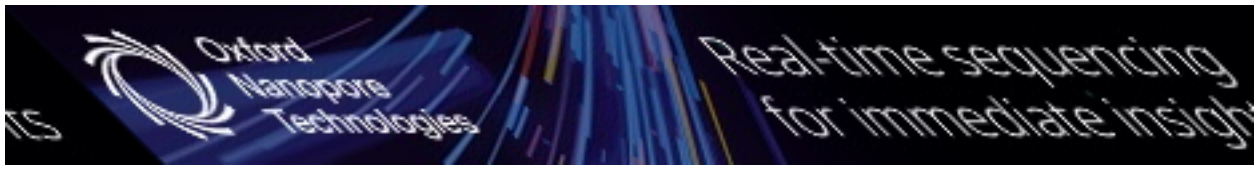
Analysis of a cell-free DNA-based cancer screening cohort links fragmentomic profiles, nuclease levels, and plasma DNA concentrations

Yasine Malki, Guannan Kang, Wai Kei Jacky Lam, et al.

Genome Res. published online November 27, 2024
Access the most recent version at doi:10.1101/gr.279667.124



Analysis of a cell-free DNA-based cancer screening cohort links fragmentomic profiles, nuclease levels and plasma DNA concentrations

Yasine Malki[1,2,3,5], Guannan Kang[1,2,3,5], W. K. Jacky Lam[1,2,3,4,5], Qing Zhou[1,2,3], Suk Hang Cheng[1,2,3], Peter P.H. Cheung[2,3], Jinyue Bai[1,2,3], Ming Lok Chan[2,3], Chui Ting Lee[2,3], Wenlei Peng[1,2,3], Yiqiong Zhang[2,3], Wanxia Gai[1,2,3], Winsome W.S. Wong[1,2,3], Mary-Jane L. Ma[1,2,3], Wenshuo Li[1,2,3], Xinzhou Xu[2,3], Zhuoran Gao[2,3], Irene O. L Tse[2,3], Huimin Shang[1,2,3], L.Y. Lois Choy[1,2,3,4], Peiyong Jiang[1,2,3,4], K. C. Allen Chan[1,2,3,4], Y.M. Dennis Lo[1,2,3,4*]

[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, Sir Y.K. Pao Centre for Cancer, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China.

[5]These authors contributed equally to this work.

*Corresponding author: loym@cuhk.edu.hk.

Manuscript type: Research Article

Running title: Fragmentomics, nucleases and cfDNA concentration

Keywords: Fragmentomics, cell-free DNA, liquid biopsy, nuclease

Abstract

The concentration of circulating cell-free DNA (cfDNA) in plasma is an important determinant of the robustness of liquid biopsies. However, biological mechanisms that lead to inter-individual differences in cfDNA concentrations remain unexplored. The concentration of plasma cfDNA is governed by an interplay between its release and clearance. We hypothesize that cfDNA clearance by nucleases might be one mechanism that contributes towards inter-individual variations in cfDNA concentrations. We performed fragmentomic analysis of the plasma cfDNA from 862 healthy individuals, with a cfDNA concentration range of 1.61 – 41.01 ng/mL. We observed an increase in large DNA fragments (231-600 bp), a decreased frequencies of shorter DNA fragments (20-160 bp), and an increased frequency of G-end motifs with increasing cfDNA concentrations. End motif deconvolution analysis revealed a decreased contribution of DNase1L3 and DFFB in subjects with higher cfDNA concentration. The five subjects with the highest plasma DNA concentration (top 0.58%) had aberrantly decreased levels of DNase1L3 protein in plasma. The cfDNA concentration could be inferred from the fragmentomic profile through machine learning, and was well correlated to the measured cfDNA concentration. Such an approach could infer the fractional DNA concentration from particular tissue types, such as the fetal and tumor fraction. This work shows that individuals with different cfDNA concentrations were associated with characteristic fragmentomic patterns of the cfDNA pool; and that nuclease-mediated clearance of DNA is a key parameter that affects cfDNA concentration. Understanding these mechanisms has facilitated the enhanced measurement of cfDNA species of clinical interest, including circulating fetal and tumor DNA.

Introduction

Much research effort has been made in obtaining diagnostic information from cell-free DNA (cfDNA) to investigate various physiological and pathological states in the context of pregnancy, oncology, and organ transplantation (Lo et al. 2021). There has also been growing research interest in understanding the biological life cycle of cfDNA molecules in circulation. Several key mechanisms of cfDNA release have been proposed, including the different modes of cell death, the release of neutrophil extracellular traps, and active cellular secretion (Grabuschnig et al. 2020; Heitzer et al. 2020; Han and Lo 2021). The rapid clearance of cfDNA has been demonstrated by studying the kinetics of fetal cfDNA in pregnant women after delivery (Lo et al. 1999; Yu et al. 2013). The measurement of cfDNA concentration in plasma at a given time reflects upon an interplay between DNA release from cells and clearance from circulation.

Many pathophysiological conditions are characterized by an altered equilibrium of plasma cfDNA concentration. An elevation of cfDNA concentration has been observed in patients with different cancers (Mattox et al. 2023), systemic lupus erythematosus (Tug et al. 2014) and infectious diseases (Han et al. 2020a; Cheng et al. 2021), when compared to healthy controls. Performing physical exercise, such as a 40-minute run, could lead to a mean increase of 18-fold in the total cfDNA concentration (Fridlich et al. 2023). The concentration of cfDNA serves as an important parameter in liquid biopsy, as it affects the sensitivity of disease diagnosis and the reproducibility of quantitative measurements. An enhanced understanding of the production and clearance of cfDNA may give rise to novel diagnostic approaches with greater sensitivity for early disease detection. Such a postulation has been recently supported by a study, in which priming agents were developed to inhibit the clearance processes of cfDNA (Martin-Alonso et al. 2024). Administration of these priming agents prior to blood collection increased the recovery of

circulating tumor DNA by over 10-fold, as demonstrated in a preclinical model of mice with lung cancer (Martin-Alonso et al. 2024).

It is currently understood that part of the clearance process of cfDNA involves enzymatic degradation by the action of DNA nucleases. It has been revealed that fragmentomic features, such as fragment sizes and end motifs, are linked to the activity of these nucleases (Serpas et al. 2019; Chan et al. 2020; Han et al. 2020b; Han and Lo 2021; Lo et al. 2021; Chen et al. 2022; Zhou et al. 2023). For instance, deoxyribonuclease 1 like 3 (DNase1L3) preferentially cleaves DNA in a manner that results in a preponderance of C-ends (Serpas et al. 2019; Chan et al. 2020; Han et al. 2020b), whereas deoxyribonuclease 1 (DNase1) preferentially cleaves DNA into T-ends (Chen et al. 2022). Our work has revealed that DNase1L3 plays a critical role in the fragmentation process of cfDNA in plasma, as evidenced by the aberrant plasma DNA fragmentation patterns in DNase1L3-deficient mouse model and human subjects (Serpas et al. 2019; Chan et al. 2020), and the presence of DNase1L3 protein in plasma (Chen et al. 2022). We also identified DNA fragmentation factor subunit beta (DFFB) as a major player in the fragmentation of newly released DNA from dying cells (Han et al. 2020b). We hypothesized that the process of nuclease-mediated fragmentation is linked to the total concentration of cfDNA.

Previous reports have shown considerable variation in the plasma cfDNA concentration in healthy individuals (Alborelli et al. 2019; Meddeb et al. 2019; Orntoft et al. 2021; Zhu et al. 2023). We reason that characteristic fragmentomic patterns in the cfDNA pool between subjects with different cfDNA concentrations may hold clues to mechanisms that regulate the levels of cfDNA in plasma. In the present study, we have analyzed the distribution of plasma cfDNA concentrations in 862 individuals. These subjects are part of a cohort of individuals who had

undergone screening for the early detection of nasopharyngeal carcinoma (NPC) in Hong Kong (Chan et al. 2017; Chan et al. 2023). The NPC screening was performed through the detection of Epstein–Barr virus (EBV) DNA in the circulating cfDNA pool. This cohort was used to investigate whether the overall concentration of circulating DNA in plasma is associated with changes in fragmentomic features, which might provide hints towards nuclease-mediated fragmentation, or other factors, that contribute to inter-individual differences in cfDNA concentrations.

Results

Subject cohort characteristics and cfDNA concentration distribution

The aim of the study was to evaluate fragmentomic characteristics of cfDNA in individuals with different concentrations of plasma cfDNA. We therefore retrieved cfDNA concentration values from individuals in a cancer screening cohort and analyzed the size profiles, end motif patterns and nuclease contributions from sequencing data of plasma cell-free DNA (Fig. 1). The individuals studied were participants who had undergone screening for NPC (Chan et al. 2017; Chan et al. 2023). These subjects underwent plasma EBV DNA testing by real-time polymerase chain reaction (PCR) for NPC screening. The screening trial was conducted between the years 2017-2020. All subjects were ethnically Chinese men.

Individuals who were tested as EBV-positive were subjected to targeted sequencing of plasma EBV DNA as a reflex test to enhance the specificity of NPC detection (Lam et al. 2018). We retrieved and analyzed the targeted sequencing data of 862 EBV-positive subjects to explore the relationship between fragmentomic features and cfDNA concentrations. To minimize the potential effects of target capture on the fragmentomic analysis, only ‘off-target’ reads were used

to study fragmentomic features of cfDNA (see Methods). The median age of the 862 subjects was 54 years (interquartile range of 49-57 years).

The distribution of cfDNA concentration is shown in Fig. 2. A large variation was observed in the cfDNA concentration among subjects, with values ranging from 1.61 – 41.01 ng/mL, and therefore a 25.5-fold difference between subjects with the highest and lowest plasma cfDNA concentration. The median cfDNA concentration was 7.39 ng/mL. The skewness of the plasma DNA concentration was 2.043 ($P < 0.001$), and the kurtosis of the distribution was 6.974 ($P < 0.001$). The high positive value of skewness (beyond the range of +1 and -1) indicates a positively skewed distribution of plasma DNA concentrations. The kurtosis value from our cohort represents a leptokurtic distribution (above 3), generally indicating a large proportion of individuals on both extremes of the distribution spectrum. The values of plasma cfDNA concentration deviates from a normal distribution (Shapiro-Wilk test: $P < 0.001$). We have demonstrated that plasma cfDNA concentrations can be fitted into a gamma distribution (Supplemental Fig. S1A). The significant skewness of plasma cfDNA concentration, particularly in subjects with higher cfDNA concentrations, may indicate potential biological or physiological factors that may play a role in modulating cfDNA concentration, resulting in large deviations of cfDNA concentrations from the median in some individuals.

To further validate the findings and conclusions, we sequenced and analyzed an additional 497 individuals from the same clinical study. These individuals had no detectable EBV-DNA in plasma, referred to as ‘EBV-negative cohort’. The subject characteristics of both cohorts are summarized in Supplemental Table 1. The cfDNA concentration ranged from 1.62 – 25.25 ng/mL, with a median of 7.56 ng/mL, consistently exhibiting a positively skewed distribution

(Supplemental Fig. S2). The skewness of the cfDNA concentration from this cohort was 1.461 ($P < 0.001$), kurtosis was 3.537 ($P < 0.001$), and the Shapiro-Wilk test resulted in $P < 0.001$. The distribution of cfDNA concentrations also fit a gamma distribution function (Supplemental Fig. S1B).

Quantification of the plasma cfDNA concentration was performed on all subjects of the cohort using fluorometric measurements (see Methods). Fluorometric methods measure the overall DNA concentration, including DNA from both nuclear and mitochondrial origins. The readings from fluorometric measurements ($n = 25$) were well correlated with results obtained from the droplet digital PCR assay, which targeted a single copy gene valosin-containing protein (VCP) (Supplemental Fig. S3). This highlights the reliability of using fluorometric methods to reflect the total cfDNA of nuclear origin.

Fragmentomic study 1 – Size profile analysis

One goal of this study was to examine how fragmentomic features, namely the size distribution and end motif patterns of cfDNA, varied among individuals with different concentrations of cfDNA in plasma. Fig. 3*A*-*B* show the overall size distribution profile of plasma cfDNA from the highest 10% of individuals (red) and lowest 10% of individuals (blue), and median distribution (black) in terms of cfDNA concentration. Plasma DNA of all subjects exhibited a modal peak size of approximately 166 bp, consistent with reports investigating the mono-nucleosome units of cfDNA. The logarithmic plot shows the presence of the di- and tri-nucleosome peaks, with sizes of approximately 350 bp and 520 bp.

Compared to subjects with low cfDNA concentrations, subjects with high cfDNA concentrations appeared to have increased frequencies of DNA fragments above 250 bp, with an enhancement in the di- and tri-nucleosomal peaks. The increased frequency for fragments above 250 bp in subjects with high cfDNA concentration resembled previous findings in DNase1L3-deficient mice, which exhibited an elevated frequency of di- and tri-nucleosomal DNA in plasma (Serpas et al. 2019). These results suggested that subjects with high cfDNA concentrations might exhibit an impaired DNase1L3-mediated fragmentation process. Short DNA fragments within the size range of 20-120 bp were decreased in subjects with high cfDNA concentrations. These shorter fragments might represent intermediate products (i.e., sub-nucleosomal DNA) derived from mono-nucleosomal DNA during the degradation process. Size distribution plots of different cfDNA concentration ranges (<5, 5-10, 10-15, ≥15 ng/mL) also show this gradient pattern of decreased frequency of DNA fragments within the 20-120 bp range, and increased frequency of DNA fragments above 250 bp (Supplemental Fig. S4).

We next studied the size profile frequencies in all 862 subjects. The frequencies of cfDNA fragments for each 10 bp bin size were measured and correlated to cfDNA concentrations. As examples, the frequency of cfDNA fragments between 81-90 bp was negatively correlated to cfDNA concentration (Pearson r = -0.65, P < 0.0001, Fig. 3*C*), while the frequency of cfDNA between 301-310 bp was positively correlated (Pearson r = 0.42, P < 0.0001, Fig. 3*D*). As shown in Fig. 3*E*, the cfDNA frequencies in bins below 160 bp were found to be negatively correlated to cfDNA concentrations, while bins above 230 bp size were positively correlated to cfDNA concentrations. The proportional quantification of DNA molecules within certain size ranges may serve as a parameter to approximate the degradation rate of cfDNA in plasma. The process of apoptosis releases DNA with a wide range of sizes (Ungerer et al. 2022; Zhu et al. 2023;

Davidson et al. 2024), while subsequent cell-extrinsic cleavage occurs via nucleases in blood, namely DNase1L3 (Serpas et al. 2019). Factors that affect the degradation of cfDNA would alter the proportion of DNA molecules within certain size ranges. The overall larger size profiles in individuals with high cfDNA concentrations might suggest decreased activity of extracellular DNA clearance in blood.

Fragmentomic study 2 – End motif analysis

We have shown that the concentration of cfDNA was associated with changes in the size distribution of the cfDNA pool. We next investigated whether the distribution of 5' end motifs varied with cfDNA concentration (see Methods for end motif analysis). To this end, we systematically studied the correlations of 256 end motifs (4-mer) to the plasma cfDNA concentrations. A total of 79 end motifs were found to be significantly correlated with cfDNA concentrations, with 34 motifs being negatively correlated and 45 being positively correlated after adjustment for multiple comparisons using Bonferroni's correction (Supplemental Tables 2-3). Motifs with a negative correlation were mostly C-end motifs (such as CACT, CATC, CACC), with several T-end and A-end motifs, while the positively correlated motifs were nearly all G-end motifs (such as GCAA, GAAC, GGCA).

We performed heatmap analysis to visualize the pattern of the significantly correlated motifs to cfDNA concentration, as shown in Fig. 4*A*. Each row corresponds to a particular 4-mer motif, with the composition of the first base indicated, while columns represent plasma DNA from each subject. For better visualization, column-wise normalization (*z*-score) was applied to the motif frequencies. A gradual change in motif frequencies was observed across the whole spectrum of cfDNA concentrations, with 4-mer end motifs starting with a 5' C nucleotide appearing to

gradually decrease as the cfDNA concentration increased. However, the 4-mer end motifs starting with a 5' G nucleotide showed a rising gradation. The most pronounced difference in end motifs frequencies was observed in cfDNA concentration ranges of < 5 ng/mL and ≥ 15 ng/mL (Supplemental Fig. S5). The observed trends in fragmentomic features were also present in the highest and lowest 10 subjects of the EBV-negative cohort (Supplemental Fig. S6). By evaluating the highest and lowest 10% of individuals from the main study cohort, the selected 79 significantly correlated motifs also showed a consistent trend when sub-classified into *Alu* regions, CpG islands and gene body regions (Supplemental Fig. S7), suggesting that these changes in end motif profiles appear to be present across different genomic regions. Furthermore, eliminating the highest and lowest 10% of the cohort yielded similar results regarding size distribution and end motif patterns (Supplemental Fig. S8), indicating that the observed patterns were not exclusively influenced by the extreme values. Of note, fragmentomic features derived from the 'off-targeted' reads in the targeted sequencing dataset were in good agreement with features in those paired samples based on genome-wide paired-end sequencing (Supplemental Fig. S9 and S10). These results demonstrated the validity of using 'off-target' reads in our fragmentomic analysis.

Motivated by the gradation patterns observed in the heatmap, we explored whether we could use regression to model the relationship between cfDNA concentration and 4-mer end motifs. We adopted a support vector regression (SVR) model, with a random selection of 50% of subjects used as the training set, and the remaining 50% as the testing set. Indeed, we observed a correlation between the cfDNA concentration predicted by the SVR model and the cfDNA concentration measured by fluorometric quantification in the validation set (Pearson $r = 0.72$, $P < 0.0001$, Fig. 4*B*). Notably, by utilizing the size profile in addition to end motif frequencies, the correlation could be further improved, with a Pearson r of 0.82 ($P < 0.0001$, Fig. 4*C*). In addition,

we used the 10 subjects with the lowest cfDNA concentrations and the 10 subjects with the highest cfDNA concentrations of the EBV-negative cohort as an additional test set for the trained model, which also showed a strong positive correlation between predicted and measured cfDNA concentration (Pearson $r = 0.94$, $P < 0.0001$, Supplemental Fig. S11). Hence, these results further validate that the overall cfDNA concentration was associated with characteristic fragmentomic patterns in cfDNA pool.

Deconvolutional analysis of end motifs

Our group has previously reported on six distinct cleavage patterns via a non-negative matrix factorization (NMF) algorithm, termed "founder" end motif profiles (F-profile), which were linked to different biological processes (Zhou et al. 2023). F-profiles I, II and III represent the contribution of DNase1L3, DNase1 and DFFB, respectively. F-profile IV is a cleavage profile with a high C-end preference (e.g., CG-end preference), while F-profile V exhibit a strong G-end preference. Profile VI represents non-specific cleavage patterns, speculated to originate from chemical factors such as oxidative stress. We investigated whether the F-profile contributions varied between subjects of different cfDNA concentrations.

We stratified the DNA fragments into three size ranges for the end motif based deconvolutional analysis, which were 20-160 bp, 161-230 bp and 231-600 bp; selected based on the correlation to cfDNA concentration in Fig 3*E*. We reasoned that each size range might result from different stages of the fragmentation process of DNA in plasma. We visualized the differences in F-profile contributions with different cfDNA concentrations using a heatmap analysis, with *z*-score normalization applied to each F-profile. Of the three size ranges, the 231-600 bp size range showed the most distinct variation in F-profile contributions with cfDNA concentration (Fig. 5).

The contribution of F-profile I (DNase1L3), II (DNase1) and III (DFFB) exhibited a gradation pattern with increased cfDNA concentration. In contrast, F-profile V showed a contrasting trend with cfDNA concentration, consistent with the higher frequency of G-end motifs associated with increased cfDNA concentration. Overall, the larger plasma DNA fragments with sizes around the di- and tri-nucleosome peaks had a higher frequency of G-end motifs, and decreased end motif frequencies from DNase1L3, DNase1 and DFFB. No particular gradation pattern was observed amongst all six F-profiles in the 21-160 bp size range (Supplemental Fig. S12). The gradation patterns in F-profiles I, III and V were also observed in the 161-230 bp size range (Supplemental Fig. S12).

A multiple linear regression analysis was performed on all six F-profiles across three size ranges to determine which profiles were significantly associated with cfDNA concentration (Supplemental Table 4). The analysis revealed that F-profiles I (DNase1L3), III (DFFB), and V (G-ends) of molecules within the 231-600 bp range, and F-profile V within the 161-230 bp range, were significantly associated with the cfDNA concentration. These results provide evidence that DNase1L3 and DFFB are factors that may regulate the concentration of circulating cfDNA in plasma.

Concentration of DNase1L3 protein in plasma

Fragmentomic analysis suggested that the activity of DNA nucleases might be at least in part responsible for regulating the concentration of cfDNA in plasma. Previous work has revealed that DNase1L3 is secreted by macrophages and dendritic cells into plasma circulation (Shiokawa and Tanuma 1998; Sisirak et al. 2016), and plays an important role in the cell-extrinsic fragmentation of cfDNA (Sisirak et al. 2016; Chan et al. 2020). We therefore developed an assay to quantify the

concentration of DNase1L3 in plasma, to determine whether subjects with different cfDNA concentrations were associated with varied levels of DNase1L3. We focused primarily on the subjects with the highest and lowest cfDNA concentrations, to evaluate whether such a difference in DNase1L3 protein levels was observed.

As a result, the highest five subjects showed a significantly decreased level of DNase1L3 protein in plasma compared to the lowest five subjects (52.5% reduction in DNase1L3 levels; $P = 0.0079$, Wilcoxon test, Supplemental Fig. S13 and S14). However, the significance of this trend was diminished when extending to the highest and lowest ten subjects (14.5% reduction in DNase1L3 levels; $P = 0.2475$, Wilcoxon test, Supplemental Fig. S13 and S14). It is possible that only subjects with particularly elevated levels of cfDNA in plasma, such as those subjects exceeding 30 ng/mL, were associated with decreased levels of DNase1L3 in plasma. Nevertheless, this analysis suggested that the levels of DNase1L3 in plasma might partially account for the variation in cfDNA concentrations among different individuals.

Tissue-of-origin analysis of cfDNA in individuals with different cfDNA concentration

As we observed that the activity of nucleases was linked to cfDNA concentration, we questioned whether the variation in cfDNA concentration might be associated with the release of cfDNA from a particular cellular source. We employed the fragmentomics-based methylation analysis (FRAGMA) to deduce the methylation status of cfDNA (Zhou et al. 2022). We selected cell-type specific hyper- and hypomethylated CpG sites for six cell types – liver, neutrophils, B cells, T cells, erythroblasts, and megakaryocytes, as they represented major cellular sources contributing to the cfDNA pool (Loyfer et al. 2023). The deduced tissue contribution was assessed based on methylation status at these CpG sites (see Methods), and was correlated to the cfDNA

concentration (Supplemental Fig. S15). The results showed that the deduced contribution of each cell type was not significantly correlated to cfDNA concentration, with a weak correlation coefficient (Pearson $r < 0.1$ for all cell types). Hence, the production of cfDNA from various cell types might not be an important factor contributing to the variation in cfDNA.

Follow-up collection of subjects with the extreme cfDNA concentrations

To investigate whether the cfDNA concentration might persist in subjects with the highest and lowest cfDNA concentrations, we contacted a total of 40 subjects (the highest and lowest 10 from both cohorts) for a follow-up blood collection. Samples from 26 individuals were obtained, as the remaining individuals either developed cancers, were lost to follow-up, or refused participation. The median interval time between blood collection from our existing cohorts for NPC screening and blood re-collection was 76 months (interquartile range of 64 - 77 months).

The concentrations of cfDNA were well correlated between the two collection points for each subject (Pearson $r = 0.75$, $P < 0.0001$, Supplemental Fig. S16A). The difference in cfDNA concentration between the high and low cfDNA groups was statistically significant ($P < 0.0001$, Supplemental Fig. S16B). A comparison of the fragmentomic profiles between paired matched samples from the two collection time points showed a significant correlation in the proportion of DNA fragments within 81-90 bp and 301-310 bp (Supplemental Fig. S17A and B), and in the pattern of end motif frequencies from the selected 79 end motifs from Fig. 4A (Supplemental Fig. S17C and D). Overall, this suggests that cfDNA concentration might be mediated by certain intrinsic physiological factors, leading to persistent extremes spanning a median collection interval of around six years.

Relationship between clinical laboratory parameters and cfDNA concentration

We further explored the possible correlation of other biochemical parameters in blood with the concentration of plasma cfDNA. Laboratory tests were performed on these individuals from the follow-up collection (n = 26 individuals), which included blood cell counts (red blood cells, neutrophils, lymphocytes, monocytes, eosinophils, and platelets), the concentration of plasma proteins [total plasma proteins, albumin, globulins, alkaline phosphatase (ALP), alanine aminotransferase (ALT), C-reactive proteins (CRP)] and non-protein components (sodium, potassium, bilirubin, urea, creatinine). Data for each parameter were correlated with the concentration of plasma cfDNA, with a false discovery rate (FDR) adjustment for multiple comparisons (Supplemental Table 5). Multiple linear regression was also performed to study the effects of plasma protein components on cfDNA concentration (Supplemental Table 6).

The correlation analyses revealed that the levels of alanine aminotransferase (ALT, Pearson r = 0.676, P = 0.0024), C-reactive protein (CRP, Pearson r = 0.595, P = 0.0168) and total plasma proteins (Pearson r = 0.549, P = 0.0296) were significantly correlated to plasma cfDNA concentrations (Supplemental Table 5, Supplemental Fig. S18). The multiple linear regression analysis revealed that only ALT (P = 0.0028) and CRP (P = 0.0095) were significant independent variables affecting plasma cfDNA concentrations (Supplemental Table 6). The levels of ALT and CRP from these subjects were within the reference intervals (ALT < 53 IU/L; CRP < 9.9 mg/L, Supplemental Fig. S18). These results demonstrate that plasma cfDNA levels were also associated with physiological factors such as liver function and inflammation.

Deduction of fractional DNA concentrations from different cell types using fragmentation patterns

As the total cfDNA concentration of individuals could be predicted using fragment sizes and end motifs, we investigated whether such fragmentomic features could be applied to subsets of cfDNA species. We used circulating tumor DNA in HCC patients and circulating fetal DNA in the plasma of pregnant women as examples, where fragmentomic features were used to predict tumoral and fetal DNA fraction, respectively (Fig. 1). Our previous work has demonstrated that fetal and tumor-derived DNA have a shorter modal size of 143 bp (Yu et al. 2014; Jiang et al. 2015; Jiang and Lo 2016). This may be attributed to differences in chromatin packing and methylation density (Sun et al. 2018; Pastor and Kwon 2022). Characteristic changes in the end motif profiles have also been observed in cancer and pregnancy; for example, decreased frequencies of DNase1L3 representative motifs such as CCCA in patients with HCC (Jiang et al. 2020). As such, an alteration in fetal or tumoral DNA fraction would lead to changes in fragmentomic features of the cfDNA pool (Yu et al. 2014; Jiang et al. 2015; Yu et al. 2017; Jiang et al. 2020). We reasoned that the integrative analysis of the whole spectrum of fragment sizes and end motifs would enable more accurate prediction of fetal or tumoral DNA fraction.

The size profile and end motif profiles of cfDNA from 30 pregnant women (including first, second and third trimester cases, all from the '1st cohort' in Supplemental Table 7) were used to train an SVR model. We adopted the leave-one-out procedure (See Methods), and compared the predicted fetal fraction (%) by fragmentomic features to the fetal fraction measured using informative single nucleotide polymorphisms (SNPs). The predicted fetal fraction was in good agreement with the SNP-based fetal DNA fraction (Pearson r = 0.81, P < 0.0001, Fig 6*A*). The magnitude of correlation exceeded a previous approach which employed the motif diversity score (Spearman r = -0.46, P = 0.01) (Jiang et al. 2020). A similar analysis was performed to predict tumor fraction from 20 patients with HCC, including individuals in early, intermediate and

advanced stages ('1st cohort' in Supplemental Table 8). The fragmentomic-based predicted tumoral DNA fraction was well correlated to the tumor fraction measured by copy number aberration (Pearson r = 0.84, P < 0.0001, Fig 6*B*). These results also lead to an enhanced correlation of tumor fraction when compared to motif diversity score (Pearson r = 0.65, P = 0.0019) (Jiang et al. 2020). We further validated our model of fetal and tumor fraction prediction by using a second cohort of pregnancy (n = 30, Supplemental Table 7) and HCC cases (n = 20, Supplemental Table 8), with an SVR model trained with the first cohort (50% training, 50% testing). Our results supported the validity and robustness of fetal fraction prediction (Pearson r = 0.81, P < 0.0001, Supplemental Fig. S19A), and tumor fraction prediction (Pearson r = 0.75, P < 0.0001, Supplemental Fig. S19B) using fragmentomic features. The use of LASSO regression on the fetal and tumor prediction allowed for the identification of features which are most informative in the model (Supplemental Tables 9-10). The most informative features for fetal fraction prediction were the end motifs ACGT, CCGA and CCGT, while that of tumor fraction prediction was the end motif ACGA. The underlying biology causing the observed results for certain end motifs warrants further investigation. Taken together, the use of fragmentomic patterns not only offered a means to predict the absolute total concentration of cfDNA, but also allowed the deduction of fractional DNA contributions from particular cell types.

Discussion

In this study, we explored the long-standing biological puzzle of the high inter-individual variability of plasma cfDNA concentrations. We showed that the total cfDNA concentration was linked to changes in the fragmentomic profiles (i.e., the size profile and end motif distribution) of the cfDNA pool. Several lines of evidence also pointed towards the involvement of nucleases, such as DNase1L3 and DFFB, in the modulation of cfDNA concentration. The use of a machine

405 learning model (i.e., the support vector regression model) allowed for the cfDNA concentration to
406 be inferred from fragmentomic features. A similar model could also be employed to predict the
407 fetal and tumoral DNA fractions in pregnant women and patients with HCC, respectively. The
408 capability to accurately estimate the fractional contributions of specific cell types to the plasma
409 DNA pool holds significant clinical relevance, for example, possibly facilitating the advancement
410 in non-invasive prenatal testing, as well as cancer detection and monitoring.
411
412 The end motif analysis revealed that 79 end motifs (4-mers) were significantly correlated to
413 cfDNA concentration. Motifs with a negative correlation to cfDNA concentration were mostly C-
414 end and T-end motifs, while those with a positive correlation were nearly all G-ends. Similarly,
415 the deconvolution analysis of the end motifs revealed the increased contribution of Profile V in
416 subjects with higher cfDNA concentrations, which is a cleavage profile characterized by a series
417 of G-end motifs (Zhou et al. 2023). The results raise the possibility that other nucleases or
418 biological processes might have a preference for the generation of these G-end motifs. It is
419 possible that some intermediate processes, or effects of DNA fragmentation during apoptosis,
420 might have increased the frequencies of these G-end motifs, such as GCAA and GAAC. Further
421 exploration of other nucleases, such as the other members of the DNase family, endonuclease G,
422 as well as potential non-enzyme mediated fragmentation processes (Zhou et al. 2023), might shed
423 mechanistic insights into the fragmentation pattern of cfDNA and the regulation of cfDNA
424 concentration in plasma.
425
426 The importance of DNase1L3 in plasma cfDNA fragmentation has previously been demonstrated
427 in DNase1L3 knockout mouse models (Serpas et al. 2019), DNase1L3-deficient patients (Chan et
428 al. 2020), and patients with hepatocellular carcinoma (Jiang et al. 2020). Our current work

revealed the link between DNase1L3-related fragmentation signatures and cfDNA concentration. Subjects with high cfDNA concentrations exhibited fragmentomic features that resembled DNase1L3-deficient mouse model and human subjects, such as the enhanced di- and tri-nucleosomal patterns in size profile, and decreased C-end motifs (Serpas et al. 2019; Chan et al. 2020). The contribution of DNase1L3 cleavage profile (Profile I) showed a decreasing frequency with cfDNA concentration, and was most noticeable in the 231-600 bp size range. It is possible that an attenuated DNase1L3 activity might hinder the degradation of longer cfDNA molecules into shorter fragments.

The reduced activity of DNase1L3 was at least partially evidenced by the decreased protein levels of DNase1L3 in plasma from subjects with the highest cfDNA concentrations (comparing the five highest- and lowest-ranked subjects). However, this trend was diminished when expanding further to the highest and lowest 10 subjects. We speculate that only individuals with extremely elevated cfDNA concentrations (i.e., approximately 30 ng/mL or above), might be associated with a deficiency in DNase1L3 concentrations in plasma. Such individuals represented the top 0.58% (top 5 of 862 individuals) in terms of plasma cfDNA concentrations from our cohort. It is worth noting that the contribution of DFFB also shows a similar gradation pattern with cfDNA concentration, with a decreased contribution in subjects with higher cfDNA concentrations. DFFB plays a major role in DNA fragmentation during apoptosis, and our previous work has demonstrated that newly released DNA exhibited strong A-end and G-end preference which was associated with DFFB activity (Han et al. 2020b). Further work is required to assess the activity of these nucleases and how it affects the process of DNA clearance.

The characteristic fragmentomic patterns observed allowed us to train a machine learning model to directly infer the cfDNA concentration. We demonstrated that such a model which integrates the size profile and end motif distribution could have useful clinical applications, such as the prediction of fetal and tumoral DNA fractions. This approach is shown to be more strongly correlated to the fetal or tumoral DNA fractions compared to a previous method of using end motif diversity scores (Jiang et al. 2020). Such a machine learning method provides a unique advantage as it incorporates both the size and end motif differences associated with a particular physiological condition or disease. These findings open up many further possibilities for analyzing the cfDNA concentration shed from various tissues using cfDNA fragmentomics. However, there is still room to enhance the current model to enable the differentiation between different pathophysiological states. For example, both pregnancy and cancer might exhibit certain overlapping fragmentation patterns, such as the shortening of the cfDNA size distribution (Jiang and Lo 2016; Lo et al. 2021). Similiarities in fragmentomic features would make it challenging for the model to differentiate between various pathophysiological statuses. Further work may involve the analysis of larger datasets of sequencing data during the model training, encompassing a variety of physiological states, such as pregnancy, cancer, and patients with other conditions like autoimmune diseases.

Of note, it is important to acknowledge several additional considerations for the current work. As our main subject cohort (n = 862) originated from a previous clinical study for NPC screening (Chan et al. 2023), information such as body weight, time of day, and the time delay between food intake, was not reported. Detailed clinical laboratory tests were not routinely performed on paired blood collections from the main study cohort. Potential effects of physiological status and lifestyle factors have not be explored in this study. It is worth noting that the current protocol of

the screening study is implemented large-scale in clinical settings, with previous reports of high negative predictive value (above 99%) and a false-positive rate of less than 0.7% (Lam et al. 2018; Chan et al. 2022; Chan et al. 2023). We believe potential pre-analytical factors may not be a substantial factor affecting our plasma processing and extraction, given such diagnostic potential and clinical utility. We have confirmed that fragmentomic patterns that vary with cfDNA concentration were also observed in a second cohort of EBV-negative individuals (Supplemental Fig. S6). However, future studies are warranted to explore the impact of pre-analytical factors on plasma cfDNA concentration and fragmentomic profiles. It is also likely that different protocols for sample processing, methods of extraction and DNA measurement, could affect the measurement of cfDNA concentration. Further investigations into the effects of different experimental techniques would facilitate comparisons across different cohorts. Our present work evaluates fragmentomic profiles of plasma DNA using short-read sequencing from Illumina platforms, which has a readout limit of ~600 bp. It would be beneficial to further expand our fragmentomic analysis to longer cfDNA molecules (>1000 bp) in our analysis of cfDNA concentration, using long-read sequencing platforms such as single molecule, real-time sequencing (by Pacific Biosciences) and nanopore sequencing (e.g., by Oxford Nanopore Technologies). In addition, different extraction methods would enable the analysis of various sub-populations of DNA in plasma. These include the bead- based approaches to isolate ultrashort single-stranded DNA (Cheng et al. 2022; Hudecova et al. 2022), which would be of great biological interest in the context of this study.

This study elucidated several potentially biologically important underpinnings for the variability of cfDNA concentration in physiologically normal individuals. The understanding of the biology governing cfDNA concentration paves the way to personalize cfDNA presentation patterns in

liquid biopsies, and facilitates technological advances that increase the sensitivity of detection in studying diseases. This was recently demonstrated by Martin-Alonso et al. (Martin-Alonso et al. 2024), where priming agents were developed to target the clearance process of cfDNA, therefore enabling increased recovery of tumor-derived cfDNA. Furthermore, it would be of great research interest to explore the biology of cfDNA concentration and associated fragmentomic features in various physiological conditions, such as in the context of pregnancy, cancer, and organ transplantation.

Methods

Study design and subject recruitment

Samples used in this study were obtained from prospectively collected plasma from previously published studies, aiming to detect the presence of Epstein–Barr Virus (EBV) DNA in plasma for the early detection of nasopharyngeal cancer (NPC) (Chan et al. 2017; Chan et al. 2023). The study was conducted in Hong Kong during 2017-2020, recruiting participants from organized public health education sessions. The exclusion criteria for the screening were individuals with cancer, autoimmune diseases, or those with symptoms of NPC, and the use of systemic glucocorticoids or immunosuppressive therapy. Blood collection from the NPC clinical study was conducted in a controlled setting, following the established protocols that ensured subjects were in a resting state prior to the blood collection. In this study, the cfDNA concentration and sequencing data from 862 subjects with detectable EBV DNA at baseline screening were used (EBV-Positive cohort). These subjects did not develop NPC or other types of cancer identified within one year of sample collection. A selection of 497 subjects with no detectable EBV DNA (EBV-Negative cohort) was used as a second cohort to validate findings from the main study cohort. The 10 subjects with the highest cfDNA concentrations and the 10 subjects with the

lowest cfDNA concentrations from both cohorts were called back for a follow-up collection for validation, with 26 individuals returning for blood collection and analysis in this study.

Sample collection and processing

Blood collection from the screening study was done using Roche Cell-Free DNA Collection Tubes (Cat No. 07832389001). Samples were stored at 4^{o}C for no longer than six hours before processing. The blood was first centrifuged at 1600 × g for 10 minutes at 4 °C. The plasma portion was further subjected to centrifugation at 16,000 × g for 10 minutes at 4 °C to pellet out residual cells and debris. Plasma was stored in aliquots at -80^{o}C until required for experimental work. Subsequent DNA extractions for all samples were carried out after a single freeze-thaw cycle, with samples frozen during plasma processing and thawed only for DNA extraction. No samples underwent multiple freeze-thaw cycles. We excluded hemolyzed plasma from further sequencing and analysis in this study cohort.

DNA extraction and measurement of cfDNA concentration

Plasma cfDNA was extracted as previously described (Chan et al. 2022; Chan et al. 2023). Briefly, 2 mL of plasma of each sample was extracted using the MagMAX Cell-Free DNA Isolation Kit in the KingFisher Flex System (ThermoFisher) according to the manufacturer's instructions. The cfDNA concentration of each extracted sample was measured using a Qubit 4 fluorometer instrument (Invitrogen).

Measurement of cfDNA concentration by digital droplet PCR (ddPCR) assay

The ddPCR assay was performed as previously described (Gai et al. 2023), for a selection of 25 samples from EBV-negative cohort as validation. We have adopted this method to target the

valosin-containing protein (VCP), using one set of primer and probe. Briefly, ddPCR reactions were performed by the QX ONE Droplet Digital PCR System (Bio-Rad). The DNA samples used were eluted with the same volume during DNA extraction. The reaction was prepared in 20 µL, with equal volumes of DNA added per reaction. Components added to the reaction include 2× ddPCR Supermix for probes (Bio-Rad), final concentration of 900 µmol/L of each primer and 250 nmol/L of the DNA probe. The ddPCR thermal profile consisted of an initial incubation at 37°C for 30 minutes, followed by the denaturation step at 95°C for 10 minutes. It was further followed by 45 cycles of amplification, with each cycle including a denaturation at 94°C for 30 seconds and annealing at 57°C for 1 minute. A final incubation at 98°C for 10 minutes was carried out. Data and calculations were Poisson corrected. To measure the cfDNA concentration, the number of DNA templates for the VCP gene was quantified from cfDNA samples. The primer sequence used for this assay was:

VCP Forward primer: 5’ GGGAGGTCTGTGGACCCTATC 3’

VCP Reverse primer: 5’ GGGAGGTCTGTGGACCCTATC 3’

Probe sequence: 5’ FAM CTTCCCCAACCATCAG MGB 3’

Target-capture enrichment and analysis

Routine clinical screening for Epstein–Barr virus (EBV) DNA for nasopharyngeal cancer (NPC) detection involves target capture enrichment of viral DNA molecules from the total pool of plasma DNA, as it improves the positive predictive value of NPC detection (Lam et al. 2018). Enrichment in the screening study was performed using a hybridization-based capture with probes that cover the entire EBV genome and selected human autosomal DNA regions. Data from the target capture of all subjects (862 subjects) were retrieved for analysis. Only off-target DNA fragments were analyzed in this study, to avoid the effects of target capture on fragmentomic

572 analysis. Off-target reads were defined as DNA fragments with no full or partial alignment to the
573 human autosomal target sites. The robustness of using 'off-target' reads for fragmentomic
574 analysis has been validated in comparison to the paired genome-wide sequencing data
575 (Supplemental Fig. S8 and S9).
576
577 **Genome-wide sequencing DNA library preparation**
578 For the selected subjects with the highest and lowest concentrations of cfDNA in the EBV-
579 positive and EBV-negative cohort, genome-wide sequencing was performed without target
580 capture. Extracted DNA from 2 mL of plasma was performed using TruSeq DNA Nano Library
581 Prep Kit (Illumina), with purification steps done using MinElute Reaction Cleanup (Qiagen)
582 according to the manufacturer's instructions. Adaptor-ligated DNA was amplified using eight
583 cycles of PCR. Quality control of the prepared libraries was done by Qubit and Agilent 4200
584 TapeStation (Agilent).
585
586 **Sequencing and alignment**
587 Target-captured DNA libraries were sequenced on the NextSeq 500 System (Illumina) with 75
588 bp × 2 (150 cycles) paired-end sequencing. Sequenced DNA was aligned to the human genome
589 (hg19). Fragmentomic analysis was performed only using the paired-end reads uniquely aligned
590 to the human autosomal chromosomes, which was consistent with our previous work (Jiang et al.
591 2020). Re-analysis of the existing data using GRCh38 (UCSC hg38) human reference genome
592 would not significantly alter the results, as the major difference between the two versions of the
593 human reference genomes is the sequence representation for highly repetitive regions and
594 centromeres. Short sequencing reads obtained from those regions would have multiple
595 alignments, and would therefore not be utilized in our downstream analysis.

Analysis of 5'-end motifs

Our analysis of end motif profiles was performed as previously described (Serpas et al. 2019; Chan et al. 2020). We analyzed the first 4-nucleotide sequence (termed the 4-mer end motif) at each 5' fragment end of plasma DNA molecules. The frequency of each of the 256 possible 4-mer end motifs (4^4 combinations) was calculated and normalized by the total number of ends. In our analysis, we correlated the frequency of each end motif to the plasma cfDNA concentration (n = 862 subjects). The p-value of each correlation was adjusted using Bonferroni's correction for multiple comparisons (Supplemental Tables 2-3). Our analysis is limited to the 5' end motif, as the identity of the original 3' end motifs will be modified during the end repair step, as reported in our previous work (Jiang et al. 2020).

Analysis of the contribution of "Founder" end-motif profiles (F-profile)

The data matrix and deduced percentage contributions of the six F-profiles were performed as previously described (Zhou et al. 2023). Briefly, the percentage contribution of each F-profile in a cfDNA sample was deduced using non-negative least squares-based deconvolution analysis of the previously established data matrix. We performed the F-profile analysis using DNA fragments stratified into different size ranges: 20 –160 bp, 161 – 230 bp, and 231 – 600 bp. Multiple linear regression was performed on all the significantly correlated F-profiles in each size range.

Deduction of the total cfDNA concentration or fractional cfDNA concentration using fragmentomic features

The SVR model was built by the 'e1071' package in R 4.1.2. The 862 subjects from the target sequencing dataset were randomly grouped into training set (n = 431 subjects) and testing set (n = 431 subjects) without overlap. The training dataset was used to build a support vector regression (SVR) model to predict the cfDNA concentration based on fragmentomic features. Frequencies of 4-mer motifs (256 motifs) and frequencies of cfDNA at different sizes (20~600 bp; 581 sizes) were used as features to build the model. The prediction target was the cfDNA concentration after logarithmic transformation. The default radial kernel was used in the SVR model, and the gamma and cost values were tuned using the function 'tune.svm'. After the model was built, the testing dataset was used to evaluate the predictive accuracy, with 20 subjects from EBV-negative cohort were used as a second testing dataset.

The SVR models were built using the leave-one-out strategy to predict fetal fraction in pregnant subjects (n = 30) and tumor fraction in patients with HCC (n = 20). The same fragmentomic profiles (i.e., end motif and size) were used as training features; with the fetal fractions and tumor fractions as target values for building the SVR model. The samples were from previously published sequencing data (Jiang et al. 2020). Raw sequencing data for HCC and pregnancy cases were obtained from a previous study (Jiang et al. 2020), with EGA accession number EGAS00001003409. The predicted fetal and tumor fraction based on fragmentomic features were correlated to the fractions quantified by FetalQuant and ichorCNA, respectively. To demonstrate the robustness of the SVR models, we included another 20 HCC patients and 30 pregnant subjects as

validation (Supplemental Fig. S19). To identify the most informative fragmentomic features in predicting fetal fraction and tumor fraction, we performed LASSO regressions using the first dataset of 30 pregnant women and 20 patients with HCC (Supplemental Tables 9-10). The LASSO regression coefficients were used to indicate the importance of each feature.

Quantitative analysis of DNase1L3 levels in plasma

DNase1L3 levels in plasma between selected samples were evaluated using the Jess automated Western Blotting system (ProteinSimple). 50 µL aliquots of plasma were diluted 200-fold using 1 × phosphate-buffered saline (PBS) and mixed with 2.5x Fluorescent Master Mix containing DTT. The samples were boiled to 95^{o}C for 5 min. The samples were loaded onto the 12 x 230 kDa Separation Module containing 25 capillary cartridges with the following settings – Separation voltage: 375 volts, 25 min. 'Primary Antibody Time' and 'Secondary Antibody Time' as 30 min. 'Detection Profile – Chemi'. DNase1L3 immuno-detected using polyclonal anti-DNASE1L3 antibodies (Thermo Fisher PA5-107113) and Secondary Anti-Rabbit antibodies (Anti-Rabbit Detection Module, DM001 Protein Simple). Chemiluminescent signals were quantified by band intensity area using Compass for SW software. Samples were run in two replicates, with the correlation between the two replicates shown in Supplemental Fig. S14.

Tissue-of-origin analysis by Fragmentomics-based Methylation Analysis (FRAGMA)

FRAGMA-based tissue deconvolution analysis is based on the deduction of the methylation status of cytosine-phosphate-guanine (CpG) sites, by the preferential cleavage of methylation sites compared to unmethylated sites, using the CGN/NCG motif ratios as previously described

(Zhou et al. 2022). Hypermethylated and Hypomethylated CpG sites are defined as CpG sites with a methylation index of above 70% and below 30%, respectively. Unique hypermethylated CpG sites for each cell type (Liver, neutrophils, B cells, T cells, erythroblasts, and megakaryocytes) were identified by bisulfite sequencing tissue references for each cell type. The percentage contribution of each tissue was deduced by the difference in CGN/NCG motif ratio between hyper- and hypomethylated CpG sites, normalized by the CGN/NCG motif ratio from the reference tissue, using the following equation:

$$\textbf{Normalized CGN/NCG Motif Ratio} = \frac{Ratio\ M - Ratio\ U}{Reference\ M - Reference\ U}$$

Where 'Ratio' denotes the raw CGN/NCG motif ratio, 'M' and 'U' represent methylation and unmethylated sites, respectively. This equation has been described in a patent application (US2023/0374601 – 'Fragmentation for measuring methylation and disease') prior to this work.

Data collection of clinical lab test parameters

Whole blood, serum, and plasma samples from individuals of the follow-up blood collection were subjected to a detailed clinical laboratory analysis in the Prince of Wales Hospital, from the Immunology, Hematology and Chemical Pathology Laboratory, run by the Hospital Authority (Hong Kong). Data from these parameters outlined in Supplemental Table 5 were provided and used in this study.

Statistical analysis

Sequencing analyses were performed by bioinformatic programs written in Perl and R languages. The *P*-values for all comparisons are stated in the figures and results section. The Pearson correlation coefficient, *r*, was used to assess correlations. Statistical comparisons between two groups were performed using the Wilcoxon rank sum test with two-tailed comparison. Statistical tests and plots were performed using R (R Core Team 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/) and GraphPad Prism 9. A value of $P < 0.05$ was considered statistically significant.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Genome-Phenome Archive (EGA), under accession number EGAC00001000078.

Competing interest statement

K.C.A.C. and Y.M.D.L. hold equities in DRA, Take2 and Insighta. P.J. and W.K.J.L. hold equities in Illumina. P.J. is a consultant to KingMed Future. W.P. is a consultant to Take2. K.C.A.C. P.J. is a director of DRA, Take2, KingMed Future, and Insighta. W.K.J.L. is a director of DRA. Y.M., G.K., W.K.J.L, P.J., K.C.A.C., and Y.M.D.L. have filed a patent application based on the data in this study. Patent incomes are received from Grail, Illumina, DRA, Take2, Insighta and Xcelom.

Acknowledgments: This work was supported by the Innovation and Technology Commission of the Hong Kong SAR Government (InnoHK initiative) and the Li Ka Shing Foundation. We would like to thank the Biomedical Technology Support Centre of the Hong Kong Science and Technology Parks for instrument use and technical support of the Jess automated Western

Blotting (Protein Simple) device. We thank Chris Kum and Xingqian Li for their technical assistance.

Author Contributions: Y.M., G.K., W.K.J.L., Q.Z., P.J., K.C.A.C., Y.M.D.L. designed research; Y.M., G.K., W.K.J.L., Q.Z., S.H.C, P.P.H.C., J.B., M.L.C., C.T.L., Y. Z., W.G., W.S.W.W., M.-J.L.M, X.X., Z.G., H.S. performed research; W.K.J.L., I.O.L.T, L.Y.L.C., K.C.A.C. took part in clinical recruitment; Y.M., G.K., W.K.J.L., Q.Z., J.B., W.P, W.S.W.W., W.L., P.J., K.C.A.C., Y.M.D.L. analyzed data and performed statistical analysis. Y.M., G.K., W.K.J.L., P.J., K.C.A.C., Y.M.D.L. wrote the paper.

Figure legends

Figure 1. Overview of the study design. The plasma cfDNA concentration of all 862 individuals from the study cohort was measured. The fragmentomic features of the cfDNA were analyzed, including the size profile and end motif distributions. The contribution of nucleases was explored using the cleavage profile end motif patterns, and protein quantification of DNase1L3 in plasma. Machine learning could be trained using fragmentomic features to predict cfDNA concentration. We further explored whether a similar model could be applied to study fractional DNA concentration in predicting the fetal and tumor fraction of plasma cfDNA in pregnant subjects and patients with HCC, respectively.

Figure 2. Frequency distribution histogram plot showing the distribution of plasma cfDNA concentration from 862 individuals of the study cohort.

Figure 3. Size profile of plasma DNA fragments in subjects of different cfDNA concentrations. The size profiles of the selected lowest and highest 10% of subjects, and the median distribution of the cohort, are shown in (A) linear scale and (B) logarithmic scale. Correlation between the frequency of DNA fragments within each 10 bp window bin and cfDNA concentration was assessed for all subjects, for the (C) 81-90 bp fragment size range, (D) 301-310 bp fragment size range, and (E) Across all 10-bp bins from 20-600 bp. Blue and red labels indicate 10-bp bins with a statistically significant negative or positive correlation to cfDNA concentration, respectively.

Figure 4. End motifs with significant correlation to cfDNA concentration. The frequencies of all 256 end motifs (4-mer) were correlated to the cfDNA concentration. The resulting analysis revealed 34 negatively correlated end motifs to cfDNA concentration, and 45 positively correlated end motifs. (A) Heatmap analysis with rows indicating a particular 4-mer motif, in which the first base of the motif highlighted by a specific color in the left-most column (A, C, G and T are colored by green, red, yellow, and blue, respectively). Each column indicates a plasma DNA sample from one subject. The frequency *z*-score, calculated for each end motif, is shown by the color scale. A support vector regression (SVR) model was trained with fragmentomic features to predict cfDNA concentration, using (B) end motif frequencies and (C) both end motif and size profile.

Figure 5. Deconvolutional analysis of end motifs to deduce the contribution of the six "founder" end motif profiles (F-profile) of plasma DNA fragments within the 231 – 600 bp size range, among subjects of different cfDNA concentrations. Heatmap analysis was performed, with each row showing the contributions (expressed as *z*-scores) of each F-profile across subjects with different cfDNA concentrations.

Figure 6. Application of fragmentomic patterns based deduction of the fractional DNA concentration from specific cell types. Fragmentomic features, including size profile and end motif distribution, were used to train SVR models in the prediction of the fractional DNA percentage, which was correlated to the proportional tissue DNA contribution. (A) Correlation between the fetal DNA fraction predicted by the fragmentomic features and SNP-based methods, using 30 pregnant subjects. (B) Correlation between the tumoral fraction predicted by fragmentomic features and copy number aberration (ichorCNA).

References

Alborelli I, Generali D, Jermann P, Cappelletti MR, Ferrero G, Scaggiante B, Bortul M, Zanconati F, Nicolet S, Haegele J et al. 2019. Cell-free DNA analysis in healthy individuals by next-generation sequencing: a proof of concept and technical validation study. *Cell Death Dis* **10**: 534.

Chan DCT, Lam WKJ, Hui EP, Ma BBY, Chan CML, Lee VCT, Cheng SH, Gai W, Jiang P, Wong KCW et al. 2022. Improved risk stratification of nasopharyngeal cancer by targeted sequencing of Epstein-Barr virus DNA in post-treatment plasma. *Ann Oncol* **33**: 794-803.

Chan KCA, Lam WKJ, King A, Lin SV, Lee PHP, B. ZCY, Chan LS, Tse IOL, Tsang FCA, Li ZJM et al. 2023. Plasma Epstein–Barr Virus DNA and Risk of Future Nasopharyngeal Cancer. *NEJM Evid 2023* **2**.

780 Chan KCA, Woo JKS, King A, Zee BCY, Lam WKJ, Chan SL, Chu SWI, Mak C, Tse IOL,
781 Leung SYM et al. 2017. Analysis of Plasma Epstein-Barr Virus DNA to Screen for
782 Nasopharyngeal Cancer. *N Engl J Med* **377**: 513-522.
783 Chan RWY, Serpas L, Ni M, Volpi S, Hiraki LT, Tam LS, Rashidfarrokhi A, Wong PCH, Tam
784 LHP, Wang Y et al. 2020. Plasma DNA Profile Associated with DNASE1L3 Gene
785 Mutations: Clinical Observations, Relationships to Nuclease Substrate Preference, and In
786 Vivo Correction. *Am J Hum Genet* **107**: 882-894.
787 Chen M, Chan RWY, Cheung PPH, Ni M, Wong DKL, Zhou Z, Ma ML, Huang L, Xu X, Lee
788 WS et al. 2022. Fragmentomics of urinary cell-free DNA in nuclease knockout mouse
789 models. *PLoS Genet* **18**: e1010262.
790 Cheng AP, Cheng MP, Gu W, Sesing Lenz J, Hsu E, Schurr E, Bourque G, Bourgey M, Ritz J,
791 Marty FM et al. 2021. Cell-free DNA tissues of origin by methylation profiling reveals
792 significant cell, tissue, and organ-specific injury related to COVID-19 severity. *Med* **2**:
793 411-422 e415.
794 Cheng J, Morselli M, Huang WL, Heo YJ, Pinheiro-Ferreira T, Li F, Wei F, Chia D, Kim Y, He
795 HJ et al. 2022. Plasma contains ultrashort single-stranded DNA in addition to
796 nucleosomal cell-free DNA. *iScience* **25**: 104554.
797 Davidson BA, Miranda AX, Reed SC, Bergman RE, Kemp JDJ, Reddy AP, Pantone MV, Fox
798 EK, Dorand RD, Hurley PJ et al. 2024. An in vitro CRISPR screen of cell-free DNA
799 identifies apoptosis as the primary mediator of cell-free DNA release. *Commun Biol* **7**:
800 441.
801 Fridlich O, Peretz A, Fox-Fisher I, Pyanzin S, Dadon Z, Shcolnik E, Sadeh R, Fialkoff G, Sharkia
802 I, Moss J et al. 2023. Elevated cfDNA after exercise is derived primarily from mature

polymorphonuclear neutrophils, with a minor contribution of cardiomyocytes. *Cell Rep Med* **4**.

Gai W, Yu SCY, Chan WTC, Peng W, Lau SL, Leung TY, Jiang P, Chan KCA, Lo YMD. 2023. Droplet digital PCR is a cost-effective method for analyzing long cell-free DNA in maternal plasma: Application in preeclampsia. *Prenat Diagn* **43**: 1385-1393.

Grabuschnig S, Bronkhorst AJ, Holdenrieder S, Rosales Rodriguez I, Schliep KP, Schwendenwein D, Ungerer V, Sensen CW. 2020. Putative Origins of Cell-Free DNA in Humans: A Review of Active and Passive Nucleic Acid Release Mechanisms. *Int J Mol Sci* **21**.

Han D, Li R, Shi J, Tan P, Zhang R, Li J. 2020a. Liquid biopsy for infectious diseases: a focus on microbial cell-free DNA sequencing. *Theranostics* **10**: 5501-5513.

Han DSC, Lo YMD. 2021. The Nexus of cfDNA and Nuclease Biology. *Trends Genet* **37**: 758-770.

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

Heitzer E, Auinger L, Speicher MR. 2020. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends Mol Med* **26**: 519-528.

Hudecova I, Smith CG, Hansel-Hertsch R, Chilamakuri CS, Morris JA, Vijayaraghavan A, Heider K, Chandrananda D, Cooper WN, Gale D et al. 2022. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. *Genome Res* **32**: 215-227.

825 Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, Wong GL, Chan SL, Mok TS,
826 Chan HL et al. 2015. Lengthening and shortening of plasma DNA in hepatocellular
827 carcinoma patients. *Proc Natl Acad Sci U S A* **112**: E1317-1325.
828 Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, Heung MMS, Xie T, Shang H, Zhou Z et
829 al. 2020. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer,
830 Pregnancy, and Transplantation. *Cancer Discov* **10**: 664-673.
831 Jiang PY, Lo YMD. 2016. The Long and Short of Circulating Cell-Free DNA and the Ins and
832 Outs of Molecular Diagnostics. *Trends in Genetics* **32**: 360-371.
833 Lam WKJ, Jiang P, Chan KCA, Cheng SH, Zhang H, Peng W, Tse OYO, Tong YK, Gai W, Zee
834 BCY et al. 2018. Sequencing-based counting and size profiling of plasma Epstein-Barr
835 virus DNA enhance population screening of nasopharyngeal carcinoma. *Proc Natl Acad*
836 *Sci U S A* **115**: E5115-E5124.
837 Lo YMD, Zhang J, Leung TN, Lau TK, Chang AM, Hjelm NM. 1999. Rapid clearance of fetal
838 DNA from maternal plasma. *Am J Hum Genet* **64**: 218-224.
839 Lo YMD, Han DSC, Jiang PY, Chiu RWK. 2021. Epigenetics, fragmentomics, and topology of
840 cell-free DNA in liquid biopsies. *Science* **372**.
841 Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S,
842 Hecht M, Pelet T et al. 2023. A DNA methylation atlas of normal human cell types.
843 *Nature* **613**: 355-364.
844 Martin-Alonso C, Tabrizi S, Xiong K, Blewett T, Sridhar S, Crnjac A, Patel S, An Z, Bekdemir A,
845 Shea D et al. 2024. Priming agents transiently reduce the clearance of cell-free DNA to
846 improve liquid biopsies. *Science* **383**: eadf2341.
847 Mattox AK, Douville C, Wang Y, Popoli M, Ptak J, Silliman N, Dobbyn L, Schaefer J, Lu S,
848 Pearlman AH et al. 2023. The Origin of Highly Elevated Cell-Free DNA in Healthy

Individuals and Patients with Pancreatic, Colorectal, Lung, or Ovarian Cancer. *Cancer Discov* **13**: 2166-2179.

Meddeb R, Dache ZAA, Thezenas S, Otandault A, Tanos R, Pastor B, Sanchez C, Azzi J, Tousch G, Azan S et al. 2019. Quantifying circulating cell-free DNA in humans. *Sci Rep* **9**: 5220.

Orntoft MW, Jensen SO, Ogaard N, Henriksen TV, Ferm L, Christensen IJ, Reinert T, Larsen OH, Nielsen HJ, Andersen CL. 2021. Age-stratified reference intervals unlock the clinical potential of circulating cell-free DNA as a biomarker of poor outcome for healthy individuals and patients with colorectal cancer. *Int J Cancer* **148**: 1665-1675.

Pastor WA, Kwon SY. 2022. Distinctive aspects of the placental epigenome and theories as to how they arise. *Cell Mol Life Sci* **79**: 569.

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

Shiokawa D, Tanuma S. 1998. Molecular cloning and expression of a cDNA encoding an apoptotic endonuclease DNase gamma. *Biochem J* **332 (Pt 3)**: 713-720.

Sisirak V, Sally B, D'Agati V, Martinez-Ortiz W, Ozcakar ZB, David J, Rashidfarrokhi A, Yeste A, Panea C, Chida AS et al. 2016. Digestion of Chromatin in Apoptotic Cell Microparticles Prevents Autoimmunity. *Cell* **166**: 88-101.

Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, Chan KCA, Leung TY, Chiu RWK, Lo YMD. 2018. Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *Proc Natl Acad Sci U S A* **115**: E5106-E5114.

Tug S, Helmig S, Menke J, Zahn D, Kubiak T, Schwarting A, Simon P. 2014. Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cell Immunol* **292**: 32-39.

Ungerer V, Bronkhorst AJ, Uhlig C, Holdenrieder S. 2022. Cell-Free DNA Fragmentation Patterns in a Cancer Cell Line. *Diagnostics (Basel)* **12**.

Yu SC, Chan KC, Zheng YW, Jiang P, Liao GJ, Sun H, Akolekar R, Leung TY, Go AT, van Vugt JM et al. 2014. Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc Natl Acad Sci U S A* **111**: 8583-8588.

Yu SC, Jiang P, Chan KC, Faas BH, Choy KW, Leung WC, Leung TY, Lo YMD, Chiu RW. 2017. Combined Count- and Size-Based Analysis of Maternal Plasma DNA for Noninvasive Prenatal Detection of Fetal Subchromosomal Aberrations Facilitates Elucidation of the Fetal and/or Maternal Origin of the Aberrations. *Clin Chem* **63**: 495-502.

Yu SC, Lee SW, Jiang P, Leung TY, Chan KC, Chiu RW, Lo YMD. 2013. High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. *Clin Chem* **59**: 1228-1237.

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

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

Zhu D, Wang H, Wu W, Geng S, Zhong G, Li Y, Guo H, Long G, Ren Q, Luan Y et al. 2023. Circulating cell-free DNA fragmentation is a stepwise and conserved process linked to apoptosis. *BMC Biol* **21**: 253.

Figure 1

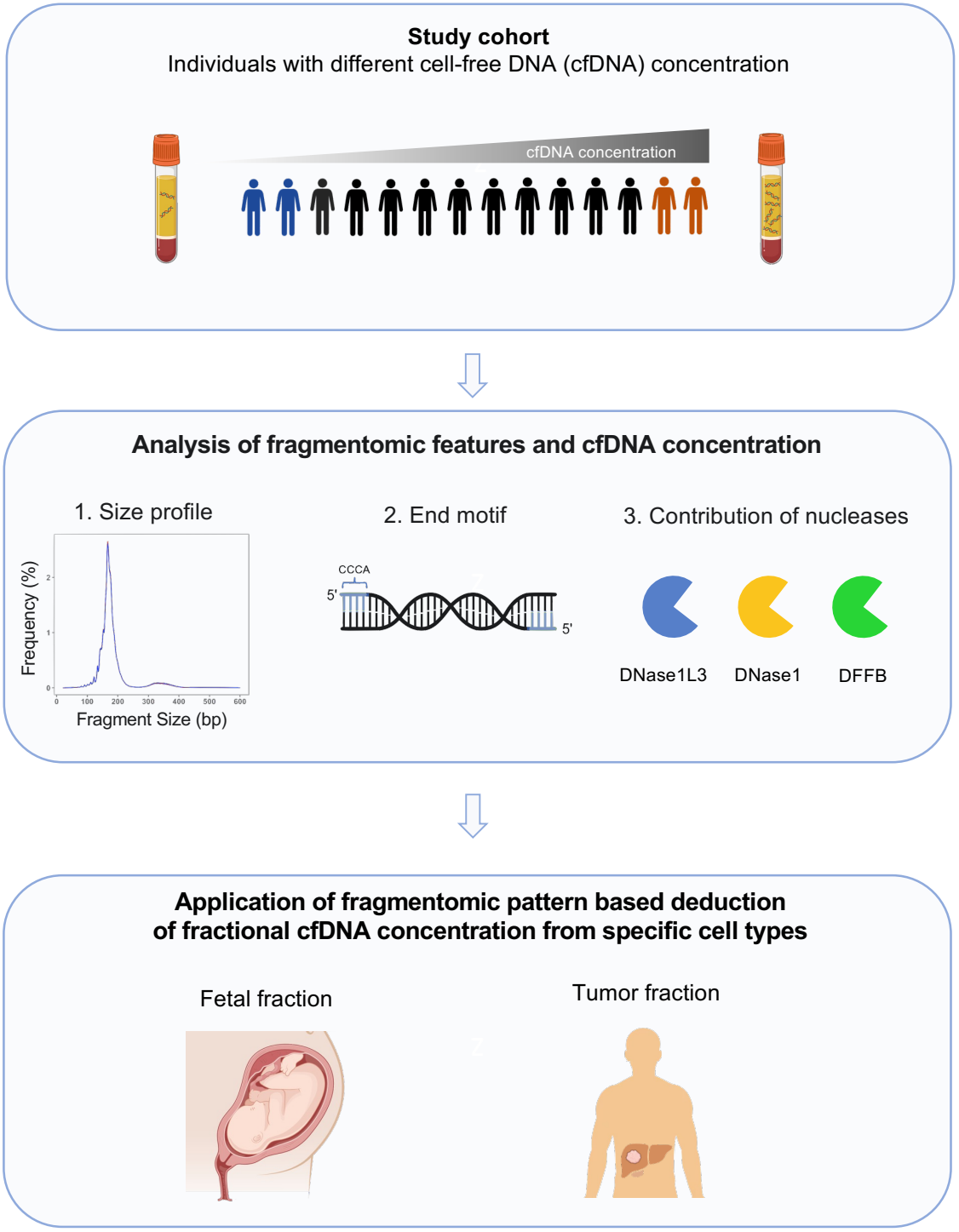


Figure 2

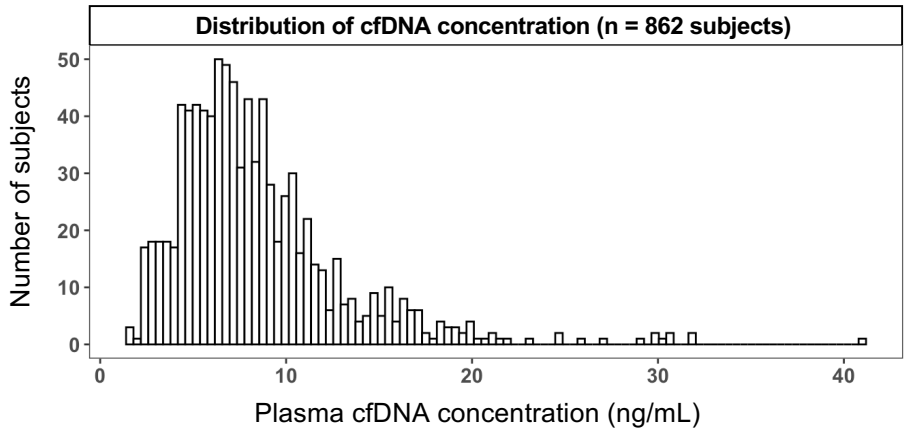


Figure 3

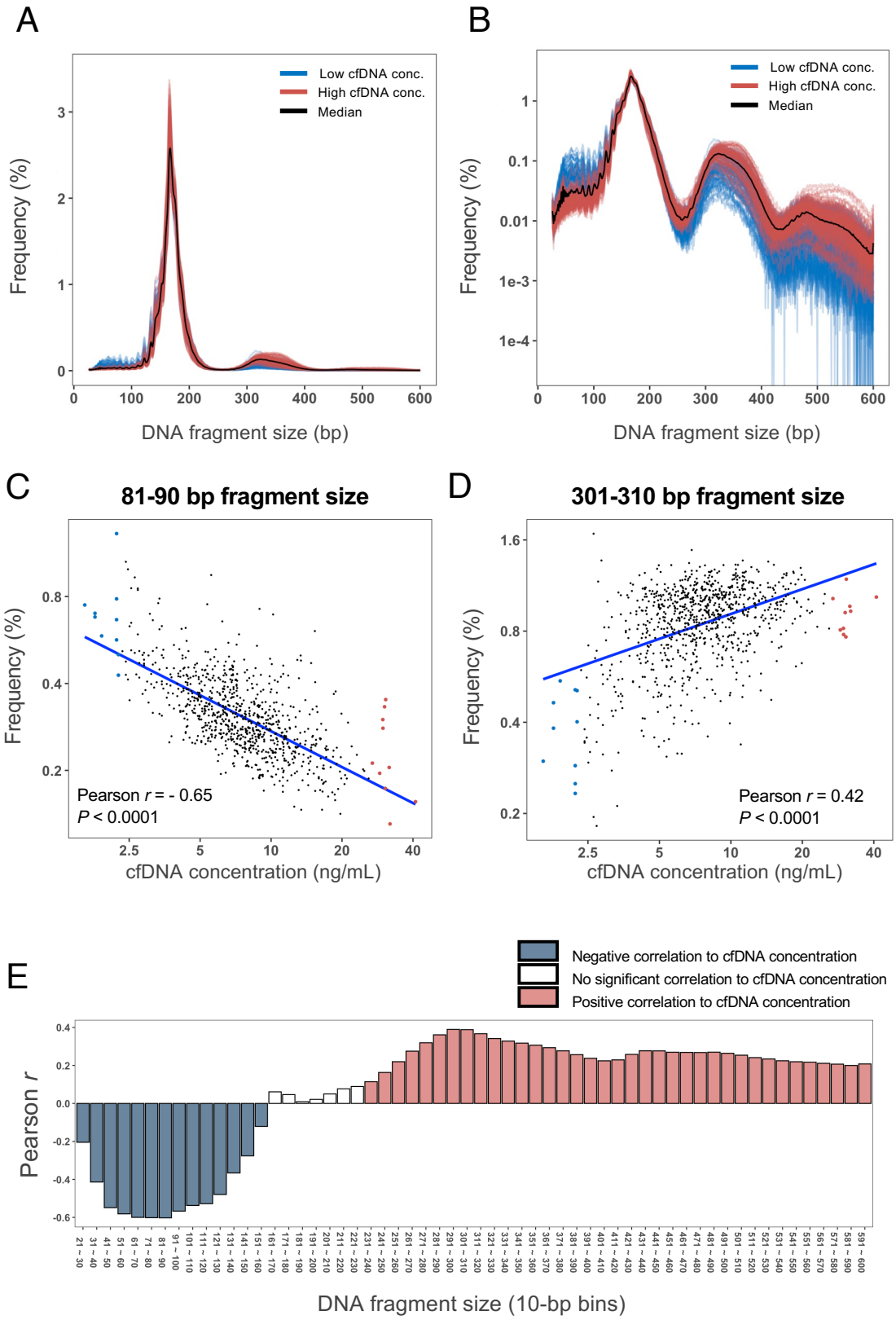


Figure 4

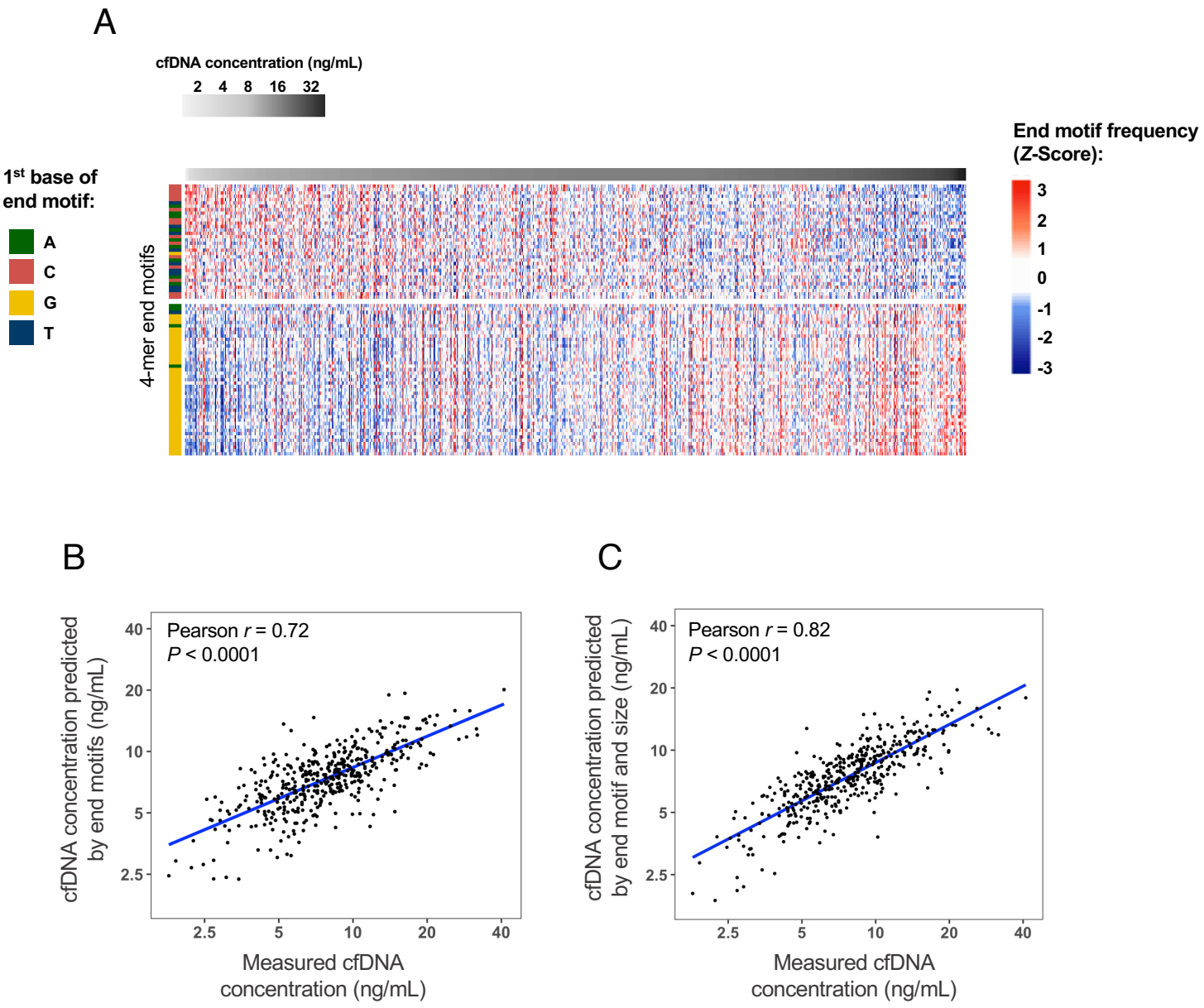


Figure 5

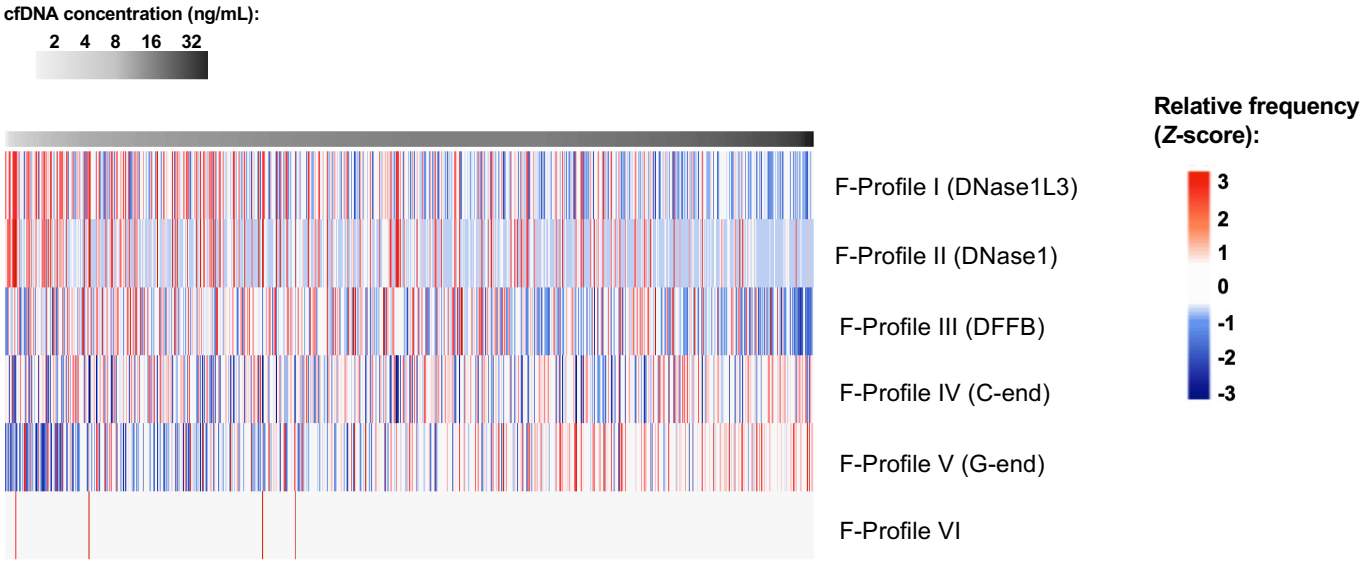


Figure 6