nature communications

Article

https://doi.org/10.1038/s41467-024-50850-8

DNA methylation and gene expression as determinants of genome-wide cell-free DNA fragmentation

Received: 24 July 2023

Accepted: 23 July 2024

Published online: 06 August 2024

Check for updates

Michaël Noë[1,2], Dimitrios Mathios[1], Akshaya V. Annapragada[1], Shashikant Koul[1], Zacharia H. Foda[1], Jamie E. Medina[1], Stephen Cristiano[1], Christopher Cherry[1], Daniel C. Bruhm[1], Noushin Niknafs[1], Vilmos Adleff[1], Leonardo Ferreira[1], Hari Easwaran[1], Stephen Baylin[1], Jillian Phallen[1], Robert B. Scharpf[1] ✉ & Victor E. Velculescu[1] ✉

Circulating cell-free DNA (cfDNA) is emerging as an avenue for cancer detection, but the characteristics of cfDNA fragmentation in the blood are poorly understood. We evaluate the effect of DNA methylation and gene expression on genome-wide cfDNA fragmentation through analysis of 969 individuals. cfDNA fragment ends more frequently contained CCs or CGs, and fragments ending with CGs or CCGs are enriched or depleted, respectively, at methylated CpG positions. Higher levels and larger sizes of cfDNA fragments are associated with CpG methylation and reduced gene expression. These effects are validated in mice with isogenic tumors with or without the mutant IDH1, and are associated with genome-wide changes in cfDNA fragmentation in patients with cancer. Tumor-related hypomethylation and increased gene expression are associated with decrease in cfDNA fragment size that may explain smaller cfDNA fragments in human cancers. These results provide a connection between epigenetic changes and cfDNA fragmentation with implications for disease detection.

Cell-free DNA (cfDNA) has been the focus of research in blood-based biomarkers for early detection and monitoring of cancer. Normally, nuclear DNA is packaged and condensed within chromosomes in part by wrapping around histone cores[1,2]. In the process of and after cellular death, DNA is digested by DNAses, in part to prevent the release of unbound DNA which can act as auto-antigens[3–6]. DNA fragments that are tightly wrapped around histone cores, collectively called nucleosomes, appear to be protected from further digestion[7]. These fragments are those that are typically represented in cfDNA and can be collected with a simple blood draw.

With the development of high-throughput sequencing methods, it has become possible to study genome-wide features of cfDNA fragmentation, or the cfDNA fragmentome, including those related to the underlying nucleosomes. In healthy individuals, the positioning of nucleosomes as well as chromatin states shows striking similarity to those of myelocytic and lymphocytic cells[8–10]. Similarly, methylation profiles of cfDNA from individuals without cancer are very similar to DNA methylation of leukocytes[11,12]. Although epigenetic changes are related to genome packaging[13] and chromatin structure[14] as well as gene expression[15], until now there have been limited studies of the connection between methylation, expression and cfDNA fragmentation[16,17]. A recent study identified an increase in cfDNA fragment ends at sites of CpG methylation[18]. However, none of these studies have examined the underlying impact of these changes on cfDNA motifs and fragment size at locations of recurrent fragment breakpoints nor demonstrated a direct connection between epigenetic changes and cfDNA fragmentation.

[1]The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [2]Present address: Netherlands Cancer Institute, Amsterdam, The Netherlands. ✉e-mail: rscharpf@jhu.edu; velculescu@jhmi.edu

In this study, we analyze the features related to motifs associated with nucleosome positioning and breakpoints of cfDNA fragments in both healthy individuals and patients with cancer. We show how epigenetic marks give rise to specific patterns of fragmentation of cfDNA and how these are related to both methylation and gene expression. Using this information, we illustrate how differentially methylated CpG's in specific sequence contexts can be used to identify differences in cfDNA fragmentation between healthy individuals and cancer patients.

Results

Enrichment of genomic sequences at preferred cfDNA fragment end positions

We investigated the frequency and composition of cfDNA start and end sequences, as these have been previously described as non-random and potentially related to cleavage by endogenous DNAses[19–21]. To rigorously identify cfDNA end positions, we pooled cfDNA sequence data from low-coverage whole-genome sequencing from a cohort of healthy individuals ($n = 543$) and investigated the frequency of cfDNA breakpoints at every evaluable position in the genome. We only considered fragment reads with high sequence and mapping quality and calculated the ratio of the number of cfDNA fragments starting or ending at a particular position compared to the number of fragments with start or end positions within 50 bp surrounding that location. We used this approach to account for differences in cfDNA fragment size, coverage across the genome, and local copy number polymorphisms[8], as well as to identify "preferred" fragment end positions which we defined as genomic locations with a higher number of fragment ends than would be expected by random fragmentation (see "Methods" section).

We evaluated cfDNA fragments with more frequently observed end-positions and found that these were enriched for specific motifs. These typically included a thymine or an adenine before the start of the cfDNA fragment and two cytosines (A/T|CC) or a cytosine followed by a guanine (A/T|CG) as the first two nucleotides of the cfDNA fragment (Fig. 1a and Supplementary Fig. 1). We reasoned that end sequences that occurred recurrently among cfDNA fragments would likely represent those locations protected by nucleosome occupancy, and found that the frequency of these base motifs increased further at preferred recurrent ends among healthy individuals (Fig. 1a, b). A/T|CC and A/T|CG preferred cfDNA fragment ends were observed at rates much higher than theoretically expected in the genome (26.5x for A/T|CC and 5.5x for A/T|CG) (Fig. 1a; $p < 0.0001$, t-test), while the frequencies of DNA ends from fragments generated through sonication of genomic DNA from lymphoblastoid cell lines were close to theoretical abundances. Consistent with the observations above, the genome-wide locations of preferred cfDNA ends were observed to be higher in the first exons of genes, regions known to have highly ordered nucleosomes, and lower in the remaining gene bodies, which are more likely to contain less consistent nucleosome structures (Fig. 1c).

To understand the underlying basis for the enrichment of A/T|CG end sequences, we examined available x-ray crystal and cryo-EM

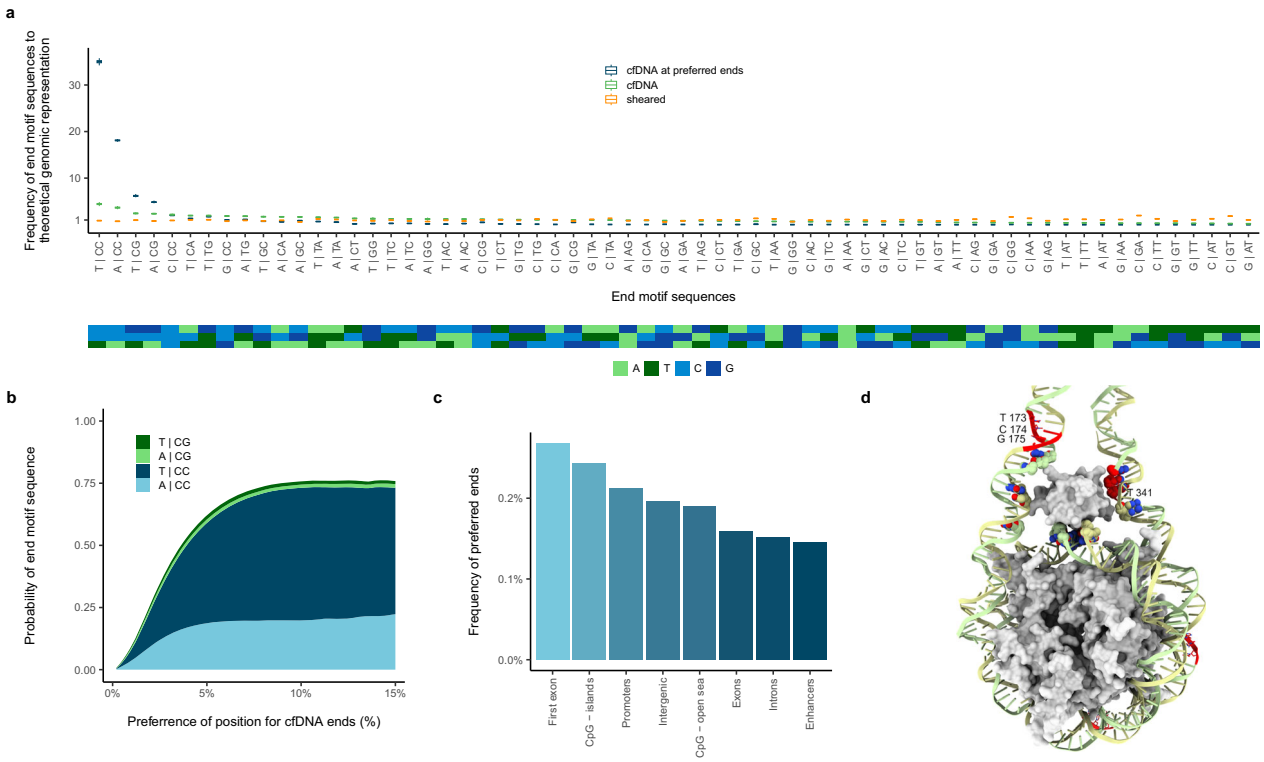

Fig. 1 | Enrichment of motifs at the ends of cfDNA-fragments and at preferred cfDNA-fragment ends. The 3 bp motifs are located around the fragment, with one base outside the fragment and the first two bases of the fragment. The vertical line indicates the start of the fragment. **a** Frequency of 3 bp DNA-motifs at the ends of DNA fragments after shearing by sonication ($n = 9$), at the ends of cfDNA fragments ($n = 543$), and at "preferred ends" of cfDNA that were observed at genomic positions where at least 5% of cfDNA fragments started at that position ($n = 543$) ("Methods" section). The relative frequencies are normalized by the observed occurrence of 3 bp motifs in the human genome. The middle hinge corresponds to the median, while the lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 × interquartile range from the hinge. All frequency box plots have error-bars, although the confidence intervals or individual data points may be too small to be visible. **b** Increased preference of motifs in preferred cfDNA-fragment end-positions, as measured by an increased ratio of the number of cfDNA-fragments starting or stopping at a certain position, over the total amount of cfDNA fragments overlapping a window of ±50 bases around that position. These included enrichment of T|CC and A|CC as well as T|CG and A|CG at these recurrent cfDNA end-positions. **c** Frequency of preferred ends of cfDNA in different functional genomic regions. **d** X-ray crystal structure of PDB entry 7COW[67]. TCG motifs colored red, nucleosome protein shown in gray surface with the histone H1 linker at the top. Bases within 5 Å of H1 linker shown as spheres.

https://doi.org/10.1038/s41467-024-50850-8

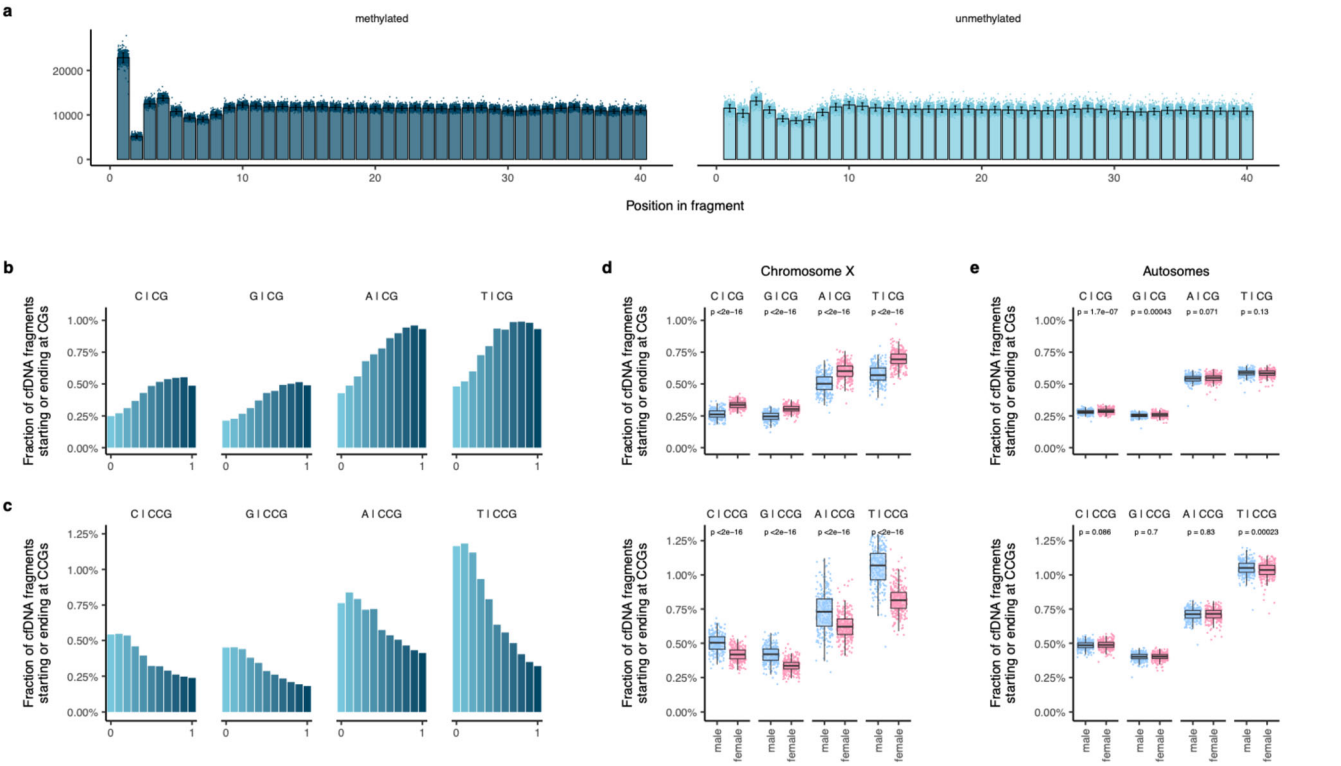

Fig. 2 | DNA methylation and cfDNA-fragmentation. a The frequency of observed CpG's at different positions in cfDNA fragments, counted from the cfDNA break, differed between methylated and unmethylated CpG sites per healthy individual ($n = 543$). The bar represents the mean value and the whiskers correspond to the standard deviation. While unmethylated CpG's showed a more equal distribution over the cfDNA fragments, methylated CpG's showed enrichment at the beginning of cfDNA-fragments. **b** The preference of fragments to start at a CpG increases with higher levels of methylation of that CpG. The preference for cfDNA-fragments to start at a CpG is measured in aggregate in 543 healthy individuals by the ratio of cfDNA-fragments starting with a CpG, over the total amount of cfDNA fragments overlapping a window of ±50 bases around the first base of that CpG. The degree of methylation at each CpG is indicated as the beta value in the horizontal axis using methylation data from cfDNA of healthy individuals at ~850 K CpG positions[11]. **c** The opposite relationship is seen when a CpG is preceded by a cytosine: there is a preference for cfDNA fragments to start with CCG, when the CpG in this motif is not methylated. **d** CpG's on chromosome X are known to be differently methylated in male ($n = 264$) and female individuals ($n = 279$). Due to X-inactivation by methylation of CpG-islands, these CpG's show a difference in the preference of cfDNA-fragment end-positions. With increased methylation, more fragments start with a CpG, while less fragments start with a CCG. *P*-values were calculated using a two-sample *t*-test, without correction for multiple testing. **e** On autosomes, we do not observe a difference in methylation between male ($n = 264$) and female individuals ($n = 279$). *P*-values were calculated using a two-sample *t*-test, without correction for multiple testing. The middle hinge in the boxplots corresponds to the median, while the lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 × interquartile range from the hinge.

structures of DNA bound to nucleosomes. We found that in 76% of unique structures an A/T|CG motif was found close to the H1 linker (within 5 Å) or centered 167 bp away, a distance that was identical to the known median fragment length of cfDNA molecules in healthy individuals (Fig. 1d)[8]. In contrast to fragment end sequences, we observed that interior regions of cfDNA fragments were enriched in adenines and thymines, with a 10–11 bp periodicity in the frequency of these nucleotides over the length of the fragment (Supplementary Fig. 1). These observations are consistent with current predictions of DNA wrapping around a histone-core and the necessity to alternate rigid DNA regions (C and G-rich) with more flexible regions (A- and T-rich) to wrap nearly two turns around the nucleosome[22].

Impact of CpG methylation on cfDNA fragmentation

Given the preponderance of fragment ends containing CGs, we wondered whether epigenetic marks of these sites could affect cfDNA fragmentation. We mapped cfDNA fragments to the genome and evaluated their ends with respect to previously identified methylated and unmethylated CpG sites of cfDNA from healthy individuals (comprised of men and women across young and old age groups) obtained from methylation arrays evaluating 850 K high-fidelity cfDNA CpG sites[11] of 97 individuals, as well as 23.6 M CpG sites from whole-genome bisulfite sequencing of cfDNA of 23 individuals[12]. We observed methylated CpGs were enriched at the ends of A/T|CG cfDNA fragments, while unmethylated CpGs were distributed evenly over the length of these molecules (Fig. 2a and Supplementary Figs. 2 and 3). To quantitatively assess the enrichment of fragment ends at CpGs, we calculated for each CpG the fraction of cfDNA fragments starting or ending at this dinucleotide position over the number of fragments with a start or end at any position within 50 bp around each CpG. We observed that the mean fraction of preferred ends increased as much as 2.4 fold with higher levels of methylation ($p < 0.0001$, *t*-test) (Fig. 2b and Supplementary Fig. 4a). Furthermore, CG cfDNA fragment ends were enriched as much as 2.2 fold at locations of methylated CpG's throughout the genome, including in CpG islands, shores, shelves and open sea regions (Supplementary Figs. 5 and 6), revealing that enrichment of methylated CpG fragments was a universal characteristic of cfDNA in these regions.

We observed that methylated CG end sequences were preferentially enriched even when they overlapped frequently observed CC fragment end sequences. When N|CC sequences were followed by guanine resulting in N|CCG, the typical N|CC end motifs were reduced in frequency as these competed with the overlapping C|CG motif that was enriched when CpG sites containing these sequences were methylated (Fig. 2b, c and Supplementary Figs. 3 and 4). The overall impact of this competition resulted in a dramatic reduction of mean N|

CCG fragment end sequences at methylated CpG positions that was even greater than the corresponding increase in fragment ends at N|CG, as seen for example, with the 3.69 fold reduction in mean T|CCG end sequences (95% CI: 3.46-3.92), compared to a 2.23 fold increase in C|CG end motifs (95% CI: 2.20-2.26). Similar results for N|CG or N|CCG end motifs were observed throughout the genome, including in CpG islands, shores, shelves, and open sea regions (Supplementary Figs. 5 and 6), as well as when using larger windows (75–125 bp) around the CpG site (Supplementary Fig. 7).

To provide additional biological evidence for the link between methylated CGs and cfDNA fragmentation, we compared cfDNA fragments arising from the X chromosome among healthy individuals, as it is well established that one copy of the two X chromosomes is inactivated by methylation of CpG islands in women, while these regions on the single X chromosome in men are not methylated[23,24]. In line with our observation of methylation-induced fragment end enrichment, cfDNA fragments ending with CG were enriched 1.23 fold in women compared to men (95% CI: 1.18–1.30), and fragment ends with CCG were preferentially reduced by 0.81 fold (95% CI: 0.74–0.87) at locations of X chromosome CpG islands, but these differences were not observed on the autosomes in men and women (median fold changes of 1.0 (95% CI: 0.99–1.03) and 1.0 (95% CI: 0.98–1.01) for CG and CCG ends, respectively) (Fig. 2d). Although this trend continued in CpG shores, we observed higher CG fragment end enrichment in men compared to women in CpG shelves and open sea, consistent with the previously reported increased methylation on the male X chromosome in these regions (Supplementary Fig. 8)[13,25].

Effects of methylation and gene expression on coverage and size of cfDNA fragments

In addition to the enrichment of cfDNA fragment-end positions at sites of epigenetic marks, we observed that cfDNA sequence coverage (the average number of cfDNA molecules overlapping a specific position) was related to methylation levels ($r = 0.6$, $p < 0.0001$, Pearson correlation test; Supplementary Fig. 9a), and was up to 1.7 fold higher across regions of CpG islands that were methylated compared to those that were not methylated ($p < 0.0001$, t-test; Fig. 3a). Given the connection between CpG island methylation and expression, we evaluated the relationship between gene expression at transcription start sites (TSS) and cfDNA fragmentation patterns. There was an inverse relationship between cfDNA coverage at TSS and expression levels of nearby genes ($r = -0.48$, $p < 0.0001$, Pearson correlation test; Supplementary Fig. 9b). Overall levels of cfDNA fragments that overlapped TSSs of unexpressed genes were up to 3.7 fold higher than at regions of expressed genes ($p < 0.0001$, t-test; Fig. 3b), likely due to the lack of destabilizing effects of transcription factors on nucleosomes and increased nucleosome occupancy[26].

Concordant with higher cfDNA coverage, we observed changes in cfDNA fragment sizes at these regions, including fragments 4-5 bp smaller at areas 800-1000 bp upstream of TSSs of highly expressed compared to unexpressed genes (164.5 bp vs 168.6 bp, respectively, $p < 0.0001$, t-test) or in regions surrounding unmethylated CpG islands compared to highly methylated CpGs (165.1 bp vs 167.3 bp, respectively, $p < 0.0001$, t-test; Fig. 3c, d). Examining broader regions surrounding TSS or CpG islands continued to reveal differences between expressed or unexpressed genes and between unmethylated or methylated islands in regions as far as 500 kb around these sites (Supplementary Fig. 10).

An analysis of cfDNA fragments adjacent to genes in KEGG and Hallmark gene sets revealed that cfDNA coverage reflected gene expression and CpG methylation[11] across all significant gene sets identified in white blood cells (WBCs; $p < 0.1$, gene set enrichment analysis; Fig. 3e and Supplementary Data 7). These included higher cfDNA coverage at regions of CpG islands and TSSs when methylation was increased and expression was decreased, and lower cfDNA coverage with decreased methylation and increased expression. For example, gene pathways not typically expressed in WBCs, including neuronal receptor-ligand interactions or olfactory receptor transduction, were typically methylated and more highly represented in cfDNA fragments at regions containing CpG islands or TSSs (Supplementary Fig. 11). In contrast, genes utilized in hematopoiesis, including in E2f transcription factor targets and blood cell metabolism genes were highly expressed, more frequently unmethylated, and represented at lower cfDNA levels at CpG or TSS regions of these genes (Fig. 3e). Overall, we found that cfDNA coverage was related to both CpG methylation and expression of nearby genes (Supplementary Fig. 12a), but that recurrent cfDNA fragment end enrichment at CpG sites was more closely related to methylation levels than gene expression (Supplementary Fig. 12b). A multivariate regression model evaluating DNA methylation (Fig. 3a, c), gene expression (Fig. 3b, d), nucleosome positioning (Supplementary Fig. 13), and the interaction of these terms revealed that each of these elements contributed independently to cfDNA coverage and fragment size (Fig. 3f and Supplementary Fig. 14). The relationship between methylation and coverage was qualitatively similar in more complex models that included additional terms for the interaction of DNA methylation and nucleosome positioning, and the three-way interaction of DNA methylation, gene expression, and nucleosome positioning (Supplementary Fig. 14). These results highlight that DNA methylation is a fundamental feature affecting cfDNA fragmentation.

To provide a direct and independent analysis of the effect of methylation or gene expression with cfDNA coverage, we assessed human cfDNA fragmentation coverage in the plasma of mice with implanted human tumors with or without a knock-in of the isocitrate dehydrogenase (IDH1) chromatin modifier with a mutation at R132 that was known to be activating through our previous work[27] and lead to widespread genome-wide methylation and expression changes[28–30]. Mice were injected with U87 glioblastoma cell line that was wild-type for IDH1 ($n = 3$) or isogenically altered to contain the R132H mutation ($n = 3$) and evaluated at 20–30 days after tumor implantation. After selecting human-derived cfDNA fragments from the mouse plasma, high coverage of human cfDNA was observed at areas of increased methylation, while low coverage of cfDNA was present at regions of increased expression (Fig. 3g) ($p < 0.053$, Monte Carlo simulation), consistent with our previous analyses. This well-controlled analysis provides a direct causal link between genome-wide changes in epigenetic features and cfDNA fragmentation.

As it has been widely reported that the overall size of cfDNA is smaller in patients with cancer compared to that of healthy individuals[31–38], we wondered whether genome-wide changes in DNA methylation and gene expression during tumorigenesis[39] may have an effect on cfDNA fragmentation in cancer patients. To unambiguously compare tumor-derived with WBC cfDNA, we examined changes in cfDNA fragment sizes of mutated tumor-derived and wild-type WBC-derived cfDNA using ultrasensitive NGS targeted sequencing from 98 patients with cancer[8,40]. We found an average shift of 3.9 bp in tumor-derived cfDNA of these patients that was similar to the observed cfDNA size differences at TSS regions of high and low expression and CpG sites in methylated versus unmethylated regions of healthy cfDNA (Fig. 3h). As we observed that tumors typically have an increased number of expressed genes and are hypomethylated compared to WBCs (Supplementary Fig. 15), consistent with previous studies[39,41,42], these results support the notion that changes in expression and methylation in cancer cells may in part be responsible for the overall smaller cfDNA fragments observed in patients with cancer.

Altered cfDNA fragmentation at regions of differential methylation to detect patients with cancer

To identify the impact of differences in CpG methylation on cfDNA fragment ends between healthy individuals and patients with cancer,

https://doi.org/10.1038/s41467-024-50850-8

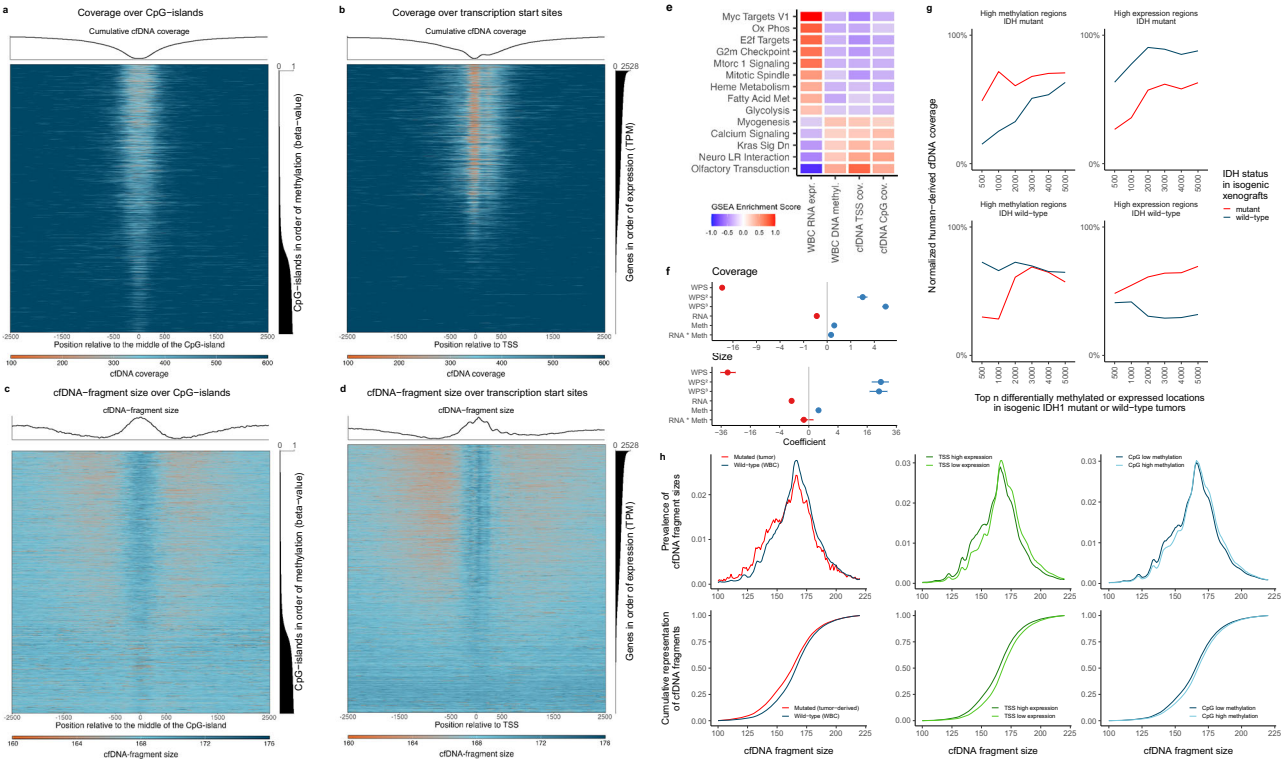

Fig. 3 | Effect of CpG methylation and gene expression on coverage and size of cfDNA fragments. a Sequence coverage in regions of CpG-islands, ordered by average methylation of the CpG-island. **b** Sequence coverage in regions of transcription start sites, ordered by the average expression of genes in myeloid cell lines. **c** Average cfDNA fragment sizes in regions of CpG-islands, ordered by average methylation of the CpG-island. **d** Average cfDNA fragment sizes in regions of transcription start sites, ordered by the average expression of genes in myeloid cell lines. **e** A correlation matrix (right) of significant gene-set enrichment analyses of gene-expression, DNA-methylation at the CpG-islands, cfDNA fragment coverage over the transcription start sites, and cfDNA coverage over the CpG-islands, all associated with the same genes. **f** (top) Average values of regression coefficients with 95% CI from a multivariate model evaluating the contribution of nucleosome positioning as portrayed by the windowed protection score (WPS, WPS^2, and WPS^3), gene expression scaled to have unit standard deviation (RNA), DNA methylation (Meth = 1 if beta ≥0.5 and 0 otherwise), and their interaction on cfDNA coverage. All coefficients have error-bars, although the confidence intervals may be too small to be visible. **f** (bottom) Average values of regression coefficients with 95% CI from a multivariate model evaluating the contribution of nucleosome positioning as portrayed by the windowed protection score (WPS, WPS^2, and WPS^3), gene expression scaled to have unit standard deviation (RNA), DNA methylation (Meth = 1 if beta ≥0.5 and 0 otherwise), and their interaction on cfDNA fragment size. All coefficients have error-bars, although the confidence intervals may be too small to be visible. **g** Data from human isogenic xenografts (IDH1 R132H mutant compared to IDH1 wild-type human glioblastoma cell lines) showing increased coverage of human cfDNA in areas of increased methylation and decreased coverage in areas of increased gene expression over the top n regions of differential methylation or gene expression. **h** cfDNA fragment size distributions (top) and cumulative representations (bottom) for fragments with a mutation (tumor-derived) compared to wild-type fragments (mainly white blood cells), fragments from regions of high expression transcription start sites compared to fragments from regions of low expression transcription start sites, and fragments from regions of methylated CpG-islands compared to fragments from regions of unmethylated CpG-islands. The cfDNA fragment profiles were obtained from targeted sequencing analyses of patients with colorectal, breast, lung, or ovarian cancer as well as individuals without cancer[40]. TSS transcription start site, IDH isocitrate dehydrogenase, TPM transcripts per million.

we evaluated regions previously identified by comparison of reduced representation bisulfite sequencing (RRBS) data from laser capture microdissected pancreatic ductal adenocarcinoma and normal pancreatic tissues, and where these regions were confirmed in cfDNA using methyl-DNA immunoprecipitation (MeDIP)[43]. We then assessed the fragment end representation at CG and CCG sites through low-coverage whole-genome cfDNA analyses of patients with pancreatic ($n = 34$), colorectal ($n = 27$), ovarian ($n = 28$), lung ($n = 39$), or breast cancer ($n = 54$) as well as from individuals without cancer ($n = 244$)[8]. In regions with increased CpG methylation in non-cancer tissues, we observed a preferential decrease in cfDNA fragments ending with N|CCG in individuals without cancer compared to the abundance of these fragments in patients with pancreatic and other cancers (Fig. 4a). In contrast, in regions with increased methylation in pancreatic cancer, we saw an increase in cfDNA fragments ending with CG in patients with cancer compared to levels in individuals without cancer (Fig. 4a). In all cases the strongest signal was observed in the patients with pancreatic cancer, suggesting that the use of tumor-specific sites of methylation resulted in improved performance in this tumor type. Incorporation of the distribution of fragment end positions at these CG and CCG sites in a gradient boosted tree machine learning model successfully distinguished individuals with pancreatic cancer compared to those without (cross-validated AUC = 0.87). Combining this approach with genome-wide fragmentation analyses (DELFI)[8] that incorporate fragment coverage and size improved the sensitivity of the combined method (AUC = 0.93, 95% CI = 0.88–0.97; Fig. 4b). These observations suggested that DNA methylation information may enhance detection of individuals with cancer using genome-wide cfDNA fragmentation.

Discussion

In this study, we have shown that specific DNA sequences at regions of CpG methylation have a profound impact on cfDNA fragmentation genome-wide. Preferred breakpoints of cfDNA fragments were associated with specific sequence motifs, both before and within the cfDNA fragment, including A/T|CC and A/T|CG. Recurrent cfDNA sequences with CG fragment ends were enriched at sites of methylation, increasing at N|CG and decreasing at competing N|CCG sites in a manner that was dependent on the level of methylation. Structural

https://doi.org/10.1038/s41467-024-50850-8

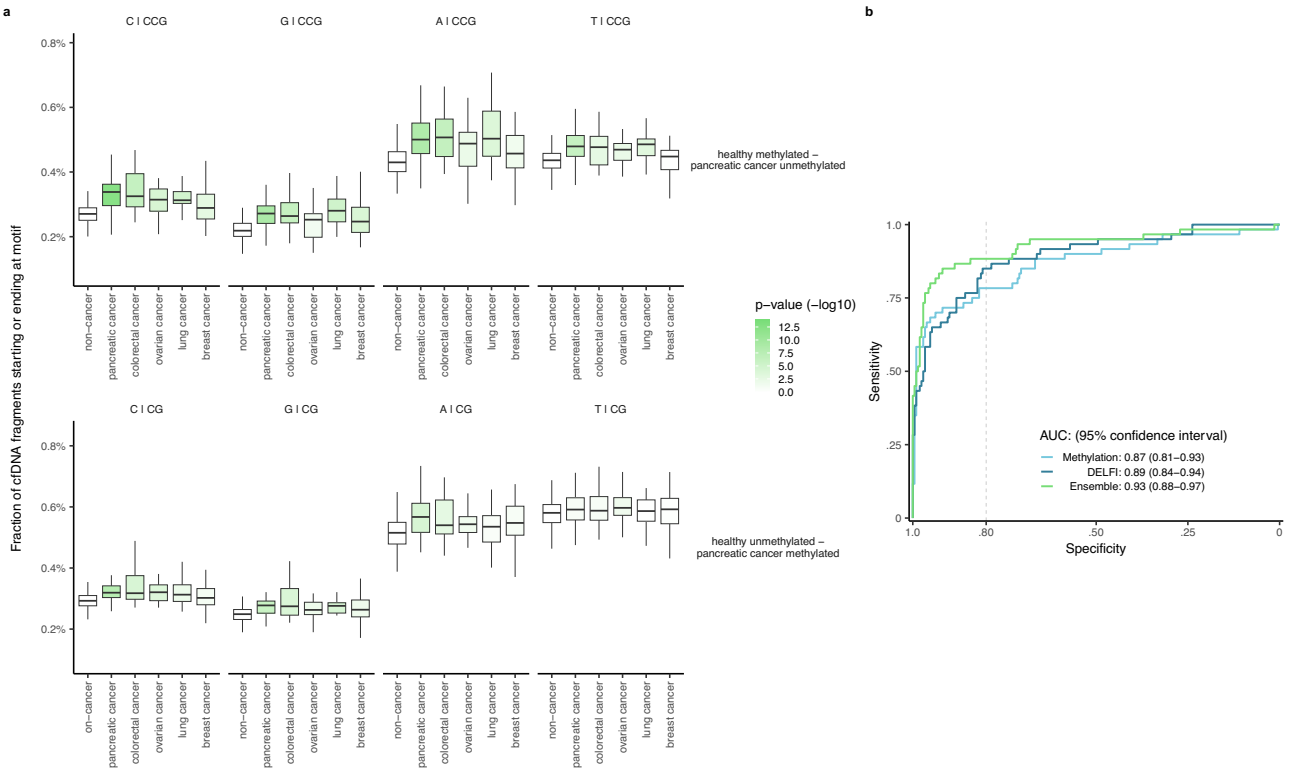

Fig. 4 | Comparison of cfDNA fragment end motifs in regions of differential methylation in individuals with and without pancreatic cancer. a Aggregated ratios of cfDNA fragments starting or ending at specific motifs containing CpG's which showed differential methylation between cfDNA of healthy individuals and pancreatic cancer tissues. The largest increase in signal was found in pancreatic cancer patients ($n = 34$) (p values across different fragment end contexts for patients with pancreatic ranged from 3.3×10^{-13} for pancreatic cancer (C|CCG) to 0.85 for breast cancer (T|CCG)), while cfDNA from patients with other cancers (colorectal $n = 27$, ovarian $n = 28$, lung $n = 39$, and breast cancer $n = 54$) showed intermediate signals compared to individuals without cancer ($n = 244$; p values across different fragment end contexts ranged from 1.4×10^{-8} for pancreatic cancer (C|CG) to 0.76 for lung cancer (T|CG)). P-values were calculated using a two-sample t-test, without correction for multiple testing. The middle hinge in the boxplots corresponds to the median, while the lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * interquartile range from the hinge. **b** The predictive value of this signal for detecting pancreatic cancer is shown as receiver operating characteristic curve in comparison with the DELFI, while combining the DELFI approach with the methylation-based signal in an ensemble-model showed the best prediction. AUC area under the curve.

analyses of DNA bound to nucleosomes showed that CG sequences were typically located close to the histone H1 linker. These observations, together with previous molecular dynamic simulations[44], suggest that methylation of CG sequences may provide a more stable interaction between the methylated DNA and the H1 linker, thereby protecting nucleosome-bound cfDNA fragments from degradation. Our observations are in part consistent with recent studies that have shown methylation associated differences in cfDNA fragmentation around CG sites[18]. However, due to the high prevalence of fragments starting with CC, we found that our newly identified methylation-associated changes at N|CCG-motifs were stronger than changes at CG-motifs alone.

Methylated CpGs affected not only fragment end positions, but also resulted in a higher amounts of circulating cfDNA at these regions compared to unmethylated CpG sites. cfDNA fragmentation was inversely affected at TSS, affecting individual genes as well as gene pathways. cfDNA fragment sizes were altered by both methylation and expression changes, and could be observed nearby CG and TSS locations, as well as at distances hundreds of thousands of bases away. Our observations of cfDNA fragmentation at regions of CpG methylation and decreased gene expression were independently validated in mouse models implanted with human tumors with or without isogenic mutant chromatin modifier IDH1, providing a direct link between epigenetic changes and cfDNA coverage.

Although our studies focused on mononucleosomal cfDNA, other studies have shown an enrichment of methylation marks in multinucleosomal cfDNA fragments in cfDNA of pregnant women[45,46]. Additionally, our analysis of gene expression used myeloid cells as a reference because these are thought to give rise to the majority of cfDNA in the circulation, but analyses of other blood cell types and cancers, as well as analyses of cfDNA fragment level-data of fragmentation and methylation, could further improve the connection between gene expression and cfDNA fragmentation.

The effect of CpG methylation and gene expression on cfDNA size, coupled with an overall increase in gene expression and decrease in methylation in human cancers observed in this study, suggest a mechanism for the global reduction of cfDNA fragment lengths observed in cancer patients. Others have suggested that changes in intracellular and plasma nucleases affect cfDNA fragmentation and may lead to specific end-motifs in cancer patients[20,47,48]. However, our observations appear to be independent of these proposed mechanisms as we showed that cfDNA from healthy individuals, largely derived from the same white blood cells with presumably the same exposure to nucleases, display cfDNA fragment length differences that were associated with sites of CpG methylation and gene expression genome-wide.

The incorporation of cfDNA fragment end features at CpG sites into a cross-validated machine-learning model suggested an approach that could be used to detect cancer independently of other cfDNA characteristics. This approach appeared complementary to the DELFI cfDNA fragmentation analyses and together resulted in a method that may improve performance. Improved analyses of genome-wide

methylation and expression differences in isolated tumor cell populations may provide avenues for better understanding cfDNA fragmentation and new methods for assessing these in the future. Concurrent sequence and methylation analyses using the same cfDNA molecules may improve these studies, although the current approach benefits by avoiding biases that may be introduced through harsh conditions of bisulfite sequencing or preferential selection of multiple adjacent methylated sites using precipitation methods[11]. Extension of these analyses to larger cohorts, including with screening populations, and other cancer types will be necessary to validate these initial observations for potential clinical applications. Integration of methylation and expression changes with other genome-wide epigenetic marks[49] in well-controlled experimental systems may provide complementary insights into the origins and mechanisms of cfDNA fragmentation.

Methods

Study populations

The data analyzed in this study were obtained from previous reports where patients provided written informed consent and their inclusion in the respective studies was performed according to the Declaration of Helsinki. These samples were originally obtained from two screening clinical trial cohorts for colorectal cancer in Denmark (Endoscopy III) and the Netherlands (COCOS, Netherlands Trial Register ID NTR182946) or as previously described[8,50]. The protocol for the Endoscopy III Project was approved by the Regional Ethics Committee and the Danish Data Protection Agency, and for the COCOS trial, ethical approval was obtained from the Dutch Health Council. For analyzing motif frequencies, recurrent ends, and the relationship with gene expression and methylation, we used low-coverage whole-genome sequencing (WGS) of cfDNA (1-2x) from 787 individuals without cancer (female $n = 442$; male $n = 345$) as well as 182 individuals with cancer (female $n = 130$; male $n = 52$)[8,50]. cfDNA methylation from 97 individuals without cancer was previously analyzed using Illumina's Infinium methylationEPIC array[11] and made available through NCBI's Gene Expression Omnibus (GEO) database (dataset identifier GSE122126). The cells that contribute to cfDNA in individuals without cancer, was used and validated previously[9,11], showing most cfDNA originates from myeloid-derived cells. Average gene expression summarized as transcripts per million bases (TPM) was obtained from 6 myeloid cell lines, as previously published[51].

Processing of cfDNA samples

Whole-genome libraries of cancer patients and cancer-free individuals were sequenced using 100 bp paired-end runs (200 cycles) on the Illumina HiSeq2500 platform at 1–2x coverage per genome. Prior to alignment, adapter sequences were filtered from reads using the fastp software[52]. Sequence reads were aligned against the hg19 human reference genome using Bowtie2[53] and duplicate reads were removed using Sambamba[54]. Only reads with a mapq score of at least 30 or greater were retained. Post-alignment, each aligned pair was converted to a genomic interval representing the sequenced DNA fragment using bedtools[55].

Frequency of motifs around the ends of cfDNA fragments

The expected frequency of 3 bp motifs in the human genome was calculated by counting the occurrence of each 3 bp motif in the human genome (hg19). For computing the empirical frequency of 3 bp motifs at the ends of sheared (sonicated) fragments, we used published sequencing data from 9 lymphoblastoid cell lines[56]. The genomic DNA from these lymphoblastoid cell lines was fragmented through sonication with a Covaris M220 Focused Ultrasonicator. For this analysis, the data were analyzed as described above for processing cfDNA samples. For the 10 lymphoblastoid cell lines and the 543 low coverage WGS of individuals without cancer, we counted the number of 3 bp motifs at the start of the fragment and the reverse complement of the 3 bp motif at the end of the fragment. The 3 bp motif contains 1 base outside the fragment, followed by the first 2 bases of the fragment. Using these absolute numbers, we calculated the fraction of each of the 64 3 bp motifs.

To quantify the preference for cfDNA fragments to end at a specific genomic location and adjust for local copy number polymorphisms in healthy individuals, we calculated the ratio of cfDNA fragments ending at this location (recurrent fragment ends) divided by the number of cfDNA fragments having a 1 bp or more overlap within ±50 bp of this position (neighboring fragments). This ratio of preferred fragment ends to neighboring fragments was computed by aggregating cfDNA fragments across all 543 individuals without cancer. We repeated this calculation for every evaluable position in the hg19 reference genome. "Preferred ends" were defined as genomic positions where at least 5% of cfDNA fragments started at that position and where 200 or more high-quality cfDNA fragments (mapq > 30) overlapped the area around that position.

X-ray crystal structure of nucleosomes

To identify structures of the nucleosome bound to DNA we searched PDB (https://www.rcsb.org/) for the term "Nucleosome" and filtered the results for those structures derived from x-ray diffraction or cryo-EM leading to 427 entries. We downloaded the DNA sequences from these structures (648) and filtered for sequences that were at least 167 bp in length. This identified 80 unique sequences from 51 PDB entries. These entries were visually inspected and those with less than 167 bases resolved or where the interaction with the H1 linker was disrupted by another DNA binding protein were removed. This left 17 structures (Supplementary Data 3). Motifs were considered well positioned if they were within 5 Å of the H1 linker or if the bases 167 bp away on the same strand were within 5 Å of the H1 linker.

Connecting cfDNA fragment patterns to CpG methylation

In order to discover whether fragmentation patterns in cfDNA from individuals without cancer were influenced by methylation, we analyzed raw data from Illumina's Infinium methylationEPIC array from 8 different cfDNA experiments, with 4 biologically different cohorts (young men, old men, young women, old women)[11]. Standard pipelines were used to process the Infinium arrays as implemented in the minfi R-package (version 1.48.0)[57]. The CpG probes were annotated for their genomic location and associated with gene names using the IlluminaHumanMethylationEPICmanifest R-package (version 0.3.0). A numeric score (beta-value) ranging between 0 (unmethylated) and 1 (methylated) was obtained at each CpG and averaged across all samples. CpG sites were labeled as unmethylated if the mean beta-value was <0.3 and methylated if the mean beta-value was > 0.7. For 543 cfDNA plasma samples of individuals without cancer (Supplementary Data 1), we recorded the position of Infinium CpG sites within the cfDNA fragments using a 1-base index. CpGs were grouped in bins of 0.1 according to their mean beta-value. We counted the number of fragments starting or ending at the CpG sites or at neighboring positions for CCG motifs and scaled this frequency by the number of fragments having any overlap within a 50 bp window of the start- or end-position. We further categorized fragments according to their 3 bp and 4 bp end motif and whether the cfDNA fragment was located in a CpG-island, shore, shelf, or open sea. We repeated this analysis using processed methylation data from whole-genome bisulfite sequencing of cfDNA (23 individuals, men and women), resulting in 23.6 million somatic CpG sites, each covered by at least 750 reads (after combining the 23 samples)[12]. A beta value was calculated for these samples by calculating the ratio of methylated reads over total coverage at each specific CpG site. To assess whether these analyses were sensitive to the initial choice of a beta value cutoff for determining methylation status at CpG sites, we repeated these analyses

with cutpoints of 0.1 and 0.9, 0.2 and 0.8, 0.3 and 0.7, 0.4 and 0.6, and 0.5 and 0.5 (Supplementary Fig. 2).

Differences in chromosome X cfDNA fragmentation between women and men

To validate the discovered differences in cfDNA fragmentation around methylated and unmethylated CpGs, we analyzed the CpGs located on the X chromosome of men and women. For each individual, we calculated the number of fragments starting at a CpG or a CCG, divided by the number of fragments having any overlap with a 50 bp window from this motif. For each individual, these locations were grouped by preceding base and by the CpG functional location (CpG-islands, shores, shelves and open sea) with the R-package 'annotatr' (version 1.28.0)[58]. The results were summarized by individual and further stratified by sex.

IDH1 isogenic xenograft model

All animal work described in this study was approved by the Johns Hopkins Animal Care and Use Committee. Three NU/J 6-8 week old female mice were injected in the flank with a human IDH1 wild-type glioblastoma cell line (U-87, HTB-14), while three other NU/J 6-8 week old female mice were injected in the flank with human IDH1 R132H mutant glioblastoma cell line (U-87, HTB-14, transfected with Crispr-Cas9). All mice were housed in ventilated cages with sterile wood shavings for bedding. Up to five mice were housed per cage with access to food and water ad libitum. Room temperature was maintained at 22 to 24 °C with a 12-h light/dark cycle. Tumors were grown for 20 – 30 days. When the tumors were no larger than 20 mm in any dimension, the mice were anesthetized with a sublethal dose of ketamine-xylazine, the blood was obtained and the mice were sacrificed through cervical dislocation. The blood was further diluted in EDTA tubes and after spinning, plasma was extracted. Libraries were created from cfDNA as previously described[59]. The sequencing data was processed with Xengsort (version 1.5.0), to separate human (hg19) and murine (mm38) cfDNA reads. Further processing was performed as described above (see *Processing of cfDNA samples*). After collecting the blood, tumor tissues were removed, and samples were processed for DNA and RNA extraction, and used for methylation (Illumina Infinium EPIC Methylation Array) and expression (RNA-seq) analyses, respectively. Methylation data were processed using standard pipelines for the Infinium arrays: R-package minfi (version 1.48.0)[57]. Differentially methylated regions were identified using R-package DMRcate (version 2.16.1)[60]. RNA-seq data was aligned to hg19 using the star aligner (version 2.7.4)[61]. FPKM gene expression values were constructed for 20,344 Ensembl gene identifiers. Differentially expressed genes were identified using DESeq2 (version 1.42.0) according to standard pipeline[62].

cfDNA sequence coverage, fragment sizes, and nucleosomes at CpG-islands and transcription start sites

To summarize cfDNA fragment lengths at one CpG island, we counted the average length of mononucleosomal (≥100 bp and ≤220 bp) cfDNA fragments around the CpG island across all 543 non-cancer samples. By convention, we referred to this as position 0. We performed this summarization step in 10 bp increments from the CpG island ranging from −500,000 bp to +500,000 bp, and repeated this procedure for each CpG island. Mean fragment lengths at TSSs were summarized in a similar manner with position 0 denoting the TSS. For cfDNA coverage, we used a similar approach, but used cumulative coverage across 543 non-cancer samples. For calculating nucleosome positions, we calculated for each position in the genome the Windowed Protection Score (WPS), as described previously[9], using cfDNA fragments pooled across all 543 non-cancer samples. Using the previously described methylation data[11] and gene-expression data from myeloid cell lines[51], we ordered the regions by decreasing levels of methylation and increasing amount of expression to visualize patterns that were associated with CpG-island methylation and gene expression.

Gene set enrichment analyses

Gene set enrichment analysis (GSEA)[63] was performed with the Hallmark[64] and KEGG[65] gene sets acquired from the Molecular Signatures Database[66] following previously described GSEA methodology[63]. Briefly, a heuristic parameter enrichment score was calculated for a subset of genes given a reference gene set based on the location of the gene set associated genes within the ranked list. Then the ranks were permuted randomly and the enrichment score was calculated from the permuted ranks to generate a null distribution that was used to determine statistical significance of the empirical enrichment score.

Transcripts were ranked by RNA expression, methylation, and total coverage. For RNA expression, transcripts were ranked by the mean TPM from 6 myeloid cell lines[51]. For methylation, the mean beta value across 97 samples was calculated for each CpG in a transcript and then averaged to obtain a transcript-level summary of methylation. Transcripts were ranked by this measure of methylation. For cfDNA coverage, we used metrics identical to those used to order genes for visualization of total cfDNA coverage at CpG islands and TSSs, including total coverage in the interval ±250 bp from the CpG and total coverage in the interval ±250 bp from the TSS. In summary, four sets of ranked transcripts were investigated by GSEA: mean RNA expression in myeloid cell lines, mean beta value for CpGs in blood, total TSS coverage in cfDNA, and total coverage of the CpG sites in cfDNA. 10,000 permutations were run for each set of ranks, leading to a minimum unadjusted p-value of 1e-4. All gene sets that were moderately significant across all of the analyzed sets of ranks in any direction (unadjusted $p < 0.1$) were selected for inclusion in the heatmap showing enrichment scores by gene set.

Multivariate model

Generalized linear models were used to evaluate the relationship between the aggregated mean cfDNA fragment size and total coverage at the transcript level with RNA expression, WPS, and methylation. For methylation, we calculated the mean beta-value at each CpG-island across 97 blood samples processed on the Infinium array (see *Study populations*). CpG-islands were mapped to transcripts by their proximity to TSSs using the R-package annotatr (version 1.28.0). A transcript was considered methylated if the mean beta value was 0.5 or higher and unmethylated otherwise. The mean RNA expression (mean TPM) across 6 myeloid cell lines were transformed as log10(mean TPM + 1) and then centered and scaled by the overall mean and standard deviation across all transcripts, respectively. WPS was summarized for each transcript in the interval +1 to +10 bases from the TSS and centered and scaled. Total cfDNA coverage across 543 non-cancers was calculated at each base in the interval −10bp to −1bp from the TSS and averaged, while mean fragment sizes were calculated in the interval from −1480bp to −1471bp from the TSS. The intervals for summarizing cfDNA coverage, fragment size, and WPS were evaluated for all 10 bp genomic intervals within 2500 bp from the TSS. The interval that yielded measurements with maximum absolute correlation to RNA expression was selected for the regression analyses. With these quantitative summaries as described above, the expected normalized coverage, DV, for transcript i is given by

$$E(DV_i|WPS_i,RNA_i,Meth_i) = \beta_0 + \beta_1 WPS_i + \beta_2 WPS_i^2 + \beta_3 WPS_i^3 + \beta_4 RNA_i + \beta_5 Meth_i + \beta_6 RNA_i \times Meth_i.$$

The expected fragment length was modeled in a similar fashion. Coefficients from these models were estimated using a generalized linear model with identity link function in R (version 4.3.2). Using analysis of variance (ANOVA), we assessed whether RNA expression

helps explain variation in coverage after adjusting for methylation and WPS by testing both the main effect for RNA expression and its interaction with methylation. We performed a similar ANOVA to evaluate whether methylation explained variation in coverage after adjusting for the effects of RNA and WPS on coverage. Forest plots were generated for each model to visualize estimated model coefficients with 95% confidence intervals using sjPlot (version 2.8.15).

Monte Carlo simulation on human cfDNA coverage in xenograft models

Coverage was calculated for the top 500, 1000, 2000, 3000, 4000, and 5000 most differentially methylated CpG-islands or most differentially expressed genes for each of the six xenografts (3 IDH1 R132H mutant xenografts and 3 IDH1 wild-type xenografts). These coverages were normalized for the total size of these regions. Comparing IDH1 mutant to wild-type, we determined whether the direction of the difference in normalized coverage agreed with our a priori expectation that we would observe higher coverage in methylated regions and lower coverage in expressed regions for each of the four possible comparisons (high methylation regions in IDH mutant, high methylation in IDH wild-type, high expression in IDH mutant, high expression in IDH wild-type). To evaluate how likely we would observe the empirical agreement under the null hypothesis that there is no difference in cfDNA coverage in mice between IDH1 mutant and wild-type tumors, we permuted the mutant and wild-type labels and evaluated the agreement as previously described. This was repeated for 10,000 iterations. We repeated this process for each of the 19 possible permutations of the sample labels, deriving a distribution of agreement under the null. The *p*-value was computed as the proportion of permutations where the agreement was as high or higher than the empirical agreement obtained from the non-permuted class labels. These analyses were repeated for each of the six region or gene list sizes indicated above for a total of 24 comparisons.

Differentially methylated CpG-based tumor-specific cfDNA methylation patterns

Using publicly available data, we evaluated differentially methylated CpGs from individuals with pancreatic cancer ($n = 60$) and without cancer ($n = 210$) from the previously published Cristiano et al. study[8]. A large cohort of differentially methylated regions was previously published for pancreatic cancers[43]. Using these differentially methylated CpGs we defined subgroups based on the direction of differential methylation (non-cancer methylated vs. pancreatic cancer unmethylated; non-cancer unmethylated vs. pancreatic cancer methylated) and based on the 3 bp and 4 bp motifs. In total, we extracted 16 different features representing frequencies of fragments ending with A|CG, C|CG, G|CG, T|CG, A|CCG, C|CCG, G|CCG, or T|CCG, at positions that were either methylated in healthy individuals and unmethylated in patients with pancreatic cancer or unmethylated in healthy individuals and methylated in patients with pancreatic cancer. For each sample, we calculated the ratio of aggregated cfDNA fragments starting or ending at these motifs, divided by the aggregated number of fragments overlapping a 101 bp window around the motif.

In addition, we computed cfDNA fragmentation features for each of these samples as previously described[8]. Briefly, we calculated GC-corrected fragment coverage for each of 473 non-contiguous 5 mb bins genome-wide, and 39 chromosomal arm-level z-scores for aneuploidy as compared to a healthy reference[50]. We cross-validated (10-fold) a gradient-boosted tree incorporating the 16 methylation features, generating a score for each sample in the held-out folds. This gradient boosted tree was trained using the R packages *caret* and *gbm* with default parameters. Using the same folds, we cross-validated a penalized logistic regression with L_1 norm penalty (LASSO with alpha=1) using the coverage and zscore features (DELFI model). This model was evaluated using the R packages *caret* and *glmnet*, and during training we performed a PCA for dimensionality reduction on the coverage features and retained only the PCs needed to explain 90% of variance. For the DELFI model, in each training fold we included the remaining cancers (not pancreatic) ($n = 128$) from the Cristiano et al. data[8] as additional training data, but performance of the DELFI score was evaluated in the held-out folds only on non-cancer and pancreatic cancer individuals.

In order to assess the combined performance of both DELFI and methylation motif features, we implemented a nested cross-validation (CV) such that the inner CV loop trains the DELFI and methylation classifiers as described above, and the outer CV loop trains an penalized logistic regression (LASSO with alpha = 1) using the DELFI and methylation scores as features. The outer cross-validation loop for the ensemble was trained with 10-fold cross validation, while the inner cross-validation loop used to train the component methylation and DELFI models used 5-fold cross-validation. We assessed performance of the methylation, DELFI and methylation-DELFI ensemble using ROC curves with 95% confidence intervals for the area under the ROC curve computed using DeLong's method.

Statistical methods

All *t*-tests were Welch two sample *t*-tests unless otherwise indicated. Statistical significance of multivariable regression models were assessed by ANOVA. Confidence intervals for fold-changes were estimated by bootstrap. All analyses were performed using R, version 4.3.2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequence data and clinical variables from the samples analyzed in this study are available in the database of Genotypes and Phenotypes (dbGaP) under study ID 34536 and in the European Genome-Phenome Archive (EGA), under accession code EGAS00001005340. Methylation data on cfDNA samples from Moss et al. are available in the NCBI Gene Expression Omnibus (GEO) database repository with the dataset identifier GSE122126[11]. Methylation data on cfDNA samples from Loyfer et al. are available in the NCBI Gene Expression Omnibus (GEO) database repository under dataset identifier GSE186458[12]. FPKM gene expression values measured for 20,344 Ensembl gene identifiers in 44 human cell lines and 32 primary tissues by the Human Protein Atlas were downloaded from https://v13.proteinatlas.org/download/rna.csv.zip[51]. The data from the TCGA and the Genotype-Tissue Expression Project (GTEx) are available from their respective website. The data from GTEx were also submitted to dbGaP under study accession phs000424.v9.p2.

Code availability

Scripts for reproducing tables and figures in the manuscript are available in the following GitHub repository (https://github.com/cancer-genomics/cfepigenetics) under the GNU GENERAL PUBLIC LICENSE Version 3.

References

1. Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311**, 532–537 (1984).
2. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
3. Ceppellini, R., Polli, E. & Celada, F. A DNA-reacting factor in serum of a patient with lupus erythematosus diffusus. *Proc. Soc. Exp. Biol. Med.* **96**, 572–574 (1957).
4. Miescher, P. & Strässle, R. New serological methods for the detection of the L. E. Factor. *Vox Sang.* **2**, 283–287 (1957).

5. Seligmann, M. [Demonstration in the blood of patients with disseminated lupus erythematosus a substance determining a precipitation reaction with desoxyribonucleic acid]. *Comptes Rendus Hebd. Des. Seances De. L'academie Des. Sci.* **245**, 243–245 (1957).
6. Robbins, W. C., Holman, H. R., Deicher, H. & Kunkel, H. G. Complement fixation with cell nuclei and DNA in lupus erythematosus. *Proc. Soc. Exp. Biol. Med.* **96**, 575–579 (1957).
7. Barra, G. B. et al. EDTA-mediated inhibition of DNases protects circulating cell-free DNA from ex vivo degradation in blood samples. *Clin. Biochem.* **48**, 976–981 (2015).
8. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
9. Snyder, M. W. et al. Comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
10. Foda, Z. H. et al. Detecting liver cancer using cell-free DNA fragmentomes. *Cancer Discov.* **13**, 616–631 (2023).
11. Moss, J. et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).
12. Loyfer, N. et al. A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
13. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
14. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* **41**, 2918–2931 (2013).
15. Keshet, I., Yisraeli, J. & Cedar, H. Effect of regional DNA methylation on gene expression. *Proc. Natl Acad. Sci. USA* **82**, 2560–2564 (1985).
16. Ulz, P. et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* **48**, 1273–1278 (2016).
17. Esfahani, M. S. et al. Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597 (2022).
18. Zhou, Q. et al. Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc. Natl Acad. Sci. USA* **119**, e2209852119 (2022).
19. Chan, K. C. A. et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc. Natl Acad. Sci. USA* **113**, E8159–E8168 (2016).
20. Serpas, L. et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl Acad. Sci. USA* **116**, 641–649 (2019).
21. Jin, C. 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).
22. Trifonov, E. N. Cracking the chromatin code: precise rule of nucleosome positioning. *Phys. Life Rev.* **8**, 39–50 (2011).
23. Norris, D. P., Brockdorff, N. & Rastan, S. Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. *Mamm. Genome* **1**, 78–83 (1991).
24. Tribioli, C. et al. Methylation and sequence analysis around Eagi sites: identification of 28 new CpG islands in XQ24-XQ28. *Nucleic Acids Res.* **20**, 727–733 (1992).
25. Duncan, C. G. et al. Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. *Sci. Rep.-UK* **8**, 10138 (2018).
26. Zhu, F. et al. The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76–81 (2018).
27. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
28. Duncan, C. G. et al. A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. *Genome Res.* **22**, 2339–2355 (2012).
29. Wei, S. et al. Heterozygous IDH1R132H/WT created by "single base editing" inhibits human astroglial cell growth by downregulating YAP. *Oncogene* **37**, 5160–5174 (2018).
30. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
31. Jiang, P. et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl Acad. Sci. USA* **112**, E1317–E1325 (2015).
32. Giacona, M. B. et al. Cell-free DNA in human blood plasma. *Pancreas* **17**, 89–97 (1998).
33. Mouliere, F. et al. High fragmentation characterizes tumour-derived circulating DNA. *PLoS ONE* **6**, e23418 (2011).
34. Lapin, M. et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. *J. Transl. Med.* **16**, 300 (2018).
35. Underhill, H. R. Leveraging the fragment length of circulating tumour dna to improve molecular profiling of solid tumour malignancies with next-generation sequencing: a pathway to advanced non-invasive diagnostics in precision oncology? *Mol. Diagn. Ther.* **25**, 389–408 (2021).
36. Mouliere, F. et al. Circulating cell-free DNA from colorectal cancer patients may reveal high KRAS or BRAF mutation load. *Transl. Oncol.* **6**, 319–IN8 (2013).
37. Thierry, A. R. Circulating DNA fragmentomics and cancer screening. *Cell Genom.* **3**, 100242 (2023).
38. Underhill, H. R. et al. Fragment length of circulating tumor DNA. *PLoS Genet* **12**, e1006162 (2016).
39. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415–428 (2002).
40. Phallen, J. et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (2017).
41. Baylin, S. B. et al. Abnormal patterns of DNA methylation in human neoplasia: potential consequences for tumor progression. *Cancer Cells Cold Spring Harb. N. Y. 1989* **3**, 383–390 (1991).
42. Gama-Sosa, M. A. et al. The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **11**, 6883–6894 (1983).
43. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
44. Li, S., Peng, Y., Landsman, D. & Panchenko, A. R. DNA methylation cues in nucleosome geometry, stability and unwrapping. *Nucleic Acids Res.* **50**, 1864–1874 (2022).
45. An, Y. et al. DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. *Nat. Commun.* **14**, 287 (2023).
46. Jensen, T. J. et al. Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biol.* **16**, 78 (2015).
47. Han, D. S. C. 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).
48. Jiang, P. et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
49. Annapragada, A. et al. Genome-wide repeat landscapes in cancer and cell-free DNA. *Sci. Transl. Med.* **16**, eadj9283 (2024).
50. Mathios, D. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).
51. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
52. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
54. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Papp, E. et al. Integrated genomic, epigenomic, and expression analyses of ovarian cancer cell lines. *Cell Rep.* **25**, 2617–2633 (2018).
57. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
58. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
59. Mathios, D. et al. Early detection of lung cancer using cfDNA fragmentation. *J. Clin. Oncol.* **39**, 8519–8519 (2021).
60. Peters, T. J. et al. De novo identification of differentially methylated regions in the human genome. *Epigenet. Chromatin* **8**, 6 (2015).
61. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
62. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
63. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
64. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
65. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
66. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
67. Adhireksan, Z. et al. Engineering nucleosomes for generating diverse chromatin assemblies. *Nucleic Acids Res.* **49**, gkab070 (2021).

Acknowledgements

We thank members of our laboratories for critical review of the manuscript. This work was supported in part by the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, SU2C in-Time Lung Cancer Interception Dream Team Grant, Stand Up to Cancer-Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Gray Foundation, the Commonwealth Foundation, the Mark Foundation for Cancer Research, the Cole Foundation, a research grant from Delfi Diagnostics, and US National Institutes of Health grants CA121113, CA006973, CA233259, CA062924, CA271896, and 1T32GM136577. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The results shown here are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

Author contributions

M.N., R.B.S., and V.E.V. designed and planned the study. M.N., D.M., A.V.A., S.K., Z.H.F., J.E.M., S.C., and C.C. performed experiments and developed experimental protocols or bioinformatic methods. M.N., D.M., A.V.A., S.K., Z.H.F., C.C., R.B.S. and V.E.V. analyzed and interpreted data. M.N., R.B.S., and V.E.V. were involved in the preparation of the manuscript, and M.N., D.M., A.V.A., S.K., Z.H.F., J.E.M., S.C., C.C., D.C.B., N.N., V.A., L.F., H.E., S.B., J.P., R.B.S., and V.E.V. read and approved the final version.

Competing interests

M.N., D.M., D.C.B., S.C., J.P., V.A., R.B.S., and V.E.V. are inventors on patent applications submitted by Johns Hopkins University related to cell-free DNA for cancer detection, including PCT/US24/34360. S.C., J.P., V.A., and R.B.S. are founders of Delfi Diagnostics, and V.A. and R.B.S. are consultants for this organization. V.E.V. is a founder of Delfi Diagnostics, serves on its Board of Directors, and owns Delfi Diagnostics stock, which is subject to certain restrictions under university policy. Additionally, Johns Hopkins University owns equity in Delfi Diagnostics. V.E.V. divested his equity in Personal Genome Diagnostics (PGDx) to LabCorp in February 2022. V.E.V. is an inventor on patent applications submitted by Johns Hopkins University related to cancer genomic analyses and cell-free DNA for cancer detection that have been licensed to one or more entities, including Delfi Diagnostics, LabCorp, Qiagen, Sysmex, Agios, Genzyme, Esoterix, Ventana and ManaT Bio. Under the terms of these license agreements, the University and inventors are entitled to fees and royalty distributions. V.E.V. is an advisor to Viron Therapeutics and Epitope. These arrangements have been reviewed and approved by the Johns Hopkins University in accordance with its conflict-of-interest policies. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-50850-8.

Correspondence and requests for materials should be addressed to Robert B. Scharpf or Victor E. Velculescu.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.