A pan-cancer compendium of 1,074 plasma cell-free DNA methylomes and fragmentomes

Yong Zeng[1, $, #], Dor D. Abelman[1, 2, $], Althaf Singhawansa[1], Nicholas Cheng[3], Emma Bell[1], Wenbin Ye[1], Sasha Main[1, 2], Ping Luo[1], Samantha L. Wilson[4], Eric Y. Stutheit-Zhao[1], Derek Wong[1], Nadia Znassi[1], Suluxan Mohanraj[1], Philip Awadalla[3], Benjamin H. Lok[1, 2], Michael M. Hoffman[1, 2, 5], Raymond H. Kim[1, 2, 6], Gelareh Zadeh[1, 7], Daniel De Carvalho[1, 2], Scott V. Bratman[1, 2], Mathieu Lupien[1, 2, 3, #], Trevor J. Pugh[1, 2, 3, #], Housheng Hansen He[1, 2, #]

[1] Princess Margaret Cancer Centre, University Health Network, Toronto, Canada

[2] Department of Medical Biophysics, University of Toronto, Toronto, Canada

[3] Ontario Institute for Cancer Research, Toronto, Ontario, Canada

[4] Department of Obstetrics and Gynecology, McMaster University, Hamilton, Ontario, Canada

[5] Vector Institute, Toronto, Ontario, Canada

[6] The Hospital for Sick Children, Toronto, Ontario, Canada

[7] Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, Canada

[$] Those authors contributed equally to this work

[#] Correspondence to: Yong.Zeng@uhn.ca; Mathieu.Lupien@uhn.ca; Trevor.Pugh@utoronto.ca; Hansen.He@uhn.ca

Abstract

Analyzing cell-free DNA (cfDNA) in blood through characterization of methylation patterns (methylome) and fragment properties (fragmentome) has advanced the minimal invasive detection of several individual cancers. Fully utilizing this approach requires identifying pan-cancer and cancer-specific cfDNA methylation and fragmentomic features, but challenges persist due to small, isolated cohorts and inconsistent data processing. In this study, we collated 1,074 cfMeDIP-seq profiles from 11 cancer types and healthy samples across nine studies. We developed a workflow for uniform data processing and quantification normalization to facilitate cross-cohort integration. Our analysis identified 24,418 pan-cancer DNA methylation markers for early cancer detection, and various cancer-specific markers for precise monitoring of particular cancer types. Fragmentomic analysis revealed distinguishing features across cancer types, such as end motifs, fragment lengths, and nucleosome footprints. Combining cfDNA methylation and fragmentomic features enhances differentiation between cancerous and healthy samples and among cancer types. This study provides a valuable pan-cancer resource for further cfDNA methylome and fragmentome investigations.

Keywords

cfMeDIP-seq, cancer, cell-free DNA, methylome, fragmentome

Introduction

Cell-free DNA (cfDNA) present in blood plasma has emerged as a promising analyte for cancer prognosis and treatment monitoring due to its non-invasive nature, with advanced genomic and epigenomic profiling techniques enabling the identification of tumor-associated signatures with high sensitivity [1]. In 2018, Shen et al. developed the cell-free DNA methylome (5-Methylcytosine) profiling technique with specialized immunoprecipitation and high-throughput sequencing (cfMeDIP-seq), demonstrating ultrasensitive tumor detection and classification capabilities [2]. This innovation has prompted an expansion in larger-scale cfDNA methylome profiling across various cancer types, utilizing refined cfMeDIP-seq protocols. Concurrently, the analysis of cfDNA fragmentation features, including fragment insert size, 5' end motifs, genome-wide fragments distribution patterns, and nucleosome footprinting, has also proven effective in cancer detection and classification [3–7]. However, despite these advancements, a comprehensive analysis that integrates both cfDNA methylome and fragmentome for a large cohort at pan-cancer level is still missing in the field.

Shen et al. reported single-end cfMeDIP-seq on approximately 400 plasma samples from seven diverse cancer types, including pancreatic cancer (PDAC: pancreatic ductal adenocarcinoma), colorectal cancer, breast cancer, lung cancer, renal cancer and bladder cancer and acute myeloid leukaemia (AML) [2]. With the evolution of cfMeDIP-seq protocols, especially the shift to paired-end sequencing, cfDNA methylome profiles have since been delineated for several other cancer types, including brain cancers (e.g. glioma, meningiomas, etc.) [8], head & neck cancer (squamous carcinoma)

[9], primary and metastatic prostate cancer (adenocarcinoma) [10], small cell lung cancer [11], uveal melanoma [12], and in individuals predisposed to various cancers due to hereditary Li-Fraumeni syndrome (LFS) [13]. Most recently, the cfDNA methylome and fragmentome have also been successfully utilized to monitor the outcomes of pembrolizumab-based treatment across multiple cancer types [14]. These studies have successfully pinpointed tumor-associated DNA methylation signatures and/or fragmentomic features, enabling the classification of cancer subtypes and differentiation from healthy controls. However, the isolation of these valuable datasets within individual research groups, coupled with diverse analytical workflows and scientific goals, has posed significant challenges for pan-cancer cfDNA methylome and fragmentome exploration.

In this study, we compiled a data resource of cfMeDIP-seq profiles derived from 1,074 plasma samples, spanning 11 major cancer types and healthy controls. These datasets were uniformly processed and quantified with batch correction and normalization, facilitating a thorough analysis of pan-cancer and cancer-specific cfDNA methylation and fragmentomic features. Moreover, we assessed the capability of these features to differentiate cancerous and healthy samples, as well as among different cancer types. Our study sets the foundation for creating a multi-cancer detection liquid biopsy-based platform that enables the early detection of cancer inception and monitoring of disease evolution.

Results

Pan-cancer liquid biopsy-based data collection, curation, uniform processing and quality assessment

We collected and curated a comprehensive dataset comprising a total of 1,074 blood plasma cfMeDIP-seq profiles sourced from nine distinct studies within The Cancer Genetics & Epigenetics (TCGE) project: TCGE-CFMe-MCA [2], TCGE-CFMe-BCA [8], TCGE-CFMe-HNSC [9], TCGE-CFMe-PRAD [10], TCGE-CFMe-AML [15], TCGE-CFM-SCLC [11], TCGE-CFMe-UM [12], TCGE-CFMe-HBC [12] and TCGE-CFMe-LFS [13]. This dataset includes liquid biopsies collected from healthy controls (N = 153) and patients diagnosed with one of 11 different cancer types, including brain cancer (N = 161), lung cancer (N = 141), prostate cancer (N = 133), AML(N = 78), pancreatic cancer (N = 71), uveal melanoma (N = 46), head & neck cancer (N = 35), breast cancer (N = 25), colorectal cancer (N = 23), bladder cancer (N = 20), renal cancer (N = 20), as well as from patients with hereditary Li-Fraumeni Syndrome (LFS), categorized into LFS-survivor (cancer-negative individuals with a cancer history, N = 75), LFS-previvor (cancer-negative individuals without a cancer history, N = 58), and LFS-positive (those with one or multiple cancer types, N = 35) (**Fig. 1A, B** and **Supplementary Table 1**).

These samples originated from a total of 918 participants, 68 of whom with multiple samples collected at different timepoints and/or health statuses (**Fig. 1C**). The samples were categorized by 373 male, and 297 female donors, with the remaining individuals lacking self-reported sex information (**Fig. 1C**). The ages at the time of blood draw, recorded for 723 samples, ranged from 1 to 92 years, with a median age of 62.5 (**Fig.**

1D). Each sample was assigned a unique ID that includes information about the source of a specific study, participant disease status (cancer type, primary versus (vs) metastatic cancer, or healthy control), and the timing order of each sample among multiple samples for the same participant (**Supplementary Table 1**).

All 1,074 plasma were processed with the cfMeDIP-seq assay and the sequencing data were processed uniformly with a standardized pipeline, MEDIPIPE [16]. Libraries were sequenced using a median of ~61 million raw sequencing reads, and about 28 million unique reads after QC trimming and duplication removal (**Fig. S1A**). Except for the TCGE-CFMe-MCA study, which utilized single-end (SE) sequencing, all samples were subjected to paired-end (PE) sequencing. The paired-end reads displayed a modal fragment length of 167 base pairs (bp), consistent with the expected size for nucleosome-associated cfDNA (**Fig. S1B**). Additionally, a median of 98.61% of paired-end reads contained CpG site(s), covering a median of 66.66% of CpG sites across the human genome (**Fig. S1B**).

We observed a strong positive correlation between the enrichment scores GoGe and relH, which indicate the enrichment of CpGs within sequencing reads compared to the reference genome [17] (**Fig. S1C, D**). However, both scores displayed a negative correlation with the fragment size in paired-end profiles (**Fig. S1D**), indicating that longer fragments may result in lower cfMeDIP signal-to-noise ratios. More than 90% of the samples reached enrichment scores of GoGe greater than 1.7, relH greater than 3.0, and a saturation score (maxEstCor) exceeding 0.9-thresholds previously suggested for

high-quality data by the inventors of the cfMeDIP-seq method [18] (**Fig. 1E** and **Fig. S1A**). We focused on 378 SE and 596 PE high-quality samples for all downstream analyses.

Establishing pan-cancer and cancer-specific cell-free DNA methylation signatures

Given the distinct DNA methylation profiles on sex chromosomes between males to females (**Fig. S2A**), and known problematic regions for genome alignment, our analyses were focused on autosomes after excluding the ENCODE blacklist region [19]. To mitigate the batch effects within and across the studies, we evaluated six different DNA methylation quantification and normalization strategies: raw read count, RPKM or FPKM, absolute methylation levels estimated by MEDEStrand and QSEA, and normalized read count by DESeq2 both without batch correction or with prior batch correction using ComBat-seq (**Methods**). We conducted this assessment within the TCGE-CFM-AML study, which included three technical replicates for each of the five participants. The results showed that the ComBat-seq + DESeq2 method was most effective in mitigating batch effects, successfully grouping technical replicates together for each participant based on batch-corrected and normalized read count (**Fig. S2B**). However, when we applied this method to all healthy samples from four studies, we found that batch effects were mitigated within the SE and PE samples but not across them (**Fig. S2C**). Additionally, the top principal components continued to exhibit a strong correlation with QC metrics such as raw and reusable read depth, and saturation scores (**Fig. S2D**). In contrast, treating the SE and PE samples separately resulted in significantly better outcomes in terms of batch effect removal (**Fig. S2C, E**). Together, these results suggest an intrinsic difference between SE and PE samples, which is likely attributed to differences in laboratory processing protocols.

Therefore, we applied the ComBat-seq + DESeq2 method to mitigate batch effects and normalize methylation quantification starting with raw read counts for the SE and PE samples, separately (**Fig. 2A, and Fig. S3A-D**). Then, we conducted differential methylation analysis and observed a greater number of hypermethylated regions compared to hypomethylated ones when combining SE and PE cancer samples against corresponding healthy samples, with fold changes of 673.3 and 244.9, respectively (**Fig. 2B).** Consistent with previous reports [2,9,10], we also observed a median of 114.5 and 8.5 across individual cancer types compared to healthy control in SE and PE studies, respectively (**Fig. S3E, F**). These hypermethylated regions tend to be enriched in the CpG islands, shores and shelves, as well as the regulatory promoter and enhancer regions (**Fig. S4A-C**). In contrast, the hypomethylated regions do not exhibit similar enrichment (**Fig. S4A-C**). We noted that the vast majority of differentially methylated regions (DMRs) identified by combined cancer samples tended to be re-identified across at least two comparisons of individual cancer types against healthy samples (**Fig. 2C**). However, only a few of these DMRs could be consistently detected in all individual cancer types vs healthy samples comparisons (**Fig. S4D**). We then focused on the 24,418 hypermethylated DMRs overlapping within the SE and PE studies as the pan-cancer signature (**Fig. 2D** and **Supplementary Data 1**). The genes associated with this pan-cancer signature are enriched in the GO term of DNA-binding transcription factor (TF) activity (**Fig. S4E**), aligning with reports that the binding efficiency of various TFs is dysregulated by hypermethylation in cancer [20].

Beyond the pan-cancer DNA methylation signature, we also identified a variable number of cancer-specific signatures, which were detectable only in the comparison of specific cancer types against all other samples in SE and PE studies in our dataset, accordingly (**Fig. S5A** and **Supplementary Data 1,** more details in **Methods**). Specifically, we observed that uveal melanoma and prostate cancer exhibited the highest number of cancer-specific hypermethylated regions (**Fig. 2E**), while pancreatic cancer and head & neck cancer had the highest number of specific hypomethylated regions (**Fig. S5B**). Intriguingly, we found that cancer-specific hypermethylated regions were more likely to be located in the promoter region and enriched around the transcription start site (TSS) region compared to cancer-specific hypomethylated ones (**Fig. 2E** and **Fig. S5B, C**). Furthermore, we found that the cancer-specific DNA methylation signatures linked to certain cancer types were related to specific GO terms and KEGG pathways (**Fig. 2F** and **Fig. S5D, E**). For instance, the prostate cancer, AML, colorectal cancer and lung cancer-specific hypermethylated regions were associated with the "DNA-binding TF activity" term (**Fig. 2F** and **Fig. S5D**). Additionally, the uveal, prostate, colorectal and renal cancer-specific hypermethylated regions were associated with metal ion transmembrane transporter activity, which plays a significant role in cancer cell biology via diverse chemical reactions and signal transduction pathways [21–23]. Meanwhile, colorectal cancer-specific hypomethylated signatures were associated with protein-membrane adaptor activity, and the bladder cancer-specific hypomethylated signatures were associated with RNA polymerase II-specific p53 binding (**Fig. S5E**). Although further investigations are needed to reveal the underlying mechanisms of how these cancer-specific DNA methylation signatures affect the biological functions and

pathways, the identification of these cancer-specific methylation signatures highlights the heterogeneity of DNA methylation captured within liquid biopsies across different cancer types.

Fragmentomic feature extraction from cfMeDIP pan-cancer data

More recently, fragmentomics [24], which probes the fragmentation patterns of cfDNA, has emerged as a promising tool for cancer detection [3–7]. While many cfDNA fragmentomic features have been studied using whole-genome sequencing (WGS) data, the exploration of these features in the context of cfMeDIP-seq has recently begun and remains a nascent area of interest [14]. Herein, we evaluated fragmentomic features of methylated cfDNA fragments, including the fragment insert size, genome-wide short to long fragment ratios (fragment ratios), nucleosome footprinting, and 5' end motif, for all uniformly processed cfMeDIP-seq paired-end samples (N = **596** and **Methods**).

Firstly, we confirmed that the fragment insert size was enriched in the range of 167 ± 15 bp (**Fig. S6A** and **Supplementary Table 2**). Our analysis also revealed that the proportion of short insert size fragments (20-150 bp / 20-600 bp) varied significantly across different cancer types **(**Kruskal-Wallis two-sided test: p-value < 1e-4) (**Fig. 3A** and **Supplementary Table 3**). All cancers, except for head & neck and brain cancers, exhibited a significantly higher proportion of short fragments compared to healthy samples (Dunn's post-hoc two-sided test, p-values ranging from 3.56e-02 to 1.23e-09). Furthermore, we derived two insert size-based fragment signatures using the Non-negative Matrix Factorization (NMF) analysis, enabling the determination of predominant fragment lengths in liquid biopsies collected in cancer and healthy samples

(**Fig. S6B**, **Supplementary Table 3** and **Methods**). These signatures, referred to as cancer-associated and healthy-associated fragment insert size profiles, allowed us to compare the extent to which fragment lengths in each sample resembled these profiles. The cancer-associated fragment insert size profiles were quantified as a weighted fragment score, providing a measure of how closely the fragment lengths in each sample matched the cancer-associated profile. Consistently, all cancer types, with the exception of head & neck and brain cancers, demonstrated a significantly higher degree of resemblance to the cancer-associated fragment insert size profile and a higher weighted fragment score compared to the healthy-associated profile, consistent with findings for the proportion of short fragments (Dunn's post-hoc two-sided test, p-values ranging from 8.76e-02 to 2.08e-11) (**Fig. S6C** and **Supplementary Table 3**).

Next, we examined the fragments ratios in 5 Mb bins across the genome using a modified version of the DELFI method (**Methods**). Each cancer type contained samples with significantly variable fragment ratios across the genome compared to healthy samples (**Fig. S7A, B** and **Supplementary Table 4**). Particularly, lung and prostate cancers had over 78% and 76% samples, respectively, that are differentiated from healthy samples based on fragment ratios (**Supplementary Table 3** and **Methods**). Subsequently, we conducted nucleosome footprinting analysis, which evaluates the proximity of likely nucleosome-bound 167 bp fragments to reference nucleosome positions [25] (**Methods**). We observed increased fragmentation within the reference nucleosome cores in different cancer types compared to healthy samples (**Fig. 3B** and **Supplementary Table 5**). Particularly, all AML samples and 56% of prostate cancer

samples showed a significant deviation from the median of the healthy samples, attributed to an increased proportion of fragment ends in the middle of expected nucleosome positions (**Supplementary Table 3** and **Methods**). These findings suggest aberrant nucleosome positioning relative to healthy controls.

We next performed fragment motif analysis by examining the frequencies of all possible four nucleotide sequences at each fragment 5' end across PE samples in our dataset (**Fig. 3C)**. We found that the fragments' 5' end motifs in the head and neck cancer samples predominantly contained combinations of adenine and thymine (A/T), while the lung cancer samples showed a higher prevalence of cytosine and guanine (C/G) compared to the healthy samples (**Fig. S8A** and **Supplementary Table 6**). And 92% of AML samples had significantly different end motifs compared to healthy controls as determined by calculation z-scores of end motif proportions (**Fig. S8B, Supplementary Table 3** and **Methods**). We found significant variations in the frequency of end motifs commonly associated with the enzyme *DNASE1L3,* which plays a crucial role in plasma DNA fragmentation during apoptosis and necrosis and has been shown to preferably cut methylated DNA [26,27]. *DNASE1L3*-associated motifs were depleted in AML and prostate cancer, while enriched in uveal melanoma, brain cancer, and LFS previvor cases relative to healthy controls (fold change of 0.71 to 1.19, p-value 2.48e-2 to 2.02 e-12, adjusted p-values from two-sided t-tests) (**Fig. S8C** and **Methods**).

Lastly, to further investigate the relationship between fragment insert size, fragment ratios, nucleosome peak distances and end motif, we computed a z-score per-sample

for each individual feature based on healthy samples (**Supplementary Table 3** and **Methods**). The strongest correlations were observed between the weighted fragment scores and the proportion of short fragments (rho = 0.87), as well as between the z-scores of nucleosome distances and fragment ratios (rho = 0.75) (**Fig. S9**). Additionally, when considering these four fragmentomic features together, we observed that healthy samples exhibited greater consistency (standard deviation (SD) = 1.83) compared to other cancer types except uveal melanoma (mean SD = 3.70 , min SD = 1.67, max SD = 11.00) (**Fig. 3D**). Across all cancer types, end motifs were most different from healthy controls (mean z-score = 2.80), followed by fragment ratios (mean z-score = 2.34), insert sizes (mean z-score = 1.74) and nucleosome peaks distances (mean z-score = 1.65). Differentiation from healthy samples also varied across cancer types: nucleosome peak distances were most distinctive for LFS positive and survivor samples, while end motifs distinguished uveal melanoma, AML, LFS previvor and head and neck cancer samples. Lung and prostate cancers were differentiated by fragment ratios, and brain cancer by insert sizes (**Fig. 3D** and **Methods**). LFS previvor, eye cancer, and brain cancer samples had the lowest average genome-wide z-scores (cohort means 1.41, 1.63, and 1.66, respectively) indicating lesser deviations from fragmentation of healthy samples. In contrast, AML, lung cancer and prostate cancer samples exhibited the highest average z-scores (cohort means 18.32, 5.78, and 4.12, respectively). In conclusion, our analysis reveals significant heterogeneity in fragmentomic features of methylated fragments across cancer types, highlighting the potential of fragmentomic profiling as biomarkers for differentiating cancer types from healthy samples.

cfDNA methylome and fragmentomic features-based cancer classifiers

Leveraging the inherent ability to extract DNA methylation and fragmentomics measurement from cfMeDIPseq data, we further explored the potential of integrating DNA methylation signatures and fragmentomic features for classifying cancer vs healthy samples. We trained individual models based on principal components (PCs) derived from nucleosome footprinting, fragment ratios, end motifs, insert sizes and methylation scores at hypermethylated DMRs using principal component analysis (PCA), separately (**Methods**). PCs derived from end motifs achieved the highest area under the curve (AUC) of 0.848, followed by methylation (0.847) and fragment ratios (0.847), insert size (0.726) and nucleosome footprinting (0.666) (**Fig. 4A** and **Fig. S10**). Combining PCs that contribute >1% of variance for end motifs and methylation together resulted in an improved AUC of 0.876, surpassing the AUCs obtained by including PCs contributing >1% of variance for fragment ratios as well (AUC = 0.855) or with all PCs contributing >5% variance across all category features (AUC = 0.817) (**Fig. 4A** and **Fig. S10**). Overall, end motifs and fragment ratios were equally effective at differentiating cancer from healthy samples as methylation scores alone. Combining fragmentation features with methylation scores resulted in an increase in classification accuracy. Regarding the classification algorithms we tested, Random Forest (RF) and Gradient Boosting Machine (GBM) consistently outperformed Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Generalized Linear Model (GLM) in discriminating cancer samples from healthy controls (**Fig. 4B**).

We subsequently developed distinct predictive models using the above selected PCs for each type of cancer, comparing each cancer type with every other cancer type as well as with the healthy controls (**Supplementary Table 7** and **Methods**). Fragment ratios were most effective at classifying lung, prostate and brain cancer samples from healthy controls, while end motifs were most effective at distinguishing LFS previvor cases. The combined end motif and methylation PCs were most effective for distinguishing LFS survivors from healthy controls (**Fig. 4C**). Overall, incorporating fragmentation features also resulted in more accurate classification of individual cancer types from healthy control samples than using methylation alone, with an average gain of 0.074 to classification accuracy (AUCs 0.887 - 0.985 with fragmentation features vs 0.756 - 0.964 with methylation alone) (**Fig. 4C**). Between cancer types, an average gain of 0.14 in AUCs were achieved when incorporating end motif to methylation alone (**Fig. 4C** and **Supplementary Table 7**). Furthermore, we also attempted to distinguish between cancer subtypes. Fragment ratios were most effective at discerning IDH wild type glioma from IDH mutant glioma (AUC = 0.790), as well as metastatic resistant prostate cancer from primary prostate cancer (AUC = 0.779) (**Fig. S11**). LFS positive cases were best distinguished from LFS previvors by end motifs (AUC = 0.833), and LFS survivors by fragment ratios (AUC = 0.873) (**Fig. S11** and **Supplementary Table 7**). LFS cases retain distinct properties to healthy samples by end motifs as previvors (AUC = 0.973) which are retained following surviving treatment (AUC = 0.889) (**Fig. S11** and **Supplementary Table 7**).

Overall, we demonstrated that both DNA methylation and fragmentomic features of methylated fragments possess classification capabilities for cancer detection. The identification of distinct end motifs and methylation signatures emphasizes their potential as biomarkers for early cancer detection and monitoring of tumor evolution.

Discussion

In this study, we compiled nine cfMeDIP-seq datasets primarily generated as part of The Cancer Genetics & Epigenetics (TCGE) Project, led by the Princess Margaret Cancer Centre/University Health Network (UHN) and the Ontario Institute for Cancer Research (OICR). Despite the continuous emergence of new cfMeDIP-seq datasets [14,28,29], our collection offers substantial resources with the largest sample size and coverage of cancer types currently available. Additionally, we evaluated several methods for quantifying and normalizing cfDNA methylation. While MeDEStrand [30] and QSEA [31] were capable of estimating absolute methylation levels for the cfMeDIP enrichment data, we observed that estimations from these tools were more consistent within the same studies or batches. In contrast, we determined that DESeq2 normalization after batch correction by ComBat-seq was generally more effective in mitigating batch effects, particularly, when integrating multiple datasets from different research groups. These assessments also lay the groundwork for incorporating future cfMeDIP-seq datasets for accurate downstream analyses.

Our analysis validated the prevalence of hypermethylation across various cancer types compared to healthy samples, as reported in previous individual studies [2,8,10–13].

However, we observed a significant variation in the number of hyper- over hypo-DMRs among different cancer types, and a varied number of cancer-specific DMRs, suggesting both common and unique cancer-specific epigenetic alterations across cancer types. The utilization and validation of these pan-cancer and cancer-specific cfDNA methylation signatures in additional independent cohorts is necessary to confirm the robustness of these biomarkers for the cancer diagnosis and prognosis. Further research could also explore the functional implications of these signatures, providing potential avenues for precision cancer diagnostic and therapeutic interventions. In addition, we acknowledge that false positive DMRs may arise from methylation signals originating in peripheral blood leukocytes (PBLs). To mitigate this issue, Burgener et al. have proposed profiling the paired PBL methylome for both cancer and healthy samples to deplete signal contaminations [9]. Subsequently, Ul Haq et al. applied this strategy of peripheral blood leukocyte methylation (PRIME) subtraction from the same 74 patients to refine profiling of the cancer-specific methylome in their cohort [11]. By incorporating this strategy, future studies may further improve the precision of these biomarkers and enhance their clinical utility.

Our investigation into cfDNA fragmentomic features revealed significant variations in insert sizes, short/long fragment ratios, nucleosome footprinting, and 5' end motifs across different cancer types. These findings are consistent with prior studies emphasizing the utility of cfDNA fragment analysis in cancer detection [3–7]. Our findings underscore the predominance of shorter fragment sizes in all cancer samples except for head & neck and brain cancers compared to healthy controls. This characteristic of low

ctDNA content in head & neck and brain cancers aligns with lower cancer probability scores from NMF analysis and minimal variations in nucleosome positioning, end motifs, and fragment ratios in these cancers. Previous studies have indicated lower ctDNA abundance in primary brain cancers such as medulloblastomas, WHO grade 2-3 gliomas, and WHO grade IV astrocytomas, with detectable levels of ctDNA found in less than 50% of medulloblastoma and less than 10% of glioma cases [32]. This reduced detectability could be due to the blood-brain barrier limiting ctDNA shedding into the bloodstream, highlighting the need for alternative strategies for ctDNA collection in central nervous system (CNS) neoplasms, such as cerebrospinal fluid sampling [33]. For head and neck cancers, the low ctDNA content could be attributed to high immune activity and a significant presence of normal cell cfDNA that might obscure the cancer signal. Head and neck tumors, both HPV-positive and HPV-negative, are among the most highly immune-infiltrated cancer types, with high levels of CD8+ T cell and CD56dim NK cell infiltration compared to other cancer types in TCGA [34]. Inflammatory responses and immune activity in head and neck cancers can lead to increased levels of cfDNA from non-cancerous cells, which may dilute detectable ctDNA [35]. These challenges may be mitigated by using saliva for ctDNA capture, as saliva is enriched for ctDNA in cancers of the oral cavity [36].

We found *DNASE1L3*-associated motifs were depleted in AML and prostate cancer, while enriched in uveal melanoma, brain cancer, and LFS previvor cases relative to healthy controls (**Fig. S8C**). These findings could be partially explained by existing literature indicating that DNASE1L3 is downregulated in several cancers and its lower

expression correlates with poorer prognosis [37]. Moreover, the distinct fragmentation patterns observed in different cancer types align with studies highlighting the role of *DNASE1L3* in generating non-random cfDNA fragmentation, which is influenced by DNA methylation status [27]. Hou et al. (2024) further support this by demonstrating that specific cfDNA fragmentation patterns, including end motifs, have diagnostic value and can differentiate cancer types when integrated with other genomic features [38]. Additionally, Han and Lo (2021) explored the balance of cfDNA generation and clearance, linking plasma nuclease activity, including *DNASE1L3*, to various pathologies. They emphasized that changes in cfDNA fragmentation patterns, influenced by nucleases like *DNASE1L3*, have significant implications for cancer diagnostics and the understanding of cfDNA biology [39]. Our results contribute to this understanding by highlighting the differential presence of *DNASE1L3*-associated end motifs across cancer types, suggesting the role of *DNASE1L3* in cancer-specific genomic instability and its potential as a biomarker for cancer detection and prognosis.

Our classification model, leveraging methylation bins at hyper DMRs and fragmentomic features, resulted in improved classification accuracy compared to using methylation alone, with average AUC gains of 0.074 for distinguishing cancer types from healthy controls and 0.14 for differentiating between cancer types when integrating fragmentation features. Particularly, end motifs and fragment ratios emerged as robust classifiers, matching the capacity of DMRs in distinguishing cancer from healthy samples, and exceeding the capacity when integrated (**Figure 4A**). This could be due to the inherent nature of end motifs and fragment ratios to capture specific fragmentation patterns associated with cancer-specific genomic instability. Irregular apoptosis and

necrosis processes often exhibited by cancer cells lead to distinct end motifs and fragment ratios compared to those in healthy cells [3,40,41]. These results highlight the potential of cfMeDIP in conjunction with fragmentomic features as a screening tool. Comparatively, traditional methods such as liquid biopsy-based mutation detection or protein biomarkers often suffer from lower sensitivity and specificity. Our findings suggest that cfDNA methylation and fragmentomic features may provide a more reliable and non-invasive alternative for cancer detection [42,43]. Furthermore, the use of cfMeDIP-seq allows for the simultaneous extraction of both methylation and fragmentomic data, streamlining the workflow and reducing costs associated with multiple testing methodologies. This integrated approach could enhance the feasibility of large-scale screening programs and personalized medicine strategies, particularly in settings where resource allocation is a concern.

Future research could aim to improve integrated models that harness the complementary strengths of cfMeDIP-seq and fragmentomics analysis. Enhancing the accuracy of classification algorithms may be achieved by exploring other clustering techniques such as neural networks and larger, more diverse datasets to better capture the heterogeneity of cancer types. Expanding the scope of analysis to encompass a broader spectrum of cancer types and stages, including early-stage cancers, will provide a more comprehensive understanding of combined methylation and fragmentome diagnostic utility. Deepening the exploration of the biological mechanisms behind observed fragmentomic patterns could reveal new insights into cancer biology, potentially identifying novel biomarkers for early detection and monitoring. Lastly,

integrating other multi-omics approaches, such as proteomics and transcriptomics, with cfDNA analysis may provide a more holistic view of tumor biology and improve classification performance.

Methods

cfMeDIP-seq data uniformly procession, QC and quantification

All centralized cfMeDIp-seq data underwent uniform processing using our previously developed pipeline MEDIPIPE [16]. Specifically, first, adapter sequence, low-quality bases and UMI barcodes (if applied) were removed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and UMI-tools [44]. Next, preprocessed reads were mapped to human genome (hg38) using BWA-MEM [45], and duplicated reads were eliminated either by SAMtools markdup [46] or UMI-tools dedup [44], depending on whether UMI barcodes were applied. Lastly, MEDIPIPE quantified methylation levels in consecutive 300 bp intervals across the genome using three different tools: MEDIPS (raw read count, and RPKM or FPKM) [17], QSEA (estimated Beta value) [31] and MeDStrand (estimated relative methylation score (rms)) [30]. Importantly, for paired-end cfMeDIP-seq reads, MEDIPIPE estimated fragment size using the Picard tool kit (http://broadinstitute.github.io/picard/). Additionally, comprehensive QC metrics (N = 21), encompassing cfMeDIP-seq raw and preprocessed read depths, saturation, coverage, specificity, enrichment scores and fragment size (for PE samples only), were calculated by MEDIPIPE as well.

cfDNA methylation quantifications normalization comparison and visualization

To evaluate the impact of the confounding factor, sex, raw read count for all consecutive 300 bp bins on chromosome X was extracted. Batch correction and normalization were performed within each study using ComBat-seq plus DESeq2. Subsequently, PCA analysis was employed to assess the effects of sex, revealing distinct differences between male and female methylation profiles on chromosome X (**Fig. S2A**). Consequently, we decided to filter out sex chromosomes. Additionally, the mitochondrial chromosome and problematic ENCODE blacklist regions [19] were excluded from our downstream analysis.

To compare methods for mitigating batch effects within (**Fig. S2B**) and across the studies (**Fig. S2C**), we assessed various approaches, including raw read count, RPKM or FPKM, absolute methylation levels estimated by MEDEStrand (rms) and QSEA (Beta). Additionally, we normalized read count using DESeq2 both without and with prior batch correction using ComBat-seq. Specifically, ComBat-seq was used to correct known library preparation or sequencing batches while persevering the sample subtype differences within each study. DESeq2 was then applied to raw read count or batch-corrected count for normalization with size factors. Although the ComBat-seq plus DESeq2 method achieved best results (**Fig. S2B, C**), it was still unable to completely mitigate the batch effect between SE and PE samples (**Fig. S2C**). Therefore, we opted to run the downstream analyses by separating the SE and PE studies.

For the DNA methylation quantifications-based PCA analyses (**Fig. 2A, Fig. S2 and Fig. S3A-D**), we required at least 1 CpG site in the 300 bp bin. The top 10k variable bins were then selected based on the interquartile range (IQR). The variance explained by the top 9 principle components (PCs) was calculated, and their Spearman correlations with 5 representative QC metrics (raw and usable read depth, saturation score, and enrichment scores (GoGe and relH)), were examined. For the corresponding UMAP plot, the top 500 PCs were utilized as the input for dimensionality reduction.

Identification of pan-cancer and cancer-specific cfDNA methylation signatures

To identify pan-cancer cfDNA methylation signatures, initially applied DESeq2 to compare each cancer against its corresponding health control based on the ComBat-seq corrected quantification within both SE and PE samples, separately. Similarly, we conducted the comparisons between combined cancer samples and healthy control samples. The identification of DMRs in each comparison was limited in those 300 bp bins with more than 5 CpG sites, and required a fold change to be greater than 2 and false discovery rate (FDR) less than 0.05. Lastly, the pan-cancer hyper- and hypo-DMRs were limited to those consistently detected in all of the above comparisons. For the cancer-specific DMRs, we compared each individual cancer type to all the rest samples within both SE and PE samples, separately, and filtered out those DMRs reported more than once (**Fig. S5A**). It's important to note that for the shared cancer type in SE and PE samples, we performed filtering based on the DMRs identified in both scenarios. In-house scripts were developed to examine the enrichment of DMRs in annotated CpG regions, as well as promoter and enhancer regions within annotatr package [47]. The ChIPseeker [48] was employed to annotate and visualize the distribution

of DMRs across genomic regions, and clusterProfiler [49] was utilized to perform functional enrichment analysis for genes associated with the corresponding DMRs.

cfMeDIP-seq fragments length and distribution across genome

The cfMeDIP-seq fragment lengths were determined using Picard's (RRID:SCR_006525) CollectInsertSizeMetrics tool (version 4.0.1.2). The proportion of each fragment length between 10 and 600 bp was standardized against the median and standard deviation of the healthy controls to derive z-scores for each sample. These z-scores were then aggregated to generate a per-sample fragment size z-score, with values exceeding 2 considered statistically significant, indicating a substantial deviation from healthy controls. The proportion of short fragments was determined as the ratio of fragments between 20 and 150 bp in length to those between 20 and 600 bp in length. Additionally, we applied Non-negative Matrix Factorization (NMF), as outlined by Vessies et al. (2022) [50], assigning a probability score to each fragment to identify if it came from cancerous or healthy samples. By splitting our dataset 70% for training, we identified two signatures through NMF: one prevalent in cancer and the other in healthy samples. We visualized fragment length frequencies and the log-ratio of signature weights for assessing fragment origin probability (**Figure S6B**). Finally, we calculated patient-specific NMF scores to determine their likelihood of indicating cancer (**Figure S6C**).

Moreover, to assess the size distribution and fragmentation ratios of cfDNA across the genome, we employed the DELFI (DNA Evaluation of Fragments for Early Interception) technique, as detailed by Cristiano et al. [4]. Briefly, this method involves calculating the

ratio of short fragments (90-150 bp) to long fragments (151-220 bp) within defined genomic windows of 5 megabase pairs (Mb) in length across the genome. This analysis includes GC content correction and read depth adjustment to ensure accurate quantification [4]. Then, the fragmentation ratio profiles were evaluated by calculating bin-wise z-scores relative to healthy controls, and the absolute value of all z-scores were summed to compute a single genome-wide z-score per sample. The Kruskal-Wallis test was applied to assess disparities since the data did not meet the criteria of normal distribution for a parametric test to be used.

cfMeDIP-seq fragments based nucleosome footprinting and 5' end motif patterns

Using the method outlined by Vanderstichele et al. (2022), we analyzed plasma cfDNA for nucleosome positioning patterns indicative of chromatin structure alterations in cancer cells [5]. By focusing on fragments which were 167 bp in length, reflective of the DNA wrapped around a single nucleosome plus linker DNA, we aimed to capture the unique "footprint" left by nucleosome organization in the tumor-derived cfDNA. This analysis was predicated on the understanding that cancer cells exhibit distinct nucleosome spacing and positioning, diverging from the patterns observed in non-cancerous cells. In order to capture this diversity, we calculated the distances of 167 bp fragments from positions of ~13 million known nucleosomes from a cfDNA healthy blood reference prepared using deep WGS of cfDNA by Snyder et al. [25]. We then calculated z-scores by comparing the median proportion of fragments at distances ± 300bp from nucleosome positions in healthy samples to cancer samples. Following the method by Jiang et al. (2020), we performed an analysis of 5' cfDNA end motifs to

identify fragmentation patterns specific to cancer cfDNA. The 5’ end was used due to the elongation or truncation of the 3’ end of fragments which occurs during sample preparation for sequencing. End motifs which could not be clearly determined by sequencing (ie, a letter indicating an uncertain base pair designation in the fragment), were removed. The proportion of fragments at each end motif was then calculated for each sample. We then calculated z-scores by comparing the proportion of fragments at each end motif between each sample and healthy controls. The z-scores were then summed using absolute values across all 256 motifs to calculate a per-sample z-score.

Genome-wide z-score calculation for fragment features

We calculated z-scores for each fragmentomic feature by comparing the proportion of fragments at each fragment length, genomic bin, distance from nucleosome core or end motif to those from healthy control samples. For each feature, we aggregated these z-scores to establish a per-sample z-score. We calculated the median and standard deviation of the fragment proportions in the healthy control group to determine the baseline distribution. For each cancer sample, we then computed the deviation from the healthy median, normalized by the healthy standard deviation, resulting in a z-score for each fragment length, bin, distance or motif. The overall z-score for each feature was the sum of the absolute values of these individual z-scores, providing a measure of how much each cancer sample deviated from the healthy baseline. Z-scores greater than 2 were considered significantly different from healthy controls, amounting to a significance level below 0.05%.

DNA methylation and fragmentomic features based classification

We first applied PCA for methylation fragmentation features dimensionality reduction, retaining the top-ranked components that together explained at least 90% of the variance for individual feature-based classification modeling. We selected the top-ranked 19, 201 and 169 components that together explained 90% of the variance for the features of methylation, fragment ratios and nucleosome peaks, respectively. For features of end motif and insert size, due to the limited number of components, we retained the top-ranked 21 and 17 components, respectively, achieving 97% explained variance. In the combined features-based classification modeling, components that explained more than either 5% or 1% of the variance were retained and tested. Nineteen components were found to explain more than 5% of the variance across all category features, while 30 components explained more than 1% of the variance across methylation, end motif and fragment ratio features.

We then performed model optimization through testing a predefined list of classification algorithms including random forests (RF), generalized linear models (GLM), gradient boosting machines (GBM), support vector machines with a radial basis function kernel (SVM), and k-nearest neighbors (KNN) with matrices containing normalized features being explored. Next, each classifier was first trained on the combined dataset of case and control groups using cross-validation. Cross-validation was employed to evaluate model performance, specifically through repeated cross-validation with 10 folds and 1 repeat. A down-sampling approach is used to balance the dataset, ensuring that each class is equally represented in the training set, thus avoiding bias towards the majority

class. This process is repeated for each of the 10 folds. Performance was then assessed using the kappa statistic (a measure of classification accuracy corrected for chance agreement). The classification algorithm which resulted in the highest kappa statistic is used for the next phase of model fitting. The total data is first split 20 times to create 20 distinct sets of data suitable for cross-validation. Each of the 20 sets is then used for a 10-fold cross-validation with 3 repeats. This strategy aims to mitigate overfitting by averaging model performance across multiple splits of the data and testing a model's ability to generalize across different subsets of the data. The outcomes of this analysis are then saved and plotted using custom R scripts to determine model accuracy. In addition, all models were developed using several training strategies to assess classification performance. Initially, a model was trained with all pooled cancer samples and healthy controls. Subsequently, separate analyses trained models to distinguish individual cancer types and subtypes from healthy samples. Further refinement involved training models to identify differences between each cancer type, as well as between each cancer subtype, in a pairwise manner. This tiered approach allowed for comparison of model effectiveness across different levels of sample specificity.

Data Availability

The raw cfMeDIP-seq data from the studies TCGE-CFMe-MCA, TCGE-CFMe-BCA, TCGE-CFMe-HNSC, TCGE-CFM-SCLC are available upon request from the corresponding authors of those studies. The other cfMeDIP-seq data was deposited in the European Genome-Phenome Archive (EGA): TCGE-CFMe-AML

(EGAS00001005069), TCGE-CFMe-PRAD (EGAS00001005522), TCGE-CFMe-UM (EGAD00001008998), TCGE-CFMe-HBC (EGAS00001006539) and TCGE-CFMe-LFS (EGAS00001006539). The uniformly processed cfMeDIP-seq results are available on Zenodo (10.5281/zenodo.12785761 via the link for reviewers: https://zenodo.org/records/12785761?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjBhM2UyNTMyLTYxMzMtNGEzOS04MDY5LWMxYTUwZjBkN2JkNiIsImRhdGEiOnt9LCJyYW5kb20iOiJjMDhiNGFiOGViODYyNThmNGFjODZlNGI0OWQyOTg2NyJ9.F-Y9Azz9k53RuERNpiHougC3PcMlVpq-Bis2ur0FiOb2A7Miu-_ainy-doCidXvFZzrf2N4KqfmUNo_tOjieOA). The remaining data are available within the Article, Supplementary Tables and Data.

Code Availability

All codes for this study have been deposited at GitHub (https://github.com/HansenHeLab/cfMeDIP-seq_Data_Resource_Codes). The analytical pipeline will be also deposited on the CoBE platform (www.pmcobe.ca). Scripts used for the fragmentomics analysis were based on scripts originally developed for these projects: https://github.com/pughlab/TGL48_Uveal_Melanoma and https://github.com/pughlab/LFS-early-detection-ctdna.

Competing interests

SVB: Stock ownership in Adela; leadership position in Adela; patents licensed to Roche, Adela; and royalties from Roche. TJP reports personal fees from AstraZeneca, Canadian Pension Plan Investment Board, Chrysalis Biomedical Advisors, Illumina, Merck, PACT Pharma, and SAGA Diagnostics, and grants from Roche/Genentech

outside the submitted work. B.H. Lok reports grants from Pfizer and grants, personal fees, and nonfinancial support from AstraZeneca and personal fees from Daiichi-Sankyo outside the submitted work. The other authors declare no competing interests.

Funding

This work was supported by the Cancer Genetics and Epigenetic Program at Princess Margaret Cancer Center. H.H.H. is supported by Canadian Institute of Health Research (CIHR) Project Grants (FRN-142246, 152863, 152864, 159567 and 438793) and Terry Fox New Frontiers Program Project Grants (1090 and 1124). H.H.H holds a tier 1 Canadian Research Chair in RNA Medicine. T.J.P. holds the Canada Research Chair in Translational Genomics and is supported by a Senior Investigator Award from the Ontario Institute for Cancer Research and the Gattuso-Slaight Personalized Cancer Medicine Fund. M.L is supported by the CIHR (FRN-153234, 158225, 168933 & 191847), the Ontario Institute for Cancer Research (OICR) Investigator Award through funding provided by the Government of Ontario (IA-031) and the Princess Margaret Cancer Foundation. M.L. holds the Joey and Toby Tanenbaum/Brazilian Ball Chair. Research in the B.H. Lok laboratory is supported by the Canada Foundation for Innovation, CIHR, National Institute of Health/National Cancer Institute (U01CA253383), Terry Fox Research Institute Program Project Grant (1124), Clinical and Translational Science Center at Weill Cornell Medical Center, MSKCC (UL1TR00457). M.M.H. is supported by a CIHR Project Grant (408773). Data used in this study were generated with the support of the Ontario Institute for Cancer Research Genomics Program (http://genomics.oicr.on.ca) and Translational Genomics Laboratory, a joint initiative between the Princess Margaret Cancer Centre and the Ontario Institute for Cancer

Research (Dax Torti, Kayla Marsh, Bernard Lam, Morgan Taschuk, Lawrence Heisler, Carolyn Ptak). These programs were enabled through funding provided by the Government of Ontario and the Princess Margaret Cancer Foundation. Additional infrastructure support to T.J.P. from the Canada Foundation for Innovation, Leaders Opportunity Fund [CFI #32383 and #38401]; Ontario Ministry of Research and Innovation, Ontario Research Fund Small Infrastructure Program; and the Ontario Institute for Cancer Research.

Acknowledgements

We would like to thank members of the cell-free Multiomics Data Coordination Centre (cfMOS-DCC) for supporting this project and facilitating access to cfMeDIP-seq data. We also acknowledge the Princess Margaret Genomics and Bioinformatics group for providing the infrastructure required to conduct analyses included in this work.

References

1. Corcoran, R. B. & Chabner, B. A. Application of Cell-free DNA Analysis to Cancer Treatment. *N. Engl. J. Med.* **379**, 1754–1765 (2018).
2. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
3. Jiang, P. *et al.* Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discov.* **10**, 664–673 (2020).
4. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
5. Vanderstichele, A. *et al.* Nucleosome footprinting in plasma cell-free DNA for the

pre-surgical diagnosis of ovarian cancer. *npj Genom. Med.* **7**, 1–9 (2022).

6. Zhou, Z. *et al.* Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2220982120 (2023).
7. Mouliere, F. *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, (2018).
8. Nassiri, F. *et al.* Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **26**, 1044–1047 (2020).
9. Burgener, J. M. *et al.* Tumor-Naïve Multimodal Profiling of Circulating Tumor DNA in Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.* **27**, 4230–4244 (2021).
10. Chen, S. *et al.* The cell-free DNA methylome captures distinctions between localized and metastatic prostate tumors. *Nat. Commun.* **13**, 6467 (2022).
11. Ul Haq, S. *et al.* Cell-free DNA methylation-defined prognostic subgroups in small-cell lung cancer identified by leukocyte methylation subtraction. *iScience* **25**, 105487 (2022).
12. Wong, D. *et al.* Integrated, Longitudinal Analysis of Cell-free DNA in Uveal Melanoma. *Cancer Res Commun* **3**, 267–280 (2023).
13. Wong, D. *et al.* Early Cancer Detection in Li-Fraumeni Syndrome with Cell-Free DNA. *Cancer Discov.* (2023) doi:10.1158/2159-8290.CD-23-0456.
14. Stutheit-Zhao, E. Y. *et al.* Early changes in tumor-naive cell-free methylomes and fragmentomes predict outcomes in pembrolizumab-treated solid tumors. *Cancer Discov.* (2024) doi:10.1158/2159-8290.CD-23-1060.

15. Wilson, S. L. *et al.* Sensitive and reproducible cell-free methylome quantification with synthetic spike-in controls. *Cell Rep Methods* **2**, 100294 (2022).
16. Zeng, Y. *et al.* MEDIPIPE: an automated and comprehensive pipeline for cfMeDIP-seq data quality control and analysis. *Bioinformatics* **39**, (2023).
17. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**, 284–286 (2014).
18. Shen, S. Y., Burgener, J. M., Bratman, S. V. & De Carvalho, D. D. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nat. Protoc.* **14**, 2749–2780 (2019).
19. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
20. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet.* **37**, 1012–1027 (2021).
21. Liu, Y., Wang, Y., Song, S. & Zhang, H. Cancer therapeutic strategies based on metal ions. *Chem. Sci.* **12**, 12234–12247 (2021).
22. Huang, L., Li, W., Lu, Y., Ju, Q. & Ouyang, M. Iron metabolism in colorectal cancer. *Front. Oncol.* **13**, 1098501 (2023).
23. Thévenod, F., Schreiber, T. & Lee, W.-K. Renal hypoxia-HIF-PHD-EPO signaling in transition metal nephrotoxicity: friend or foe? *Arch. Toxicol.* **96**, 1573–1607 (2022).
24. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16 Suppl 13**, S1 (2015).

25. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).
26. Chan, R. W. Y. *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).
27. An, Y. *et al.* DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. *Nat. Commun.* **14**, 287 (2023).
28. Lu, H. *et al.* Detection of ovarian cancer using plasma cell-free DNA methylomes. *Clin. Epigenetics* **14**, 74 (2022).
29. Janke, F. *et al.* Longitudinal monitoring of cell-free DNA methylation in ALK-positive non-small cell lung cancer patients. *Clin. Epigenetics* **14**, 163 (2022).
30. Xu, J., Liu, S., Yin, P., Bulun, S. & Dai, Y. MeDEStrand: an improved method to infer genome-wide absolute methylation levels from DNA enrichment data. *BMC Bioinformatics* **19**, 540 (2018).
31. Lienhard, M. *et al.* QSEA-modelling of genome-wide DNA methylation from sequencing enrichment experiments. *Nucleic Acids Res.* **45**, e44 (2017).
32. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
33. De Mattos-Arruda, L. *et al.* Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat. Commun.* **6**, 8839 (2015).
34. Mandal, R. *et al.* The head and neck cancer immune landscape and its

immunotherapeutic implications. *JCI Insight* **1**, e89829 (2016).

35. Zwirner, K. *et al.* Circulating cell-free DNA: A potential biomarker to differentiate inflammation and infection during radiochemotherapy. *Radiother. Oncol.* **129**, 575–581 (2018).
36. Cui, Y. *et al.* Longitudinal detection of somatic mutations in saliva and plasma for the surveillance of oral squamous cell carcinomas. *PLoS One* **16**, e0256979 (2021).
37. Deng, Z. *et al.* DNASE1L3 as a Prognostic Biomarker Associated with Immune Cell Infiltration in Cancer. *Onco. Targets. Ther.* **14**, 2003–2017 (2021).
38. Hou, Y., Meng, X.-Y. & Zhou, X. Systematically Evaluating Cell-Free DNA Fragmentation Patterns for Cancer Diagnosis and Enhanced Cancer Detection via Integrating Multiple Fragmentation Patterns. *Adv. Sci.* e2308243 (2024).
39. Han, D. S. C. & Lo, Y. M. D. The Nexus of cfDNA and Nuclease Biology. *Trends Genet.* **37**, 758–770 (2021).
40. Budhraja, K. K. *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).
41. Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).
42. Furuki, H. *et al.* Evaluation of liquid biopsies for detection of emerging mutated genes in metastatic colorectal cancer. *Eur. J. Surg. Oncol.* **44**, 975–982 (2018).
43. Luchini, C. *et al.* Liquid Biopsy as Surrogate for Tissue for Molecular Profiling in Pancreatic Cancer: A Meta-Analysis Towards Precision Medicine. *Cancers* **11**, (2019).

44. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/arXiv.1303.3997.
46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
48. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
49. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
50. Vessies, D. C. L. *et al.* Combining variant detection and fragment length analysis improves detection of minimal residual disease in postsurgery circulating tumour DNA of stage II–IIIA NSCLC patients. *Mol. Oncol.* **16**, 2719–2732 (2022-7).

Figure Legend

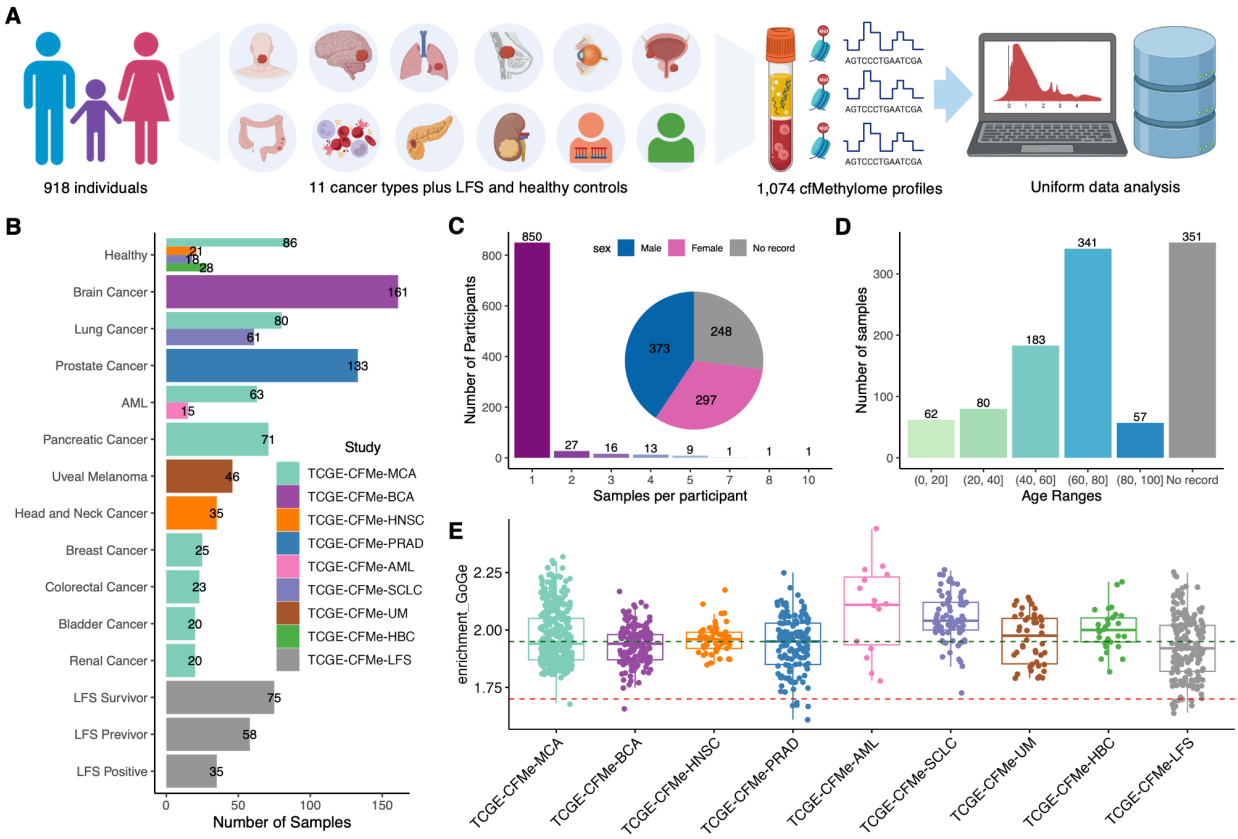


Figure 1: Collection, curation and quality control of cfMeDIP-seq datasets. A) Overview of the centralized cell-free methylome profiles encompassing multiple cancer types and healthy control samples, which was created with BioRender.com. **B)** Number of samples categorized by sample type, along with the breakdown of sample numbers from individual studies. **C)** Distribution of the number of samples per participant, accompanied with the sex composition for all participants. **D)** Number of participants in different age ranges. **E)** Distribution of samples' cfMeDIP-seq enrichment scores (, GoGe), grouped by study. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × interquartile range (IQR).

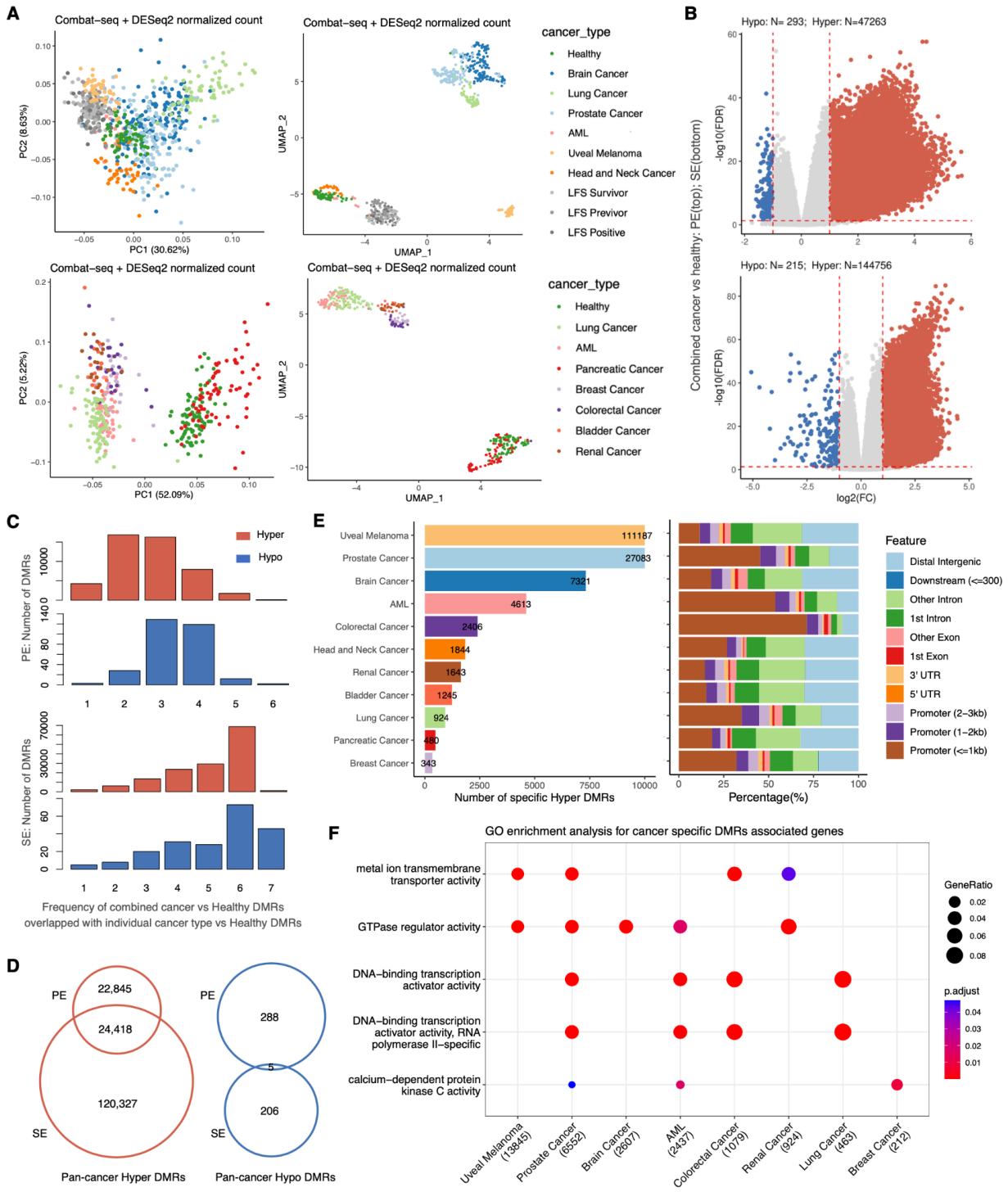


Figure 2: Pan-cancer and cancer-specific cell-free DNA methylation signatures. A) The PCA (left) and UMAP (right) plots using ComBat-seq + DESeq2 normalized count count for all PE samples (top) and SE samples (bottom), with colors indicating different sample types. **B)** Volcano plots for differentially methylated regions in the comparison of

combined cancer samples against the healthy control from PE (top) and SE (bottom) studies. **C)** Histogram of overlapping hyper- and hypomethylated regions between the comparisons of combined cancer vs health control and individual cancer vs healthy control for PE (top 2 rows) and SE (bottom 2 rows) studies. **D)** Venn plots for the overlapped pan-cancer hyper- (left) and hypomethylated DRMs (right) between PE and SE samples based comparisons. **E)** The number of cancer-specific hypermethylated regions, where the bars for uveal melanoma, AML and lung cancer have been scaled down to 10,000 along with exact count labeling (left), and the distribution of cancer-specific hypermethylated regions among the annotated genomic regions (right). **F)** The GO terms enriched for the cancer-specific hypermethylated regions associated genes.

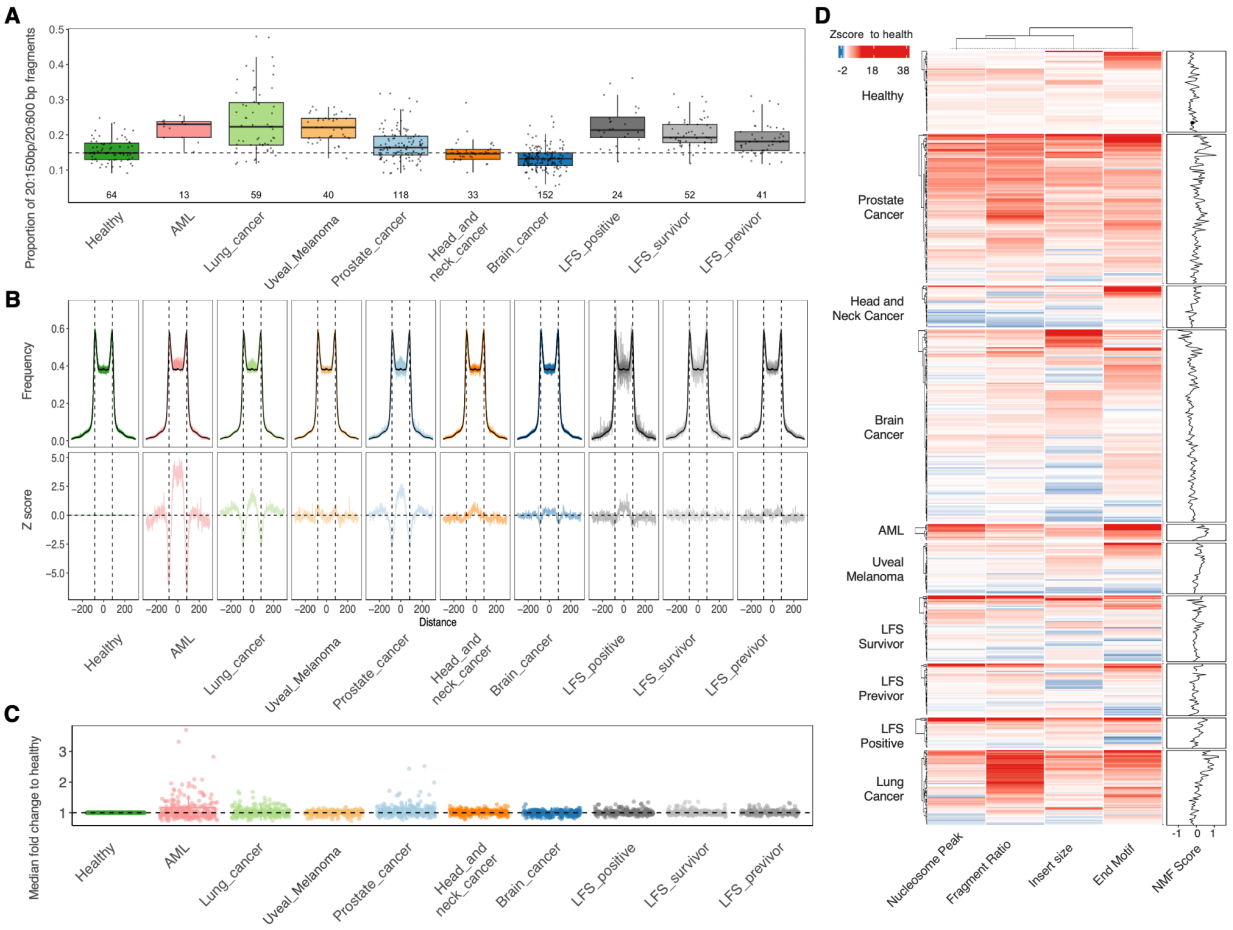


Figure 3: Overview of pan-cancer cell-free DNA fragmentomic features. A) The proportion of short fragments between 20:150bp/20:600bp in length across cancer types. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **B)** Top: line plot showing the difference between the median proportion of likely nucleosome-bound fragments (167 bp in length) from expected nucleosome positions in healthy blood (in black) and cancer samples (colored). Each colored line represents the proportion of 167 bp fragments which ended at that position from the nucleosome. Vertical dotted black lines represent the expected positions of fragments if they were correctly bound to a nucleosome (±83-84 bp from the middle of an expected nucleosome position). Bottom: z-scores calculated as the difference in fragment frequencies between cancer types and healthy controls at each position. **C)** Median fold changes of end motifs relative to median frequencies of healthy controls. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **D)** Heatmap showing the z-scores of different fragmentation features of

methylated fragments relative to healthy controls across cancer types. Right panel: the weighted fragmentation score, calculated using NMF. Larger differences from 0 indicate greater deviation from expected insert sizes in healthy controls.

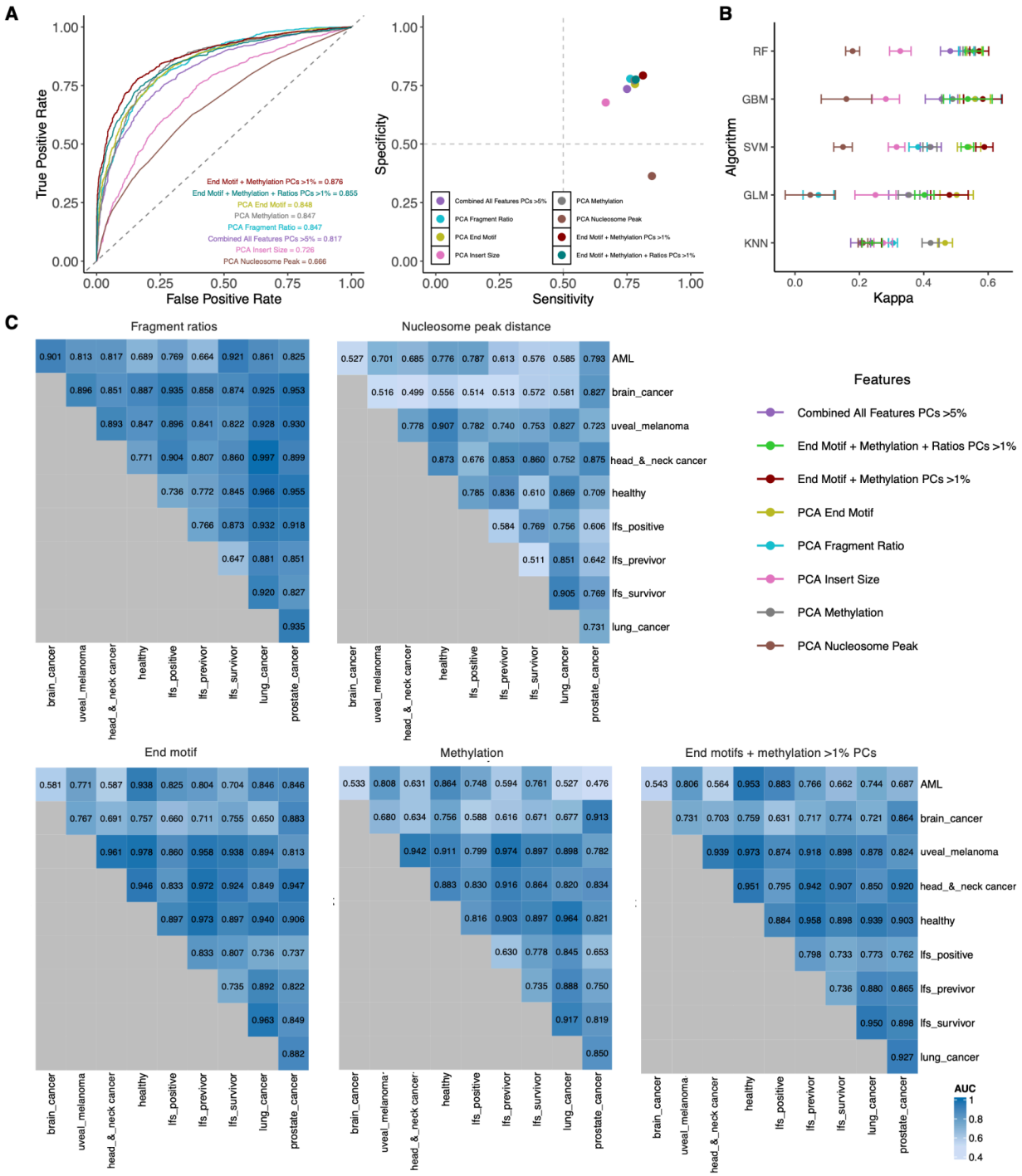


Figure 4: Pan-cancer cell-free DNA classification features. A) Left: ROC curves indicating classification accuracy for varying classification approaches based on methylation, individual fragmentation features of methylated fragments, and feature combinations. Right: Scatterplot showing specificity and sensitivity metrics for each classifier used. **B)** Plot showing the kappa score generated using different clustering

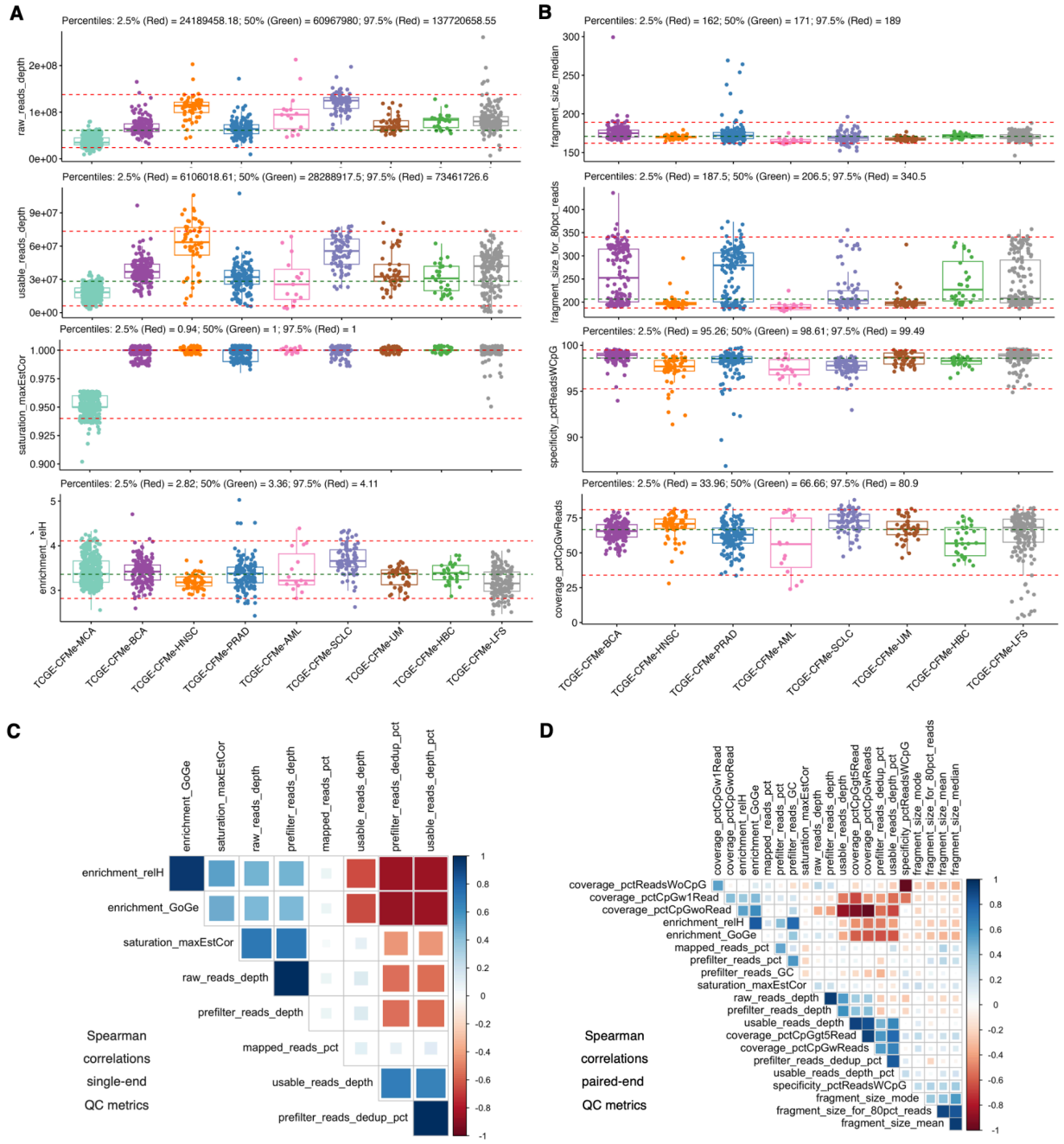
algorithms. The highest kappa score achieved was used for model training (results shown in A). **C)** Heatmap of AUCs for models trained between different cancer types and healthy controls, by fragmentation or methylation features.

Extend Data

A pan-cancer compendium of 1,074 plasma cell-free DNA methylomes and fragmentomes

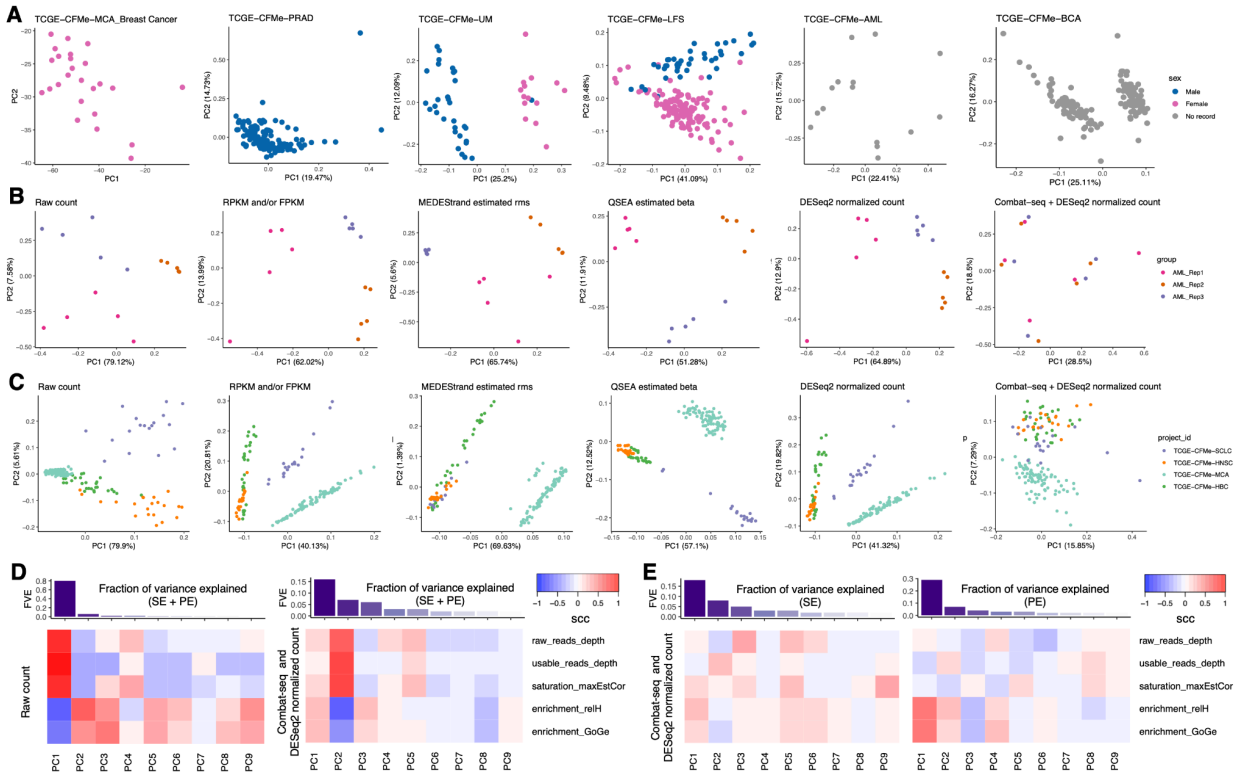
Yong Zeng[1, $, #], Dor D. Abelman[1, 2, $], Althaf Singhawansa[1], Nicholas Cheng[3], Emma Bell[1], Wenbin Ye[1], Sasha Main[1, 2], Ping Luo[1], Samantha L. Wilson[4], Eric Y. Stutheit-Zhao[1], Derek Wong[1], Nadia Znassi[1], Suluxan Mohanraj[1], Philip Awadalla[3], Benjamin H. Lok[1, 2], Michael M. Hoffman[1, 2, 5], Raymond H. Kim[1, 2, 6], Gelareh Zadeh[1, 7], Daniel De Carvalho[1, 2], Scott V. Bratman[1, 2], Mathieu Lupien[1, 2, 3, #], Trevor J. Pugh[1, 2, 3, #], Housheng Hansen He[1, 2, #]

Supplementary Figures



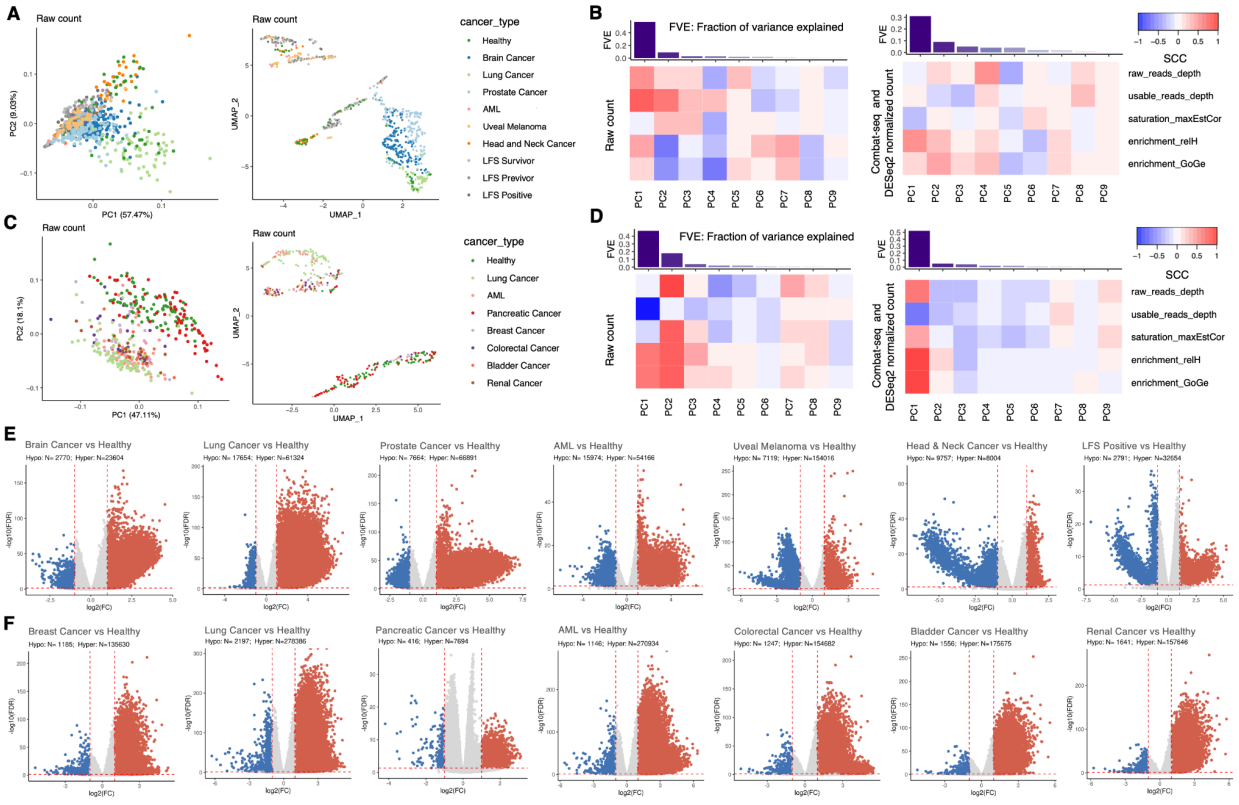
Supplementary Figure 1: Distribution and correlation of cfMeDIP-seq QC metrics. **A)** The distribution of raw read depth, usable read depth, saturation scores (maxEstCor), and enrichment score (relH) across all studies. The 2.5%, 50% and 97.5% percentiles are shown with dashed lines accordingly. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **B)** The distribution of

median fragment size, the fragment range that cover 80% of sequencing reads, cfMeDIP-seq specificity (measured by the percentage of reads with CpG), and cfMeDIP-seq coverage (measured by the percentage of human genome CpG sites covered by at least a single sequencing read). The SE samples were excluded in this figure. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **C)** The spearman correlation among the QC metrics within SE samples. **D)** The spearman correlation among the QC metrics within PE samples, including those fragmentomic QC metrics specific for PE sequencing data.



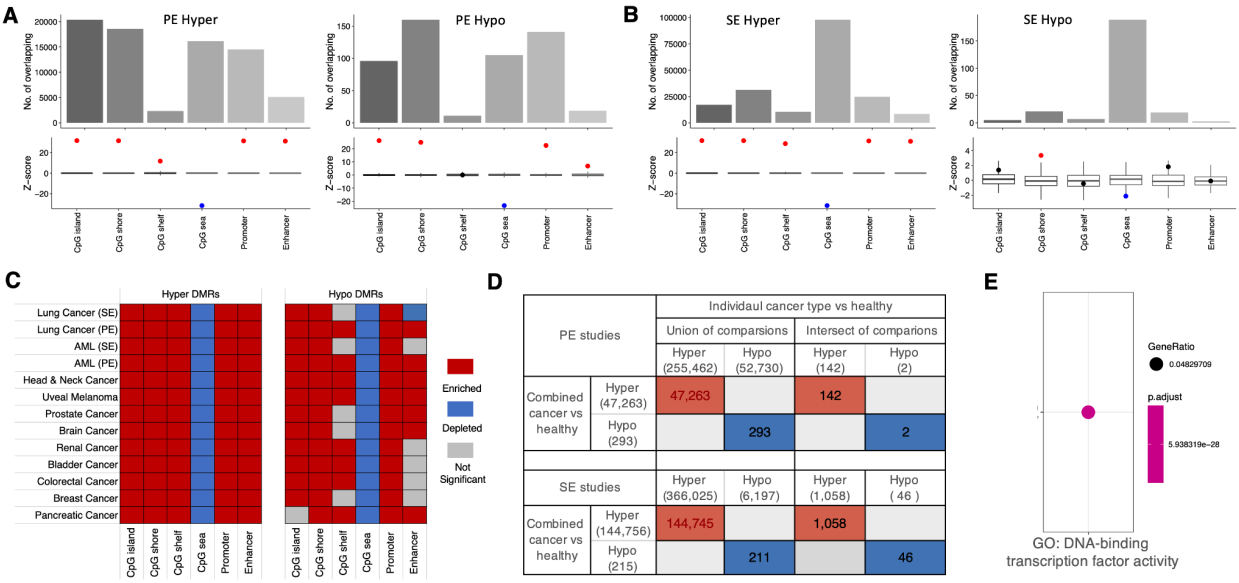
Supplementary Figure 2: Comparisons of methods for DNA methylation quantification and normalization. A) PCA plots based on the methylation signals on the chromosome X, including the studies with only females (TCGE-CFMe-MCA:breast cancer), only males (TCGE_CFMe-PRAD: Prostate cancer), mixed samples with known male and female labels (TCGE-CFMe-UM and TCGE-CFMe-LFS), and samples without avail sex labels (TCGE-CFMe-AML and TCGE-CFMe-BCA). **B)** PCA plots for the TCGE-CFMe-AML study, illustrating different outcomes for different DNA methylation

quantification and normalization methods, including raw read count, RPKM or FPKM, absolute methylation levels estimated by MEDEStrand and QSEA, normalized read count by DESeq2 without and with prior batch correction using ComBat-seq. Different colors indicate the replicates of samples. **C)** PCA plots showing outcomes for different DNA methylation quantification and normalization methods with all healthy control samples from four independent studies. **D)** Fractions of variance explained by the top 9 PCs and correlations between these top 9 PCs and 5 main QC metrics for combined SE and PE healthy control samples using raw read count (left) and ComBat-seq + DESeq2 normalized count (right). **E)** Fractions of variance explained by the top 9 PCs and correlations between these top 9 PCs and 5 main QC metrics for SE (left) and PE (right) samples using ComBat-seq + DESeq2 normalized count, respectively.

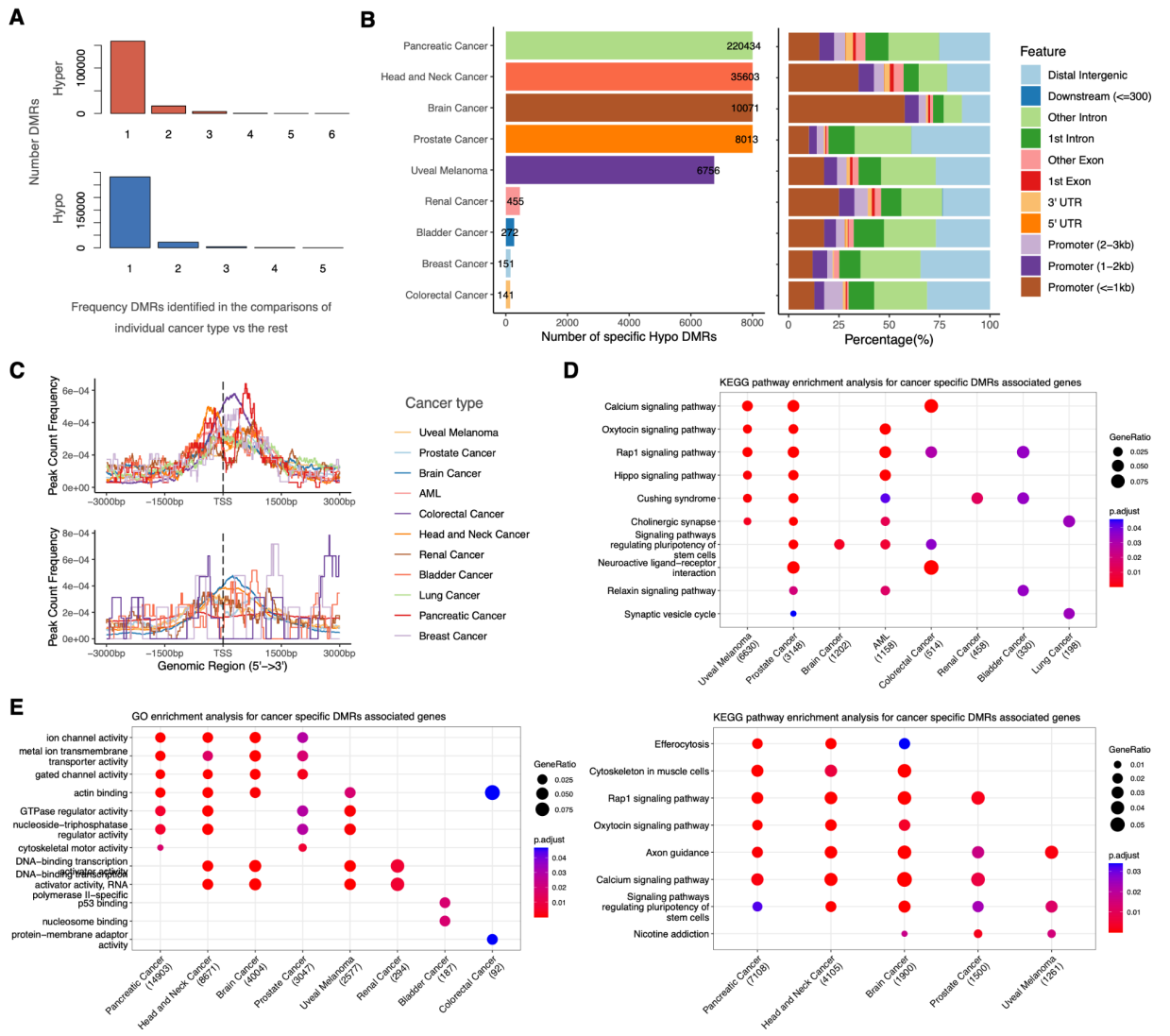


Supplementary Figure 3: Differential methylation analysis between individual cancer types against healthy control. The PCA (left) and UMAP (right) plots use raw read count for all PE samples (**A**) and SE samples (**C**), with colors indicating different

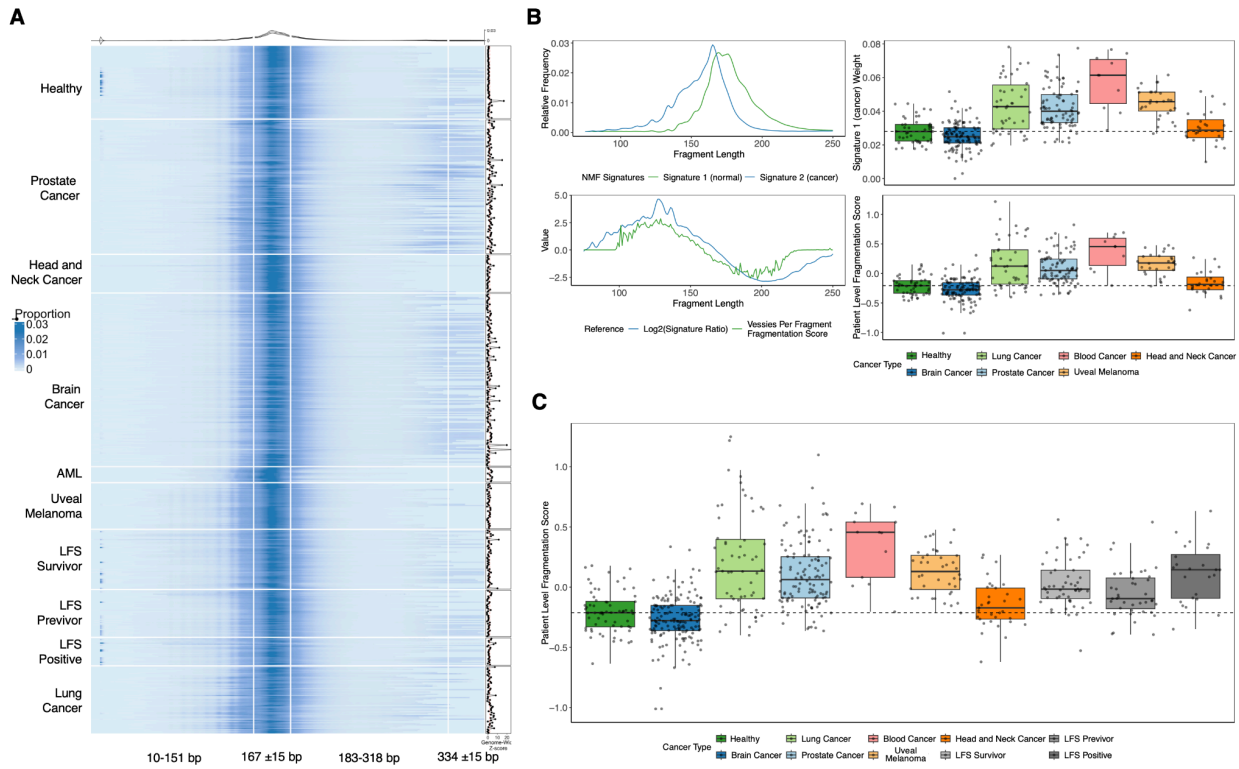
sample types. Fractions of variance explained by the top 9 PCs and correlations between these top 9 PCs and 5 main QC metrics using raw read count (left) and ComBat-seq + DESeq2 normalized count (right) for PE (**B**) and SE (**D**) samples, respectively. The volcano plots for differentially methylated regions in the comparison of each individual cancer type against the healthy control for PE studies (**E**) and SE studies (**F**).



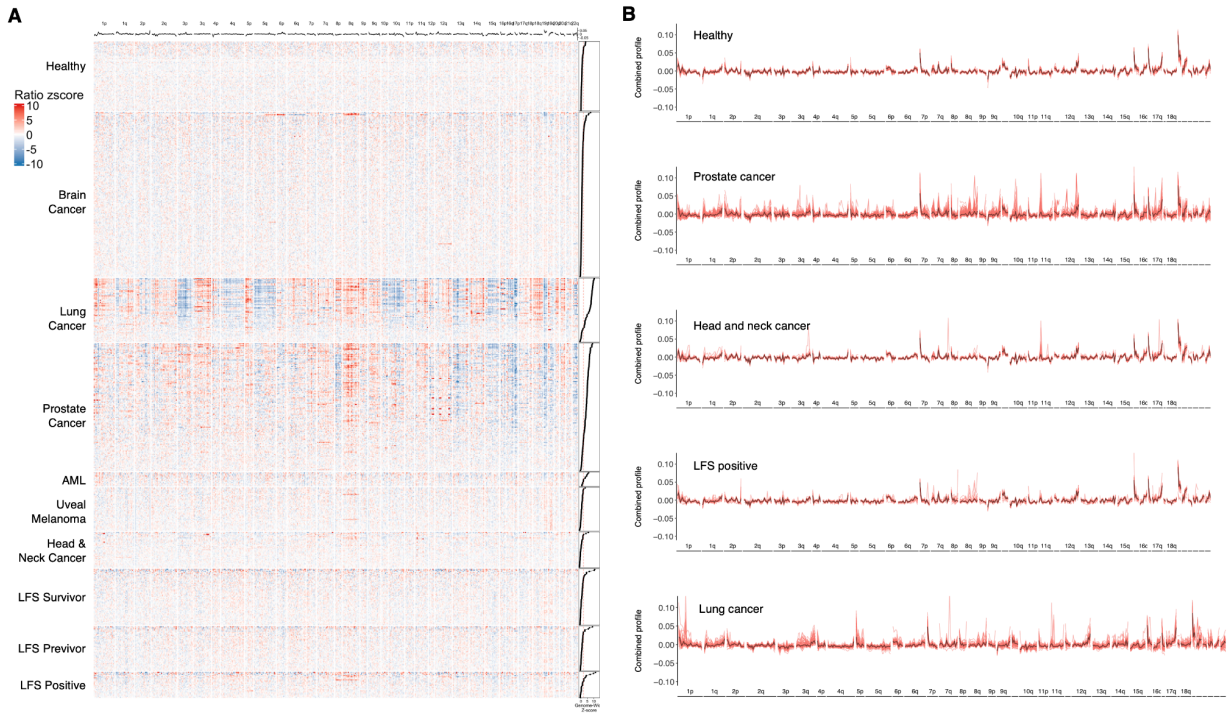
Supplementary Figure 4: Characterization of pan-cancer methylated regions. The enrichment analysis of hyper- (left) and hypomethylated DMRs, identified using PE (**A**) and SE (**B**) samples, in annotated CpG regions, as well as the promoter and enhancer regions. The top barplots show the number of corresponding DMRs within annotated regions, while the bottom plots depict the permutation test results, with red, blue, gray indicating DMRs enriched, depleted, and not significant difference within annotated regions, respectively. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **C)** The enrichment analysis for the hyper- (left) and hypomethylated (right) DMRs in annotate CpG regions, promoter and enhancer regions. **D)** Overlapping of DMRs identified in combined cancer vs health control and the union/common of DMRs across all individual cancer vs healthy control for PE (top 2 rows) and SE (bottom 2 rows) studies. **E)** GO term enriched for the pan-cancer hypermethylated regions associated genes.



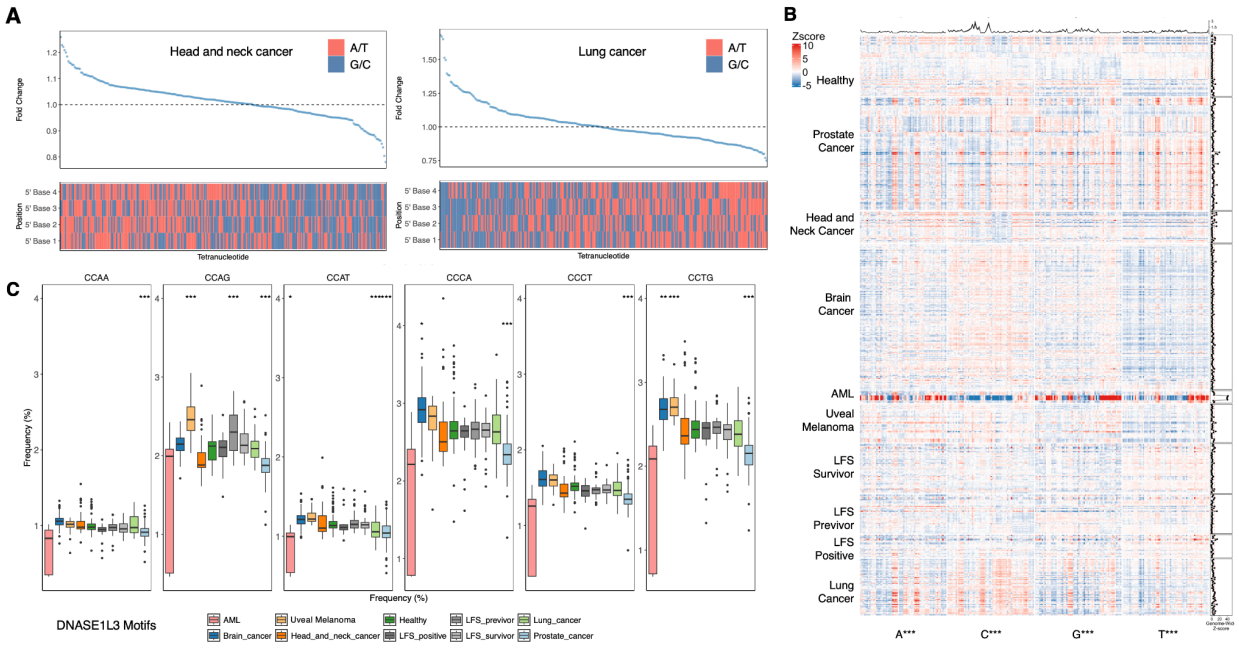
Supplementary Figure 5: Characterization of cancer-specific methylated regions. **A)** The histogram of frequency of cancer-specific hyper- (top) and hypomethylated (bottom) regions across the comparisons. **B)** The number of cancer-specific hypomethylated regions, where the bars for lung cancer and AML have been scaled down to 8,000 along with exact count labeling (left), as well as the distribution of corresponding regions among the annotated genomic regions (right). **C)** The distribution of cancer-specific hyper- (top) and hypomethylated (bottom) regions around the TSS. **D)** The KEGG terms enriched for the cancer-specific hypermethylated regions associated genes. **E)** GO (left) and KEGG (right) terms enriched for the cancer-specific hypomethylated regions associated genes.



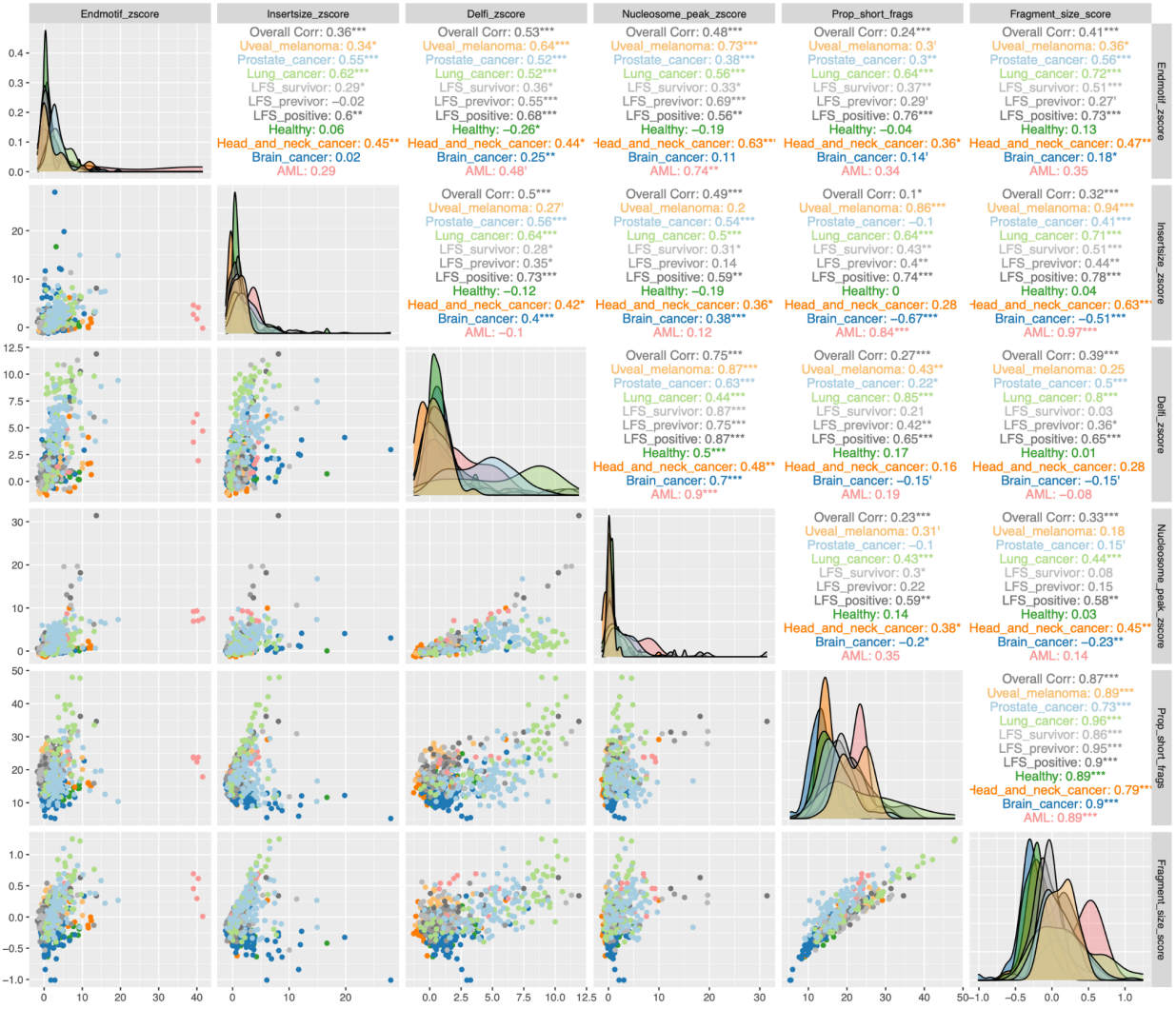
Supplementary Figure 6: Evaluation of insert sizes and weighted fragment length scores. A) Heatmap showing the proportion of fragments of various insert size lengths. The line plot on the top of the heatmap represents the median frequency in healthy controls with error bars. The points on the right of the heatmap represent a genome-wide z-score summed across each frequency. **B)** Comparison of NMF training on a 70% split of data. Top left: a line plot showing the relative frequency of fragment lengths between the two signatures that were discovered by NMF in a training cohort. Top right: Signature weights across cancer and healthy samples. Top right: the log change of the signature across fragment lengths, indicating the probability of a fragment being cancer or healthy across each length. Bottom right: A patient level score as determined by the proportion of short fragments multiplied by the probability of that value being cancer or healthy. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **C)** Patient level fragmentation scores across all samples. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR.



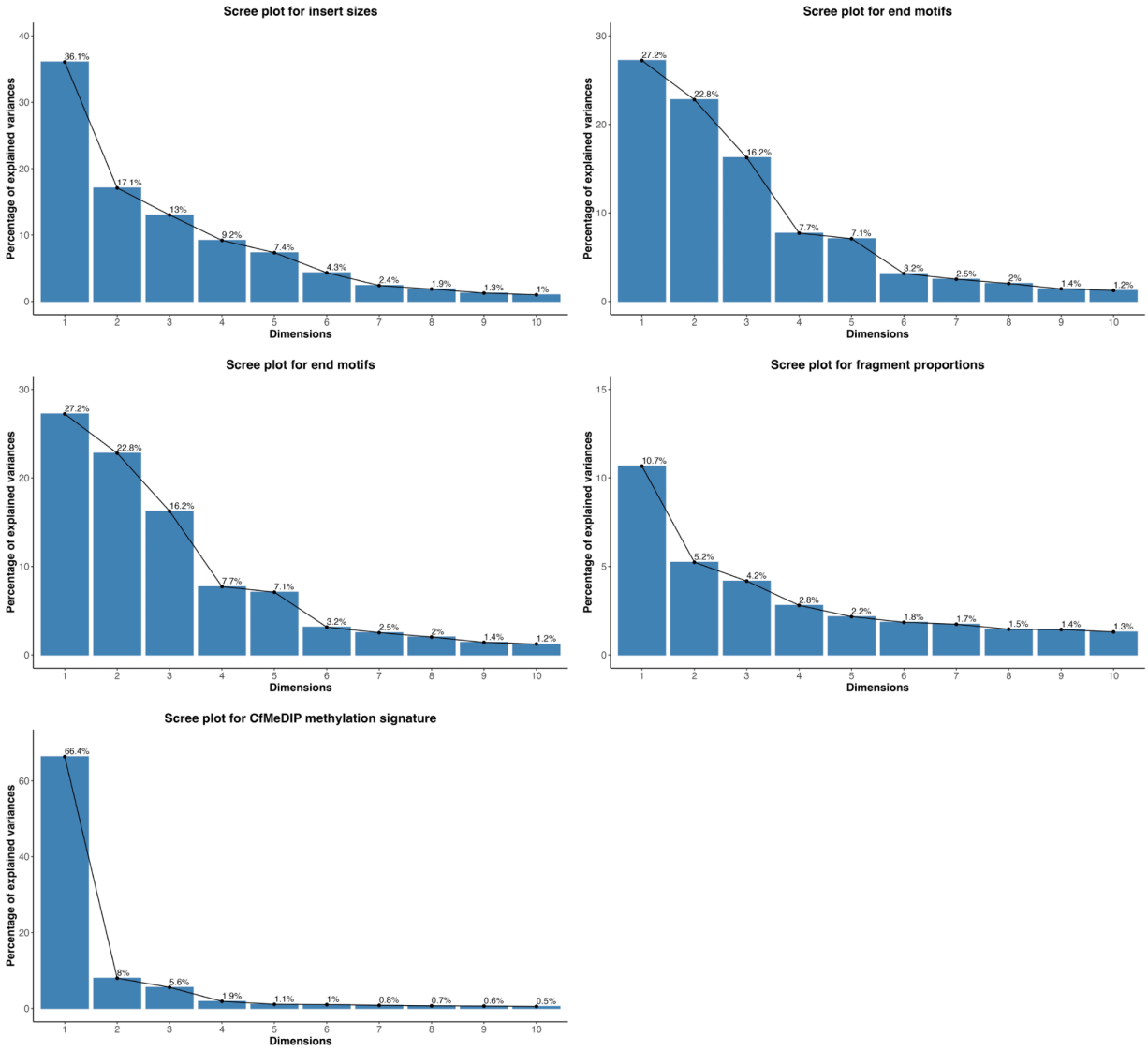
Supplementary Figure 7: Comparison across genome-wide short/long fragment ratios by 5 Mb bin. A) Heatmap showing z-score differences in the proportion of short/long fragments across 5 Mb bins between cancer and LFS types and healthy controls. The line plot on the top of the heatmap represents the median frequencies of short/long fragments in healthy controls. The points on the right of the heatmap represent a genome-wide z-score summed across each 5 Mb bin. Bins are ordered in order of genome positions from chromosome 1p to 22q. **B)** Line plot showing differences in short/long fragment profiles across sample groups. The black line represents the median proportion of short/long fragments at that bin. Each red line represents the proportion of short/long fragments in a sample. Values are multiplied by differences in coverage relative to the expected coverage at that position, with values that have a significant amount of short fragments and higher than expected coverage in healthy controls appearing higher.



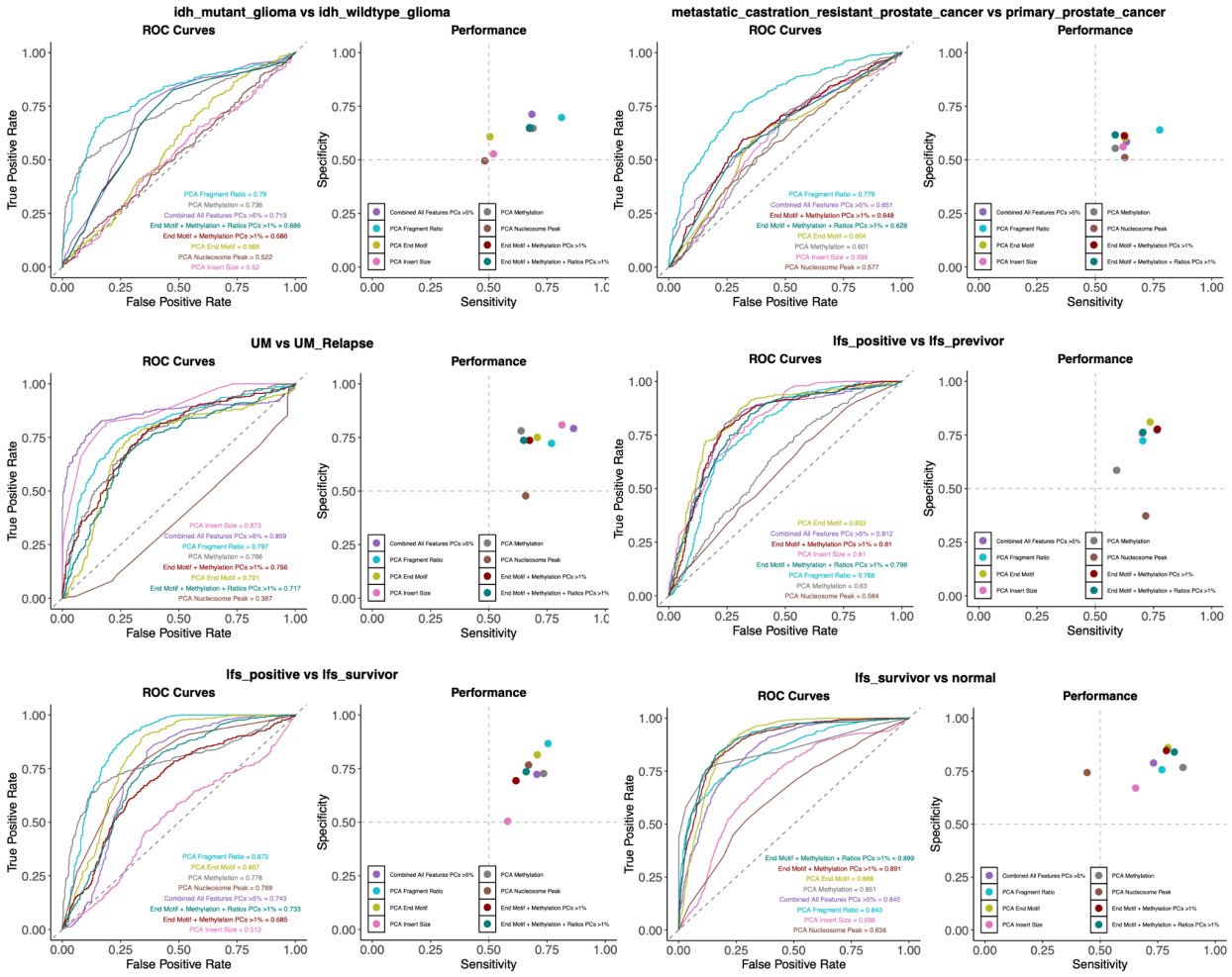
Supplementary Figure 8: Characterization of end motif frequencies across cancer types. A) Median fold changes between head & neck cancer vs healthy controls and lung cancers vs healthy controls. The specific end motif being explored is plotted underneath and colored by base as either A/T or C/G. **B)** Heatmap showing z-score differences in 5-prime 4-mer end motif frequencies between cancer and LFS types and healthy controls. The line plot on the top of the heatmap represents the median frequencies of end motifs in healthy controls. The points on the right of the heatmap represent a genome-wide z-score summed across each tetranucleotide combination between each sample type and healthy controls. End motifs are ordered by alphabetical order from AAAA to TTTT. **C)** End motif frequencies between healthy controls and cancer types at motifs associated with *DNASE1L3*. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR.



Supplementary Figure 9: Correlations between fragmentomic features across cancer types. Correlation matrix showing the Spearman correlation between different fragmentomic features across all high quality paired end samples (N = 596). Features include genome-wide z-scores for end motifs, insert sizes, fragment ratios, nucleosome peak distances, as well as the proportion of short fragments and fragmentation scores calculated by non-matrix factorization (NMF). Each scatter plot and density plot represents the relationship and distribution of fragmentomic features among cancer types. Correlation coefficients for each cancer type are displayed, with asterisks indicating statistical significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Spearman correlation was used since not all groups achieved the normality conditions required for Pearson correlation to be used.



Supplementary Figure 10: Principal component analysis (PCA) Scree plots for fragmentation and methylation features. Scree plots showing the percentage of explained variance by each principal component for different fragmentation features, including insert sizes, end motifs, methylation signatures, and fragment ratios. The scree plots are used to determine the number of principal components to retain for downstream analysis by showing the variance explained by each component. This is used to identify the most significant components that capture the majority of the variance in the dataset for dimensionality reduction and feature selection.



Supplementary Figure 11: Pan-cancer cell-free DNA classification features between cancer types. ROC curves and specificity/performance scatter plots demonstrating the classification accuracy between cancer types and subtypes based on fragmentation features using the top performing from either random forests, generalized linear models, gradient boosting machines, support vector machines with a radial basis function kernel, and k-nearest neighbors classification algorithms. This analysis includes comparisons of IDH mutant glioma vs. IDH wildtype glioma, metastatic castration-resistant prostate cancer vs. primary prostate cancer, uveal melanoma primary vs relapse cases, and comparisons between LFS positive, previvor, and survivors. The performance of classifiers is shown as AUCs achieved for each comparison on the bottom right of the ROC plot. Different color lines represent different classifiers of features used in the analysis.

Supplementary Tables and Data

Supplementary Table 1. Meta information for centralized and curated cfMeDIP-seq samples.

Supplementary Table 2. Genome-wide distribution of insert sizes.

Supplementary Table 3. Genome-wide Z-scores per fragmentomic feature, proportion of short fragments, and weighted fragment scores.

Supplementary Table 4. Proportions of short / long fragment ratios by 5Mbp bins.

Supplementary Table 5. Distance of likely nucleosome-bound (167bp) fragments to expected nucleosome positions.

Supplementary Table 6. Genome-wide proportions of end motifs.

Supplementary Table 7. AUCs for cancer type classification models.

Supplementary Data 1. Bed files for pan-cancer and cancer-specific DMRs.