGAGATAAGCC	HHV8	qPCR
HHV8 2	CTGGGCAAGCAGTTTTTCA	TTCGGTAGGTGAGGTCTTTGTC	TCAGGAAGACGGCCTAGAGCGATAC	HHV8	qPCR
HHV8 3	ATCTCGTCTTCCATCCTTTCC	GTTCCAGTTACCCACCCAAA	TTGTGGCCTAGCTTTCGCAGGAC	HHV8	qPCR
HHV8 4	GTCTTCCATCCTTTCCCAAAA	ACCGGTCCCTTACAGACAGATA	TGGCCTAGCTTTCGCAGGACA	HHV8	qPCR
HMBS	GTCCAGCCTGTTGGTGAG	GGCCACAATTCAGATCTTCCTA	CTCTTGCAGTCATTCCAGAGCCCT	human	qPCR
RHO	CCACACAGAAGGCAGAGAAG	TGGTGGGTGAAGATGTAGAATG	TCATCGCTTTCCTGATCTGCTGGG	human	qPCR
HPV E1		Vi03453396_s1		HPV16	qPCR
ANKRD49		Hs06279676_cn		human	qPCR
18S		Hs03928990_g1		human	RT-qPCR
ACTB		Hs01060665_g1		human	RT-qPCR
CCNA2		Hs00996788_m1		human	RT-qPCR
CCNE1		Hs01026536_m1		human	RT-qPCR
TERT		Hs00972656_m1		human	RT-qPCR
GLI1		Hs00171790_m1		human	RT-qPCR
KMT2B		Hs00207065_m1		human	RT-qPCR
TNFSF10		Hs00921974_m1		human	RT-qPCR
Primer name	**Sequence 5'-3'**			**Target**	**used for**
AAV2 tot 6 F b	ACAGTACTCCACGGGACAGG			AAV	ITRs junction PCR and seque
AAV2 epi R 2	GCTGGGGACCTTAATCACAA			AAV	ITRs junction PCR and seque
AAV2 6F NEW	CCTGACTCGTAATCTGTAATTGC			AAV	ITRs junction PCR and seque
AAV2 ITR5 F int	GTGGCCAACTCCATCACTAG			AAV	ITRs junction PCR and seque

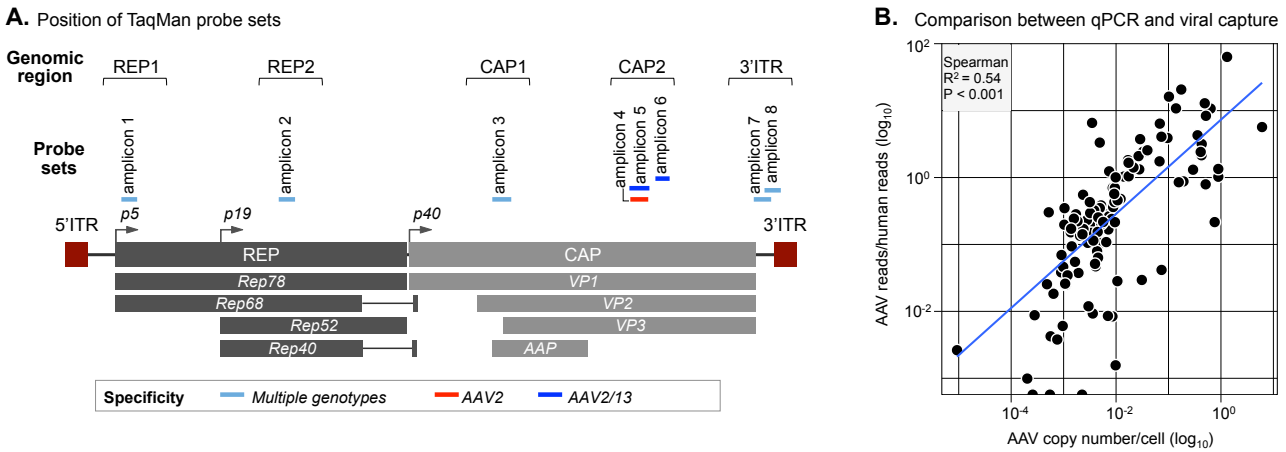
Supplementary table 3. Results of the viral DNA screening

		AAV positivity	Copy nb/cell (mean [sd])	Episomal AAV	Expressed episomal AAV	AAV subtype			HHV6 co-infection
						AAV2	AAV2-AAV3/13	Unknown	
All samples	**NT**	17.7% (233/1319)	0.007 [± 0.021]	4.6% (60/1304)	2.1% (28/1304)	29.2% (68/233)	32.2% (75/233)	38.6% (90/233)	7.9% (104/1318)
	T	7.8% (109/1406)	0.141 [± 0.617]	0.6% (8/1397)	0.2% (3/1397)	22.9% (25/109)	31.2% (34/109)	45.9% (50/109)	1.1% (15/1405)
Benign	**NT**	27% (78/289)	0.008 [± 0.021]	12% (34/284)	6% (17/284)	37.2% (29/78)	38.5% (30/78)	24.3% (19/78)	12.5% (36/289)
	T	9.4% (31/331)	0.048 [± 0.234]	1.8% (6/327)	0.9% (3/327)	41.9% (13/31)	38.7% (12/31)	19.4% (6/31)	1.8% (6/331)
Malignant	**NT**	15% (155/1030)	0.007 [± 0.021]	2.5% (26/1020)	1.1% (11/1020)	25.2% (39/155)	29% (45/155)	45.8% (71/155)	6.6% (68/1029)
	T	7.3% (78/1075)	0.178 [± 0.712]	0.2% (2/1070)	0% (0/1070)	15.4% (12/78)	28.2% (22/78)	56.4% (44/78)	0.8% (9/1074)
HCA	**NT**	25.3% (47/186)	0.007 [± 0.017]	11% (20/182)	6% (11/182)	36.2% (17/47)	42.5% (20/47)	21.3% (10/47)	12.9% (24/186)
	T	7.9% (17/215)	0.006 [± 0.013]	1.9% (4/215)	0.9% (2/215)	23.5% (4/17)	52.9% (9/17)	23.5% (4/17)	1.4% (3/215)
FNH	**NT**	32.1% (27/84)	0.012 [± 0.029]	14.5% (12/83)	7.2% (6/83)	40.7% (11/27)	29.6% (8/27)	29.6% (8/27)	14.3% (12/84)
	T	12.5% (12/96)	0.106 [± 0.362]	2.2% (2/92)	1.1% (1/92)	75% (9/12)	25% (3/12)	0% (0/12)	2.1% (2/96)
HCC	**NT**	15.4% (135/875)	0.004 [± 0.010]	2.2% (19/868)	0.8% (7/868)	24.4% (33/135)	28.9% (39/135)	46.7% (63/135)	6.5% (57/875)
	T	7.6% (69/909)	0.200 [± 0.755]	0.2% (2/906)	0% (0/906)	15.9% (11/69)	26.1% (18/69)	58% (40/69)	0.8% (7/908)
HB/TLCT	**NT**	9.7% (7/72)	0.020 [± 0.048]	1.4% (1/69)	1.4% (1/69)	42.9% (3/7)	28.6% (2/7)	28.6% (2/7)	2.8% (2/72)
	T	6.7% (5/75)	0.013 [± 0.017]	0% (0/73)	0% (0/73)	20% (1/5)	20% (1/5)	60% (3/5)	2.7% (2/75)
CCK	**NT**	16.3% (7/43)	0.004 [± 0.006]	4.7% (2/43)	2.3% (1/43)	0% (0/7)	28.6% (2/7)	71.4% (5/7)	11.6% (5/43)
	T	4.4% (2/45)	0.001 [± 0.001]	0% (0/45)	0% (0/45)	0% (0/2)	50% (1/2)	50% (1/2)	0% (0/45)
FLC	**NT**	19.4% (6/31)	0.056 [± 0.080]	12.9% (4/31)	3.2% (1/31)	33.3% (2/6)	50% (3/6)	16.7% (1/6)	10% (3/30)
	T	6.2% (2/32)	0.001 [± 0.001]	0% (0/32)	0% (0/32)	0% (0/2)	100% (2/2)	0% (0/2)	0% (0/32)
others	**NT**	14.3% (4/28)	0.002 [± 0.002]	7.1% (2/28)	3.6% (1/28)	50% (2/4)	25% (1/4)	25% (1/4)	3.6% (1/28)
	T	5.9% (2/34)	0.001 [± 0.0002]	0% (0/34)	0% (0/34)	0% (0/2)	0% (0/2)	100% (2/2)	2.9% (1/34)

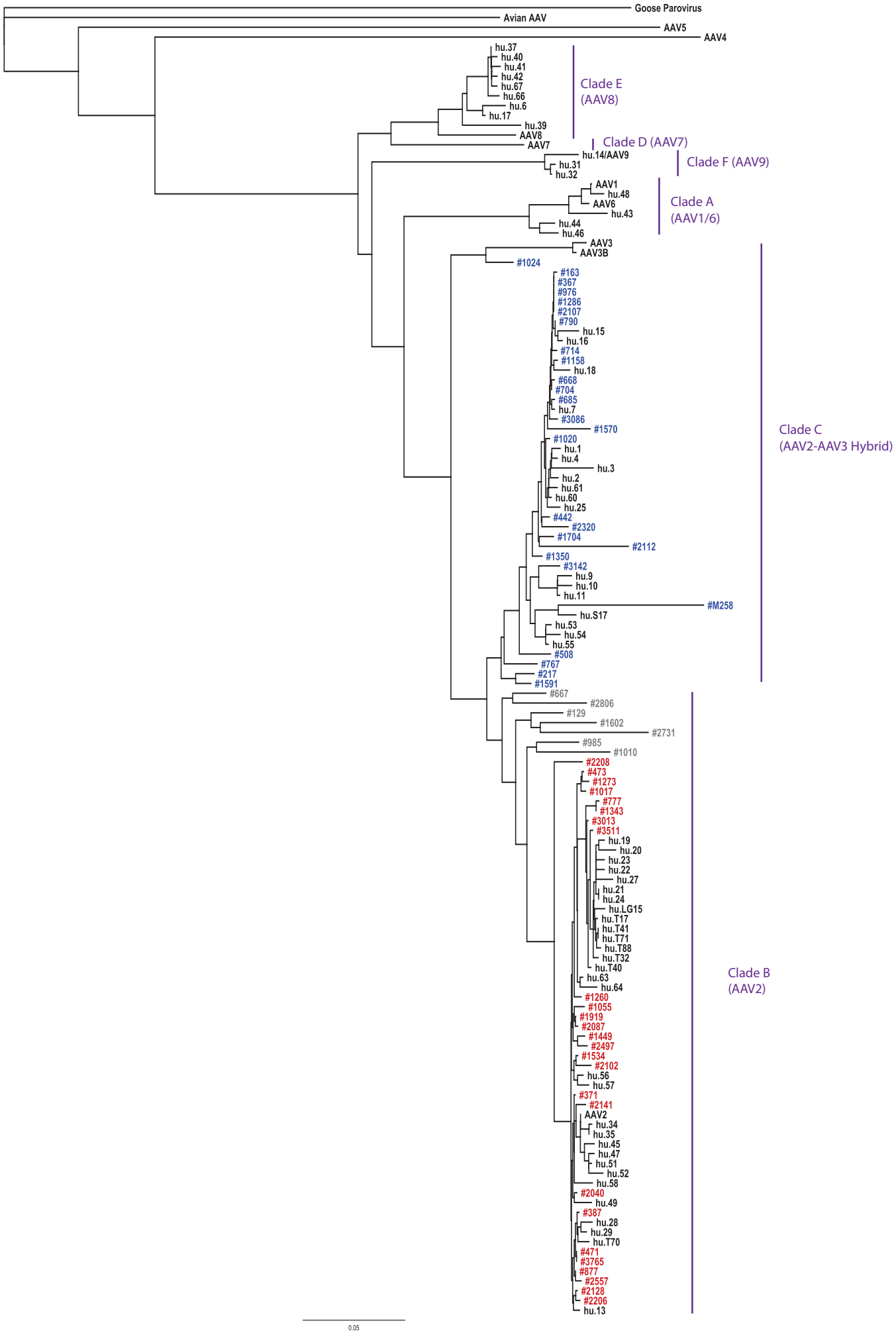
Supplementary table 4. List of AAV clonal inserted sites in liver tumors

Series	Sample	ıan target ç	Virus	Breakpoint 1 on human genome (hg19)	Breakpoint 2 on human genome (hg19)	Breakpoint 1 on viral genome	Breakpoint 2 on viral	Orientation (5'>3')	Orientation (Flip/Flop)	Genotype	Publication
LiC1162	#2128T	CCNA2	AAV2	chr4:122,743,475	chr4:122,743,452	2934	1493	3>5	ND	AAV2	Nault *et al.* 2015
LiC1162	#313T	CCNA2	AAV2	chr4:122,742,417	chr4:122,742,409	2611	4585	5>3	Flop	AAV2	Nault *et al.* 2015
LiC1162	#2206T	CCNA2	AAV2	chr4:122,742,478	chr4:122,742,456	3560	4596	5>3	Flip	AAV2	Nault *et al.* 2015
LiC1162	#2848T	CCNA2	AAV2	chr4:122,742,948	chr4:122,742,929	4630	3730	3>5	Flip	AAV2/13	
LiC1162	#129T	CCNA2	AAV2	chr4:122,743,844	chr4:122,743,835	4597	4316	3>5	Flip	AAV2/13	Bayard et *al*. 2018
LiC1162	#2112T	CCNA2	AAV2	chr4:122,742,667	chr4:122,742,660	4389	4608	5>3	Flip	AAV2/13	Nault *et al.* 2015
TCGA	TCGA-G3-AAUZ	CCNA2	AAV2	ND	chr4:122,743,570	ND	4613	5>3	Flop	AAV2/13	Bayard et al. 2018
TCGA	TCGA-ZS-A9CF	CCNA2	AAV2	chr4:122,742,400	chr4:122,742,166	3407	4514	5>3	ND	AAV2	Bayard et al. 2018
TCGA	TCGA-BC-A10Y	CCNA2	AAV2	chr4:122,743,756	chr4:122,743,742	4064	4586	5>3	ND	AAV2/13	Bayard et al. 2018
LiC1162	#2557T, #2558T	TNFSF10	AAV2	chr3:172,224,150	chr3:172,302,191	4388	4597	5>3	Flop	AAV2	Nault *et al.* 2015
LiC1162	#1602T	TNFSF10	AAV2	chr3:172,224,027	chr3:172,224,026	4270	4571	5>3	Flop	AAV2/13	Nault *et al.* 2015
LiC1162	#1185T	KMT2B	AAV2	chr19:36,212,635	chr19:36,212,644	4602	4247	3>5	Flip	ND	Nault *et al.* 2015
LiC1162	#2141T	CCNE1	AAV2	chr19:30,287,313	chr19:30,287,316	4599	4379	3>5	Flop	AAV2	Nault *et al.* 2015
LiC1162	#1591T	CCNE1	AAV2	chr19:30,303,035	chr19:30,303,053	4340	4598	5>3	Flip	AAV2/13	Nault *et al.* 2015
LiC1162	#2208T	CCNE1	AAV2	chr19:30,304,532	chr19:30,304,542	4604	4237	3>5	Flop	AAV2	Nault *et al.* 2015
LiC1162	#M257T	CCNE1	AAV2	chr19:30,303,740	chr19:30,303,743	4590	4253	3>5	Flop	AAV2/13	Bayard et al. 2018
LiC1162	#3641T	CCNE1	AAV2	chr19:30,291,515	chr19:30,291,518	4239	4534	5>3	ND	ND	Bayard et al. 2018
TCGA	TCGA-BC-A10T	CCNE1	AAV2	chr19:30,302,864	ND	4626	ND	ND	ND	ND	Bayard et al. 2018
LiC1162	#985T	TERT	AAV2	chr5:1,295,308	chr5:1,295,291	4390	4597	5>3	Flip	AAV2/13	Nault *et al.* 2015
LiC1162	#2102T	TERT	AAV2	chr5:1,295,238	chr5:1,295,235	4288	4597	5>3	Flip	AAV2	
LiC1162	#1919T	GLI1	AAV2	chr12:57,856,202	chr12:57,856,230	4619	4108	3>5	Flip	AAV2	
LiC1162	#1920T	GLI1	AAV2	chr12:57,856,195	chr12:57,856,212	4202	4042	3>5	Flip	AAV2	
LiC1162	#1921T	GLI1	AAV2	chr12:57,854,317	chr12:57,854,407	3940	4679	3>5	Flip	AAV2	
LiC1162	#3765T	GLI1/INHBE	AAV2	chr12:57,850,368	chr12:57,850,370	4071	4679	3>5	Flip	AAV2	
LiC1162	#1920T	H2AFY2	AAV2	chr10:71,864,210	chr10:71,864,240	4579	3629	3>5	Flip	AAV2	
LiC1162	#1344T	IGR chr10	AAV2	chr10:61,519,721	chr10:61,520,465	4342	2447	3>5	ND	AAV2	
LiC1162	#2141T	NKAIN3	AAV2	chr8:63,681,461	chr8:63,681,478	4608	4570	3>5	Flop	AAV2	
ICGC-JP	HX032	CCNE1	AAV2	chr19:30,304,511	chr19:30,304,517	4600	4386	3>5	ND	ND	Fujimoto *et al.*, 2016
ICGC-JP	RK112	KMT2B	AAV2	chr19:36,213,603	chr19:36,213,955	3320	4438	5>3	ND	ND	Fujimoto *et al.*, 2016
ICGC-JP	RK236	IGR chr5	AAV2	chr5:18,813,512	chr5:18,813,398	82	2982	3>5	ND	ND	Fujimoto *et al.*, 2016

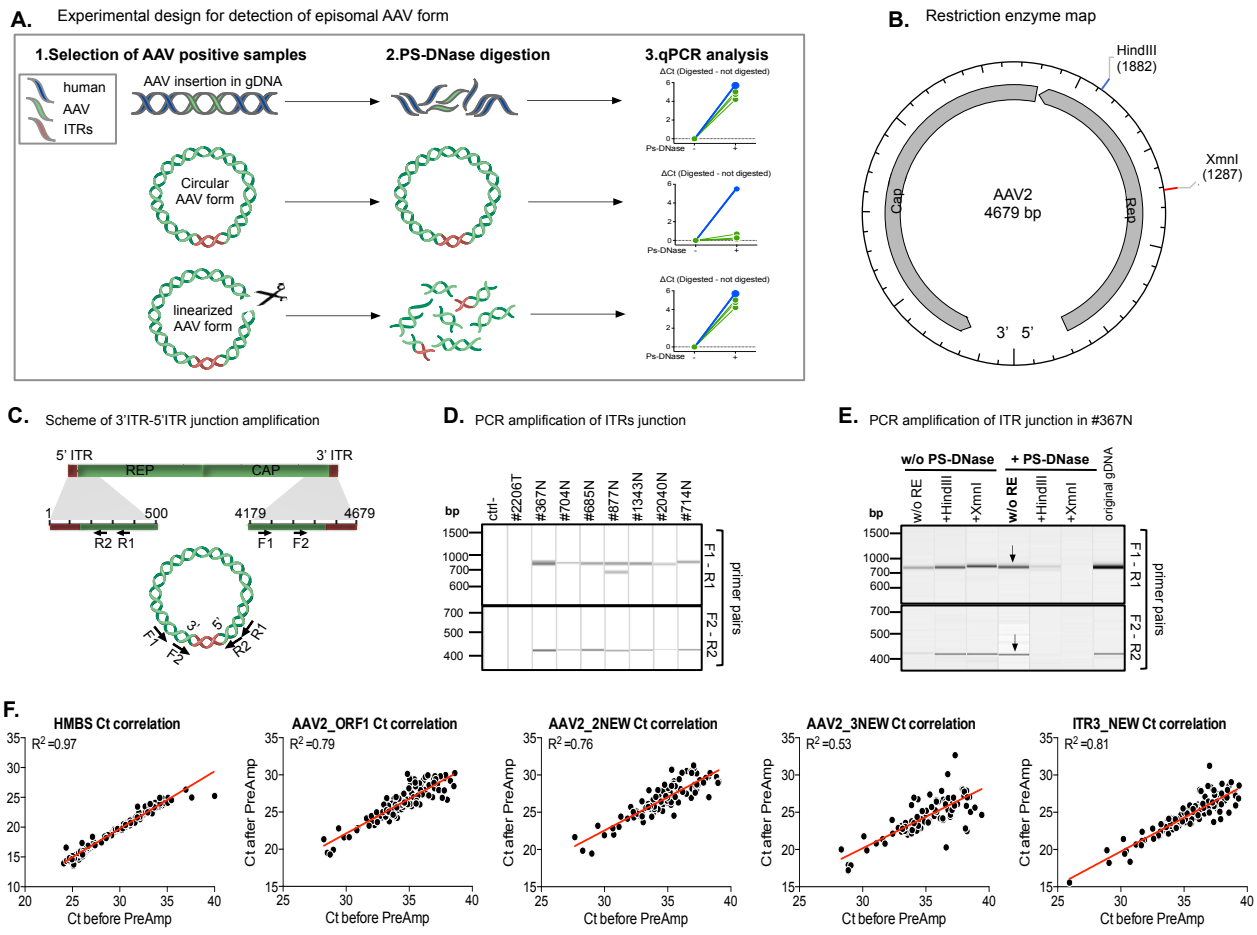
ND: not determined



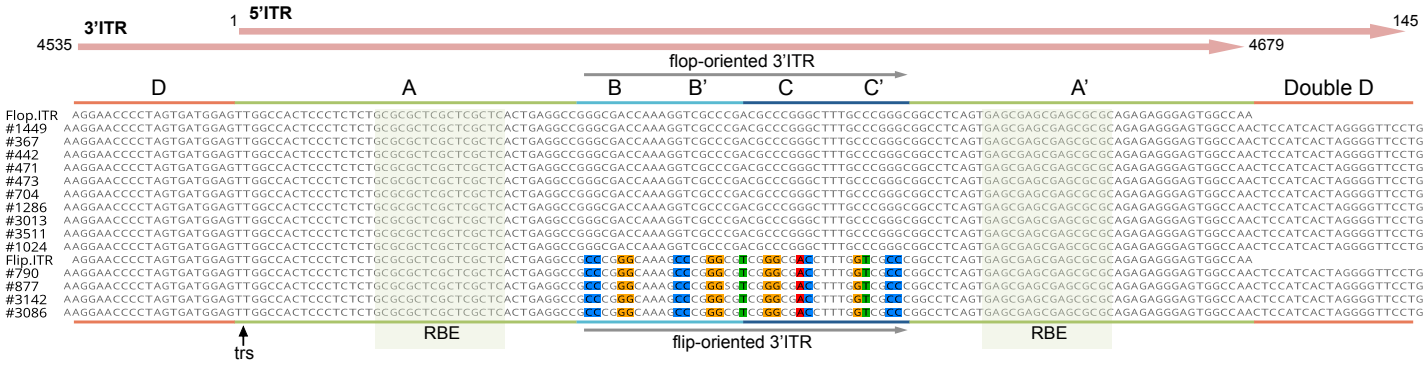
Supplementary figure 1. AAV detection using quantitative PCR and viral capture. A) Schematic representation of the 8 TaqMan probe sets used in qPCR screening mapped on AAV2 reference genome (NC_001401). The 5 genomic regions targeted by the probe sets are defined at the top of the scheme. The color of the probe sets indicates the specificity for the viral genotypes. B) Correlation between the number of AAV copies per cell determined using qPCR and the number of AAV reads in viral capture normalized on human reads.


Goose Parovirus
Avian AAV
AAV5
AAV4
hu.37
hu.40
hu.41
hu.42
hu.67
hu.66
hu.6
hu.17
hu.39
AAV8
AAV7
Clade E
(AAV8)
Clade D (AAV7)
hu.14/AAV9
Clade F (AAV9)
hu.31
hu.32
AAV1
hu.48
AAV6
hu.43
hu.44
hu.46
Clade A
(AAV1/6)
AAV3
AAV3B
#1024
#163
#367
#976
#1286
#2107
#790
hu.15
hu.16
#714
#1158
hu.18
#668
#704
#685
hu.7
#3086
#1570
#1020
hu.1
hu.4
hu.3
hu.2
hu.61
hu.60
hu.25
#442
#2320
#1704
#2112
#1350
#3142
hu.9
hu.10
hu.11
#M258
hu.S17
hu.53
hu.54
hu.55
#508
#767
#217
#1591
Clade C
(AAV2-AAV3 Hybrid)
#667
#2806
#129
#1602
#2731
#985
#1010
#2208
#473
#1273
#1017
#777
#1343
#3013
#3511
hu.19
hu.20
hu.23
hu.22
hu.27
hu.21
hu.24
hu.LG15
hu.T17
hu.T41
hu.T71
hu.T88
hu.T32
hu.T40
hu.63
hu.64
#1260
#1055
#1919
#2087
#1449
#2497
#1534
#2102
hu.56
hu.57
#371
#2141
AAV2
hu.34
hu.35
hu.45
hu.47
hu.51
hu.52
hu.58
#2040
hu.49
#387
hu.28
hu.29
hu.T70
#471
#3765
#877
#2557
#2128
#2206
hu.13
Clade B
(AAV2)
0.05

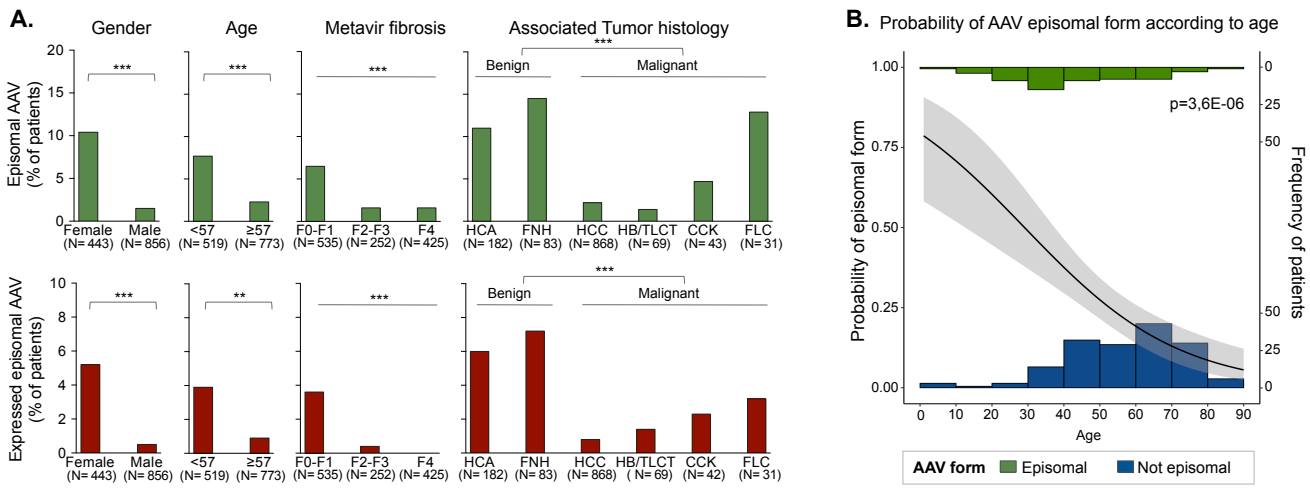
Supplementary figure 2. Phylogenetic tree on VP1 coding protein region. This was constructed with (1) in black 73 human AAVs, 1 avian AAV and a Goose Parvovirus, described in Chen et al.[42], (2) all AAV sequences described therein that to the AAV2 group of sequences in red or (3) to the AAV2/13 group of sequences in blue. A neighbor-joining method on the basis of the Jules-Cantor model was used to derive phylogenetic distances based on 735 aa of VP1 sequence. Goose parvovirus was used as the outgroup. Previously described AAV clade nomenclature described in Gao et al.[14] was adopted and organized by vertical brackets. The scale for genetic distance is indicated in the bottom left corner.



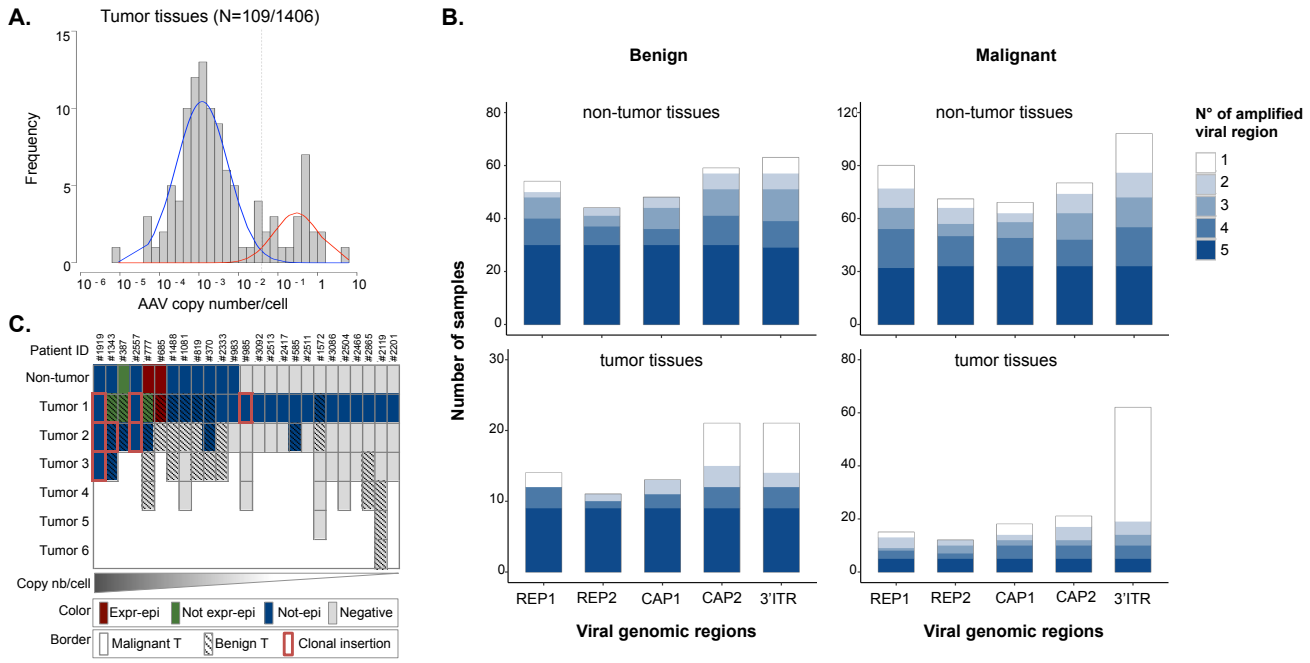
Supplementary figure 3. Viral episomal form investigation. A) Experimental design of the DNase/TaqMan assay for episomal form detection. The ΔCt (Digested – Not digested DNA) is used to define the presence or absence of episomal AAV form. An increased ΔCt indicates the digestion of AAV DNA due to the presence of an integrated not-episomal form (top), whereas a low ΔCt results from the protection of the viral DNA due to the presence of an episomal form (middle). An increased ΔCt is also observed as consequence of linearization of the episomal AAV form prior to DNase digestion (bottom). A probe set (HMBS) targeting the human genome (in blue) is used as control of DNase digestion efficiency. B) AAV map with positions of restriction enzymes (HindIII and XmnI) used to linearize viral episomal form. C) Schematic representation of the primers used to amplify the 3'ITR-5'ITR junction. Two different couples of primers were used: F1-R1 and F2-R2. D) Junction's amplification in 7 patients with confirmed viral episomal form (from line 3 to 9) using 2 different couples of primers. #2206 is used as negative control. E) PCR amplification of the 3'ITR-5'ITR junction before and after DNase digestion in patient #367. When the PCR is preceded by the linearization of episomal AAV, no amplification is detected after DNase digestion. The arrows indicate the 3'ITR-5'ITR junction in digested DNA without linearization of the episomal form. High-resolution capillary electrophoresis was performed using Qiaxcel with individual lanes of the same migration figured on the photo in D and E. F) Correlation between Ct value of HMBS and AAV probe sets before and after DNA pre-amplification.



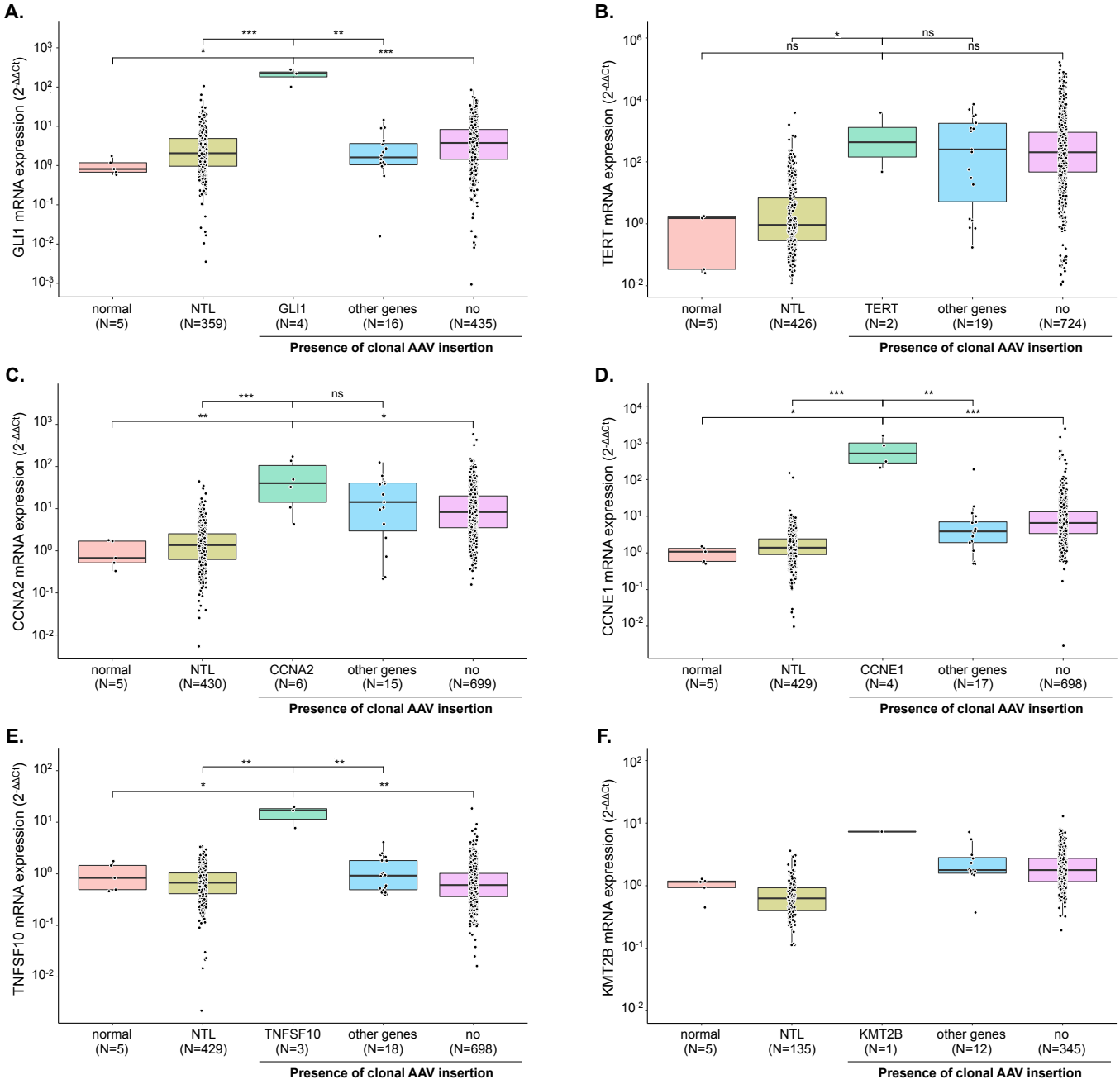
Supplementary figure 4. Viral 3'ITR-5'ITR episomal junction determined by *in silico* analysis in 14 patients. ITR is composed of seven regions: A, A', B, B', C, C' and D. The B-B‘ and C-C' regions exist in two palindromic configurations, "flip" (10 patients) and "flop" (4 patients) oriented DNA (grey arrows). Rep-binding element (RBE) motif is boxed; the arrow indicates the positions of the terminal resolution site (trs).



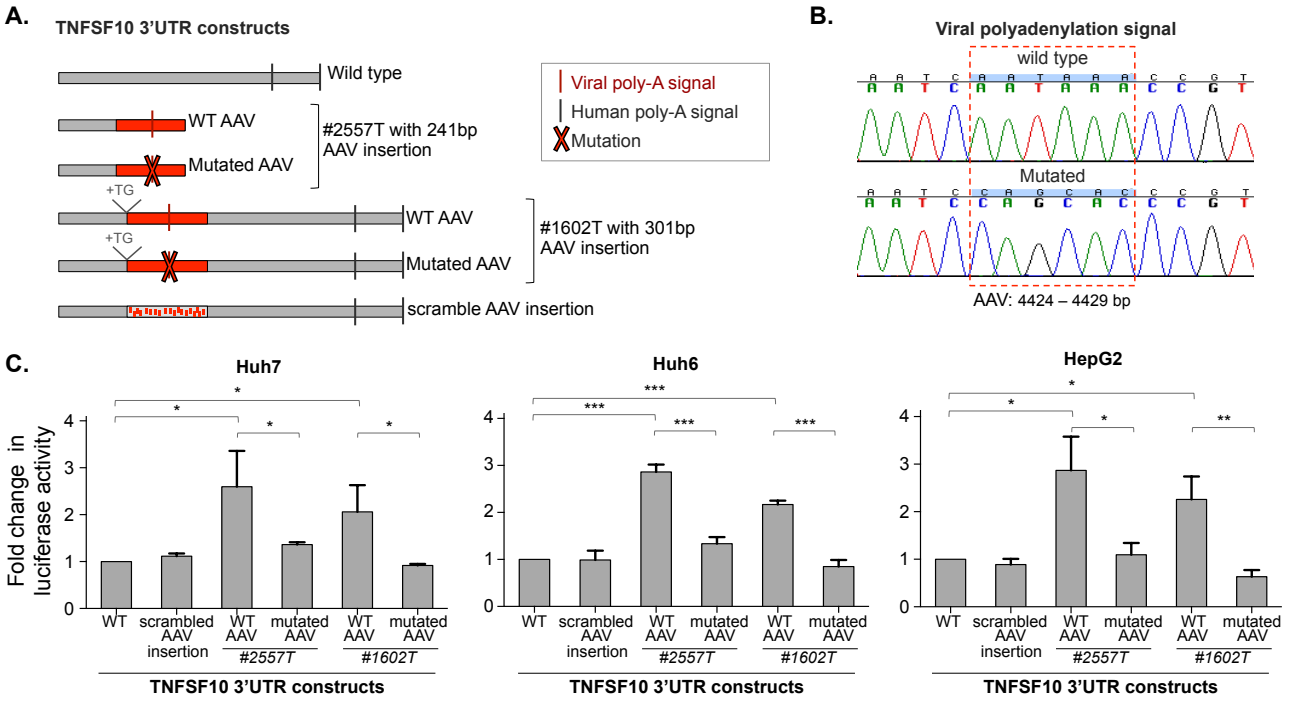
Supplementary figure 5. Correlation between AAV form and features of the patients. A) Frequency of episomal AAV (top) and episomal and expressed AAV form (bottom) according to gender, age, Metavir fibrosis score and associated tumor histology. The histological tumor types are grouped according to their malignancy. Frequency of AAV positive patients is displayed (χ^2 test with Monte Carlo simulation and χ^2 test for trend in proportions for Metavir score). ***P < 0.001; **P < 0.01; *P < 0.05. B) Frequency of episomal and not-episomal AAV form according to age of the patients. Logistic regression model was used to predict the probability to have an episomal form according to the age, where 1 indicates the highest probability and 0 the lowest. Regression curve of the probability with standard deviation is represented.



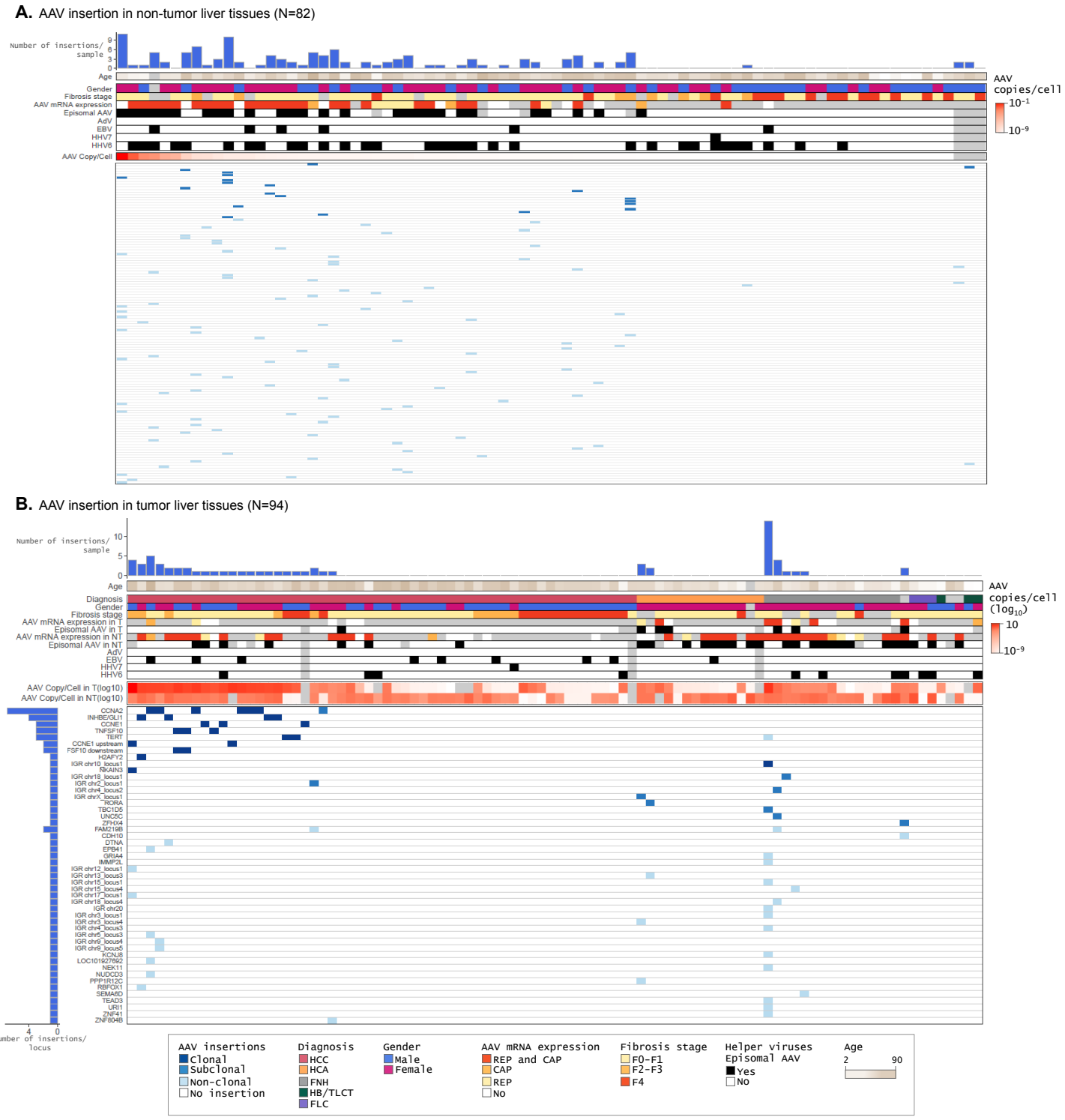
Supplementary figure 6. AAV presence in tumor tissues and comparison with non-tumor counterparts. A) Viral copy number/cell distribution in 109 tumor samples. The density line defines the low and high positivity groups in blue and red respectively. B) Number of amplified viral regions according to the 5 different genomic regions in non-tumor (top) and tumor tissues (bottom) of patients with benign (left) and malignant (right) tumors. C) Inter-tumor heterogeneity in patients with multiple nodules analyzed for AAV presence. Each column represents one patient. The 26 patients are ordered according to the copy number/cell in the non-tumor tissues. In case of negative non-tumor tissue, the order is determined by the copy number/cell in tumor. The color indicates the molecular form of the virus: episomal and expressed AAV (Expr-epi; red), not-expressed episomal AAV (Not expr-epi; green), not-episomal (Not-epi; blue) and negative (grey). The border and the pattern indicates the tumor type (malignant or benign) and the presence of clonal insertions.



Supplementary figure 7. Impact of clonal AAV integration on the expression of the target oncogenes: GLI1 (A), TERT (B), CCNA2 (C), CCNE1 (D), TNFSF10 (E) and KMT2B (F). The level of expression was assayed using qPCR in HCC with AAV insertions, within the tested gene and other target genes, in comparison to HCC without AAV insertion and non-tumor liver (NTL) tissues. Gene expression is presented relative to the expression in normal liver tissue on the *y* axis (log_{10}). The black line in each boxplot corresponds to the mean values. A significant difference in expression was defined by Wilcoxon rank-sum test: ***P < 0.001; **P < 0.01; *P < 0.05.



Supplementary figure 8. Viral polyadenylation signal usage in TNFSF10. A) TNFSF10 3'UTR constructs into a pmirGLO vector controlling the expression of Firefly luciferase gene under PGK promoter. Site of insertions with #2557T and #1602T AAV sequences (in red), with and without mutated viral poly-A signal, or with the scrambled viral sequence are represented. B) Sanger validation of the mutations in viral poly-A signal. Nucleotides position refers to AAV2 (NC_001401). C) Impact of AAV integration in the 3'UTR of TNFSF10 in presence of wild type and mutated viral poly-A signal was evaluated using luciferase assays in Huh7, Huh6 and HepG2 liver cell lines. The 3′UTR of TNFSF10 with either of the two AAV insertions identified in #2557T and #1602T or a scrambled sequence cloned into the pmirGLO vector was compared to vector encoding the wild-type 3′UTR (WT). Error bars, s.d. of triplicate experiments corresponding to three independent transfections for each plasmid in each cell line. t tests were performed; ***P < 0.001; **P < 0.01; *P < 0.05.



Supplementary figure 9. Description of AAV insertions in non-tumor (A) **and tumor liver tissues** (B). AAV insertions are represented according to the type (clonal, subclonal and non-clonal). Number of samples refers to multi-samples. Annotations related to patients features, AAV presence and helper virus infections are reported. The histograms indicate the number of AAV insertion per sample (on the top) and the number of AAV insertion per genomic locus (on the left). The latter histogram is not shown for in the non-tumor tissues graph because each locus presents only unique insertion event.

Supplementary Materials and Methods

Computational analysis of human-AAV sequences

Paired-end read sets, for which samples we have confirmed the presence of the virus, were further consider to assemble the whole viral genome using an overlap layout consensus (OLC) assembling strategy. Both circular and linear contigs were generated with the Genious Pro 11.1.5 (http://www.geneious.com, Biomatters Ltd, Auckland, New Zealand) using the de novo assembler with high sensitivity threshold, considering a minimum overlap of 25 bases and 80% of identity, a maximum gap size of 5 bases, less than 20% gaps per read, and maximum ambiguity of 16 bases. Consensus sequences were generated, based on the major base distribution, and confirmed on reads alignment. Sequences have been deposited in the Genbank database.

From all sequence contigs, VP1 open reading frames were used to segregate our human isolates into individual molecular forms (or clades) as described in Chen et al.[1] The coding protein region was derived from the nucleotide sequences. A neighbor-joining method[2] on the basis of the Jules-Cantor model was used to derive phylogenetic distances based on 735 amino acids of VP1 sequence from our 58 human isolates and 73 human AAV plus 1 avian AAV. Goose parvovirus was used as the outgroup.

The whole circular and linear AAV DNA structures (4679 bp) from respectively 14 and 57 isolates were compared using multi-alignment with the ClustalW algorithm[3] using default parameter. A Neighbor-Joining method[2] on the basis of the Tamura-Nei method[4] was used to derive phylogenetic distances based on the nucleotides sequences. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in Genious Pro 11.1.5 software using Treebuilder. Sequence variations with wild-type AAV2 sequences and among human AAV isolates were identified considering a minimum variant frequency of 20%. The approximate *p*-value represents the probability of a sequencing error, which is calculated using the binomial distribution, resulting in observing bases with at least the given sum of qualities.

The assessment of the circular structures of AAV on episomal junction, allowed us to define the overlapping region of 3' and 5' inverted terminal repeats (ITRs) and determine the existence of a flip or flop sequence in each human isolate.

Experimental procedure to search for episomal forms of AAV

240ng of genomic DNA were digested by incubation with 10U of Plasmid-Safe DNase (Epicentre) in a final volume of 16µL for 16h at 37°C, followed by heat inactivation of the nuclease. 20ng of DNase-digested and not-digested control DNA were pre-amplified by TaqMan PreAmp Master Mix (Applied Biosystem) using 14 preamplification PCR cycles. Real time quantitative PCR was performed on 10ng of DNA before preamplification and 5µL of preamplification product (diluted 1:10 in TE buffer) using custom made AAV probes (Supplementary Table 2) and AB 7900HT Fast Real-Time PCR System (Applied Biosystem). Expression data (Ct values) were acquired using SDS Software (v2.3). The efficiency of the nuclease digestion was determined using HMBS probe as control of the genomic DNA digestion. The difference between AAV Ct without and with PS-DNase digestion allowed to determine the presence of viral episomal form, or the digestion of the linear viral DNA.

In order to validate the results, an additional step of linearization of the viral genome was added before PS-DNase digestion. 600ng of gDNA were first digested in a double enzymatic digestion (1h using 12U of enzyme and 1h with additional 10U at 37°C) with restriction enzyme (XmnI or HindIII from New England Biolabs) that cuts only one site within AAV sequence. A specific probe set (Supplementary Table 2) that overlapped the cutting site was used to check the efficiency of the digestion. Digested DNA was tested for the presence of episomal form with the DNAse/TaqMan based assay previously described (Supplementary Figure 3).

A first set of 10 samples including 8 tissues with at least 4 amplified genomic regions and 2 controls positive only for 2 regions was tested and used to set up the protocol. In addition, a single cutter restriction enzyme was used to linearize the viral DNA prior to DNase digestion. The combination of the DNase digestion with or without the previous linearization of the viral DNA allowed to determine the presence of an episomal form according to the final TaqMan results on digested DNA. If the virus was in the episomal form, the DNase was not able to digest the viral genome and the sample was still positive for the virus in the TaqMan assay. Conversely, if the virus was in linear form, the viral DNA was digested giving a negative TaqMan results. We found the presence of episomal form in 7/8 non-tumor tissues and none of the controls was positive. In all the cases we observed a digestion of the linearized AAV DNA demonstrated by the increased Ct value.

The sample #367N was tested for 3'ITR-5'ITR junction amplification after DNase treatment with or without previous linearization of the viral genome. Interestingly, the ITR junction was protected from the DNase treatment showing that the 3'ITR-5'ITR junction belonged to the circular AAV episomal form (Supplementary Figure 3E).

Sure of the reliability of the technique, we enlarged the investigation of viral episome in all non-tumor samples with 3 or more amplified genomic regions, which represented the 58% of AAV positive non-tumor samples. A step of pre-amplification of viral DNA after DNase digestion was added in order to assess the episomal status even in case of low number of viral copies. Ct values of each probe set before and after pre-amplification were strongly correlated (Supplementary Figure 3F), therefore qPCR results after pre-amplification were used to determine the episomal status.

References Supplementary Materials and Methods

1. Chen CL, Jensen RL, Schnepp BC, et al. Molecular characterization of adeno-associated viruses infecting children. J Virol 2005;**79**:14781-92.
2. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;**4**:406-25.
3. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;**23**:2947-8.
4. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;**10**:512-26.