Supplementary Materials for

**A case study of HBV integrations in HBV-HCC by HBV genome-enriched Single cell sequencing**

**Author Names:**

Wenhui Wang1,2†: wenhui.wang@mssm.edu

Yan Chen3† :c\_yan66@yahoo.com

Liang Wu4: wuliang@genomics.cn

Yi Zhang5 : zhaqi1972@163.com

Seungyeul Yoo1,2,6: seungyeul.yoo@mssm.edu

Quan Chen1,2,6: quan.chen@mssm.edu

Shiping Liu4: liushiping@genomics.cn

Yong Hou4: houyong@genomics.cn

Xiao-ping Chen3: chenxp@medmail.com.cn

Qian Chen7\*: chenqian201579@yahoo.com

Jun Zhu1,2,6,8\*: jun.zhu@mssm.edu

1Department of Genetics and Genomic Sciences, 2Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY;

3The Hepatic Surgery Centre at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (HUST), Wuhan, China

4BGI, Shenzhen, China;

5Department of Mathematics, Hebei University of Science and Technology, Shijiazhong, Hebei, China;

6Sema4, a Mount Sinai venture, Stamford, Connecticut, USA

7The Division of Gastroenterology, Department of Internal Medicine at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (HUST), Wuhan, China;

8The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY;

†co-first authors with equal contribution

\*Corresponding author:

Q.C. (chenqian201579@yahoo.com)

J.Z. (jun.zhu@mssm.edu) (will handle correspondence at all stages of refereeing and publication, also post-publication):

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

**Supplementary Tables:**

Supplementary Table S1. Correlation of CNVs detected with HGE-scSeq, WGS and SNP array based on Pearson, Spearman and Cosine.

Supplementary Table S2. Reads distribution for HGE-scSeq from tumor, bulk tissue data from adjacent normal and HGE-scSeq from MHCC97H cells including number of raw reads, number of reads after filtering, number of reads pair-ended mapped to human genome, number of reads pair-ended mapped to HBV genome and number of soft clipped reads covering HBV integration site.

**Supplementary Table S3.** HBV integrations detected with WGS from MHCC97H. At least one soft clipped read and two adjacent reads are required to call HBV integration.

**Supplementary Table S4.** HBV integrations detected with the HGE-scSeq of 5 MHCC97H cells. Highlighted integrations are matched with HBV integrations detected with WGS in 5000 bp range.

**Supplementary Table S5**. All detected Integration Events. All the detected HBV integration events for single cells from tumor and bulk tissue from adjacent normal. There are totally 471 HBV integrations observed from single cells in tumor and 17 HBV integrations observed from adjacent normal.

**Supplementary Table S6**. All unique integration sites. HBV integrations are merged if their position is within 20bp. They are totally 164 unique HBV integrations for single cells from tumor and 13 unique HBV integrations for bulk tissue from normal. The gene annotation is provided by running ANNOVAR.

**Supplementary Table S7**. Integration hot spots supported by known fusion events. Hot spots genes are reported as cancer fusion gene by both cancer cell line and TCGA for different kinds of cancers.

**Supplementary Table S8.** Genome regions where CNV amplification is significantly associated with decreasing rate of rare HBV integration carrying cells, when focusingon chr11, whose CNV differentiated clone 1 and clone 2, 3, 4,

**Supplementary Table S9**. Genome regions where CNV amplification is significantly associated with decreasing rate of rare HBV integration carrying cells, when focusingon chr8:118268310-146364022, whose CNV differentiated clone 2 and clone 3, 4,

**Supplementary Table S10.** Clinicopathological information of the patient. HCC, hepatocellular carcinoma; HBsAg, hepatitis B virus surface antigen; HBsAb, hepatitis B virus surface antibody; HBcAb, hepatitis B core antibody; HBeAb, hepatitis B e antibody; HCV Ab, hepatitis C virus antibody; AFP, alpha-fetoprotein; PVTT, portal vein tumor thrombosis; IVCTT, inferior vena cava tumor thrombosis. Hepatitis serology testing showed that the patient was HBsAb positive, HBsAg negative, HBcAb positive, HBeAb positive, HCV Ab negative and had no detectable blood HBV DNA copy number.

**Supplementary Table S11.** Coverage and width information for reads pair-ended mapped to human genome for HGE-scSeq from tumor, bulk tissue data from adjacent normal and HE-scSeq from from MHCC97H cells.

**Supplementary Table S12.** Match with Poisson distribution. Reads distribution on human genome is tested against Poisson distribution. The null hypothesis is the reads mapped to human genome following Poisson distribution. The test consistently fails until the corresponding region covering 88% of human genome.

Supplementary Table S13. Detected HBV virus and corresponding number of cells. The number of singles for each detected HBV sub strain is collected. The top 3 major HBV sub strains are HBV G247-B3 (GI121485896; GeneBank:EF134945.1), HBV strain Whutj-37 (GI38147024; GeneBank:AY293309.1) and HBV isolate G247-B5(GI:121485902;GeneBank:EF134946.1).

**Supplementary Table S14**. Pairwise alignment result with blat for the reference of top 5 most enriched HBV sub strain.

**Supplementary figure:**

**Supplementary Figure S1**. The location of tumors and thrombi on liver. A. Magnetic

resonance imaging (MRI) shows a 15 cm x 10 cm larger lesion in the left hepatic lobe and

multiple smaller lesions in the right hepatic lobe, all less than 3 cm in diameter. Yellow

arrows indicate multiple tumor foci of various sizes. B. MRI with contrast enhancement reveals tumor thrombosis involving the inferior vena cava (IVCTT), and the right portal vein branch

(PVTT), indicated by the red arrows, respectively, suggesting intrahepatic and extrahepatic

vascular spread of HCC.

Supplementary Figure S2. MHCC97H’s CNV profile generated by enriched single cell sequencing, whole genome sequencing and SNParray.

**Supplementary Figure S3**. Distribution of number of cells with readscovering the each loci for MHCC97H. Each bin corresponds to the fraction of human genome is successfully sequenced in a number of cells. If the reads distribute randomly on human genome, the distribution follows Poisson distribution. Chi-square test against Poisson distribution producing p-value 0.98.

**Supplementary Figure S4**. Distribution of repeatedly covered loci across the copy number amplified region called from Whole genome sequence data for MHCC97H.

**Supplementary Figure S5.** A. Compare HBV sequence and human genome sequence with Fisher values. B. Fisher values from Human mapped region for MHCC97H. C. Fisher Values from Human unmapped region for MHCC97H.

**Supplementary Figure S6**. A. HBV reads pileup results for an example cell with IGV. The reference genome is G247-B3. HBx-protein region is labeled as red. B. HBV reads pileup results comparing between tumor tissues and adjacent normal tissues. The upper panel is for all the HBV reads in adjacent normal tissues. The lower panel is for all the HBV reads in tumor tissues

**Supplementary Figure S7.** A. Distribution of HBV integrations across HBV proteins of P, S, X, C. HBV integrations are located on S, C and X. B. Distribution of HBV integrations across HBV genome.

Supplementary Figure S8. Compare the reads throughput on the two clustered sets of cells from Figure 2C. Histograms of reads throughput from these two sets of cells are almost overlapped. K.S. test shows no significant difference between these two distributions. The set of cells carrying extra integrations other than hot spot integrations are not benefit from higher throughput of reads.

**Supplementary Figure S9**. Label the phylogenetic tree in Figure 4A by carrying only hot spot integrations, extra rare integrations and no integrations. We can find that with dynamic clonal evolution. The rate of rare integration is becoming less and less.

**Supplementary Figure S10**. Expression of the hot spot genes from ICGC and TCGA. Hot spot genes *CSMD2, MED30*, and *EXT1* are find expressed significantly higher in tumor samples then adjacent normal samples.

**Supplementary Figure S11.** Data analysis flow chat. A. General analysis flow chat. After filtering low quality raw reads and detecting the HBV sub strain. HBV integrations and single cell CNV are called separately. B. Pipeline for detecting HBV integration. C. Pipeline for detecting single cell CNV.

**Supplementary Figure S12.** Histograms of number of human reads (A), number of HBV reads (B), number of inter chromosome chimera reads (C), number of intra chromosome chimera reads (D), number of softclipped reads (E). The average chimera reads ratio is 0.025 % which is lower than the reported chimera reads ratio of 6.19% by Tu et.al and 2%/3% by Xie’s group. F) Correlation coefficients between number of human reads, number of inter chromosome chimera reads, number of intra chromosome chimera reads, number of HBV integrations and number of HBV reads. Number of chimera reads for inter and intra chromosome are highly correlated. Number of chimera reads are not correlated with number of reads on HBV, number of soft clipped reads and number of reads on human. Number of reads on HBV and number of soft clipped reads are correlated.

**Supplementary Figure S13.** Linear correlation between inter chromosome chimera reads, intra chromosome chimera reads and length of chromosomes. Scatter plots (A, C) and boxplot (B, D) of number of chimera reads and length of chromosome for both inter and intra chromosome cases. The blue triangles indicate Chr1 and Chr8. The numbers in A and C are (correlation between chromosomes’ length and mean # of chimera reads | p-value) and (correlation between chromosomes’ length and median # of chimera reads | p-value). The correlations between numbers of chimera reads and length of chromosome are significant.

**Supplementary Figure S14.** Distribution of number of cells with reads covering the each loci. Red line indicates the mean. Each bin corresponds to the fraction of human genome is successfully sequenced in a number of cells. If the reads distribute randomly on human genome, the distribution follows Poisson distribution.

**Supplementary Figure S15**. A. Compare HBV sequence and human genome sequence with Fisher values. B. Fisher values from Human mapped region. C. Fisher Values from Human unmapped region.

**Supplementary Figure S16**. A. Find the best tuning parameter for the pseudo count weight adjustment. B. Select the best cutoff for the selected best tuning parameter.

**Supplementary Figure S17.** Quality of Bin's read count correction. A. Fold enrichment of top x% bins carrying HBV integration before correction. Bins are sorted by the number of reads mapped in the bin. B. Fold enrichment of top % bins carrying HBV integration after correction. Bins are sorted by corrected reads. C. MAPD and MAD before batch effect correction. D. MAPD and MAD after batch effect correction.

**Supplementary Figure S18.** Comparison of the number reads between normal bulk tissues and tumor single cells. Histogram shows the distribution for tumor single cells while vertical color lines show the corresponding quantity of normal control tissue. A: comparison of numbers of filtered reads; B: comparison of percentage of reads mapped to human genome; C: comparison of coverage on human genome; D: comparison of width on human genome.

**Supplementary Figure S19.** A. Comparison dispersion of binned reads count after mappability and GC content correction between the smallest one in single tumor cells and the four normal control tissue. B. CNV results on normal tissues.

.