










# Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs

Ze Zhou<sup>a,b,c,1</sup>, Mary-Jane L. Ma<sup>a,b,c,1</sup>, Rebecca W. Y. Chan<sup>a,b,c</sup>, W. K. Jacky Lam<sup>a,b,c,d</sup>, Wenlei Peng<sup>a,b,c</sup>, Wanxia Gai<sup>a,b,c</sup>, Xi Hu<sup>a,b,c</sup>, Spencer C. Ding<sup>a,b,c</sup>, Lu Ji<sup>a,b,c</sup> , Qing Zhou<sup>a,b,c</sup> , Peter P. H. Cheung<sup>b,c</sup>, Stephanie C. Y. Yu<sup>a,b,c</sup>, Jeremy Y. C. Teoh<sup>e</sup>, Cheuk-Chun Szeto<sup>f</sup>, John Wong<sup>g</sup> , Vincent W. S. Wong<sup>f</sup>, Grace L. H. Wong<sup>g</sup>, Stephen L. Chan<sup>d,h</sup> , Edwin P. Hui<sup>d,h</sup>, Brigette B. Y. Ma<sup>d,h</sup>, Anthony T. C. Chan<sup>d,h</sup>, Rossa W. K. Chiu<sup>a,b,c</sup> , K. C. Allen Chan<sup>a,b,c,d</sup>, Y. M. Dennis Lo<sup>a,b,c,d,2</sup> , and Peiyong Jiang<sup>a,b,c,d,2</sup> 

Contributed by Y.M. Dennis Lo; received December 20, 2022; accepted March 20, 2023; reviewed by Muhammed Murtaza and Xianghong J. Zhou

Cell-free DNA (cfDNA) fragmentation is nonrandom, at least partially mediated by various DNA nucleases, forming characteristic cfDNA end motifs. However, there is a paucity of tools for deciphering the relative contributions of cfDNA cleavage patterns related to underlying fragmentation factors. In this study, through non-negative matrix factorization algorithm, we used 256 5' 4-mer end motifs to identify distinct types of cfDNA cleavage patterns, referred to as “founder” end-motif profiles (F-profiles). F-profiles were associated with different DNA nucleases based on whether such patterns were disrupted in nuclease-knockout mouse models. Contributions of individual F-profiles in a cfDNA sample could be determined by deconvolutional analysis. We analyzed 93 murine cfDNA samples of different nuclease-deficient mice and identified six types of F-profiles. F-profiles I, II, and III were linked to deoxyribonuclease 1 like 3 (DNASE1L3), deoxyribonuclease 1 (DNASE1), and DNA fragmentation factor subunit beta (DFFB), respectively. We revealed that 42.9% of plasma cfDNA molecules were attributed to DNASE1L3-mediated fragmentation, whereas 43.4% of urinary cfDNA molecules involved DNASE1-mediated fragmentation. We further demonstrated that the relative contributions of F-profiles were useful to inform pathological states, such as autoimmune disorders and cancer. Among the six F-profiles, the use of F-profile I could inform the human patients with systemic lupus erythematosus. F-profile VI could be used to detect individuals with hepatocellular carcinoma, with an area under the receiver operating characteristic curve of 0.97. F-profile VI was more prominent in patients with nasopharyngeal carcinoma undergoing chemoradiotherapy. We proposed that this profile might be related to oxidative stress.

non-negative matrix factorization | oxidative stress | fragmentomics | cancer detection | liquid biopsy

Cell-free DNA (cfDNA) is a mixture of DNA fragments released from different tissues (1, 2). cfDNA fragmentation is nonrandom (3, 4), at least in part mediated by various DNA nucleases, such as deoxyribonuclease 1 (DNASE1), deoxyribonuclease 1 like 3 (DNASE1L3), and DNA fragmentation factor subunit beta (DFFB) (5, 6). Han et al. revealed that the generation of cfDNA molecules might intracellularly and extracellularly involve a series of nuclease-directed fragmentation processes in a stepwise manner (7). Such a stepwise fragmentation model suggests that cfDNA might be initially cleaved intracellularly by DFFB and DNASE1L3, preferentially forming A-end and C-end fragments, respectively, followed by the extracellular cleavages mediated by DNASE1L3 and T-end preferred DNASE1 (7). Thus, the compositions of nucleotides at the end of cfDNA molecules (i.e., k-mer end motifs; k indicates the length of nucleotides) are believed to be associated with the DNA nuclease activities. Based on murine models with the knockout of different nuclease genes, Serpas et al. reported that DNASE1L3 might be associated with the generation of “CCCA” end motif (8), which was subsequently confirmed in human data (9). On the other hand, DNASE1 might be associated with the generation of the motif “TGTG” (10). These studies suggest that DNA nuclease activities are involved in cfDNA fragmentation.

Recently, many studies demonstrated that the use of plasma end motifs was able to inform the presence of various diseases ranging from autoimmune diseases (9) to multiple cancer types (11–13). *Dnase1l3*-deficient mice rapidly developed autoantibodies to DNA and chromatin, followed by the development of systemic lupus erythematosus (SLE) like diseases (14). DNASE1L3 deficiency could be restored by adeno-associated virus (AAV) based transduction of *Dnase1l3* into *Dnase1l3*-deficient mice (14). Interestingly, the AAV-based transduction of *Dnase1l3* could restore the aberrant end-motif profiles to those normally present in the plasma cfDNA of wild-type (WT) mice, suggesting that

## Significance

Cell-free DNA (cfDNA) fragmentation patterns carry a wealth of information related to DNA nuclease activities and tissues of origin. However, there is a lack of tools for obtaining a bird's eye view of the involvement of nucleases and other fragmentation mechanisms in the process of cfDNA generation. Using mathematical analysis of short terminal nucleotide sequences of cfDNA, called end motifs, six distinct types of cfDNA fragmentation patterns were observed. Several patterns were associated with nucleases such as DFFB, DNASE1, and DNASE1L3. These patterns shed light on the spectrum of processes in cfDNA fragmentation. Aberrations in these patterns could be used as markers for cancer and immune diseases. One aberrant pattern appeared to be associated with increased oxidative stress during cancer.

Reviewers: M.M., University of Wisconsin-Madison; and X.J.Z., University of California Los Angeles.

Competing interest statement: The authors have organizational affiliations to disclose. K.C.A.C. and Y.M.D.L. hold leadership positions in Centre for Novostics. R.W.K.C., K.C.A.C., and Y.M.D.L. hold equities in Take2. Z.Z., M.-J.L.M., K.C.A.C., Y.M.D.L., and P.J. have filed a patent application on the described technology.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

<sup>1</sup>Z.Z. and M.-J.L.M. contributed equally to this work.

<sup>2</sup>To whom correspondence should be addressed. Email: loym@cuhk.edu.hk or jiangpeiyong@cuhk.edu.hk.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2220982120/-/DCSupplemental>.

Published April 19, 2023.

end motifs might be a biomarker for monitoring therapeutic response following restoration of nuclease activities (9).

We reasoned that it would be clinically meaningful to holistically determine the distinct types of cfDNA cleavage patterns. Such an approach might allow us to identify previously unknown mechanisms for cfDNA fragmentation. To this end, we attempted to use 256 4-mer end motifs of cfDNA molecules to deconvolute distinct types of cfDNA cleavages using non-negative matrix factorization (NMF) (Fig. 1). The 4-mer end-motif profiles, which were obtained from 93 murine cfDNA samples from WT mice and knockout mice with the deletion of different DNA nuclease genes, were subjected to NMF analysis. Such factorization analysis yielded six major types of end-motif components representing the distinct types of cfDNA cleavages, referred to as “founder” end-motif profiles (F-profiles). An observed motif profile could be deconvoluted into different F-profiles by iteratively adjusting the proportional contribution of each F-profile. With the use of the F-profiles generated from cfDNA of mice, the proportional contributions of F-profiles could be determined for human cfDNA samples. Such an approach could potentially be used for developing biomarkers for assessing physiological and pathological conditions.

## Results

**cfDNA End-Motif Landscape Profiles.** To determine the distinct types of cfDNA cleavages, we first calculated the frequencies for each 4-mer end motif in cfDNA samples. The 4-mer end motif was defined as the terminal 4 nucleotides at each 5′ fragment end of cfDNA molecules, totaling 256 categories of 4-mer end motifs (i.e., 4<sup>4</sup>). To make the profiles of motif patterns comparable between human and mouse, the frequencies of 4-mer end motifs related to the human and murine cfDNA were normalized by the genomic contexts of the human and mouse genomes, respectively (see details in *Materials and Methods*). As shown in Fig. 2, the frequencies of 256 end motifs in both plasma and urinary cfDNA of mice with different nuclease-knockout genotypes were organized in alphabetical order, forming the end-motif profile. Motifs starting with adenine (A), cytosine (C), guanine (G), and thymine (T) were highlighted in blue, red, green, and yellow, respectively. We observed certain distinct patterns in end-motif profiles across different mice. For instance, compared with WT mice, the plasma cfDNA of the *Dnase1l3*<sup>-/-</sup> mice showed periodic spikes in frequencies of the end motifs, typically at those end motifs with A ends, C ends, and G ends. For urinary cfDNA of the *Dnase1*<sup>-/-</sup> mice, the abundance of motifs with T ends was reduced significantly ( $P < 0.0001$ , Mann–Whitney *U* test) compared with the WT mice. Although it was visually hard to discern the difference when comparing plasma cfDNA of WT mice vs. *Dnase1*<sup>-/-</sup> mice, or urinary cfDNA of WT mice vs. *Dnase1l3*<sup>-/-</sup> mice, we hypothesized that the subtle differences in these end-motif profiles could be discerned when an appropriate analytical algorithm was adopted. Hence, in the following sections, we used an algorithm called NMF (15–19) to holistically analyze the 256 motifs as a whole instead of focusing on one or a few specific motif species. NMF is commonly used in processing audio spectrograms, e.g., deconvoluting acoustical signals into different acoustic events (20).

**Deconvolutional Analysis of End-Motif Profiles.** We applied NMF analysis to decompose end-motif profiles into several F-profiles. A total of 93 murine cfDNA samples with different genotypes of DNA nuclease knockouts were used for such NMF analysis, including plasma and urinary cfDNA (see details in *Materials and*

*Methods*). The optimal number of F-profiles was determined to be 6, with a high reproducibility and a low error (*SI Appendix, Fig. S1 and Table S1*). These profiles were named F-profiles I, II, III, IV, V, and VI.

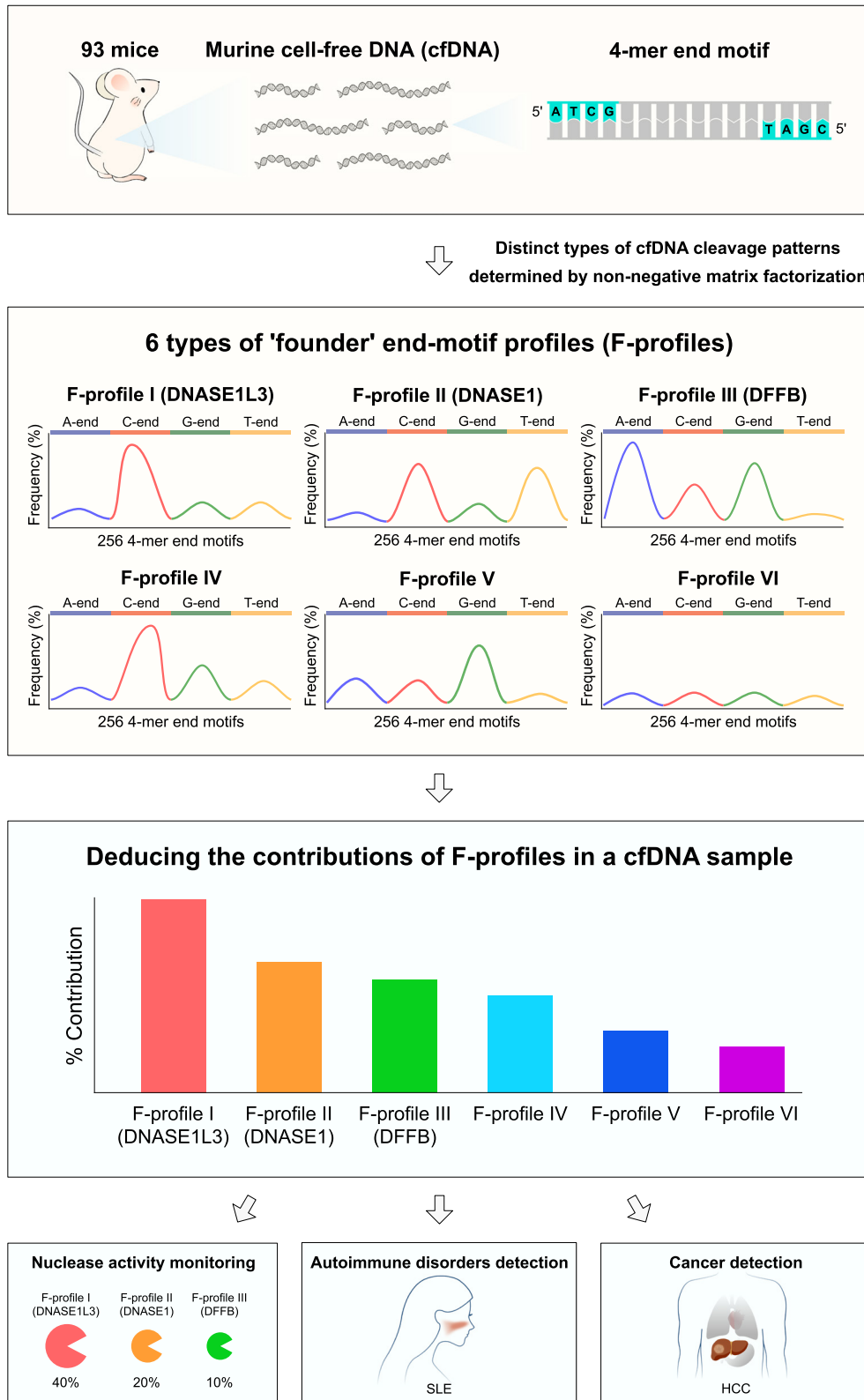
We could determine the proportional contribution of each F-profile, which was deduced by NMF, in an individual cfDNA sample, when the minimal error was achieved between an observed end-motif profile and the sum of F-profiles weighted by their proportional contributions (Fig. 3*A*). Such a mathematical process was referred to as F-profile-based deconvolutional analysis of end motifs in this study. To biologically link the F-profiles to possible DNA nuclease cleavages, we investigated the typical end motifs in an F-profile and measured its alteration in proportional contribution when depleting or enhancing a particular nuclease activity.

F-profile I displayed a predominance of C-end motifs (55%) and was characterized by the “CC” motifs (Fig. 3*B*), which was in line with DNASE1L3-cutting properties demonstrated in our previous studies (8, 9). We observed that the contributions of F-profile I in the plasma cfDNA of *Dnase1l3*<sup>-/-</sup> mice were significantly lower compared with that in WT mice (median: 2.7% vs. 35.4%; range: 0.0 to 4.6% vs. 19.5 to 47.9%) ( $P < 0.0001$ , Mann–Whitney *U* test). Hence, F-profile I was deemed to be a DNASE1L3-associated F-profile, which could be used to reflect the nuclease usage level of DNASE1L3.

F-profile II exhibited a major preference for T-end motifs (51%), with a significant enrichment observed for “TG” motifs (Fig. 3*C*). Such a preference was coincided with the DNASE1-cutting motifs (10). In WT mice, F-profile II contributions were significantly higher in urinary cfDNA in comparison with plasma cfDNA (median: 43.4% vs. 11.6%; range: 31.8 to 50.1% vs. 0.0 to 22.1%) ( $P < 0.0001$ , Mann–Whitney *U* test). Of note, the DNASE1 activity was known to be much higher in urine than in plasma for WT mice (10). Furthermore, a median of approximately eightfold reduction for F-profile II contributions was observed in both plasma and urinary cfDNA of *Dnase1*<sup>-/-</sup> mice when compared with the WT counterparts. Thus, F-profile II was deduced to be related to the DNASE1 activity. It was worth noting that 3 out of 10 *Dnase1*<sup>-/-</sup> mice still showed considerable contributions of F-profile II (DNASE1) in plasma cfDNA (Fig. 3*A*), perhaps implying that some other enzymes might play a complementary role to the DNASE1 and might yield a degree of compensation when cells lacked DNASE1. This hypothesis would need further experimental validation in future studies.

F-profile III comprised a substantial proportion of A-end motifs (40%) and was characterized by the preference for C and T nucleotides at the third and fourth positions in the 4-mer motifs, respectively, in the 5′ to 3′ direction (Fig. 3*D*). Notably, F-profile III showed concordant pattern with DFFB-cutting signatures (7). The contributions of F-profile III diminished significantly in the plasma cfDNA of *Dffb*<sup>-/-</sup> mice (median: 0.0%; range: 0.0 to 0.5%), compared with WT mice (median: 10.1%; range: 0.0 to 26.9%) ( $P < 0.001$ , Mann–Whitney *U* test). We noticed that the contributions of F-profiles IV (median: 0.0% vs. 2.3%) and V (median: 0.0% vs. 15.2%) also diminished in the plasma cfDNA samples of 6 *Dffb*<sup>-/-</sup> mice compared with WT mice. However, we no longer observed significant changes of F-profiles IV and V in another dataset (7, 21) comprising 11 *Dffb*<sup>-/-</sup> mice and 12 WT mice, whereas the significant reduction of F-profile III in *Dffb*<sup>-/-</sup> could be validated (*SI Appendix, Fig. S2*). Taken together, we concluded that only F-profile III was reproducibly associated with DFFB activity.

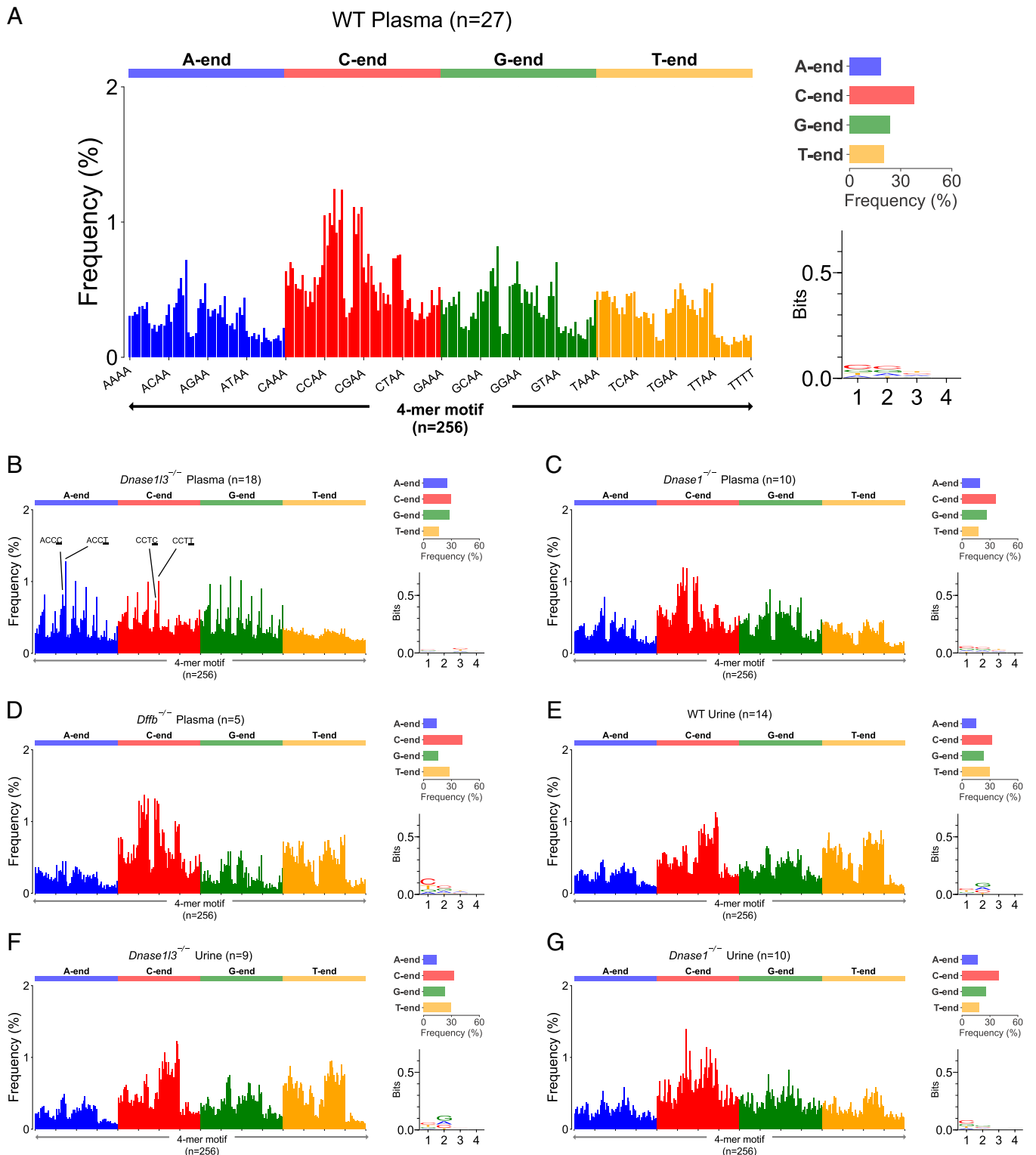
Although F-profile IV exhibited a high C-end preference (50%) which was to some extent reminiscent of F-profile I, it had several



**Fig. 1.** Schematic of distinct types of cfDNA cleavage analysis for cfDNA molecules. The terminal 4 nucleotides at each of the 5' fragment ends (i.e., 4-mer end motifs;  $n = 256$ ) were determined from 93 murine cfDNA samples, including WT mice and nuclease-deficient mice. Six categories of distinct types of cfDNA cleavage patterns were found, referred to as "founder" end-motif profiles (i.e., F-profiles), by applying NMF analysis to the 4-mer end-motif profiles. F-profiles I, II, and III were associated with the cutting preference of DNASE1L3, DNASE1, and DFFB, respectively. The distinct types of cfDNA cleavage patterns learned from murine cfDNA could be extrapolated to human cfDNA for informing the proportional contributions of F-profiles in both mouse and human cfDNA samples (referred to as deconvolutional analysis of end motifs), allowing the detection of immune diseases and cancers.

unique characteristics, for example, the absence of CC-end preference. F-profile IV also exhibited "G" base preferences at the second, third, and fourth positions in 4-mer motifs (Fig. 3E). F-profile V exhibited

a strong G-end preference (50%) (Fig. 3F). These results suggested that F-profile IV and V were not directly attributed to the previously established nucleases involved in cfDNA fragmentation (7), implying

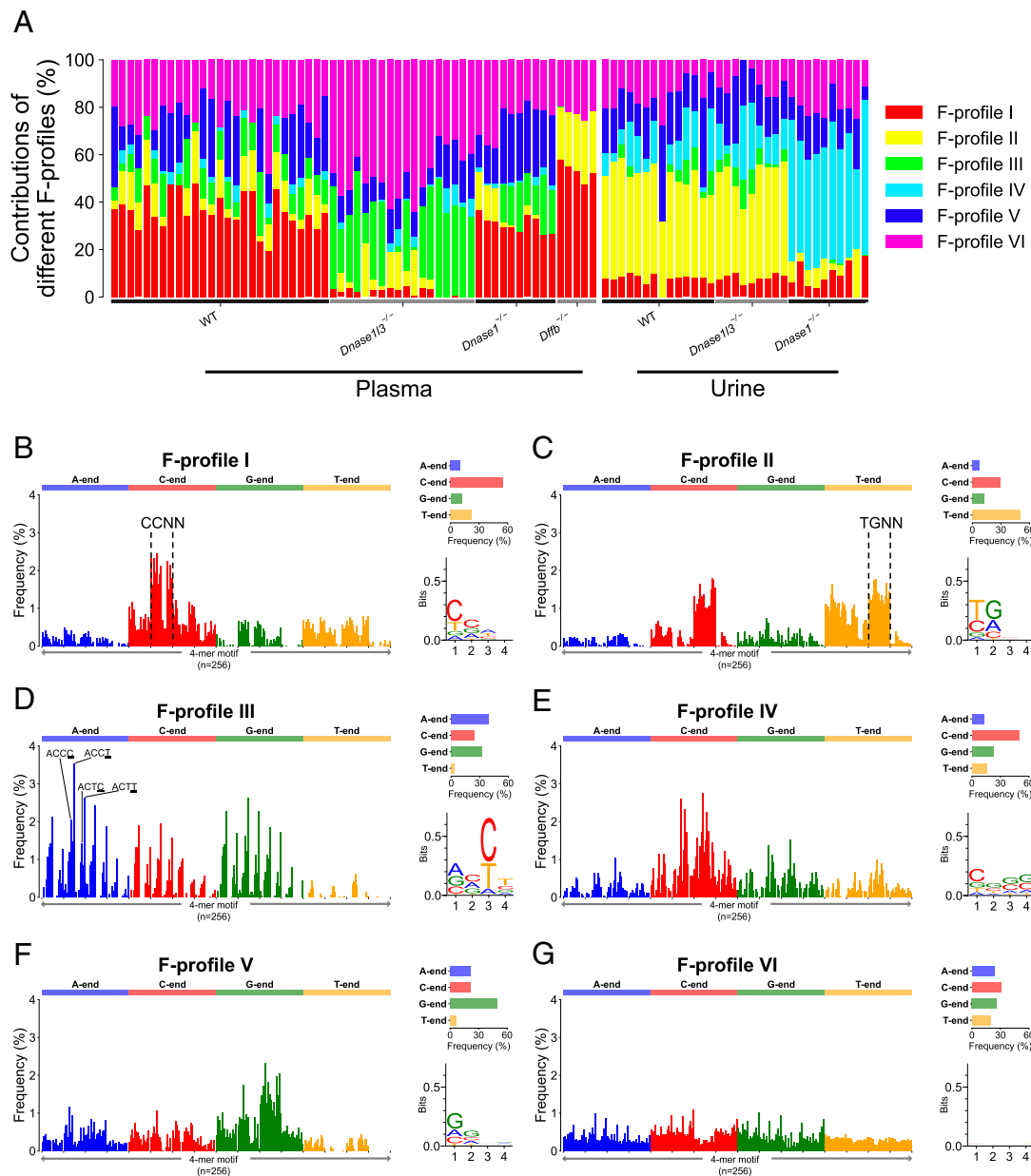


**Fig. 2.** The observed end-motif profiles of mouse plasma and urinary cfDNA molecules. The observed end-motif frequencies of plasma cfDNA from (A) WT mice, (B) *Dnase113*<sup>-/-</sup> mice, (C) *Dnase1*<sup>-/-</sup> mice, and (D) *Dffb*<sup>-/-</sup> mice, respectively. The observed end-motif frequencies of urinary cfDNA from (E) WT mice, (F) *Dnase113*<sup>-/-</sup> mice, and (G) *Dnase1*<sup>-/-</sup> mice, respectively.

that some other cleavage pathways might play roles in cfDNA fragmentation. Notably, F-profile VI showed a relatively even distribution across 256 motifs without obvious end motif preference (Fig. 3G).

**Deconvolutional Analysis of End Motifs during In Vitro Incubation of Mouse Plasma.** Han et al. previously demonstrated the stepwise fragmentations mediated by different DNA nucleases, such

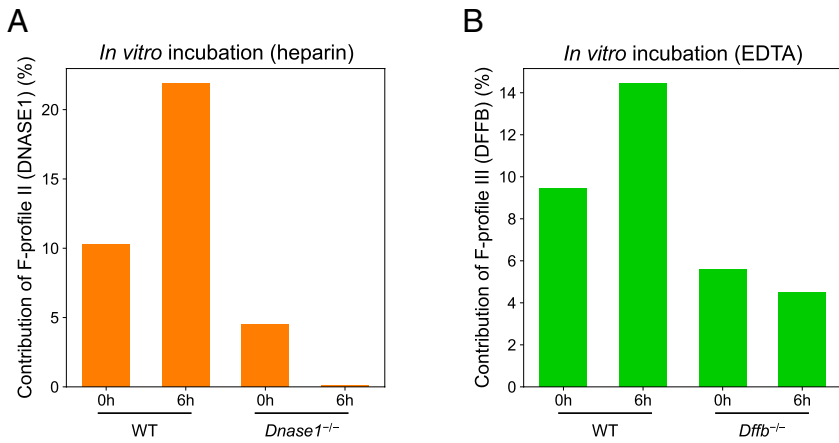
as DFFB, DNASE1L3, and DNASE1, based on different in vitro incubation conditions with the presence of EDTA or heparin (7). We tested whether F-profile could be used to reflect the degree of nuclease involvement by applying F-profile-based deconvolutional analysis to those independent samples. The previous study (7) has indicated that heparin could disrupt the nucleosomal structures and enhance DNASE1 cleavage (7). As shown in Fig. 4A, after 6-h



**Fig. 3.** Six F-profiles deduced from mouse plasma and urinary cfDNA using NMF analysis. (A) Proportional contribution of each F-profile in murine cfDNA samples with different knockout genotypes. (B–G) Plots for the six F-profiles.

incubation of whole blood in the presence of heparin, the average F-profile II (DNASE1) levels in WT mice increased 2.1 times from 10.3 to 21.9%, compared to the data at the time point of 0 h. In contrast, there was a reduction in F-profile II (DNASE1) level in plasma cfDNA of *Dnase1*-deficient mice after 6-h incubation [F-profile II level: 4.5% (6 h) vs. 0.0% (0 h)]. On the other hand, after the 6-h incubation of whole blood with the presence of EDTA, the mean F-profile III (DFFB) levels in the plasma of WT mice increased 1.5 times (9.4% vs. 14.4%) (Fig. 4B), compared to the data at the time point of 0 h. In contrast, F-profile III (DFFB) showed no apparent change in the plasma cfDNA of *Dffb*<sup>-/-</sup> mice (5.6% vs. 4.5%) between these two time points. The observation agreed with the previous conclusion that DFFB-preferred cleavage was enriched in the plasma cfDNA newly released from cells at the 6-h time point (7). These results further demonstrated the feasibility and biological relevance of revealing the linkage between nucleases and F-profile-based analysis. The other relevant F-profiles were shown in *SI Appendix*, Fig. S3.

**Deconvolutional Analysis of End Motifs for Human Plasma and Urinary cfDNA.** As the homology of amino acid sequences between human and mouse nucleases are 82%, 79%, and 76% for DNASE1L3 (11), DNASE1, and DFFB, respectively, we hypothesized that the deconvolutional analysis of end motifs established from the mouse data could be extrapolated to human cfDNA. To test this hypothesis, we estimated the contributions of distinct types of cfDNA cleavage patterns by applying the deconvolution algorithm on cfDNA from 18 human plasma samples and paired urine samples (Fig. 5A). As shown in Fig. 5B, F-profile I levels (DNASE1L3) in plasma cfDNA (median: 42.9%; range: 33.2 to 48.7%) were significantly higher than that in urinary cfDNA (median: 5.2%; range: 0.3 to 22.1%) ( $P < 0.0001$ , Mann–Whitney  $U$  test). Conversely, F-profile II (DNASE1) showed significantly higher proportional contributions in urinary cfDNA (median: 43.4%; range: 20.0 to 55.3%) than that in plasma cfDNA (median: 12.5%; range: 6.2 to 18.4%) ( $P < 0.0001$ , Mann–Whitney  $U$  test) (Fig. 5C). These data suggested that



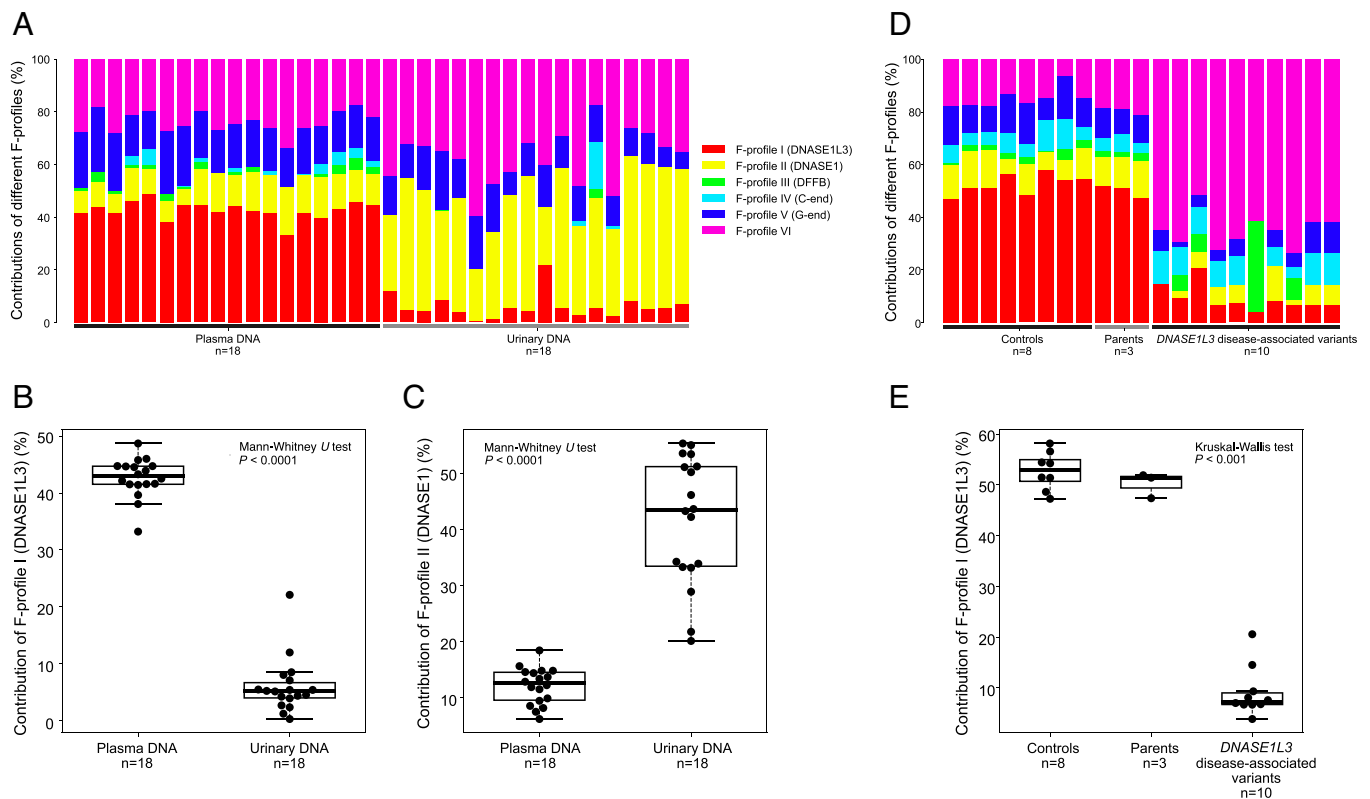
**Fig. 4.** Deconvolutional analysis of end motifs of mouse plasma cfDNA samples that were subjected to whole blood *in vitro* incubation. (A) F-profile II (DNASE1) levels in plasma cfDNA from WT mice before and after 6 h incubation in heparin-contained tube, and from *Dnase1*<sup>-/-</sup> mouse before and after 6 h heparin incubation. (B) F-profile III (DFFB) levels in plasma cfDNA from WT mice before and after 6 h incubation in EDTA-contained tube, and from *Dffb*<sup>-/-</sup> mice before and after 6 h EDTA incubation.

DNASE1L3 and DNASE1 played major roles in shaping plasma and urinary cfDNA fragmentation patterns, respectively. This conclusion was in agreement with previous reports (7, 8, 10, 11), suggesting the F-profile analysis could be generalizable between murine and human cfDNA samples.

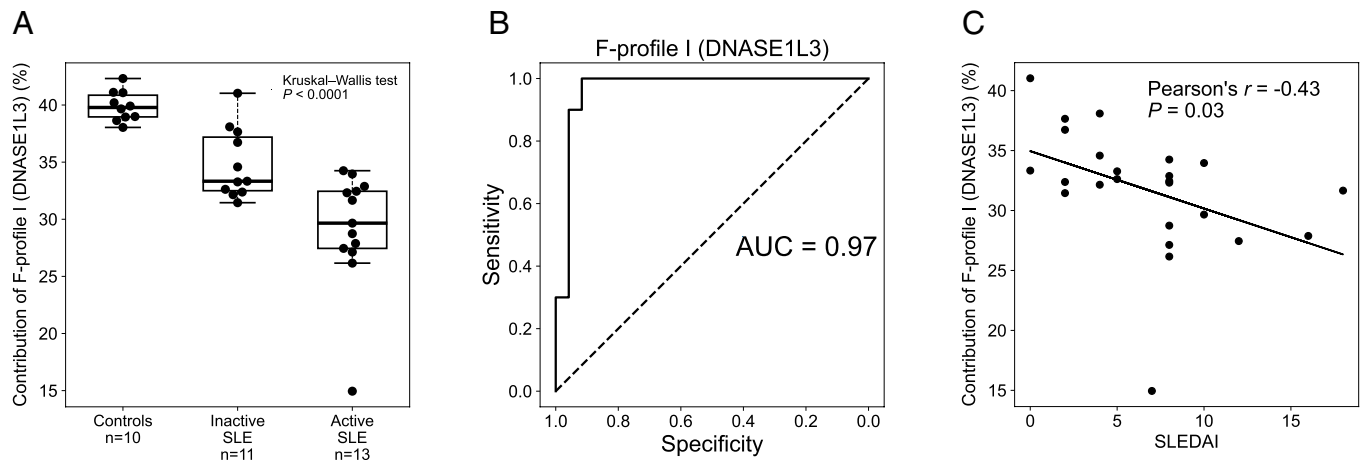
**Distinct Types of cfDNA Cleavages in Human Subjects.** Human subjects with DNASE1L3 deficiency would develop SLE-like symptoms with childhood onset, which was also referred to as familial SLE (9). We investigated the contributions of distinct types of cfDNA cleavages by analyzing plasma cfDNA from patients with both copies of *DNASE1L3* gene carrying genetic mutations (i.e., *DNASE1L3*-deficient) ( $n = 10$ ), parents of these patients ( $n = 3$ ) carrying one copy of a mutant *DNASE1L3* gene (the other copy was able to function), and healthy control subjects ( $n = 8$ ) (9). F-profile I (DNASE1L3) levels in plasma cfDNA

of patients with DNASE1L3 deficiency appeared to diminish significantly (median: 7.3%; range: 3.8 to 20.5%) compared with their parents (median: 51.4%; range: 47.4 to 51.9%) and healthy subjects (median: 52.9%; range: 47.3 to 58.2%) ( $P < 0.001$ , Kruskal–Wallis test) (Fig. 5 D and E).

Moreover, in a cohort comprising 10 healthy controls, 11 and 13 patients with inactive and active sporadic SLE (22), respectively, we observed that the DNASE1L3 usage levels gradually decreased across healthy subjects (median: 39.8%; range: 38.0 to 42.3%), patients with inactive SLE (median: 33.3%; range: 31.4 to 41.0%), and patients with active SLE (median: 29.7%; range: 14.9 to 34.2%) ( $P < 0.0001$ , Kruskal–Wallis test) (Fig. 6A and SI Appendix, Fig. S4A). The metric of DNASE1L3 usage level (F-profile I) enabled the differentiation between human individuals with and without SLE, with an AUC of 0.97 (Fig. 6B). The use of F-profiles II, III, IV, V, and VI resulted in worse performance, with



**Fig. 5.** Deconvolutional analysis of end motifs in paired human plasma and urinary cfDNA samples, and human plasma cfDNA of subjects with and without DNASE1L3 deficiency. (A) Bar chart of results of deconvolutional analysis between human plasma and urinary cfDNA. (B–C) Boxplots of F-profile I and II levels between human plasma and urinary cfDNA. (D) Bar chart of results of deconvolutional analysis in plasma cfDNA between subjects with and without *DNASE1L3* disease-associated variants. (E) Boxplot of F-profile I levels for healthy subjects, patients with DNASE1L3 deficiency, and parents of the patients.



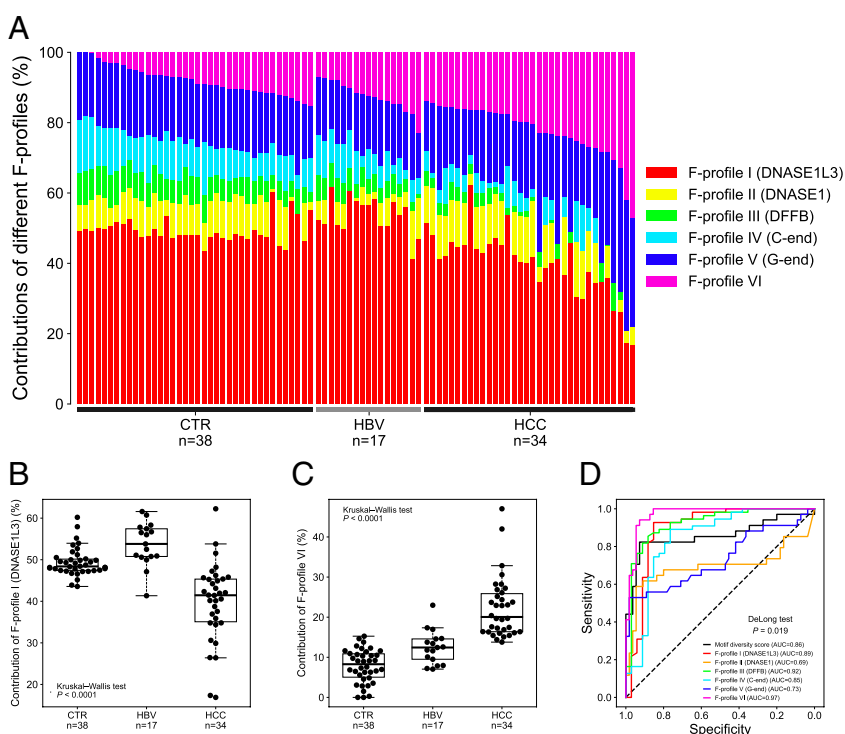
**Fig. 6.** Deconvolutional analysis of end motifs in plasma cfDNA of human subjects with and without SLE. (A) Boxplot of F-profile I levels (DNASE1L3) in plasma cfDNA across healthy control subjects, patients with inactive SLE, and patients with active SLE. (B) Area under the receiver operating characteristic (ROC) curve (AUC) for differentiation between patients with and without SLE using F-profile I. (C) Correlation between the SLEDAI and F-profile I levels in patients with SLE.

AUC values of 0.56, 0.75, 0.73, 0.61, and 0.87, respectively. The DNASE1L3 usage levels showed a negative correlation with the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) (Pearson's  $r = -0.43$ ;  $P = 0.03$ ) (Fig. 6C). Hence, the metric of F-profile I, which was linked to DNASE1L3, would inform the presence of autoimmune diseases, as well as facilitate the monitoring of disease progression.

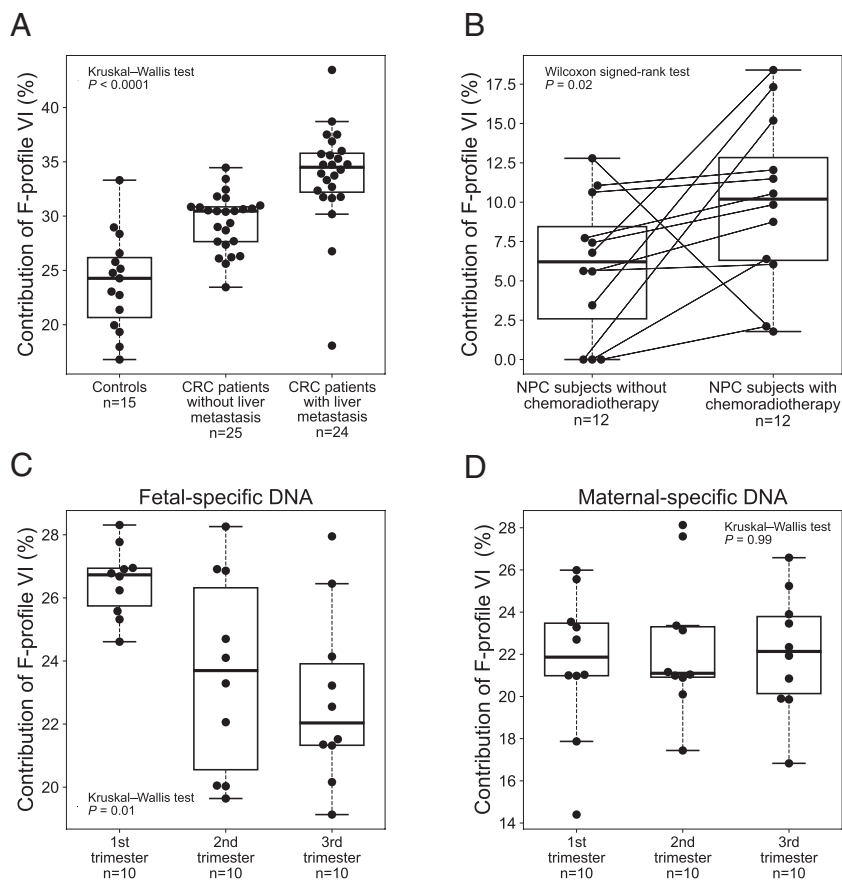
Besides the autoimmune disease model, patients with hepatocellular carcinoma (HCC) were reported to have aberrant DNASE1L3 activities (11). We further analyzed the contributions of distinct types of cfDNA cleavages in a cohort consisting of 38 healthy controls (CTR), 17 HBV carriers without HCC (HBV), and 34 patients with HCC from a previous study (HCC) (11) (Fig. 7A). Compared with healthy controls, F-profile I level was indeed found to be decreased by a median of 6.9% in HCC patients, whereas no appreciable change was observed in HBV carriers (Fig. 7B). Interestingly, among the 6 F-profiles, the most

discriminative power in detecting patients with HCC was F-profile VI (AUC: 0.97). F-profile VI was distinct from the other F-profiles, for seemingly lacking specific preference across the 256 end motifs (Fig. 7C and D). As a tool in discriminating HCC patients from HBV carriers, F-profile VI was superior to the previously reported motif diversity score (AUC: 0.86) ( $P = 0.019$ , DeLong test) (11). The motif diversity score was essentially a measure of the evenness of overall end-motif frequencies.

**Potential Biological Significance of F-profile VI.** As F-profile VI showed a promising differentiation power between the patients with and without HCC, we wondered whether any biological process was linked to F-profile VI. Because F-profile VI displayed a lack of obvious preference in the frequencies across 256 4-mer motifs, we hypothesized that cfDNA fragmentation occurring in patients with cancer might in part be induced by DNA damages via a mechanism which was distinct from the known apoptotic



**Fig. 7.** The distinct types of cfDNA cleavage analysis in plasma cfDNA of human subjects with and without HCC. (A) Bar chart of F-profile levels in plasma cfDNA of patients with and without HCC. (B) Boxplots of F-profile I and (C) VI levels in plasma cfDNA of patients with and without HCC. (D) ROC curves for the differentiation between non-HCC and HCC groups using different metrics, including motif diversity score and six F-profiles.



**Fig. 8.** The F-profile VI levels in plasma cfDNA of human subjects under different oxidative stress. (A) Boxplot of F-profile VI contributions in healthy control subjects, CRC patients without and with liver metastasis. (B) Boxplot of F-profile VI contributions in patients with NPC who were subjected to chemoradiotherapy or not. Boxplots of F-profile VI contributions in the (C) fetal- and (D) maternal-specific DNA in plasma cfDNA of pregnant women across first, second and third trimesters.

pathways, for example, DFFB and/or DNASE1L3-mediated DNA fragmentation (7, 23). Considering that oxidative stress was implicated in many steps in carcinogenesis, and cancer cells produced more oxidants than normal cells (24), including HCC (25), colorectal cancer (CRC) (26), and nasopharyngeal carcinoma (NPC) (27), we hypothesized that F-profile VI might be associated with oxidative stress. In the previous section, we observed that F-profile VI contribution significantly increased in patients with HCC (Fig. 7C). To further investigate this mechanism beyond HCC, we analyzed a cohort comprising 15 human healthy control subjects, 25 CRC patients without liver metastasis, and 24 CRC patients with liver metastasis. We observed that F-profile VI contributions were increased in CRC patients without liver metastasis (median: 30.5%; range: 23.4 to 34.5%) and with liver metastasis (median: 34.5%; range: 18.1 to 43.5%), compared with the healthy control group (median: 24.3%; range 16.8 to 33.3%) ( $P < 0.0001$ , Kruskal–Wallis test) (Fig. 8A and *SI Appendix, Fig. S4B*). Such a finding coincided with the report that oxidative stress increased in patients with CRC and further enhanced in CRC patients with liver metastasis (26, 28).

In addition, the increase of oxidative stress could be observed during chemotherapy (29) and radiotherapy (30). The chemoradiotherapy (e.g., concurrently treated by ionizing radiation and platinum coordination complexes such as cisplatin and carboplatin) would be expected to generate high levels of reactive oxygen species (ROS), which would likely induce oxidative stress and damage DNA (31). Indeed, the F-profile VI contributions in plasma were elevated in patients with NPC ( $n = 12$ ) who were subjected to cisplatin- or carboplatin-based chemoradiotherapy (median: 10.2%; range: 1.8 to 18.4%), compared with cfDNA from the same patient collected before the treatment (median:

6.2%; range: 0.0 to 12.8%) ( $P = 0.02$ , Wilcoxon signed-rank test) (Fig. 8B and *SI Appendix, Fig. S4C*).

On the other hand, the oxidative stress in the placenta was reported to decline as the gestational age increased (32). In this regard, we observed that the F-profile VI contributions in fetal-specific DNA molecules in maternal plasma, which were essentially of placental origin, exhibited a downward trend across the first, second and third trimesters, with median F-profile VI contributions of 26.7%, 23.7%, and 22.0%, respectively ( $P = 0.01$ , Kruskal–Wallis test) (Fig. 8C and *SI Appendix, Fig. S5*). However, such a decline trend was not observed in maternal-specific DNA molecules mainly of hematopoietic origin ( $P = 0.99$ , Kruskal–Wallis test) (Fig. 8D). Taken together, these observations suggested that F-profile VI might be at least in part associated with oxidative stress.

## Discussion

In this study, we developed an approach for discerning the distinct types of cfDNA cleavage patterns, referred to as F-profiles, using the NMF algorithm. The proportional contribution (i.e., weight) for each F-profile in a cfDNA sample could be determined when the product of F-profiles and their weights were closest to the observed end-motif profile of that sample. In contrast to the previous studies that focused on one specific nuclease activity each time using one end motif or several top-ranked end motifs (8–11), the approach investigated in this study could simultaneously assess a number of nuclease activities as well as other possible non-enzymatic fragmentation processes. We identified six distinct types of F-profiles. Based on mice with different nuclease-knockout genotypes, we had annotated the biological meanings of a number of F-profiles. F-profiles I, II, and III were linked to DNASE1L3,



DNASE1, and DFFB-mediated cleavages, respectively. Hence, the proportional contributions of F-profiles I, II, and III could reflect the nuclease usage levels related to these three nucleases.

The analytic framework for the distinct types of cfDNA cleavage analysis established from murine models was then extended to human cfDNA samples. Among those distinct types of cfDNA cleavage patterns, three of them were associated with known DNA nucleases, thus allowing the assessment of their nuclease usage levels. F-profile I (DNASE1L3) level was drastically reduced in patients with homozygous DNASE1L3 mutations who developed SLE-like symptoms. In addition, F-profile I (DNASE1L3) level was higher in human plasma cfDNA than in human urinary cfDNA, whereas F-profile II (DNASE1) level showed the opposite pattern. The observation was in agreement with the fact that the DNASE1L3 concentration was higher in human plasma than urine, whereas the DNASE1 concentration was higher in human urine (10). The nuclease activities across serial plasma samples that underwent in vitro incubation could be monitored using the NMF-based deconvolutional analysis of end motifs in this study. In addition, F-profile I (DNASE1L3) level declined in patients with HCC in which the *DNASE1L3* RNA expression was downregulated (11), compared with patients without HCC. These results suggested that it would be feasible to use deconvolutional analysis of end motifs to examine the nuclease activities in liquid biopsy, by making use of the founder end-motif profiles established from mice with various DNASE-knockout genotypes.

F-profile I (DNASE1L3) level appeared to be effective in differentiating the patients with and without sporadic SLE, with an AUC of 0.97. However, for the detection of patients with HCC, the best performance was achieved when using F-profile VI (AUC: 0.97), rather than those linked to the currently known DNA nucleases. F-profile VI did not exhibit an obvious link to certain end motifs. In other words, F-profile VI represented a nonspecific cleavage pattern. The increase in F-profile VI level was observed in patients with HCC compared to individuals without HCC. We conjectured that the increase in F-profile VI level might be associated with oxidative stress, providing that the various types of cancer cells would be subjected to the increase of oxidative stress (24). One possible mechanism would be that the oxidative stress could induce free radicals that might exert oxidative damage to DNA, thus likely at least in part causing DNA breaks distinct from the cleavages mediated by those well-studied nucleases such as DNASE1, DNASE1L3 and DFFB (6, 7). Such a hypothesis was in part supported by the observation that an increased level of F-profile VI was present in plasma of patients with NPC subjected to chemoradiotherapies that were expected to enhance the oxidative stress. Of note, mechanistically, ionizing radiation and chemotherapy drugs might cause DNA damages in various ways. For example, ionizing radiation could directly induce DNA double-strand breaks and indirectly introduce abasic sites and single-strand breaks via induced ROS (30). The cisplatin used in chemotherapy could be a ROS inducer, and it could also directly cause DNA intrastand diadducts (33). During chemoradiation therapy, the exact impact of radiological and chemical factors on F-profile VI remains to be explored. On the other hand, the decrease in F-profile VI contribution for fetal DNA molecules was seen in plasma of pregnant women, coincidentally in line with the fact that the oxidative stress in the placenta was expected to be downregulated at advanced gestational ages (32). Such a hypothesis regarding the biological linkage between F-profile VI and oxidative stress would require further experimental validation, for example, using the serial plasma DNA samples of a mouse model

in which the oxidative stress could be gradually induced by chemotherapeutic drugs. Nonetheless, the metric of the F-profile-based analysis led to a better performance in cancer detection than the motif diversity score (11). One possible reason might be that the cfDNA cleavage analysis was conducted in a way that multiple nucleases as well as other underlying fragmentation factors were holistically examined under one analytic framework. Such analysis may improve the signal-to-noise level, allowing for more precise quantification of distinct types of cfDNA cleavages involved in cfDNA fragmentation.

In addition to different DNA nucleases and other non-specific mechanisms of DNA degradation, differential chromatin accessibility may be another contributor to the generation of characteristic patterns of cfDNA end motifs. Genomic DNA in various tissues is characterized by specific nucleosomal patterns, such as histone-bound regions and open chromatin regions (34, 35). Such differential chromatin structures across tissues have been reported to be associated with cfDNA fragmentation patterns (36). For instance, it has been demonstrated that the nucleosomal footprints (i.e., recurrently protected regions) are often present in cfDNA molecules (3, 4, 37), which are associated with the tissues of origin (36, 38). The complex interplay of factors involving DNA fragmentation and differential chromatin accessibility is worthy of future exploration.

One question that remained to be solved in this study was that the biological meanings of F-profiles IV and V were still elusive. For F-profile IV, the predominant motif preference was at C ends, followed by G ends, T ends, and A ends, seemingly similar to F-profile I (DNASE1L3), but was characterized by several notable differences from F-profile I. For example, F-profile IV preferred “CG”-started ends instead of “CC” termini of F-profile I. Of note, compared with the plasma of WT mice, the contributions of F-profiles IV and V diminished in one dataset of *Dffb*<sup>-/-</sup> mice but did not show significant change in another dataset of *Dffb*<sup>-/-</sup> mice. We cannot rule out the possibility that the interplay among different nucleases as well as other biological factors such as chromatin remodeling factors involved in cfDNA fragmentation may contribute towards inter-batch variability among these *Dffb*<sup>-/-</sup> mice. It is worth noting the potential limitation of the current study that as a certain number of F-profiles may not be identified during the factorization step due to the limited number of nuclease-depleted mouse types available, the incomplete list of F-profiles may also affect the accuracy of F-profile contribution deduction. This issue could be alleviated with the availability of other types of nuclease knockout mice in the future. F-profile-based cfDNA fragmentation analysis is still at a nascent stage. It would be interesting to explore how sample heterogeneities, the complex interplay between nonenzyme/enzyme-mediated fragmentations and chromatin accessibility, and preanalytical experimental process would impact F-profile analysis in future studies. The potential room for improvement of F-profile analysis is to eliminate the need of prior information of nucleases responsible for 1-mer motif in the future version, it may broaden the applicability of this technology and unveil more insights for those F-profiles altered in mice with the deletion of nucleases.

In summary, this study provided an approach for studying the mechanisms involved in cfDNA fragmentation. This approach can be used for generating testable hypotheses that could yield further light on the mechanisms of cfDNA fragmentation. In addition to plasma, this approach can also be used in multiple bodily fluids, e.g., urine. This method is also of value in developing biomarkers for pregnancy complications, autoimmune diseases, and cancer.

## Materials and Methods

**Datasets of Murine cfDNA.** Paired-end massively parallel sequencing data of 93 murine cfDNA samples were obtained from previous studies (8–10, 39), including 60 plasma cfDNA samples and 33 urinary cfDNA samples. The mouse plasma cfDNA samples were taken from 27 WT mice, 10 mice with *Dnase1* gene deletion (*Dnase1*<sup>-/-</sup>), 18 mice with *Dnase1/3* gene deletion (*Dnase1/3*<sup>-/-</sup>), five mice with *Dffb* gene deletion (*Dffb*<sup>-/-</sup>), with a median number of paired-end reads of 50 million (range: 16 to 243 million). In addition, whole-genome sequencing data of mouse urinary cfDNA samples were obtained from 14 WT mice, 10 *Dnase1*<sup>-/-</sup> mice, and 9 *Dnase1/3*<sup>-/-</sup> mice (median number of paired-end reads: 43 million; range: 2 to 134 million). Furthermore, we analyzed 30 mouse plasma samples obtained from a previous study (7) that were subjected to in vitro incubation experiments for 0 h and 6 h in the conditions with EDTA or heparin (median number of paired-end reads: 37 million; range: 22 to 55 million). All animal studies were approved by the Animal Experimentation Ethics Committee of The Chinese University of Hong Kong.

**Datasets of Human cfDNA.** Paired-end sequencing data of 192 human plasma and 18 urinary cfDNA were obtained from previous studies (9, 39–41), including plasma cfDNA from eight healthy individuals, 10 patients with *DNASE1L3* disease-associated variants, three parents of the patients with mutant *DNASE1L3* gene (median number of paired-end reads: 108 million; range: 40 to 162 million); plasma cfDNA from 24 SLE patients and 10 healthy individuals (median paired-end reads: 120 million; range: 18 to 208 million); plasma cfDNA from 38 healthy individuals, 17 patients with chronic HBV infection but without HCC (i.e., HBV carriers), and 34 patients with HCC (median paired-end reads: 38 million; range: 18 to 65 million); and plasma cfDNA from 30 pregnant women across first trimester (12 to 14 wk; n = 10), second trimester (20 to 24 wk; n = 10), and third trimester (38 to 40 wk; n = 10) (median number of paired-end reads: 103 million; range: 52 to 186 million); and paired plasma and urinary cfDNA samples collected from 18 individuals without cancer (median paired-end reads: 126 million; range: 21 to 205 million). In addition, paired-end sequencing data of 94 human plasma was generated, including target-capture sequencing of plasma cfDNA from 15 healthy subjects, 25 patients with CRC but without liver metastasis, and 24 CRC patients with liver metastasis (median number of paired-end reads: 40 million; range: 16 to 89 million); and plasma cfDNA from 12 patients with NPC subjected to cisplatin- or carboplatin-based chemoradiotherapy, as well as paired plasma cfDNA samples before the treatment (median number of paired-end reads: 58 million; range: 20 to 113 million). The detailed clinical information for the cancer patients was summarized in *SI Appendix, Tables S2–S4*. All recruited human subjects gave written informed consent, and the study was approved by The Joint Chinese University of Hong Kong–Hospital Authority New Territories East Cluster Clinical Research Ethics Committee under the Declaration of Helsinki.

**End-Motif Frequency Calculation and Normalization.** End motifs were determined from the terminal 4-nucleotide sequence, i.e., 4-mer end motif, at each 5′ fragment end of cfDNA molecules (8). The observed frequency (O) of each of the motifs (i.e., a total of 256 motifs) was determined from the total number of fragment ends. As the different sequence contexts between the mouse and human genomes could cause biases when using the end motif patterns of murine cfDNA to interpret the data from human cfDNA, we performed a reference genome context-based normalization for end motif measurement. An expected 4-mer end-motif frequency (E) was introduced for this normalization step, which was determined by simulating 4-mer end motifs from a reference genome using a 4-nucleotide sliding window across each chromosome. For the data generated using target-capture sequencing, the reference genome herein refers to the probe-targeted regions (42). The normalized end motif frequency was calculated as a ratio of observed and expected frequencies (O/E ratio) and then divided by the sum of all 256 normalized motif frequencies. The end motif frequency mentioned in this study was termed the normalized end motif frequency of which the sum is equal to 100%.

**Defining the “Founder” End-Motif Profiles (F-profiles).** After obtaining the end-motif frequencies, a data matrix (*M*) was constructed in a way that each row indicates a cfDNA sample (a total of 93 murine cfDNA samples) and each column represents a type of end motif (a total of 256 end motifs), thus having the dimension of 93 × 256. The data matrix was subjected to NMF analysis (15, 16) to obtain two matrices *W* and *F*. The mathematical relationship among *M*, *W*, and *F* were shown below:

$$M = WF.$$

*M* was the result of the product of *W* and *F*, where *W* was the relative weight for each F-profile in a 93 × *n* matrix, where *n* corresponded to the number of F-profiles. *F* represented F-profiles in a *n* × 256 matrix. *W* and *F* were determined by minimizing the objective function below:

$$\|M - WF\|, \text{ subject to } W \geq 0 \text{ and } F \geq 0.$$

Singular value decomposition (SVD) was used to initialize the procedure of NMF. Such factorization analysis was implemented in the Python language by using the function of `sklearn.decomposition.NMF` (v1.1.1) (43).

To estimate the optimal number of F-profiles, a fivefold cross-validation pre-analysis was performed. Six F-profiles (*SI Appendix, Table S1*) were determined by considering the tradeoff between the reproducibility of factorized components and the value of objective function (i.e., end-motif profile reconstruction error) (*SI Appendix, Fig. S1*).

**Deducing Percentage Contributions of F-profiles.** The six F-profiles were deduced via NMF as mentioned above. The percentage contribution of each F-profile in a cfDNA sample could be determined using non-negative least square (NNLS) based deconvolution analysis. We let a matrix of *F* represent the deduced F-profiles. The end-motif frequencies of cfDNA molecules were represented by a vector of *X*. The F-profile level was denoted as *P* which could be determined by NNLS:

$$X = \sum_i (P_i \times F_i).$$

where *i* represented an integer index of a particular F-profile, ranging from 1 to 6. Furthermore, all the F-profile levels would be required to be non-negative with a sum of 100%:

$$P_i \geq 0, \forall i;$$

$$\sum_i P_i = 100\%.$$

NNLS was implemented based on the Python function of `scipy.optimize.nnls` (v1.8.1).

**Data, Materials, and Software Availability.** Raw sequencing data can be accessed in European Genome-Phenome Archive (EGA) (<https://www.ebi.ac.uk/ega/>), with the accession numbers EGAS00001000962 (44), EGAS00001003174 (45), EGAS00001003409 (46), EGAS00001003514 (47), EGAS00001004080 (48), EGAS00001004342 (49), EGAS00001005563 (50), EGAS00001006700 (51), EGAS00001006701 (52); and Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>), with the accession number PRJNA842499 (53). All study data are included in the article and *SI Appendix*. Previously published data were used for this work (7–11, 22, 39–41).

**ACKNOWLEDGMENTS.** This work was supported by Innovation and Technology Commission of the Hong Kong SAR Government (InnoHK initiative) and the Li Ka Shing Foundation.

Author affiliations: <sup>a</sup>Centre for Novostics, Hong Kong Science Park, Pak Shek Kok, Hong Kong SAR, China; <sup>b</sup>Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>c</sup>Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>d</sup>State Key Laboratory of Translational Oncology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>e</sup>Department of Surgery, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>f</sup>Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>g</sup>Medical Data Analytics Centre, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; and <sup>h</sup>Department of Clinical Oncology, Sir Y. K. Pao Centre for Cancer, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

Author contributions: Z.Z., M.-J.L.M., R.W.K.C., K.C.A.C., Y.M.D.L., and P.J. designed research; Z.Z., M.-J.L.M., R.W.Y.C., W.K.J.L., W.G., S.C.D., Q.Z., P.P.H.C., S.C.Y.Y., J.Y.C.T., C.-C.S., J.W., V.W.S.W., G.L.H.W., S.L.C., E.P.H., B.B.Y.M., and A.T.C.C. performed research; Z.Z., M.-J.L.M., W.P., X.H., L.J., and P.J. analyzed data; R.W.Y.C., W.K.J.L., W.G., S.C.Y.Y., J.Y.C.T., C.-C.S., J.W., V.W.S.W., G.L.H.W., S.L.C., E.P.H., B.B.Y.M., and A.T.C.C. performed case recruitment; and Z.Z., M.-J.L.M., Y.M.D.L., and P.J. wrote the paper.

1. K. Sun *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.* **112**, E5503–E5512 (2015).
2. W. Li *et al.*, CancerDetector: Ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* **46**, e89 (2018).
3. Y. M. D. Lo *et al.*, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
4. H. Markus *et al.*, Analysis of recurrently protected genomic regions in cell-free DNA found in urine. *Sci. Transl. Med.* **13**, eaaz3088 (2021).
5. Y. M. D. Lo, D. S. C. Han, P. Jiang, R. W. K. Chiu, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021).
6. D. S. C. Han, Y. M. D. Lo, The nexus of cfDNA and nuclease biology. *Trends Genet.* **37**, 758–770 (2021).
7. D. S. C. Han *et al.*, The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106**, 202–214 (2020).
8. L. Serpas *et al.*, Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 641–649 (2019).
9. R. W. Y. Chan *et al.*, Plasma DNA profile associated with DNASE1L3 gene mutations: Clinical observations, relationships to nuclease substrate preference, and in vivo correction. *Am. J. Hum. Genet.* **107**, 882–894 (2020).
10. M. Chen *et al.*, Fragmentomics of urinary cell-free DNA in nuclease knockout mouse models. *PLoS Genet.* **18**, e1010262 (2022).
11. P. Jiang *et al.*, Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
12. L. Chen *et al.*, Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. *Cell Res.* **31**, 589–592 (2021).
13. C. Jin *et al.*, Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection. *Mol. Oncol.* **15**, 2377–2389 (2021).
14. V. Sisirak *et al.*, Digestion of chromatin in apoptotic cell microparticles prevents autoimmunity. *Cell* **166**, 88–101 (2016).
15. D. Lee, S. H. Sebastian, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
16. G. L. Stein-O'Brien *et al.*, Enter the matrix: Factorization uncovers knowledge from omics. *Trends Genet.* **34**, 790–805 (2018).
17. S. Zhang, X. J. Zhou, "Matrix factorization methods for integrative cancer genomics" in *Cancer Genomics and Proteomics: Methods and Protocols*, N. Wajapeyee, Ed. (Springer, New York, 2014), pp. 229–242.
18. S. Zhang *et al.*, Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* **40**, 9379–9391 (2012).
19. S. Zhang, Q. Li, J. Liu, X. J. Zhou, A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* **27**, i401–i409 (2011).
20. P. Smaragdis, J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription" in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, (IEEE Conference, New Paltz, NY, USA, 2003), pp. 177–180.
21. S. C. Ding *et al.*, Jagged ends on multinucleosomal cell-free DNA serve as a biomarker for nuclease activity and systemic lupus erythematosus. *Clin. Chem.* **926**, 917–926 (2022).
22. R. W. Y. Chan *et al.*, Plasma DNA aberrations in systemic lupus erythematosus revealed by genomic and methylomic sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5302–E5311 (2014).
23. T. Watanabe, S. Takada, R. Mizuta, Cell-free DNA in blood circulation is generated by DNase1L3 and caspase-activated DNase. *Biochem. Biophys. Res. Commun.* **516**, 790–795 (2019).
24. H. J. Forman, H. Zhang, Targeting oxidative stress in disease: Promise and limitations of antioxidant therapy. *Nat. Rev. Drug Discov.* **20**, 689–709 (2021).
25. M. Jo *et al.*, Oxidative stress is closely associated with tumor angiogenesis of hepatocellular carcinoma. *J. Gastroenterol.* **46**, 809–821 (2011).
26. S. B. Singh, K. Dahiya, A. Bharti, Lipid peroxidation and antioxidant status in colorectal cancer. *JK Pract.* **11**, 403–406 (2005).
27. Y. J. Huang *et al.*, Nitrate and oxidative DNA damage as potential survival biomarkers for nasopharyngeal carcinoma. *Med. Oncol.* **28**, 377–384 (2011).
28. L. M. van der Waals *et al.*, Increased levels of oxidative damage in liver metastases compared with corresponding primary colorectal tumors: Association with molecular subtype and prior treatment. *Am. J. Pathol.* **188**, 2369–2377 (2018).
29. N. I. Weijl *et al.*, Cisplatin combination chemotherapy induces a fall in plasma antioxidants of cancer patients. *Entomol. Exp. Appl.* **103**, 1331–1337 (1998).
30. E. I. Azzam, J. P. Jay-Gerin, D. Pain, Ionizing radiation-induced metabolic oxidative stress and prolonged cell injury. *Cancer Lett.* **327**, 48–60 (2012).
31. V. Sosa *et al.*, Oxidative stress and cancer: An overview. *Ageing Res. Rev.* **12**, 376–390 (2013).
32. J. Basu *et al.*, Placental oxidative status throughout normal gestation in women with uncomplicated pregnancies. *Obstet. Gynecol. Int.* **2015**, 276095 (2015).
33. J. Hu, J. D. Lieb, A. Sancar, S. Adar, Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11507–11512 (2016).
34. R. Sadeh *et al.*, ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* **39**, 586–598 (2021).
35. W. Meuleman *et al.*, Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
36. M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
37. K. K. Budhreja *et al.*, Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. *Sci. Transl. Med.* **15**, eabm6863 (2023).
38. K. C. A. Chan *et al.*, Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E8159–E8168 (2016).
39. P. Jiang *et al.*, Detection and characterization of jagged ends of double-stranded DNA in plasma. *Genome Res.* **30**, 1144–1153 (2020).
40. T. H. T. Cheng *et al.*, Noninvasive detection of bladder cancer by shallow-depth genome-wide bisulfite sequencing of urinary cell-free DNA for methylation and copy number profiling. *Clin. Chem.* **65**, 927–936 (2019).
41. P. Jiang *et al.*, Gestational age assessment by methylation and size profiling of maternal plasma DNA: A feasibility study. *Clin. Chem.* **63**, 606–608 (2017).
42. M. J. L. Ma *et al.*, Topologic analysis of plasma mitochondrial DNA reveals the coexistence of both linear and circular molecules. *Clin. Chem.* **65**, 1161–1170 (2019).
43. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. R. W. Y. Chan, Plasma DNA aberrations in systemic lupus erythematosus revealed by genomic and methylomic sequencing. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001000962>. Accessed 26 November 2014.
45. L. Serpas, Dnase1l3 knockout causes aberrations in plasma DNA fragmentation. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001003174>. Accessed 28 December 2018.
46. P. Jiang, Plasma DNA motif analysis. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001003409>. Accessed 5 January 2020.
47. D. S. C. Han, The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3 and DFFB. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001003514>. Accessed 30 January 2020.
48. P. Jiang, End structure of DNA in plasma: detection, characterization and diagnostic applications. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001004080>. Accessed 14 August 2020.
49. R. W. Y. Chan, Plasma DNA profile in DNASE1L3 deficiency. European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001004342>. Accessed 5 October 2020.
50. S. C. Ding, Jagged ends of plasma DNA (mouse). European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001005563>. Accessed 19 May 2022.
51. Z. Zhou, Plasma DNA end analysis (mouse). European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001006700>. Accessed 20 December 2022.
52. Z. Zhou, Cell-free DNA cleavages analysis (human). European Genome-Phenome Archive (EGA). <https://ega-archive.org/studies/EGAS00001006701>. Accessed 20 December 2022.
53. M. Chen, Fragmentomics of urinary cell-free DNA in nuclease knockout mouse models. Sequence Read Archive (SRA). <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA842499>. Accessed 25 May 2022.