HBV Genome-Enriched Single Cell Sequencing Revealed Heterogeneity in HBV-Driven HCC

Wenhui Wang
Icahn School of Medicine at Mount Sinai
Yan Chen
Huazhong University of Science and Technology Tongji Hospital Hepatic Surgery Center
Liang Wu
Beijing Genomics Institute: BGI Group
Yi Zhang
Hebei University of Science and Technology
Seungyeul Yoo
Icahn School of Medicine at Mount Sinai
Quan Chen
Icahn School of Medicine at Mount Sinai
Shiping Liu
Beijing Genomics Institute: BGI Group
Yong Hou
Huazhong University of Science and Technology Tongji Hospital Hepatic Surgery Center
Xiao-ping Chen
Huazhong University of Science and Technology Tongji Hospital Hepatic Surgery Center
Qian Chen
Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Jun Zhu (jun.zhu@mssm.edu)
Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0003-0834-8178

Research article

Keywords: Hepatocellular Carcinoma, Hepatitis B virus integration, Enriched single cell sequencing, Copy number variation, Clonal evolution.

Posted Date: May 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-537064/v1

HBV Genome-Enriched Single Cell Sequencing Revealed Heterogeneity in HBV-Driven HCC

Wenhui Wang[1,2†], Yan Chen[3†], Liang Wu[4], Yi Zhang[5], Seungyeul Yoo[1,2,6], Quan Chen[1,2,6], Shiping Liu[4], Yong Hou[4], Xiao-ping Chen[3], Qian Chen[7*], Jun Zhu[1,2,6,8*],

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[3]The Hepatic Surgery Centre at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (HUST), Wuhan, China
[4]BGI, Shenzhen, China
[5]Department of Mathematics, Hebei University of Science and Technology, Shijiazhong, Hebei, China
[6]Sema4, a Mount Sinai venture, Stamford, Connecticut, United States
[7]The Division of Gastroenterology, Department of Internal Medicine at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (HUST), Wuhan, China.
[8]The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[†]co-first authors with equal contribution
[*]Corresponding Authors:
Dr. Jun Zhu
Department of Genetics and Genomic Sciences,
Icahn School of Medicine at Mount Sinai,
1425 Madison Ave. New York, NY, United States. 10029
Tel: 212-659-8942
Email: jun.zhu@mssm.edu

Dr. Qian Chen
The Division of Gastroenterology, Department of Internal Medicine at Tongji Hospital,
Tongji Medical College, Huazhong University of Science and Technology (HUST),
Wuhan, China
Email: chenqian201579@yahoo.com

Abstract

Background: HBV-HCC is heterogeneous and frequently contains multifocal tumors. To interrogate heterogeneity of HBV-HCC, we developed a HBV genome enriched single cell sequencing (HGE-scSeq) procedure and a computational method to identify HBV integration sites and infer DNA copy number variations (CNVs).

Results: We performed HGE-scSeq on 269 cells from 4 tumor sites and 2 tumor thrombi of a HBV-HCC patient. HBV integrations were identified in 142 out of 269 (53%) cells sequenced, and were enriched in two HBV integration hot spots chr1:34,397,059 (*CSMD2*) and chr8:118,557,327 (*MED30/EXT1*). There were also 162 rare integration sites. HBV integration sites were enriched in DNA fragile sites and sequences around HBV integration sites were enriched for microhomologous sequences between human and HBV genomes. Cells were grouped into 4 clonal groups based on CNVs. The HBV integration heterogeneity was associated with single cell's CNVs. All of 269 cells carried chromosome 1q amplification, a recurrent feature of HCC tumors, suggesting that 1q amplification occurred before HBV integration events in this case study. Further, we performed simulation studies to demonstrate that the sequential events (HBV infecting transformed cells) could result in the observed phenotype with biologically reasonable parameters.

Conclusion: Our HGE-scSeq data reveals high heterogeneity of HCC tumor cells in terms of both HBV integrations and CNVs.

Keywords: Hepatocellular Carcinoma, Hepatitis B virus integration, Enriched single cell sequencing, Copy number variation, Clonal evolution.

Background

Hepatocellular carcinoma (HCC) is ranked as the third most lethal cancer worldwide [1], and 54% of HCC cases originate from chronic Hepatitis B Virus (HBV) infection [2]. During HBV infection, a small fraction of viral replication is in double-stranded linear DNA (dslDNA) form, which can be inserted into the host genome at double-stranded break points [3]. HBV integrations only occur in early phase of HBV infection [3, 4]. HBV integration into the human genome is one of the most important etiological mechanisms of HBV inducing HCC [5]. Recurrent HBV integrations are identified by sequencing studies [6-11].

HBV-HCC tumors are of high heterogeneity in terms of HBV DNA integration patterns and somatic genomic alterations, and the heterogeneity associates with prognosis and drug response in HBV-HCC [12]. Both empirical and simulation studies show that only integration events of high allele frequency can be detected at a given sequencing depth [9, 13]. It is expensive to implement whole genome sequencing (WGS) with high sequencing depth in a large scale study in general. HIVID (high-throughput Viral Integration Detection) by Li *et al*. [14] describes an efficient way to accurately detect HBV integration in the whole genome. Regions containing virus genome sequences are enriched in the process of preparing DNA library so that the genomic regions to be sequenced are remarkably smaller than the whole human genome. Recently, HIVID has been applied in sequencing of a large number of HBV-HCC samples [15] as well as in detecting Human papillomavirus (HPV) integration sites [16].

DNA single cell sequencing has demonstrated its power in studying tumor clonal expansion and tumor heterogeneity. Navin *et al.* [17] first introduce DNA single cell sequencing technique in tumor evolution study. Although only 6% of genome is covered due to limitation of whole genome amplification technique (Sigma-Aldrich GenomePlex WGA4 kit), computational methods have been developed to accurately estimate DNA copy number variations (CNVs). Zong *et al.* [18] proposed the multiple annealing and looping-based amplification cycles (MALBAC) for whole genome sequencing.

Both GenomePlex and MALBAC are extensively reviewed and compared with multiple displacement amplification (MDA) under different circumstances [19-26] due to the vital importance of Whole Genome Amplification (WGA) in DNA single cell sequencing. However, none of WGA methods perform consistently best in all situations. Some studies suggest MDA as the approach [20, 21, 23] while other studies disagree [19]. In general, studies [24-26] indicate that MDA performs well in terms of single-nucleotide variations detection and CNVs detection. Single cell sequencing has been used in studying human brain cells [27], kidney cancer [28], lung cancer [29], bladder cancer [30], JAK2-negative myeloproliferative neoplasm [31], and individual's gamete genomes [32]. More recently, Wang *et al.* [33] and Leung *et al.* [34] significantly improved the WGA technique by sequencing cells under G2/M stage when the cell has two times DNA material comparing to other stages. The coverage width has been increased to 91%, which makes it possible to study the single nucleotide variation at single cell level [33, 34].

With single cell sequencing technology advances, several open questions of HBV-HCC tumorigenesis need to be re-examined. (1) What is the frequency of HBV integration? The frequency of HBV integration is estimated in the range of 1 in per 1000 hepatocytes [35, 36]. The expected frequency of 2 HBV integrations in one hepatocyte is ~10^{-6} , a unlikely event under a normal condition as suggested in literature [3]. As HBV integrations occur in early phase of HBV infection [3, 4], HBV integrations will not increase during tumorigenesis. Thus, multiple HBV integrations occurred in one hepatoma cell is highly unlikely as well. However, there are HBV-HCC cell lines with multiple integrations [37, 38]. A single cell genome sequencing study also indicates that there are 5-6 HBV integrations in a cell, which are also identified by bulk tissue WGS [39]. It is shown that tumor-initiating cells are more prone to HBV integration due to genome instability [40]. It is possible that integration frequency is much higher in cells prone to double-stranded breaks [41]. (2) What is the role of HBV integrations, initiating tumorigenesis or accelerating clonal expansion of tumor-initiating cells? (3) Clonal relationship among multifocal HBV-HCC tumors in terms of HBV integrations and CNV patterns?

To address these questions, we aimed to examine cells from a HBV-HCC patient with multifocal tumors and metastasis using single cell genomic sequencing technique. One of the limitations with sequencing studies on HBV integration is the low coverage of HBV reads. It is expensive to sequencing whole genomes at high depth for a large number of single cells. In this paper, we present a proof-of-concept study based on HBV genome-enriched Single cell sequencing (HGE-scSeq) approach to identify the heterogeneity of HBV integrations and genomic alterations in HBV-HCC tumor cells at the single cell level (Figure 1). The approach has potential to achieve two goals: to yield a high coverage at HBV-integration sites for HBV integration identification and to generate shallow genome-wide sequencing for CNV estimation. We performed HGE-scSeq and WGS on a HBV-HCC cell line MHCC97H, and validated HGE-scSeq results using the WGS data. We then performed HGE-scSeq on 269 cells from 4 independent tumor sites and 2 tumor thrombi from a HBV-HCC patient (Supplementary Figure S1) and HBV genome-enriched bulk sequencing data on 4 corresponding adjacent normal tissue samples. HBV virus sequences were detected in 205 of 269 cells, and HBV integrations were identified in 142 cells. In total, 471 integration events corresponding to 164 unique integration sites were observed. Two integration hot spots, chr1:34,397,059 (*CSMD2*) and chr8:118,557,327 (*MED30/EXT1*), were detected in 100 and 121 cells, respectively. And we showed that these genes affected tumor cell growth experimentally [42]. In addition, we estimated CNVs based single cell sequencing data and showed that all 269 cells contained 1q amplification, a recurrent feature of advanced HCC [43]. Based cell's CNVs, we further clustered 269 cells into 4 clonal groups. Our result of this HBV-HCC case suggests that (1) there were multiple HBV integrations per cell, the integration frequency was much higher than 1 integration per 1000 cells, suggesting there were cells of genome instability before HBV infection; (2) there were heterogeneities in both HBV integrations and CNVs at the single cell level; (3) multifocal tumors and tumor thrombi had a common origin with common CNV and HBV integration patterns. To show the scenario is reasonable, we performed a series of simulation studies to demonstrate that the proposed sequential events resulted in observed phenotypes with biologically reasonable parameters.

Results

Identification of HBV integration sites and estimation of CNVs in HBV-HCC cell line MHCC97H

MHCC97 is a HBV positive, highly metastatic HCC cell line [44]. MHCC97H is further isolated from MHCC97 due to its higher metastatic potential [45]. We characterized MHCC97H by WGS with 1,485,306,632 100bp pair-end reads. After read QC [13] (Methods), 1,308,162,600 reads were mapped to human genome with average 42.2 folds coverage. CNVs of MHCC97H were estimated based on the WGS data. Read counts were normalized and corrected for GC content. CBS [46] was used to infer the segmentation. CNVs of MHCC97H were also measured using SNP arrays (GSE38326 [47]). The copy number amplifications based on WGS and SNP arrays were similar (correlation γ=0.96, Supplementary Table S1, Supplementary Figure S2).

We performed HGE-scSeq on 5 MHCC97H cells. For each cell, 32,253,536 (in average) reads were generated (Supplementary Table S2). After read QC (Methods), 10,336,455 (in average) reads were included in further analysis. Among them, 19,717 (in average) contained sequences in HBV genome, and 5,452,432 (in average) were mapped to human genome (Supplementary Table S2).

HBV integration sites

For the WGS data, we applied the pipeline as described previously [13] and set the threshold of supporting reads (one soft clipped read or 2 adjacent reads). Total 5 HBV integration sites were identified (Supplementary Table S3). For HGE-scSeq data, 22-69 integration sites were identified in each cell, resulting total 176 unique integration sites (Methods, Supplementary Table S4). If treating WGS and HGE-scSeq data derived integration sites that were within 5000bp as the same sites, 57 of HBV integration sites based on single cell data matched with 4 integration sites based on WGS (Highlighted in Supplementary Table S4). Each cell had 2-4 integrations common with the integrations identified by WGS. Among 176 HBV integration sites, 41 were identified in at least 2 cells (Supplementary Table S4).

CNV estimations

Even though sequencing libraries were enriched for HBV genome sequences, on average 52.97% of reads were mapped to human genome and 2.68% of human genome covered with at least one read. Some regions were covered by multiple reads. Numbers of reads at each locus across human genome followed a Poisson distribution (Supplementary Figure S3, chi-seq test, p-value 0.98). And the loci covered by reads in multiple cells were enriched in copy number amplified regions defined by WGS (Supplementary Figure S4). To check whether there were any genome feature differences between human genome regions with and without mapped reads, we first constructed a Fisher machine prediction model [48] to distinguish HBV and human genomes (Supplementary Figure S5A, Methods). Then, we applied the Fisher machine to quantify sequence feature differences between genome regions with and without mapped reads. There was no clear difference between human genome regions with and without mapped reads (Supplementary Figure S5 B&C). These results together suggest that HGE-scSeq reads dispersed randomly across human genome.

We developed a method to infer CNVs based on HGE-scSeq data (Methods) and applied it to infer CNVs of MHCC97H cell line. The inferred CNVs based on HGE-scSeq data were consistent with WGS and SNP array data (correlation γ=0.85-0.92 and 0.8-0.88, respectively, Supplementary Table S1, Supplementary Figure S2).

Heterogeneity of MHCC97H cells

A single cell genomic sequencing study of HepG2 cells suggests that HepG2 cells are heterogeneous in term of CNVs [49], and the variation of CNVs among cells are consistent with transcription level variations at the CNV regions, suggesting the variations are unlikely due to random errors in single cell sequencing. Our HE-scSeq data of MHCC97H cells were not identical in terms of both HBV integrations and CNVs at the single cell level. The variations could be due to errors introduced during genome multiplication and sequencing or due to true heterogeneity of cells in a cell line. Multiple HBV

integrations were identified in more than one cell, suggesting these HBV integrations were unlikely resulted from random sequencing errors. As HBV integrations only occur in early phase of HBV infection [3, 4], the results suggest that the rare integrations might not have any impact on cell proliferation so that the composition of cells with different HBV integrations was stable during cell passage.

HGE-scSeq of multifocal HBV-HCC tumors

HGE-scSeq was applied to 269 cells from 6 sites (Supplementary Figure S1). HBV virus sequence reads were detected in 205 out of 269 cells (detailed in Methods, Figure 2A). HBV assemblies were close to HBV isolate G247-B3 (an example of pileup of sequencing reads is shown in Supplementary Figure S6). It is worth to note that HBV sequencing reads from normal tissues contained reads covering the whole HBV genome (Supplementary Figure S6A). In contrast, the HBV virus assemblies from all single cells missed most of the HBV genomic region encoded for X protein (Supplementary Figure S6B).

Heterogeneity of HBV integrations

Among the 205 cells with HBV sequence reads detected, HBV integrations were detected in 142 cells (detailed in Methods). A total 471 integration events were identified (Supplementary Table S5, which corresponds to 164 unique integration sites (Supplementary Table S6). The HBV integration sites were not evenly distributed across the human genome (Figure 2B). There were two integration hot spots, chr1: 34,397,059 (*CSMD2*) and chr8:118,557,327 (*MED30/EXT1*), where the integration events were identified in 100 and 121 cells, respectively (Figure 2B). With regard to HBV genome, most of HBV integrations located in HBVgp2_S, HBVgp3_X and HBVgp4_Precore/Core proteins (Supplementary Figure S7A) with the integrations at the hot spot on human chr1 mapped to HBVgp3_X while the ones at the hot spot on chr8 mapped to HBVgp4_Precore/Core. The distribution of HBV integration sites across HBV genome is shown in Supplementary Figure S7B. On average 3.32 integration events were detected in each

cell. Based on HBV integration profile, cells were clustered into two groups with one group only carrying integrations at the hot spots and the second group carrying extra rare integrations (Figure 2C). Numbers of sequencing reads for cells in the two groups were similar (Supplementary Figure S8). Most of integration sites were detected in only in 1 cell. Only 39 integration sites were detected in multiple cells or multiple tumor sites. The heterogeneity on frequency of HBV integrations across cells and tissues was observed. All the cells with HBV integration carried at least one of the hotspot integrations. The HBV integration sites were distributed across 46 genes or gene pairs based on UCSC known genes. The integration sites at the two hot spots, chr1: 34,397,059 (*CSMD2*) and chr8:118,557,327 (*MED30/EXT1*), were not reported in previous HBV integration studies (except in this dataset as we previously reported [42]), but overlapped with multiple fusion events from both cancer cell lines and TCGA [50] (Supplementary Table S7).

Next, we compared HBV integration patterns in adjacent normal tissues close to the 4 tumor sites. In total, 17 integration events (Supplementary Table S5) were detected at 13 loci (Supplementary Table S6) in the four adjacent normal tissues. The numbers of HBV integrations in adjacent normal tissues and in tumors were not directly comparable as one based on bulk tissue sequencing and one based on single cell genomic sequencing. In a loose sense, there were more integration events in tumors than in normal tissues than tumors, consistent with previous reports [51]. The integration sites at the two hot spots were also detected in each adjacent normal tissue except that the integration site at chr1 hot spot was not detected in N1 and chr8 hot spot integration was not detected in N2 (in which only one soft clipped read was detected and less than the minimum threshold of two soft clipped reads). The integration events at the two hot spots were the only two recurrent events across 4 adjacent normal tissues. The available information is not sufficient to distinguish whether HBV integrations at the two hot spots in adjacent normal tissues were results of clone expansion or diffusion from tumor tissues. Additional information is needed to inform clonal relationship between cells with HBV integrations at the two hot spots in adjacent normal and tumor tissues.

Properties of HBV integration sites

The mechanism of how HBV genome is integrated into the human genome is still under explored. Hu *et al*. [16] observed significant enrichment of microhomologous (MH) sequences at or near 120 HBV integration sites detected from 31 liver samples from Sung *et al* [8]. Recently, Zhao *et al*. [52] sequenced 426 HBV-HCC patients and shows enrichment of microhomologous sequences around the HBV integration sites as well. All these observations suggest the potential involvement of MH mediated mechanism in the process of HBV integration. Based on single cell sequencing data we identified 164 unique integration sites. Microhomologous sequences between the human genome and HBV genome (an example shown in Figure 2D) were enriched at the HBV integration sites (Figure 2E). We also found the enrichment of integration sites within the common and rare fragile regions [53] (Figure 2F). The enrichment of microhomologous sequences near HBV integration and enrichment of HBV integration on fragile region elucidate that the HBV integration is a physical driven process, which is highly related with the sequence content and corresponding physical characteristics of host genome sequence.

HBV integration hot spots

The two integration hot spots, chr1: 34,397,059 and chr8:118,557,327 locate at the intronic region of *CSMD2* and the intergenic region of *MED30-EXT1*, respectively. The chr1 hot spot could partially be explained by microhomology (Figure 2D), which led to loss of *CSMD2* expression. The integration at the chr8 hot spot resulted over expression of *EXT1*, which promoted cell growth in vitro and in vivo [42].

Heterogeneity of CNVs

In addition to HBV integration, we estimated each cell's CNVs based on the HGE-scSeq data (Methods). The whole human genome was divided into 5000 bins which were then group into 49 super-bins of equal sizes for visualization purpose. As expected, most of the bins had normal copy number of DNA (Figure 3A). All cells carried a DNA copy number amplification at chromosome 1q, which is a recurrent feature of HCC [43] (Figure 3A). The cells were clustered into 4 clone groups based on CNVs (Figure 3A), each clone had a distinct pattern of DNA copy number amplifications. And each clone group contained cells

with different types of HBV integrations (Figure 3B). From clones 1 to 4, the ratio of cells carrying rare integrations decreased.

Clonal evolution and its relationship with HBV integration

Based on CNV pattern, we constructed a phylogenetic tree (detailed in Methods, Figure 4A), which suggests that clone 1 directly developed from the ancestor. Clone 2 and clones 3&4 were derived from clone 1, suggesting there were two different evolution directions. The inner node corresponding to the origin of clone 2 and clones 3&4 as well as the inner node corresponding to the split between clone 2 and clones 3&4 were annotated in Figure 4A. These inner nodes can be directly linked to CNVs on a specified region. The root node in the phylogenetic tree corresponded to the cells with CNVs of 1q. The common origin of clones 2, 3, and 4 had Chr11 amplification. The regions differentiating clones 2-4 from clone 1 contained potential genomic regions that may associate with the decreasing ratio of rare integration carrying cells. Cells in clone 2 contained CNVs on Chr11 while cells in clones 3 and 4 contained additional CNVs at Chr8:118,268,000-146,364,000. More CNVs split clones 3 and 4. Supplementary Figure S9 is the same as Figures 4A except nodes colored according to cells with hot spot and rare HBV integrations. It is clear that rare HBV integrations were not randomly distributed in the phylogenetic trees.

To identify the potential CNV regions associated with decreasing number of rare HBV integrations, we tested the association between CNV and HBV integrations in the clone evolution process from clone 1 to clones 2-4 and the split between clone 2 and clones 3&4 separately. The significant regions (Supplementary Table S8) associated with the HBV integration difference between clone 1 vs. clones 2-4 were enriched for immune related genes (Table 1). Genes encoding for secretoglobin family proteins (*SCGB1A1*, S*CGB1D1*, *SCGB1D2*, *SCGB1D4*, *SCGB2A1*, and *SCGB2A2*) were enriched in the regions (Fold change=50.9, p-value=5.8E-8). Secretoglobin family 1 proteins have anti-inflammation and immunomodulation property [54] and are inducible by interferon-gamma [55]. Members (*APOA1*, *APOA4*, *APOA5*, *SAA1*, *SAA2*, and *SAA4*) of high density liporprotein (HDL) were significantly enriched

in the regions (Fold change=32, p-value=8.5E-7). It has been shown that serum HDL level is reversely associated with serum HBV DNA level [56]. Similarly, AIM2 inflammasome complex was enriched (p-value=2.8E-5, Fold change=58.4), which contains genes *CASP1*, *CASP4*, *CASP4* and *CASP12*. In addition, *AIM2* locates in chromosome 1q, which was amplified in all cells (Figure 3). AIM2 inflammasome complex is reported contributing to the defend against bacterial and double-stranded viral DNA [57]. Another annotated inflammasome IPAF complex was enriched (p-value=1.4E-5, Fold change=70.1). Inflammasome has been shown to relate to both cancer suppression and promotion under different context, which is described as "double-edged sword" for cancer development [58]. Serum amyloid A (SAA) proteins, which were also significantly enriched in the regions (Fold change=115.6, p-value=2.4E-6), interact with inflammasome [59]. For the evolution process separating clone 2 and clone 3&4, significant regions consisting of 48 genes (Supplementary Table S9) were identified. These genes were enriched for genes in Urokinase-type plasminogen activator receptor (uPAR) complex (p-value=1.1E-6, Fold change=182.4, Table 2), which expression is elevated during inflammation and tissue remodeling [60], again suggesting that tumor cells of different genomic features may have different ability against HBV replication and HBV insertion. Also, uPAR expression is associated with invasiveness of malignant tumor cells [61], which is consistent with the observation that more than 50% cells in the two tumor thrombi were clones 3 and 4 (Figure 4B).

Clone 2 vs. other clones

Somatic mutation patterns were derived from bulk tissue whole genome sequencing T1-4 tumors, 2 thrombi and a normal tissue [42]. A phylogenetic tree was constructed based the somatic mutation patterns, which suggested that the largest tumor T1 was the primary tumor and other tumors were derived from T1 [42]. Even though all tumors were from the same origin, the clonal composition of each tumor was different. The proportion of clone 2 cells was significantly higher in T1 than in other tumors (Figure 4B). To identify differences between clone 2 cells and other cells, we compared CNVs across all bins and identified 282 bins (consisting of 2246 genes) where clone 2 cells had lower CNVs comparing to cells of

other clones. These genes were enriched in the GO term calcium-dependent cell-cell adhesion (p-value=9.6E-7, Fold enrichment=3.7, 0 3) and chemokine activity (p-value=6.2E-6, Fold enrichment=4.0, Table 3). N-cadherin promotes cancer cell invasion [62]. Chemokines and their receptors involved in tumor immunogenicity and aggressiveness [63, 64]. Lower abundance of chemokines and their receptors might lead to lower potential to metastasis. These may explain why the fraction of clone 2 cells in the primary tumor T1 was higher than the fractions in other tumors (Figure 4B).

Simulation of clonal evolution

To assess different clonal evolution scenarios, we performed cell simulations according the birth-death [65, 66]. We tested a wide range of parameter space, then calculated the posterior of parameters based on the distance of simulated distribution and the observed data. A simulation starts from a cell after malignant transformation. In the observed data (Figure 4A), the root node had to carry the chromosome 1q amplification. Otherwise, no simulation resulted in the scenario that 100% cells carried the chromosome 1q amplification. In each replication cycle, a cell divided or died at the probability P_{div} and Q_{death}, respectively (Figure 5A). The simulations stopped when the total number of cells reached to 10^7, corresponding to a tumor of size 0.5cm x 0.5cm x 0.5cm. First, we simulated clonal evolution due to CNV changes without HBV integrations. Each novel CNV change likely alters the fitness of the cell and increases the probability of cell division over the probability of cell death, and the selection coefficient was noted as SC (Figure 5A). With n CNVs acquired addition to the root event, the division probability was $P(1+SC)^n$, and the corresponding death probability was $1-P(1+SC)^n$. In a normal cell, the DNA copy number mutation rate (MR) per cell per division is in the range from 10^{-10} to $3.4*10^{-6}$ [67] . We simulated HCC cells with the copy number mutation rate (5e-6,1e-5,5e-5,1e-4,5e-4,1e-3) and the selection coefficient (0.01, 0.05, 0.1, 0.2, 0.3) for each additional CNV. For each simulation, a CNV among the CNVs in Figure 3A was randomly draw and introduced to the cell according the mutation rate. With10,000 cell populations simulated and compared with the observed one, the posterior of parameters (Figure 5B) indicated the parameter combination SC=0.01 and MR=0.001 fitted the observation the best.

Next, we performed simulations to examining HBV integrations with the parameter combination for CNVs fixed as SC=0.01 and MR=0.001 estimated above. We assumed HBV infection occurred when the tumor grew to 10^5 cells and random HBV integrations occurred in 1 out of 50 HCC cells in the tumor (Figure 6A). Among the HBV integrations, 1% were hot spot integrations, and only cells with hot spot integrations gained cell growth advantage with the selection coefficient SC_{HBV} in (0.01,0.05, 0.075,0.1,0.2,0.3). Similar as above, the simulations stopped when the total number of cells in the tumor reached to 10^7 cells. For each SC_{HBV} we simulated 2000 cell populations/tumors. Then, we compared the ratios of cells with HBV integrations among cells in tumors at the end of simulation (Figure 6B). After HBV acute infection, 2% of cells in the simulated tumor carried HBV integrations (blue line in Figure 6B). When the simulated tumors reached to 10^7 cells, around 50% of cells carried HBV integrations with SC_{HBV} in the range between 0.075 and 0.1, close to the ratio 53% observed in the patient data (red line in Figure 5B). Similarly, after HBV acute infection, $2x10^{-4}$ of cells in the simulated tumor carried HBV integrations (blue line in Figure 6C). When the simulated tumors reached to 10^7 cells, around 50% of cells carried hot spot HBV integrations with SC_{HBV} in the range between 0.075 and 0.1, close to the ratio 52% observed in the patient data (red line in Figure 6C), indicating the ratios of cells with hot spot HBV integrations vs. cells with HBV integrations were close to 1 (Figure 6D).

Discussion

HBV genome-enriched single cell sequencing approach can efficiently identify HBV integration sites and genomic alterations. We developed a data analysis pipeline for HBV genome enriched single cell sequencing data. Our analyses revealed both highly recurrent and rare HBV integrations in cells. Especially, a large number of rare HBV integrations were identified in the single cell sequencing study, and these rare HBV integrations suggest that HBV genome was randomly integrated at sites according to physical properties (Figures 2E&2F). The MH enrichment around the HBV integration sites indicates that MH mediates HBV integrations in general. The HBV integration frequency here was much higher than 1 integration expected per 1000 cells [35, 36], suggesting that cells of genome instability (which leading to

higher HBV integration frequency [40]) existed before HBV infection. The result is consistent with the result that all tumor cells had 1q amplification but not all tumor cells had HBV integrations. Our simulation studies demonstrated that the event sequence was possible under biological possible parameters (Figures 6).

There were two HBV integration hot spots (Figure 2C). The integration hot spot chr1: 34,397,059 (*CSMD2*) could partially be explained by microhomology (Figure 2D). For the HBV integration hot spot at chr8, *EXT1* expressed significantly higher in tumor tissues than in adjacent nonneoplastic liver tissues (Supplementary Figure S10), and high expression of *EXT1* was associated poor prognosis in lung, thyroid, and cervical cancers in TCGA. *EXT1* promoted cell growth in vitro and in vitro [42]. These results together suggest that the hot spots chr1: 34,397,059 (*CSMD2*) chr8:118,557,327 (*MED30/EXT1*) were likely due to proliferation advantage of cells with these integrations over other cells.

Our analyses indicate that not only HBV integration sites but also CNVs can be identified from HBV genome-enriched single cell sequencing data. Both CNV analysis and cell evolution analysis suggested that 1q amplification was one of driver alternations [43] (Figures 3A and 4A). The 1q amplification common among all tumor cells based on HBV genome-enriched single cell sequencing data was consistent with bulk tissue whole genome sequencing data [42]. Although T2-4 and the two thrombi were derived from T1 [42], our CNV analyses revealed heterogeneity of clone composition at each tumor site. All clones were likely to share the same clone origin (Figure 4A). The clone compositions at T2-4 and the two thrombi sites were different from the composition of the original tumor site T1 (Figure 4B), likely due to different invasive potentials of each clone, e.g. clones 3 and 4 cells were more likely to invade to other sites.

There are multiple limitations of the HGE-scSeq approach. The sensitivity of the approach is hard to estimate unless an extensive single cell whole genome sequencing was performed as the ground truth to compare with, which is expensive to do. Given the uncertainty of the sensitivity, it is not clear

whether some tumor cells lacking chr1 or chr8 hot spot integrations were due to capture/sequencing sensitivity or due to clonal expansion. We compared two scenarios: (1) the root clone had HBV integration, which drive tumorigenesis. In this scenario, all clones should have exact same HBV integration pattern (as HBV integration occurs only in early phase of HBV integration [3, 4]), which contradicts with our observation that some clones had more HBV integrations than other (Figure 3B). (2) the root clone had 1q amplification, and the root clone cells were of genome instability. Then, HBV infection occurred and HBV integration in each cell occurred at different sites and at different frequency depending on cell's molecular state and genome stability, consistent with our observation (Figure 3B). The cells in clone 4 were more likely missing the hot spot integrations than the cells in clone 1, suggesting that no hot spot integration in these cells was unlikely due to the sensitivity of the assay but due to molecular differences between the clones.

The relationship between CNVs and HBV integrations observed in this case study needs to be considered as anecdotal until the relationship can be replicated in more patient samples or validated in vitro experiments that exceeds the scope of this study.

Conclusion

We developed a data analysis pipeline for HBV genome-enriched single cell sequencing data. HCC tumor cells were heterogeneous in terms of both HBV integration sites and CNVs. The frequency of HBV integration observed in the study was much higher than expected. For the HBV-HCC case in the study, multifocal tumors and tumor thrombi shared common HBV and CNV patterns, suggesting that they shared the same tumor origin.

Methods

Patient and tissue samples

The study of tumor cell heterogeneity was approved by the Institutional Review Board of Tongji Hospital, Tongji Medical College of HUST, in Hubei province, China. The signed written informed consent was obtained before patients' recruitment, according to the regulations of the institutional ethics review boards. The patient and sample information was detailed in Chen et al. [42]. The clinicopathological information of the patient is summarized in Supplementary Table S10. In brief, a 47-year-old patient matched with the research design. Obtained medical history indicated that he had no history of alcohol abuse, recognized acute hepatitis, mother-to-child transmission of HBV, blood transfusion, or injection drug use. Tests indicated the patient had a resolved HBV infection (HBs Ab level 884.5 mIU/mL, HBs Ag-negative, HBc Ab-positive, HBe Ab-positive, HCV Ab-negative, and blood HBV undetectable). MRI revealed a 15cm x 10cm main lesion in the left hepatic lobe and multiple smaller lesions in the right hepatic lobe, all under 3 cm in diameter (Supplementary Figure S1A). Tumor thrombi involved in the right portal vein branch (PVTT) and inferior vena cava (IVCTT) were revealed by MRI with contrast enhancement, indicating the intrahepatic and extrahepatic vascular spreading of HCC (Supplementary Figure S1B). Tumor tissues from the 4 tumor sites (noted as T1-4) and corresponding adjacent normal tissues as well as tissues from two tumor thrombi were collected after surgery.

HBV genome enriched single cell sequencing (HGE-scSeq)

The fresh (with 1 hour after surgery) frozen (stored in -80°C) tumor tissue samples were thaw in water bath at room temperature and digested into cell solution by collagenase as previously described [68]. With sufficient collagenase dissociation and dilution, the cancer tissues were separated into single cells solution, cell clusters and cell debris. Suspension was filtered by injecting into the membrane filter (pore size = 20μm) to filter out the massive cell clusters. To avoid contamination of cell debris, suspension was re-suspended and centrifuged in Phosphate Buffered Saline (PBS) for 5 times. After filtration, cell suspension was added into a PBS droplet containing 0.5% BSA. Single cell isolation was performed using micro pipette as previously described [68] under microscope and cells with intact cell membrane were randomly selected for single cell sequencing.

For each cell, WGA was performed with MDA using REPLI-g Mini Kit (QIAGEN, Inc.) according to the instructions of the manufacturer as previously described [68]. HIVID [14] procedure was then used to enrich for sequences containing HBV genome sequence. DNA library from amplified single cell genome was hybridized with the biotinylated HBV probe to enrich DNA fragments containing HBV DNA sequences. Then, the enriched libraries were quantified and subjected to 101 cycles paired-end index sequencing in Illumina HiSeq 2000 sequencer according to manufacturer's instructions (Illumina Inc., San Diego, CA). The raw data are deposited at NIH SRA (BioProject: PRJNA553308) (reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA553308?reviewer=rmut731nv0i3cor179v2vr0g47).

Mapping HGE-scSeq reads

On average, 17.39M (17,393,993) reads were generated for each cell. Low quality reads were filtered out according the following criteria. If any single read in a read pair had more than half base of quality less than 5, the corresponding read pair was filtered (Supplementary Figure S11A). If a read pair was contaminated by adaptor sequences, it was filtered. If two read pairs were the same, only one copy was kept in further analysis. After quality filtration 5.49M (5,494,183) reads were kept in further analyses. Among them 77.13% and 0.24% were aligned to human and HBV genome, respectively on average. With paired-end assembly and re-mapping, reads supporting virus integration were identified. The number of reads supporting HBV virus integration in each cell was in a range of 0 to 53,290. The average percentage of human genome covered by sequencing reads was 3.13% with average depth of coverage 3.14. The detailed information of reads distribution can be found in Supplementary Table S2 and Supplementary Table S11.

Bulk tissue HBV enriched DNA sequencing

Corresponding adjacent non-neoplastic liver tissues for the four independent tumor sites, noted as N1-4, were collected for bulk tissue HBV enriched DNA sequencing. For the 4 adjacent normal tissues, the HIVID procedure was directly applied to the extracted DNA without the WGA step, followed by the same 101 cycles paired-end index sequencing. On average, 45.96M reads were generated for each tissue sample. After quality filtration 12.13M reads were kept for further analyses. Among them 78.48 % reads were mapped to the human genome, and 0.013% reads were mapped to HBV genome. On average only 50 reads supporting HBV integration were detected for each control tissue sample. The average percentage of human genome covered by reads was 6.9% with average depth of coverage 1.272. The detailed information of reads distribution can be found in Supplementary Table S2 and Supplementary Table S11.

Quality check of whole genome sequencing reads

Our previously described pipeline [13] was used to process the whole genome sequencing data. In brief, prinseq-lite [69] was used to filter the reads that were exactly the same or of mean reads quality lower than 20 and more than 10% Ns. The remaining reads were mapped to human genome with Bowtie2 (-D 15 -R 2 -N 0 -L 22 -i S,1,1.15) [70]. Duplicated reads after alignment were filtered using Picard.

Quality check of chimera reads in HGE-scSeq data

Limited amount of input material from a single cell for WGA causes a lot of technical errors, including low physical coverage, non-uniform coverage, allelic dropout events, false positive and false negative errors due to insufficient coverage [18-21, 23, 26, 71-73]. Chimera reads, which can be partially mapped to different parts of the genome that are not physically linked [26], are common artifacts of single cell WGA [26], which can interfere our ability to identify HBV-human genome chimera sequences. The frequency of chimera reads (identified following the standard protocol [26, 74]) was 0.025 % which is much lower than 6.19% reported by Tu *et al.* [74] and 2-3% by Huang *et al.* [26] for MDA. Also the number of chimera reads from both inter chromosome and intra chromosome were independent from

number of HBV-Human soft clipped reads, HBV reads and Human reads (Supplementary Figure S12). Numbers of inter chromosome and intra chromosome chimera reads were correlated, and they both correlated with the length of chromosome, consistent with random nature of human chimera reads (Supplementary Figure S13). The numbers chimera reads did not correlate with the number of reads on HBV or soft clipped reads, nor correlate with number of reads on human genome, suggesting that chimera reads had no impact on the HBV integration detection and copy number variation detection. The number of soft clipped reads and the number of HBV reads were strongly correlated (Supplementary Figure S12F), which suggests that we need to correct the number of HBV reads when identifying HBV integration.

Quality check of reads mapped to human genome

Although the regions containing HBV sequences were enriched in the sequencing library preparation step, around 77.13% of the sequencing reads on average were mapped to human genome. The reads distributions were summarized in Supplementary Table S2. The average percentage of human genome covered by sequencing reads was 3.13% with average depth of coverage 3.14. For the adjacent normal tissues, the average percentage of human genome covered by reads was 6.9% with average depth of coverage 1.272. The coverage and width of bulk tissue data were comparable with the ones in the single cell data.

To check whether loci covered by sequencing reads were randomly distributed across human genome, for each locus, we counted the number of cells with reads covering the locus. If the reads mapped to human genome were randomly distributed, then the number of cells with reads at each locus is expected to follow a poison distribution. The largest number of cells with reads covering a locus was 209, the mean was 8.2277, and the fraction of loci not covered by reads in any cell was 11.22% (Supplementary Figure S14). The observed distribution was tested against a Poisson distribution with chi square test on range of [1, k] (k indicates a locus covered by reads in k cells, which corresponds to the kth bar in Supplementary Figure S14) with k from 15 to 37 (Supplementary Table S12). The distribution

matched with a Poisson distribution until k=28, which corresponding to 87.97% of human genome. When k>=29, the distribution was not a Poisson distribution anymore. Thus, the mapped reads on the majority of human genome following a Poisson distribution, except the region consistently missed by all cells and 0.81% of human genome covered by reads from a number of cells significantly more than expected by chance. These observations suggested that CNV profile at single cell level could be accurately estimated with appropriate normalization method.

Comparing human genome regions with and without HGE-scSeq reads

To infer CNVs from reads mapped to the human genome, these reads should be evenly distributed across the human genome and there should be no systematic difference between the regions covered with sequencing reads and the regions without reads. To investigate the property of the regions with and without read sequence coverage, we first constructed a Fisher machine prediction model [48] to distinguish HBV and human genome sequences by randomly sampling 10,000 sequences of 100bp length from HBV and human genomes. Then, we applied to Fisher machine to test whether the sequences in the human genome regions with or without HGE-scSeq reads were similar to HBV or human genome sequences. For each cell, 10,000 sequences of 100bp length were randomly samples from human genome regions with and without mapped reads, and input them to the Fisher machine. There was no difference between scores of regions with and without mapped reads (Wilcox rank sum test p-value=0.3636, Supplementary Figure S15).

Mapping to HBV virus genomes

The filtered reads were aligned to UCSC hg19 with soap2 [75] (Version: 2.20) in paired-end mode (Supplementary Figure S11B). The parameters used were "-s 85 -l 50 -v 2 –r 1 -p 6 -m 100 -x 500". If any read in a pair was not mapped to human genome, the pair was kept as a candidate for virus detection. These reads were collected and transformed from FASTQ to FASTA format. The virus detection part in VirusFinder [76, 77] was used to detect the virus. The reads not mapped to human genome were aligned

493 to a virus database which contains genomes of all known viruses (32102 in total) [78]. The reads aligned
494 to virus genome were *de novo* assembled into contigs. Then, the contigs were aligned to human genome
495 and virus database. The contigs that can be aligned to human genome were filtered out. If the percentage
496 of identity between the contig and virus' genome was less than 85% or less than 75% of the contig was
497 aligned to a reference genome, the alignment was filtered out. The alignment score of contigs was defined
498 as the multiplication of the mapped length of the contig and percentage of identity between the mapped
499 region of the contig and the virus genome. The virus substrains were ranked by the maximum alignment
500 score of contigs aligned to its genome. The top ranked virus substrain was reported as the matched virus
501 substrain in the cell (Supplementary Table S13). The top common substrains were all HBV B subtype and
502 were similar in sequences (Supplementary Table S14).
503
504 ***Detecting HBV integration sites***
505 The reads not mapped to the human genome were aligned to the detected virus genome using soap2
506 (Version: 2.20 with the following parameters "-s 85 -l 50 -v 5 –r 1 -p 6 -m 100 -x 500"). The paired-end
507 reads not mapped to the human genome and virus genome were collected and assembled to long reads
508 using flash (with parameters "-m 5 x 0.2 –p 64") [79]. The designed smaller insertion size compared to
509 the total length of a pair of reads enabled most read pairs to be assembled into one read of much longer
510 length. The assembled reads were aligned to the human genome and virus genome using bwa and bwasw
511 [80] (-a 1 –b 2 –q 5 –r 2). The soft clipped reads with at least 30bp aligned to the human genome and at
512 least 30bp aligned to virus genome were collected for identifying the integration sites. If the distance
513 between two breakpoints was less than 20bp on both the human genome and HBV genome, we defined
514 them as one breakpoint which was supported by reads combined from the two breakpoints. In order to
515 make the predicted integration events between different cells comparable, we also merge integration sites
516 within 20bp when collecting the predicted integration sites across different cells. The number of soft
517 clipped reads was tightly correlated with the number of HBV reads (Supplementary Figure S12F). We
518 normalized soft clipped read against the number of HBV reads. The Optimal threshold of soft clipped

reads for HBV integration was selected to minimize the correlation between numbers of HBV reads and detected HBV integrations. Further refinement based on Bayesian model was used to identify recurrent HBV integrations. Steps are detailed as following:

1. We collect soft clipped reads into a matrix A of $m \times n$, m is the number of cells, n is the number of candidate integration sites. $A(i, j)$ is the number of soft clipped reads for cell i on site j. Only cell i and site j was included if $\sum_{i=1}^{m} A(i, j) > 0$ and $\sum_{j=1}^{n} A(i, j) > 0$. As result, n=1108 sites and m=189 cells were included for our data.
2. We collect the number of HBV reads $H_{m\times 1}$ in each cell. H_i is the number of HBV reads in cell i.
3. We normalize the number of soft clipped reads against the number of HBV reads as $A^1_{i,j} = \frac{A_{i,j}}{(H_i \times 100)/\mathrm{L_{HBV}}} 10000$, where the denominator is the estimation of HBV load. The number of soft clipped reads is normalized with the same load of HBV, or the same scale of HBV probe enrichment.
4. We rank order $A^1[A \geq 2]$ and convert them into 1 to 100 quantiles as the probability of integration as $P^1_{i,j}$. At this step, we only consider integrations supported with at least 2 soft clipped reads in a cell.
5. Next, we search for the optimal cutoff value at which the total number of integration sites is least depended on the total number of HBV reads. For each quantile cutoff k, calculate the correlation coefficient between the number of integrations and number of HBV reads as the following:

$$r(k) = corr(\sum_{j} (P^1_{i,j} > k/100), H_i) \ .$$

At the cutoff k corresponding to lowest absolute correlation coefficient, we identified 141 cells carrying HBV integrations on 164 unique sites, which are denoted as *INT_SITE1*. Among them, 31 integrations were identified in more than one cell. There are totally 427 HBV integration events, which are denoted as *INT_EVENT1*.

6. Then, we tried to identify more recurrent integrations. We implemented a Pseudo Count Weight Adjustment (PCWA) to rescue the potential false negative integrations given supporting reads in other cells. The pseudo count weight matrix $W_{m\times n}$ is defined as: $W_{i,j} = sqrt(\sum_{i\neq j, i\leq m} A^1_{i,j})$.

   The pseudo count weight adjusted score $S_{m\times n}$ is defined as: $S_{i,j} = 1 - e^{-A^1_{i,j}(1+\alpha W_{i,j})}$. α is the tuning parameter used to adjust the contribution from pseudo count weight.

7. For a given α, a score matrix $S_{m\times n}(\alpha)$ is generated. Following step 5, we convert $S_{m\times n}(\alpha)$ into 1 to 100 quantiles as the probability of integration as $P^2_{i,j}(\alpha)$. A cutoff is needed to call integrations.

8. In order to find the optimal cutoff *k* which recovers most repeated integrations while incorporates random integrations and misses detected integrations as few as possible, we select the following optimization criteria:

$$val_\alpha(k) = \frac{N\_gain(k)}{(N\_gain\_random(k)+1)\times(N\_loss(k)+1)}$$, where

$N\_gain_\alpha(k) = \#\{P^2_{i,j}(\alpha) > k/100, j \in INT\_STIE1\} \cap \overline{INT\_EVENT1}$ is the number of new integrations above threshold at the sties included in the first round;

$N\_gain\_random_\alpha(k) = \#\{P^2_{i,j}(\alpha) > k/100, j \notin INT\_STIE1\} \cap \overline{INT\_EVENT1}$ is the number of new integrations above threshold at the sites NOT included in the first round;

561 $N\_loss_{\alpha}(k) = \#\{P_{i,j}^{2}(\alpha) < k/100\}$ I INT_EVENT1 is the number of integrations below
562 threshold for integrations included in the first round. And k is selected as the cutoff producing the
563 maximum validation score $val_{\alpha}(k)$ as $k = \arg\max_{1 \le k \le 100}\{val_{\alpha}(k)\}$

564

565 9. For a given α , the maximum validation score $\max_{1 \le k \le 100}\{val_{\alpha}(k)\}$ indicates the best performance with
566 the α . Therefore, we define the validation score for α as $VAL(\alpha) = \max_{1 \le k \le 100}\{val_{\alpha}(k)\}$.

567 10. We searched α on $[10^{-6}, 2\times10^{-6}...10^{-4}, 2\times10^{-4}, ..., 0.01, 0.02, ...1, 2...100]$. We find that the
568 best alpha arrive at 0.26 (Supplementary Figure S16A). Corresponding to the best cutoff,
569 N_gain=44, N_gain_random=4, N_loss=1 (Supplementary Figure S16B).
570 11. The 44 new recovered HBV integration events are merged to the results from step 5. Finally,
571 HBV integrations are detected on 164 unique sites and 142 cells. 39 of HBV integrations are
572 repeatedly discovered.

573

574 ***Estimating copy number variations***

575 Reads mapped to human genome were randomly distributed (Supplementary Figure S14), which enabled
576 us to estimate DNA copy numbers across human genome. Because the sequencing data was based on an
577 enriched single cell sequencing protocol [14], the existing pipelines for detecting copy number variations
578 in single cell sequencing data [17, 81, 82] are not applicable directly. If applied directly, more regions of
579 copy number aberration than the regions of normal copy number were identified, which is counter
580 intuitive. Therefore, a new pipeline was needed for analyzing the data set. Based on reads mapped to
581 human genome, we developed a pipeline for inferring copy number variations by modifying the method
582 reported by Baslan *et al* [82].

583 Kuilman *et al.* [83] shows that off-target reads from enriched sequencing can be used to obtain
584 DNA copy number profiles by removing peaks of mapped reads and compensating according to the size
585 of peaks and average local coverage. However, Kuilman *et al.*'s method [83] was not directly applicable
586 for this data set due to the sparsity of reads covered region originated from single cell whole genome
587 amplification. Baslan *et al.* [82] describes a procedure characterizing single cell copy number variation
588 based on flow sorting of single nuclei, whole genome amplification and next generation sequencing. An
589 informatics workflow of inferring CNV from the raw single cell sequencing data is outlined in
590 Supplementary Figure S11C. In addition to correction for mappability, removal of duplication, and GC
591 content normalization, we used soap [84] as the alignment software to be consistent with the alignment
592 software used in HIVID.

593 There are following steps in the pipeline for calling CNV:

594 1). The sequence of pseudo autosomal regions on chrY was changed to N. The sequence of pseudo
595 autosomal regions on chrY is exactly the same as the corresponding regions on chrX. Generate the index
596 of the reference genome for soap with 2bwt-builder.

597 2). Sequencing reads were simulated based on the reference genome. Starting from the first position of a
598 chromosome a fragment of 100 bases was extracted to generate a sequencing read. The step continued at
599 the following positions until the end of the chromosome.

600 3). The simulated sequencing reads were mapped to the modified human reference genome with soap (-s
601 85 -l 50 -v 2 -r 1 -p 6 -m 100 -x 500)

602 4) Genome positions with simulated sequencing reads mapped uniquely back to where they were
603 extracted from were defined as mappable positions. All the mappable positions were collected, and the
604 number of mappable positions on each chromosome was counted.

5) Mappable positions were grouped into 5000 bins. Number of bins allocated to each chromosome was proportional to the number of mappable positions on that chromosome. The number of mappable positions for each bin was computed by dividing the number of mappable positions on the chromosome with the number of allocated bins on that chromosome. The boundaries of bins were decided by the number of mappable positions in each bin sequentially on the chromosome. The average length of bin is 560485.9 (sd: 989.8). If more than 50% of a bin overlaps with bad bins reported in the paper [82], the bin is defined as a bad bin. These bad bins are mainly located at centromere regions of chromosomes. As a result 11 bad bins were filtered out..

6) The GC content in each bin was calculated.

7) The filtered reads were mapped to reference human genome with soap (-s 85 -l 50 -v 2 -r 1 -p 6 -m 100 -x 500)

8) The result of pair end aligned reads was converted to sam format with soap2sam.pl, then to bam format with samtools. The duplicated reads were removed with Picard.

9) The number of reads in each bin was counted and normalized by dividing the mean read count of a cell.

10) GC content was normalized using LOWESS smoothing. In brief, a regression model was constructed by regressing the read count against the GC content percentage with LOWESS regression. Then, the corrected read count is calculated by minus the input read count with the one predicted by the regression model on the corresponding GC percentage.

11) The read counts after mappability and GC content normalization were collected into a 269 by 4989 matrix M, where 269 was the number of cells and 4989 was the number of bins. The index of dispersion for each cell was calculated as the ratio between standard deviation and mean of bin's read count. As suggested in Garvin *et al.* [81], the sample with lowest index of dispersion is mostly likely to be the cell

with the most balanced ploidy. Therefore, the read count profiles of all cells were normalized against the one having lowest index of dispersion with LOWESS.

12) Similar to Gao *et al.* [85], outliers were removed with R function *winsorize*. Then, a multiple sample population segmentation algorithm with default parameter [86] was used to call the segments under the condition that these cells are related. Segments with less than 10 bins were removed and the neighbor segments were joined or separated in the middle of removed segment if they differed significantly. At the end, bins were merged into a total of 49 segments.

13) A least-square rounding method was used to get the optimum scaling factor that had the least sum of deviations from the closet integer after rounding. Integer copy number status was further classified into 3 cases of loss, normal and amplification and denoted as -1, 0, 1.

Evaluation of read count correction

Because sequences containing HBV sequence were enriched at the DNA library preparation step, we need to correct read count bias due to enrichment sequencing. For each cell, we collected the rank of read counts for bins with HBV integrations detected. Then, for all the bins with HBV integrations, we calculated the fraction of bins with rank higher than $\alpha\%$, $f(\alpha)$, where α is ranging from 1 to 100. Then the fold enrichment of top α % ranked bins among bins with HBV integration is defined as $100 \times f(\alpha) / \alpha$. In an ideal case where there is no bias from enrichment sequencing, the fold enrichment should be around 1.

Two metrics to characterize the overall quality of binned reads were introduced. Assuming there are *n* bins, the number of reads in each bin is *count*[*k*] (*k*=1..*n*). The average number of reads across bins is *C*. Garvin et. al [81] introduce median absolution deviation (MAD) to quantify the uniformity of bin's read count. For each cell, MAD is defined as $MAD = median\{|\frac{count[k]}{C} - \frac{count[k-1]}{C}|, k = 2...n\}$. MAD is expected to reflect the bin count dispersion due to technical noise. Another metric named as

MAPD[87] is defined as $MPAD = median\{|\log_2(\frac{count[k]}{C}) - \log_2(\frac{count[k-1]}{C})|, k = 2...n\}$, which is originally used as a QC metric for microarray data. Cai et al. [27] propose to utilize MAPD [87] to measure the quality of read counts in bins. MAPD is shown to be more robust to identify true CNVs [27]. An optimal threshold of 0.45 is suggested by Cai et al. [27], which is also used in other studies [19].

Supplementary Figure S17 A&B are the bar plots of fold enrichment of top α% ranked bins according to reads count across all the bins with HBV integration. The bins with HBV integration were enriched for top ranked bins according to raw read count (Supplementary Figure S17A). The highest fold enrichment was for top 1% ranked bins, which was as high as 19.48. The reads counts after normalization are showed in Supplementary Figure S17B. The highest fold enrichment was 2.03, which means the bias of reads count from enrichment sequencing had been successfully corrected. The Supplementary Figure S17C &D show the box plot of MAD and MAPD. The Supplementary Figure S17C shows MAPD and MAD for the reads count after mappabiliy and GC content correction, and Supplementary Figure S17D shows the read count after further corrected by the reads count from the cell with least dispersion. Both MAD and MAPD were significantly decreased after the corrections. At the threshold of 0.45 suggested by Cai *et al.* [27] for filtering the cells with low reads quality, only 3 cells passed the threshold before the least dispersion sample correction, while all the cells passed the threshold after the least dispersion sample correction.

Evaluating the CNV pipeline with reads from normal control

As our CNV pipeline was modified from a CNV pipeline for single cell sequencing data, which takes full consideration of correcting for bias incorporated from WGA [88]. Whether our modified pipeline can handle bulk tissue enrichment sequencing data needs to be evaluated.

As shown above, the reads mapped to human genome for normal control tissue resulted in higher average coverage and width comparing to ones for tumor single cells. However the improvement on coverage and width was not large. The cases for normal control tissue and tumor single cell were

674 comparable (Supplementary Figure S18). Thus, the sequencing data for normal tissue samples can be
675 used for evaluating the performance of our CNV pipeline in correcting the bias generated from enriched
676 sequencing.

677 The dispersion of the binned reads counts for the four adjacent normal liver tissue samples after
678 mappability and GC content correction was lower than the smallest corresponding dispersion in tumor
679 single cells (Supplementary Figure S19A). Therefore, mappability correction and GC content correction
680 for the normal control tissue data were necessary. Most of the regions across the 4 adjacent normal tissue
681 samples were of normal copy number (Supplementary Figure S19B). Therefore, our CNV pipeline for
682 correcting the potential bias due to enriched sequencing step was validated.

683

684 ***Association between clone evolution and HBV integrations***

685 Parsimony method is mostly recommended for constructing phylogenetic trees from single cell CNV
686 profiles [85, 89]. The distance based tree building method generally assumes that evolution drives by
687 mutations independently accumulated one at a time. CNVs estimated in our study contain multiple
688 alterations. Therefore, in this study, we used a parsimony method [85] to build phylogenetic trees based
689 on CNVs at the 49 identified CNV segments.

690 We identified 4 putative clones according to copy number profiles (Figure 3A). Meanwhile, there
691 were two categories for cells based on HBV integration, cells with only hot spot integrations and cells
692 with extra rare integrations. There was a clear trend that the ratio of cells carrying rare integrations
693 decreased when number of regions with DNA copy number amplification increased (Figure 3B).

694 A phylogenetic tree based test of association method was used to decode the association between
695 specific CNV and ratio for cells with rare HBV integrations. First, inner node in the phylogenetic tree that
696 separate parent and child clones which are identified by hierarchical clustering was detected. Second,
697 copy number variations at the genomic region corresponding to the detected inner node for the cells
698 belonging to parent and child clones were collected. Third, a contingency table was built based on

numbers of cells with extra rare HBV integrations or only HBV integrations at hot spots, copy number amplification, copy number normal. Fourthly, Fisher's exact test p-value corrected by multiple testing was used to assess the association. Last, functional enrichment analysis by DAVID [90] was used to annotate genes in the CNV bins that were significantly associated with rare HBV integration events.

For example, CNVs on Chr11 which differentiated Clone 1 and Clones 2-4 were identified based on the phylogenetic tree shown in Figure 4A. Next, the cells from related clones (Clone 1 and Clones 2-4) were compared at bins within the genomic signaling region. For each bin in the region, a contingency table was built to test the association between hot spot integration vs. rare integration and copy number amplification vs. normal. Bins of FDR<0.05 were collected and annotated for potential enriched functions. Last, genes located on the significant regions were used as input for functional enrichment analysis.

List of abbreviations:

HBV: Hepatitis B Virus; HCC: Hepatocellular carcinoma; dslDNA: double-stranded linear DNA; WGS: whole genome sequencing; high-throughput Viral Integration Detection: HIVID; HPV: Human papillomavirus; CNVs: copy number variations; MALBAC: multiple annealing and looping-based amplification cycles; MDA: multiple displacement amplification; WGA: Whole Genome Amplification; HGE-scSeq: HBV genome-enriched Single cell sequencing;

Declarations:

Ethics approval and consent to participate

The study of tumor cell heterogeneity was approved by the Institutional Review Board of Tongji Hospital, Tongji Medical College of HUST, in Hubei province, China.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the NIH SRA (BioProject: PRJNA553308) (reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA553308?reviewer=rmut731nv0i3cor179v2vr0g47)

Funding

This work was partially supported by National Institutes of Health [grant numbers: U011HG008451 and U19 AI118610].

Conflicts of Interest Statement

LW, SL, YH were employed by BGI-Shenzhen, SY, QC, JZ were employee of Sema4. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contribution

QC and JZ designed the experiment, XPC and QC lead clinical design, YC, XPC and QC contributed biological data, , LW, SL, and YH contributed genomic data generation, WW, YZ, SY, QC, and JZ contributed data analyses, WW and JZ wrote the manuscript. All authors reviewed and commented the manuscript.

Acknowledgements

Not applicable.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D: **Global cancer statistics**. *CA: a cancer journal for clinicians* 2011, **61**(2):69-90.
2. El-Serag HB: **Epidemiology of viral hepatitis and hepatocellular carcinoma**. *Gastroenterology* 2012, **142**(6):1264-1273 e1261.
3. Tu T, Budzinska MA, Shackel NA, Urban S: **HBV DNA Integration: Molecular Mechanisms and Clinical Implications**. *Viruses* 2017, **9**(4).
4. Tu T, Budzinska MA, Vondran FWR, Shackel NA, Urban S: **Hepatitis B Virus DNA Integration Occurs Early in the Viral Life Cycle in an In Vitro Infection Model via Sodium Taurocholate Cotransporting Polypeptide-Dependent Uptake of Enveloped Virus Particles**. *J Virol* 2018, **92**(11).
5. Ishikawa T: **Clinical features of hepatitis B virus-related hepatocellular carcinoma**. *World journal of gastroenterology* 2010, **16**(20):2463-2467.
6. Paterlini-Brechot P, Saigo K, Murakami Y, Chami M, Gozuacik D, Mugnier C, Lagorce D, Brechot C: **Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene**. *Oncogene* 2003, **22**(25):3911-3916.
7. Gozuacik D, Murakami Y, Saigo K, Chami M, Mugnier C, Lagorce D, Okanoue T, Urashima T, Brechot C, Paterlini-Brechot P: **Identification of human cancer-related genes by naturally occurring Hepatitis B Virus DNA tagging**. *Oncogene* 2001, **20**(43):6233-6240.
8. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C *et al*: **Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma**. *Nat Genet* 2012, **44**(7):765-769.
9. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S: **The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients**. *Genome research* 2012, **22**(4):593-601.
10. Miao R, Luo H, Zhou H, Li G, Bu D, Yang X, Zhao X, Zhang H, Liu S, Zhong Y: **Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis**. *Journal of hepatology* 2014, **61**(4):840-849.
11. Jhunjhunwala S, Jiang Z, Stawiski EW, Gnad F, Liu J, Mayba O, Du P, Diao J, Johnson S, Wong KF *et al*: **Diverse modes of genomic alteration in hepatocellular carcinoma**. *Genome biology* 2014, **15**(8):436.
12. Lu LC, Hsu CH, Hsu C, Cheng AL: **Tumor Heterogeneity in Hepatocellular Carcinoma: Facing the Challenges**. *Liver cancer* 2016, **5**(2):128-138.
13. Yoo S, Wang W, Wang Q, Fiel MI, Lee E, Hiotis SP, Zhu J: **A pilot systematic genomic comparison of recurrence risks of hepatitis B virus-associated hepatocellular carcinoma with low- and high-degree liver fibrosis**. *BMC Med* 2017, **15**(1):214.
14. Li W, Zeng X, Lee NP, Liu X, Chen S, Guo B, Yi S, Zhuang X, Chen F, Wang G *et al*: **HIVID: an efficient method to detect HBV integration using low coverage sequencing**. *Genomics* 2013, **102**(4):338-344.
15. Zhao LH, Liu X, Yan HX, Li WY, Zeng X, Yang Y, Zhao J, Liu SP, Zhuang XH, Lin C *et al*: **Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma**. *Nature communications* 2016, **7**:12992.
16. Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L *et al*: **Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a**

potential microhomology-mediated integration mechanism. *Nature genetics* 2015, **47**(2):158-163.

17. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D *et al*: **Tumour evolution inferred by single-cell sequencing**. *Nature* 2011, **472**(7341):90-94.
18. Zong C, Lu S, Chapman AR, Xie XS: **Genome-wide detection of single-nucleotide and copy-number variations of a single human cell**. *Science* 2012, **338**(6114):1622-1626.
19. Ning L, Li Z, Wang G, Hu W, Hou Q, Tong Y, Zhang M, Chen Y, Qin L, Chen X *et al*: **Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons**. *Scientific reports* 2015, **5**:11415.
20. Navin NE: **The first five years of single-cell cancer genomics and beyond**. *Genome research* 2015, **25**(10):1499-1507.
21. Navin NE: **Cancer genomics: one cell at a time**. *Genome Biol* 2014, **15**(8):452.
22. Macaulay IC, Voet T: **Single cell genomics: advances and future perspectives**. *PLoS genetics* 2014, **10**(1):e1004126.
23. Wang Y, Navin NE: **Advances and applications of single-cell sequencing technologies**. *Molecular cell* 2015, **58**(4):598-609.
24. Hou Y, Wu K, Shi X, Li F, Song L, Wu H, Dean M, Li G, Tsang S, Jiang R: **Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing**. *GigaScience* 2015, **4**(1):1-16.
25. de Bourcy CF, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR: **A quantitative comparison of single-cell whole genome amplification methods**. *Plos One* 2014, **9**(8):e105585.
26. Huang L, Ma F, Chapman A, Lu S, Xie XS: **Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications**. *Annu Rev Genomics Hum Genet* 2015, **16**:79-102.
27. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA: **Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain**. *Cell Rep* 2014, **8**(5):1280-1289.
28. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H *et al*: **Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor**. *Cell* 2012, **148**(5):886-895.
29. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, Zong C, Bai F, Wang J, Xie XS: **Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients**. *Cancer Research* 2014, **74**(19 Supplement):3577-3577.
30. Li Y, Xu X, Song L, Hou Y, Li Z, Tsang S, Li F, Im KM, Wu K, Wu H: **Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer**. *GigaScience* 2012, **1**(1):1-14.
31. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D: **Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm**. *Cell* 2012, **148**(5):873-885.
32. Wang J, Fan HC, Behr B, Quake SR: **Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm**. *Cell* 2012, **150**(2):402-412.
33. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H: **Clonal evolution in breast cancer revealed by single nucleus genome sequencing**. *Nature* 2014, **512**(7513):155-160.
34. Leung ML, Wang Y, Waters J, Navin NE: **SNES: single nucleus exome sequencing**. *Genome Biol* 2015, **16**:55.

35. Summers J, Jilbert AR, Yang W, Aldrich CE, Saputelli J, Litwin S, Toll E, Mason WS: **Hepatocyte turnover during resolution of a transient hepadnaviral infection**. *Proc Natl Acad Sci U S A* 2003, **100**(20):11652-11659.
36. Mason WS, Gill US, Litwin S, Zhou Y, Peri S, Pop O, Hong ML, Naik S, Quaglia A, Bertoletti A *et al*: **HBV DNA Integration and Clonal Hepatocyte Expansion in Chronic Hepatitis B Patients Considered Immune Tolerant**. *Gastroenterology* 2016, **151**(5):986-998 e984.
37. Bowcock AM, Pinto MR, Bey E, Kuyl JM, Dusheiko GM, Bernstein R: **The PLC/PRF/5 human hepatoma cell line. II. Chromosomal assignment of hepatitis B virus integration sites**. *Cancer Genet Cytogenet* 1985, **18**(1):19-26.
38. Tay N, Chan SH, Ren EC: **Detection of integrated hepatitis B virus DNA in hepatocellular carcinoma cell lines by nonradioactive in situ hybridization**. *Journal of medical virology* 1990, **30**(4):266-271.
39. Duan M, Hao J, Cui S, Worthley DL, Zhang S, Wang Z, Shi J, Liu L, Wang X, Ke A *et al*: **Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing**. *Cell research* 2018, **28**(3):359-373.
40. Podlaha O, Wu G, Downie B, Ramamurthy R, Gaggar A, Subramanian M, Ye Z, Jiang Z: **Genomic modeling of hepatitis B virus integration frequency in the human genome**. *PLoS One* 2019, **14**(7):e0220376.
41. Budzinska MA, Shackel NA, Urban S, Tu T: **Cellular Genomic Sites of Hepatitis B Virus DNA Integration**. *Genes (Basel)* 2018, **9**(7).
42. Chen XP, Long X, Jia WL, Wu HJ, Zhao J, Liang HF, Laurence A, Zhu J, Dong D, Chen Y *et al*: **Viral integration drives multifocal HCC during the occult HBV infection**. *J Exp Clin Cancer Res* 2019, **38**(1):261.
43. Wong N, Lai P, Lee SW, Fan S, Pang E, Liew CT, Sheng Z, Lau JW, Johnson PJ: **Assessment of genetic changes in hepatocellular carcinoma by comparative genomic hybridization analysis: relationship to disease stage, tumor size, and cirrhosis**. *Am J Pathol* 1999, **154**(1):37-43.
44. Tian J, Tang ZY, Ye SL, Liu YK, Lin ZY, Chen J, Xue Q: **New human hepatocellular carcinoma (HCC) cell line with highly metastatic potential (MHCC97) and its expressions of the factors associated with metastasis**. *British journal of cancer* 1999, **81**(5):814-821.
45. Li Y, Tang ZY, Ye SL, Liu YK, Chen J, Xue Q, Chen J, Gao DM, Bao WH: **Establishment of cell clones with different metastatic potential from the metastatic hepatocellular carcinoma cell line MHCC97**. *World J Gastroenterol* 2001, **7**(5):630-636.
46. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data**. *Biostatistics* 2004, **5**(4):557-572.
47. Wang K, Lim HY, Shi S, Lee J, Deng S, Xie T, Zhu Z, Wang Y, Pocalyko D, Yang WJ *et al*: **Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma**. *Hepatology* 2013, **58**(2):706-717.
48. Zhang Y, Wang X, Kang L: **A k-mer scheme to predict piRNAs and characterize locust piRNAs**. *Bioinformatics* 2011, **27**(6):771-776.
49. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y *et al*: **Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas**. *Cell research* 2016, **26**(3):304-319.
50. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J *et al*: **A comprehensive transcriptional portrait of human cancer cell lines**. *Nature biotechnology* 2015, **33**(3):306-312.
51. Zhao LH, Liu X, Yan HX, Li WY, Zeng X, Yang Y, Zhao J, Liu SP, Zhuang XH, Lin C *et al*: **Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma**. *Nature communications* 2016, **7**:12992.

52. Zhao LH, Liu X, Yan HX, Li WY, Zeng X, Yang Y, Zhao J, Liu SP, Zhuang XH, Lin C *et al*: **Erratum: Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma**. *Nature communications* 2016, **7**:13591.
53. Debacker K, Kooy RF: **Fragile sites and human disease**. *Human molecular genetics* 2007, **16 Spec No. 2**:R150-158.
54. Wu CJ, Chen LC, Huang WC, Chuang CL, Kuo ML: **Alleviation of lung inflammatory responses by adeno-associated virus 2/9 vector carrying CC10 in OVA-sensitized mice**. *Human gene therapy* 2013, **24**(1):48-57.
55. Choi MS, Ray R, Zhang Z, Mukherjee AB: **IFN-gamma stimulates the expression of a novel secretoglobin that regulates chemotactic cell migration and invasion**. *J Immunol* 2004, **172**(7):4245-4252.
56. Mohamadkhani A, Sayemiri K, Ghanbari R, Elahi E, Poustchi H, Montazeri G: **The inverse association of serum HBV DNA level with HDL and adiponectin in chronic hepatitis B infection**. *Virology journal* 2010, **7**:228.
57. Sharma D, Kanneganti TD: **The cell biology of inflammasomes: Mechanisms of inflammasome activation and regulation**. *J Cell Biol* 2016, **213**(6):617-629.
58. Kolb R, Liu GH, Janowski AM, Sutterwala FS, Zhang W: **Inflammasomes in cancer: a double-edged sword**. *Protein Cell* 2014, **5**(1):12-20.
59. Niemi K, Teirila L, Lappalainen J, Rajamaki K, Baumann MH, Oorni K, Wolff H, Kovanen PT, Matikainen S, Eklund KK: **Serum amyloid A activates the NLRP3 inflammasome via P2X7 receptor and a cathepsin B-sensitive pathway**. *J Immunol* 2011, **186**(11):6119-6128.
60. Smith HW, Marshall CJ: **Regulation of cell signalling by uPAR**. *Nat Rev Mol Cell Biol* 2010, **11**(1):23-36.
61. Boonstra MC, Verspaget HW, Ganesh S, Kubben FJ, Vahrmeijer AL, van de Velde CJ, Kuppen PJ, Quax PH, Sier CF: **Clinical applications of the urokinase receptor (uPAR) for cancer patients**. *Curr Pharm Des* 2011, **17**(19):1890-1910.
62. Berx G, van Roy F: **Involvement of members of the cadherin superfamily in cancer**. *Cold Spring Harbor perspectives in biology* 2009, **1**(6):a003129.
63. Marra F, Tacke F: **Roles for chemokines in liver disease**. *Gastroenterology* 2014, **147**(3):577-594 e571.
64. Cui X, Li Z, Gao J, Gao PJ, Ni YB, Zhu JY: **Elevated CXCL1 increases hepatocellular carcinoma aggressiveness and is inhibited by miRNA-200a**. *Oncotarget* 2016, **7**(40):65052-65066.
65. Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, Suarez CJ, Berghoff AS, Cremolini C, Falcone A *et al*: **Quantitative evidence for early metastatic seeding in colorectal cancer**. *Nature genetics* 2019, **51**(7):1113-1122.
66. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics**. *Genetics* 2002, **162**(4):2025-2035.
67. Lauer S, Avecilla G, Spealman P, Sethia G, Brandt N, Levy SF, Gresham D: **Single-cell copy number variant detection reveals the dynamics and diversity of adaptation**. *PLoS biology* 2018, **16**(12):e3000069.
68. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D *et al*: **Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm**. *Cell* 2012, **148**(5):873-885.
69. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011, **27**(6):863-864.
70. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.

71. Zhang CZ, Adalsteinsson VA, Francis J, Cornils H, Jung J, Maire C, Ligon KL, Meyerson M, Love JC: **Calibrating genomic and allelic coverage bias in single-cell sequencing**. *Nat Commun* 2015, **6**:6822.
72. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H *et al*: **Clonal evolution in breast cancer revealed by single nucleus genome sequencing**. *Nature* 2014, **512**(7513):155-160.
73. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, Zong C, Bai H, Chapman AR, Zhao J *et al*: **Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(52):21083-21088.
74. Tu J, Guo J, Li J, Gao S, Yao B, Lu Z: **Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis**. *Plos One* 2015, **10**(10):e0139857.
75. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.
76. Wang Q, Jia P, Zhao Z: **VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data**. *PLoS One* 2013, **8**(5):e64465.
77. Wang Q, Jia P, Zhao Z: **VERSE: a novel approach to detect virus integration in host genomes through reference genome customization**. *Genome Med* 2015, **7**(1):2.
78. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA: **Rapid identification of non-human sequences in high-throughput sequencing datasets**. *Bioinformatics* 2012, **28**(8):1174-1175.
79. Magoc T, Salzberg SL: **FLASH: fast length adjustment of short reads to improve genome assemblies**. *Bioinformatics* 2011, **27**(21):2957-2963.
80. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
81. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC: **Interactive analysis and assessment of single-cell copy-number variations**. *Nature methods* 2015, **12**(11):1058-1060.
82. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B *et al*: **Genome-wide copy number analysis of single cells**. *Nat Protoc* 2012, **7**(6):1024-1041.
83. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, Nevedomskaya E, Xu G, de Ruiter J, Lolkema MP *et al*: **CopywriteR: DNA copy number detection from off-target sequence data**. *Genome biology* 2015, **16**:49.
84. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.
85. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai PC, Casasent A, Waters J, Zhang H *et al*: **Punctuated copy number evolution and clonal stasis in triple-negative breast cancer**. *Nature genetics* 2016, **48**(10):1119-1130.
86. Nilsen G, Liestol K, Van Loo P, Moen Vollan HK, Eide MB, Rueda OM, Chin SF, Russell R, Baumbusch LO, Caldas C *et al*: **Copynumber: Efficient algorithms for single- and multi-track copy number segmentation**. *BMC genomics* 2012, **13**:591.
87. Affymetrix: **Median of the absolute values of all pairwise differences and quality control on Affymetrix genome-wide human SNP array 6.0.** In.; 2008.
88. Gawad C, Koh W, Quake SR: **Single-cell genome sequencing: current state of the science**. *Nature reviews Genetics* 2016, **17**(3):175-188.
89. Schwartz R, Schaffer AA: **The evolution of tumour phylogenetics: principles and practice**. *Nature reviews Genetics* 2017, **18**(4):213-229.

90. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.

Figure Legends

Figure 1 Overview of the study. 269 cells from 4 tumor tissues and 2 thrombi tissues were extracted. HBV genome sequence enrichment was performed after whole genome amplification on the single cell DNA genome. Pair-end sequencing was used. A pipeline was developed for HBV integration identification and CNV inference. Tumor clones were inferred based CNV profile. Association between HBV integration and CNV was assessed based on clone inference and phylogenetic tree. Key CNVs differentiate two clones were identified with phylogenetic tree. Statistical test was performed on the key genetic regions while considering only cells belonging to related clones.

Figure 2 HBV integration heterogeneity and mechanisms of HBV integration. A) Fractions of cells in each tissue with or without HBV sequences detected. B) Circos map of integration; each circle indicates integrations identified in a tumor tissue. C) HBV integration distribution across the human genome. Each row represents the integration profile of a cell. The cells are labeled by its tissue source. The columns are loci with HBV integrations along chromosomes. The cells were clustered by hierarchical clustering. D) An example of Microhomolog between sequences of the human genome and HBV genome at an HBV integration hot spot site Chr1 34,307,059. There are two 4bp homologs between human genome and HBV genome (AGAG and TGAA) with 1bp mismatch in the middle. E) Microhomology enrichment. Numbers of HBV integrations carrying different length of homology sequences between human genome and HBV genome near the HBV integration sites were collected (blue). The observed numbers were significantly different from the numbers based on random simulations (red). F) Fragile region enrichment. Both common and rare fragile regions on the human genome were enriched for HBV integrations.

Figure 3. **Copy number variation heterogeneity at single cell level.** A) CNV profiles. Each row corresponds to a cell. The cells are labeled with regard to source tissue, clone annotation, HBV integration category and HBV sequence detection result. Each column corresponds to a bin. The bins are ordered by their chromosome locations (chromosome 1 to 22). Cells can be categorized into 4 groups corresponding to 4 clones. White means normal copy number, blue indicates copy number loss, red indicates copy number amplification. B) Composition of cells with no HBV detected, cells with HBV sequence detected but no integration, cells carrying rare integration, and cells carrying only hot spot integration only in each clone. The frequency of cells carrying rare HBV integrations is highest for clone 1 and lowest for clone 4. The frequencies for clones 2 and 3 were comparable, and both were lower than the one for clone 1.

Figure 4. **Clonal relationship of cells from different tumor sites.** A) A phylogenetic tree built based on single cell CNV profiles. Each node corresponds to a cell. The cells are colored according clone annotation. Splitting nodes are marked as squared nodes. The scale of splitting node correlates to the number of its decedent nodes. B) Clone composition of each tumor tissue**.** Pie plots for each bin on the fractions of 4 clones. Each tumor tissue had one major clone and 3 minor clones. There was no single major clone in the thrombus tissues, but clones 3 and 4 together accounted for more than 50% of cells in the tumor thrombi, suggesting the two clones were more invasive.

Figure 5. **Simulation of clonal evolution with only CNVs.** A) The scheme of birth-death clonal evolution model. Cells accumulated CNVs during cell growth. Each additional CNV increased cell's probability to divide over to die. B) Cell populations/tumors were simulated with different combinations of mutation rates (MRs) and selection coefficients (SCs). The posterior probability of each parameter combination was calculated.

Figure 6. **Simulation of clonal evolution with both CNVs and HBV integrations.** A) The scheme of birth-death clonal evolution model with HBV integration. At the tumor size of 10^5 cells, cells were

infected with HBV and HBV integration events occurred. Simulations were generated with the selection coefficient of the hot spot integrations SC_{HBV} in a wide range. B) The frequency of cells with HBV integrations in the simulated cell populations. The red line is the observed frequency of cells with HBV integrations in the patient data (142/269) and the blue line marks the initial frequency of HBV integration (2%). C) The frequency of cells with the hot spot HBV integrations in the simulated cell populations. The red line is the observed frequency of cells with the hot spot HBV integrations in the patient data (139/269) and the blue line marks the initial frequency of the hot spot HBV integrations (0.02%). D) The ratio of cells with the hot spot HBV integrations versus cells with HBV integrations. The red line is the observed ratio in the patient data (139/142) and the blue line marks the initial ratio (1%).

Table Legends

Table 1. Functional enrichment of genes in the CNV blocks that were significantly different between clone1 and clones 2-4. A total of 370 genes were in the regions. DAVID [90] was used to test functional enrichment.

Table2. Functional enrichment of genes in the CNV blocks that were significantly different between clone 2 and clones 3 and 4. A total of 48 genes were in the regions. DAVID [90] was used to test functional enrichment.

Table 3. GO enrichment of genes in the CNV bins where cells of clone 2 had consistently lower CNVs than clones 1, 3, and 4 cells.

Figures

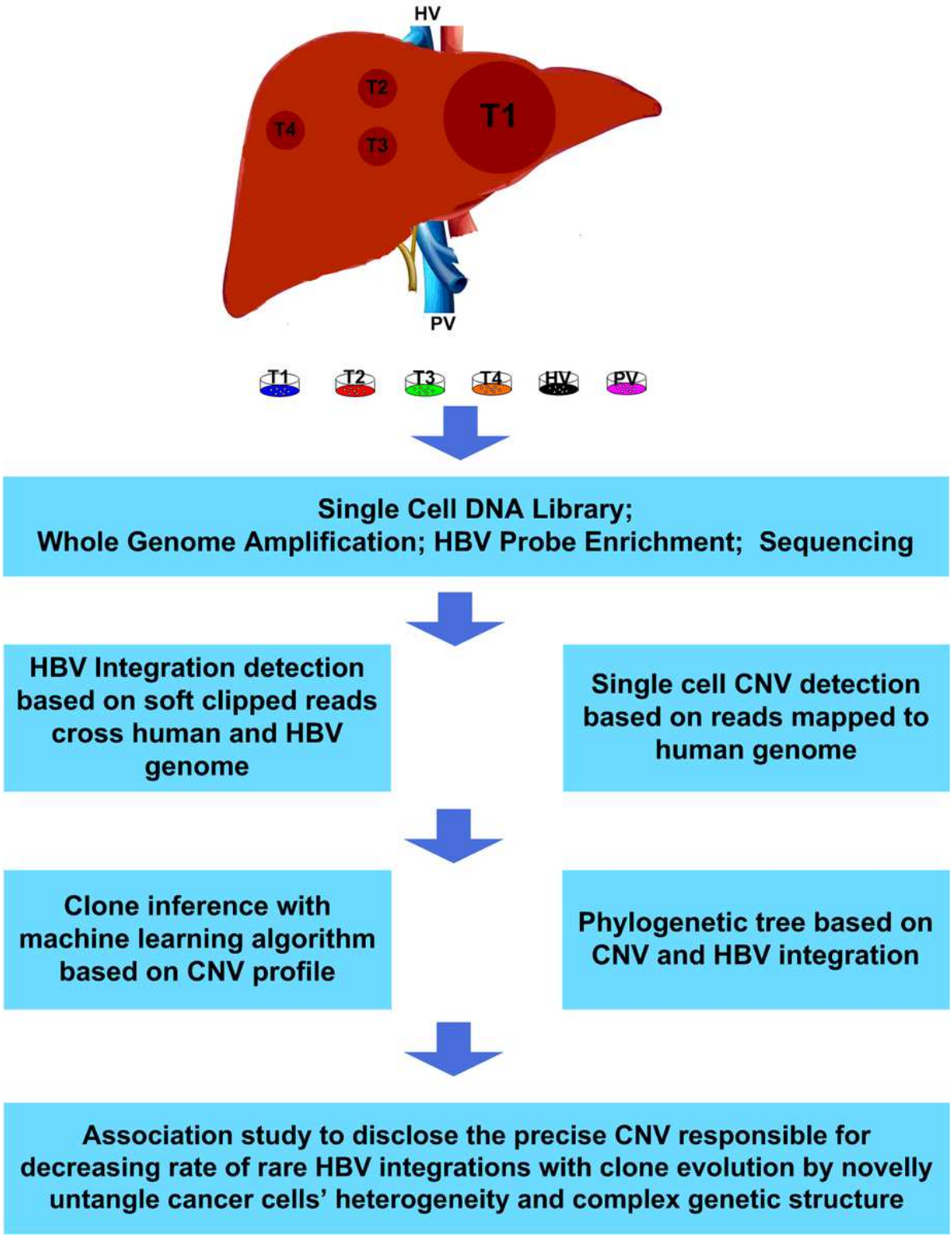

Figure 1

Overview of the study. 269 cells from 4 tumor tissues and 2 thrombi tissues were extracted. HBV genome sequence enrichment was performed after whole genome amplification on the single cell DNA genome. Pair-end sequencing was used. A pipeline was developed for HBV integration identification and CNV

inference. Tumor clones were inferred based CNV profile. Association between HBV integration and CNV was assessed based on clone inference and phylogenetic tree. Key CNVs differentiate two clones were identified with phylogenetic tree. Statistical test was performed on the key genetic regions while considering only cells belonging to related clones.

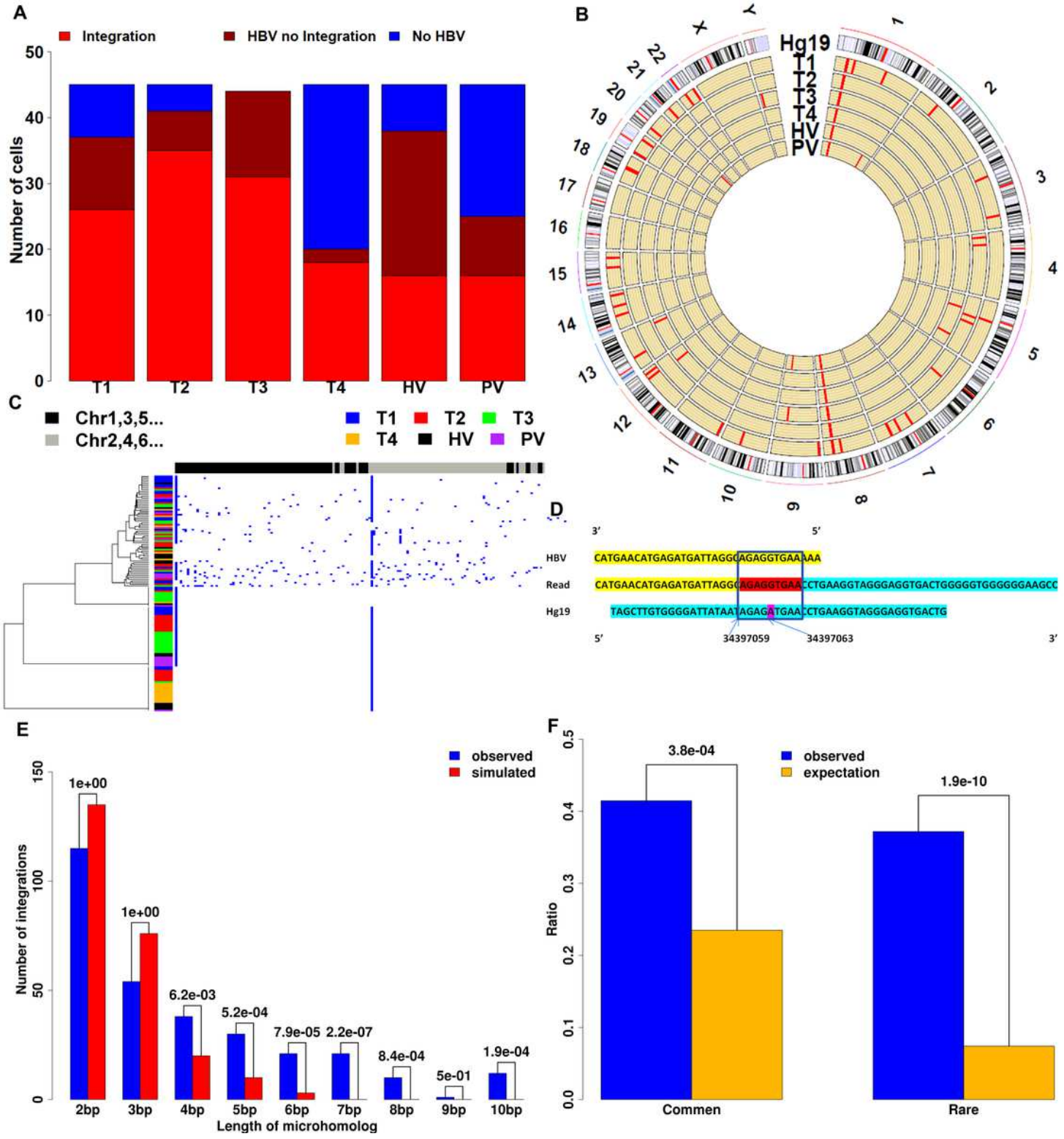

Figure 2

HBV integration heterogeneity and mechanisms of HBV integration. A) Fractions of cells in each tissue with or without HBV sequences detected. B) Circos map of integration; each circle indicates integrations identified in a tumor tissue. C) HBV integration distribution across the human genome. Each row represents the integration profile of a cell. The cells are labeled by its tissue source. The columns are loci with HBV integrations along chromosomes. The cells were clustered by hierarchical clustering. D) An example of Microhomolog between sequences of the human genome and HBV genome at an HBV integration hot spot site Chr1 34,307,059. There are two 4bp homologs between human genome and HBV genome (AGAG and TGAA) with 1bp mismatch in the middle. E) Microhomology enrichment. Numbers of HBV integrations carrying different length of homology sequences between human genome and HBV genome near the HBV integration sites were collected (blue). The observed numbers were significantly different from the numbers based on random simulations (red). F) Fragile region enrichment. Both common and rare fragile regions on the human genome were enriched for HBV integrations.

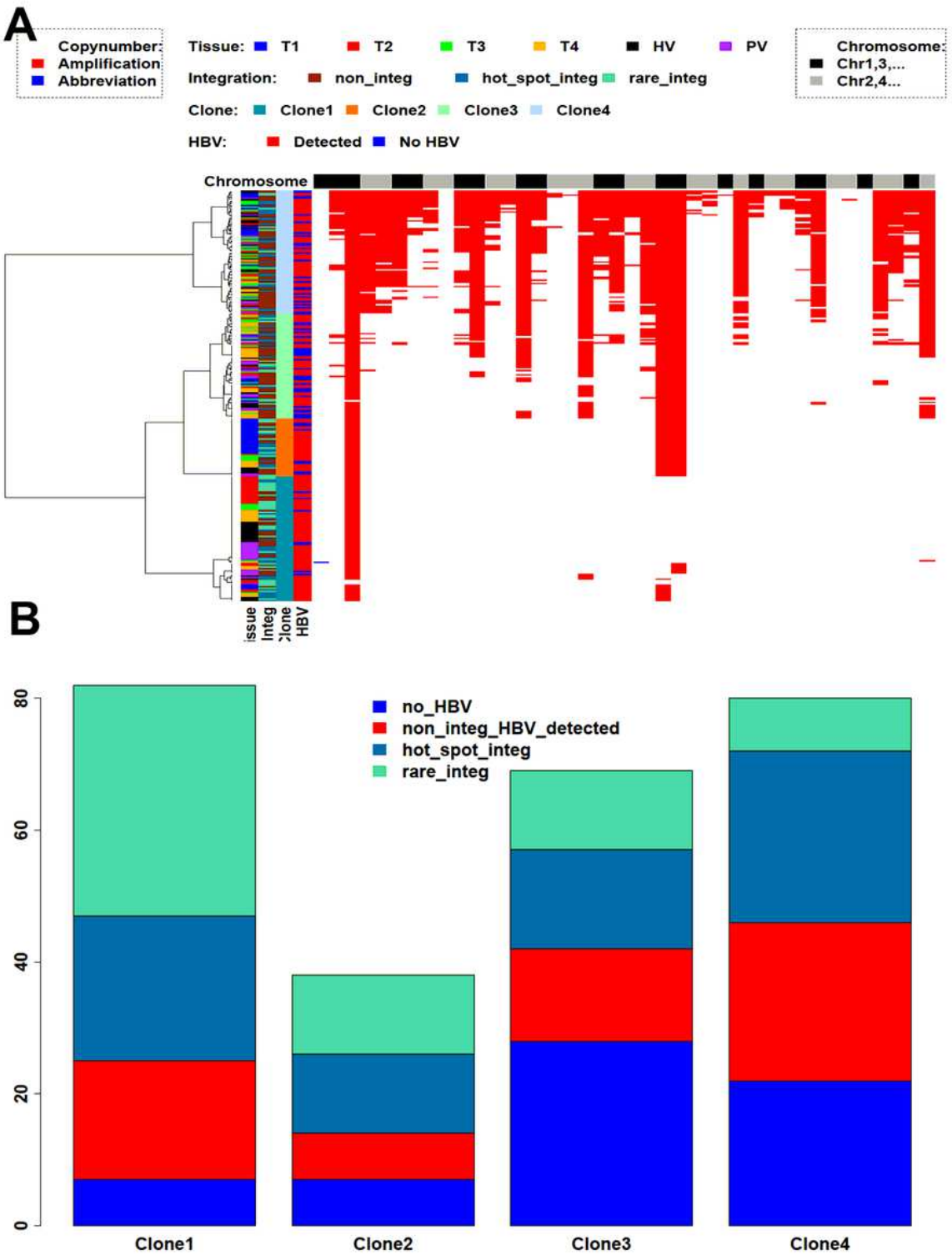

Figure 3

Copy number variation heterogeneity at single cell level. A) CNV profiles. Each row corresponds to a cell. The cells are labeled with regard to source tissue, clone annotation, HBV integration category and HBV sequence detection result. Each column corresponds to a bin. The bins are ordered by their chromosome locations (chromosome 1 to 22). Cells can be categorized into 4 groups corresponding to 4 clones. White means normal copy number, blue indicates copy number loss, red indicates copy number amplification.

B) Composition of cells with no HBV detected, cells with HBV sequence detected but no integration, cells carrying rare integration, and cells carrying only hot spot integration only in each clone. The frequency of cells carrying rare HBV integrations is highest for clone 1 and lowest for clone 4. The frequencies for clones 2 and 3 were comparable, and both were lower than the one for clone 1.

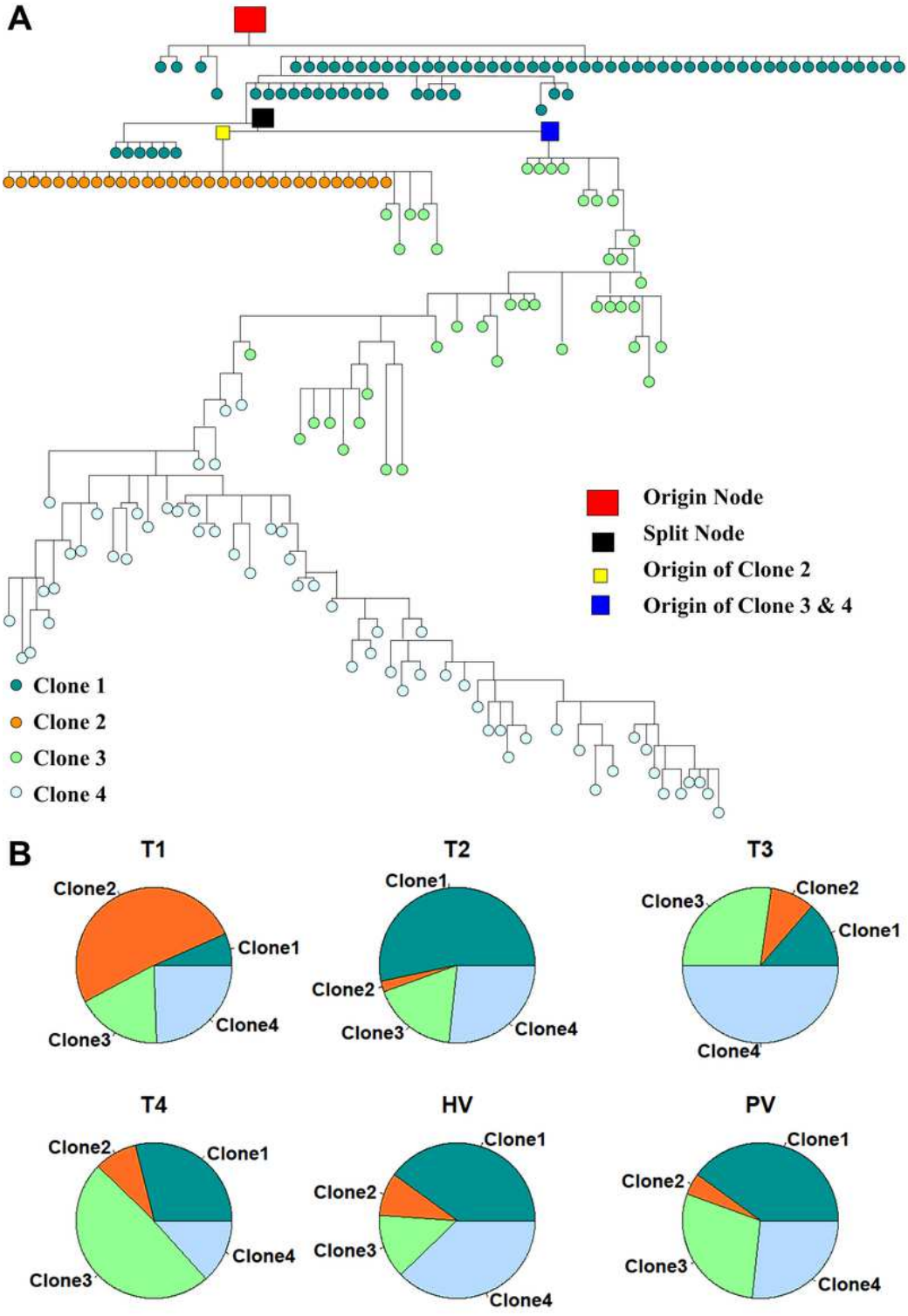

Figure 4

Clonal relationship of cells from different tumor sites. A) A phylogenetic tree built based on single cell CNV profiles. Each node corresponds to a cell. The cells are colored according clone annotation. Splitting nodes are marked as squared nodes. The scale of splitting node correlates to the number of its decedent nodes. B) Clone composition of each tumor tissue. Pie plots for each bin on the fractions of 4 clones. Each tumor tissue had one major clone and 3 minor clones. There was no single major clone in the thrombus tissues, but clones 3 and 4 together accounted for more than 50% of cells in the tumor thrombi, suggesting the two clones were more invasive.

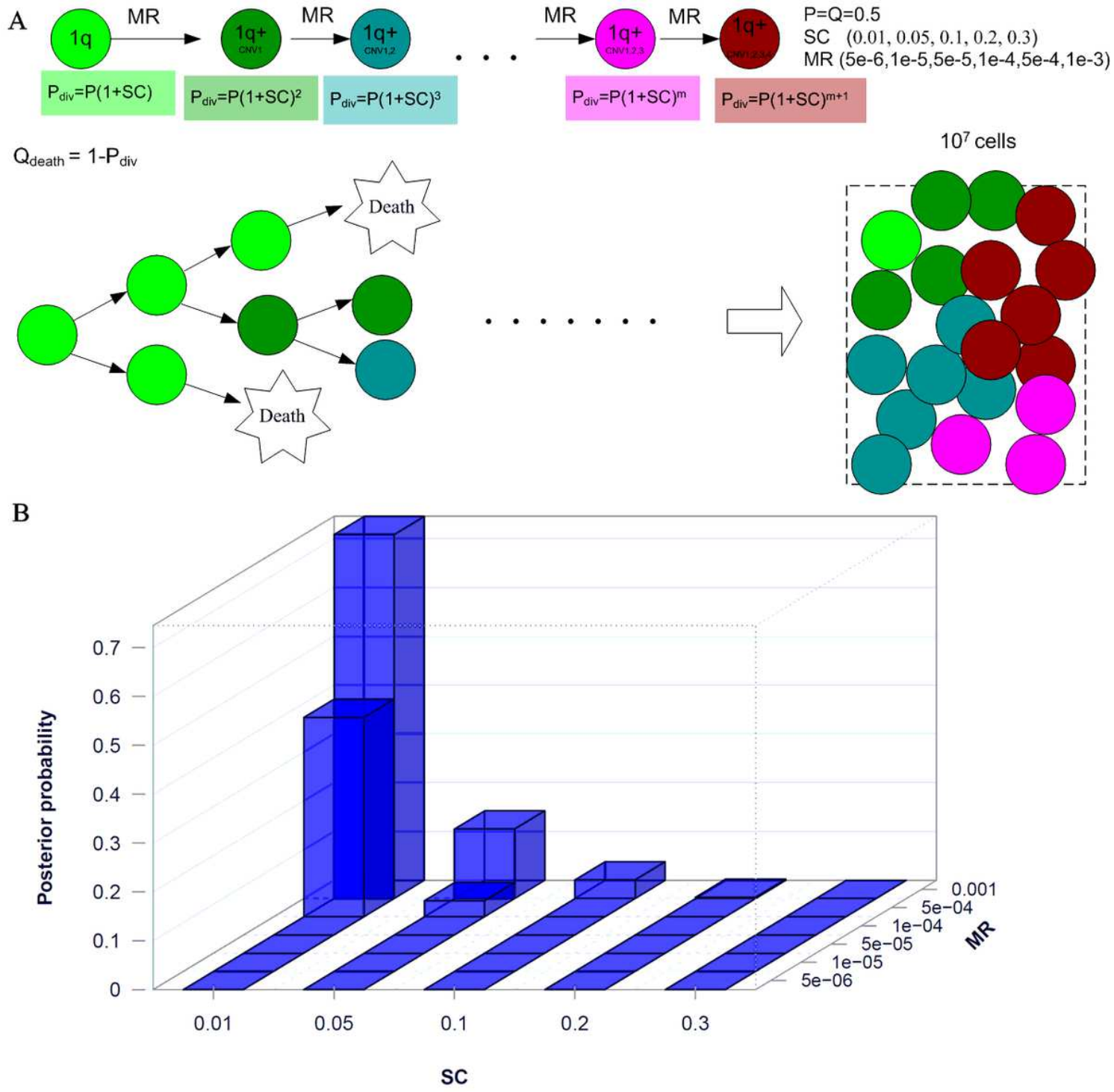

Figure 5

Simulation of clonal evolution with only CNVs. A) The scheme of birth-death clonal evolution model. Cells accumulated CNVs during cell growth. Each additional CNV increased cell's probability to divide over to die. B) Cell populations/tumors were simulated with different combinations of mutation rates (MRs) and selection coefficients (SCs). The posterior probability of each parameter combination was calculated.

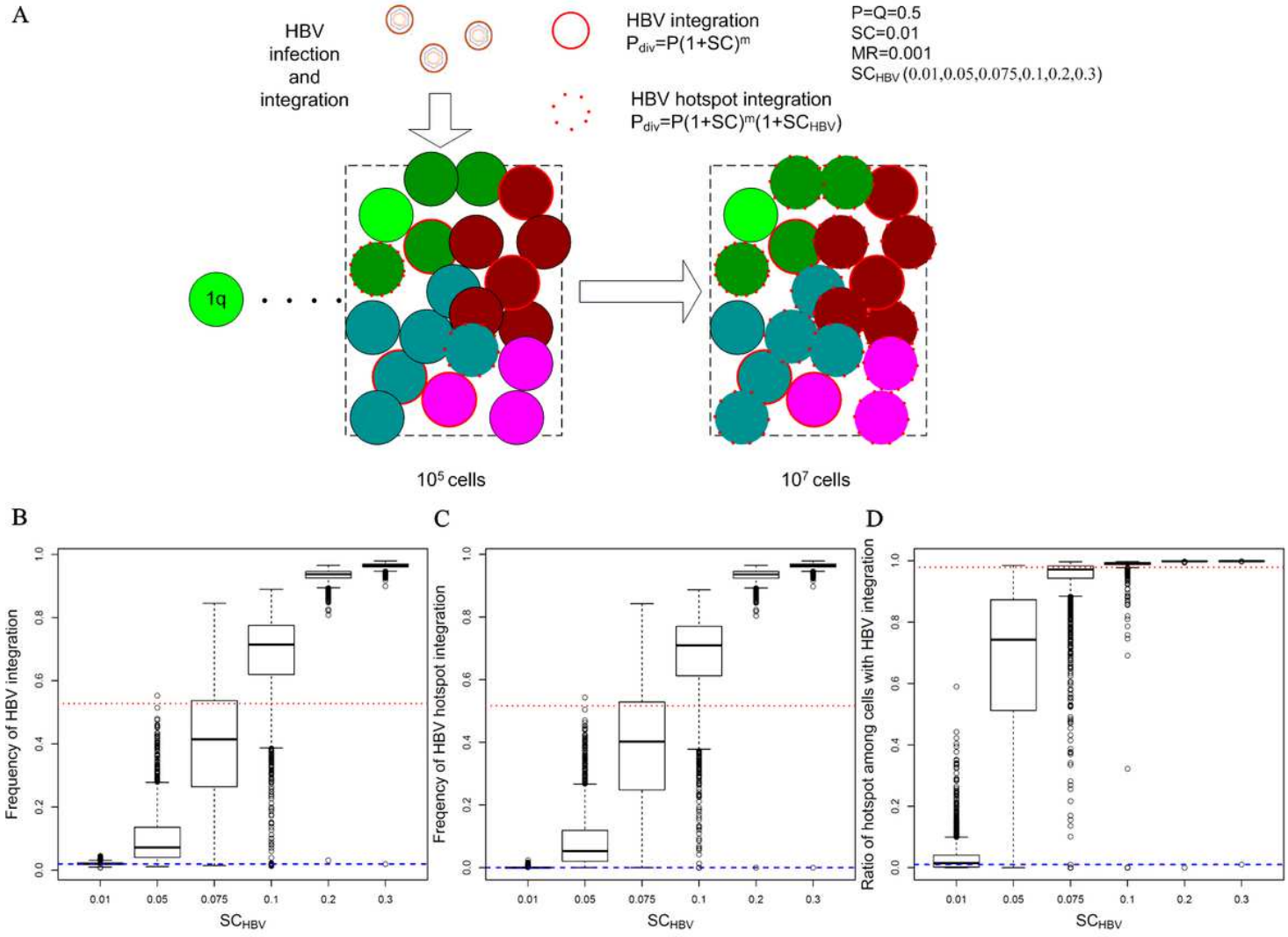

Figure 6

Simulation of clonal evolution with both CNVs and HBV integrations. A) The scheme of birth-death clonal evolution model with HBV integration. At the tumor size of 105 cells, cells were infected with HBV and HBV integration events occurred. Simulations were generated with the selection coefficient of the hot spot integrations SCHBV in a wide range. B) The frequency of cells with HBV integrations in the simulated cell populations. The red line is the observed frequency of cells with HBV integrations in the patient data (142/269) and the blue line marks the initial frequency of HBV integration (2%). C) The frequency of cells with the hot spot HBV integrations in the simulated cell populations. The red line is the observed frequency of cells with the hot spot HBV integrations in the patient data (139/269) and the blue line marks the initial frequency of the hot spot HBV integrations (0.02%). D) The ratio of cells with the hot spot HBV integrations versus cells with HBV integrations. The red line is the observed ratio in the patient data (139/142) and the blue line marks the initial ratio (1%).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1.tiff
- SupplementaryFigureS10.tiff
- SupplementaryFigureS11.tiff
- SupplementaryFigureS12.tiff
- SupplementaryFigureS13.tiff
- SupplementaryFigureS14.tiff
- SupplementaryFigureS15.tiff
- SupplementaryFigureS16.tiff
- SupplementaryFigureS17.tiff
- SupplementaryFigureS18.tiff
- SupplementaryFigureS19.tiff
- SupplementaryFigureS2.tiff
- SupplementaryFigureS3.tiff
- SupplementaryFigureS4.tiff
- SupplementaryFigureS5.tiff
- SupplementaryFigureS6.tiff
- SupplementaryFigureS7.tiff
- SupplementaryFigureS8.tiff
- SupplementaryFigureS9.tiff
- SupplementaryMaterialsubmission.docx
- SupplementaryTableS1.xlsx
- SupplementaryTableS10.docx
- SupplementaryTableS11.xlsx
- SupplementaryTableS12.xlsx
- SupplementaryTableS13.docx
- SupplementaryTableS14.docx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS6.xlsx

- SupplementaryTableS7.docx
- SupplementaryTableS8.xlsx
- SupplementaryTableS9.xlsx