Genomic origin, fragmentomics, and transcriptional correlation of long cell-free DNA molecules in human plasma

Huiwen Che, Peiyong Jiang, Lois L Y Choy, et al.

Genome Res. published online February 26, 2024
Access the most recent version at doi:10.1101/gr.278556.123

Genomic origin, fragmentomics and transcriptional correlation of long cell-free DNA molecules in human plasma

Running title: Origin and characteristics of long cfDNA

Huiwen Che[1,2,3], Peiyong Jiang[1,2,3,4], L.Y. Lois Choy[1,2,3,4], Suk Hang Cheng[1,2,3], Wenlei Peng[1,2,3], Rebecca W.Y. Chan[1,2,3], Jing Liu[1,2,3], Qing Zhou[1,2,3], W.K. Jacky Lam[1,2,3,4], Stephanie C.Y. Yu[1,2,3], So Ling Lau[5], Tak Y. Leung[5], John Wong[6], Vincent Wai-Sun Wong[7], Grace L.H. Wong[7], Stephen L. Chan[4,8], K.C. Allen Chan[1,2,3,4], Y.M. Dennis Lo[1,2,3,4,#]

AFFILIATIONS

[1]Centre for Novostics, Hong Kong Science Park, Pak Shek Kok, Hong Kong SAR, China.

[2]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

[3]Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

[4]State Key Laboratory of Translational Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China.

[5]Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

[6]Department of Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China.

[7]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China.

[8]Department of Clinical Oncology, Sir Y.K. Pao Centre for Cancer, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China.

CORRESPONDENCE

[#]Correspondence to Prof. Y.M. Dennis Lo

Address: Centre for Novostics, 1/F, 18 Science Park East Ave., Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong SAR, China.

Email: loym@cuhk.edu.hk

ABSTRACT

Recent studies have revealed an unexplored population of long cell-free DNA (cfDNA) molecules in human plasma using long-read sequencing technologies. However, the biological properties of long cfDNA molecules (> 500 bp) remain largely unknown. To this end, we investigated the origins of long cfDNA molecules from different genomic elements. Analysis of plasma cfDNA using long-read sequencing revealed uneven distribution of long molecules from across the genome. Long cfDNA molecules showed overrepresentation in euchromatic regions of the genome, in sharp contrast to short DNA molecules. We observed a stronger relationship between the abundance of long molecules and mRNA gene expression levels, compared with short molecules (Pearson's r = 0.71 *versus* -0.14). Moreover, long and short molecules demonstrated distinct fragmentation patterns surrounding CpG sites. Leveraging the cleavage preferences surrounding CpG sites, the combined cleavage ratios of long and short molecules could differentiate patients with hepatocellular carcinoma (HCC) from non-HCC subjects (AUC = 0.87). We further investigated knockout mice in which selected nuclease genes had been inactivated, in comparison with wild-type mice. The proportion of long molecules originating from transcription start sites were lower in *Dffb*-deficient mice but higher in *Dnase1l3*-deficient mice, compared to that of wild-type mice. This work thus provides new insights into the biological properties and potential clinical applications of long cfDNA molecules.

INTRODUCTION

Fragmentation patterns of cell-free DNA (cfDNA) are demonstrated to be linked to a myriad of biological characteristics, including nuclease activities (Serpas et al. 2019; Han et al. 2020), DNA methylation (Zhou et al. 2022), nucleosome structures (Ivanov et al. 2015; Snyder et al. 2016; Sun et al. 2019), mRNA expression levels (Ulz et al. 2016), and DNA-binding transcription factor activity (Ulz et al. 2019). These characteristics have spurred much research efforts in understanding the underlying biological mechanisms. For example, size analysis of cfDNA molecules reflects characteristic patterns of cfDNA fragmentation. Through short-read sequencing (Illumina), a typical plasma cfDNA fragment size distribution could be depicted with a dominant peak around 166 base pairs (bp), with 10-bp periodicities below 150 bp, suggesting a nucleosomal origin (Lo et al. 2010; Ivanov et al. 2015). Thus, cfDNA is generally believed to be a population of short DNA molecules. Moreover, cfDNA fragments from different tissues may bear information tracing to their tissues of origin. As examples, fetal cfDNA molecules in pregnant women and tumoral cfDNA molecules in patients with cancer are generally shorter than background hematopoietic cell-derived cfDNA (Chan et al. 2004; Lo et al. 2010; Jiang et al. 2015). These findings have catalyzed the use of size information of cfDNA for disease detection (Yu et al. 2014; Jiang et al. 2015; Mouliere et al. 2018).

Single-molecule sequencing technologies, including Pacific Biosciences single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) sequencing, have opened up the possibility of detecting and characterizing long cfDNA molecules. Recent studies using these platforms have uncovered a population of long cfDNA molecules up to tens of kilobases, in the plasma DNA of healthy, pregnant individuals and patients with cancer (Yu et al. 2021; Choy et al. 2022; Yu et al. 2023b, 2023a). Depending on the sequencing platforms used, long molecules account for about a median of 15% and 5% of the total molecules from SMRT and ONT platforms, respectively (Yu et al. 2023b). Hence,

single-molecule, long-read sequencing technologies appear to be capable of sequencing a much wider spectrum of cfDNA molecules.

Studies have investigated the quantity, size distribution, methylation patterns and end motifs of cfDNA molecules generated by long-read sequencing (Yu et al. 2021; Choy et al. 2022; Katsman et al. 2022; Lau et al. 2023). As single-molecule sequencing technologies are able to measure methylation patterns (Tse et al. 2021; Lau et al. 2023), one can use methylation patterns to explore the tissues of origin of long cfDNA molecules. Additional biological properties of long molecules, however, remain largely unexplored. Hence, in this study, we investigated whether long and short cfDNA molecules might originate from different genomic elements, including euchromatin and heterochromatin. We aggregated cfDNA data generated by SMRT or ONT sequencing to profile the genomic representations and studied the correlation between long cfDNA molecules and gene transcriptional activity. We examined end frequencies of long molecules derived from regulatory regions such as DNase I hypersensitive sites (DHSs) and CCCTC-binding factors (CTCF). Moreover, we analyzed cleavage profiles surrounding cytosine-phosphate-guanine sites (CpGs) to investigate long cfDNA fragmentomics in the context of cancer detection. We generated additional cfDNA data from various nuclease-knockout mice using SMRT sequencing to gain biological insights regarding the role of nucleases in the generation of long cfDNA molecules.

RESULTS

Long cfDNA molecules originate unevenly across the human genome

A preponderance of molecules with 5' end-nucleotide of base adenine (A), namely A-end, has been observed in cfDNA molecules longer than 500 bp (Yu et al. 2021). We therefore defined long cfDNA molecules as those of > 500 bp. We investigated whether long cfDNA molecules originated evenly across the genome. By pooling sequenced data from previous studies (Yu et al. 2021; Choy et al. 2022; Yu et al. 2023b) using the SMRT and ONT platforms (**Supplemental Table S1**), respectively, we first compared the genomic representation profiles of long (>500 bp) and short (<= 500bp) molecules, across 100-kilobase (kb) non-overlapping bins (Methods). Both genomic origins of long and short molecules displayed unevenly along the whole genome. We observed differential genomic representations between long and short molecules in some regions, with Chromosome 10 representation shown as an example (**Fig. 1; Supplemental Fig. S1**). The differential signals in genomic representation were found in both SMRT (**Fig. 1A**) and ONT (**Fig. 1B**) sequencing data, with platform-specific differences. Specifically, we observed that long molecules tended to be overrepresented in regions that overlapped with light bands on Giemsa-stained chromosomes and underrepresented in dark-banded regions shown in ideograms (**Fig. 1**). The light and dark bands typically correspond to GC-rich euchromatic and AT-rich heterochromatic regions, respectively (Bickmore and Sumner 1989; BAC Resource Consortium et al. 2001). Thus, long molecules were preferentially derived from euchromatic regions compared to short molecules. This differential representation was particularly prominent on Chromosome 19 (**Supplemental Fig. S2**), which is well-known for high GC-content and considered as the most gene-rich chromosome (Grimwood et al. 2004).

As euchromatin generally tends to be gene-rich, we wondered if the gene density of a genomic region might positively correlate with the difference between long and short cfDNA molecules from that region. Indeed, the difference was significantly correlated to autosomal gene density (Pearson's r = 0.56, $P < 2.2 \times 10^{-16}$) and the gene-rich Chromosome 19

showed a correlation of 0.7 using non-pregnant controls from SMRT sequencing data. The correlation to Chromosome 19 gene density was 0.76 ($P < 2.2 \times 10^{-16}$) for pregnant samples using ONT sequencing data. We further examined the potential relationship between genomic representation of cfDNA molecules and presence of DNA double-strand breaks. Previous studies have found that transcribed genes are hotspots for endogenous double-strand breaks (Crosetto et al. 2013; Lensing et al. 2016; Ballarino et al. 2022). A weak positive correlation (Pearson's r = 0.43, $P < 2.2 \times 10^{-16}$; **Supplemental Fig. S3**) between overrepresentation of long molecules and double-strand breaks detected from a lymphoblastoid cell line sample was observed (Methods).

Association between long molecules and gene expression in human samples

Motivated by the observation that long molecules appeared to be enriched in euchromatin that is associated with active transcription, we assessed the potential relationship between abundance of such molecules and transcriptional activity. First, we analyzed the abundance of cfDNA molecules originating from unexpressed and housekeeping genes (Methods). Long cfDNA molecules showed higher abundance over housekeeping genes and lower abundance over unexpressed genes in non-pregnant controls and pregnant subjects (**Fig. 2A, B**). We further examined the abundance of cfDNA molecules originating from gene bodies of differentially expressed genes. Autosomal protein-coding genes were ranked based on their median expression across tissues and grouped into five sets, from a lowly expressed EXP1 to a highly expressed EXP5 set. In line with the enrichment in housekeeping genes, the analysis of non-pregnant controls revealed a stepwise increase of long cfDNA molecules abundance as expression levels increased from EXP1 to EXP5 (P = 0.01374, Mann-Kendall trend test; Pearson's r = 0.78), while no such trend was found in short molecules (P = 0.4032, Mann-Kendall trend test; Pearson's r = -0.28; **Supplemental Fig. S4A**). We confirmed the observation that long molecule abundance was positively associated with gene expression from SMRT in pregnancies (**Supplemental Fig. S4B**) and

ONT sequencing data (**Supplemental Fig. S4C, D**). Analyzing the relationship at a higher resolution, we showed that the median abundances of long molecules were positively correlated (Pearson's r = 0.71, $P < 2.2 \times 10^{-16}$) with median transformed expression scores across tissues based on 15,556 expressed genes. In contrast, such positive correlation was not observed in short molecules (Pearson's r = 0.14, P = 0.003; **Fig. 2C**). The conclusion was further confirmed by the ONT data (**Fig. 2D**). Moreover, investigating overall methylation levels revealed lower methylation in long molecules compared with short molecules (**Fig. 2E**, **F**).

As the abundance of long cfDNA molecules in human plasma demonstrated a relationship with gene expression, we explored whether such signature could be used to distinguish patients with hepatocellular carcinoma (HCC) from subjects without HCC. Thus, we collected SMRT sequencing data using plasma samples from 20 healthy individuals, 19 hepatitis B virus (HBV) carriers and 48 patients with HCC (**Supplemental Table S2**). Median values of 1,021,412, 286,492, and 712,223 high-quality circular consensus sequencing reads were obtained for healthy individuals, HBV carriers and patients with HCC, respectively. Long cfDNA molecules accounted for a median of 21.3%, 19.7% and 23.7% of total molecules in healthy subjects, HBV carriers and patients with HCC, respectively. As the number of molecules was relatively low, we attempted to examine a group of genes. The top 5000 expressed genes in HCC tumor from The Cancer Genome Atlas dataset (Methods) were identified. We compared the abundance of long and short molecules over these HCC-associated genes. Long molecules showed significantly higher ($P = 5.814 \times 10^{-5}$, Kruskal-Wallis test) abundance in patients with HCC compared to non-HCC groups, whereas short molecules did not exhibit significant differences (P = 0.06504, Kruskal-Wallis test) in this dataset (**Fig. 3A**). A receiver operating characteristic (ROC) curve based on the abundance of long molecules yielded an area under the curve (AUC) of 0.76 to distinguish 39 non-HCC individuals and 48 patients with HCC. When the number of total molecules increased to 1

million for each sample, the AUC improved to 0.9 to distinguish 12 non-HCC individuals and 18 patients with HCC (**Fig. 3B**).

In addition to the abundance of molecules, we assessed molecule ends in specific regions. End frequencies of cfDNA at transcription start sites (TSSs) and regulatory regions that include DHSs and CTCF binding sites (Methods) were examined. As the expression level rose, the normalized end frequencies of long molecules gradually increased in the vicinity of TSSs (e.g. 50-bp upstream and downstream of TSSs), rising from 1.01-fold in EXP1 to 1.7-fold in EXP5. Conversely, for short molecules, the normalized end frequencies had a decreasing tendency (**Fig. 4A**; **Supplemental Fig. S4E**). Long molecules pooled from the non-pregnant controls displayed higher end frequencies near the peak of tissue-invariant DHSs (**Fig. 4B**). We observed that the end frequencies of long cfDNA molecules were two times that of short molecules in DHSs (Methods). Analyzing end frequencies near tissue-specific DHSs revealed effects of tissue types on long cfDNA fragmentations. Specifically, at liver-specific and myeloid-specific DHSs, long cfDNA in non-pregnant controls tended to have stronger end preferences (1.4-fold) comparing to short cfDNA (**Supplemental Fig. S4F**). As a confirmation, the same pattern around tissue-invariant DHSs was observed in plasma of pregnant women from ONT sequencing (**Supplemental Fig. S4G**). At CTCF binding sites, long molecules exhibited higher cfDNA end frequencies as well, in both SMRT and ONT sequencing data (**Fig. 4C**; **Supplemental Fig. S4H**). These results have further highlighted that long molecules were preferentially derived from open chromatin regions and their abundance was correlated with transcriptional activities.

Differential cleavage profiles between long and short molecules surrounding CpGs

Studies have demonstrated that the cfDNA cleavage preferences occurred at methylated CpG sites (Han et al. 2021; Zhou et al. 2022). Beyond the fragmentation around regulatory regions, we wondered whether fragmentation of long molecules at CpGs would be different from those of short molecules. We analyzed the cleavage patterns at commonly methylated

and unmethylated CpGs in pooled healthy data (Methods). The cleavage of short molecules recapitulated earlier findings that nuclease-mediated preferential cleavage occurs at methylated CpGs in healthy subjects analyzed by Illumina sequencing (Zhou et al. 2022). Specifically, the cfDNA cleavage was relatively enriched at the cytosine of methylated CpGs (position 0), followed by a rapid decrease at the nucleotide 1 position immediately preceding methylated CpGs. We found that the cleavage of long molecules did not show such a pattern at methylated cytosines but exhibited the preferred cutting at positions several nucleotides away from the methylated CpGs (e.g. positions -4, -2 and 4) (**Supplemental Fig. S5A**).

To explore whether the cleavage profile around CpGs for long molecules would be associated with diseases, we evaluated cleavage patterns using the HCC cohort data mentioned above. Considering the relatively low number of molecules, we evaluated cleavage patterns related to all autosomal CpGs. As the human genome is highly methylated, with 70%-80% of CpGs being methylated (Ziller et al. 2013), the cleavage profiles of short molecules showed a cutting preference at position 0 cytosine (**Fig. 5A**). Due to the low number of sequenced molecules, long molecules showed higher variability in cleavage profiles. The cleavages of long molecules, however, consistently showed significant higher proportions ($P = 2.6 \times 10^{-6}$, paired Wilcoxon test, *P* values adjusted by Benjamini-Hochberg method) than the adjacent downstream nucleotide at positions -4, -2, 1 and 4 in healthy subjects and HBV carriers (**Fig. 5B; Supplemental Fig. S5B**). The ratio of 5' CGN to NCG end motifs (CGN/NCG motif ratio) has been shown to inform methylation level and can be used to distinguish patients with HCC from non-HCC individuals using Illumina sequencing (Zhou et al. 2022). In this SMRT sequencing dataset, patients with HCC showed overall lower CGN/NCG motif ratios ($P = 2.081 \times 10^{-7}$, Kruskal-Wallis test) for short molecules. On the contrary, for long molecules, CGN/NCG motif ratios of patients with HCC showed no significant difference ($P = 0.05137$, Kruskal-Wallis test) among the three groups (**Fig. 5C**).

We further attempted to use the ratio between the molecules ending at positions -4, -2, 1 and 4 and those ending at position -1. Using this cleavage ratio, both long and short molecules displayed significant differences ($P = 3.893 \times 10^{-8}$ and $P = 8.036 \times 10^{-8}$, Kruskal-Wallis test) between patients with HCC and non-HCC subjects (**Fig. 5D**). As a result, long molecules alone achieved an AUC of 0.85 to distinguish 39 non-HCC individuals and 48 patients with HCC. Combining the ratio of long and short molecules, the AUC increased to 0.87 (**Fig. 5E**).

Investigation of long molecules generation using knock-out mice

The amount of cfDNA molecules with 5' ending in an A or G was reported to increase as the size of cfDNA molecules increased (Yu et al. 2021) and the nuclease DNA fragmentation factor subunit beta (DFFB) exhibited a preference to cut 5' to an A or G nucleotide (Han et al. 2020). We hypothesized that long molecules were predominantly generated by DFFB. To gain potential mechanistic insights on generating long cfDNA molecules in human plasma, we studied mouse models in which different nuclease genes had been knocked out. Mice of the C57BL/6 background were used. We sequenced 4 wild-type (WT), 5 *Dffb*$^{-/-}$, 5 *Dnase1*$^{-/-}$, 5 *Dnase1l3*$^{-/-}$, and 5 *Dnase1*$^{-/-}$ and *Dnase1l3*$^{-/-}$ double knockout mice samples, with each sample pooled from 3-5 mice subjects, using SMRT sequencing (**Supplemental Table S3**). We first analyzed the size profile of plasma cfDNA from these mice by pooling sequenced data of the same genotype. Compared with wild-type mice, the frequencies of long cfDNA molecules above 1 kb in *Dffb*$^{-/-}$ mice were consistently lower, whereas the corresponding frequencies in *Dnase1*$^{-/-}$ mice were higher, and such an increasing pattern was further enhanced in *Dnase1l3*$^{-/-}$ mice and mice with double deletion of *Dnase1* and *Dnase1l3* (*Dnase1*$^{-/-}$ & *Dnase1l3*$^{-/-}$) (**Fig. 6A**). For instance, the percentages of cfDNA molecules with a size of > 500 bp were 21.7%, 16.8%, 29.3%, 37.2%, and 45% in wild-type, *Dffb*$^{-/-}$, *Dnase1*$^{-/-}$, *Dnase1l3*$^{-/-}$, and *Dnase1*$^{-/-}$ & *Dnase1l3*$^{-/-}$ mice, respectively. The percentages of long molecules were significantly different ($P = 0.002053$, Kruskal-Wallis test) between different

types of mice, with *Dnase1l3*$^{-/-}$, and *Dnase1*$^{-/-}$ & *Dnase1l3*$^{-/-}$ mice higher than wild-type mice (**Supplemental Fig. S6A**). Meanwhile, the frequencies of short molecules with a size < 250 bp decreased in *Dnase1l3*$^{-/-}$ and *Dnase1*$^{-/-}$ & *Dnase1l3*$^{-/-}$ mice, with a relatively more pronounced reduction in mononucleosomal size (**Fig. 6B**). The data suggested that DFFB was at least in part responsible for the generation of long cfDNA molecules. The removal of DNASE1L3 and DNASE1 would lead to an enhanced effect in generating long cfDNA molecules, as DFFB might, in the absence of DNASE1L3 and DNASE1, act as a dominant nuclease in contributing cfDNA into circulation. The deletion of *Dffb* resulted in the reduction of long molecules originating from regions around TSSs which normally represent open chromatin states (**Fig. 6C**), further providing evidence that DFFB might be responsible for the patterns related to long cfDNA molecules. The increase of long cfDNA molecules in regions nearby TSSs was elevated in mice where DFFB functioned prominently by knocking out *Dnase1* and *Dnase1l3*. These patterns could be reproduced in open chromatin regions comprising DHSs and CTCF binding sites (**Supplemental Fig. S6B, C**). Moreover, the highly expressed genes tended to harbor more long molecules in *Dffb*-competent mice. However, such enrichment of long molecules was absent in *Dffb*-deficient mice (**Fig. 6D**).

DISCUSSION

This study explores previously unknown properties of long cfDNA molecules in human plasma. First, long cfDNA molecules have been found to be preferentially originated from transcriptionally active regions of the genome, with overall lower methylation and the end frequencies of long cfDNA molecules being more enriched in transcriptional start sites and open chromatin regions. Second, long molecules exhibit a distinctive cleavage profile surrounding CpG sites. The cleavage ratio that makes use of multiple positions surrounding CpG sites can be used to distinguish non-HCC individuals from patients with HCC. In contrast to the previous report that the generation of short cfDNA molecules might be largely attributed to DNASE1L3 (Serpas et al. 2019; Chan et al. 2020), we demonstrate that the characteristic distribution of long cfDNA molecules in the genome might at least in part be attributed to the DFFB enzyme activity using genetically modified mice. This observation suggests that different nucleases involved in the fragmentation of long and short cfDNA molecules in human plasma.

By using knockout mice, a plasma DNA fragmentation model based on the end motifs of cfDNA molecules has been proposed, in which DFFB, DNASE1L3 and possibly other nucleases first digest DNA intracellularly, and then extracellular DNASE1L3 and DNASE1 further degrade the plasma DNA (Han et al. 2020). DNASE1L3, DNASE1 and DFFB preferentially cut at C-end, T-end and A-end nucleotides of cfDNA fragments, respectively (Han et al. 2020). The long molecules uncovered via single-molecule sequencing platforms are likely to be, at least in part, the end product of DFFB other than DNASE1L3, which is supported by the observation of enrichment of A-end fragments in long cfDNA molecules (Yu et al. 2021). Of note, following knocking out of the *Dnase1l3* and/or *Dnase1* genes in mice, DFFB appears to take on the leading role in cfDNA fragmentation. Additionally, no clear correlation was found between total cfDNA concentration and proportion of long molecules. The higher proportions of long molecules originated from accessible chromatin regions in *Dnase1l3*$^{-/-}$ and *Dnase1*$^{-/-}$ & *Dnase1l3*$^{-/-}$ double knockout mice were reminiscent of the

characteristic of long molecules in human plasma. The long cfDNA molecules in human plasma are therefore likely to be the digestion products of DFFB, which have escaped further or secondary digestion by DNASE1L3. The exact biological mechanisms for the escape of cleavage for long cfDNA remain elusive. *Dnase1l3*$^{-/-}$ mice have been shown to develop features of systemic lupus erythematosus on both the 129 and B6 genetic backgrounds (Sisirak et al. 2016). DNASE1L3 deficiencies in human are associated with systemic lupus erythematosus (Al-Mayouf et al. 2011). Long cfDNA that are generated by DFFB might have potential immunomodulatory effects to the innate immune responses.

Prior studies systematically profiled plasma DNA end motifs using mice of different genotypes and found altered end motif frequencies in nuclease-deficient mice (Serpas et al. 2019; Chan et al. 2020). By linking the end motifs and nuclease-deficient genotypes, a non-negative matrix factorization algorithm-based approach was developed to deconvolute contributions of nucleases. The work has demonstrated the feasibility of inferring nuclease activities in liquid biopsy and that DNASE1L3 is a leading contributor to fragmentation in the plasma cfDNA (Zhou et al. 2023). One reason for previous studies focusing on the DNASE1L3 activity is largely because that the Illumina sequencing platform is designed for short read sequencing, with an upper limit of sequenceable fragments being at approximately 600 bp. Long cfDNA molecules investigated in this study are more likely to reflect the DFFB activity. The genome-wide representational analysis of long cfDNA molecules provides another means to dissect the DFFB activity and link the enzyme activity to gene transcriptional activity. Further investigations to elucidate cfDNA fragmentation are warranted. A more comprehensive picture of various nuclease activities could be uncovered when examining the full spectrum of cfDNA molecules in different nuclease-knockout mouse models.

We have uncovered genomic representation difference between long and short cfDNA molecules in human plasma. The enrichment of long molecules in transcriptionally active

regions may have important implications. It has been suggested that DFFB first cleaves at nuclear scaffold attachment sites (Nagata et al. 2003) to unfold the chromatin. DFFB induces DNA breaks as a signal for cell fate (Larsen and Sørensen 2017). Though direct evidence of locations of DFFB-induced DNA breaks is lacking, we observed a weak positive correlation between overrepresentation of long cfDNA molecules and DNA double-strand breaks detected in a lymphoblastoid cell line sample. Future examinations at higher resolutions to address the relationship between transcription-associated DNA double-strand breaks and long DNA generation will be needed. Whether DFFB and other nucleases preferentially cleave at those damaged DNA breakpoints and thereafter induce downstream DNA fragmentation remain to be investigated. Additionally, the nuclear morphology at the time of cell death might be another dimension that affects the fragmentation and hence results in overrepresentation of long molecules in euchromatin (Toné et al. 2007). Beyond the mechanism, the enrichment of long molecules in genic regions highlights potential diagnostic value. Retaining characteristics of chromatin structure, non-random fragmentation of cfDNA reflects transcriptomic and epigenomic status of dying cells (Ivanov et al. 2015; Ulz et al. 2016; Snyder et al. 2016; Lo et al. 2021). A number of studies showed that coverage profiles near TSSs correlate with gene expression and that accessibilities near the transcription factor binding sites reflect the transcription factor activity (Ulz et al. 2016, 2019). Notably, the normalized end frequencies at TSSs in long cfDNA molecules were higher than those of short cfDNA molecules, suggesting a potential higher utility for long cfDNA molecules for inferring gene expression when compared with their shorter cfDNA counterparts. Moreover, the abundance of long molecules demonstrated discriminating power in cancer detection.

Compared with sequencing-by-synthesis, the current throughput of single-molecule sequencing is relatively low, which presents challenges to analyses that require a larger number of molecules. The low number of long cfDNA molecules sequenced creates obstacles for analyzing the association with gene expression at a higher resolution, especially in the context of measuring expression of a single gene. Nevertheless, with

technological advances, such as with the recent launch of a higher throughput version of the SMRT-based sequencer, and further optimized protocols, the throughput is likely to improve, potentially facilitating the elucidation of further biological insights and the development of clinical tools. Alternatively, enriching the long molecules of interest may help to adequately harness the advantages for diagnostic purposes. The performance of cancer detection using long molecule abundance and fragmentation patterns can be further enhanced. In addition, a larger cohort that allows validation of these biomarkers is desired. Platform-specific differences were observed from SMRT and ONT sequencing data. Future investigations into the use of a specific platform to maximize the use of long cfDNA are warranted.

In summary, this work expands the current knowledge of fragmentation properties of long cfDNA molecules. The characteristic distribution of long cfDNA molecules along the human genome is correlated to transcriptional activities and disease states, potentially opening up possibilities for biomarker discovery and new clinical applications.

METHODS

Human sample collection

The study was approved by the Joint Chinese University of Hong Kong – Hospital Authority New Territories East Cluster Clinical Research Ethics Committee. PacBio SMRT sequencing data of plasma cfDNA, including healthy individuals (N=15), pregnancies of different trimesters (N=28), HBV carriers (N=13) and patients with HCC (N=13), were obtained from previous studies (Yu et al. 2021; Choy et al. 2022). Additionally, we collected blood samples from healthy individuals (N=5), HBV carriers (N=6) and patients with HCC (N=35) from the Prince of Wales Hospital, Hong Kong, with written informed consent. Nanopore sequencing data of plasma cfDNA from pregnancies of different trimesters (N=31) were obtained from the prior study (Yu et al. 2023b). Additionally, plasma cfDNA from non-pregnant healthy individuals (N=5) were collected and subjected to ONT sequencing. The blood samples were processed by a double centrifugation protocol (1,600 x *g* for 10 min at 4°C, followed by further centrifugation of the plasma at 16,000 x *g* for 10 min at 4°C) as previously described (Chiu et al. 2001).

Murine Models

Mice with a CRISPR-Cas9-targeted deletion of exon 5 in *Dnase1l3* of the C57BL/6 (B6) background were generated by The Jackson Laboratory. Mice carrying a targeted allele mutation of *Dnase1* (*Dnase1*$^{tm1.1(KOMP)Vlcg}$) and mice carrying a targeted allele of *Dffb* (*Dffb*$^{C57BL/6N-Dffbem1Wtsi}$), both on the B6 background were obtained from the Knockout Mouse Project Repository of the University of California, Davis. These mice were obtained under a third-party distribution agreement and are nontransferable. *Dnase1l3*$^{-/-}$ mice were cross-bred with *Dnase1*$^{-/-}$ mice to produce *Dnase1*$^{-/-}$ *& Dnase1l3*$^{-/-}$ *mice* in the Laboratory Animal Services Centre of The Chinese University of Hong Kong (CUHK). Mice including wild-type B6 were maintained in the same facility. All experimental procedures were approved by the Animal Experimentation Ethics Committee of CUHK and performed in compliance with the

Guide for the Care and Use of Laboratory Animals (8th edition) established by the National Institutes of Health.

Murine sample collection

Mice were sacrificed and exsanguinated by cardiac puncture. Blood was transferred into EDTA-containing collection tubes. The blood samples were processed by a double centrifugation protocol (1,600 x *g* for 10 min at 4°C, followed by further centrifugation of the plasma at 16,000 x *g* for 10 min at 4°C) as previously described (Chiu et al. 2001). The resulting plasma was collected.

Library preparation and sequencing

Plasma DNA was extracted from 1.2 to 4 mL of human plasma and 1.1 to 1.4 mL of pooled mice plasma with the QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's instructions. SMRTbell Express Template Prep Kit 2.0 (PacBio) was used for the library preparation of plasma DNA. Briefly, DNA molecules were ligated with hairpin adaptors to form a circularized template. Sequencing Primer v4 was annealed to the sequencing template, followed by binding of polymerase to templates using a Sequel II Binding Kit 2.1 and Internal Control Kit 1.0 (PacBio). SMRT Cell 8M was used for sequencing with a Sequel II Sequencing 2.0 Kit (PacBio), and sequencing movies were collected for 30 hours. Nanopore library preparation and sequencing of pregnant samples are as described before (Yu et al. 2023b). For healthy samples, nanopore libraries were prepared using Native Barcoding Kit (SQK-NBD114.96, ONT) according to the manufacturer's instructions except that a bead-to-sample ratio of 3 was used in all clean-up steps using AMPure XP beads with prolonged incubation time for DNA repair, end-prep, barcode and adapter ligation. Short Fragment Buffer, which retained DNA fragments of all sizes, was used in adapter ligation. Each library was loaded onto a PromethION Flow Cell R10.4.1 and sequenced on a PromethION device for 72 hours.

Sequence alignment

Circular consensus sequencing (CSS) reads that were generated from at least 3 PacBio subreads were kept. CCS reads from human and mice data were aligned to hg19 and mm10 respectively using blasr (Chaisson and Tesler 2012). Nanopore-generated sequences were aligned to hg19 with minimap2 (Li 2018) as described in (Yu et al. 2023b). The size distributions of the raw sequenced molecules before and after alignment were examined. SMRT sequencing data showed highly overlapped size distributions, while ONT sequencing data showed deviations comparing before and after alignment (**Supplemental Fig. S7**). Chimeric reads were filtered out prior to downstream analyses. For human sequence data, re-analysis of the data using the GRCh38 human reference genome would not affect the results significantly, as the major difference between these two versions of the human reference genome is the sequence representation for centromeres and highly repetitive regions. Similarly, for mouse sequence data, re-analysis of the data using the GRCm39 mouse reference genome would not affect the results significantly because only uniquely aligned reads were included in downstream analyses.

Genomic representation analysis

The human genome was partitioned into non-overlapping 100 kb bins. Sequences that mapped to blacklist regions (Amemiya et al. 2019) were removed. Long and short molecules were counted in each bin. The bin counts were smoothed by a 1 Mb moving average window and normalized by the median bin count of autosomes. Cytobands information for chromosome ideograms were obtained from UCSC Genome Browser (http://genome.ucsc.edu/). Gene densities were estimated by number of genes in a 100-kb window. The correlations between normalized bin counts and gene densities were calculated using Pearson's method and smoothed data applying 1Mb-window were used.

To examine the correlation between genomic representation of cfDNA molecules and presence of DNA double-strand breaks, we obtained double-strand breaks detected from a lymphoblastoid cell line sample (Bouwman et al. 2020). DNA double-strand breaks were counted in each 100-kb bin and normalized by the median bin count.

To calculate the proportions of long molecules originated from TSSs, DHSs and CTCF binding sites in mice, we counted long and short molecules overlapped with these regions. 2000 bp upstream and downstream of transcription start sites or peak points of DHSs and CTCF binding sites were regarded as regions of interest. At least 50% size of a molecule needs to be overlapped with these regions of interest to be counted. The proportions of long molecules were computed as the number of long molecules divided by the total number of molecules overlapped with the regions of interest.

Genic abundance analysis

To quantify molecule abundance over genes, we adopted similar quantification of transcripts using reads per kilobase per million mapped reads. We counted molecules overlapped with gene bodies and normalized by the total number of molecules and the total length of the genes. For both long and short molecules, we required at least 50% size of a molecule to be overlapped with a gene to be counted. To calculate the Pearson's correlation between molecule abundance and gene expression, we used the median molecule abundance of samples and correlated with the median gene expression of each gene set. The Mann-Kendall trend test was performed using the median molecule abundance of samples.

HCC cancer-associated genes were identified using curated HCC expression data from The Cancer Genome Atlas (Cancer Genome Atlas Research Network et al. 2013). Expressions were averaged across 356 HCC tumor tissue samples and ranked by the averaged expression. Top 5000 expressed genes were selected and used as HCC cancer-associated

genes. The abundance of long and short molecules over these 5000 genes were calculated. AUCs were calculated based on the long molecule index.

Single cell gene expression data were downloaded from The Human Protein Atlas (https://www.proteinatlas.org/) and The Tabular Muris project (Tabula Muris Consortium et al. 2018) for human and mouse, respectively. In human analysis, housekeeping genes were retrieved from a previous study (Eisenberg and Levanon 2013) and 3510 autosomal housekeeping genes were used. 2154 genes with a mean of less than 1 normalized transcript per million across 54 tissues were defined as unexpressed genes. Autosomal protein coding genes were stratified by median expression across 79 cell types. Gene groups of EXP1-5 correspond to genes with median normalized transcript per million in the range of [0, 0.001), [0.001, 1), [1, 8), [8, 30), [30, 11800), with each containing 4836, 3163, 4187, 4816, and 3088 genes, respectively. To correlate the abundance of molecules with gene expression at relatively high resolution in human samples, median expression level of a gene across cell types were log-transformed and scaled to a score ranging from 0 to 1000. Median molecule abundance of genes at each scaled expression level was used. Genes with expression scores of lower and upper quantiles (outliers) were discarded from the correlation analysis.

Murine genes were stratified by median expression across 20 tissue types. The low, medium and high expression gene groups in mice correspond to autosomal protein coding genes with median normalized transcript count in the range of [0, 50), [50,180) and [180,450), with each containing 10987, 3883 and 5195 genes, respectively. The percentage of change in long molecule abundance between the knockout and wild-type mice was computed using the difference of long molecule abundance between knockout and wild-type mice, and divided by the abundance of long molecules in the wild-type mice.

End frequency analysis

End frequency analysis was performed surrounding the defined region of interests. The region of interests, including CpGs, DHSs, CTCF binding sites and TSSs, were extracted as follows.

Transcription start sites were retrieved using NCBI RefSeq data (https://www.ncbi.nlm.nih.gov/refseq/). DHSs and CTCF binding sites were identified from publicly available data. Tissue-invariant DHSs from the study (Meuleman et al. 2020) were identified and genomic coordinates were converted from hg38 to hg19 using the UCSC liftOver tool. 44,997 tissue-invariant DHSs were used in the analysis. 157,556 CTCF binding sites from the Encyclopedia of DNA Elements (ENCODE) uniformly processed transcription factor binding stie clusters in human were used for analysis (Dunham et al. 2012). 107,227 DHSs (ENCFF855RCO) and 62,461 CTCF binding sites (ENCFF883UPM) from CH12.LX mouse cell line available via ENCODE were used in mice analysis.

The peak or mid-point of the region of interests was regarded as position 0. The end frequencies within +/- 2000 bp were aggregated and normalized with median count of the region. For regulatory regions (DHS and CTCF) and TSS, we applied a 10 bp and 20 bp window to smooth the count. Fold change of end frequencies around DHSs was measured as follows. End frequencies within a DHS region (a median size of 211 bp) were aggregated. In parallel, end frequencies of the equivalent region that are 3000 bp away upstream and downstream were aggregated and used as a normalization factor. Fold change of end frequencies around TSSs was measured using a ratio between long and short in human. Normalized end frequencies from position -50 to 50 were aggregated to compute the ratio.

Cleavage profiles surrounding CpGs

For human data analysis, commonly methylated and unmethylated CpG sites were defined as methylation level > 70% and < 30% respectively in human buffy coat, liver and placental tissues. Whole genome bisulfite sequencing subjected to the Illumina platform from the three

tissues available from a previous study (Sun et al. 2015) were used to identify 2,422,965 commonly methylated and 558,267 commonly unmethylated CpGs. For analysis of all CpGs, 26,752,699 autosomal CpGs from the human genome hg19 were obtained. The cleavage proportion to measure the cutting frequency were measured as the number of molecules ends at a particular site divided by sequencing depth at the site, as described in (Zhou et al. 2022). The CGN/NCG motif ratio was calculated using the number of molecules ends at position 0 divided by the number of molecules ends at position -1. The cleavage ratio was calculated using the aggregated number of molecules ends at position -4, -2, 1 and 4 divided by the number of molecules ends at position -1.

Statistical analysis

The Mann-Kendall trend test was used for testing monotonic trend in molecule abundance over gene expression groups. The alternative hypothesis that monotonic increasing trend is present was assumed. The Pearson's correlation test was used to measure relations between the abundance of molecules and gene expression. The Wilcoxon rank-sum test was used to compare two groups at a significance level of 0.05. The Kruskal-Wallis test or Friedman test in case of dependent groups was used to compare three or more groups at a significance level of 0.05. Post hoc pairwise Wilcoxon tests were performed with Benjamini-Hochberg adjustment to yield pairwise *P* values.

DATA ACCESS

The sequencing data generated in this work have been submitted to the European Genome-phenome Archive (EGA; https://web2.ega-archive.org/) under accession number EGAS00001005515.

COMPETING INTEREST

K.C.A.C. and Y.M.D.L. hold equities in DRA, Take2, Grail/Illumina, and Insighta. P.J. and W.K.J.L hold equities in Grail/Illumina. P.J. is a consultant to KingMed Future. W.P. is a consultant to Take2. K.C.A.C. and P.J. are Directors of DRA, Take2, KingMed Future and Insighta. W.K.J.L is a Director of DRA. S.C.Y.Y. received financial support from Oxford Nanopore for attending meetings. H.C., P.J., K.C.A.C. and Y.M.D.L. filed a US patent application (No. 63/544,014), entitled “Genomic origin, fragmentomics, and transcriptional correlation of long cell-free DNA” on October 13, 2023, based on the data in this study. Patent royalties are received from Grail, Illumina, DRA, Take2, and Xcelom.

ACKNOWLEDGEMENTS

The work was supported by the Innovation and Technology Commission (InnoHK Initiative). Y.M.D.L. received an endowed chair from the Li Ka Shing Foundation. We would like to thank Ms. Angel Lai, Mr. Chris Kum, Mr. Saravanan Ramakrishnan, Mr. Xingfu Qin and Mr. Danny Wong for their technical assistance.

583 **FIGURE LEGENDS**

584

585 **Figure 1. Distribution of long and short molecules from human plasma DNA.** (A)
586 Comparison of genomic representation on Chromosome 10 between long and short DNA
587 molecules in 15 non-pregnant controls using SMRT sequencing. Overrepresentation and
588 underrepresentation of long cfDNA molecules with respect to short molecules are indicated
589 in blue and red, respectively. The genomic representation was determined based on 100-kb
590 bins, and was further smoothed by a 1-Mb moving average sliding window. The horizontal
591 solid lines indicate normalized median differences between long and short molecules. The
592 dashed rectangular boxes indicate one euchromatic (i) and one heterochromatic (ii) region.
593 The track in between overrepresentation and underrepresentation of long molecules shows
594 the chromosome ideogram. The ideogram band colors correspond to cytogenetic bands in
595 UCSC Genome Browser. Darker bands are AT-rich and lighter bands are GC-rich.
596 Centromeric regions are indicated in dark green. The bottom track displays gene densities
597 estimated by number of genes in 100-kb windows. (B) Comparison of genomic
598 representation on Chromosome 10 between long and short DNA molecules in 31 pregnant
599 samples using ONT sequencing.

600

601 **Figure 2. The abundance of long cfDNA molecules exhibits a positive correlation with**
602 **transcriptional activity.** (A, B) The abundance of long and short molecules on gene bodies
603 of unexpressed and housekeeping genes for non-pregnant controls and pregnant subjects.
604 (A) Data from SMRT sequencing. (B) Data from ONT sequencing. (C, D) The correlation
605 between gene expression and molecules abundance. Median expression level of a gene
606 across tissues was log-transformed and scaled. The median molecule abundance was
607 derived from scaled expression levels. (C) Pooled data of 15 non-pregnant controls from
608 SMRT sequencing. (D) Pooled data of 31 pregnant samples from ONT sequencing. *P* values
609 by Pearson's correlation test (C, D). (E, F) Comparisons of DNA methylation between long
610 and short molecules. Data points from one sample are connected with a black line. (E) Data

611 from SMRT sequencing. (F) Data from ONT sequencing. *P* values by Wilcoxon rank-sum
612 test (E, F).
613
614 **Figure 3. The abundance of long cfDNA molecules for HCC detection.** (A) Comparison
615 of the abundance of SMRT sequencing molecules among healthy individuals, HBV carriers
616 and patients with HCC. The abundance of long and short molecules was measured using
617 the top 5000 expressed genes in HCC tumor tissues. The Kruskal-Wallis test *P* value for
618 differences among groups. Post hoc pairwise Wilcoxon rank sum test *P* values with
619 Benjamini-Hochberg adjustment are shown above horizontal lines. (B) ROCs of long
620 molecule abundance measured in (A) for distinguishing individuals without HCC, including
621 healthy subjects and HBV carriers, and with HCC. Multiple thresholds, including 0.1 million
622 (Total > 0.1M), 0.3 million (Total > 0.3M), 0.5 million (Total > 0.5M) and 1 million (Total > 1M)
623 molecules from a sample, were used to include samples for constructing ROCs.
624
625 **Figure 4. Normalized end frequencies of SMRT sequencing data.** (A) Normalized end
626 frequencies of SMRT long and short molecules pooled from plasma of healthy individuals at
627 TSSs of expression-stratified gene groups EXP1 to EXP5, corresponding to low to high
628 expression. Transcription start positions are denoted as position 0. All transcription start
629 sites were strand-adjusted so that positive positions are in the direction of transcription. (B, C)
630 Normalized end frequencies of SMRT long and short molecules pooled from plasma of
631 healthy individuals at DHSs (B) and CTCF binding sites (C). DHSs or CTCF binding sites
632 peaks are denoted as position 0; downstream and upstream 2000 bp are shown.
633
634 **Figure 5. Cleavage profiles of long and short molecules surrounding CpGs.** (A, B)
635 Cleavage profiles surrounding all autosomal CpGs for short (A) and long (B) cfDNA
636 molecules. Each line represents one sample. A cleavage window of 11 bases is shown.
637 Position 0 and 1 indicate cytosine and guanine, respectively. (C) Box plot of CGN/NCG motif
638 ratios for long and short molecules. (D) Box plot of cleavage ratios between aggregating

639 position -4, -2, 1 and 4 and position -1. (E) AUCs for distinguishing patients with HCC from
640 non-HCC subjects using cleavage ratios in (D). *P* values of differences among groups by
641 Kruskal-Wallis tests. Post hoc pairwise *P* values by Wilcoxon rank sum tests with Benjamini-
642 Hochberg adjustment shown above horizontal lines (C, D).
643
644 **Figure 6. Nuclease-mediated fragmentation in knockout mice.** (A) Pooled molecule size
645 distributions of wild-type and nuclease-deficient mice from SMRT sequencing. Visualization
646 of size in the range of 0 to 5000 bp and $\log_{10}$-transformed frequencies were used. (B) Zoom-
647 in plot of (A) showing size in the range of 0 to 250 bp on the linear scale. (C) Boxplot
648 showing the proportions of long molecules originated from 2000 bp upstream and
649 downstream of transcription start sites. (D) Percent of changes in pooled long molecules
650 (>500 bp) abundance relative to wild-type mice on low, medium and high expression gene
651 groups.

REFERENCES

Al-Mayouf SM, Sunker A, Abdwani R, Abrawi SA, Almurshedi F, Alhashmi N, Al Sonbul A, Sewairi W, Qari A, Abdallah E, et al. 2011. Loss-of-function variant in DNASE1L3 causes a familial form of systemic lupus erythematosus. *Nat Genet* **43**: 1186–1188.

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* **9**: 9354.

BAC Resource Consortium T, Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen X-N, Furey TS, Kim U-J, Kuo W-L, et al. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**: 953–958.

Ballarino R, Bouwman BAM, Agostini F, Harbers L, Diekmann C, Wernersson E, Bienko M, Crosetto N. 2022. An atlas of endogenous DNA double-strand breaks arising during human neural cell fate determination. *Sci Data* **9**: 400.

Bickmore WA, Sumner AT. 1989. Mammalian chromosome banding — an expression of genome organization. *Trends Genet* **5**: 144–148.

Bouwman BAM, Agostini F, Garnerone S, Petrosino G, Gothe HJ, Sayols S, Moor AE, Itzkovitz S, Bienko M, Roukos V, et al. 2020. Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nat Protoc* **15**: 3894–3941.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**: 1113–1120.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.

Chan KCA, Zhang J, Hui ABY, Wong N, Lau TK, Leung TN, Lo K-W, Huang DWS, Lo YMD. 2004. Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem* **50**: 88–92.

Chan RWY, Serpas L, Ni M, Volpi S, Hiraki LT, Tam L-S, Rashidfarrokhi A, Wong PCH, Tam LHP, Wang Y, 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**: 882–894.

Chiu RWK, Poon LLM, Lau TK, Leung TN, Wong EMC, Lo YMD. 2001. Effects of blood-processing protocols on fetal and total DNA quantification in maternal plasma. *Clin Chem* **47**: 1607–1613.

Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, Olivia Tse OY, Wong J, Wong VWS, Wong GLH, et al. 2022. Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. *Clin Chem* **68**: 1151–1163.

Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, Karaca E, Chiarle R, Skrzypczak M, Ginalski K, et al. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10**: 361–365.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569–574.

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, et al. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428**: 529–535.

Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, Lo YMD. 2020. The Biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am J Hum Genet* **106**: 202–214.

Han DSC, Ni M, Chan RWY, Wong DKL, Hiraki LT, Volpi S, Jiang P, Lui KO, Chan KCA, Chiu RWK, et al. 2021. Nuclease deficiencies alter plasma cell-free DNA methylation profiles. *Genome Res* **31**: 2008–2021.

Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. 2015. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16**: S1.

Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, Wong GLH, Chan SL, Mok TSK, Chan HLY, et al. 2015. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **112**: E1317–E1325.

Katsman E, Orlanski S, Martignano F, Fox-Fisher I, Shemer R, Dor Y, Zick A, Eden A, Petrini I, Conticello SG, et al. 2022. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from nanopore sequencing. *Genome Biol* **23**: 158.

Larsen BD, Sørensen CS. 2017. The caspase-activated DNase: apoptosis and beyond. *The FEBS Journal* **284**: 1160–1170.

Lau BT, Almeda A, Schauer M, McNamara M, Bai X, Meng Q, Partha M, Grimes SM, Lee H, Heestand GM, et al. 2023. Single-molecule methylation profiles of cell-free DNA in cancer with nanopore sequencing. *Genome Med* **15**: 33.

Lensing SV, Marsico G, Hänsel-Hertsch R, Lam EY, Tannahill D, Balasubramanian S. 2016. DSBCapture: in situ capture and sequencing of DNA breaks. *Nat Methods* **13**: 855–857.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, Zheng YW, Leung TY, Lau TK, Cantor CR, et al. 2010. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* **2**: 61ra91-61ra91.

Lo YMD, Han DSC, Jiang P, Chiu RWK. 2021. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616.

Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Teodosiadis A, et al. 2020. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**: 244–251.

733 Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R,
734 Goranova T, Marass F, Heider K, et al. 2018. Enhanced detection of circulating tumor
735 DNA by fragment size analysis. *Sci Transl Med* **10**, eaat4921.

736 Nagata S, Nagase H, Kawane K, Mukae N, Fukuyama H. 2003. Degradation of
737 chromosomal DNA during apoptosis. *Cell Death Differ* **10**: 108–116.

738 Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A, Soni C, Sisirak V, Lee W-S,
739 Cheng SH, et al. 2019. Dnase1l3 deletion causes aberrations in length and end-motif
740 frequencies in plasma DNA. *Proc Natl Acad Sci U S A* **116**: 641–649.

741 Sisirak V, Sally B, D'Agati V, Martinez-Ortiz W, Özçakar ZB, David J, Rashidfarrokhi A, Yeste
742 A, Panea C, Chida AS, et al. 2016. Digestion of chromatin in apoptotic cell
743 microparticles prevents autoimmunity. *Cell* **166**: 88–101.

744 Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. 2016. Cell-free DNA comprises an in
745 vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**: 57–68.

746 Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, Chan W, Ma ESK, Chan SL,
747 Cheng SH, et al. 2015. Plasma DNA tissue mapping by genome-wide methylation
748 sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc*
749 *Natl Acad Sci U S A* **112**: E5503–E5512.

750 Sun K, Jiang P, Cheng SH, Cheng THT, Wong J, Wong VWS, Ng SSM, Ma BBY, Leung TY,
751 Chan SL, et al. 2019. Orientation-aware plasma cell-free DNA fragmentation analysis
752 in open chromatin regions informs tissue of origin. *Genome Res* **29**: 418–427.

753 Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and
754 processing, Library preparation and sequencing, Computational data analysis, Cell
755 type annotation, Writing group, Supplemental text writing group, Principal
756 investigators. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula
757 Muris. *Nature* **562**: 367–372.

758 Toné S, Sugimoto K, Tanda K, Suda T, Uehira K, Kanouchi H, Samejima K, Minatogawa Y,
759 Earnshaw WC. 2007. Three distinct stages of apoptotic nuclear condensation
760 revealed by time-lapse imaging, biochemical and electron microscopy analysis of
761 cell-free apoptosis. *Exp Cell Res* **313**: 3635–3644.

762 Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, Chan SL, Poon LCY, Leung TY,
763 Chan KCA, et al. 2021. Genome-wide detection of cytosine methylation by single
764 molecule real-time sequencing. *Proc Natl Acad Sci U S A* **118**: e2019768118.

765 Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, Wölfler A, Zebisch A, Gerger A,
766 Pristauz G, et al. 2019. Inference of transcription factor binding from cell-free DNA
767 enables tumor subtype prediction and early detection. *Nature Commun* **10**: 4666.

768 Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, Abete L, Pristauz G, Petru E,
769 Geigl JB, et al. 2016. Inferring expressed genes by whole-genome sequencing of
770 plasma DNA. *Nat Genet* **48**: 1273.

771 Yu SCY, Chan KCA, Zheng YWL, Jiang P, Liao GJW, Sun H, Akolekar R, Leung TY, Go ATJI,
772 Vugt JMG van, et al. 2014. Size-based molecular diagnostics using plasma DNA for
773 noninvasive prenatal testing. *Proc Natl Acad Sci U S A* **111**: 8583–8588.

774 Yu SCY, Choy LYL, Lo YMD. 2023a. 'Longing' for the next generation of liquid biopsy: the
775 diagnostic potential of long cell-free DNA in oncology and prenatal testing. *Mol Diagn*
776 *Ther* **27**: 563-571.

777 Yu SCY, Deng J, Qiao R, Cheng SH, Peng W, Lau SL, Choy LYL, Leung TY, Wong J, Wong
778 VW-S, et al. 2023b. Comparison of single molecule, real-time sequencing and
779 nanopore sequencing for analysis of the size, end-motif, and tissue-of-origin of long
780 cell-free DNA in plasma. *Clin Chem* **69**: 168–179.

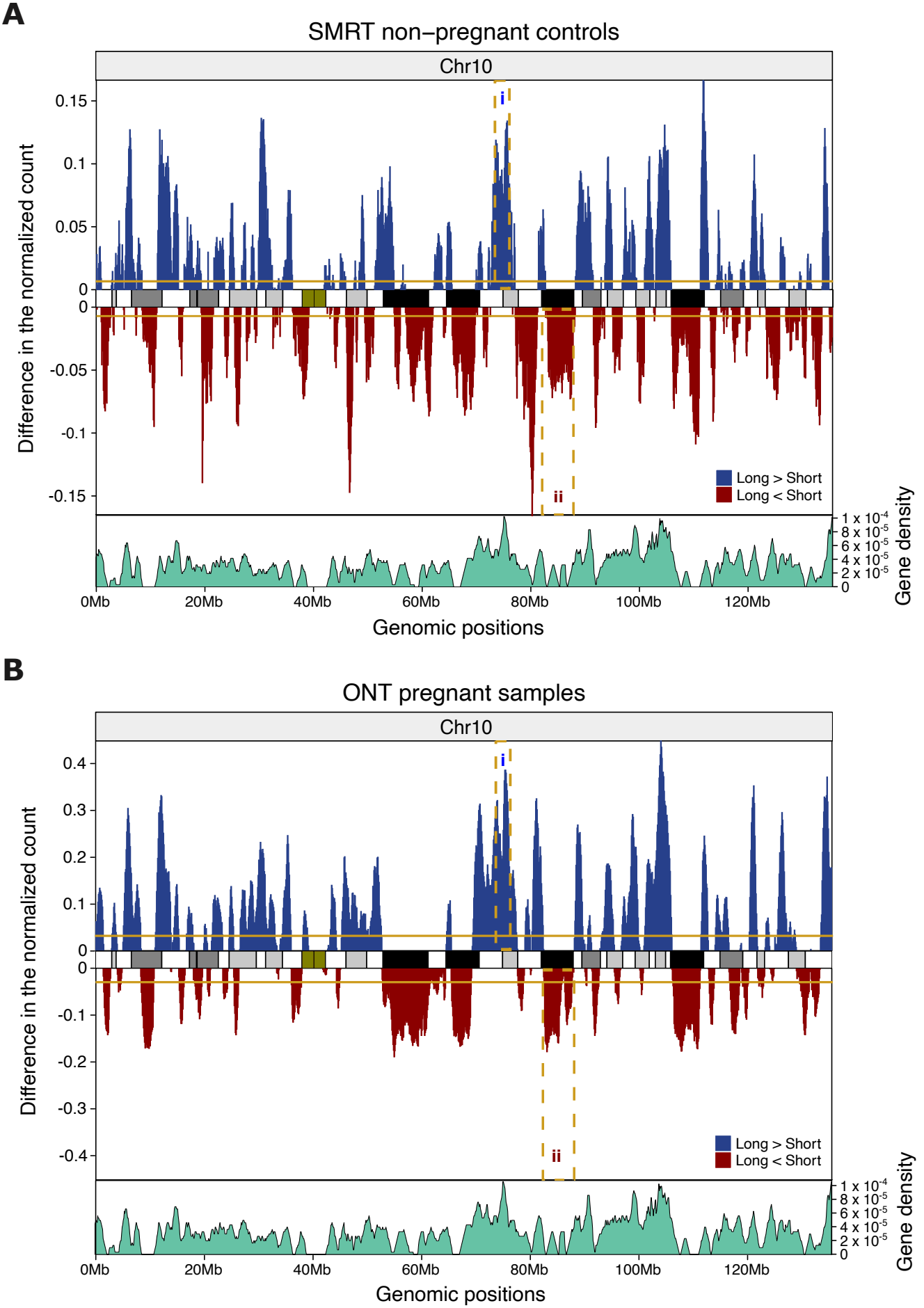
781 Yu SCY, Jiang P, Peng W, Cheng SH, Cheung YTT, Tse OYO, Shang H, Poon LC, Leung TY,
782 Chan KCA, et al. 2021. Single-molecule sequencing reveals a large population of
783 long cell-free DNA molecules in maternal plasma. *Proc Natl Acad Sci U S A* **118**:
784 e2114937118.

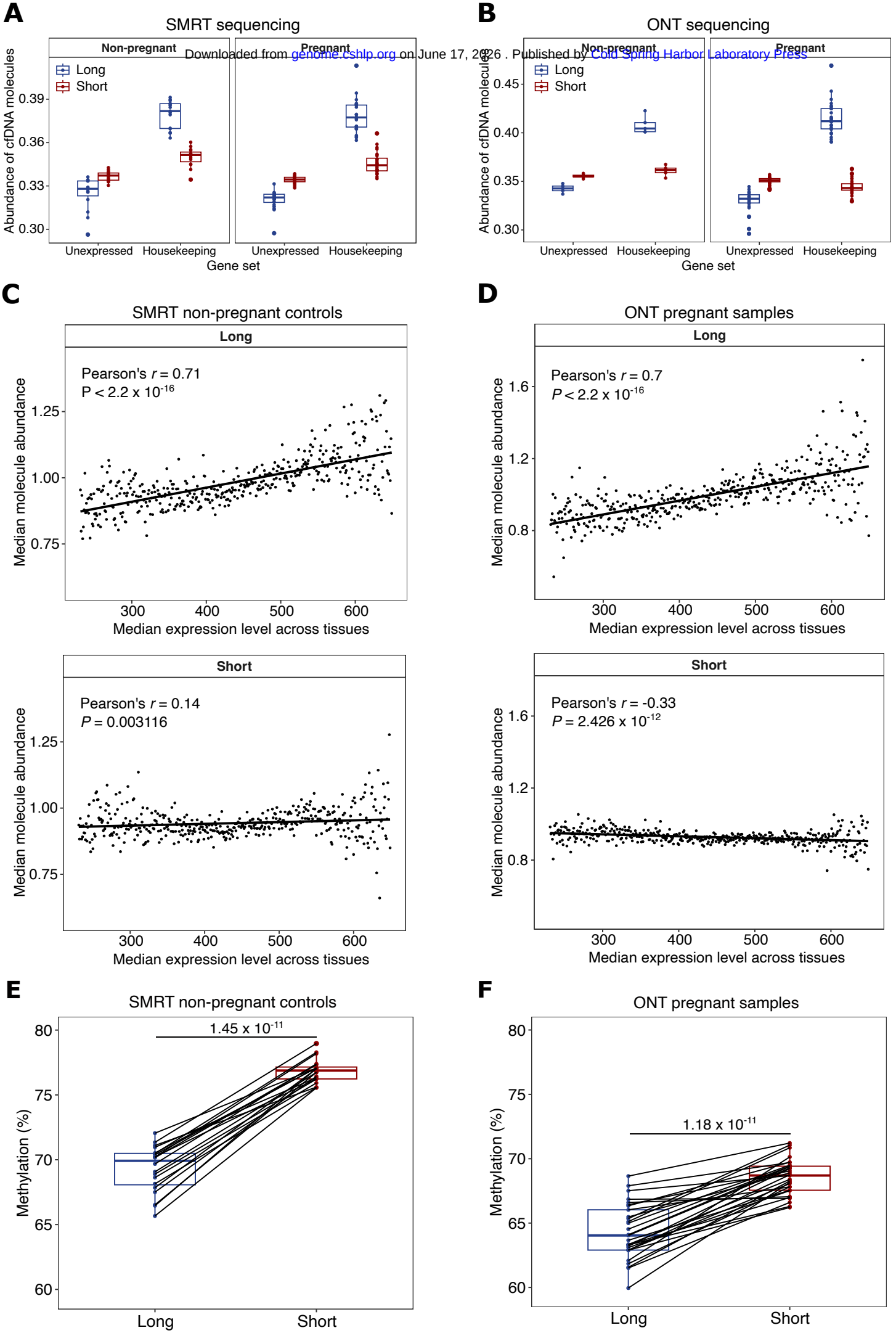
785 Zhou Q, Kang G, Jiang P, Qiao R, Lam WKJ, Yu SCY, Ma M-JL, Ji L, Cheng SH, Gai W, et al.
786 2022. Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc Natl Acad*
787 *Sci U S A* **119**: e2209852119.

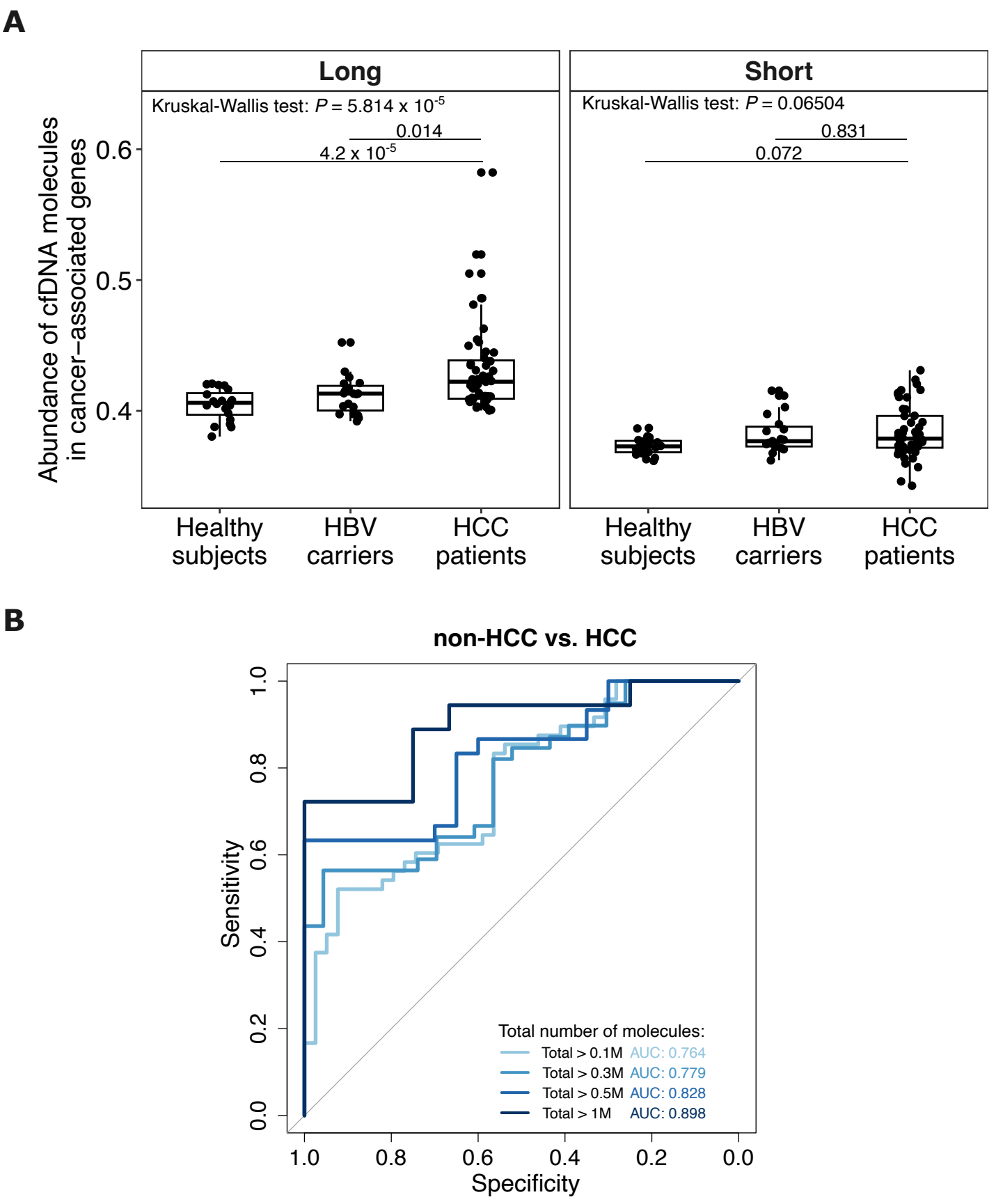
788 Zhou Z, Ma M-JL, Chan RWY, Lam WKJ, Peng W, Gai W, Hu X, Ding SC, Ji L, Zhou Q, et al.
789 2023. Fragmentation landscape of cell-free DNA revealed by deconvolutional
790 analysis of end motifs. *Proc Natl Acad Sci U S A* **120**: e2220982120.

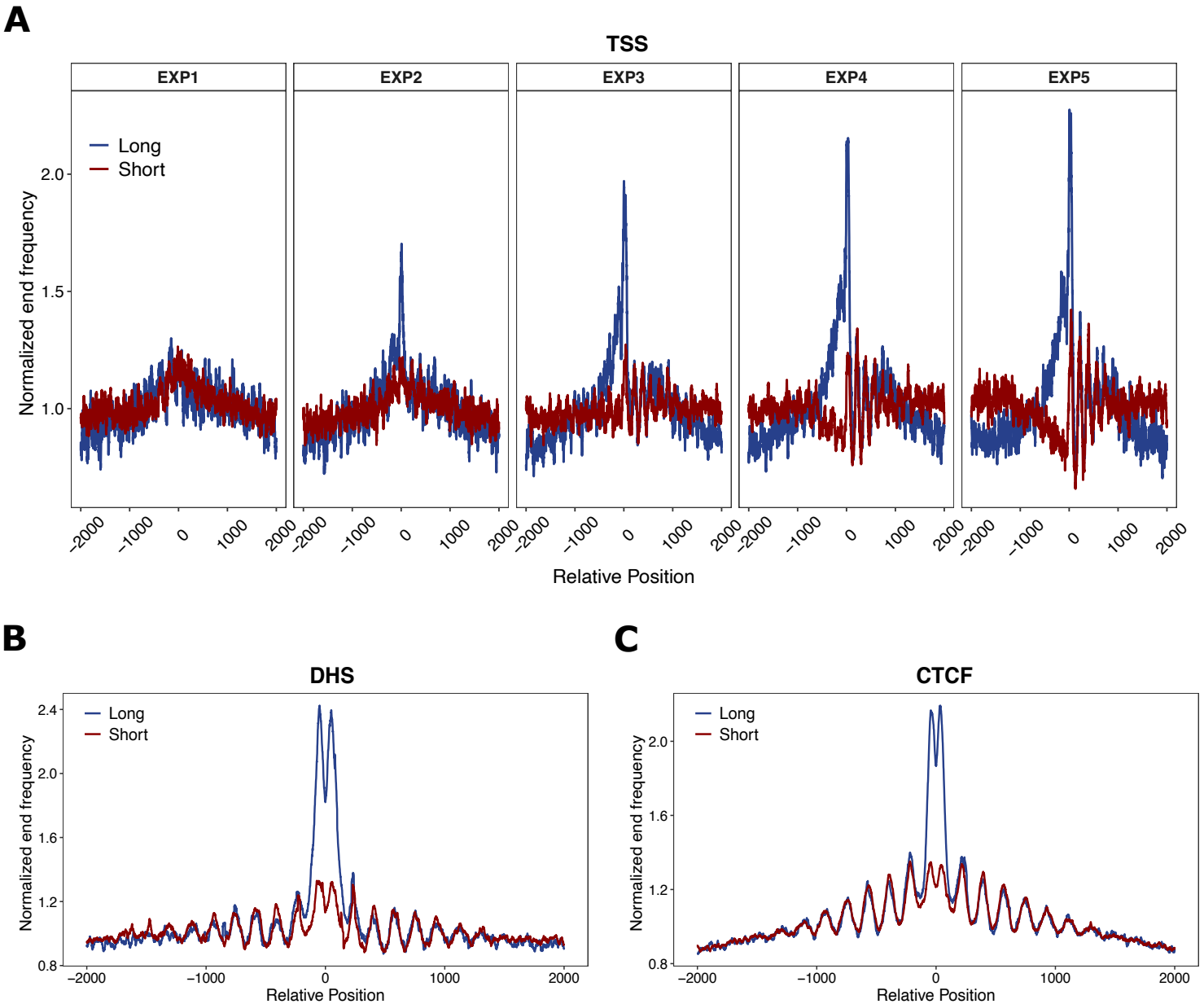
791 Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED,
792 Bennett DA, Bernstein BE, et al. 2013. Charting a dynamic DNA methylation
793 landscape of the human genome. *Nature* **500**: 477–481.

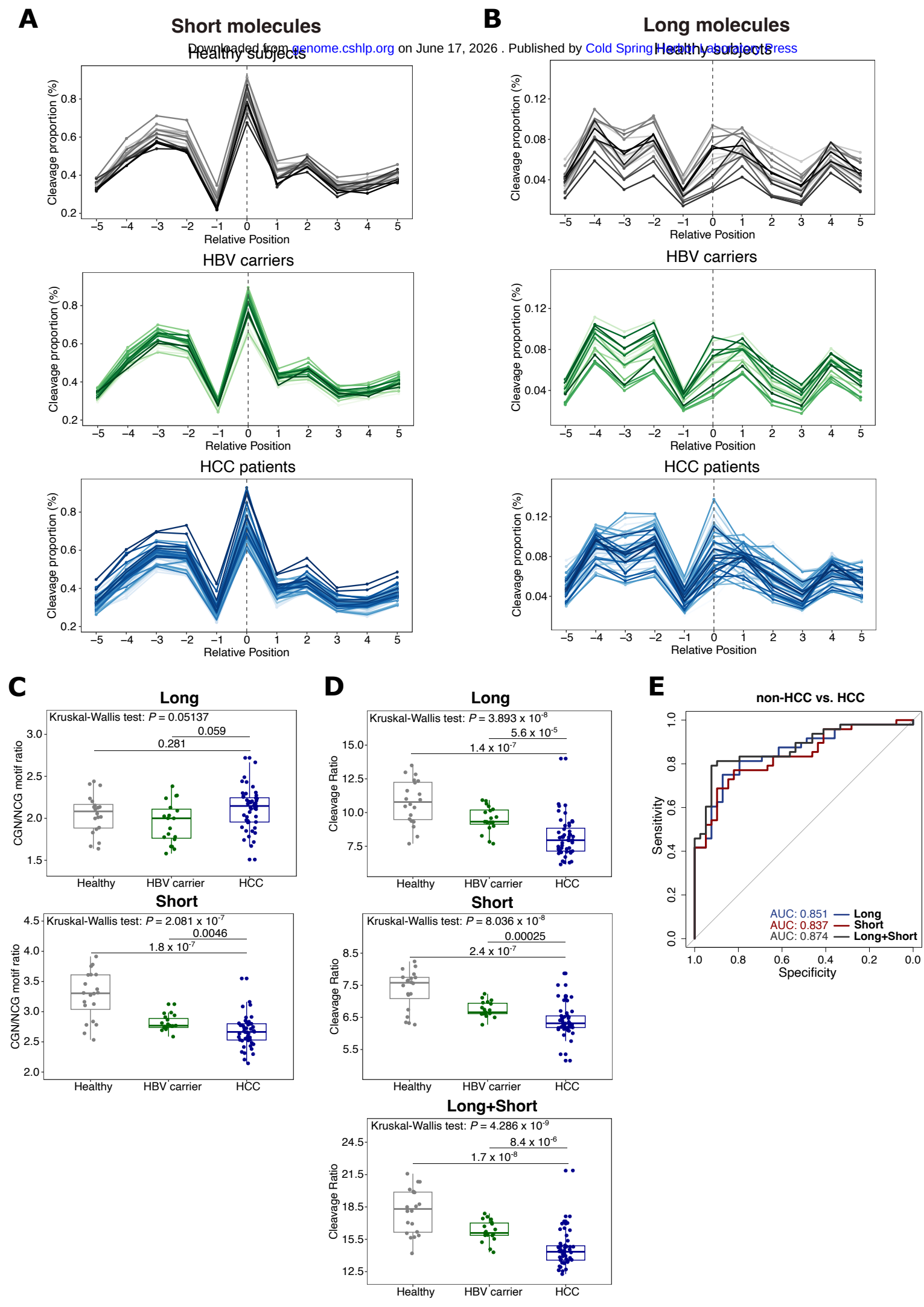
794

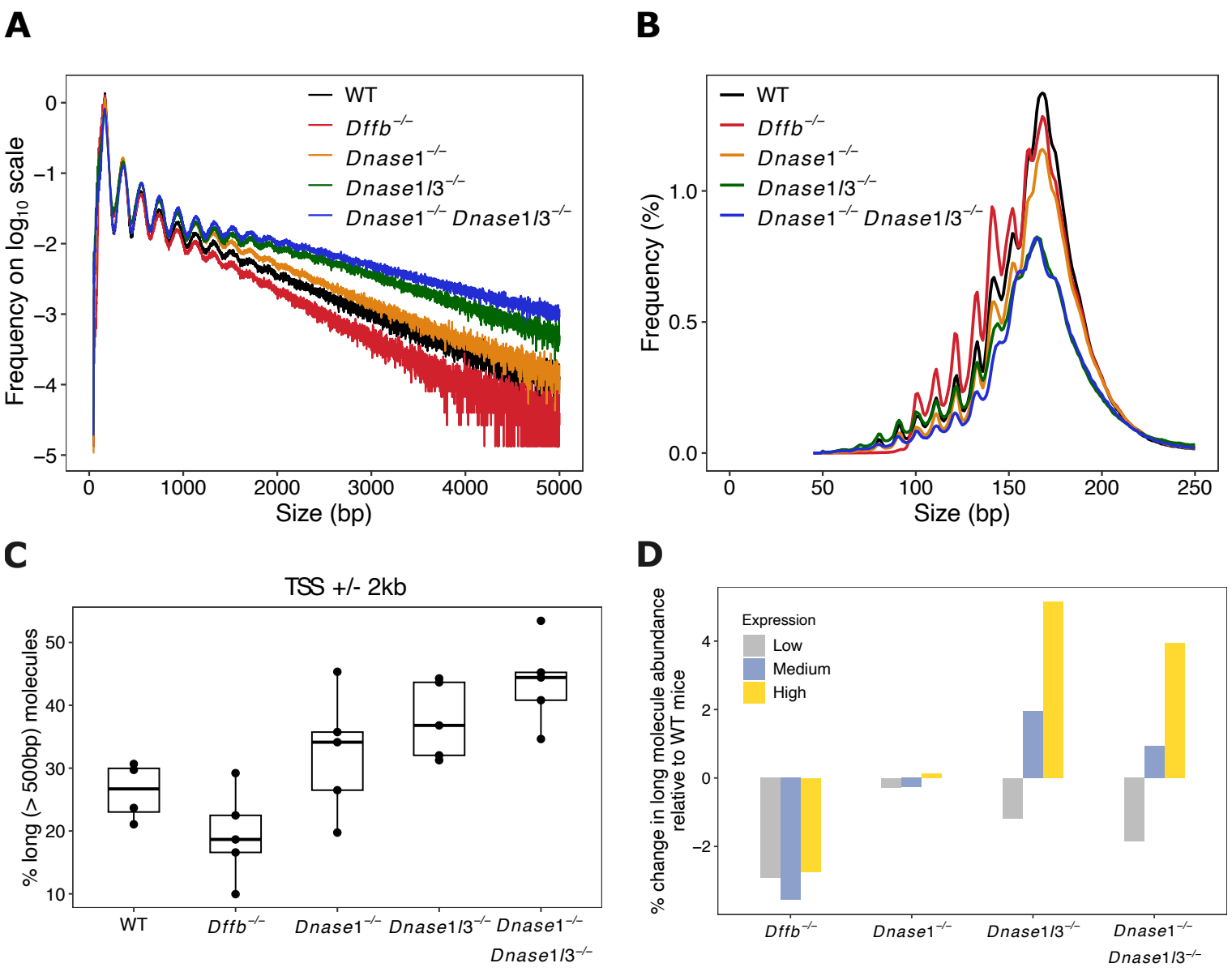

A
SMRT non−pregnant controls
Chr10
i
ii
0.15
0.1
0.05
0
0
-0.05
-0.1
-0.15
Difference in the normalized count
Long > Short
Long < Short
1 x 10-4
8 x 10-5
6 x 10-5
4 x 10-5
2 x 10-5
0
Gene density
0Mb
20Mb
40Mb
60Mb
80Mb
100Mb
120Mb
Genomic positions
B
ONT pregnant samples
Chr10
i
ii
0.4
0.3
0.2
0.1
0
0
-0.1
-0.2
-0.3
-0.4
Difference in the normalized count
Long > Short
Long < Short
1 x 10-4
8 x 10-5
6 x 10-5
4 x 10-5
2 x 10-5
0
Gene density
0Mb
20Mb
40Mb
60Mb
80Mb
100Mb
120Mb
Genomic positions




A
Long
Kruskal-Wallis test: P = 5.814 x 10-5
0.014
4.2 x 10-5
Short
Kruskal-Wallis test: P = 0.06504
0.831
0.072
Abundance of cfDNA molecules in cancer−associated genes
0.6
0.5
0.4
Healthy subjects
HBV carriers
HCC patients
B
non-HCC vs. HCC
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Total number of molecules:
Total > 0.1M AUC: 0.764
Total > 0.3M AUC: 0.779
Total > 0.5M AUC: 0.828
Total > 1M AUC: 0.898


A
TSS
EXP1
EXP2
EXP3
EXP4
EXP5
Long
Short
Normalized end frequency
2.0
1.5
1.0
−2000
−1000
0
1000
2000
Relative Position
B
DHS
Long
Short
Normalized end frequency
2.4
2.0
1.6
1.2
0.8
−2000
−1000
0
1000
2000
Relative Position
C
CTCF
Long
Short
Normalized end frequency
2.0
1.6
1.2
0.8
−2000
−1000
0
1000
2000
Relative Position




A
Frequency on log10 scale
0
−1
−2
−3
−4
−5
0
1000
2000
3000
4000
5000
Size (bp)
WT
Dffb−/−
Dnase1−/−
Dnase1/3−/−
Dnase1−/− Dnase1/3−/−
B
Frequency (%)
1.0
0.5
0.0
0
50
100
150
200
250
Size (bp)
WT
Dffb−/−
Dnase1−/−
Dnase1/3−/−
Dnase1−/− Dnase1/3−/−
C
TSS +/- 2kb
% long (> 500bp) molecules
50
40
30
20
10
WT
Dffb−/−
Dnase1−/−
Dnase1/3−/−
Dnase1−/− Dnase1/3−/−
D
% change in long molecule abundance relative to WT mice
4
2
0
−2
Expression
Low
Medium
High
Dffb−/−
Dnase1−/−
Dnase1/3−/−
Dnase1−/− Dnase1/3−/−