PLOS One

RESEARCH ARTICLE

Dissecting cell-free DNA fragmentation variation in tumors using cell line-derived xenograft mouse

Ruiqing Fu[1☯], He Amy Su[2☯], Yi Zhao[3], Yafei Tian[4], Hongyan Chen[4], Daru Lu[4,5*]

1 Singlera Genomics (Shanghai) Ltd., Shanghai, China, **2** WLSA Shanghai Academy, Shanghai, China, **3** School of Medicine, Hangzhou City University, Hangzhou, China, **4** MOE Engineering Research Center of Gene Technology, Obstetrics and Gynecology Hospital, Fudan University, Shanghai, China, **5** State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China

☯ These authors contributed equally to this work.
* drlu@fudan.edu.cn

Check for updates

OPEN ACCESS

Citation: Fu R, Su HA, Zhao Y, Tian Y, Chen H, Lu D (2025) Dissecting cell-free DNA fragmentation variation in tumors using cell line-derived xenograft mouse. PLoS One 20(7): e0327483. https://doi.org/10.1371/journal.pone.0327483

Editor: Elizabeth Christie, Peter MacCallum Cancer Institute, AUSTRALIA

Received: August 13, 2024

Accepted: June 16, 2025

Published: July 15, 2025

Copyright: © 2025 Fu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The datasets generated have been deposited into the Genome Sequence Archive for Human (GSA-Human) of China National Genomics Data Center (NGDC) under accession number PRJCA026522 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA026522).

Abstract

Cell-free DNA (cfDNA) is increasingly studied for its diverse applications in non-invasive detection. Non-randomly cleaved by nucleases and released into the bloodstream, cfDNA exhibits a variety of intrinsic fragmentation patterns indicative of cell status. Particularly, these fragmentation patterns have recently been demonstrated to be effective in predicting cancer and its tissue-of-origin, owing to increased variation of fragmentation features observed in tumor patients. However, there remains a lack of detailed exploration of altered cfDNA fragmentation profiles in tumors, which consist of a mixture of both non-tumor cfDNA and circulating tumor DNA (ctDNA). Hence, we leveraged the human tumor cell line-derived xenograft (CDX) mouse model, where different tumor cell lines were implanted into different anatomical sites, to isolate pure ctDNA and separately investigate the fragment properties of CDX-cfDNA and ctDNA. We found an enrichment of short cfDNA fragments in both CDX-cfDNA and ctDNA compared to normal plasma cfDNA, with more elevated short fragments in ctDNA. Moreover, the CDX-cfDNA fragmentation features distinguished between CDX models of different tumor cell lines, while the ctDNA fragmentation features conversely discriminate between CDX models of different anatomical sites. The results suggested that both non-tumor cfDNA and ctDNA contribute to the increased variation observed in tumors, and that cfDNA fragmentation may be highly variable and susceptible to regulations by both original cells and cells within the local niche.

Introduction

Liquid biopsies, particularly the blood-based tests, are recently actively developed, for its promising applications in clinical oncology [1–3], prenatal testing [4,5], and organ transplant monitoring [6,7]. Among a range of blood biomarkers, the plasma

Funding: This work was supported by National Key R&D Program of China (2023YFC2705600). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: RF is employees of Singlera Genomics (Shanghai) Ltd., China. All other authors have declared no conflicts of interest. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

cell-free DNA (cfDNA) is widely studied and applied, especially for early cancer detection [8–11], owing to its non-invasive accessibility, and rich molecular characteristics, including genomic mutations [12,13], methylation alterations [14,15], and fragmentomics signatures [16,17], which could inform cell pathological status (*e.g.*, cancer cells). The genomic mutations and methylation alterations of cfDNA have been intensively investigated due to their roles in tumorigenesis [18,19] and cancer prediction, while fragmentomics is a novel feature that encompasses various properties of cfDNA fragments, including their size, end motifs and orientation, nucleosome occupancy, and topology [20–22].

The cfDNA molecule typically wraps around a nucleosome core, corresponding to ~147 bp, with an unbound linker (~20 bp) interspersed between nucleosomes [23]. During various cell death processes [24], cfDNA undergoes non-random cleavage by a series of nucleases [25], where the DNA is naked, and is released into the bloodstream, thus resulting in a distinctive distribution of fragment size, with a mode of ~167 bp [26], and preferred frequencies of fragment end motifs [27]. Meanwhile, nucleosomes that are wrapped by DNA and transcription factors that bind to DNA can protect it from cleavage, imprinting the nucleosome occupancy [26]. Similarly, other factors that modify DNA conformation, such as DNA methylation, also play a role in cfDNA fragmentation [28]. In other words, cfDNA fragments from open regions or regions depleted of other modification factors may be cleaved too shortly to be captured for sequencing by conventional sequencing approaches, thus resulting in very low genome coverage [17,29]. The majority of cfDNA originates from blood cells [5,26], while in cancer patients, it contains circulating tumor DNA (ctDNA) derived from tumor cells [30]. As a result, fragmentomics signatures can be utilized to predict cancer, as well as tissue-of-origin (TOO) in principle [26]. Additionally, copy number alterations (CNAs), which are genomic duplications or deletions frequently observed in tumor cell genomes, are basically reflected by the sequencing coverages patterns of cfDNA fragments, and thus it is commonly incorporated into fragmentomics-based approaches to detect cancer and localize TOO [31,32]. Studies have manifested an enrichment of shorter cfDNA fragments in the presence of tumors [33], accompanied by increased variation in fragment profiles among cancer patients [16]. Therefore, despite the alterations in genomic variation and methylation, the presence of tumors may also change the composition of cfDNA fragments. However, it remains to be elucidated whether such tumor-derived variation is solely contributed by ctDNA or by both ctDNA and non-tumor cfDNA. Several studies have attempted to infer TOO of cancers using fragmentomics signatures [32,34,35], but these efforts have been limited by small sample sizes and specific cancer types. There has been a lack of detailed exploration into using cfDNA fragmentation to differentiate among cell types. The natural mixture of cfDNA fragments and the low concentrations of ctDNA pose significant challenges in dissecting their properties. Thus, we leveraged cell line-derived xenograft (CDX) mouse models to isolate nearly pure ctDNA from cfDNA and separately investigated their fragmentation properties, demonstrating the influence of tumor cells on the non-tumor cfDNA, and their efficiency in differentiating cancer origins.

Results

Isolating ctDNA from non-cancer cfDNA by CDX models

To separately and accurately characterize the properties of plasma cfDNA and ctDNA, we established 12 CDX mouse models, by implanting A549 lung carcinoma (A549) and MHCC-97H liver carcinoma (M97H) cell lines into both the pancreas and rectum of nude mice (n = 3 per group). The CDX models allowed us to delineate variations in cfDNA fragmentation attributed to tumor cell lines and anatomical sites of tumorigenesis (Fig 1A). Whole blood was collected from the CDX models for plasma cfDNA extraction, followed by deep whole genome sequencing (WGS) for each model (mean coverage >38 ×, with a range of 33.18–42.78×). Additionally, three normal mice were used as controls to provide the normal background of cfDNA profile. The extracted plasma cfDNA from the CDX models comprised a mixture of non-tumor cfDNA released from mouse cells (*i.e.*, cfDNA from CDX mice, referred to as 'CDX-cfDNA') and tumor cfDNA derived from human tumor cell lines (*i.e.*, ctDNA). Several tools have been developed to analyze xenograft sequencing data, with a variable of sensitivities to detect xenograft reads [36]. With the aim of minimizing interference between CDX-cfDNA and ctDNA, while maximizing the retrieval of human-derived reads, we built a bioinformatic pipeline to distinguish between the two sources of cfDNA (see Methods). Briefly, we first generated the blacklisted regions of the human genome, which were highly similar to the mouse genome, by mapping human-prone reads from control mouse samples to the human genome. Subsequently, sequencing reads identified as human genomic fragments by either xenome [37] or XenofilteR [38] were re-aligned to the human reference genome (hg19), after which high-quality alignments that did not fall within the blacklisted regions were then used for ctDNA analysis (Fig 1 in S1 Text). With this elaborate subtraction process, we were able to isolate pure human-derived ctDNA.

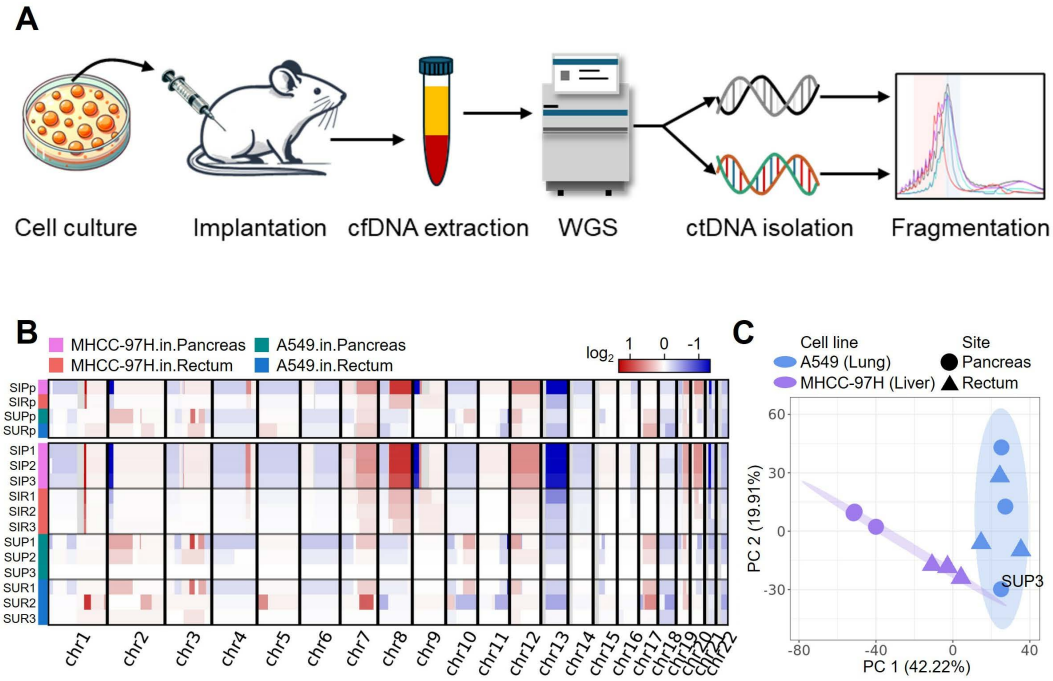


Fig 1. Schematic illustration of the study and CNA patterns in tumor cells. (A) The overview of the study design. **(B)** The CNAs detected by isolated ctDNA fragments in human-derived tumor cells from each CDX model (bottom panel) and in pooled representative samples (top panel). **(C)** Principal component analysis (PCA) on the CNA patterns for CDX models, displaying the variance explained by each PC. The colored ovals represent 95% confidence ellipses for the corresponding groups.

https://doi.org/10.1371/journal.pone.0327483.g001

The fractions of ctDNA ranged from 0.21% to 18.96% (Table A in S2 Table), corresponding to ctDNA fragments (50–1000 bp) ranging from 2.99×10^5 to 424.57×10^5 across the CDX models, with the M97H cell in pancreas model exhibiting the highest proportion of ctDNA. Human tumor cell lines are associated with distinctive CNAs [39]. Therefore, we first used ctDNA fragments to infer the large-scale CNAs, to examine the differentiation among CDX models (Table B in S2 Table). In addition to individual samples, ctDNA fragments from the replicated samples of the same CDX mouse model were pooled together to generate four representative samples, one for each cancer cell line and implantation site, to enhance the signal of ctDNA and ensure robust detection of CNAs. As expected, the ctDNA fragments manifested a diverse spectrum of CNAs across CDX models (Fig 1B), with the exception of SUP3, possibly due to its enrichment in very short fragments (Fig 7 in S1 Text), which may not be suitable for large-scale CNA detection. Interestingly, despite substantial differences between the two tumor cell lines, primarily attributing to genomic alterations of tumor cell lines, we also observed considerable variations between the two anatomical site models within each implanted cell line. Based on the CNA patterns, the models were first clustered by different cell lines (PC 1 explaining 42.22% of the variance) and then by different implantation sites (PC 2 explaining 19.91%) (Fig 1C and Fig 2 in S1 Text), suggesting potential differentiation of ctDNA fragments properties post-implantation, as CNAs were mainly inferred from the sequencing coverage of ctDNA fragments, which were susceptible to local regulations.

Fragmentation variations revealed by cfDNA from CDX models

To explore the properties of cfDNA from CDX mice potentially influenced by tumorigenesis, we first analyzed mouse cfDNA. All mouse cfDNA data were down-sampled to ~200 million reads, corresponding to an average of 995.92×10^5 and 993.87×10^5 cfDNA fragments for normal plasma samples and xenograft plasma samples, respectively. With the protection of mono-nucleosomes, the majority of cfDNA fragments typically exhibited a size of ~167 bp [26], with periodic decreases of ~10 bp (Fig 2A and Fig 3 in S1 Text). However, compared to normal plasma cfDNA, the CDX models exhibited an increased proportion of short cfDNA fragments. Therefore, we calculated the ratio between short fragments (80−160 bp) and long fragments (161−200 bp) to characterize the enrichment of short fragments of cfDNA, referred to as the S2L ratio. Interestingly, CDX models showed significantly increased short cfDNA fragments ($p = 0.048$ and 0.048 for A549 and M97H CDX models, with 95% confidence interval (95% CI) of [−0.702, −0.042] and [−0.880, −0.075], respectively; U-test) (Fig 2B and Fig 4A in S1 Text). Furthermore, the clustering of mouse cfDNA also revealed distinct fragment distributions between normal plasma samples and xenograft plasma samples, and even between the two CDX models implanted with different tumor cell lines (Fig 2C and Fig 4B and C in S1 Text). Thus, although both derived from mouse cells, cfDNA from CDX mice exhibited considerable differences from cfDNA from control mice.

To further investigate the differences between CDX models based on different tumor cell lines and different anatomical sites, we calculated several fragmentomics features of the cfDNA fragments, including breakpoint motif (BPM), end motif (EDM), and fragment size distribution (FSD). Due to the limited sample size, no significant differences ($p > 0.05$; U-test) were observed either between anatomical sites within each cell line, or between cell lines within each anatomical site. However, we found that CDX models can generally be clustered by different cell lines, using these features (Fig 3A and Fig 5 in S1 Text). Therefore, we grouped the CDX models either by tumor cells or by anatomical sites and attempted to detect informative markers (*i.e.*, motifs or FSD bins) between different groups by calculating the area under curve (AUC) values. Interestingly, we found that the distribution of marker AUCs between tumor cell lines were significantly different from that between anatomical sites ($p < 1e^{-4}$; U-test) (Fig 3B), with a significantly greater number of informative motifs and FSD bins ($AUC > 0.9$) detected between cell lines compared to those between anatomical sites ($p < 1e^{-4}$; $\chi 2$-test) (Table 1). The feature of fragment size, namely the FSD, exhibited an enhanced ability to distinguish between cell lines rather than anatomical sites. To further assess the significance of FSD in distinguishing samples, we performed permutations ($n = 1{,}000$) of the CDX models (see Methods), resulting in a significant differentiation between two tumor cell lines ($p = 0.03$) but not between two anatomical sites ($p = 0.57$) (Fig 6A in S1 Text). The FSD feature consisted of 5 bp-bins

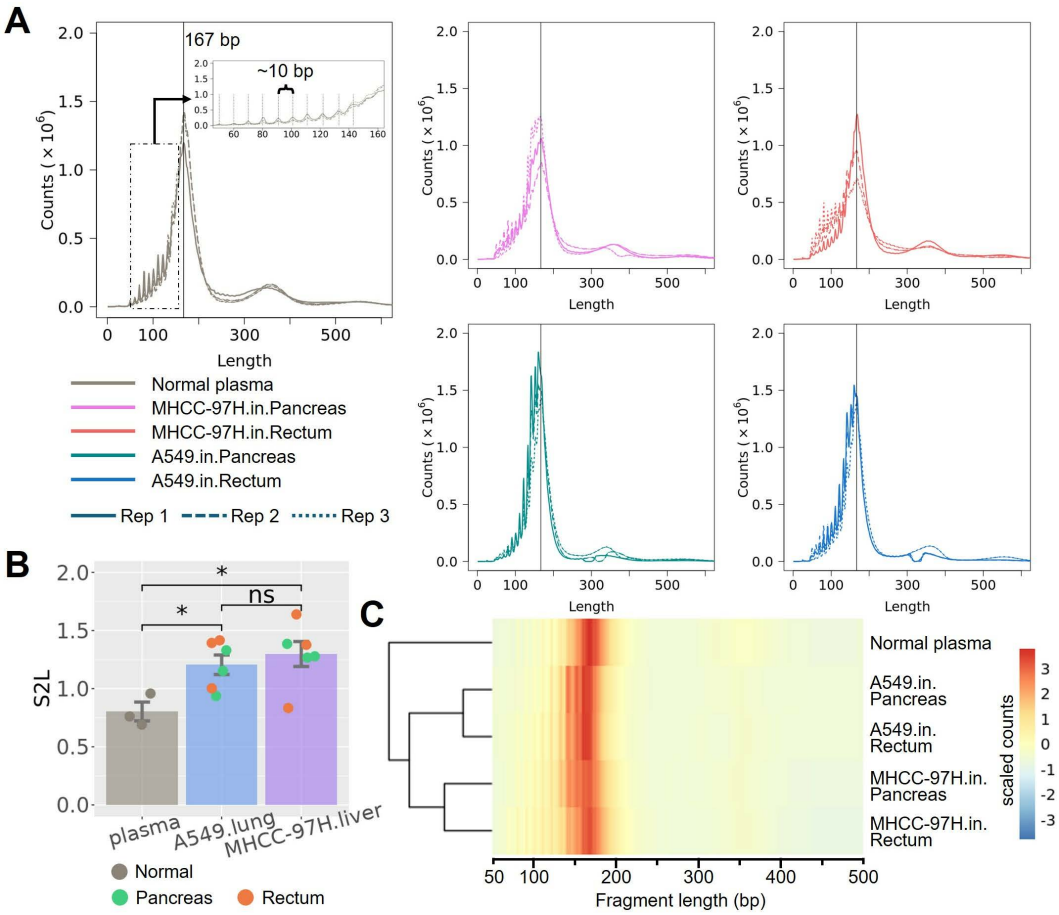


Fig 2. Fragment profile of mouse plasma cfDNA and the enrichment of short fragments. (A) The fragment size distribution of normal control plasma cfDNA, and CDX-cfDNA from four types of CDX models, with a major peak at 167 bp. **(B)** Enrichment of short fragments in CDX-cfDNA from CDX models, indicated by S2L ratio. **(C)** Hierarchical clustering of cfDNA fragment size in pooled samples. (significance: '*', $p < 0.05$; U-test).

https://doi.org/10.1371/journal.pone.0327483.g002

ranging from 65 bp to 399 bp and was calculated by counting the cfDNA fragments falling within the corresponding bin size for each individual chromosome (see Methods). Notably, we found that M97H cell CDX models exhibited enriched cfDNA fragments in the size range of 220–309 bp, while A549 cell CDX models involved more shorter cfDNA fragments ranging from 150 bp to 174 bp (Fig 3C). The results indicated that different tumor cells may result in microenvironments with considerable differences between each other.

Fragmentation variations revealed by human-derived ctDNA

We sought to investigate the potential differences between original tissues or cell types, possibly revealed by ctDNA fragments. Unlike cfDNA fragments, ctDNA was enriched in shorter fragments, with a peak at ~143 bp (Fig 4A and Fig 7 in S1 Text), consistent with previous studies [33,40]. The periodicity was also significantly shorter than that in cfDNA (7.17 bp and 9.98 bp; $p = 0.021$; paired t-test). The S2L ratio demonstrated that ctDNA involved a significantly higher proportion of short fragments than both normal plasma cfDNA ($p = 0.024$ and 0.024 for M97H cell and A549 cell CDX models, with 95% CI of [−19.381, −1.238] and [−6.567, −1.621], respectively; U-test) and CDX-cfDNA ($p = 0.031$ and 0.031 for M97H cell and A549 cell CDX models, with 95% CI of [−18.758, −0.653] and [−5.997, −1.380], respectively; paired U-test).

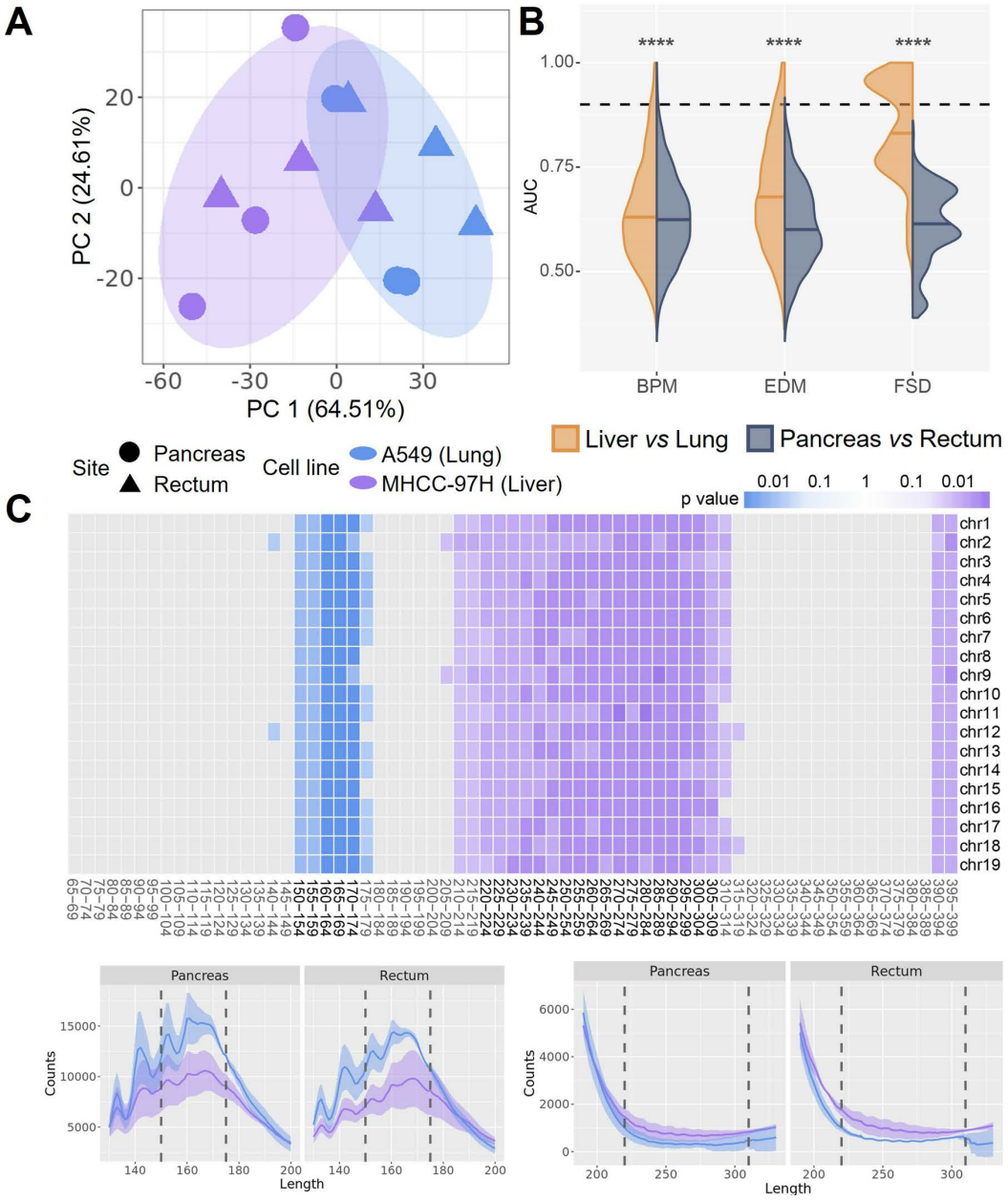


Fig 3. CDX-cfDNA distinguished between different cell line CDX models. (A) Different cell line CDX models are distinguished mainly by CDX-cfDNA fragmentation features like fragment size distribution (FSD), shown by PCA plot with the variance explained by each PC. The colored ovals represent 95% confidence ellipses for the corresponding groups. **(B)** Comparison of AUC value distributions between samples grouped by tumor cell line and by anatomical site, for fragmentation features including breakpoint motif (BPM), end motif (EDM), and FSD, respectively. The dashed line indicates AUC = 0.9.(significance: '****', $p < 1e^{-4}$; U-test) **(C)** The fragment bin sizes of FSD that are enriched in M97H CDX models (purple) and A549 CDX models (blue) (top panel), and the differentiation of the fragment profiles corresponding to the bin sizes, with the line indicates the mean value of the replicates and the shadow indicates the standard error (bottom panel). The mosaic was colored if the AUC of the corresponding FSD bin size was larger than 0.9, with the darkness indicating the p value calculated by the U-test.

https://doi.org/10.1371/journal.pone.0327483.g003

Table 1. Number of informative fragmentation features (AUC > 0.9).

	CFDNA			CTDNA		
FEATURE	BPM (n = 4096)	EDM (n = 4096)	FSD (n = 1407)	BPM (n = 4096)	EDM (n = 4096)	FSD (n = 2680)
MHCC-97H VS. A549	163	257	589	55	71	4
PANCREAS VS. RECTUM	26	3	0	157	185	278

Abbreviations: BPM, breakpoint motif; EDM, end motif; FSD, fragment size distribution.

https://doi.org/10.1371/journal.pone.0327483.t001

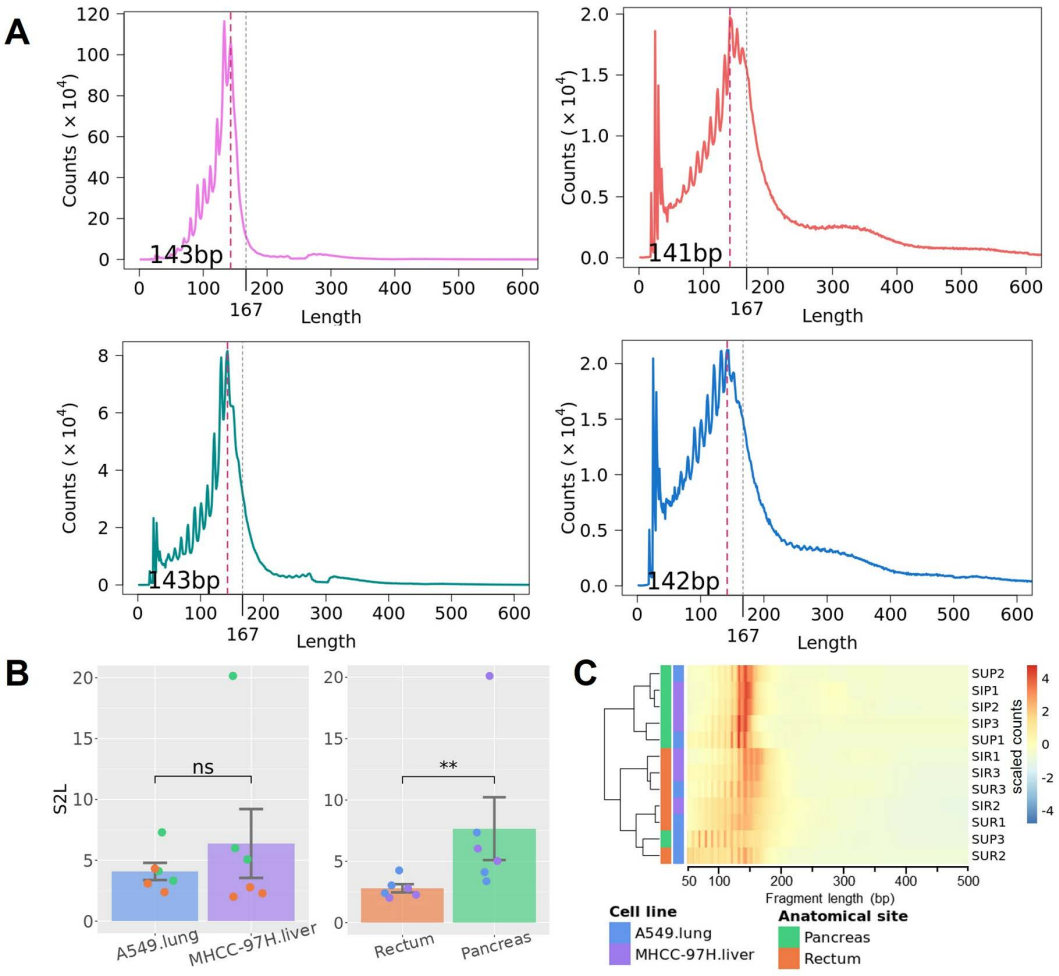


Fig 4. Fragment profile of human tumor cell-derived ctDNA and the enrichment of short fragments. (A) The fragment size distribution of human tumor cell-derived ctDNA from pooled samples representing each of the four types of CDX models, with a major peak around 143 bp. **(B)** Short ctDNA fragments are significantly more enriched in pancreas CDX models, indicated by S2L ratio. **(C)** Hierarchical clustering of ctDNA fragment size in CDX models. (significance: '*', $p < 0.05$; '**', $p < 0.01$; U-test).

https://doi.org/10.1371/journal.pone.0327483.g004

Notably, there were no significant differences between CDX models of the two cell lines, but ctDNA from pancreas models was significantly more enriched in short fragments than that from rectum models ($p = 0.0087$ and 95% CI of [−15.862, −0.969]; U-test) (Fig 4B). We further conducted clustering on the ctDNA from CDX models and found that samples from

the same anatomical sites clustered together, rather than from the same cell line (Fig 4C and Fig 8 in S1 Text). The ctDNA fragments from pancreas models were enriched in a relatively narrow range of length, while the fragments from rectum models exhibited a much wider range of length. The SUP3 sample, which was not observed with detectable large-scale CNAs, displayed an outlier pattern characterized by the highly enriched presence of very short fragments (less than 90 bp) (Fig 7 in S1 Text).

It's interesting to observe larger variations of ctDNA fragments between anatomical sites than that between cell lines. Distinct fragmentation patterns may result in different fragmentomics features, thus we also calculated BPM, EDM, and FSD for ctDNA fragments. Conversely to CDX-cfDNA, samples were mainly clustered by anatomical sites, rather than by cell lines (Fig 5A and Fig 9 in S1 Text). We also observed significantly different AUC distributions when comparing cell lines versus comparing anatomical sites ($p < 1e^{-4}$; U-test) (Fig 5B), resulting in significantly more informative motifs and FSD bins (AUC > 0.9) between pancreas and rectum models than between M97H cell and A549 cell models ($p < 0.001$; χ2-test) (Table 1). The permutation test also demonstrated a significant distinguishment between anatomical sites ($p = 0.019$) but not between tumor cell lines ($p = 0.361$) (Fig 6B in S1 Text). CtDNA from pancreas models included a higher proportion of fragments with sizes between 120–129 bp, while ctDNA from rectum models was sparsely enriched in fragments ranging from 185 bp to 214 bp (Fig 5C). These results implied that ctDNA could undergo specific fragmentation in distinct niches during tumorigenesis, resulting in distinguished fragmentomics features likely indicating its anatomical site.

Discussion

While plasma cfDNA has been widely exploited for cancer detection via genomic and epigenetic alterations, its intrinsic fragmentation features remain underexplored. Utilizing CDX mouse models, we isolated human-derived ctDNA from CDX-cfDNA, and separately investigated their properties, demonstrating that both CDX-cfDNA and ctDNA contributed to the variations of cfDNA fragments observed in tumor samples.

Based on the CDX models and control mouse plasma samples, our bioinformatic pipeline generated nearly pure mouse cfDNA fragments and human ctDNA fragments, respectively, minimizing interference with each other. To validate the robustness of the results, we also aligned the mouse cfDNA fragments to the human reference genome, generating similar fractions (2.65% *vs* 2.70%) of human-prone reads for CDX models and control mice, respectively (Table A in S1 Table), suggesting such fractions of fragments were likely to derive from homologous regions in mouse genome and/or mapping artifacts. Therefore, these fragments have a minimal influence on our analysis. The concentration of ctDNA varied largely among CDX models but had little influence on the differentiation with respect to fragmentation patterns (Fig 10 in S1 Text). Previous studies have reported an increase in short cfDNA in tumor samples but were unable to tell the source of the elevated short fragments. In the present study, it's remarkably interesting to observe an enrichment of short fragments in not only human-derived ctDNA but also mouse cfDNA in CDX models. This result suggested that the cells in the tumor niche, regardless of tumor cells, possibly underwent some modifications, which increased the accessibility of DNA and further resulted in short cfDNA fragments from non-tumor cells. Furthermore, considering the lack of significant fragmentomic differences between implantation sites, the cfDNA alterations were more likely derived from hematopoietic cells [41,42], responding to cancer-type specific signals, rather than local tissue cells at the implantation sites, which would be expected to exhibit site-specific responses if they were the dominant contributors..

It's more interesting to observe a larger variation between anatomical sites than that between tumor cell lines in ctDNA fragmentation. Tumor cells implanted in pancreas showed significantly elevated short ctDNA fragments, likely attributable to the aggressiveness of cancers promoted by the tumor microenvironment within the pancreas [43]. Notably, the CNA patterns in ctDNA mainly exhibited the differentiation between tumor cell lines, as expected, while the fragmentation features distinguished the samples between different anatomical sites, implying that fragmentation features might be more variable than genomic alterations. Therefore, they are likely influenced by post-implantation regulation of the tumor cells, or other factors in the tumor microenvironment. Furthermore, we observed converse results that CDX-cfDNA and human

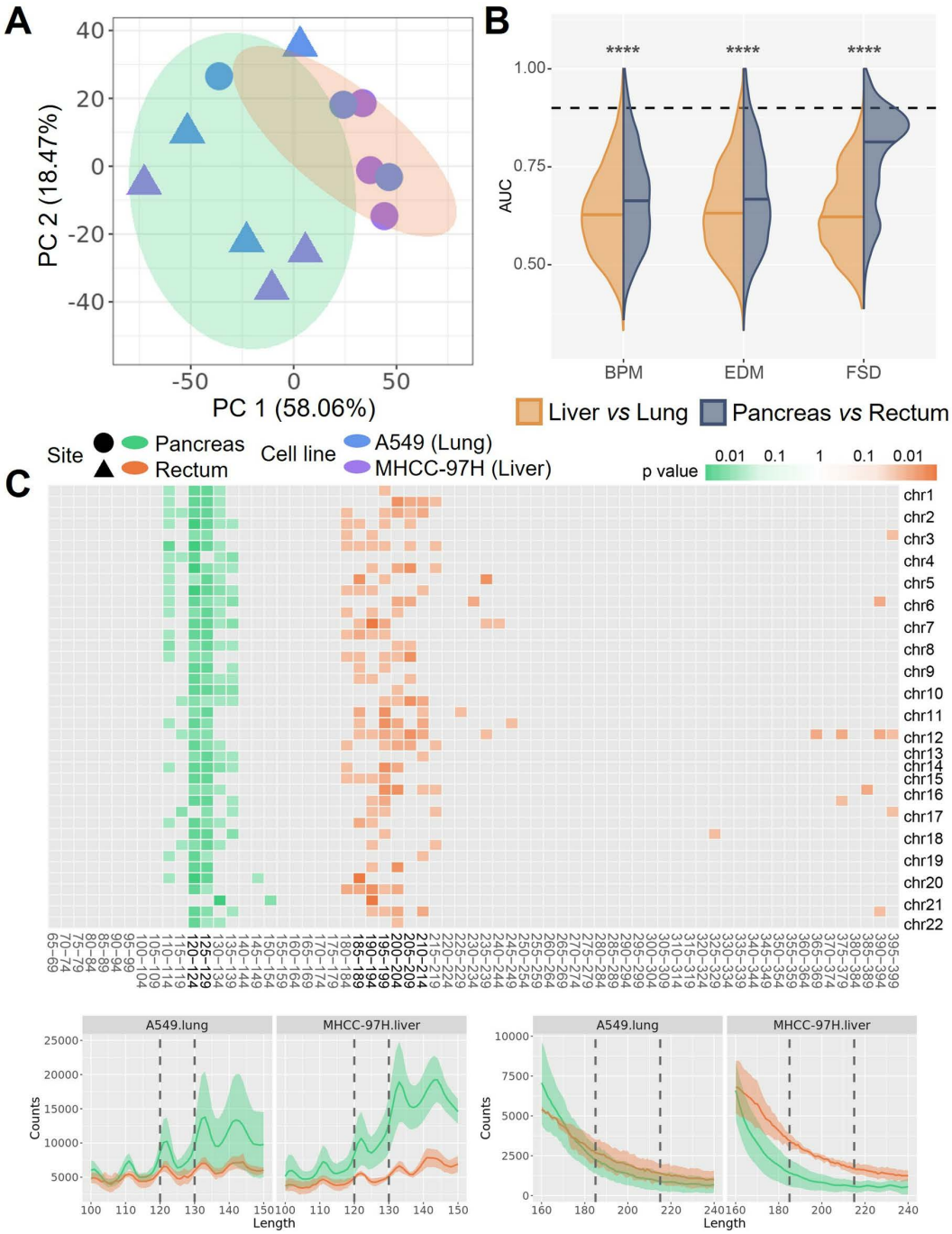


Fig 5. Human tumor cell-derived ctDNA distinguished between CDX models of different anatomical sites. (A) CDX models of different anatomical sites are distinguished by ctDNA fragmentation features like FSD, shown by PCA plot with the variance explained by each PC. The colored ovals represent 95% confidence ellipses for the corresponding groups. **(B)** Comparison of AUC value distributions between samples grouped by tumor cell line and by anatomical site, for fragmentation features including BPM, EDM, and FSD, respectively. The dashed line indicates AUC = 0.9. (significance: '****', $p < 1e^{-4}$; U-test) **(C)** The fragment bin sizes of FSD that are enriched in pancreas CDX models (green) and rectum CDX models (orange) (top panel), and the differentiation of the fragment profiles corresponding to the bin sizes, with the line indicates the mean value of the replicates and the shadow indicates the standard error (bottom panel). The mosaic was colored if the AUC of the corresponding FSD bin size was larger than 0.9, with the darkness indicating the p value calculated by the U-test.

https://doi.org/10.1371/journal.pone.0327483.g005

ctDNA could distinguish samples between different tumor cells and samples between different anatomical sites, respectively, which were confirmed by multiple fragmentation features. Hence, we hypothesized that tumor-induced short cfDNA and ctDNA might stem from different cell types and could be subject to different regulatory mechanisms, producing distinct fragmentation patterns. However, further studies are required to validate the differences between tumor-induced cfDNA and ctDNA, since the sample size was limited in the present study. Other approaches, like methylation sequencing, may be applied to confirm the molecular changes of the DNA fragments, but this is beyond the scope of this study. Besides, we admit that cfDNA fragments are vulnerable to pre-analytical factors, such as sample processing and storage. While standardized and validated protocols were employed to minimize the influence, subtle variations such as plasma preparation delays may still impact fragmentation profiles. Future studies integrating stringent quality control measures across multi-site collaborations are warranted to address these challenges comprehensively.

Altogether, taking advantage of CDX models, we were able to dissect the fragmentation features of cfDNA from CDX mice and ctDNA from human tumor cells, which was also provided as an ideal resource to study the properties of ctDNA and develop new analytical tools. We demonstrated that both the CDX-cfDNA and ctDNA contributed to the fragmentation variations observed in tumor samples, offering a promising metric for cancer prediction. However, caution is warranted in inferring TOO, since the fragmentation features may be more variable and susceptible to local regulations in tumor niches.

Methods

Establishment of cell line-derived xenograft models

Human MHCC-97H cells (liver carcinoma with high metastatic potential) and A549 cells (lung carcinoma) were cultured in DMEM and DMEM/F-12 medium, respectively, supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin and maintained at 37°C with 5.0% CO_2. The cell cultures sustained at suitable concentrations so that 5×10^6 cells and 2×10^6 cells were *in situ* injected into pancreas and rectum per mouse, respectively. A total of 12 BALB/C male nude mice, aged 6 weeks, were used for cell line implantation, with 3 replicates for each cell line and each anatomical site for injection. Tumor sizes were monitored twice per week until the mice were sacrificed for blood collection, which occurred 25–50 days post-implantation, depending on tumor growth (Fig 11 in S1 Text). Mice were deeply anesthetized using Zoletil 50 (VIRBAC Trading (Shanghai) Co., Ltd.), followed by blood collection via cardiac puncture to obtain at least 500 μl of blood, and then humanely euthanized using CO_2. All experiment procedures were conducted with the approval of the Ethics Committee of the School of Life Sciences at Fudan University.

Plasma cfDNA extraction and sequencing

Whole blood from the xenograft mice and three normal male BALB/C mice was collected in cfDNA preservation tubes (HiGiA) and immediately stored in 4°C until centrifugation. The plasma isolation was processed as previous described [44], using a two-step centrifugation method and stored at −80°C. The cfDNA (>20 ng) was then extracted from plasma using HiPure Circulating DNA Midi Kit C (Magen) according to the manufacturer's protocol and stored at −20°C. Sequencing library was constructed with all the extracted cfDNA per sample, using KAPA Hyper Prep Kit (Roche), according to the manufacturer's instructions, during which a 10 bp-UMI was liganded to 3'end of each cfDNA fragment. The Illumina NovaSeq was employed to conduct WGS, with 150 bp paired-end reads.

Sequencing data processing and ctDNA isolation

The entire WGS sequencing data (fastq) consisted of a mixture of mouse cfDNA fragments and human tumor cell-derived ctDNA fragments. A pipeline was developed to precisely separate the two types of cfDNA fragments. Raw reads were first trimmed by fastp [45] (version 0.20.1). We then employed xenome [37] and XenofilteR [38] to retrieve high-confidence human-derived reads. The high-quality reads were mapped to the mouse reference genome (mm10) and the human reference genome (hg19) using bwa [46] (version 0.7.5a-r405), respectively, after which XenofilteR was used to extract

human-derived reads by comparing human alignments to mouse alignments. The results were subsequently converted into fastq files. Meanwhile, xenome was applied to directly cluster human reads and mouse reads from the trimmed reads, separately. The resulting human-derived reads from both methods were merged and then re-aligned to the human reference genome. Additionally, to build the blacklisted regions of the human reference genome, we also attempted to generate human-alignments of the three normal mouse samples using both XenofilteR and xenome. Regions that were supported by either at least two samples or two methods were used to compile the blacklisted regions. These regions were then utilized to filter the re-alignments of the human-derived reads to generate high-confidence ctDNA fragments. All the alignments were filtered by mapping quality ≥ 30. CfDNA fragments were inferred from the paired-end information from the bam files and included in the analysis if their length fell within the range of 50–1000 bp. Reads from the replicated samples of the same tumor cell line and the same implanting anatomical site were pooled together to create four representative samples.

CNA detection and fragmentomics feature calculation

The CNAs of the tumor cell lines were inferred from ctDNA using ichorCNA [31,47], separately for each CDX model, as well as the four pooled samples. The bin size was set to 1Mb, and the $\log_2$ ratio of each bin was used to generate the CNA feature matrix, which excluded regions in a blacklist (Duke Excludable Regions) and genomic gap list. The results were in Table B in S1 Table. Other fragmentomics features, including the BPM, EDM, and FSD, were calculated following a previous study [32]. Briefly, the BPM and EDM were calculated for frequencies of all combinations of 6-bp motif (n = 4,096) flanking 3 bp of both ends and inner ends of each cfDNA fragment, respectively. The FSD was defined as the fraction of fragment sizes falling within different bins ranging from 65 bp to 399 bp, with a step of 5 bp. It was calculated for each chromosome arm of the human genome (except for chr13p, chr14p, chr15p, chr21p, chr22p and chrY). For the mouse plasma cfDNA, only q arms of the mouse genome were included for calculation.

Permutation and statistical analysis

Permutation (n = 1,000 times) was conducted on the CDX models to randomly generate two groups, each consisting of six samples. During each time, group labels (of either cell lines or anatomical sites) for the 12 CDX models were shuffled, and AUC values were calculated for each marker (*e.g.*, FSD bin) between groups based on the shuffled labels across samples. Markers with AUC > 0.8 were defined as informative, indicating excellent discriminatory capability. This process generated a null distribution of informative marker counts across 1,000 permutations. The observed number of informative markers, based on the real labels, either by tumor cell line or anatomical site, was compared to this null distribution, and the times of permutation that generate informative marker counts no less than the observed value were used to calculate the p value of the permutation test (Fig 6 in S1 Text). The 95% left-side CI for informative marker counts from the permutation was calculated using the 95% quantile.

All the statistical analyses were performed in R. The rank sum test (U-test) or t-test was used to compare two groups of samples as indicated in the main text, with BH correction applied when necessary. AUC values were calculated using 'pROC' package (version 1.18.0) in R. Specifically, for each marker, feature values and group labels across samples were supplied to *roc* and *auc* functions to generate AUC values. For each fragmentomics feature, such as FSD, the AUC value of each marker was calculated both between two cell lines and between two anatomical sites, generating two distributions of AUC values according to cell line discrimination and anatomical site discrimination. The two distributions were compared using U-test.

Supporting information

S1 Text. Supplementary figures.
(DOCX)

S1 Table. Supplementary tables.
(XLSX)

Acknowledgments

The authors thank Hua Chen, Chengcheng Ma, Minjie Xu, Wei Li, Chengxiang Gong and Zhixi Su for helpful discussions.

Author contributions

Conceptualization: He Amy Su, Yi Zhao, Daru Lu.

Data curation: He Amy Su, Yafei Tian, Hongyan Chen.

Formal analysis: Ruiqing Fu, He Amy Su.

Funding acquisition: Daru Lu.

Investigation: Daru Lu.

Methodology: Yi Zhao.

Resources: Daru Lu.

Supervision: Daru Lu.

Writing – original draft: Ruiqing Fu, He Amy Su, Daru Lu.

Writing – review & editing: Ruiqing Fu, He Amy Su, Yi Zhao, Yafei Tian, Hongyan Chen, Daru Lu.

References

1. Chan KCA, Jiang P, Chan CWM, Sun K, Wong J, Hui EP, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci U S A. 2013;110(47):18761–8. https://doi.org/10.1073/pnas.1313995110 PMID: 24191000
2. Mo S, et al. Early detection of molecular residual disease and risk stratification for stage I to III colorectal cancer via circulating tumor DNA methylation. JAMA Oncol. 2023.
3. Heitzer E, Perakis S, Geigl JB, Speicher MR. The potential of liquid biopsies for the early detection of cancer. NPJ Precis Oncol. 2017;1(1):36. https://doi.org/10.1038/s41698-017-0039-5 PMID: 29872715
4. Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, Leung TY, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A. 2008;105(51):20458–63. https://doi.org/10.1073/pnas.0810641105 PMID: 19073917
5. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112(40):E5503-12. https://doi.org/10.1073/pnas.1508736112 PMID: 26392541
6. Lo YM, Tein MS, Pang CC, Yeung CK, Tong KL, Hjelm NM. Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. Lancet. 1998;351(9112):1329–30. https://doi.org/10.1016/s0140-6736(05)79055-3 PMID: 9643800
7. Gielis EM, Ledeganck KJ, Dendooven A, Meysman P, Beirnaert C, Laukens K, et al. The use of plasma donor-derived, cell-free DNA to monitor acute rejection after kidney transplantation. Nephrol Dial Transplant. 2020;35(4):714–21. https://doi.org/10.1093/ndt/gfz091 PMID: 31106364
8. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31(6):745–59. https://doi.org/10.1016/j.annonc.2020.02.011 PMID: 33506766
9. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 2020;11(1):3475. https://doi.org/10.1038/s41467-020-17316-z PMID: 32694610
10. Liang N, Li B, Jia Z, Wang C, Wu P, Zheng T, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. Nat Biomed Eng. 2021;5(6):586–99. https://doi.org/10.1038/s41551-021-00746-5 PMID: 34131323
11. Zhang K, Fu R, Liu R, Su Z. Circulating cell-free DNA-based multi-cancer early detection. Trends Cancer. 2024;10(2):161–74. https://doi.org/10.1016/j.trecan.2023.08.010 PMID: 37709615
12. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med. 2017;9(403):eaan2415. https://doi.org/10.1126/scitranslmed.aan2415 PMID: 28814544
13. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359(6378):926–30. https://doi.org/10.1126/science.aar3247 PMID: 29348365

14. Stackpole ML, Zeng W, Li S, Liu C-C, Zhou Y, He S, et al. Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. Nat Commun. 2022;13(1):5566. https://doi.org/10.1038/s41467-022-32995-6 PMID: 36175411
15. Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet. 2017;49(4):635–42. https://doi.org/10.1038/ng.3805 PMID: 28263317
16. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385–9. https://doi.org/10.1038/s41586-019-1272-6 PMID: 31142840
17. Zhou X, Zheng H, Fu H, Dillehay McKillip KL, Pinney SM, Liu Y. CRAG: de novo characterization of cell-free DNA fragmentation hotspots in plasma whole-genome sequencing. Genome Med. 2022;14(1):138. https://doi.org/10.1186/s13073-022-01141-8 PMID: 36482487
18. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022;12(1):31–46.
19. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44. https://doi.org/10.1016/j.cell.2014.06.049 PMID: 25109877
20. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. BMC Genomics. 2015;16 Suppl 13(Suppl 13):S1. https://doi.org/10.1186/1471-2164-16-S13-S1 PMID: 26693644
21. Liu Y. At the dawn: cell-free DNA fragmentomics and gene regulation. Br J Cancer. 2021.
22. Ding SC, Lo YMD. Cell-Free DNA Fragmentomics in Liquid Biopsy. Diagnostics (Basel). 2022;12(4):978. https://doi.org/10.3390/diagnostics12040978 PMID: 35454026
23. Sanchez C, Roch B, Mazard T, Blache P, Dache ZAA, Pastor B, et al. Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight. 2021;6(7):e144561. https://doi.org/10.1172/jci.insight.144561 PMID: 33571170
24. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372(6538):eaaw3616. https://doi.org/10.1126/science.aaw3616 PMID: 33833097
25. Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, et al. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. Am J Hum Genet. 2020;106(2):202–14. https://doi.org/10.1016/j.ajhg.2020.01.008 PMID: 32004449
26. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1–2):57–68. https://doi.org/10.1016/j.cell.2015.11.050 PMID: 26771485
27. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discov. 2020;10(5):664–73. https://doi.org/10.1158/2159-8290.CD-19-0622 PMID: 32111602
28. An Y, Zhao X, Zhang Z, Xia Z, Yang M, Ma L, et al. DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. Nat Commun. 2023;14(1):287. https://doi.org/10.1038/s41467-023-35959-6 PMID: 36653380
29. Doebley AL, et al. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. Nat Commun. 2022;13(1):7475.
30. Luo H, Wei W, Ye Z, Zheng J, Xu R-H. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. Trends Mol Med. 2021;27(5):482–500. https://doi.org/10.1016/j.molmed.2020.12.011 PMID: 33500194
31. Wan N, Weinberg D, Liu T-Y, Niehaus K, Ariazi EA, Delubac D, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer. 2019;19(1):832. https://doi.org/10.1186/s12885-019-6003-8 PMID: 31443703
32. Bao H, Wang Z, Ma X, Guo W, Zhang X, Tang W, et al. Letter to the Editor: An ultra-sensitive assay using cell-free DNA fragmentomics for multi-cancer early detection. Mol Cancer. 2022;21(1):129. https://doi.org/10.1186/s12943-022-01594-w PMID: 35690859
33. Thierry AR. Circulating DNA fragmentomics and cancer screening. Cell Genom. 2023;3(1):100242.
34. Siejka-Zielińska P, Cheng J, Jackson F, Liu Y, Soonawalla Z, Reddy S, et al. Cell-free DNA TAPS provides multimodal information for early cancer detection. Sci Adv. 2021;7(36):eabh0534. https://doi.org/10.1126/sciadv.abh0534 PMID: 34516908
35. Helzer KT, et al. Fragmentomic analysis of circulating tumor DNA-targeted cancer panels. Ann Oncol. 2023.
36. Jo S-Y, Kim E, Kim S. Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis. Genome Biol. 2019;20(1):231. https://doi.org/10.1186/s13059-019-1849-2 PMID: 31707992
37. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, et al. Xenome--a tool for classifying reads from xenograft samples. Bioinformatics. 2012;28(12):i172-8. https://doi.org/10.1093/bioinformatics/bts236 PMID: 22689758
38. Kluin RJC, Kemper K, Kuilman T, de Ruiter JR, Iyer V, Forment JV, et al. XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. BMC Bioinformatics. 2018;19(1):366. https://doi.org/10.1186/s12859-018-2353-5 PMID: 30286710
39. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3 PMID: 31068700
40. Jiang P, Lo YMD. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends Genet. 2016;32(6):360–71. https://doi.org/10.1016/j.tig.2016.03.009 PMID: 27129983
41. Mattox AK, et al. The origin of highly elevated cell-free DNA in healthy individuals and patients with pancreatic, colorectal, lung, or ovarian cancer. Cancer Discov. 2023.
42. Li S, et al. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proc Natl Acad Sci U S A. 2023;120(28):e2305236120.

43. Espinet E, Klein L, Puré E, Singh SK. Mechanisms of PDAC subtype heterogeneity and therapy response. Trends Cancer. 2022;8(12):1060–71. https://doi.org/10.1016/j.trecan.2022.08.005 PMID: 36117109

44. Mo S, Dai W, Wang H, Lan X, Ma C, Su Z, et al. Early detection and prognosis prediction for colorectal cancer by circulating tumour DNA methylation haplotypes: A multicentre cohort study. EClinicalMedicine. 2023;55:101717. https://doi.org/10.1016/j.eclinm.2022.101717 PMID: 36386039

45. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086

46. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698 PMID: 20080505

47. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8(1):1324. https://doi.org/10.1038/s41467-017-00965-y PMID: 29109393