## Response to Reviewer #1 Comments:

### 1. —kindly include full form of HIVID

We added the full form in the text. HIVID stands for High-throughput Viral Integration Detection. Thank you for pointing this out.

### 2. In discussion—- Integration at 11q13.3 is a potential driver event[19], but its frequency is not high.——not in line with reference 19—-kindly check

Thank you for finding this issue. The information in this text is not accurate and requires correction. According to reference 19, integration into *CCND1*, not 11q13.3, is mentioned as a potential driver event. Here is a quote from reference 19:

*Among the non-recurring HBV insertional events associated with very high levels of RNA transcription, potential driver events were observed. These include known oncogenes CCND1, CCNE1, and GLI2 (a sonic hedgehog transcription factor). Thus, the effect of HBV on transcriptional levels of key oncogenes demonstrated potential driver events affecting a number of patients.*
(Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017)

The integration into CCND1 is also listed in "Supplementary Table S5C" in reference 19.

Supplemental Table 5C. Identification of HBV RNA fusions with likely cancer driver activities in 196 TCGA HCC

| Human Gene | # Insertions | Mean RNA Expression* | Predicted Oncogene/Tumor Suppressor | Evidence for Cancer Association | Relevant Papers-1** | Relevant Papers-2** |
|---|---|---|---|---|---|---|
| *MLL4* | 5 | 97.8 | Oncogene | chromatin regulator and frequent target of HBV insertion in HCC | 25901726 | 22634754 |
| *TERT* | 2 | 88.2 | Oncogene | telomerase; established cancer driver gene in HCC | 26099527 | |
| *CCNE1* | 1 | 99.4 | Oncogene | cyclin E1; key cell cycle progression gene overexpressed in cancers | 16043362 | |
| *GLI2* | 1 | 100 | Oncogene | transcription factor that regulates sonic hedgehog genes | 17440069 | 21695716 |
| *CCND1* | 1 | 97.9 | Oncogene | cyclin D1; cell cycle progression gene amplified/overexpressed in some HCC | 21397858 | 23505090 |
| *ST8SIA1* | 1 | 97.1 | Oncogene | regulates stem cell function and activates c-Met; regulates cell adhesion | 20889649 | 25109336 |

Therefore, we have corrected the text to be more accurate as follows:

“Integration into CCND1, located at 11q13.3, is a potential driver event[23], but its frequency is not high.”

Reference 19 suggests that the integration of HBV into 11q13.3 increases the expression of *CCND1*.

Our analysis revealed a diverse pattern of HBV integration into 11q13.3. It is possible that the mechanism behind oncogene activation is copy number amplification, rather than activation of promoters such as *TERT* leading to increased gene expression. We conducted a search for studies exploring the relationship between the increased copy number of the cancer-driving gene FGF19 on 11q13.3 and HBV-infected liver cancer. We found two reports that showed a significant correlation between these two factors. These studies were cited in the DISCUSSION section.

“HBV integration at this locus may be linked to cancer gene activation, as FGF19 amplification was associated with chronic HBV infection.[28,29].”

### 3. several studies have detected these events[4,5,20]————- not in line with

**quoted references—-kindly check**

Thank you for checking the references. Here, “these events” meant the detection of HBV DNA integration into 11q13.3. Integration into 11q13.3 is not mentioned in the text because it is not the topic of these papers and is only shown in the supplemental data. We have revised the text to clarify that we refer to supplemental data.

“several studies have detected these events only as supplementary data[4,5,24]”

## Response to Reviewer #2 Comments:

**It’s worth noting that there is no line numbers in the manuscript, and line number have been added, the following points listed according to that line numbers.**

Thank you for bringing this to our attention. We apologize for not including line numbers in the manuscript. We understand the importance of adhering to academic conventions and will include line numbers in future submissions to make it easier for reviewers. We appreciate your feedback and will consider it for future writing.

### Major point:

**1. Line 104-105: As we all known, marker genes are critical and often play the most important roles in the pathway. In this paper, frequent integration breakpoints were similar to that marker genes, however, the data in the Table2 suggested that the original study worked better than the current study, so what’s the point of this study?**

We would like to take this opportunity to clarify the key points of our study, as it may not have been sufficiently explained in the previous manuscript. We would like to clarify three key differences between our study and the original one.

1. Detection of integrations in a larger number of samples:

In our study, we detected HBV integrations in a larger number of samples compared to the original study. We have revised the "Comparison of HBV integration sites" section to clarify this point. We modified Table 1 to show the number of samples and breakpoints detected and added the number of samples in Table 2 by gene. Additionally, we added a bar graph on the right side of Figure 1A to visualize the increase in the number of samples.

Table1

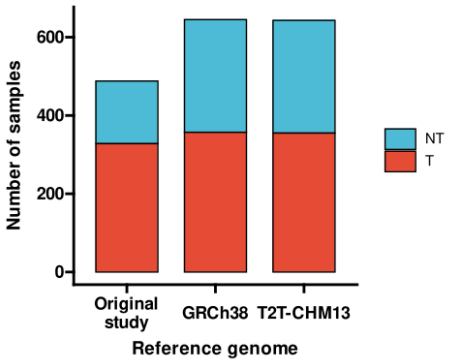
| | GRCh38 | T2T-CHM13 | Original |
|---|---|---|---|
| Tumor | | | |
| Number of breakpoints | 2439 | 2487 | 3486 |
| Number of samples | 357 | 355 | 328 |
| Non-tumor | | | |
| Number of breakpoints | 2759 | 2874 | 739 |

| | | | |
|---|---|---|---|
| Number of samples | 288 | 288 | 160 |

Table2

| Gene | GRCh38 | | Original | |
|---|---|---|---|---|
| | Breakpoints (n) | Samples (n) | Breakpoints (n) | Samples (n) |
| Tumor | | | | |
| *TERT* | 150 | 105 | 160 | 95 |
| *KMT2B* | 56 | 33 | 55 | 30 |
| *DDX11L1* | 0 | 0 | 36 | 23 |
| *CCNA2* | 12 | 7 | 14 | 8 |
| *CCNE1* | 13 | 9 | 14 | 7 |
| Non-tumor | | | | |
| *FN1* | 97 | 56 | 19 | 17 |
| *TERT* | 12 | 10 | 8 | 3 |
| *IQGAP2* | 7 | 5 | 1 | 1 |
| *KMT2B* | 7 | 4 | 5 | 3 |

Figure1A (right side)




2. Use of a more comprehensive reference genome:

In this study, we used a newer reference genome, GRCh38, and T2T-CHM13. Although T2T-CHM13 includes repetitive regions, GRIDSS has been verified using T2T-CHM13 and can accurately detect single breakend SVs for repetitive sequences.

3. Integration detection using neworithms and software:

Our analysis used a structural variant caller based on an advanced algorithm that assembles soft-clipped reads, reads with unmapped or ambiguously mapping mates, and assemblies with unmapped or ambiguously mapping breakend sequence.

HIVID assembles paired-end reads and detects integration sites, but it has the disadvantage of not using reads with an insert size greater than 200 bp for detection.

The "insert size" of the reads obtained from HIVID is shorter than that of normal sequences, but it still contains a substantial number of reads with an "insert size" of 200 bp or more. For example, we plotted the "insert size" of the randomly selected sample SRR3104550 using Picard. (We also measured the "insert size" of several other samples randomly selected, and the same trend was observed.) The green frame indicates reads with an insert size of 200 or more.

GRIDSS creates a more complete breakend assembly for detecting integrations by utilizing discordant read pairs, soft-clipped reads, split reads, reads with unmapped mates, and indel-containing reads. In other words, it can effectively utilize the reads that were not utilized by HIVID.

As a result, Breakend is expected to detect breakpoints that were not detected by HIVID and avoid false positives.

Supplementary Figure 4 displays examples of newly detected integrations in our analysis and provides an explanation of potential causes in Supplementary Figure 5. Furthermore, Supplementary Figure 2 shows examples that were not detected in our study but were found in the original study, which may be false positives.

**2. Line 34, 182: These two conclusion seem to exist some ambiguity, please make sure your conclusions are consistent.**

We have made improvements to make two CONCLUSIONs more consistent.

CONCLUSION in the abstract

"GRIDSS VIRUSBreakend using T2T-CHM13 was accurate and sensitive in detecting HBV integration. Re-analysis provides new insights into the regions of HBV integration and their potential roles in HCC development."

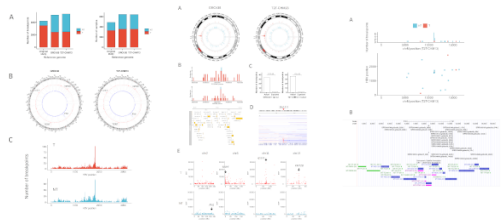
CONCLUSION in the main text

"HBV integration into HCC samples was characterized using the complete human reference. GRIDSS VIRUSBreakend using T2T-CHM13 was accurate and sensitive in detecting HBV integration. HBV frequently integrates at the 11q13.3 region, where the CCND1 gene is located, and this region is frequently amplified in several types of cancer, including HCC. Further research is needed to examine how HBV integration interacts with driver gene expression and copy number alteration."

**3. Figures should be combined into one figure, such as figure 1A-1C should be put together, but not interspersed.**

We have taken your feedback into consideration and combined the figures into one figure, labeling each section as A, B, C, and D. We apologize for any inconvenience caused by the

lack of integration in the previous figures. Thank you for taking the time to review our work.



## Minor point:

Thank you for pointing out these minor points. We have made the necessary corrections.

**1. Line 12: "hepatocellular carcinoma" should be written as "HCC".**

Changed "hepatocellular carcinoma" to "HCC"

**2. Line 22: what is the hg19? Please write full name on its first occurrence.**

We have corrected it as follows.

"Most previous studies have used Genome Reference Consortium Human build 37 (GRCh37) or human genome 19 (hg19) as the reference genomes."

**3. Line 52-53: when HBV infects liver cells, is HBV DNA integrated into the human genome certainly? If not, please do a accurate description.**

As you mentioned, when HBV infects liver cells, HBV DNA is not always integrated into the human genome. We have corrected the text below.

"When HBV infects liver cells, HBV DNA can be integrated into the human genome."

**4. Line 57: HBx, Please write full name on its first occurrence.**

We have corrected as follows.

"(4) inducing expression of HBV X protein (HBx) or HBx fusion proteins that contribute to carcinogenesis."

**5. Line 63-64: "HBV-infected HBV-infected" should be written as "HBV-infected".**

We have corrected as follows.

"In an examination of an HBV-infected human-hepatocyte chimeric mouse model, mitochondrial DNA (mtDNA) was thought to be a frequent site of integration[1]."

**6. Line 64: "mitochondria" should be written as "mitochondria DNA".**

We have corrected as follows.

"In an examination of an HBV-infected human-hepatocyte chimeric mouse model, mitochondrial DNA (mtDNA) was thought to be a frequent site of integration[1]."

**7. Line 74: The word "breakends" has not an exact meaning. Please use the exact word.**

Thank you for pointing this out. We have revised the part as follows.

"VIRUSBreakend utilizes a virus-centric variant calling and assembly approach to identify viral integrations with high sensitivity and low false discovery rate, allowing the identification of integrations in repetitive host regions"

"Breakend" is a term used in the VCF file format. The VCF file format can represent Single Nucleotide Variants, Deletions, Insertions, and others (DEL, INS, DUP, INV, CNV, BND) in one record. Breakend (BND) means that a part of the genome is connected to another part. The breakpoint has two breakends (human genome side and virus genome side). GRIDSS is different from other structural variation detection software in that it reports all records as breakends. This allows for structural variations and copy number amplification/decrease to be treated separately. These features are particularly useful for detecting virus genome integration. GRIDSS VIRUSBreakend first detects a single breakend on the virus genome and then detects the corresponding breakend on the human genome. This allows for detection even in cases where the human side is difficult to map to a repeating sequence.

**8. Line 95-96: "In total" and "Overall" have similar meaning, this sentence should be rewritten.**

We have corrected as follows.

"In total, 5361 and 5198 integration breakpoints were detected with T2T-CHM13 and GRCh38, respectively. The breakpoints were similar between the references using GRCh38 and T2T-CHM13 (Figures 1A and B)."

**9. Line 101-103: "For example,….. in the original analysis", these data were not showed up in this paper.**

Thank you for pointing this out. We have created a supplemental table that can be opened in Microsoft Excel so that readers of the Journal can access all the results. We also prepared Table 3, which summarizes the number of HBV DNA incorporated into FN1 and TERT in non-tumor samples.

**10. Line 112, 141, 156 et al: Chr11q13.3 and 11q13.3 should mean the same thing, please unified their writting.**

Thank you for pointing this out. Unified to 11q13.3.

**11. Line 134: "the GCCXTTCTCATC sequence", please explain what does X stand for in the DNA sequence.**

Thank you for pointing this out. We have revised the part as follows:

"For example, the GCCNTTCTCATC sequence, where N represents any nucleotide or gap,

was observed at the junction of the *ND4* gene (Chromosome M:11079) and the HBV genome (HBV:1559). In contrast, the GCTTCACC sequence was observed at the junction of the *ND4* gene (Chromosome M:11104) and the HBV genome (HBV:1590)."

**12. Line 167, 172: "reported" is used many times in the manuscript, please replace other similar words.**

We used the services of an English editing service to improve the overall English text. We believe that these English problems have been corrected.

**13. The references need to be edited and sorted using reference management software, such as ENDNOTE.**

Thank you for pointing this out. We use Zotero and Pandoc's coross-refefence to sort out references.

**14. Editing for English grammar and usage is needed.**

Thank you for pointing this out. I submitted my manuscript to a specialized English proofreading company and received correction.