Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection

Chao Jin¹#, Xiaonan Liu²#, Wenyuan Zheng³#, Liping Su⁴, Yang Liu⁴, Xu Guo⁴, Xiaoming Gu⁴, Hongping Li³, Bo Xu³, Gang Wang⁴, Jiyan Yu³, Qiong Zhang³, Dengke Bao², Shaogui Wan⁶, Fei Xu⁷, Xiaohuan Lai⁷, Jiayun Liu⁸*, Jinliang Xing⁴*

¹ Department of General Surgery, Tangdu Hospital, Fourth Military Medical University, Xi'an, Shaanxi 710032, China
² Ambulatory Surgery Center of Xijing Hospital, Fourth Military Medical University, Xi'an, Shaanxi 710032, China

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/1878-0261.13041

Molecular Oncology (2020) © 2020 The Authors. Published by FEBS Press and John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Chao Jin, Xiaonan Liu, and Wenyuan Zheng contributed equally to this study

*Correspondences:

Jinliang Xing, M.D., Ph.D., State Key Laboratory of Cancer Biology and Department of Physiology and Pathophysiology, Fourth Military Medical University, Xi’an, Shaanxi 710032, China. Tel: (86)-29-84774764; Email: xingjl@fmmu.edu.cn.

Jiayun Liu, M.D., Department of Clinical Laboratory, Xijing Hospital, Fourth Military Medical University, Xi’an, Shaanxi 710032, China. Tel: (86)-29-84775455; Email: jiayun@fmmu.edu.cn.
Abbreviations

HCC: hepatocellular carcinoma
HBV: hepatitis B virus
cfDNA: circulating cell-free DNA
WGS: whole-genome sequencing
SNVs: single-nucleotide variants
CNV: copy number variation
TF: tumor fraction

Key words: Hepatocellular carcinoma; circulating cell-free DNA; tumor fraction; copy number variation; fragment sizes; end motifs
Abstract
Circulating cell-free DNA (cfDNA) fragmentomics, which encompasses the measurement of cfDNA length and short nucleotide motifs at the ends of cfDNA molecules, is an emerging field for cancer diagnosis. The utilization of cfDNA fragmentomics for the diagnosis of patients with hepatocellular carcinoma (HCC) caused by hepatitis B virus (HBV) is currently limited. In this study, we utilized whole-genome sequencing data of cfDNA in samples from patients with HCC (n=197) and HBV (n=187) in order to analyze the relationships of fragment size selection (<150 bp) with tumor fraction (TF), copy number variation (CNV) alterations, and the proportional change of 4-mer end motifs in HCC and HBV samples. Our analyses identified five typical CNV markers (i.e. loss in chr1p, chr4q and chr8p, and gain in chr1q and chr8q) in cfDNA with a cumulatively positive rate of ~95% in HCC samples. Size selection (<150 bp) significantly enhanced TF and CNV signals in HCC samples. Additionally, three 4-mer end motifs (CCCA, CCTG, and CCAG) were identified as preferred end motifs in HCC samples. We identified 139 end motifs significantly associated with fragment size that showed similar patterns of associations between patients with HCC and HBV, suggesting that end motifs might be inherently coupled with fragment size by an ubiquitous mechanism. Here we conclude that CNV markers, fragment size selection and end motif pattern in cfDNA, hold the potential for effective detection of patients with HCC.

1. Introduction
Hepatocellular carcinoma (HCC) is the sixth most common cancer and fourth leading cause of cancer death with 841,000 new cases and 781,000 death worldwide in 2018 [1]. Hepatitis B virus (HBV) as one of the main risk factors of HCC, is responsible for 35% death of HCC [2]. HBV may contribute to HCC through integrating HBV DNA into host genome to mediate genetic abnormality and influence the expression of HCC-related genes [3]. HBV infection is the leading cause of HCC in Asian and African countries, especially in China [2, 4]. Early diagnosis and effective control are important to reduce the social burden induced by HBV-infected HCC. To
date, although tumor biopsy remains the gold standard for HCC diagnosis, controversies are raised with less reliable in small nodule cases and adverse consequences such as bleeding and intraprocedural hematogenous dissemination [5]. For HBV infected HCC with poor liver function, it is urgent to search for new potential biomarkers based on noninvasive methods.

Circulating cell-free DNA (cfDNA) is a kind of short fragments detectable in blood or other body fluids shed from different cell types. Recent advances in combined cfDNA analysis and whole-genome sequencing (WGS) method contribute to the development of noninvasive liquid biopsy for cancer diagnosis as a surrogate of standard tumor biopsies [6]. So far, the characterization of cfDNA in HCC patients has been well reported and laid the foundation for HCC diagnosis by using genomic alteration and DNA methylation derived markers, such as single-nucleotide variants (SNVs) and copy number variations (CNVs) [7-9]. However, the biomarkers excavated for the diagnosis of HCC patients with HBV infection are still limited.

Recently, cfDNA fragmentomics, also named as properties or patterns of cfDNA fragmentation, which encompasses fragment sizes, ends points and nucleosome footprints, is an emerging field for cancer diagnosis [10-12]. Cristiano et al applied machine learning model based on incorporated fragmentation features to improve sensitivities and specificity for cancer detection [13]. DNA fragment selection of 90-150 bp was reported to improve the detection of mutated DNA fraction of ovarian cancer up to 11-fold [14]. However, the cfDNA fragment patterns in cancers, especially in HBV-HCC, are still poorly understood. Therefore, a comprehensive understanding of the fragmentation features, mechanisms, and patterns of cfDNA is important for discovering promising cancer biomarkers in HCC.

A latest effort by Jiang et al has exploited the DNA end-motif profiles of HCC samples using massively parallel sequencing and mentioned a significant increase in the diversity of DNA end motifs in HCC samples [15]. Additionally, they reported that HCC subjects had some preferential patterns of 4-mer end motifs such as CCCA, compare to controls. However, in their study, they focused on HCC patients, not HCC patients with HBV infection. Additionally, as both fragment size and motifs are associated with ctDNA, it remains unclear whether they are independent
biomarkers or not to enhance cancer detection. As fragment sizes and end motifs are likely to be involved in DNA cleavage, investigation of fragment sizes and end motifs is of critical importance in further dissecting the mechanisms of cfDNA patterns.

Thus, in this study, we analyzed WGS data (~5X depth coverage on average) of cfDNA from 197 HCC with HBV infection and 187 non-cancer HBV samples [16]. We aimed to screen and identify representative CNV markers for distinguishing the HBV-HCC samples from controls with HBV infection, and further select appropriate fragment sizes to improve the tumor fraction (TF) of ctDNA in HCC samples. We also attempted to explore the end motifs patterns as a biomarker for HCC detection, and in particular, evaluate the relationships of fragment sizes and 4-mer end motifs in HCC and HBV samples.

2. Methods and methods

2.1. Sample characteristics

The WGS data of 197 HBV-related HCC samples and 187 non-HCC HBV samples from our previous investigation [16] were utilized in this work. The characteristics of samples were shown in Table 1 of the previous article [16]. The study was approved by the local Ethics Committees of the involved hospitals, and all the subjects had written informed consents. This study methodologies conformed to the Declaration of Helsinki.

2.2 Cell-free DNA isolation, library construction and WGS sequencing

Plasma cfDNA from fresh whole blood was extracted using QIAamp Circulating Nucleic Acid Kit (Qiagen) following the manufacturer’s protocols. cfDNA concentration and quality were assessed using Qubit 3.0 (Thermo Fisher) and Bioanalyzer 2100 (Agilent Technologies), respectively. cfDNA libraries for WGS sequencing were prepared using about 20ng cfDNA as previously described [16]. Briefly, cfDNA was processed by end repairing, dA tailing and ligation to loop adapter, followed by size selection using AMPure XP beads (Beckman). The adaptor-ligated DNA fragments were amplificated by a 14-cycle PCR using two Illumina p5 and p7 primers. Subsequently, libraries were quantified by KAPA Library Quantification kit (KAPA Biosystems)
and library fragment size was determined by Bioanalyzer 2100. WGS sequencing was then conducted on the Illumina HiSeq X10 platform using a paired-end 150 bp protocol.

2.3 WGS data processing
Raw paired-end reads were pre-processed using fastp (version 0.20.0) to remove adaptor, low quality bases and consensus bases. Quality control of reads was evaluated using QPLOT [17]. Then, the trimmed reads were aligned to the reference human genome hg19 using BWA (version 0.7.17, https://sourceforge.net/projects/bio-bwa/files/) [18] and GATK4 MarkDuplicates (version 4.1.2.0) was used for duplicate reads masking of BAM files.

2.4 CNV and TF analyses
In order to estimate the CNVs and TFs of HCC samples, ichorCNA was used to analyze fractions of tumor in cfDNA from WGS samples [19]. Briefly, the genome was divided into non-overlapping 1-Mb bins, and reads with mapping quality of less than 20 were filtered. The mapped reads within each bin were calculated by HMMcopy software in R package (http://bioconductor.org/packages/2.11/bioc/html/HMMcopy.html). After conducting GC content and mappability bias correction, CNAs prediction and TF estimation in cfDNA were performed using ichorCNA. Segmentation of CNVs profiling data was performed using ichorCNA, in which a HETD CNV region is defined as loss and a GAIN or AMP CNV region is defined as gain. The maps of genome-scale overview of cfDNA CNVs were drawn by Matplotlib (version 3.1.1) package.

2.5 cfDNA fragment size analysis
The insert size of an assigned read pair from BAM file was considered as the fragment size of cfDNA fragment. Fragment size distributions of cfDNA were drawn using Matplotlib (version 3.1.1) package. Mean proportions of cfDNA fragments less than 150 bp in all fragment sizes from 20 HCC samples with TF>0.2, 43 HCC samples with 0<TF≤0.2, and 187 HBV samples were analyzed.
compared using one-sized t-tests and the results represented by box plots were drawn using the Matplotlib package. The effects of selecting fragments of size ≤150 bp on TF and CNV signal changes in HCC samples were further evaluated. Firstly, read pairs with insert sizes of ≤150 bp were extracted to rebuild a new BAM file. Then, the TF after ≤150 bp size selection was recalculated through ichorCNA software, and the linear least-squares regression was conducted using scipy.stats.linregress function in Python to investigate the correlation between TFs values before and after size selection.

2.6 cfDNA 4-mer end motifs analysis

cfDNA end motifs were defined using the first 4-nucleotide sequence of each 5’ end of aligned reads as previously described [15], and only properly paired reads with mapping quality >20 were used for downstream analyses. In total, there were 256 possible 4-mer motifs, among which the frequencies of six representative motifs (CCCC, CCTG, CCAG, TAAA, AAAA and TTTT) at the end of sequenced fragments in HCC-high (TF>0.2), HCC-low (0<TF≤0.2) and HBV samples were calculated, and compared using one-sized t-tests. The results shown by box plots were drawn by scipy software. The fragment size distributions of reads with these six representative end motifs were analyzed using Matplotlib. Next, the proportion changes of 256 4-mer end motifs before and after selecting fragments shorter than 150 bp in 197 HCC samples and 187 HBV samples were calculated. The associations of the proportions of 256 4-mer end motifs before and after size selection were investigated by linear least-squares regression. Furthermore, hierarchical clustering was conducted to analyze the associations of fragment sizes and 70 end motifs using clustermap function in seaborn software (version 0.9.0), and only end motifs with their proportions greater than 0.005 were used.

3. Results

3.1 CNV alterations of cfDNA in HCC samples based on WGS

The cfDNA profiles of 63 HCC samples with estimated TF>0 and 187 HBV samples were
analyzed. As a result, the CNVs of chr1, chr4 and chr8 were commonly affected in HCC, including loss in chr1p, chr4q and chr8p, and gain in chr1q and chr8q (Fig. 1). Our data revealed that 59 of 63 (93.7%) HCC samples exhibited at least one of the above five CNVs, and 18 of 19 (94.7%) HCC samples with high TF (TF>0.2) had at least one of these CNVs alterations. For HBV samples, 185 of 187 cases presented none of the above CNVs alterations except for two cases: one (HBV-275) was detected with chr1q and chr8q amplifications (chr1q: \( P=3.02\times10^{-10} \); chr8q: \( P=4.37\times10^{-3} \); one-sided t-test comparing the q-arm and the p-arm), and another (HBV-116) was detected with the same CNV aberrations (chr1q: \( P=7.59\times10^{-4} \); chr8q: \( P=8.31\times10^{-4} \)) (Table 1). Unfortunately, both of these two subjects were subsequently diagnosed with HCC during follow-up, with HBV-275 being diagnosed at 3-month and HBV-116 being diagnosed at 6-month follow-ups. The results suggested that the HCC samples had high positive rates of these five typical CNVs markers across patient samples, providing valuable biomarkers in distinguishing HCC samples from HBV controls, and for early detection of undiagnosed HBV patients 6 months before the clinical diagnose.

3.2 Utilization of cfDNA fragment size to enhance the detection of CNV markers

After CNV analysis, the size distribution of cfDNA of HBV-HCC samples with TF>0.2 and control samples with HBV infection were explored. The cfDNA of all samples showed a most prominent peak at 167 bp in the size distribution plot, for both HBV-HCC and HBV samples (Fig. 2A). A distinct difference of fragment lengths was observed between cfDNA derived from HCC and HBV samples, in that cfDNA fragment lengths in HBV-HCC samples were obviously enriched in shorter lengths (< 150 bp) than that in HBV samples (Fig. 2A). Box-plot analysis showed that the proportions of cfDNA fragments < 150 bp in high-TF (TF>0.2) (mean=0.39, \( P=7.37\times10^{-7} \)) and low-TF (0<TF≤0.2) (mean=0.27, \( P=5.29\times10^{-7} \)) HCC samples were all significantly higher than that in HBV samples (mean=0.2) (Fig. 2B). These results demonstrated that cfDNA fragments shorter than 150bp could well detect the CNV markers, consistent with previous reports [20]. This feature enables us to further explore the utility of fragment size to
enhance the detection of cancer using CNV aberrations as a biomarker.

3.3 ctDNA enrichment with cfDNA fragment selection increases TF and CNV signals

In light of the above findings, we sought to evaluate whether the selection of cfDNA fragment size < 150bp could enrich ctDNA fragments against the large background of cfDNA fragments. Before fragment selection, there were 63 HCC samples with TF>0 and 134 HCC samples with TF=0, while we observed a significantly increased number (N=107) of HCC samples with estimated TF>0 after selecting fragments of <150bp, corresponding to a 70% increase in detection rate. We further quantified the magnitude of ctDNA enrichment, and observed that the average tumor purity of cfDNA in HCC samples was enhanced by approximately 6.5% (Fig. 3A). For instance, prior to fragment size selection, the TF of one HCC sample (HCC-1366) was estimated to be 0, and the CNV signal was undetectable. Of note, after the fragment size enrichment, the TF of this sample was enhanced to 0.08, and the CNV signal could be readily detectable (Fig. 3B), demonstrating enriched ctDNA of HCC samples by utilizing the biomarker of fragment sizes.

For further analysis, we selected 43 HCC samples with ctDNA TF between 0 and 0.2 to investigate the strength of CNV signal alterations before and after <150 bp fragment selection. Although partial CNV signals could be observed prior to size selection, the CNV signals were markedly enhanced after fragment size enrichment (Fig. 3C). Interesting, we found that the CNV signals after size selection were highly consistent with the signal features in Fig. 1, representing commonly occurring aberrant CNVs that we identified in this study, e.g. chr1, chr4 and chr8 (Fig. 3C). Based on the findings that the selection of less than 150 bp fragment size was able to improve TF of ctDNA and theoretical hypothesis that the strong CNV signal was associated with the high TF of ctDNA, we wondered whether the CNV signals in the samples with lack of detectable CNVs could be detected through increasing TF after fragment size selection. As for this point, the other 44 HCC samples with low ctDNA proportions in cfDNA (TF=0) were chosen, and it was clear that CNV signals, including aberrant CNVs in chr1, chr4 and chr8, emerged after fragment
size selection (Fig. 3D). These findings demonstrated that using less than 150 bp fragments selection of cfDNA, coupled with the CNV aberration biomarkers identified in this study, was able to enhance HCC-specific CNV signals for cancer detection.

3.4 cfDNA end motifs determination in HCC and HBV samples

It has been reported that cfDNA fragmentation is a nonrandom course and there are a class of ctDNA signatures with preferred DNA ends (e.g. CCCA, CCTG, CCAG, TAAA, AAAA and TTTT) related to HCC [15, 21]. To investigate the end motif patterns in our data, we first counted the proportions of six representative end motifs (CCCA, CCTG, CCAG, TAAA, AAAA and TTTT), reported previously [15], in high-TF (TF>0.2) HCC, low-TF (0<TF≤0.2) HCC, and HBV groups. As shown in Fig. 4A, the proportions of three motifs CCCA, CCTG, and CCAG were higher in contrast to the other three motifs (TAAA, AAAA and TTTT) in all three groups. We observed significantly reduced proportions of the three motifs, i.e. CCCA (P=5.42e-07), CCTG (P=6.21e-08), and CCAG (P=1.25e-07), in high-TF HCC group vs. HBV group, which was consistent with a previous report [15], although the magnitudes of differences in our study were smaller than that report. In particular, with the increase of cfDNA TF in HCC samples, the proportions of end motifs CCCA, CCTG, and CCAG were significantly reduced (Fig. 4A). However, inconsistent with the previous report [15], we did not observe significant differences in the proportions of three motifs (TAAA, AAAA and TTTT) in HCC compared to HBV samples, regardless of increasing the TF of the HCC samples. These data demonstrated that the end motifs CCCA, CCTG, and CCAG were HCC-associated cfDNA preferred ends, while it was inconclusive whether or not TAAA, AAAA and TTTT were associated with ctDNA fragment ends, and the clarification of which requires future investigation.

Having established that both fragment sizes and end motifs are characteristics of ctDNA, and realizing that both may be involved in the ctDNA cleavage process, we wondered whether the two characteristics are independent events. To this end, we investigated the fragment size distribution.
of cfDNA sequencing reads, separately for reads containing each of these six end motifs. We found that the fragment size distributions were different for different end motifs, for both HCC and HBV samples (Fig. 4B). In particular, the difference was more profound in fragment sizes of <150bp, corresponding to the fragment sizes that were strongly associated with ctDNA (Fig. 4B). Among these six motifs, reads with end motif of AAAA were enriched in shorter reads of <150bp, while reads with the CCAG motif were opposite (Fig. 4B). We found that the fragment size distributions of reads with these end motifs in HBV samples were similar to that in HCC samples (Fig. 4B), indicating that the association of fragment size with end motifs was not unique to HCC.

To further quantify the magnitude of the association for different motifs, we investigated the effects of fragment size selection on the proportions changes of these six end motifs, for both HCC and HBV samples, respectively. The results demonstrated that the proportions of end motifs CCCA, CCAG, and CCTG were consistently decreased in HCC and HBV samples after fragment size selection of <150bp, compared to that before fragment size selection (Fig. 4C). Conversely, the proportions of end motifs TAAA, AAAA, and TTTT were consistently increased in HCC and HBV after fragment size selection (Fig. 4C). These findings were consistent with results in Fig. 4A and Fig. 4B, and could be explained by a larger fragment size of end motifs CCCA, CCAG, and CCTG than end motifs TAAA, AAAA, and TTTT. Of particular importance, we observed that the change patterns of the proportions of these six end motifs in HCC and HBV samples after size selection were largely indistinguishable, suggesting that the coupling of fragment size and end motifs is likely to be a general mechanism involved in the cfDNA cleavage process, rather than a process specific to HCC (Fig. 5).

In addition to the above six representative 4-mer end motifs, we were interested in comprehensively investigating the proportions of all 256 4-mer cfDNA end motifs in reads of HCC and HBV samples before and after fragment size selection, respectively. We found that the proportions of 139 4-mer cfDNA end motifs were significantly changed after fragment size selection, and the changes were consistent in both HCC and HBV samples, compared to those
prior to fragment size selection (Fig. S1), indicating the complex coupling of fragment sizes and end motifs. Among the significantly associated motifs, the proportions of 69 end motifs were significantly increased and 70 end motifs were significantly decreased after fragment size selection, in both HCC and HBV samples (Table S1). The top 10 most significant (smallest $P$ values) end motifs are listed in Table 2. Interestingly, 80% (8/10) of the top 10 end motifs (e.g., AAAA, GAAT, GAAA, and TCCA) were shared in both HCC and HBV samples. We did not, however, find the proportion changes of any end motif in HCC samples after fragment size selection that were different from that in HBV samples.

In consideration of the nonrandomness of cfDNA fragmentation, we wondered whether there are some finer-scale relationships between fragment size of cfDNA and end motifs. To that end, we explored the clustering of end motifs and the fragment sizes. A total of 70 motifs with their proportions greater than 0.005 in both HCC and HBV samples were selected for this analysis. The heat map demonstrated that there were groups of motifs that were associated with shorter fragment sizes, while some motifs were clustered with longer fragments. For example, AAAA, TTTT, AAAT, TATT, CATT, GAAA, GAAT, and TCCA were consistently coupled with fragments shorter than 150bp, in both HCC and HBV samples, and end motifs (e.g. CCAA, CCAG, CTGA, GGAG, CCAC, TGGG, CAGA, CAGG, AGAG, TGAG, CAAG, and GGGA) were consistently coupled with fragments larger than 150bp, in both HCC and HBV samples (Fig. 4D, Fig.5, and Fig. S2-3). The strong coupling of groups of motifs with either short or long fragments suggests a mechanism in which these motifs participate in the fragmentation of cfDNA.

4. Discussion
In this study, we comprehensively investigated the fragmentomics features of cfDNA derived from WGS data of HCC and HBV, as well as the CNV aberrations that could be used for cancer detection. We identified commonly occurring CNVs alterations in HCC, and in particular, we identified five typical CNVs as biomarkers for HCC detection. We found the positive rate of the
five CNVs was higher in HCC samples with high TF compared with that with low TF, and abnormal CNVs might be considered as preclinical signals to boost the early detection of HCC. In addition, our result showed that the proportion of fragments with length less than 150bp was obviously higher in HCC samples with high TF than that in low TF and HBV samples, and the size less than 150bp selection was not only able to enhance the TF to improve the clinical utility of ctDNA detection, but also provide a new strategy for detecting CNVs more accurately for HCC samples with lower CNVs signals. For motif analyses, three 4-mer end motifs (CCCA, CCTG, and CCAG) were identified as preferred end motifs of HCC, while we were not able to replicate the HCC-associated motifs (TAAA, AAAA and TTTT) reported previously [15]. Our study was the first, to our knowledge, to investigate the relationship of fragment sizes and 4-mer end motifs in HCC and HBV, and identify a group of 139 end motifs that were significantly associated with fragment sizes. In addition, we found that the size characteristics of these end motifs in HCC samples were similar to that in HBV samples.

CNVs detected by next-generation sequencing account for important types of genomic abnormalities in cancers, from sub-microscopic events to complete chromosomal aneuploidies [22, 23]. Several most common chromosomal arm alterations in both HCC and HBV-infected HCC tissue specimens have been revealed, including loss in chr1p, chr8p and chr17p, and gain in chr1q and chr8q [24, 25]. Interesting, chromothripsis that refers to a phenomenon with a large number of rearrangements clustered in a chromosomal region may drive chr 1q and 8q amplifications to contribute to hepatocarcinogenesis [26]. In addition, significant changes of CNVs (gain in chr1q, chr7q, and chr19q, and loss in chr1p, chr9q, and chr14q) are detected in blood samples of HCC with chronic liver diseases [27]. Consistently, in this work, some typical CNVs changes comprising of loss in chr1p, chr4q and chr8p, and gain in chr1q and chr8q were frequently detected in HBV-HCC blood samples. Of note, 2 of 187 HBV subjects exhibited abnormal CNVs and were diagnosed with HCC several months later. It is possible that abnormal CNVs may be considered as preclinical signals to boost the early detection of HCC. Notably, our results
regarding the landscape of genomic CNVs of HCC based on liquid biopsy were well consistent with the corresponding results derived from tissue-based analyzed by zhang et al [25], suggesting that our data were reliable and WGS sequencing of cfDNA was able to detect CNVs harbored in ctDNA. As the five CNV aberrations were derived from cfDNA in HCC samples, the high coverage (~95% of cfDNA samples from HCC patients) of these CNV biomarkers provides promising targets for early detection of HCC.

In our work, the cfDNA size distribution plot of HCC and HBV samples showed a prominent peak at 167 bp. This length is in close proximity in the length of DNA wrapped around a nucleosome and its linker [28]. This fragmentation pattern may be resulted from liberation of cfDNA via cell apoptosis and necrosis, during which histones complex binding to nuclear DNA acts as a main degradation type to protect DNA from cleavage [29, 30]. In addition, we found cfDNA fragments less than 150bp were more likely to enrich in HCC samples compared with that in HBV samples, which was similar to previous findings in animal and human cancers [14, 31]. Of note, we found the proportion of lengths less than 150bp was obviously higher in HCC samples with high TF than in HCC samples with low TF and HBV samples, supported by findings that shorter fragments were enriched in HCC patients with higher levels of ctDNA [32]. These findings motivated us to select cfDNA fragment size < 150bp to enrich ctDNA of HCC samples and observe the CNV signals. As a result, tumor purity of cfDNA of HCC samples was enhanced approximately 6.5% by selecting the fragment size of <150bp. Similarly, it has been reported that the selection of 90–150bp fragment size may improve more than 2-fold enrichment of tumor DNA in larger than 95% cases [20], and on average 44% greater of CNAs was detectable in cancer patients after selecting fragments with sizes up to 142bp (+/-15bp) [33].

Evidence has showed that cfDNA has a higher predilection of specific end motif sequence as a result of being cut at specific genomic regions or elements [34]. A class of cfDNA with preferred ends is found to be selectivity associated with fetal- or maternal-derived DNA, and the number
ratio of fetal preferred ends to those with maternal preferred ends in maternal plasma presents a correlation with the fetal DNA fraction [35]. Similarly, another key observation of DNA end characteristics in HCC is that there are 5.4 million HCC associated cfDNA preferred ends and 4.4 million preferred ends shared in HCC and chronic hepatitis B patients [15]. In addition, they suggest the abundance of HCC-related DNA is associated with DNA TF. Recently, Jiang et al have firstly identified a group of 4-mer end motifs showing significant differences between HCC and non-HCC subjects, in which the frequencies of three motifs (CCCA, CCAG, and CCTG) are significantly decreased in HCC subjects, and the frequencies of three motifs (TAAA, AAAA, and TTTT) are significantly increased in HCC subjects, as compared to non-HCC subjects [21]. However, in our work, no significant difference of frequencies of motifs (TAAA, AAAA and TTTT) among high-TF HCC, low-TF HCC and HBV groups was detected. Although we found that the lower proportions of three end motifs (CCCA, CCTG, and CCAG) were significantly related with HBV-HCC and higher cfDNA TF, the magnitude of the differences among groups was not as dramatic as the results of Jiang et al [21]. The differences might be caused by different studied subjects. We only collected HCC with HBV infection samples, while Jiang et al focused on HCC subject without HBV infection. HBV-infected individuals are a group of high-risk populations to develop HCC and may also harbor driver mutations prevalent in HCC. Collectively, our results showed that the motifs CCCA, CCTG, and CCAG might be HCC-related preferred end motifs, while future researches are still needed to investigate the associations of motifs TAAA, AAAA and TTTT with HCC.

There is evidence to show short and long plasma DNA molecules are associated with different preferred DNA end sites [36]. In our work, we investigated whether there were some motifs among the 256 4-mer end motifs related with fragments size specific in HCC. We found that the proportions of 69 4-mer end motifs (eg., TAAA, AAAA, and TTTT) were significantly increased and 70 4-mer end motifs (eg., CCCA, CCTG, and CCAG) were significantly decreased in both HCC and HBV samples after the selection of a fragment size of
less than 150bp. Notably, the proportion changes of any end motif after fragment size selection in HCC samples were not found to be different from that in HBV samples. Clustering analysis revealed a more pervasive coupling of end motifs with fragment sizes (Fig. 4D and Fig. 5). These findings together suggested that end motifs might be inherently coupled with fragment sizes, in both HCC and HBV, as a ubiquitous mechanism (Fig. 5). Further investigations are needed to dissect the mechanisms behind this coupling to reveal fine-scale characteristics of fragmentomic features to facilitate more powerful detection of HCC, as well as other cancer types.

5. Conclusions

In summary, we identified representative abnormal CNV alterations related to HCC and found selecting fragment sizes shorter than 150bp was an effective strategy for detecting CNVs more accurately and improving the clinical utility of ctDNA detection of HCC samples. The study discovered strong coupling of end motifs with fragment sizes, and revealed similar fragment size characteristics of 4-mer end motifs between HCC and HBV subjects.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grants 81830070) and the Open Project Program of the State Key Laboratory of Cancer Biology (grants CBSKL2019ZDKF05 and CBSKL2019ZDKF10).

Conflict of interest

WYZ, HPL, BX, JYY, and QZ are employees of Oriomics Biotech Inc. No other disclosures were reported.

Data Availability statement

The data that support the findings of this study are openly available in BIG Sub at https://bigd.big.ac.cn/gsub/, accession number subCRA00289, and other data generated during this

Molecular Oncology (2020) © 2020 The Authors. Published by FEBS Press and John Wiley & Sons Ltd.
study are included in Supporting information file.

**Author contributions**

CJ, XNL, WYZ, JYL, and JLX conceived and supervised the study. LPS, JYY, and QZ conducted data collection, and WYZ, YL, and BX participated in the data analysis. CJ, and XNL wrote the first draft of the manuscript. JYL, JLN, XG, XMG, HPL, GW, DKB, and SGW provided main edits of manuscript, and JYL and JLX had responsibility for final content. All authors read and approved the final manuscript.
References

1 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424.

2 Wild C, Weiderpass E, Stewart B (2020) World cancer report: cancer research for cancer prevention. *Lyon: International Agency for Research on Cancer*, 1-613.

3 Levrero M, Zucman-Rossi J (2016) Mechanisms of HBV-induced hepatocellular carcinoma. *J Hepatol* **64**, S84-s101.

4 Sarin SK, Kumar M, Lau GK, Abbas Z, Chan HL, Chen CJ, et al. (2016) Asian-Pacific clinical practice guidelines on the management of hepatitis B: a 2015 update. *Hepatol Int* **10**, 1-98.

5 Jain D (2014) Tissue diagnosis of hepatocellular carcinoma. *J Clin Exp Hepatol* **4**, S67-S73.

6 Gorgannezhad L, Umer M, Islam MN, Nguyen NT, Shiddiky MJA (2018) Circulating tumor DNA and liquid biopsy: opportunities, challenges, and recent advances in detection technologies. *Lab on a chip* **18**, 1174-96.

7 Kaseb AO, Sánchez NS, Sen S, Kelley RK, Tan B, Bocobo AG, et al. (2019) Molecular Profiling of Hepatocellular Carcinoma Using Circulating Cell-Free DNA. *Clin Cancer Res* **25**, 6107-18.

8 Hlady RA, Zhao X, Pan X, Yang JD, Ahmed F, Antwi SO, et al. (2019) Genome-wide discovery and validation of diagnostic DNA methylation-based biomarkers for hepatocellular cancer detection in circulating cell free DNA. *Theranostics* **9**, 7239-50.

9 Cai Z, Chen G, Zeng Y, Dong X, Li Z, Huang Y, et al. (2019) Comprehensive Liquid Profiling of Circulating Tumor DNA and Protein Biomarkers in Long-Term Follow-Up Patients with Hepatocellular Carcinoma. *Clin Cancer Res* **25**, 5284-94.

10 Chan RWY, Serpas L, Ni M, Volpi S, Hiraki LT, Tam LS, et al. (2020) Plasma DNA Profile Associated with DNASE1L3 Gene Mutations: Clinical Observations, Relationships to Nuclease Substrate Preference, and In Vivo Correction. *Am J Hum Genet* **107**, 882-94.

11 Jiang P, Xie T, Ding SC, Zhou Z, Cheng SH, Chan RWY, et al. (2020) Detection and Molecular Oncology (2020) © 2020 The Authors. Published by FEBS Press and John Wiley & Sons Ltd.
characterization of jagged ends of double-stranded DNA in plasma. *Genome Res* **30**, 1144-53.

12 Zhang Z, Zhang W (2019) Fragmentation patterns of circulating cell-free DNA demonstrate biomarker potential for human cancers. *Biotarget* **3**, 16.

13 Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385-9.

14 Mouliere F, Piskorz AM, Chandrananda D, Moore E, Morris J, Smith CG, et al. (2017) Selecting short DNA fragments in plasma improves detection of circulating tumour DNA. *bioRxiv*, 134437.

15 Jiang P, Sun K, Peng W, Cheng SH, Ni M (2020) Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discov* **10**, 664-73.

16 Tao K, Bian Z, Zhang Q, Guo X, Yin C, Wang Y, et al. (2020) Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumor DNA for early detection of hepatocellular carcinoma. *EBioMedicine* **56**, 102811.

17 Li B, Zhan X, Wing M-K, Anderson P, Kang HM, Abecasis GR (2013) QPLOT: a quality assessment tool for next generation sequencing data. *BioMed research international* **2013**, 865181-.

18 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

19 Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG (2017) Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324.

20 Mouliere F, Chandrananda D (2018) Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med* **10**.

21 Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, et al. (2018) Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc Natl Acad Sci U S A* **115**, E10925-e33.

22 Macé A, Kutalik Z, Valsesia A (2018) Copy Number Variation. *Methods Mol Biol* **1793**, Molecular Oncology (2020) © 2020 The Authors. Published by FEBS Press and John Wiley & Sons Ltd.
23 Shen W, Szankasi P, Durtschi J, Kelley TW, Xu X (2019) Genome-Wide Copy Number Variation Detection Using NGS: Data Analysis and Interpretation. *Methods Mol Biol* **1908**, 113-24.

24 Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu CGARN (2017) Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-41.e23.

25 Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, et al. (2013) Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* **23**, 1422-33.

26 Fernandez-Banet J, Lee NP, Chan KT, Gao H, Liu X, Sung WK, et al. (2014) Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma. *Genomics* **103**, 189-203.

27 Xu H, Zhu X, Xu Z, Hu Y, Bo S, Xing T, et al. (2015) Non-invasive Analysis of Genomic Copy Number Variation in Patients with Hepatocellular Carcinoma by Next Generation DNA Sequencing. *J Cancer* **6**, 247-53.

28 Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* **2**, 61ra91.

29 Mouliere F, Rosenfeld N (2015) Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc Natl Acad Sci U S A* **112**, 3178-9.

30 Heitzer E, Auinger L, Speicher MR (2020) Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends Mol Med* **26**, 519-28.

31 Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN (2016) Fragment Length of Circulating Tumor DNA. *PLoS genetics* **12**, e1006162.

32 Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, et al. (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **112**, E1317-25.

33 Holden KA FK, McLennan G, Faulkner N, Guan X, Eisenberg M, Jensen TJ (2020) Size
selection of cell-free DNA increases the proportion of tumor specific variants in cancer patients. 

*AACR Annual Meeting June 22-24.*

34 Chiu RWK, Heitzer E, Lo YMD, Mouliere F, Tsui DWY (2020) Cell-Free DNA Fragmentomics: The New “Omics” on the Block. *Clin Chem* 66, 1480-4.

35 Chan KC, Jiang P, Sun K, Cheng YK, Tong YK, Cheng SH, et al. (2016) Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci U S A* 113, E8159-e68.

36 Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, et al. (2018) Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *Proc Natl Acad Sci U S A* 115, E5106-E14.
Figure legends

Fig. 1 Genome-scale overview of cfDNA copy number variation (CNV) in 63 HCC samples with HBV infection detected by WGS
X axis represents 1-22 and X chromosomes and each line in Y axis represents one HCC sample. CNVs amplifications are presented in red, while CNVs deletions are presented in blue. The darker the color is, the greater the amplitudes/deletions are.

Fig. 2 A distinct difference of fragment lengths in HCC and HBV samples
A: Fragment size density distribution of DNA fragments of 20 HCC samples with TF>0.2 and 187 HBV samples. X axis represents the length of fragments, and Y axis represents the proportion of fragments with the corresponding fragment size. HCC samples are indicated by blue, and HBV samples are revealed by red. B: Box plots of the mean proportions of cfDNA fragments less than 150 bp in all fragment sizes from 20 HCC samples with TF>0.2, 43 HCC samples with 0<TF≤0.2, and 187 HBV samples. The mean proportions of cfDNA fragments less than 150 bp among three groups were compared using one-sized t-tests. Orange line represent mean and whiskers represent range. TF: tumor fraction.

Fig. 3 cfDNA fragment selection increases TF and CNV signals
A: The effect of selecting fragments shorter than 150 bp on tumor purity of HCC samples using linear least-squares regression. The tumor purities of each sample after size selection (Y axis) and before size selection (X axis) are shown. B: CNV profiles of one HCC sample (HCC-1366) before (at the bottom) and after fragment size enrichment (on the top). CNVs amplifications are presented in red, while CNVs deletions are presented in green. The darker the color is, the greater the amplitudes/deletions are. C: CNV profiles of 43 HCC samples with ctDNA TF between 0 and 0.2 to show the CNV signal alterations before (on the top) and after selecting <150 bp fragments (at the bottom). D: CNV profiles of 44 HCC samples with low ctDNA proportion in cfDNA (TF=0) to show the CNV signal alterations before (on the top) and after <150 bp fragment selection (at the bottom).
bottom). CNVs amplifications are presented in red, while CNVs deletions are presented in blue. The darker the color is, the greater the amplitudes/deletions are.

Fig. 4 cfDNA end motifs determination and the association of fragment sizes and end motifs in HCC (TF>0, N=63) and HBV samples (N=187)

A: Box plot analysis of differential frequencies of six representative motifs (CCCA, CCTG, CCAG, TAAA, AAAA and TTTT) among HCC-high TF (TF>0.2) (N=20), HCC-low TF (0<TF≤0.2) (N=43) and HBV groups (N=187). The comparisons among three groups were analyzed by one-sized t-tests. Orange line represent mean and whiskers represent range. * indicates p<0.05 and *** indicates p<0.001. B: Fragment size distributions of six motifs from HCC samples with TF>0.2 (on the top) and HBV samples (at the bottom). C: The effect of selection of fragments shorter than 150bp on motif proportions using linear least-squares regression. The proportions of each motif in all cfDNA reads after size selection (Y axis) and before size selection (X axis) are shown. Dark blue dots represent HCC samples, while orange dots represent HBV samples. D: Heat map analysis of the association of fragment sizes and 70 motifs with their proportions greater than 0.005 from HCC_highTF (TF>0.2) samples. X axis represents the length of fragments, and Y axis represents the proportion of the corresponding motif. Each matrix represents the mean proportion of reads count of the corresponding end motif from all samples, and the data is revised by z score.

Fig. 5 The inherently coupling of cfDNA fragment size and end motifs that may be associated with cfDNA cleavage process, in both HCC and HBV samples, as a ubiquitous mechanism.
Supplementary information

**Table S1** The significance of proportion changes of 139 4-mer cfDNA end motifs in HCC and HBV samples, respectively, after fragment size selection.

**Fig. S1** The effect of selection of fragments shorter than 150 bp on the proportion changes of 256 4-mer end motifs using linear least-squares regression.

**Fig. S2** Heat map analyses of the associations of fragment size and 70 motifs with their proportions greater than 0.005 from low TF (0<TF≤ 0.2) HCC samples.

**Fig. S3** Heat map analyses of the association of fragment size and 70 motifs with their proportions greater than 0.005 from HBV samples.

**Supplementary figure legends**

**Fig. S1** The effect of selection of fragments shorter than 150 bp on the proportion changes of 256 4-mer end motifs using linear least-squares regression. The proportions of each motif in all cfDNA reads after size selection (Y axis) and before size selection (X axis) are shown. Dark blue dots represent HCC samples, while orange dots represent HBV samples.

**Fig. S2** Heat map analyses of the associations of fragment size and 70 motifs with their proportions greater than 0.005 from low TF (0<TF≤ 0.2) HCC samples. X axis represents the length of fragment, and Y axis represents the proportion of corresponding motif. Each matrix represents the mean proportion of reads count of the corresponding end motif from all samples, and the data is revised by z score.
Fig. S3 Heat map analyses of the association of fragment size and 70 motifs with their proportions greater than 0.005 from HBV samples. X axis represents the length of fragment, and Y axis represents the proportion of corresponding motif. Each matrix represents the mean proportion of reads count of the corresponding end motif from all samples, and the data is revised by z score.
| Subject category | chr1p | chr1q | chr4q | chr8p | chr8q | Any of chr1p/1q/4q/8p/8q |
|------------------|-------|-------|-------|-------|-------|-------------------------|
| HCC (n=63)       | 26/63 | 50/63 | 44/63 | 31/63 | 44/63 | 59/63 (93.7%)           |
| HCC-TF high (n=19) | 11/19 | 13/19 | 15/19 | 17/19 | 12/19 | 18/19 (94.7%)           |
| HBV (n=187)      | 0/187 | 2/187 | 0/187 | 0/187 | 2/187 | 2/187 (1.07%)           |

Chr, chromosome.
Table 2 Top 10 most significant (smallest $P$ values) end motifs showing proportion changes in HCC and HBV samples after fragment size selection

| Motif | HCC Mean | HCC Difference | HCC $P$ value |
|-------|----------|----------------|---------------|
| TTCC  | 0.001734 | 0.759349       | 5.50E-136     |
| GAAT  | 0.005042 | 0.299973       | 3.59E-110     |
| ATGG  | 0.001753 | 0.496929       | 1.24E-104     |
| GAAAA | 0.007313 | 0.203002       | 1.87E-98      |
| AAAA  | 0.009739 | 0.273082       | 4.14E-92      |
| TCCA  | 0.005373 | 0.181188       | 1.57E-80      |
| AATG  | 0.003325 | 0.272515       | 1.90E-80      |
| ATTC  | 0.001841 | 0.375516       | 5.94E-77      |
| AATT  | 0.002967 | 0.280476       | 8.92E-71      |
| GGAC  | 0.003352 | -0.10491       | 4.27E-65      |

| Motif | HBV Mean | HBV Difference | HBV $P$ value |
|-------|----------|----------------|---------------|
| GAAAA | 0.00734  | 0.216998       | 8.98E-153     |
| GAAT  | 0.005054 | 0.314193       | 5.12E-152     |
| TTCC  | 0.001652 | 0.82927        | 7.68E-141     |
| ATGG  | 0.001693 | 0.531796       | 6.93E-140     |
| AAAA  | 0.009874 | 0.289954       | 7.95E-117     |
| CGTA  | 0.000953 | -0.13613       | 3.05E-116     |
| TCCA  | 0.005338 | 0.187821       | 3.17E-113     |
| GATT  | 0.003373 | 0.228652       | 7.96E-110     |
| ATTC  | 0.001849 | 0.380833       | 3.90E-107     |
| AATG  | 0.003363 | 0.27518        | 5.70E-107     |
mol2_13041_f2.tif
mol2_13041_f5.tif