

Aberrant integration of Hepatitis B virus DNA promotes major restructuring of human hepatocellular carcinoma genome architecture

Eva G. Álvarez^{1,2}, Jonas Demeulemeester^{3,4†}, Clemency Jolly^{3, ‡}, Daniel García-Souto^{1, ‡}, Paula Otero^{1,2, ‡}, Ana Pequeño¹, Jorge Zamora¹, Marta Tojo⁵, Javier Temes¹, Adrian Baez-Ortega⁶, Bernardo Rodríguez-Martín^{1,2}, Yilong Li⁷, Ana Oitaben¹, Alicia L. Bruzos¹, Mónica Martínez-Fernández¹, Kerstin Haase³, Martin Santamarina^{1,2}, Sonia Zumalave^{1,2}, Rosanna Abal¹, Jorge Rodríguez-Castro¹, Aitor Rodriguez-Casanova^{8,9}, Angel Diaz-Lagares^{8,10}, Keiran Raine⁷, Adam P. Butler⁷, Atsushi Ono¹¹, Hiroshi Aikata¹¹, Kazuaki Chayama¹¹, Masaki Ueno¹², Shinya Hayami¹², Hiroki Yamaue¹², Miguel G. Blanco¹, Xavier Forns¹³, Carmen Rivas¹, Sofía Pérez-del-Pulgar¹³, Raúl Torres-Ruiz^{14,15}, Sandra Rodríguez Perales¹⁴, Urtzi Garaigorta^{16,†}, Hidewaki Nakagawa^{17,†}, Peter J. Campbell^{7,18,†}, Peter Van Loo^{3,†} & Jose M. C. Tubio^{1,2,7,*}

¹Molecular Medicine and Chronic Diseases Research Centre (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain.

²Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain.

³The Francis Crick Institute, London NW1 1AT, UK.

⁴Department of Human Genetics, University of Leuven, Leuven B-3000, Belgium.

⁵The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo 36310, Spain.

⁶Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK.

21 ⁷Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

22 ⁸Cancer Epigenomics, Translational Medical Oncology Group (Oncomet), Health Research
23 Institute of Santiago (IDIS), University Clinical Hospital of Santiago (CHUS/SERGAS), 15760
24 Santiago de Compostela, Spain.

25 ⁹Roche-Chus Joint Unit, Translational Medical Oncology Group (Oncomet), Health Research
26 Institute of Santiago (IDIS), 15760 Santiago de Compostela, Spain.

27 ¹⁰Centro de Investigación Biomédica en Red Cáncer (CIBERONC), 28029 Madrid, Spain

28 ¹¹Department of Gastroenterology and Metabolism, Graduate School of Biomedical and Health
29 Sciences, Hiroshima University, Hiroshima, Japan.

30 ¹²Department of Surgery, Wakayama Medical University, Wakayama, Japan.

31 ¹³Liver Unit, Hospital Clínic, University of Barcelona, IDIBAPS, CIBERehd, Barcelona, Spain.

32 ¹⁴Molecular Cytogenetics and Genome Engineering Group, Centro Nacional de Investigaciones
33 Oncológicas (CNIO), Madrid, Spain.

34 ¹⁵Division of Hematopoietic Innovative Therapies, Centro de Investigaciones Energéticas,
35 Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain.

36 ¹⁶Department of Molecular and Cellular Biology, Centro Nacional de Biotecnología – Consejo
37 Superior de Investigaciones Científicas (CNB – CSIC), Madrid 28049, Spain.

38 ¹⁷RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan.

39 ¹⁸Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK.

40 ‡Equal contribution

41 †Equal contribution

42

43 *Correspondence to:

44 Dr Jose M. C. Tubio,

45 Genomes and Disease,

46 Molecular Medicine and Chronic Diseases Research Centre (CIMUS),

47 University of Santiago de Compostela,

48 Santiago de Compostela 15706,

49 Spain

50 Phone: +34 881 815 382

51 e-mail: jmctubio@gmail.com

Most cancers are characterized by the somatic acquisition of genomic rearrangements during tumour evolution that eventually drive the oncogenesis. There are different mutational mechanisms causing structural variation, some of which are specific to particular cancer types. Here, using multiplatform sequencing technologies, we identify and characterize a remarkable mutational mechanism in human hepatocellular carcinoma caused by Hepatitis B virus, by which DNA molecules from the virus are inserted into the tumour genome causing dramatic changes in its configuration, including non-homologous chromosomal fusions and megabase-size telomeric deletions. This aberrant mutational process, present in at least 8% of all HCC tumours, is active early during liver cancer evolution and can provide the driver rearrangements that a cancer clone requires to survive and grow.

Human hepatocellular carcinoma (HCC) is the most common primary liver malignancy, resulting in over 700,000 deaths globally every year ¹. Previous studies indicate that the disease has a complex genomic landscape, with frequent copy number changes and interchromosomal rearrangements ^{2,3}. Hepatitis B virus (HBV) infection – a condition affecting 240 million people worldwide – is the second most frequent cause of cancer after tobacco, and a major cause of HCC. HBV infection has been associated with chromosomal instability in cancerous and non-cancerous liver genomes, and HBV DNA integration is known to be the cause of chromosomal rearrangements in HCC ⁴⁻¹¹. However, we still ignore the full extent to which HBV DNA integrations impact the structure (i.e., patterns and mechanisms of mutation) and function (i.e., driver events) of HCC genomes ¹², which may have important consequences for the diagnosis, prognosis and treatment of the disease. Here, we harness recent advances in DNA sequencing

technologies using short and long-reads to characterize patterns of structural variation associated with HBV DNA integration in human HCC. Our analyses further illuminate a remarkable mutational mechanism, present in at least 8% of all HCC tumours, by which somatic integration of HBV DNA promotes non-homologous interchromosomal rearrangements coupled with telomeric (i.e., that includes the telomere) deletions in one or two of the chromosomes involved, occasionally representing tumour driver events in HCC.

We run our bioinformatic algorithms (**Online Methods**) to explore the landscape of HBV DNA integrations acquired somatically on Illumina paired-end whole-genome sequencing data from 286 HCC tumours from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project¹³. Their matched-normal samples derived from blood, were also sequenced. This analysis retrieved a total of 148 somatic HBV integration events in 51 tumour samples (**Fig. 1a; Supplementary Table 1**). Forty-two of these events represent canonical viral DNA insertions where the paired-end mapping data shows a classical pattern, characterized by two reciprocal – face-to-face oriented – read clusters delimiting the integration site, and whose mates support the presence of viral DNA (**Fig. 1b**). This result is consistent with an alternative study on the same dataset carried out by others¹⁴. However, in addition to these canonical insertions, our analysis revealed that a majority (72%, 106/148) of events followed an unexpected, non-canonical pattern. Here, paired-end mapping data showed single clusters of reads whose mates identify one extreme of the somatic viral integration only, while the reciprocal cluster supporting the other extreme of the insertion appeared to be missing. For instance, in one HCC tumour, SA501453, paired-end reads show a single cluster supporting one extreme of an HBV insertion event on chromosome 19, with no reciprocal cluster in the proximity of the integration site (**Fig. 1c**). Our algorithms successfully reconstructed the

ends of these 106 non-canonical insertion events, confirming that they match HBV sequences
(**Supplementary Table 1**).

Similar paired-end mapping patterns were previously identified in cancer genomes with high retrotransposition rates¹⁵, where this type of events represented hidden genomic rearrangements mediated by aberrant DNA integrations. This suggested that our findings could represent cryptic somatic rearrangements mediated by HBV DNA insertion. Actually, somatic rearrangements linked to HBV insertion sites have been recently identified using long-read sequencing technologies in human HCC cell-lines⁵ and primary tumours¹⁰. Hence, to illuminate the genuine configuration of the relevant rearrangement involved in the patterns described above, we performed long-read whole-genome sequencing on the affected tumour, SA501453, using Oxford Nanopore Technologies (ONT) to a final coverage of 13.5X (median read size = 12 kb). The long reads revealed a cryptic translocation linking chromosomes 19 and 11, which is bridged by a 640 bp HBV DNA insert (**Fig. 2**). Although our algorithms had initially identified the missing reciprocal cluster on chromosome 11 in the paired-end data, the interchromosomal rearrangement remained undetectable due to size constraints of the Illumina sequencing library, which was too short to span the HBV insertion. Notably, the genomic breakpoints of this translocation remained unnoticed to a set of four different structural variation calling pipelines, which were employed in the identification of genomic rearrangements and in the PCAWG dataset^{13,16}.

Many non-canonical HBV insertions occur in association with megabase-size deletions that remove telomeric regions of a chromosome. For instance, in HCC tumour SA529726, the paired-end sequencing data revealed one single cluster of an HBV insertion on the short arm of

chromosome 3. Here, the insertion boundary is associated with a large copy number loss (**Fig. 3a**), suggesting that the insertion event occurred in conjunction with a telomeric deletion that removed 21 Mb of chromosome 3p. We performed long-read sequencing on this sample, which revealed that the telomeric deletion occurred due to an unbalanced translocation between chromosomes 3 and X bridged by a 3.3 kb HBV insertion that shows a classical fragmented and rearranged form^{5,17} (**Fig. 3a; Supplementary Fig. 1a**). In the same sample, the ONT data showed a second, unrelated HBV insertion (3.5 kb long) that bridges a translocation between chromosomes 4 and 7, associated with a 20 Mb telomeric deletion on the 4q (**Fig. 3a; Supplementary Fig. 1b**). Similarly, in another remarkable HCC tumour, SA501511, up to three different HBV insertions were found associated with large deletions – 20.5, 33.6, and 76.7 Mb long – removing the telomeres on chromosome arms 10p, 4p and 13q, respectively (first circos plot in **Fig. 3b**). This time, the long-read sequencing data revealed three cryptic HBV-mediated translocations between the long arm of chromosome 8 and the relevant deletion breakpoints on chromosomes 4, 10 and 13 (**Fig. 3b; Supplementary Fig. 2a-c**).

We looked in the PCAWG HCC dataset for other HBV insertions demarcating huge telomeric copy number loss events, which could involve the same mutational mechanism, finding 26 additional events in 19 different HCC tumours (**Supplementary Fig. 3**). We analysed three of these samples (SA501424, SA501481 and SA529830) by whole-genome long-read sequencing with ONT, which confirmed cryptic interchromosomal rearrangements linked to telomeric deletion breakpoints in all of them (**Fig. 3b**), demonstrating that this aberrant mutational mechanism mediated by HBV insertions is recurrent in human HCC. Notably, in two of the samples sequenced with ONT (SA501481 and SA529830) the configuration of the rearrangement

found supports a derived chromosomal fusion that generates a dicentric chromosome (i.e., a chromosome with two centromeres; **Fig. 3b**). These chromosomes are known to represent potential source for breakage-fusion-bridge (BFB) repair^{15,18}, unless they become stabilized due to reduced intercentromeric distance or by means of inactivation of one of the two centromeres¹⁹. Here, the absence of copy number profiles and chromosomal rearrangements typically associated with BFB cycles supports the last scenario.

Our results illuminate a scenario where rearrangements mediated by viral DNA integration are important remodelers of human HCC genomes. The analysis of copy number profiles revealed that many HBV-mediated rearrangements occurred in chromosomes with copy number gains, providing opportunities for timing analyses^{20,21}. To pinpoint these rearrangements on a timeline from the fertilized egg to tumour diagnosis, we performed a modification of current timing algorithms to operate with single read-clusters only (**Online Methods**). The method revealed that somatic insertions of HBV DNA in HCC are typically clonal events that have been acquired early in tumour evolution. For example, in one notable HCC, SA269680, that underwent whole-genome duplication, we identified eleven viral insertion events. All but one were catalogued as early events (**Fig. 4a; Supplementary Table 2**), and five of these early events corresponded to single clusters associated with megabase-size copy number losses (**Supplementary Fig. 3**), supporting that these large-scale rearrangements may be important in the initial stages of liver oncogenesis.

To further investigate the clinical relevance of HBV integration in HCC, we employed real time estimation data of whole-genome duplication (WGD) events from PCAWG²⁰ to perform a more precise timing estimation of HBV events along patients' lifetime. The method is based on the

analysis of mutational clock signatures that correlate with patient age at diagnosis ²², which can be used for timing of WGDs and their associated variants ²⁰. This approach allowed real-time timing of 37 HBV insertions (8 canonical and 29 non-canonical) embedded in whole-genome duplications (**Fig. 4b; Supplementary Table 3**), and revealed some of these rearrangements may appear many years before diagnosis. For instance, in HCC SA501645, a cryptic HBV-mediated rearrangement in chromosome 10, coupled with a 7.3 Mb telomeric deletion on 10p, occurred over 21 years before the patient was diagnosed with HCC (**Fig. 4b-c**).

We found instances having integrations of HBV involved in cancer driver rearrangements in which essential tumour suppressor genes are lost. In one remarkable HCC, SA529830, we identified one paired-end single cluster supporting an HBV insertion on the short arm of chromosome 17. The insertion occurred in conjunction with a 14.9 Mb clonal telomeric deletion at the integration site, which removed one copy of tumour suppressor gene *TP53* (**Fig. 5**). Notably, the second copy of *TP53* in this tumour is inactivated by the missense point mutation C242S ^{23,24} (**Supplementary Fig. 4**). The paired-end data showed a similar pattern on the short arm of chromosome 8, where a single cluster supporting an HBV insertion occurred together with a loss of the first 41 Mb of the chromosome. The patterns suggested that an HBV DNA molecule could be bridging an unbalanced translocation between chromosomes 17 and 8 that would generate a dicentric chromosome (see circos plot in **Fig. 3b**). We carried out whole-genome long-read sequencing, which confirmed the expected configuration of this relevant rearrangement (long-read plot in **Fig. 5**). In addition, we performed in-situ hybridization to identify the loss of *TP53* and the chromosomal fusion between chromosomes 17 and 8, which further validated these concomitant events (**Supplementary Fig. 5**).

190

191 Similarly, in one additional HCC tumour, SA501481, we identified an HBV insertion into
 192 chromosome 1 associated with the deletion of one copy of tumour suppressor gene *ARID1A* (**Fig.**
 193 **6a**). Here, paired-end data shows a single cluster of reads, whose mates support the HBV insertion,
 194 demarcating a copy number loss of the first 57.2 Mb of 1p including *ARID1A*. Again, in this case,
 195 we initially lacked the DNA region on the other side of the rearrangement mediated by the virus,
 196 due to Illumina library insert size constraints. The paired-end data showed an analogous pattern in
 197 chromosome 9, with an independent cluster supporting an HBV insertion that occurred together
 198 with a telomeric deletion of the first 41 Mb of 9p at the integration site. This scenario suggested a
 199 cryptic unbalanced translocation between 1p and 9p, generating a dicentric chromosome (see
 200 circos in **Fig. 3b**), which was confirmed by long-read sequencing with ONT (long-read plot in **Fig.**
 201 **6a**).

202

203 *ARID1A* is a relevant cancer gene harboring monoallelic loss-of-function mutations in 10-15% of
 204 human HCC samples ²⁵. Notably, in a different HCC, SA501424, we found a similar scenario to
 205 the one described above. This time, an HBV insertion demarcates a deletion of the first 31.5 Mb
 206 of chromosome 1p, which again involved loss of one copy of *ARID1A* (**Fig. 6b**). Hence, we
 207 performed long-read sequencing with ONT, which revealed a cryptic interchromosomal
 208 rearrangement between chromosomes 1p and 11q bridged by HBV (see circos in **Fig. 3b**). The
 209 recurrence of these patterns, involving deletion of tumour suppressor genes *TP53* and *ARID1A* in
 210 three different samples, demonstrates a mutational mechanism mediated by aberrant integration of
 211 HBV DNA, which likely contributes to the development of human HCC. The deletion and the
 212 chromosomal fusion were also validated by in-situ hybridization (**Supplementary Fig. 5**).

213

214 Most cancers are characterized by somatic acquisition of genomic rearrangements during tumour
 215 evolution that, eventually, drive the oncogenic process ²⁶. These structural aberrations are caused
 216 by different mutational mechanisms that generate particular patterns or signatures in the DNA ²⁷.
 217 Identification of these mechanisms and their associated patterns is necessary to understand the
 218 dynamic processes shaping the cancer genome. Here we described the patterns of a recurrent, quite
 219 remarkable mutational mechanism occurring in the early stages of human HCC development
 220 whereby HBV DNA integration mediates interchromosomal rearrangements contributing to
 221 megabase-size telomeric deletions, which may lead to loss of tumour suppressor genes. Our results
 222 demonstrate that the consequences of this mutational mechanism are dramatic for the architecture
 223 of HCC genomes and, on occasion, the resulting structural configuration can drive the oncogenic
 224 process.

225

226 MAIN FIGURES

Figure 1

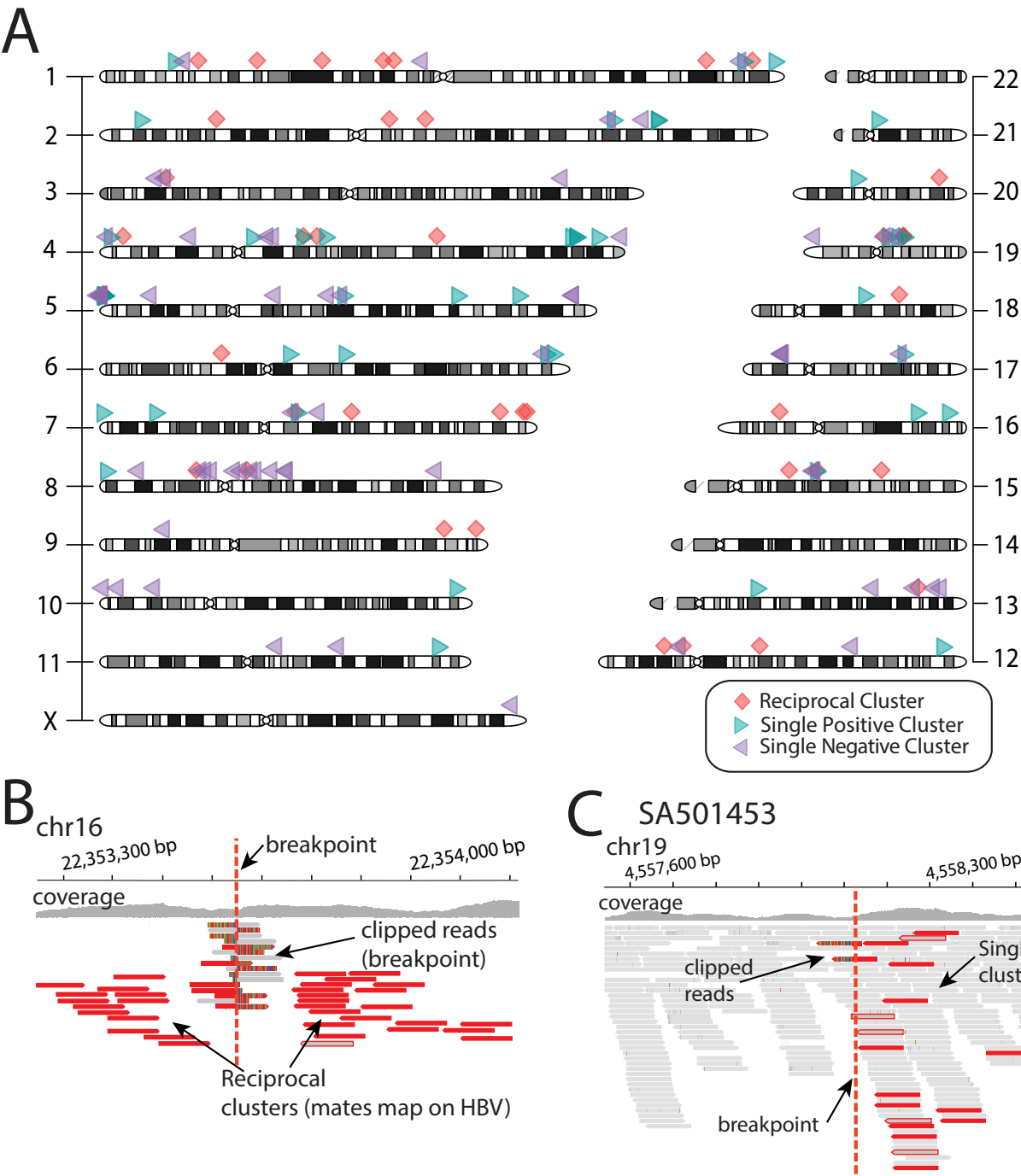


Fig. 1. The landscape of HBV insertions in 286 HCCs from the PCAWG dataset. (a)

Canonical (reciprocal) insertions are represented as red diamonds, and non-canonical insertions (single-clusters) as purple and green triangles for positive and negative clusters, respectively. In total, 148 integration events are shown of which 72% represent non-canonical events. **(b)** Classical pattern of canonical HBV insertions identified with Illumina paired-end mapping data are characterized by two reciprocal clusters of discordant reads, and clipped reads, in face-to-face orientation, demarcating the boundaries of the genomic integration. The mates of these reads map onto HBV consensus sequences. Clipped reads span the insertion site allowing base-pair resolution of the insertion breakpoints. **(c)** Most HBV insertion events in HCC tumours show a non-canonical pattern in which a single cluster of paired-end reads (short-reads in red) demarcates one of the two boundaries of the insertion only, while the second cluster is missing.

Figure 2
SA501453

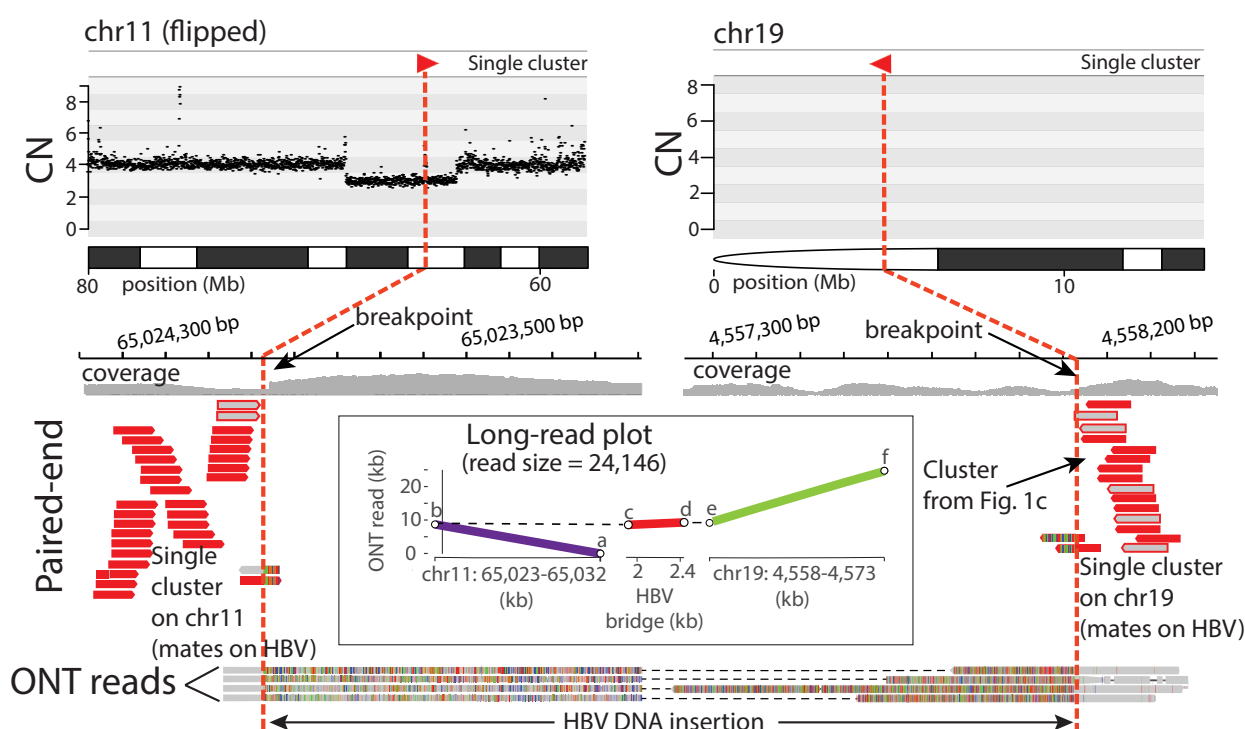


Fig. 2. Long-read sequencing reveals cryptic HBV-mediated translocations in human HCC.

In HCC SA501453, a hidden interchromosomal rearrangement between chromosomes 11 and 19 is identified using Oxford Nanopore Technologies (ONT). The copy number plot (CN) at the top shows the copy number profiles of the chromosomes involved in the rearrangement (note that the CN plot on chromosome 11 is flipped for illustrative purposes). The Illumina paired-end sequencing data (short-reads in red) shows two single clusters of discordant read pairs, one on 11q and a second on 19p, pointing to HBV insertion events that cannot be bridged due to Illumina library size constraints. The bottom shows four long-reads obtained with ONT that reveals the real configuration of the hidden rearrangement, consisting of a 640 bp HBV DNA insertion bridging a translocation between 11q and 19p. The long-read plot represents the alignment of one ONT read

251 – 24 kb long – to chromosomes 11 and 19 of the human reference genome and to an HBV
252 consensus sequence.

253

Figure 3

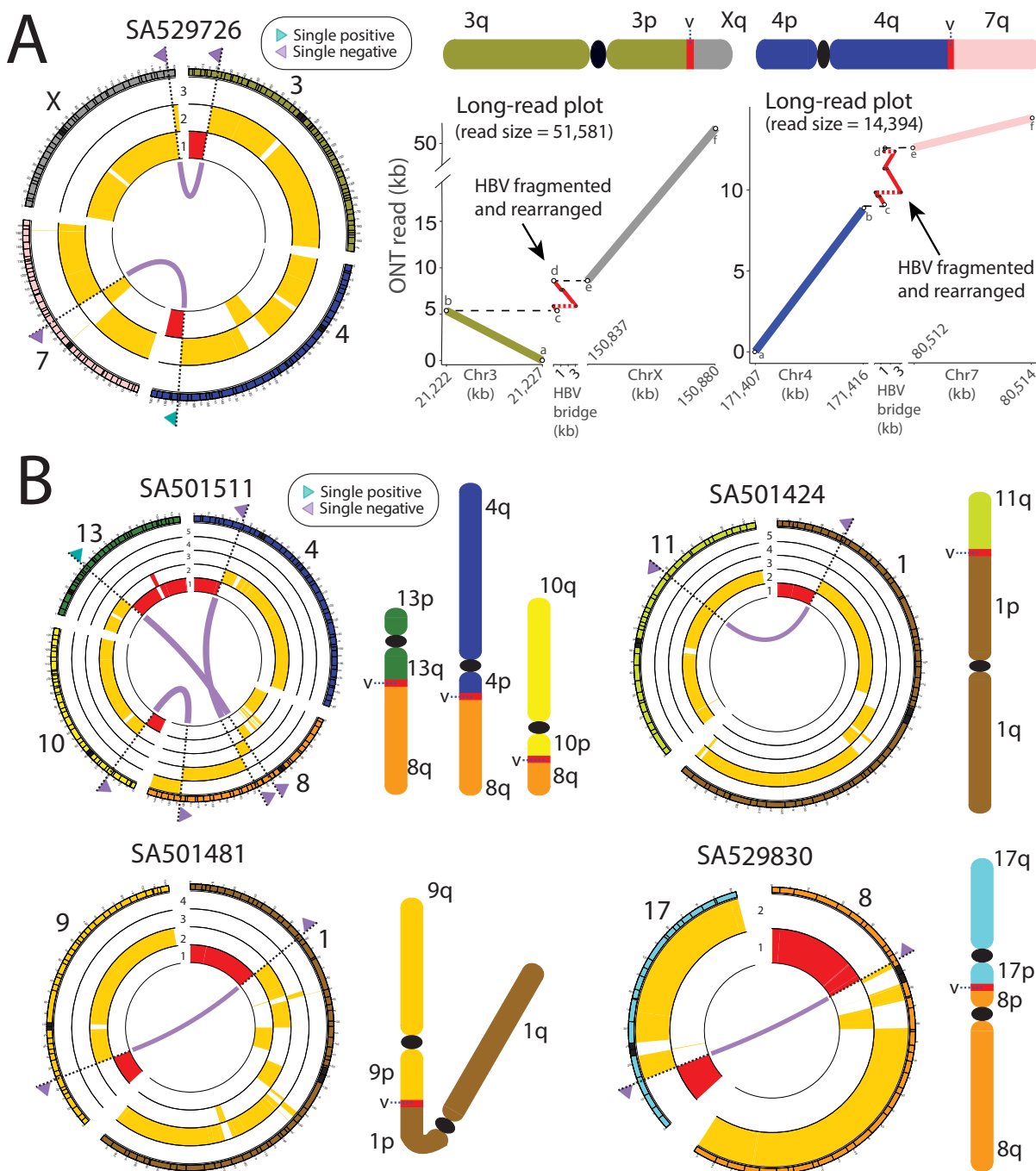


Fig. 3. HBV DNA integration mediates interchromosomal genomic rearrangements that lead to megabase-size telomeric deletions in HCC. (a) In tumour SA529726, two unrelated HBV-mediated interchromosomal rearrangements between chromosomes 3 and X, and between

chromosomes 4 and 7, promote 21.2 Mb and 19.8 Mb telomeric deletions on the 3p and the 4q, respectively. The circos plot (left) represents the translocations (purple lines) revealed by ONT data. Single clusters identified with paired-end mapping data are denoted as triangles (green for positive orientation, purple for negative) on the chromosome ideograms. The copy number profiles are shown in yellow below the chromosome ideograms, with relevant telomeric deletions highlighted in red. The long-read plots (right) represent the alignment of one ONT long-read to chromosomes 3 and X, and chromosomes 4 and 7, of the human reference genome and an HBV consensus sequence, which validates the interchromosomal rearrangements mediated by the virus shown in the circos plot. Here, the analysis of the long-reads supporting the HBV events showed an HBV DNA insertion in a classical fragmented and rearranged form ^{5,17}. The expected configuration of the rearranged chromosome is shown above each long read plot (the ideograms are for illustrative purpose only); ‘v’ denotes the HBV insertion. **(b)** Circos plots and chromosome diagrams of similar HBV-mediated non-homologous translocations promoting megabase-size telomeric deletions in four additional HCC tumours. Again, the expected configuration of the rearranged chromosome is shown next to each circos (the ideograms are for illustrative purpose only). In SA501511, three unrelated HBV-mediated translocations involving different loci on chromosome 8 promote huge deletions involving telomeric regions on chromosomes 13q, 4p and 10p. In SA501424, one HBV insertion bridges a genomic translocation between chromosomes 1 and 11 that generates a terminal deletion at 1p. In SA501481 and SA529830, HBV-mediated translocations generate dicentric chromosomes and promote megabase-size terminal deletions.

Figure 4

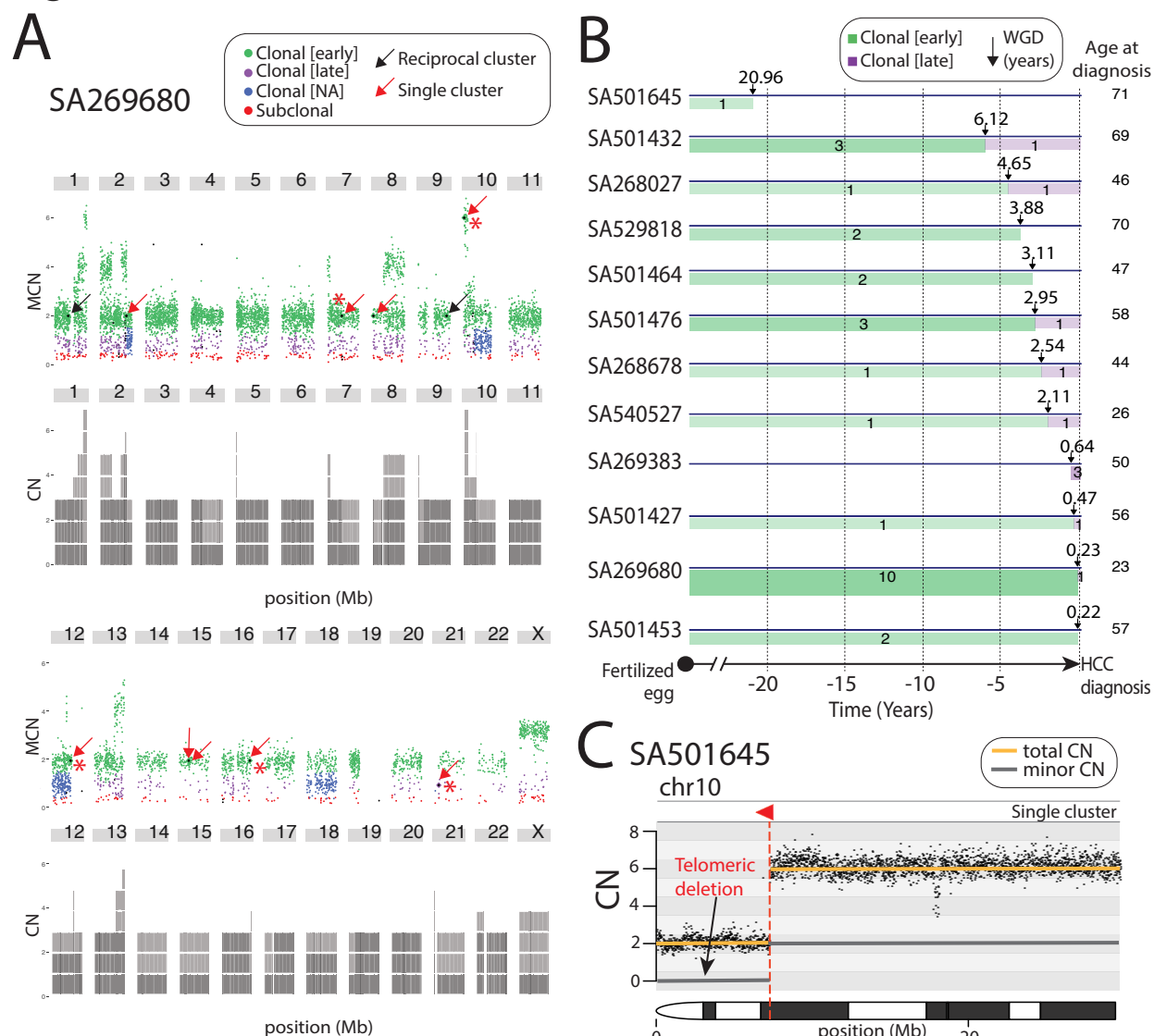


Fig. 4. HBV-mediated rearrangements are early clonal events in HCC evolution. (a) In SA269680, an HCC with a whole-genome duplication, HBV insertions are shown in the context of point mutation burden for that sample. Colored dots above chromosome ideograms represent point mutations with different clonality or timing: early clonal (before the whole-genome duplication; green), late clonal (after the whole-genome duplication; purple), clonal (blue), subclonal (red). We identified nine HBV single clusters (black dots with red arrows), all but one

catalogued as early clonal events. Five of these early HBV insertions (marked with red asterisks) are associated with megabase-size telomeric deletions (see copy number plots in **Supplementary Fig. 3**). The same sample bears two additional early clonal HBV canonical insertions (black dots with black arrows). Grey blocks below chromosome ideograms represent the copy number profile.

(b) Real-time timing estimation of HBV insertions along patients' lifetime in samples with whole-genome duplication events. The X axis shows the time interval when – before (green) and after (purple) – the somatic HBV insertions took place relative to the WGD event; thickness and strength of the green and purple bars correlates with the number of events. Black arrows represent when a WGD event took place, and numbers above arrows show the time – in years – before HCC diagnosis when the WGD event has occurred. Numbers within green and purple timelines represent number of insertion events. Numbers at the end of the timeline represent the age of patient at diagnosis. **(c)** Copy number plot showing a single cluster that supports an HBV insertion event (red triangle) associated with a 7 Mb telomeric deletion on chromosome 10 in SA501645 that, according to **Fig. 4b**, occurred at least 20.96 years before HCC diagnosis. Gold line is the total chromosome CN, and grey line is the minor chromosome CN.

Figure 5

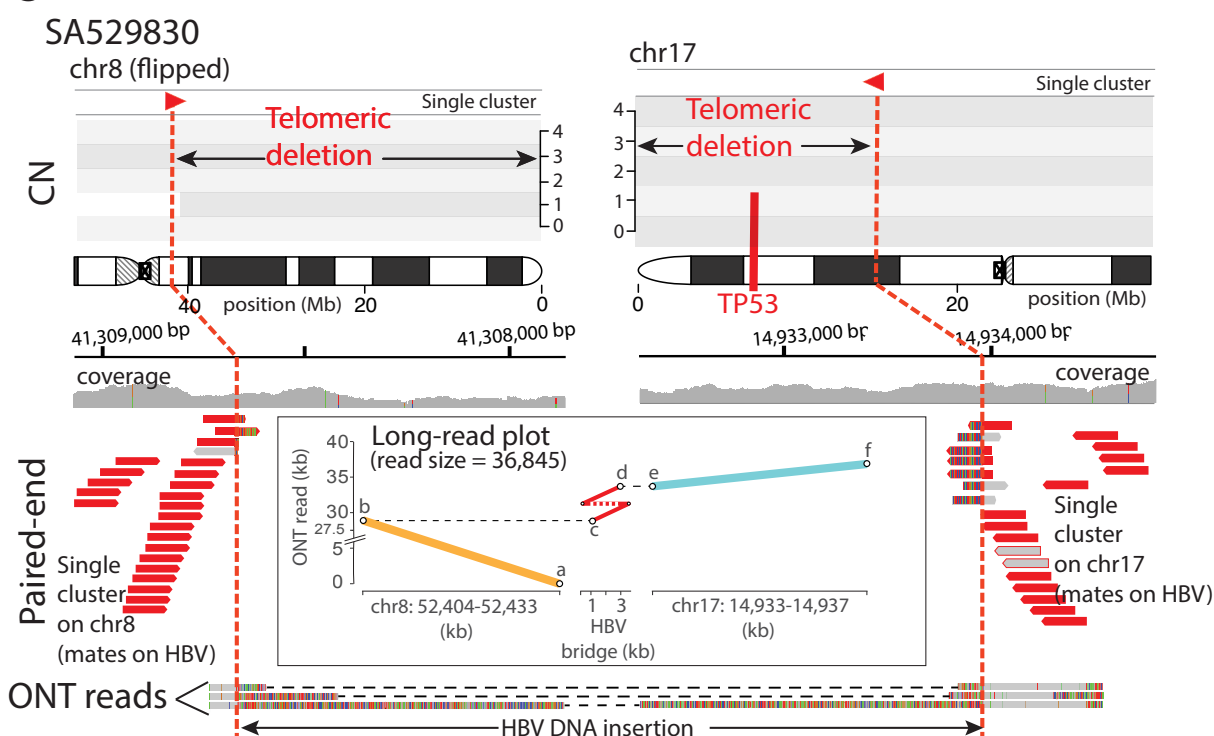


Fig. 5. HBV-mediated translocations may lead to loss of tumour suppressor gene *TP53*. In HCC SA529830, a cryptic interchromosomal rearrangement between chromosomes 17 and 8 is bridged by a 4,829 bp HBV insertion associated with a 14.9 Mb telomeric deletion on chromosome 17 that removes one copy of the tumour suppressor gene *TP53*, and a second 43 Mb telomeric deletion on chromosome 8. Note that the CN plot on chromosome 8 is flipped for illustrative purposes. Single paired-end clusters (short-reads in red) on chromosomes 17 and 8 demarcate the boundaries of both deletions and support the insertion of HBV DNA. One ONT read of 36,845 bp evidences the extent of the rearrangement, whose alignment onto the reference genome – chromosomes 8 and 17 – and to a consensus HBV sequence is shown in the long-read plot. The configuration of the rearrangement predicts the formation of a dicentric chromosome (**Fig. 3b**).

Figure 6

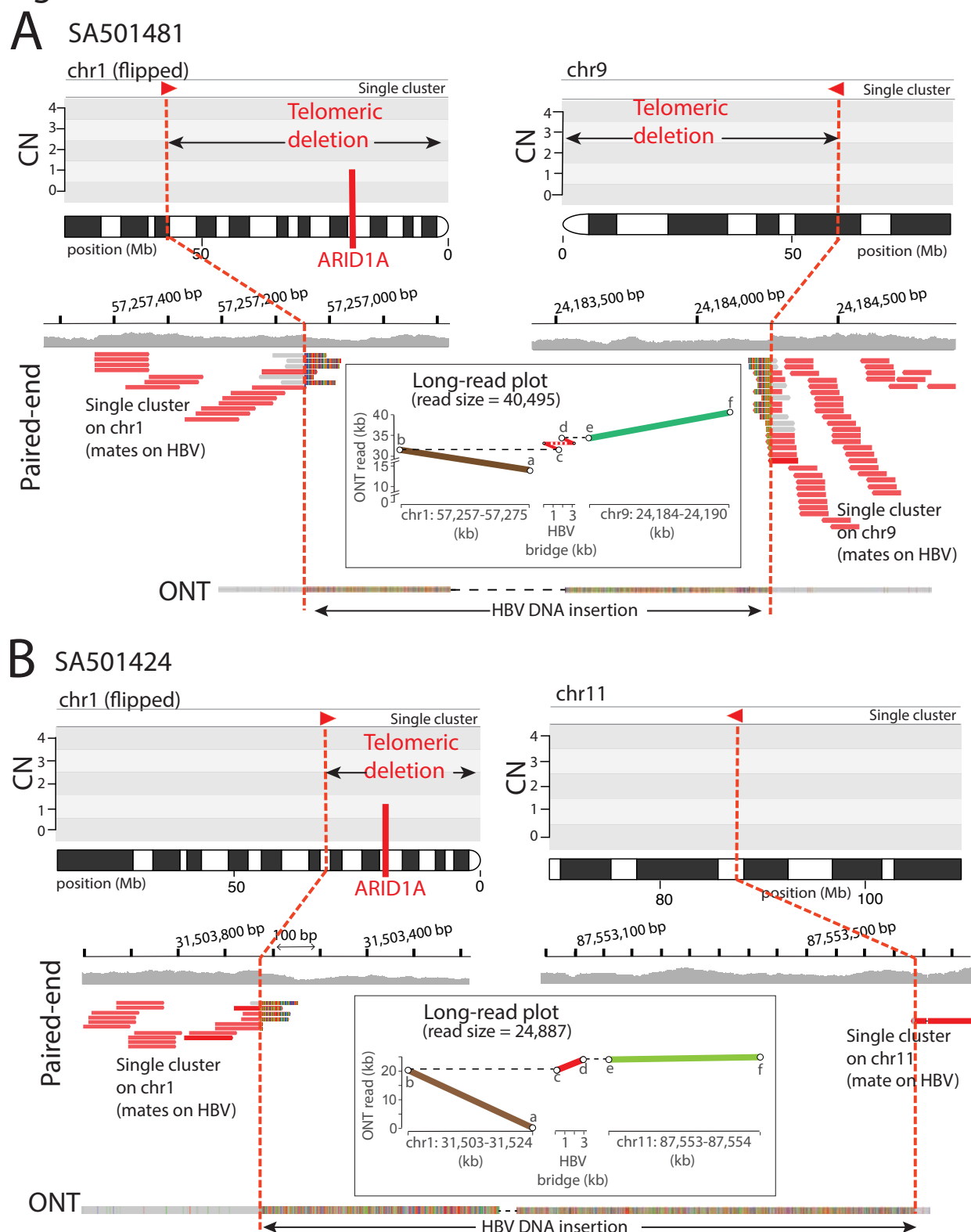


Fig. 6. HBV-mediated translocations lead to recurrent loss of tumour suppressor gene *ARID1A*. (a) In HCC tumour SA501481, the Illumina paired-end data (short-reads in red) shows two clusters, one on chromosome 1 and another on chromosome 9, which point to both extremes of an HBV insertion. The copy number (CN) plot at the top shows the total (gold line) and minor (grey line) chromosomes' copy number profiles. The CN plot reveals two telomeric deletions associated with HBV events, one that removes 57.2 Mb on 1p, including one copy of the *ARID1A* tumour suppressor gene, and a second deletion that removes 21.2 Mb on 9p. Note the CN plot from chromosome 1 is flipped for illustrative purposes. The long-read plot shows a 2,688 bp HBV insertion that bridges an interchromosomal rearrangement between chromosomes 1p and 9p. The configuration of the rearrangement predicts the formation of a dicentric chromosome (**Fig. 3b**).

(b) A similar scenario, in tumour SA501424, where an HBV DNA insertion induces an interchromosomal translocation between chromosomes 1 and 11. The Illumina paired-end data (short-reads in red) shows two single clusters, one on chromosome 1 and another on chromosome 11, which point to both extremes of an HBV insertion. The CN plot at the top reveals a 31.5 Mb telomeric deletion on 1p associated with the HBV insertion event (note that the CN plot from chromosome 1 is flipped for illustrative purposes). Here, the associated telomeric deletion on chromosome 1 removes one copy of tumour suppressor gene *ARID1A*. The long-read alignment plot demonstrates an interchromosomal rearrangement between chromosomes 1 and 11 mediated by an HBV insertion.

ONLINE METHODS

Materials and Methods

Illumina genomes dataset and DNA sources

We analyzed Illumina whole-genome paired-end sequencing reads, 100-150 bp long, from 286 hepatocellular carcinoma (HCC) tumours and their matched-normal samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project dataset ²⁸. The tumour specimens consisted of a fresh frozen sample, while the normal specimens consisted of a blood sample. The average coverage was 44 reads per bp for tumour samples and 33 reads per bp for matched normal samples. BWA-mem algorithm ²⁹ was used to align sequencing reads to human reference genome build GRD37, version hs37d5. Tumour's DNA for additional long-read sequencing and FISH were transferred from the HCC tumours collection at RIKEN (Japan) within the framework of the International Cancer Genome Consortium (ICGC).

Detection of viral insertions using v-TraFiC

v-TraFiC represents a modified version of former algorithm TraFiC ³⁰, for the identification of somatic insertion events of viral DNA using paired-end sequencing data in three main steps: (i) selection of candidate reads; (ii) reads clustering; and (iii) identification of viral DNA events.

1) Candidate reads selection

v-TraFiC identifies reads from BWA-mem mapping that are likely to provide information pertaining to viral DNA site inclusion. Two different read-pair types are considered for the identification of viral insertions, named SINGLE_END (i.e., one end of the pair – called anchor is mapped onto the reference genome while the other is unmapped), and ABERRANT (i.e., both reads of the pair are improperly mapped to a chromosome, where the read with the highest MAPQ

is considered the anchor). In both cases, the anchor's MAPQ must be higher than zero, its mapping pattern must not be 'soft clipping–alignment match–soft clipping' (i.e., CIGAR string must not be #S#M#S, where # represents the number of nucleotides), and must not map onto decoy sequences, mitochondrial DNA or Y chromosome. In addition, the pair is also excluded if any of the reads is not a primary alignment, fails platform/vendor quality checks, or is PCR or optical duplicate. Non-anchor reads must not contain unsequenced nucleotides ('N') and MAPQ of non-anchor ABERRANT reads must be < 20 . The algorithm dustmasker³¹ is used to identify non-anchor read-pairs containing low complexity sequences, which are later discarded.

2) Clustering

Anchor reads are clustered together if (i) they share the same orientation, and (ii) the distance relative to the nearest mapped read of the same cluster is ≤ 200 bp. Two main cluster categories are defined, namely POSITIVE and NEGATIVE (i.e., anchor reads are mapped onto the positive and negative strand, respectively). A preliminary range of genome coordinates is associated with each single cluster – final breakpoint coordinates are refined in a later step –. Ranges are defined by a lower (left) coordinate (P_L_POS and N_L_POS, respectively for positive and negative clusters) and an upper (right) coordinate (P_R_POS and N_R_POS, for positive and negative). Only clusters consisting of ≥ 4 supporting reads are considered for further analysis. To avoid miscalls due to alignments in complex regions, the full set of reads mapping within cluster coordinates [P_L_POS, N_R_POS] are further analysed, and clusters are removed if: (i) the proportion of reads with MAPQ ≤ 10 relative to the total reads mapped within cluster boundaries represents > 0.3 (30%), and/or (ii) the proportion of reads with CIGAR string #S#M#S relative to the total reads mapped within cluster boundaries represents < 0.15 (15%). Clusters in the tumour are removed if a syntenic cluster in the matched-normal sample is detected with the same

orientation and mapping the same locus less than 500 bp away. Finally, one positive and one negative clusters are reciprocal if $P_R_POS \geq N_L_POS$ and $\text{abs}(N_L_POS - N_R_POS) \leq 350$ bp, otherwise clusters are catalogued as single (or independent).

3) Identification of viral DNA events

Non-anchored reads from each cluster were de novo assembled using Velvet³², and contigs were used as queries of BLAST searches against the RVDB Reference viral database³³ v12.2 containing 2,467,269 viral DNA sequences, of which 91,455 correspond to human Hepatitis B virus (HBV). Only contigs matching human HBV DNA are considered, and reciprocal clusters pointing to HBV DNA are catalogued as canonical HBV DNA insertion events, while single, independent clusters are catalogued as candidates for aberrant HBV DNA integration events. Finally, we used the algorithm MEIBA¹⁵, to identify and reconstruct HBV DNA insertion breakpoints to base-pair resolution, with the following non-default parameters: ‘Maximum number of clipped read clusters in the insertion region’ = 20 (default = 10), and ‘Window size to search for clipped read clusters from discordant read-pair clusters ends’ = 100 bp (default = 50 bp).

Identification of HBV-mediated translocations and validation of v-TraFiC calls using single-molecule sequencing with Oxford Nanopore

We performed long-read whole-genome sequencing with Oxford Nanopore Technologies (ONT) on nine native HCC tumours with relevant HBV DNA insertion events (i.e., SA501491, SA529726, SA529759, SA529830, SA501424, SA501453, SA501481, SA501511, SA501534). Libraries were constructed using the Oxford Nanopore Sequencing 1D ligation library preparation kit (SQK-LSK109, Oxford Nanopore Technologies Ltd) according to the manufacturer's protocol, including an initial DNA repair step with NEBNext FFPE DNA Repair Mix (New England

BioLabs) and NEBNext Ultra II Ligation Module (New England BioLabs). Two low DNA yield samples (SA529726 and SA501481) were whole-genome amplified using ϕ 29 DNA polymerase (REPLI-g midi kit, Qiagen) prior library construction. Amplified DNA was then digested with *t7* endonuclease I (New England BioLabs) for linearization of branched amplicons and deproteinized with Proteinase K (New England BioLabs). Next, unbranched DNA underwent size selection of fragments longer than 20 Kb by means of a Short Read Eliminator buffer (Circulomics) precipitation step and was further purified with Ampure XP Beads (Beckman Coulter Inc). Then, libraries were obtained according the manufacturer's protocol as described above.

Sequencing was performed onto MinION R9.4 flowcells (FLO-MIN106 rev-D, Oxford Nanopore Technologies Ltd), controlled by the Oxford Nanopore MinKNOW software. Base-calling and post-processing of the ONT raw fast5 files was conducted with ONT software Albacore or Guppy to obtain fastq files. Files with quality scores below the recommended values were dropped at this point from further analysis. Reads for each library were then independently mapped to the hs37d5 human reference genome with minimap²³⁴ and the resulting SAM files were converted to BAM files, sorted and indexed using Samtools³⁴. All partial BAM files were merged, sorted and indexed to the final BAM files.

We performed validation of 47 putative somatic HBV insertion events (36 single clusters and 11 reciprocal insertions) identified with v-TraFiC in the 9 HCC tumours that were sequenced using Illumina paired-end and ONT long-reads. For each one of the HBV events we interrogated the long-read tumour BAM file to seek for long-reads validating the event. Two types of supporting reads were employed, namely (i) 'spanning-reads', composed of ONT reads completely spanning the HBV insertion, hence they can be identified as a standard insertion on the reference genome, and (ii) 'clipped-reads', composed of ONT reads spanning only one of the HBV insertion ends,

hence they get clipped the alignment onto the reference genome. HBV events supported by at least one ONT read were considered true positive events, while those not supported by such reads were considered false positive calls. Overall, we find ~10% (5/47) of false positive events (note that this rate could be overestimated due to low coverage in the ONT data). Spanning-reads were used to identify 11 cryptic translocations.

Copy-number dataset

We analyzed copy number profiles obtained by the PCAWG Working Group 11 (PCAWG-11) using a consensus approach combining six different state-of-the-art copy number calling methods³⁵. GC content corrected LogR values were extracted from intermediate Battenberg algorithm results³⁶, smoothed using a running median, and transformed into copy number space according to $n = [2^{\log R}(2(1 - \rho) + \psi\rho) - 2(1 - \rho)] / \rho$, where ρ and ψ are the PCAWG-11 consensus tumour purity and ploidy, respectively.

Identification of telomeric deletions associated to HBV insertion events

Single read clusters, identified with v-TraFiC, supporting an HBV insertion event (i.e., clusters of discordant read-pairs – Illumina – with apparently no reciprocal cluster within the proximal 500 bp, and whose mates support a somatic HBV event), were interrogated for the presence of associated telomeric deletions. Briefly, we looked for copy number loss calls from PCAWG (see “Copy number dataset” above) where: (i) the copy number loss extends from the HBV insertion breakpoint up to the end of the chromosomal arm, involving the telomere, and (ii) one independent cluster, which supports the integration of the HBV event, unequivocally demarcates the copy number loss boundary. We used MEIBA¹⁵ to reconstruct the relevant insertion breakpoint.

449

450 Microhomologies search at breakpoint boundaries

451 Integrated HBV DNA molecules were subjected to microhomology search looking for
452 homologous motifs at the insertion site in the human reference hg19. Briefly, after reconstruction
453 of the HBV-hg19 junctions with MEIGA¹⁵, HBV DNA junctions were mapped onto a database
454 containing a set of HBV-strains consensus sequences using the Biopython PairwiseAligner in
455 ‘local’ mode with a -10 gap penalization. This allowed the identification of the DNA region from
456 the HBV consensus sequence flanking the HBV insertion junction sequence. Then, we compared
457 these HBV flanking sequences with their corresponding breakpoint sequences at the reference
458 genome insertion site.

459

460 Timing of viral insertions

461 We inferred the timing of HBV DNA insertion events that occurred in the context of copy number
462 gains. We employed the SVclone algorithm²⁰ to obtain the number of reads supporting and non-
463 supporting HBV DNA insertions. To deal with HBV insertions supported by single-read clusters
464 only, a modification of the method was implemented to accept structural variants with only one
465 break-end side as follows: (i) relevant filters were switched off in order to allow insertion events
466 with one breakpoint only to be considered by SVclone; (ii) only two types of reads were extracted
467 from the BAM file: split reads (soft-clipped reads that cross each break-end) and normal reads
468 (reads that cross or span the break-ends but match the reference), being spanning reads removed
469 (read pairs that align either side of the break-ends but match the reference). Read counts from
470 SVclone, together with tumour purity and copy number states, were used as input of
471 MutationTime.R²⁰ for the classification of HBV insertions into four different timing categories,

namely clonal [early], clonal [late], clonal [NA] or subclonal. Then, real-time estimates for whole-genome duplication (WGD) events, based on CpG>TpG mutations analysis²⁰, were used to place particular HBV insertion within a chronological time-frame – in years – during a patient’s lifespan, depending on whether mutations occurred before or after a WGD event.

Probe synthesis and fluorescence in situ hybridization

Two sets of bacterial artificial chromosome (BAC) clones (RP5-1125N11 and RP11-891N16 for t(1;11); and RP11-125F4 and RP11-652N13 for t(8;17)) were obtained from the BACPAC Resources Center (<https://bacpacresources.org/>) to develop two-color single-fusion FISH probes to detect chromosome translocations. *ARID1A* deletion probe was developed with RP5-696E2 and RP11-372B18 BAC clones, and Metasystems #D-5103-100-OG probe was used to study *TP53* gene deletion. RP5-1125N11, RP11-125F4, and RP5-696E2 BACs were labelled with Spectrum-Orange, and RP11-891N16, RP11-652N13 and RP11-372B18 with Spectrum-Green. FISH analyses were performed using the Histology FISH Accessory Kit (DAKO) following the manufacturer’s instructions (PMID: 25798834 DOI: 10.1038/onc.2015.70) on 5mm TMA sections mounted on positively charged slides (Thermo Scientific). Briefly, the slides were first deparaffined in xylene and rehydrated in a series of ethanol. Slides were pre-treated in 2-[N-morpholino]ethanesulphonic acid (MES), followed by a 30 min protein digestion performed on proteinase-K solution. After dehydration, the samples were denatured in the presence of the specific probe at 73°C for 5 min and left overnight for hybridization at 37°C. Finally, the slides were washed with 20×SSC (saline-sodium citrate) buffer with detergent Tween-20 at 63°C, and mounted on fluorescence mounting medium (DAPI in antifade solution). Cells were imaged with a Leica DM 5500B fluorescence microscope equipped with a 100x oil-immersion objective, Leica

DM DAPI, Green and Orange fluorescence filter cubes and a CCD camera (Photometrics SenSys camera) connected to a PC running the Zytovision image analysis system (Applied Imaging Ltd., UK) with Z stack software. The z-stack images were manually scored by two independent investigators by counting the number of co-localized signals, representing fused transcripts, or missing signals, representing deletions, all over the tissue.

ACKNOWLEDGEMENTS

We thank the Supercomputing Centre of Galicia (CESGA) for providing complementary computational resources. **Funding:** J.M.C.T is supported by European Research Council Grant ERC-2016-STG – ERC Starting Grant 716290, and Ministerio de Ciencia, Innovación y Universidades, Grants PGC2018-102245-B-I00 and RYC-2014-14999. E.G.A. and P.O. are supported by Ministerio de Educacion Cultura y Deporte, fellowships FPU17/05396 and FPU18/03421, respectively. B.R-M., M.S. and S.Z. are supported by Xunta de Galicia fellowships ED481A-2016/151, ED481A-2017/306, and ED481A-2018/199, respectively. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (grant number MR/L016311/1). M.M-F. is funded by the Spanish Association for Cancer Research (AECC, grant207mart). S.R-P. is supported by grants from the Spanish National Research and Development Plan, Instituto de Salud Carlos III, and FEDER (PI17/02303, PI20/01837 and DTS19/00111); AEI/MICIU EXPLORA Project BIO2017-91272-EXP and AECC_Lab_2020

Project (Asociación Española Contra el Cáncer). J.D. is a postdoctoral fellow of the Research Foundation – Flanders (FWO). PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation’s support towards the establishment of The Francis Crick Institute. This research was funded in part by the Wellcome Trust (FC001202). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

E.G.A., P.V.L., P.J.C., J.M.C.T. conceived the study. E.G.A., J.D., J.Z., C.J., A.B-O., B.R-M., K.H., M.S., S.Z., A.P.B., P.V.L., J.M.C.T. contributed to pipelines. E.G.A., J.D., J.Z., C.J., M.T., K.H., P.V.L., P.J.C., J.M.C.T. analyzed the sequencing data. J.T., D.G-S., J.R-C performed Oxford Nanopore sequencing. M.T., M.M-F., R.A., A.O., A.LB. performed laboratory experiments. M.T., H.N., X.F., S.P.P, A.O., H.A., K.C., M.U., S.H., H.Y. provided tumour specimens and/or performed pathological diagnosis. D.G-S., U.G., M.G-B., J.M.C.T. build a comprehensive biological model for HBV-mediated rearrangements. R.T., S.R.P. performed cytogenetics. E.G.A., P.V.L., P.J.C., J.M.C.T. wrote the manuscript with assistance from all authors. M.T., P.V.L., P.J.C., J.M.C.T. supervised the project.

COMPETING INTERESTS

The authors declare no competing interests.

CODE AVAILABILITY

A preliminary version of the code v-TraFiC for the identification of somatic HBV insertions, is available at <http://gitlab.com/mobilegenomesgroup/v-TraFiC>. A final version of the code will be available together with complete documentation and tutorials after publication.

DATA AVAILABILITY

Sequencing data has been generated in the framework of the Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative have been deposited in a public databases¹³. All data generated for this manuscript is available in the Supplementary Tables.

549 REFERENCES

- 550
- 551 1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and
552 mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424
553 (2018).
 - 554 2. Midorikawa, Y. *et al.* Molecular karyotyping of human hepatocellular carcinoma using
555 single-nucleotide polymorphism arrays. *Oncogene* **25**, 5581-90 (2006).
 - 556 3. Wong, N. *et al.* A comprehensive karyotypic study on human hepatocellular carcinoma
557 by spectral karyotyping. *Hepatology* **32**, 1060-8 (2000).
 - 558 4. Meyer, M., Wiedorn, K.H., Hofschneider, P.H., Koshy, R. & Caselmann, W.H. A
559 chromosome 17:7 translocation is associated with a hepatitis B virus DNA integration in
560 human hepatocellular carcinoma DNA. *Hepatology* **15**, 665-71 (1992).
 - 561 5. Meng, G. *et al.* TSD: A Computational Tool To Study the Complex Structural Variants
562 Using PacBio Targeted Sequencing Data. *G3 (Bethesda)* **9**, 1371-1376 (2019).
 - 563 6. Wang, Y. *et al.* Characterization of HBV integrants in 14 hepatocellular carcinomas:
564 association of truncated X gene and hepatocellular carcinogenesis. *Oncogene* **23**, 142-8
565 (2004).
 - 566 7. Rogler, C.E. *et al.* Deletion in chromosome 11p associated with a hepatitis B integration
567 site in hepatocellular carcinoma. *Science* **230**, 319-22 (1985).
 - 568 8. Pineau, P. *et al.* A t(3;8) chromosomal translocation associated with hepatitis B virus
569 intergration involves the carboxypeptidase N locus. *J Virol* **70**, 7280-4 (1996).
 - 570 9. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of
571 hepatocellular carcinoma patients. *Genome Res* **22**, 593-601 (2012).
 - 572 10. Peneau, C. *et al.* Hepatitis B virus integrations promote local and distant oncogenic driver
573 alterations in hepatocellular carcinoma. *Gut* (2021).
 - 574 11. Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of
575 noncoding and structural mutations in liver cancer. *Nat Genet* **48**, 500-9 (2016).
 - 576 12. Tu, T., Budzinska, M.A., Shackel, N.A. & Urban, S. HBV DNA Integration: Molecular
577 Mechanisms and Clinical Implications. *Viruses* **9**(2017).
 - 578 13. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-
579 93 (2020).
 - 580 14. Zapatka, M. *et al.* The landscape of viral associations in human cancers. *Nat Genet* **52**,
581 320-330 (2020).
 - 582 15. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver
583 rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* **52**, 306-319 (2020).
 - 584 16. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature*
585 **578**, 112-121 (2020).
 - 586 17. Nagaya, T. *et al.* The mode of hepatitis B virus DNA integration in chromosomes of
587 human hepatocellular carcinoma. *Genes Dev* **1**, 773-82 (1987).
 - 588 18. Li, Y. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute
589 lymphoblastic leukaemia. *Nature* **508**, 98-102 (2014).
 - 590 19. Sullivan, B.A. & Willard, H.F. Stable dicentric X chromosomes with two functional
591 centromeres. *Nat Genet* **20**, 227-8 (1998).
 - 592 20. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv* (2018).

21. Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol* **19**, 95 (2018).
22. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-7 (2015).
23. Jordan, J.J. *et al.* Altered-function p53 missense mutations identified in breast cancers can have subtle effects on transactivation. *Mol Cancer Res* **8**, 701-16 (2010).
24. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
25. Sun, X. *et al.* Arid1a Has Context-Dependent Oncogenic and Tumor Suppressor Functions in Liver Cancer. *Cancer Cell* **32**, 574-589 e6 (2017).
26. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719-24 (2009).
27. Li, Y. *et al.* Patterns of structural variation in human cancer. *bioRxiv* (2017).
28. Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O. & Stein, L.D. Pan-cancer analysis of whole genomes. *bioRxiv* (2017).
29. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
30. Tubio, J.M. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
31. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028-40 (2006).
32. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
33. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A.S. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**(2018).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
35. Dentre, S.C. *et al.* Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv* (2018).
36. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).