nature cancer

Resource

https://doi.org/10.1038/s43018-026-01116-3

A pan-cancer compendium of 1,294 plasma cell-free DNA methylomes and fragmentomes enabling multicancer detection

Received: 6 July 2024

Accepted: 9 January 2026

Published online: 19 February 2026

Check for updates

Yong Zeng[1,13] ✉, Dor D. Abelman[1,2,13], Althaf Singhawansa[1], Nicholas Cheng[3], Yuanchang Fang[1,2], Sasha C. Main[1,2], Emma Bell[3], Wenbin Ye[1], Ping Luo[1], Samantha L. Wilson[4], Eric Y. Stutheit-Zhao[1], Derek Wong[1,5], Nadia Znassi[1], Kui Chen[1,6], Suluxan Mohanraj[1], Enrique Sanz-Garcia[1,7], Faiyaz Notta[1,2,3], Anand Ghanekar[6,8], Philip Awadalla[9], Benjamin H. Lok[1,2], Michael M. Hoffman[1,2,10], Raymond H. Kim[1,3,11], Gelareh Zadeh[1,12], Daniel D. 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] ✉

Cell-free DNA analysis via methylation and fragmentation profiling has advanced minimally invasive cancer detection; however, broader application has been limited by small cohorts and inconsistent data processing. Here we collated 1,074 cfMeDIP-seq profiles across 9 studies, comprising cancer samples from 11 cancer types, carriers of Li-Fraumeni syndrome and healthy controls. We developed a uniform computational workflow to mitigate technical and biological confounders across cohorts. This analysis identified 14,202 pancancer differentially methylated regions for cancer detection, along with cancer-specific markers for subtype monitoring. Fragmentomic profiling revealed distinguishing differences in 5′ end motifs, fragment lengths and nucleosome footprints across cancers. Integrating methylome and fragmentome features enhanced cancer detection and classification. Validation in 220 independent samples, including 3 cancer types absent from the primary dataset, confirmed the robustness of our findings. Altogether, this work provides a pancancer cell-free DNA resource of 1,294 samples to support future methylome and fragmentome studies.

Cell-free DNA (cfDNA) present in blood plasma has emerged as a promising analyte for cancer prognosis and treatment monitoring due to its noninvasive nature[1]. In 2018, Shen et al. developed the cfDNA 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 fragmentomic 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 the pancancer 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, colorectal cancer, breast cancer, lung cancer, renal

A full list of affiliations appears at the end of the paper. ✉e-mail: Yong.Zeng@uhn.ca; Mathieu.Lupien@uhn.ca; Trevor.Pugh@utoronto.ca; Hansen.He@uhn.ca

cancer, bladder cancer and acute myeloid leukemia (AML)[2]. With the evolution of cfMeDIP-seq protocols, especially the shift to paired-end (PE) sequencing, cfDNA methylome profiles have since been delineated for several other cancer types, including brain cancers[8], head and neck cancer[9], primary and metastatic prostate cancer[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]; however, the isolation of these valuable datasets within individual research groups, coupled with diverse analytical workflows and scientific goals, has posed substantial challenges for pancancer cfDNA methylome and fragmentome exploration.

In this study, we initially compiled cfMeDIP-seq profiles derived from 1,074 plasma samples, spanning 11 major cancer types, LFS and healthy controls. These datasets were uniformly processed and quantified with correction and normalization to minimize technical biases and enable a comprehensive analysis of pancancer and cancer-specific cfDNA methylation and fragmentomic features. We further evaluated the capability of these features to differentiate cancerous and healthy samples, as well as among different cancer types. To validate our findings, we incorporated an independent set of 220 cfMeDIP-seq samples, encompassing five cancer types, three of which were not presented in the initial dataset, and healthy controls[14,15]. Our study presents the largest compiled multicancer plasma cfMeDIP-seq dataset (n = 1,294) to date, establishing a foundational resource for advancing methods aimed at detecting cancer inception and monitoring disease evolution.

Results

Pancancer 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 and Epigenetics (TCGE) project: TCGE-CFMe-MCA[2], TCGE-CFMe-BCA[8], TCGE-CFMe-HNSC[9], TCGE-CFMe-PRAD[10], TCGE-CFMe-AML[16], TCGE-CFMe-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 (n = 753), as well as from patients with hereditary LFS (n = 168), categorized into LFS-survivor (cancer-negative individuals with a cancer history), LFS-previvor (cancer-negative individuals without a cancer history) and LFS-positive (those with one or multiple cancer types) (Fig. 1a,b and Supplementary Table 1).

These samples originated from a total of 918 participants, 68 of whom had multiple samples collected at different time points and/or health statuses (Fig. 1c). The participants included 390 male donors and 305 female donors. The remaining 223 individuals lacked study-reported sex information (Fig. 1c and Supplementary Table 2). The ages at the first time of blood draw, recorded for 674 participants, ranged from 1 to 92 years, with a median age of 64 years (Fig. 1d and Supplementary Table 2). Each sample was assigned a unique ID that includes information about the source of a specific study, participant disease status (cancer type, primary versus metastatic cancer or healthy control), and the timing order of each sample among multiple samples for the same participant (Supplementary Table 1).

All cfMeDIP-seq data for 1,074 plasma samples were processed uniformly with a standardized pipeline, MEDIPIPE[17]. Libraries were sequenced with a median of ~61 million raw sequencing reads and about 28 million unique reads after quality control (QC) trimming and duplication removal (Extended Data Fig. 1a). Except for the TCGE-CFMe-MCA study, which utilized single-end (SE) sequencing, all samples were subjected to PE sequencing. The PE reads displayed a modal fragment length of 167 base pairs (bp), consistent with the expected size for nucleosome-associated cfDNA (Extended Data Fig. 1b). Additionally, a median of 98.61% of PE reads contained CpG site(s), covering a median of 66.66% of CpG sites across the human genome (Extended Data Fig. 1b).

We observed a strong positive correlation between the enrichment scores GoGe and relH, which indicates the enrichment of CpGs within sequencing reads compared to the reference genome[18] (Extended Data Fig. 1c,d); however, both scores displayed a negative correlation with the fragment size in PE profiles (Extended Data Fig. 1d), 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[19] (Fig. 1e and Extended Data Fig. 1a). For downstream analysis, we focused on 974 high-quality samples (SE, 378; PE, 596) that met all QC criteria (Extended Data Fig. 1e).

Evaluation and mitigation of technical and biological confounding effects in cfDNA methylation signature identification

To mitigate the technical and biological confounding effects, we focused our analyses on autosomes and excluded the ENCODE blacklist regions (Extended Data Fig. 2a)[20]. Among six tested DNA methylation quantification and normalization strategies (Methods), the ComBat-seq + DESeq2 method most effectively reduced batch effects, especially when SE and PE samples were processed separately (Extended Data Fig. 2b–e); however, SE and PE samples showed intrinsic differences beyond sequencing, as confirmed by mimicked and original SE samples comparison (Extended Data Fig. 2f), which also influenced differentially methylated regions (DMRs) identification (Extended Data Fig. 2g). We also found that samples collected at multiple time points from the same participant affected DMR identification due to individual variability[21] (Extended Data Fig. 2h). Therefore, ComBat-seq + DESeq2 was applied separately to SE and PE samples (Fig. 2a and Extended Data Fig. 3a–d), and DMR identification was restricted to baseline time point samples for both SE and PE datasets (Extended Data Fig. 1e, labeled as 'A' in the vial column of Supplementary Table 1). Uniformly processed data from sequential samples were provided as a resource.

As age and sex have also been reported to be associated with DNA methylation alterations in cancer[22–25], we compared the DMRs identified with and without these factors as confounders across cancer type to healthy controls (Extended Data Fig. 3e–i and Methods). In the LFS cohort, 78.0% of hyper-DMRs and 51.7% of hypo-DMRs remained consistent after accounting for age and sex (Extended Data Fig. 3h and Supplementary Table 3). Across other cancer types, an average of 93.5% (range 81.0–98.9%) of hyper-DMRs and 85.8% (range 72.2–98.8%) of hypo-DMRs were more consistently identified after adjustment (Extended Data Fig. 3h,i and Supplementary Table 3). Additionally, DMRs identified without adjustment showed minimal overlap with age- (an average of 0.48% of hyper-DMRs and 0.40% of hypo-DMRs) and sex-associated (an average of 0.13% of hyper-DMRs and 0.17% of hypo-DMRs) signatures defined from this dataset and previous studies[26,27] (Methods and Supplementary Table 4 and 5), indicating these effects are largely controlled when age distributions are matched. To further mitigate peripheral blood leukocyte (PBL) contamination[9,11], we utilized a modified in-silico PBL depletion strategy using 20 healthy PBL samples (Methods and Extended Data Fig. 3j). We focused on DMRs that overlapped with PBL-depleted bins (n = 2,128,687), keeping on average 22.9% (range 2.4–67.8%) of hyper-DMRs and 19.4% (range 1.1–38.7%) of hypo-DMRs across individual cancer type versus healthy control (Supplementary Table 6), excluding DMRs overlapping blood- or endothelial cell-associated regions[28] 0.52% (range 0.16–1.36%) hyper-DMRs and 0.88% (range 0–9.09%) hypo-DMRs (Supplementary Table 7).

https://doi.org/10.1038/s43018-026-01116-3

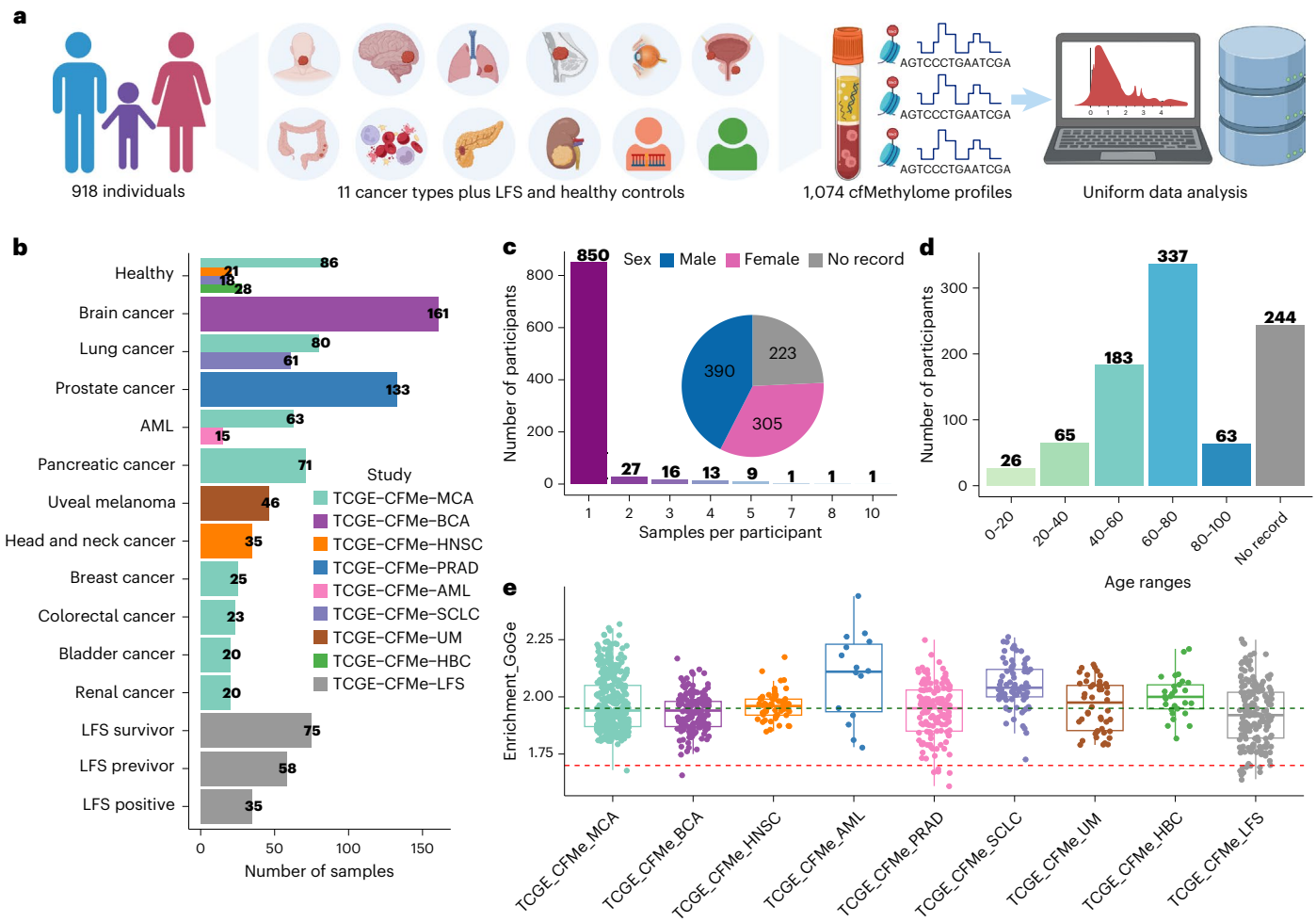


Fig. 1 | Collection, curation and quality control of cfMeDIP-seq datasets. **a**, Overview of the centralized cell-free methylome profiles encompassing multiple cancer types and healthy control samples. **b**, Number of samples categorized by sample type, along with the breakdown of sample numbers from individual studies ($n_{\text{Healthy\_TCGE-CFMe-MCA}} = 86$, $n_{\text{Healthy\_TCGE-CFMe-HNSC}} = 21$, $n_{\text{Healthy\_TCGE-CFMe-SCLC}} = 18$, $n_{\text{Healthy\_TCGE-CFMe-HBC}} = 28$, $n_{\text{Brain cancer}} = 161$, $n_{\text{Lung cancer\_TCGE-CFMe-MCA}} = 80$, $n_{\text{Lung cancer\_TCGE-CFMe-SCLC}} = 61$, $n_{\text{Prostate cancer}} = 133$, $n_{\text{AML\_TCGE-CFMe-MCA}} = 63$, $n_{\text{AML\_TCGE-CFMe-AML}} = 15$, $n_{\text{Pancreatic cancer}} = 71$, $n_{\text{Uveal melanoma}} = 46$, $n_{\text{Head and neck cancer}} = 35$, $n_{\text{Breast cancer}} = 25$, $n_{\text{Colorectal cancer}} = 23$, $n_{\text{Bladder cancer}} = 20$, $n_{\text{Renal cancer}} = 20$, $n_{\text{LFS survivor}} = 75$, $n_{\text{LFS previvor}} = 58$, $n_{\text{LFS positive}} = 35$). **c**, Distribution of the number of samples per participant (participants: $n_1 = 850$, $n_2 = 27$, $n_3 = 16$, $n_4 = 13$, $n_5 = 9$, $n_7 = 1$, $n_8 = 1$, $n_{10} = 1$), accompanied with the sex composition for all participants ($n_{\text{Male}} = 390$, $n_{\text{Female}} = 305$, $n_{\text{No record}} = 223$). **d**, Number of participants in different age ranges ($n_{(0,20)} = 26$, $n_{(20,40)} = 65$, $n_{(40,60)} = 183$, $n_{(60,80)} = 337$, $n_{(80,100)} = 63$, $n_{\text{No record}} = 244$). **e**, Distribution of samples' cfMeDIP-seq enrichment scores (GoGe), grouped by study ($n_{\text{TCGE-CFMe-MCA}} = 388$, $n_{\text{TCGE-CFMe-BCA}} = 161$, $n_{\text{TCGE-CFMe-HNSC}} = 56$, $n_{\text{TCGE-CFMe-PRAD}} = 133$, $n_{\text{TCGE-CFMe-AML}} = 15$, $n_{\text{TCGE-CFMe-SCLC}} = 79$, $n_{\text{TCGE-CFMe-UM}} = 46$, $n_{\text{TCGE-CFMe-HBC}} = 28$ and $n_{\text{TCGE-CFMe-LFS}} = 168$). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × interquartile range (IQR). Panel **a** created in BioRender; Zeng, Y. https://biorender.com/9bjvp4l (2025).

In summary, to maximize sample use while minimizing confounders, we initially identified pancancer and cancer-specific DMRs using DESeq2 without adding age and sex as covariates, restricting the analysis to the baseline time point for SE and PE samples separately. We then filtered out DMRs that overlapped with age- and sex-associated signatures, non-PBL-depleted regions and regions linked to blood and endothelial cells.

Pancancer and cancer-specific cfDNA methylation signatures

To establish a pancancer cfDNA methylation signature, we first identified 24,301 and 19,289 filtered hyper-DMRs, and 196 and 38 filtered hypo-DMRs by comparing the combined cancer samples against healthy samples from PE and SE studies, respectively (Fig. 2b and Methods). The majority of these DMRs were reproducibly detected in multiple cancer versus healthy comparisons, with only a few excluded due to lack of recurrence (Extended Data Fig. 4a–c). Hyper-DMRs greatly outnumbered hypo-DMRs in both combined and individual cancer types compared to healthy controls (Fig. 2b and Supplementary Table 8). Additionally, we found that both hyper- and hypo-DMRs were consistently enriched in CpG islands, shores and promoters (Extended Data Fig. 4d,e). Hyper-DMRs were also enriched in enhancers, whereas hypo-DMRs showed variable enhancer enrichment (Extended Data Fig. 4d,e). This aligns with previous reports that cfMeDIP-seq preferentially captures high-CpG-density regions, often hypermethylated in cancer[2,29,30].

We then focused on the 14,202 common hyper-DMRs between the PE and SE studies as the pancancer cfDNA methylation signature (Fig. 2c). Of these, 99.8% appeared in multiple cancer types, and 66.4% in at least four (Fig. 2d). On average, 53.2% of the hyper-DMRs identified in individual cancer type comparisons against healthy controls overlapped with this signature (Extended Data Fig. 4f and Supplementary Table 9). Furthermore, the common hyper-DMRs tend to have significantly higher fold change and lower *P* values compared to PE- or SE-unique DMRs (Fig. 2e, one-sided Wilcoxon rank-sum test: all *P* values $< 2.2 \times 10^{-16}$). Then, we performed functional enrichment analysis by focusing on 7,698 (54.2%) regions located within the promoter

regions (Extended Data Fig. 4g,h and Methods). Top-ranked Gene Ontology (GO) terms included DNA-binding transcription factor (TF) activator and repressor activities (Fig. 2f and Supplementary Table 10), consistent with hypermethylation disrupting TF binding in cancer[29], and gated channel activity, reflecting role in cancer cell proliferation and metastasis[31]. Top-enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, including neuroactive ligand–receptor interaction, calcium and cAMP signaling pathways (Fig. 2f and Supplementary Table 10), were also implicated in various cancers[32–36]. Together, these findings confirm that this signature represents widespread cancer-associated cfDNA methylation changes.

Beyond the pancancer cfDNA methylation signature, we also identified cancer-specific hypermethylated and hypomethylated signatures by comparing each cancer type against all other samples in SE and PE studies, respectively (Fig. 2g,h and Methods). We focused on those hyper-DMRs (67.7%) and hypo-DMRs (64.3%) that were uniquely detectable in individual cancer types (Extended Data Fig. 4i). Prostate cancer and uveal melanoma had the most specific hyper-DMRs, while head and neck and brain cancers had the most hypo-DMRs (Fig. 2g,h). Additionally, on average, 57.1% of cancer-specific hyper-DMRs and 47.7% of hypo-DMRs were located in the promoter region (Fig. 2g,h) and enriched around the transcription start site (Extended Data Fig. 4j). Furthermore, we consistently found that prostate cancer and lung cancer-specific hyper-DMRs, as well as head and neck, brain and pancreatic cancer-specific hypo-DMRs, were associated with the 'DNA-binding TF activity' (Fig. 2i,j). We also observed that cancer-specific cfDNA methylation signatures enriched in distinct GO terms or KEGG pathways (Fig. 2i,j, Extended Data Fig. 4k,l and Supplementary Table 11). For instance, colorectal hyper-DMRs linked to tyrosine kinase activity and PI3K-Akt signaling, bladder hypo-DMRs to (metallo)aminopeptidase activity and brain hypo-DMRs to axon guidance and neuroactive ligand–receptor pathways (Fig. 2i,j and Extended Data Fig. 4k,l). These results highlight cfDNA methylation heterogeneity across cancer types.

Fragmentomic features of methylated cfDNA across different cancers

More recently, fragmentomics[37], which probes the fragmentation patterns of cfDNA, has emerged as a promising tool for cancer detection[3–7,38]. 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 motifs, for 473 PE samples at the baseline time point (Methods).

Insert sizes showed the expected nucleosomal peak (~167 ± 15 bp) (Extended Data Fig. 5a and Supplementary Table 12). The proportion of short fragments (20–150 bp within 20–600 bp) differed by cancer type (two-sided Kruskal–Wallis test: P value < 2.2×10^{-16}) and was broadly elevated relative to healthy; head and neck and brain were the main exceptions (two-sided Dunn's post hoc test, P values ranging from 0.0222 to 1.45×10^{-7}) (Fig. 3a and Supplementary Table 13). To summarize length patterns, we factorized insert size distributions into two reference profiles using non-negative matrix factorization (NMF) and computed a weighted fragment score (FS) measuring similarity to the cancer-associated profile. Using this method, AML, lung cancer, prostate cancer, uveal melanoma and LFS-positive demonstrated a significantly higher degree of resemblance to the cancer-associated fragment insert size profile and a higher weighted FS compared to the healthy-associated profile (two-sided Dunn's post hoc test, P values ranging from 0.003 to 9.58×10^{-10}) (Extended Data Fig. 5b,c and Supplementary Table 13).

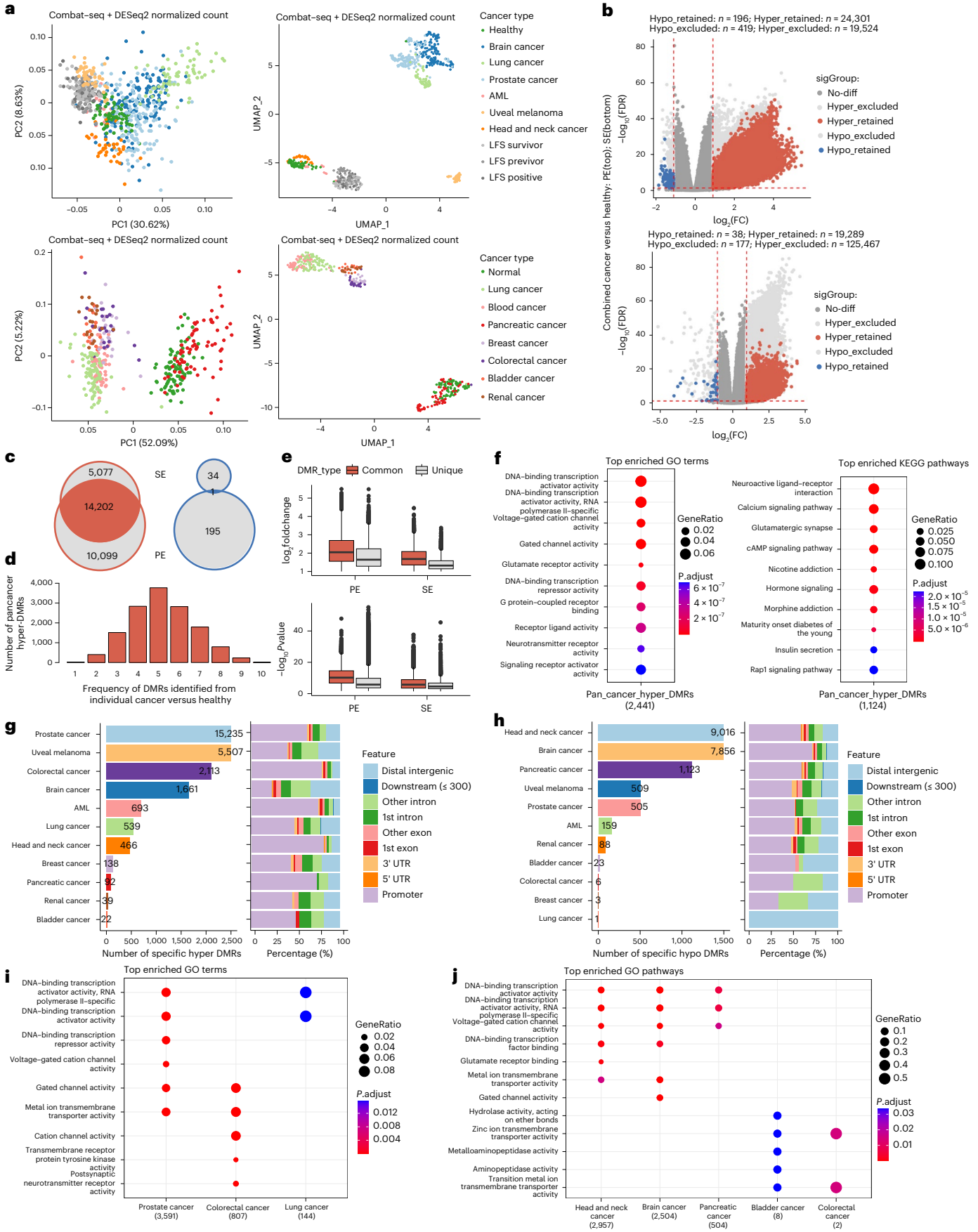
Genome-wide fragment ratios (5-Mb bins; DELFI-style[4]) varied more across the genome in cancers than in healthy samples (Extended Data Fig. 5d,e, Supplementary Table 14 and Methods). By this measure, the majority of AML (>80%), lung (77%) and prostate (73%) samples were distinguishable from healthy (Supplementary Table 13 and Methods). Subsequently, we conducted nucleosome footprinting analysis, which evaluates the proximity of likely nucleosome-bound 167 bp fragments to reference nucleosome positions (nucleosome peak distances)[39] (Methods). Nucleosome footprinting showed increased cutting within nucleosome cores, consistent with altered chromatin accessibility; all AML and 54% of prostate samples deviated from healthy medians (Fig. 3b and Supplementary Table 15).

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). Head and neck tumors were enriched for A/T-rich ends, whereas lung showed more C/G-rich ends relative to healthy samples (Extended Data Fig. 6a and Supplementary Table 16). Using z-scores versus healthy, 5′ end-motif profiles were significantly altered in most samples within AML (80%), lung (76%) and prostate (68%) (Extended Data Fig. 6b and Supplementary Table 13). Motifs associated with DNASE1L3 cleavage[40,41] (linked to methylated DNA) shifted by cancer group: they were depleted in prostate, lung and LFS cases, and enriched in uveal melanoma, brain and LFS-previvors (Extended Data Fig. 6c). Consistently, TCGA showed *DNASE1L3* downregulation in several tumor types including lung and prostate, with no difference in glioblastoma (GBM) (Extended Data Fig. 6d). Motif abundance inversely correlated with tumor cfDNA load ($R = -0.396$, $P = 3.54 \times 10^{-19}$) (Extended Data Fig. 6e), supporting a link between DNASE1L3 activity and tumor-derived cfDNA content.

To assess confounding, we z-scored each feature versus healthy PE and fit linear models with age, sex and cancer type. Age was not significant after accounting for cancer type; sex effects were small and largely limited to fragment ratios (Extended Data Fig. 6f, Supplementary Table 13 and Methods). These findings suggest that

Fig. 2 | Pancancer and cancer-specific cell-free DNA methylation signatures. **a**, The PCA (left) and UMAP (right) plots using ComBat-seq + DESeq2 normalized count for all PE samples (top) and SE samples (bottom), with colors indicating different sample types. **b**, Volcano plots for DMRs in the comparison of combined cancer samples against the healthy controls from PE (top) and SE (bottom) studies after filtering. **c**, Venn plots for the overlapped hyper-DMRs (left) and hypo-DMRs (right) between PE and SE samples based on comparisons of combined cancer samples versus healthy controls. **d**, Histogram of common hyper-DMRs identified in **c** across individual cancer versus healthy control comparisons in both PE and SE studies. The union of DMRs was counted for the overlapped cancer types, including AML and lung cancer. **e**, Box plots comparing methylation levels fold change (top) and P values (bottom) between the common hyper-DMRs (n = 14,202) and either PE- (n = 10,099) or SE-unique (n = 5,077) hyper-DMRs using one-sided Wilcoxon rank-sum test. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **f**, Top ten enriched GO terms (left) and KEGG pathways (right) (hypergeometric test with Benjamini–Hochberg-adjusted P values) for genes associated with pancancer hyper-DMRs located in promoter regions. **g**,**h**, The number of cancer-specific hyper-DMRs (**g**: $n_{\text{Prostate cancer}}$ = 15,235, $n_{\text{Uveal melanoma}}$ = 5,507, $n_{\text{Colorectal cancer}}$ = 2,113, $n_{\text{Brain cancer}}$ = 1,661, n_{AML} = 693, $n_{\text{Lung cancer}}$ = 539, $n_{\text{Head and neck cancer}}$ = 466, $n_{\text{Breast cancer}}$ = 138, $n_{\text{Pancreatic cancer}}$ = 92, $n_{\text{Renal cancer}}$ = 39, $n_{\text{Bladder cancer}}$ = 22) and hypo-DMRs (**h**: $n_{\text{Head and neck cancer}}$ = 9,016, $n_{\text{Brain cancer}}$ = 7,856, $n_{\text{Pancreatic cancer}}$ = 1,123, $n_{\text{Uveal melanoma}}$ = 509, $n_{\text{Prostate cancer}}$ = 505, n_{AML} = 159, $n_{\text{Renal cancer}}$ = 88, $n_{\text{Bladder cancer}}$ = 23, $n_{\text{Colorectal cancer}}$ = 6, $n_{\text{Breast cancer}}$ = 3, $n_{\text{Lung cancer}}$ = 1), with bars for prostate cancer and uveal melanoma scaled down to 2,500 and exact counts labeled (**g** left), head and neck and brain cancer scaled down to 1,500 along with exact count labeling (**h** left). The right panels of **g** and **h** display the distribution of cancer-specific hypermethylated regions across annotated genomic regions, with a legend specific to this subpanel only. **i**,**j**, Top-enriched GO terms (hypergeometric test with Benjamini–Hochberg-adjusted P values) for genes associated with cancer-specific hyper-DMRs (**i**) and hypo-DMRs (**j**) located in promoter regions.

https://doi.org/10.1038/s43018-026-01116-3



https://doi.org/10.1038/s43018-026-01116-3

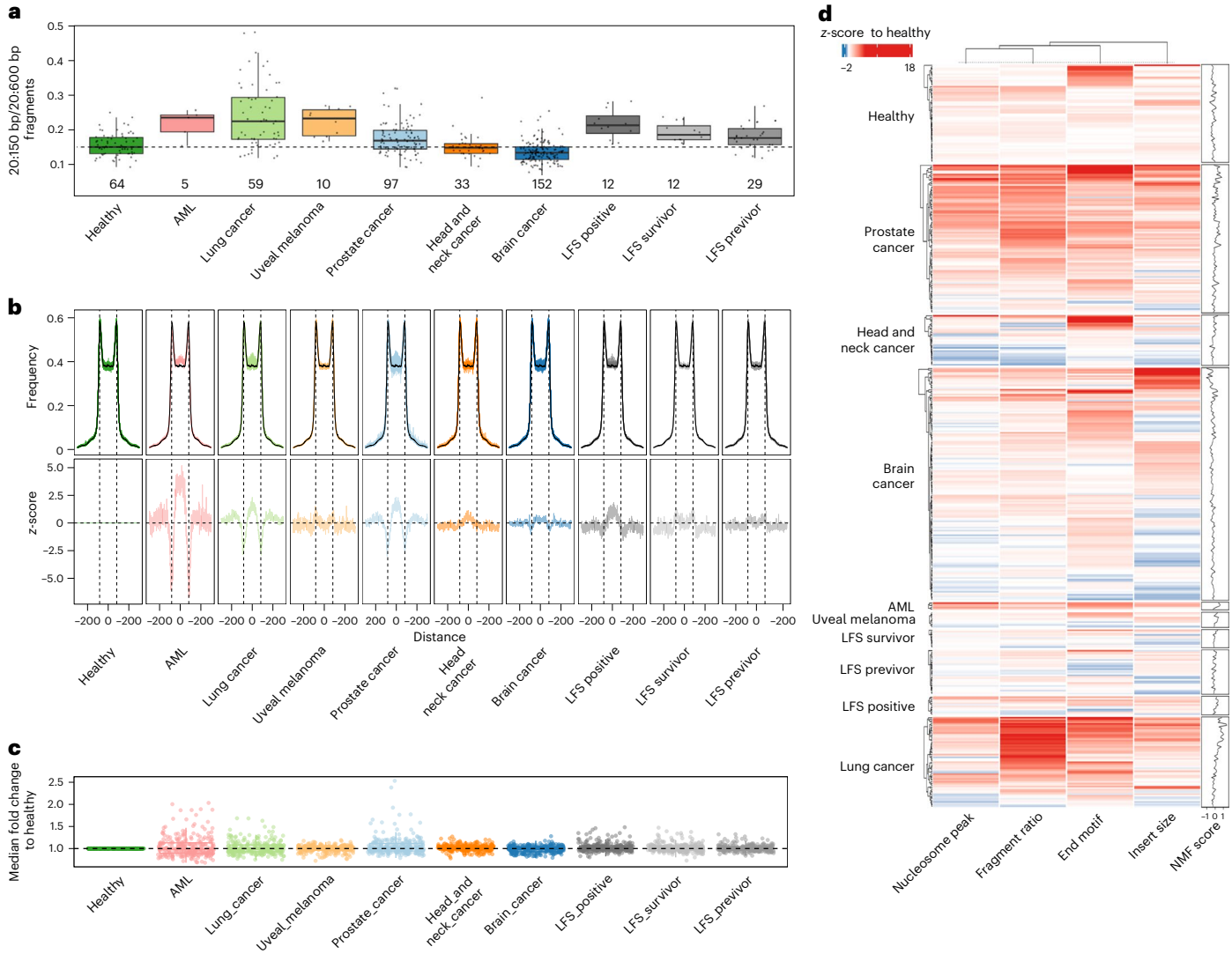


Fig. 3 | Overview of pancancer cfDNA fragmentomic features. a, The proportion of short fragments between 20:150 bp/20:600 bp in length across cancer types (samples: $n_{\text{Healthy}} = 64$, $n_{\text{AML}} = 5$, $n_{\text{Lung cancer}} = 59$, $n_{\text{Uveal melanoma}} = 10$, $n_{\text{Prostate cancer}} = 97$, $n_{\text{Head and neck cancer}} = 33$, $n_{\text{Brain cancer}} = 152$, $n_{\text{LFS positive}} = 12$, $n_{\text{LFS survivor}} = 12$, $n_{\text{LFS previvor}} = 29$). 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 individual samples (colored). Each colored line represents the proportion of 167 bp fragments ending at that position from the nucleosome. Vertical dashed 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 5′ 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. The sample size includes all 256 possible 4-mer end-motif combinations per cancer type, which were calculated on all PE samples which passed qc (detailed counts in **a**). **d**, Heatmap showing the *z*-scores of different fragmentomic features of methylated fragments relative to healthy controls across cancer types. Right: the weighted fragmentation score, calculated using NMF. Larger differences from 0 indicate greater deviation from the expected insert size in healthy controls.

the observed differences in fragmentation measures are primarily driven by cancer type rather than by demographic factors like age or sex, consistent with previous reports on cfDNA fragmentation using shallow whole-genome sequencing[42]. To further explore the relationship among these fragmentomic features, we performed a pairwise Spearman correlation analysis. Features were partially correlated: FS tracked the short-fragment proportion (Spearman $\rho = 0.89$), and nucleosome peak-distance *z*-scores correlated with fragment ratios ($\rho = 0.73$) (Extended Data Fig. 6g). Considering the four features, motifs deviated most from healthy (mean $z = 2.44$), followed by fragment ratios (2.35), insert size (1.68), and nucleosome peak distances (1.31) (Fig. 3d). Healthy samples were more homogeneous (s.d. = 1.83) than most cancer cohorts (mean cohort s.d. = 1.97, range 0.99–3.18), indicating increased fragmentation heterogeneity in cancer (Fig. 3d). Differentiation from healthy controls varied by cancer type and feature. AML was most distinctive by nucleosome peak distances; uveal melanoma, LFS-positive/survivor, and head and neck were most distinctive by 5′ end motifs; lung, prostate and LFS-previvor were most distinctive by fragment ratios; brain was most distinctive by insert size (Fig. 3d and Methods). Average genome-wide *z*-scores were lowest for LFS-previvor (0.44), LFS-survivor (0.51), uveal melanoma (0.77) and head and neck (0.98), indicating smaller deviations from healthy, and highest for AML (3.68), lung (3.46) and prostate (3.23).

In conclusion, our analysis reveals substantial heterogeneity in fragmentomic features of methylated cfDNA fragments across cancer types, highlighting the potential of fragmentomic profiling as a biomarker for differentiating cancer types from healthy samples.

cfDNA methylation and fragmentomic features-based cancer versus healthy classification

We trained classifiers on principal components (PCs) of cfDNA methylation and fragmentomic features (5′ end motifs, fragment ratios, insert size and nucleosome metrics), benchmarking seven algorithms by nested cross-validation, and tested on predefined, independent datasets (Extended Data Fig. 7a and Methods). Power analyses demonstrated sufficient sample sizes for the detection of cancer versus healthy status across individual and integrated feature sets (Extended Data Fig. 7b–d and Methods).

For the independent validation dataset, we first included the TCGE-CFMe-LFS cohort and treated the LFS-positive (n = 12), LFS-survivor (n = 12) and LFS-previvor (n = 29) separately, considering their distinct clinical status[43]. In addition, we incorporated 198 cancer cfMeDIP-seq samples from the TCGE-CFMe-INSPIRE study[14] and 22 healthy cfMeDIP-seq samples from the TCGE-CFMe-HCC study[15] (Extended Data Fig. 7e–g and Supplementary Tables 1 and 2). We were restricted to 95 baseline samples that passed the QC for our validation (Extended Data Fig. 7h,i). On average, 49.4% of the hyper-DMRs identified in individual cancer type comparisons against healthy controls overlapped with our pancancer cfDNA methylation signature (Extended Data Fig. 8a,b and Supplementary Table 17), consistent with our previous observations (Fig. 2b and Supplementary Table 9). Moreover, the fragmentomic features of these 95 samples (Extended Data Fig. 8c–h and Supplementary Tables 18–21) were similar to those in our primary PE dataset (Fig. 3 and Extended Data Figs. 5 and 6), except for the 5′ end-motif distribution in 17 healthy controls from the TCGE-CFMe-HCC study, which utilized a different laboratory protocol (Extended Data Fig. 8g).

We first trained and evaluated the classification workflow on PE samples, excluding all LFS cases. Among individual features, the classifiers based on PCs derived from methylation alone achieved the highest mean area under the curve (mAUC) across model folds (0.986), followed by 5′ end motifs (0.930) and fragment ratios (0.923), whereas insert size and nucleosome peak distances contributed less (mAUCs of 0.748 and 0.689, respectively) (Fig. 4a, Extended Data Fig. 9a,b and Methods). Integrating PCs that contributed >1% of the variance for methylation, 5′ end motifs and fragment ratios resulted in an mAUC of 0.974 (Fig. 4a and Extended Data Fig. 9b). Per-type accuracies were uniformly high: uveal melanoma, lung and brain reached 100%, whereas head and neck cancer was lower yet high (93.9% methylation; 84.8% integrated) (Extended Data Fig. 9c,d and Supplementary Tables 22 and 23). We next evaluated model sensitivities at 99% specificity. The methylation-based classifiers attained the highest mean sensitivity (94.9 ± 8.7% across folds; Fig. 4b), followed by the combined classifiers incorporating methylation, 5′ end motifs and fragment ratios (89.6 ± 8.6%), with lower sensitivities observed for individual fragmentomic features (Fig. 4b).

We next assessed the generalization of these models in our predefined independent validation dataset. The combination of fragment ratios and methylation achieved the highest mAUC of 0.954 (Fig. 4c), close to the primary PE results of 0.961 (Fig. 4a and Extended Data Fig. 9e). However, fragment ratios and methylation alone produced mAUCs of 0.848 and 0.674, respectively (Fig. 4c). Specifically, among cancer types, the combined fragment ratio and methylation classifier achieved the highest accuracies in LFS-positive cases (97.5%, n = 12) and mixed cancers (92.3%, n = 13), while melanoma showed the lowest accuracy (33.3%, n = 9) (Extended Data Fig. 9f and Supplementary Tables 22 and 23). On average, incorporating fragmentomic features improved accuracy by 6.8% compared to using methylation alone (Extended Data Fig. 9f,g and Supplementary Table 23). At 99% specificity, sensitivity was 76.6 ± 4.0% for the combined model versus 63.3 ± 1.9% for methylation; at 90% specificity, 84.3 ± 5.7% versus 65.2 ± 2.0% (Fig. 4d). These results are comparable to those from GRAIL's targeted methylation assay (54.9% sensitivity at 99.3% specificity)[44] and the cfMethyl-seq assay (80.7% sensitivity at 97.9% specificity)[45]. Differences in 5′ motif baselines for the HCC healthy controls, likely protocol-related (Extended Data Figs. 6b and 8g), reduced mAUCs and 99%-specificity sensitivities for motif-containing classifiers (Fig. 4c). Even so, 5′ motif-only models averaged 87.6% accuracy (range 77.8–98.3%) across cancer types, including previously unseen types (Extended Data Fig. 9h and Supplementary Table 23), demonstrating the strong discriminatory power of 5′ motifs. These results highlight the advantages of combining fragmentomic and methylation features to balance model fitting and generalization.

We further assessed our classification workflow on the SE dataset using methylation, 5′ end motif, and their integration (Fig. 4e–h and Extended Data Fig. 10a–e). In our primary SE dataset, methylation achieved an mAUC of 0.938 across folds, whereas 5′ end motifs outperformed methylation with an mAUC of 0.965. Combining methylation and 5′ end motifs further improved classification, resulting in an mAUC of 0.971 (Fig. 4e). At 99% specificity, the mean sensitivity of the combined model was 87.9 ± 9.7%, compared to 90.5 ± 7.7% for 5′ end motifs alone and 79.0 ± 26.5% for methylation alone (Fig. 4f). We observed consistent results within the SE validation dataset (Fig. 4g,h and Methods), where the combined model had the highest mAUC of 0.986, compared to 0.983 for 5′ end motifs and 0.921 for methylation alone (Fig. 4g). Sensitivity was similarly enriched, with 93.0 ± 4% for the combined classifier versus 89.3 ± 3.4% for 5′ end motifs alone and 72.7 ± 23% for methylation (Fig. 4h). Across cancer types, the combined model incorporating methylation and 5′ end motifs achieved the highest mean accuracy in both the primary (95.1%, s.d. = 6.9%) and validation dataset (94.2%, s.d. = 11.6%) SE datasets (Extended Data Fig. 10d,e). These results confirm that our classification workflow is reproducible in the SE dataset, further supporting the added value of integrating methylation with fragmentomic features.

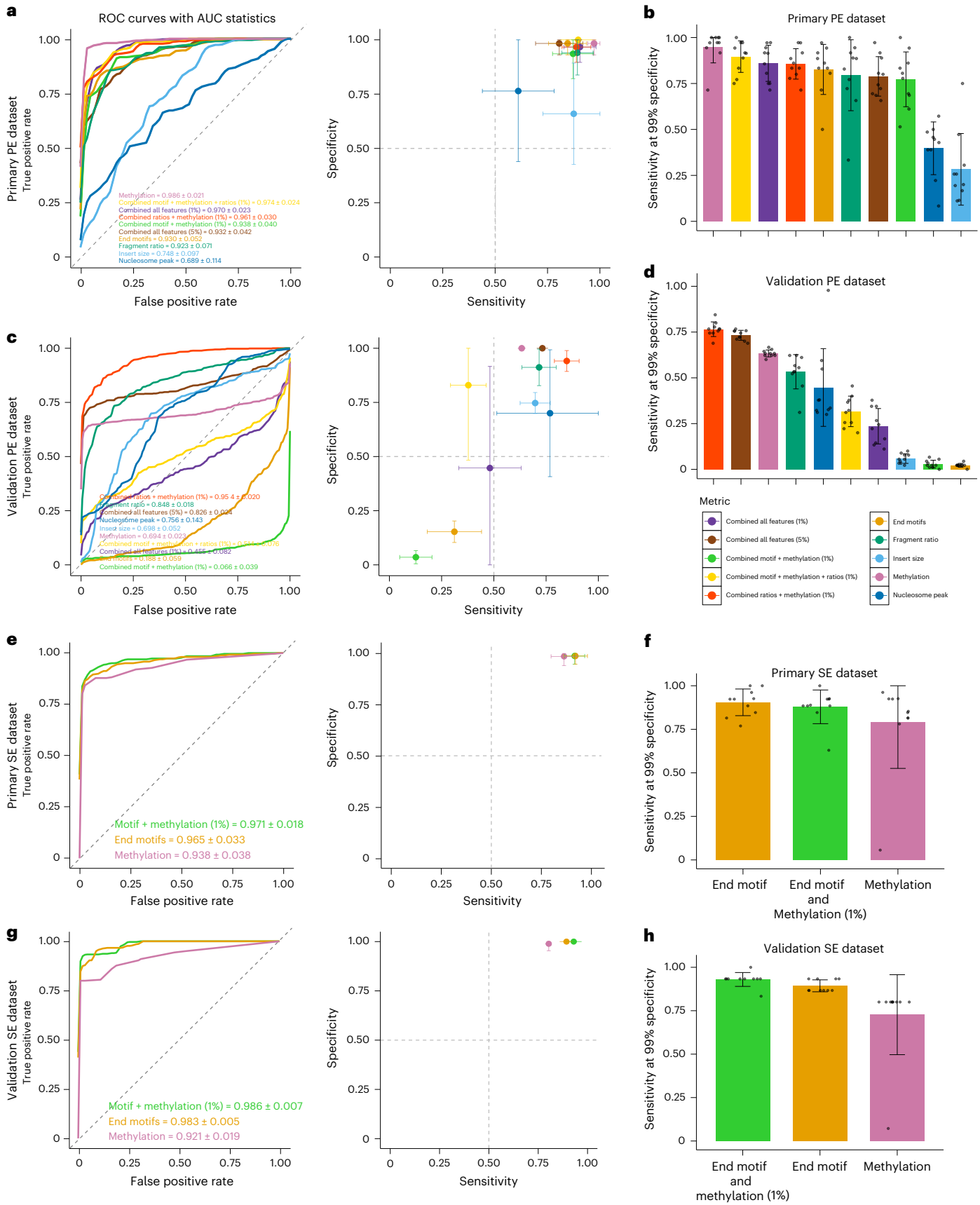
In summary, these results confirm that integrating methylation and fragmentomic features, particularly fragment ratios and 5′ end motifs from cfDNA, improves classification performance in our primary and validation datasets, further reinforcing the robustness and clinical applicability of our classification workflow to distinguish cancer from healthy samples.

cfDNA methylation and fragmentomic features-based cancer type and cancer subtype classification

We subsequently developed distinct predictive classifiers for each cancer type, comparing them against all other cancer types to assess

Fig. 4 | cfDNA methylation and fragmentomic features-based cancer versus healthy classification. a, Left: receiver operating characteristic (ROC) curves illustrating cancer versus healthy classification accuracy in the primary PE dataset. Each curve depicts the mean ROC across cross-validation folds, corresponding to classifiers using methylation, individual fragmentomic features of methylated fragments or combinations thereof. Right: scatter-plot summarizing specificity and sensitivity metrics for each classifier. **b**, Bar chart showing classifier sensitivities at 99% specificity in the primary PE dataset. **c**, ROC curves illustrate cancer versus healthy classification accuracy in the validation dataset, using the same classifiers as **a**. Right: corresponding scatter-plot summarizing sensitivity and specificity metrics. **d**, Bar chart showing classifier sensitivities at 99% specificity in the validation dataset. **e**, ROC curves comparing methylation-only, 5′ end motifs only and combined (5′ end motifs + methylation) classifiers for the primary (original) SE dataset. The scatter-plot depicts each model's sensitivity and specificity. **f**, Bar chart shows sensitivity at 99% specificity for each classifier. **g**, ROC curves comparing methylation-only, 5′ end motifs-only and combined (5′ end motifs + methylation) classifiers for the validation SE dataset. The scatter-plot depicts each model's sensitivity and specificity. **h**, Bar chart shows sensitivity at 99% specificity for each classifier. Error bars in **a**,**c**,**e** and **g** represent ±1 × s.d. across cross-validation folds (n = 10). In **b**,**d f** and **h**, dots show per-fold means (n = 10), and error bars indicate the s.e.

https://doi.org/10.1038/s43018-026-01116-3



https://doi.org/10.1038/s43018-026-01116-3

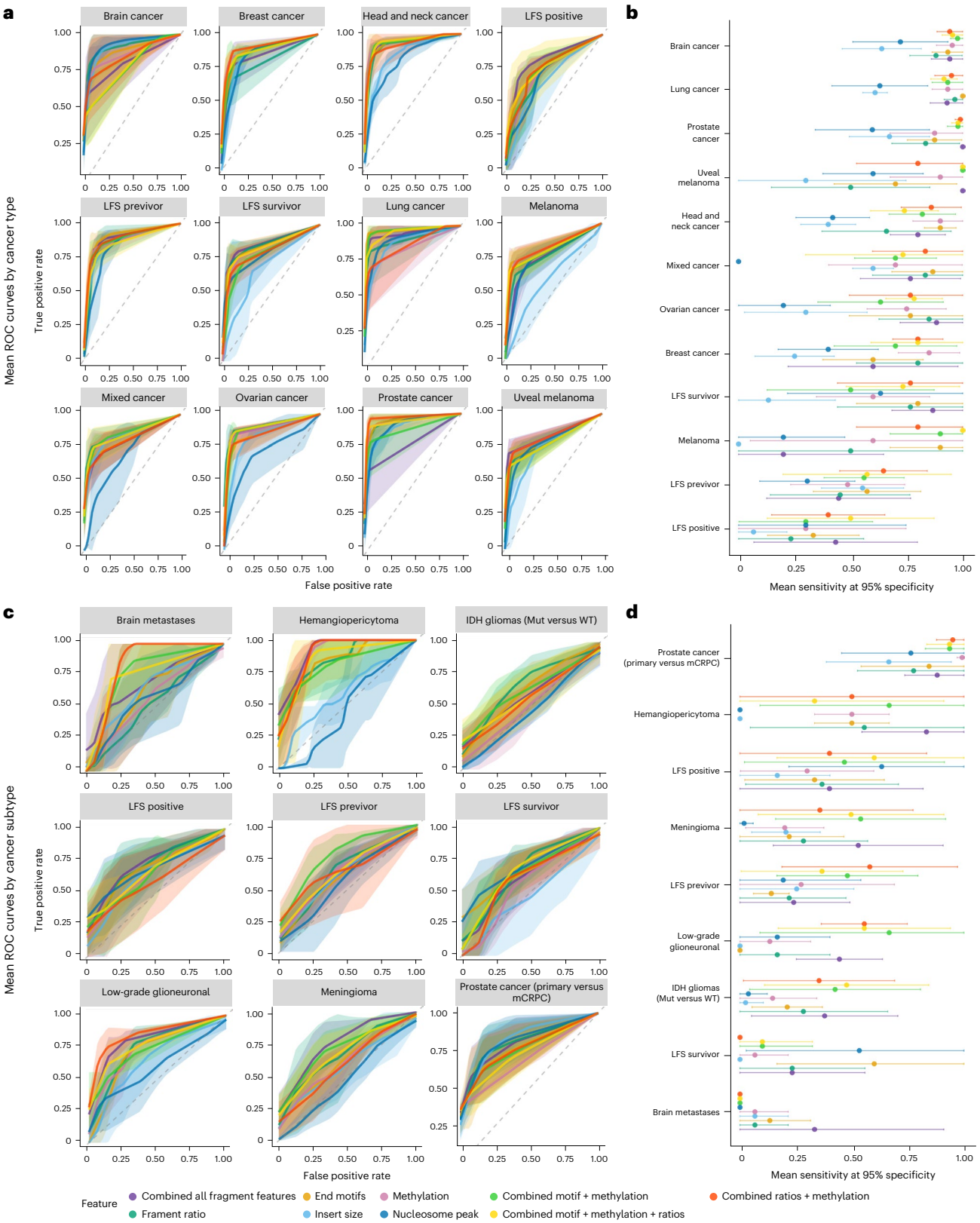


Fig. 5 | cfDNA methylation and fragmentomic features-based cancer type and subtype classification. a, ROC curves illustrate multicancer classification accuracy, averaged across held-out test sets (outer folds of nested *k*-fold cross-validation), in which each cancer type is classified against all others. Shaded areas indicate ±1 × s.d. from the mean ROC curve. **b**, Sensitivity at 95% specificity by cancer type and feature set, illustrating performance variation of each classification approach across different cancer types. Shown is the mean and s.d. of the sensitivities across model folds ($n = 10$ for brain cancer and $n = 5$ for all other cancer types), with error bars representing s.d. **c**, Mean ROC curves for each cancer subtype, stratified by methylation and fragmentomic features alone or their combination (colors), and averaged across the held-out test sets (the outer folds from nested *k*-fold cross-validation). Shaded areas around each curve indicate the s.d. across test sets. **d**, Corresponding sensitivities at 95% specificity, displayed for each cancer subtype and feature set, including error bars reflecting the s.d. across test sets ($n = 3$ for hemangiopericytoma and $n = 5$ for all other evaluated cancer subtypes).

their classification performance beyond a binary cancer versus healthy distinction (Methods). Models used both primary and validation datasets to maximize the number of cancer types available for interpretation (Extended Data Fig. 10f and Methods). Top mAUCs were achieved in prostate (0.972), lung (0.968) and brain cancers (0.968) (Fig. 5a), suggesting that these cancers were particularly well distinguished, potentially due in part to the increased statistical power associated with their larger sample sizes. When the performance was averaged across all cancer types, the highest-performing model was the combined methylation, fragment ratios, and 5′ end-motifs classifiers, which achieved an mAUC of 0.960 and a mean sensitivity of 81% at 95% specificity across cancer types, with the fragment ratios and methylation classifier following with an mAUCs of 0.958 and a mean sensitivity of 80% at 95% specificity (Fig. 5b). Adding fragmentomics to methylation increased mAUC by 0.003 (range −0.014 to 0.042) and mean sensitivity by 3.6% (range −12.5 to 16.7%) across all models (Fig. 5a,b). The combined 5′ end motif, methylation and fragment ratios classifier had an average mAUC gain of 0.007 (range −0.019 to 0.04) and an average sensitivity gain of 7.2% at 95% specificity (range −16 to 40%) (Fig. 5a,b). The performance of other feature sets varied across cancer type, for example, fragment ratios alone yielded the highest performance for lung cancer, while methylation alone performed best for head and neck cancer (Fig. 5a). These findings underscore that integrating fragmentomic features with methylation improves model performance across a range of cancer types and enhances sensitivity without compromising specificity.

Furthermore, we used a similar approach to distinguish between cancer subtypes, but limited each comparison to a single cancer type containing multiple labeled subtypes (Extended Data Fig. 10g and Methods). For prostate cancer subtypes, primary prostate cancer and metastatic castration-resistant prostate cancer (mCRPC) were highly distinguishable from each other, with an mAUC of 0.996 using methylation alone (Fig. 5c). Within central nervous system tumors, low-grade glioneuronal tumors and hemangiopericytoma showed moderate to strong classification performance, with mAUCs of 0.941 and 0.969 (Fig. 5c), respectively. Among LFS subtypes, nucleosome peak distances best captured LFS-survivor, whereas combinations of motifs, methylation and fragment ratios plus nucleosome peak distances aided LFS-previvor and LFS-positive (AUCs ranging from 0.817 to 0.849) (Fig. 5c). Across eight out of nine cancer subtype comparisons, models that used nonmethylation features (either combined with methylation or based on fragmentation features alone) outperformed those based on methylation alone. In these cases, the average improvement was 0.159 in mAUC (ranging from 0.06 in hemangiopericytoma to 0.262 in brain metastases) and 25.5% in sensitivity at 95% specificity (ranging from a decrease of 6.7% in brain metastases to an increase of 53.3% in low-grade glioneuronal tumors) (Fig. 5c,d). These findings highlight that even among genetically related subtypes, different cfDNA features may capture distinct aspects of disease or predisposition state.

Overall, we further demonstrated that both cfDNA methylation and fragmentomic features of methylated fragments possess classification capabilities for cancer type and subtype detection. The identification of distinct fragmentation and methylation signatures emphasizes their potential as biomarkers for cancer detection and monitoring of tumor evolution.

Discussion

Despite the continuous emergence of new cfMeDIP-seq datasets[46–48], our collection that compiled 11 cfMeDIP-seq datasets offers substantial resources with the largest sample size and coverage of cancer types. While this study focused on baseline time point samples for each participant to avoid mixed effects[21], we provided uniformly processed data for all samples as a resource. Longitudinal cohorts such as TCGE-CFMe-LFS[13], TCGE-CFMe-UM[12] and TCGE-CFMe-INSPIRE[14] provide opportunities to study cfDNA methylome and fragmentome dynamics, with mixed-effects methods, such as linear mixed models[49], will facilitate tracking these changes and estimating contributions to clinical outcomes.

Sex-related effects were minimal across autosomes in our dataset; however, a more focused investigation of sex-specific DNA methylation alterations, particularly on the X chromosome[24,50], may reveal distinct methylation signatures. In contrast, age has been reported as a critical confounding factor of DNA methylation alterations[22,23]. Despite the vast majority of our hyper- and hypo-DMRs remaining consistently identified before and after the adjustment of age, we acknowledge that gathering age information is crucial for better controlling age-related confounding effects. In addition, we applied an adapted in-silico PBL depletion approach proposed by Burgener et al.[9] by focusing on DMRs with very low methylation signals in healthy PBLs. Pairing PBL methylome profiling for cancer samples could further reduce false positives[9,11].

Our investigation into cfDNA fragmentomic features revealed substantial variations in insert size, short/long fragment ratios, nucleosome footprinting and 5′ end motifs across different cancer types. These observations are concordant with previous cfDNA fragmentation studies that support diagnostic utility[3–7]. Short fragments predominated in all cancers except head and neck and brain relative to healthy controls. Motif shifts were consistent with DNASE1L3 activity and tumor cfDNA load, and with reports of *DNASE1L3* downregulation in several tumors[51]; however, mechanistic links require functional validation. Because cfMeDIP-seq enriches methylated fragments, our fragmentation readouts arise primarily from methylation-dense regions rather than total cfDNA. This likely accentuates cancer-relevant biology but may not capture genome-wide fragmentation.

cfMeDIP-seq yields methylation and fragmentation in a single assay, which streamlines workflows and may reduce costs for large studies. Integrating 5′ end motifs and fragment ratios with methylation improved discrimination of cancer versus healthy and aided cancer-type and subtype classification. Motifs and ratios contributed most, whereas insert size and nucleosome footprints were informative in specific settings, in line with previous work[3,7,52], while also corroborating previous evidence that epigenetic modifications are highly tissue- and tumor-specific[6]. These results highlight the potential of cfMeDIP in conjunction with fragmentomic features as a screening tool[53,54].

Methods

All samples obtained in this study complied with the relevant ethical regulations approved by the institutional ethics committee and Research Ethics Board at the University Health Network (UHN) with informed and written consent from all participants. For samples collected in this study, all sex and age information was self-reported. Although potential confounding effects of age and sex were evaluated, these variables were not included in our analyses because records were missing for over 240 samples. Informed consent was obtained at clinical follow-up and was consistent with the local Research Ethics Board.

cfMeDIP-seq data processing, QC, quantifications normalization comparison and visualization

All centralized cfMeDIP-seq data underwent uniform processing using our previously developed pipeline MEDIPIPE (v.1.1.0)[17]. 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[17]. To evaluate the impact of sex as a confounding factor, raw read counts for all consecutive 300 bp bins on chromosome X were extracted. Batch correction and normalization were performed within each study using ComBat-seq (sva v.3.52.0) plus DESeq2 (v.1.44.0). Subsequently, principal-component analysis (PCA) was employed to assess the effects of sex, revealing distinct differences between male and female methylation profiles on chromosome X (Extended Data Fig. 2a). Consequently, we excluded bins located on sex chromosomes to minimize potential biases. Additionally, bins located on mitochondrial chromosome and

ENCODE blacklist regions[20] were excluded. A total of 7,445,098 bins containing at least one CpG site were retained for downstream analysis.

To mitigate the batch effects within and across the studies, we first performed PCA to compare six different approaches, including raw read count, RPKM or FPKM, absolute methylation levels estimated by MEDEStrand (RMS, v.0.0.0.9000) and QSEA (beta, v.1.20.0), as well as DESeq2 normalized read counts with and without previous batch correction using ComBat-seq. ComBat-seq was applied to raw read counts and used to correct labeled library preparation or sequencing batches while preserving sample subtype differences, followed by DESeq2 normalization. These assessments were conducted within the TCGE-CFMe-AML study, which included three technical replicates for each of the five participants, and across all healthy controls from four studies (Extended Data Fig. 2b,c). We determined that the ComBat-seq + DESeq2 method was most effective in mitigating batch effects within the SE and PE samples but not across them (Extended Data Fig. 2b–e). To further investigate the difference between SE and PE samples, we extracted R1 reads to mimic SE samples from the TCGE-CFM-AML study, and combined them with original PE samples as well as SE samples for the same five participants from the TCGE-CFMe-MCA study (Extended Data Fig. 2f). The SE and PE labels were then used for an additional round of ComBat-seq batch correction for the known sequencing difference.

For the DNA methylation quantification-based PCA (Fig. 2a and Extended Data Figs. 2a–c,f, 3a–d and 7i), the top 10,000 variable bins were selected based on the interquartile range (IQR). The variance explained by the top nine PCs was calculated, and their Spearman correlations with five representative QC metrics (raw and usable read depth, saturation score and enrichment scores (GoGe and relH)) were examined. For the corresponding Uniform Manifold Approximation and Projection (UMAP) plot, the top 500 PCs were utilized as the input for dimensionality reduction.

Mitigating biases during the cfDNA methylation signatures identification

We restricted our identification of DMRs to the baseline time point samples to avoid mixed effects due to multiple samples from the same participant. Additionally, we filtered the 300-bp bins to include only those containing more than five CpG sites to enhance computing efficiency (n = 1,279,093). DMRs were identified using DESeq2 on ComBat-seq corrected data, requiring an absolute $\log_2$ fold change >1 and a false discovery rate (FDR) <0.05.

To assess the potential variance introduced by age and sex, we compared the DMRs identified with and without these covariates in DESeq2 when comparing each cancer type against corresponding healthy controls, limited samples with available information (Extended Data Fig. 3e–i). AML, bladder, renal and brain cancer were excluded from this analysis due to missing age and sex records for all or a majority (>50%) of samples (Extended Data Fig. 3g and Supplementary Table 2). For establishing age-associated signature, we extended 1,542 unique CpG sites to their centered 300-bp regions, which were previously linked to age across eight DNA methylation clocks compiled in the methylclock (v.1.4.0) R package[26]. Additionally, we performed a Spearman correlation analysis between ComBat-seq + DESeq2 normalized DNA methylation levels and age within each cancer type in our dataset, and identified 576 bins significantly associated with age with an absolute correlation coefficient greater than 0.6 and an FDR < 0.05. To establish a sex-associated signature, we extended 396 autosomal CpG sites previously linked to sex in whole blood samples to their centered 300-bp regions[27]. Furthermore, we performed the DMRs analysis between males and females within each cancer type in our dataset, and identified 232 significantly sex-associated bins with a more relaxed threshold of absolute $\log_2$ fold change > 0 and FDR < 0.05.

To mitigate the potential issue of methylation signals originating from PBLs, we utilized PBL samples from 20 healthy individuals and adopted a modified in silico PBL depletion strategy from the TCGE-CFMe-HNSC study[9]. Specifically, PBL-depleted bins (2,128,686 out of 7,445,098) were defined based on the criterion that the median MEDStrand-estimated absolute methylation value across healthy PBLs was less than 0.1 (Extended Data Fig. 3j). Instead of restricting DMR identification to PBL-depleted bins, we first performed the DMR identification using DESeq2 and then focused on DMRs that overlapped with PBL-depleted regions to suppress false positives. Furthermore, we excluded DMRs that overlapped with regions extending 300 bp from 1,623 cell-type-specific CpG sites linked to blood cells as well as vascular endothelial cells, based on a cell-free DNA tissue origin study[28]. Notably, for AML, we selectively excluded DMRs overlapping 1,065 regions associated with lymphoid cells and vascular endothelial cells, thereby preserving myeloid signals.

Identifying and characterizing the pancancer and cancer-specific cfDNA methylation signatures

For pancancer DMRs, we compared each individual cancer type against healthy controls and all cancer samples combined against healthy controls within SE and PE samples, separately. Only DMRs consistently detected in both individual and combined cancer comparisons were retained as initial pancancer hyper- or hypo-DMRs. For initial cancer-specific DMRs, we compared each cancer type to all the rest of the samples within both SE and PE samples, separately, and filtered out those DMRs reported more than once (Extended Data Fig. 4i). Notably, for cancer types present in both SE and PE samples, we retained only DMRs identified in both comparisons. Then, we filtered out initially identified DMRs that overlapped with age- and sex-associated signatures and further removed DMRs that overlapped with non-PBL-depleted regions and regions reported linked to blood cells and vascular endothelial cells as detailed in the previous section of Methods. In-house scripts were developed to examine the enrichment of DMRs in annotated CpG regions, including CpG island, shore, shelf and sea, as well as promoter and enhancer regions using the annotatr (v.1.24.0) package[55]. To test for gene-set enrichment more robustly, we applied a two-step method using ClusterProfiler (v.4.6.2)[56] with ChIPseeker (v.1.34.1)[57], followed by missMethyl (v.1.32.1)[58], focusing on DMRs located in promoter regions (1,500 bp upstream and downstream of the transcription start site) to establish a more direct link between methylation changes and potential functional consequences. We retained ClusterProfiler-enriched and simplified GO and KEGG terms (FDR < 0.05) that were consistently identified using missMethyl with all EPIC array probes as the background for testing (P value < 0.05). For visualization, we selected the top ten overlapping enriched terms for pancancer DMRs and the top five overlapping enriched terms for cancer-specific DMRs.

cfMeDIP-seq fragment length and distribution across the genome

The cfMeDIP-seq fragment lengths were determined using Picard's (RRID:SCR_006525) CollectInsertSizeMetrics tool (v.4.0.1.2). For fragment-length bins (10–600 bp), per-bin z-scores were computed relative to healthy controls (z = (p_sample−median_healthy)/sd_healthy) and a per-sample score was defined as the sum of per-bin z-scores (Σz). Thresholds ($|z| > 2$ per bin) were used for flagging deviation, not for formal hypothesis testing. 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 NMF, as outlined by Vessies et al.[52], to fragment-length distributions, yielding two latent signatures (cancer-enriched versus healthy-enriched). Per-sample cancer-likelihood was summarized using the ratio of signature weights (Extended Data Fig. 5b).

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[4]. DELFI fragment ratios

were computed as short (90–150 bp) to long (151–220 bp) within 5 Mb windows genome-wide, with GC and depth correction as in Cristiano et al.[4]. Window-wise *z*-scores versus healthy controls were computed; per-sample genome-wide deviation was Σz. The Kruskal–Wallis test was applied to assess disparities as the data was not normally distributed.

cfMeDIP-seq fragment-based nucleosome footprinting and 5′ end-motif patterns

Using the method outlined by Vanderstichele et al.[5], we analyzed plasma cfDNA for nucleosome positioning patterns indicative of chromatin structure alterations in cancer cells[5]. By focusing on fragments that 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 cancer-derived cfDNA. 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.[39]. At each ±1-bp offset within ±300 bp of reference nucleosome dyads, we computed per-offset *z*-scores versus healthy controls and summarized per-sample deviation by Σz. Following the method by Jiang et al.[3], we analyzed 5′ cfDNA end motifs to identify fragmentation patterns specific to cancer cfDNA in both PE and SE data. The 5′ end was used due to the elongation or truncation of the 3′ end of fragments, which occurs during sample preparation for sequencing. Ambiguous bases (N) were excluded. Per-motif proportions (256 motifs) were *z*-scored versus healthy; per-sample motif deviation was Σz.

For each DNASE1L3-associated motif, we computed the fold change by dividing the mean normalized frequency in cancer samples by the mean frequency in healthy controls. Two-sided *t*-tests were then performed to compare motif frequencies between each cancer type and the filtered healthy control group, with *P* values adjusted for multiple comparisons using the Benjamini–Hochberg (BH) method. This approach allowed us to quantify motif enrichment (fold change > 1) or depletion (fold change < 1) in cancer relative to healthy samples. *DNASE1L3* expression across TCGA tumors and normals was retrieved from University of California, Santa Cruz Xena (https://gdc-hub.s3.us-east-1.amazonaws.com/download/GDC-PANCAN.htseq_fpkm.tsv). Sample-level DNASE1L3-associated motif fold changes (versus median healthy) were averaged across DNASE1L3 motifs and correlated with weighted fragment size by Pearson's *r* (Extended Data Fig. 6e).

Confounding effects evaluation for fragmentomic features

To evaluate the effects of age and sex as confounding factors on fragmentomic features, we used a filtered dataset that excluded observations with missing values for age, sex or cancer type. For each feature (weighted FS, fragmentation ratios, short-fragment proportion, nucleosome peak distances, 5′ motifs and fragment length), we fit linear models with age and sex as covariates and cancer type as an adjustment factor. Diagnostics (residuals, Q–Q) and variance inflation factor were used to assess assumptions and multicollinearity. The resulting regression coefficients and their confidence intervals provided estimates of the independent effects of age and sex on each response variable.

Power and sensitivity analysis

A power analysis was conducted to ensure our study was adequately powered to detect differences between samples. For univariate features, including specific 5′ end motifs, insert size, 5-Mb fragment ratios and individual DMRs, we computed Cohen's *d*. The median absolute effect size was used as a representative metric for further sensitivity analysis. Power estimates were derived by varying the sample size in a two-sample *t*-test framework at a 0.05 significance level with pwr.t.test (v.1.3-0). To assess the sample size requirements for our integrated cfDNA classifier, we performed a non-parametric, permutation-based multivariate power analysis using PERMANOVA (adonis2, vegan v.2.6-10). Analyses used PCA scores from the training PE dataset labeled cancer versus healthy. Multivariate normality was assessed with Shapiro–Wilk (mvnormtest v.0.1.9.3) and was violated for the high-dimensional PCA scores. We therefore used a permutation-based approach using the adonis2 function (vegan v.2.6-10) to simulate power. For each dataset, equal numbers of samples were drawn from each group (without replacement) across a range of sample sizes (10–500 per group), and PERMANOVA was performed with 999 permutations. The empirical power was estimated as the proportion of replicates achieving a P value < 0.05. These estimates defined power curves and sample size requirements for detecting group differences from PCA-transformed features.

Our power analysis showed that detecting cancer from healthy samples based on individual fragmentation features reached 80% power with the sample sizes ranging from approximately 50 to 280 (Extended Data Fig. 7b,c). For pancancer DMRs identification in PE and SE samples, we achieved a moderate power of 73.7% with the sample sizes of 378 (SE) and 420 (PE) (Extended Data Fig. 7b,c). Integrative classification models combining PCA-transformed methylation and fragmentomic features achieved >80% power with as few as 15–20 samples per group (Extended Data Fig. 7d).

DNA methylation and fragmentomic features-based classification

To develop classifiers based on cfDNA methylation and fragmentomic features, we first applied PCA to reduce feature dimensionality while retaining the major sources of variation. PCA was performed on the training set and the learned transformations were applied to the validation set to prevent data leakage. For each feature set (methylation, 5′ end motifs, fragment ratios, insert size, nucleosome peak distances), PCA retained components explaining ≥80–95% of variance, selecting thresholds to optimize high explained variance while minimizing the number of features. For integrative models, we also evaluated component cutoffs contributing >5% and >1% of total variance. Standardization and PCA were fitted on training data only (within folds) and applied to validation folds to prevent leakage.

We benchmarked multiple classification algorithms, including LASSO, support vector machines, random forest, gradient-boosting machines and XGBoost (we also evaluated linear discriminant analysis and *K*-nearest neighbors in sensitivity analyses). For binary cancer versus healthy classification tasks, models were evaluated using repeated tenfold cross-validation with performance assessed by Cohen's kappa. Sampling was stratified by class, with class balance enforced by random down-sampling within resamples. The best-performing algorithm per feature set was then optimized using nested cross-validation, with inner/outer fold counts scaled to dataset size (for $n < 10$, leave-one-out cross-validation was used; for $10 \le n \le 100$, we used five outer and five inner folds; for $101 \le n \le 250$, ten outer and five inner folds; and for $n > 250$, ten outer and ten inner folds). External validation was performed using either an independent cohort (for PE data) or a 10% hold-out set (for SE data), which was excluded from PCA, training and hyperparameter tuning.

Cancer-type classification used a one-versus-all framework, while subtype classification compared biologically or clinically meaningful pairs, such as primary versus metastatic prostate cancer or IDH-mutant versus IDH-wildtype gliomas. Given label imbalance, the area under the curve was the selection metric (Cohen's kappa for cancer versus healthy). For subtypes, PCA variance thresholds (60–95%) were included in the inner-loop tuning grid to accommodate varying signal and sample sizes. Performance was summarized with fold-averaged receiver operating characteristic (ROC) and sensitivity plots; error bars denote across-fold s.d. In cancer versus healthy classification (Fig. 4c), LFS-previvor and LFS-survivor cases were excluded from training but scored post hoc (Extended Data Fig. 9f–h); for sensitivity summaries, they were labeled as 'cancer' given their elevated risk.

Statistics and reproducibility

Data collection and analysis were not blinded and randomized. No statistical methods were used to predetermine sample sizes. Sample size was based on sample availability, and power analyses confirmed that the sample sizes were sufficient to detect cancer versus healthy status across individual and integrated feature sets. We analyzed all eligible samples after prespecified QC. From the primary datasets, 100 samples were excluded due to stringent QC, and 123 samples from patients with multiple samplings were removed to avoid mixed effects. In validation datasets, 19 failed QC and 106 multiple samplings were excluded. Full datasets for all samples are provided as a resource.

All data were assumed to be not normally distributed unless proven otherwise. Unless stated otherwise, tests were two-sided and *P* values were FDR-adjusted. Parametric pairwise comparisons used two-sample *t*-tests with BH or Holm correction as noted; Levene's test assessed homoscedasticity. When assumptions were violated, we used Wilcoxon rank-sum (pairwise) or Kruskal–Wallis (multigroup) with Dunn's post hoc tests (BH/Holm-adjusted). Rank associations among fragmentomic metrics used Spearman's ρ (BH-adjusted). Chi-squared tests assessed motif-frequency shifts relative to expected counts. Gene-set enrichment used the hypergeometric test with BH correction. DMR enrichment across annotations was evaluated by a permutation test ($n = 1{,}000$). Where per-bin/feature *z*-scores were computed, $|z| > 2$ was treated as a descriptive deviation flag and inferential claims relied on the tests above. Test choices and any figure-specific deviations are detailed in figure legends and relevant Methods subsections. QC/feature generation ran under R/4.0.0; machine learning pipelines under R/4.1; and summarization/plotting under R 4.2.2–4.3.3.

Reporting summary

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

Data availability

The raw cfMeDIP-seq data that support the findings of this study from TCGE-CFMe-MCA, TCGE-CFMe-BCA, TCGE-CFMe-HNSC and TCGE-CFMe-SCLC were obtained directly from the corresponding authors of the respective cohorts upon request. TCGE-CFMe-MCA and TCGE-CFMe-BCA data can be requested from D.D.D.C. (ddecarv@uhnresearch.ca), TCGE-CFMe-HNSC data from S.V.B. (scott.bratman@rmp.uhn.ca), and TCGE-CFMe-SCLC data from B.H.L. (benjamin.lok@rmp.uhn.ca) via submission of a data access application. The other cfMeDIP-seq datasets were deposited in the European Genome-Phenome Archive (EGA): TCGE-CFMe-AML (EGAS00001005069), TCGE-CFMe-PRAD (EGAS00001005522), TCGE-CFMe-UM (EGAD00001008998), TCGE-CFMe-HBC (EGAS00001006539), TCGE-CFMe-LFS (EGAS00001006539), TCGE-CFMe-INSPIRE (EGAD00001011312) and TCGE-CFMe-HCC (EGAD50000000652). Access to all these seven UHN-generated datasets is made available upon completion of the required data access agreement, which will be reviewed by the UHN genomics data access committee (dac@uhn.ca). Data access will be granted to qualified investigators for appropriate and compliant use. The source data for DMRs analyses (Fig. 2b and Extended Data Figs. 3e,f, 4d,e,g and 8a) and methylation-based PCA and UMAP plots (Fig. 2a and Extended Data Figs. 2a–c,f, 3a,c and 7i), together with BED files for pancancer and cancer-specific DMRs as well as age- and sex-associated regions, are available on Zenodo (https://doi.org/10.5281/zenodo.15191455 (ref. 59)). All other source data are provided in this paper. The remaining data are available within the article and Supplementary Tables. Source data are provided with this paper.

Code availability

All codes for this study have been deposited on GitHub (https://github.com/HansenHeLab/cfMeDIP-seq_Data_Resource_Codes). The analytical pipeline will also be deposited on the CoBE platform (www.pmcobe.ca). Scripts used for the fragmentomic 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.

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**, 30 (2022).
6. Zhou, Z. et al. Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. *Proc. Natl Acad. Sci. USA* **120**, e2220982120 (2023).
7. Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (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*. https://doi.org/10.1158/2159-8290.CD-23-0456 (2023).
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*. https://doi.org/10.1158/2159-8290.CD-23-1060 (2024).
15. Chen, K. et al. Plasma cell-free DNA methylomes for hepatocellular carcinoma detection and monitoring after liver resection or transplantation. *Ann. Surg.* https://doi.org/10.1097/SLA.0000000000007003 (2025).
16. Wilson, S. L. et al. Sensitive and reproducible cell-free methylome quantification with synthetic spike-in controls. *Cell Rep. Methods* **2**, 100294 (2022).
17. Zeng, Y. et al. MEDIPIPE: an automated and comprehensive pipeline for cfMeDIP-seq data quality control and analysis. *Bioinformatics* **39**, btad423 (2023).
18. 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).
19. 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).
20. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).

21. Gomes, D. G. E. Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ* **10**, e12794 (2022).
22. Lau, C.-H. E. & Robinson, O. DNA methylation age as a biomarker for cancer. *Int. J. Cancer* **148**, 2652–2663 (2021).
23. Reale, A., Tagliatesta, S., Zardo, G. & Zampieri, M. Counteracting aged DNA methylation states to combat ageing and age-related diseases. *Mech. Age. Devel.* **206**, 111695 (2022).
24. Yu, C. et al. Epigenetic drift association with cancer risk and survival, and modification by sex. *Cancers* **13**, 1881 (2021).
25. Kiesel, B. et al. Sex-specific differences in DNA methylation defining prognostically relevant subgroups in glioblastoma. *J. Neurosurg.* https://doi.org/10.3171/2024.9.jns24665 (2025).
26. Pelegí-Sisó, D., de Prado, P., Ronkainen, J., Bustamante, M. & González, J. R. methylclock: a Bioconductor package to estimate DNA methylation age. *Bioinformatics* **37**, 1759–1760 (2021).
27. Grant, O. A., Wang, Y., Kumari, M., Zabet, N. R. & Schalkwyk, L. Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. *Clin. Epigenetics* **14**, 62 (2022).
28. 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).
29. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet.* **37**, 1012–1027 (2021).
30. Huang, J. et al. Cancer detection and classification by CpG island hypermethylation signatures in plasma cell-free DNA. *Cancers* **13**, 5611 (2021).
31. Rao, V. R., Perez-Neut, M., Kaja, S. & Gentile, S. Voltage-gated ion channels in cancer cell proliferation. *Cancers* **7**, 849–875 (2015).
32. Ji, X. et al. Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nat. Commun.* **9**, 3221 (2018).
33. Zhang, Y., Chen, Q., Gong, M., Zeng, Y. & Gao, D. Gene regulatory networks analysis of muscle-invasive bladder cancer subtypes using differential graphical model. *BMC Genom.* **22**, 863 (2021).
34. Chen, B., Chakrobortty, N., Saha, A. K. & Shang, X. Identifying colon cancer stage related genes and their cellular pathways. *Front. Genet.* **14**, 1120185 (2023).
35. Ahmed, M. B., Alghamdi, A. A. A., Islam, S. U., Lee, J.-S. & Lee, Y.-S. cAMP signaling in cancer: a PKA-CREB and EPAC-centric approach. *Cells* **11**, 2020 (2022).
36. Wu, L., Lian, W. & Zhao, L. Calcium signaling in cancer progression and therapy. *FEBS J.* **288**, 6187–6205 (2021).
37. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genom.* **16**, S1 (2015).
38. Bruhm, D. C. et al. Genomic and fragmentomic landscapes of cell-free DNA for early cancer detection. *Nat. Rev. Cancer* https://doi.org/10.1038/s41568-025-00795-x (2025).
39. 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).
40. 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).
41. An, Y. et al. DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. *Nat. Commun.* **14**, 287 (2023).
42. van der Pol, Y. et al. The effect of preanalytical and physiological variables on cell-free DNA fragmentation. *Clin. Chem.* **68**, 803–813 (2022).
43. Guha, T. & Malkin, D. Inherited *TP53* mutations and the Li-Fraumeni syndrome. *Cold Spring Harb. Perspect. Med.* **7**, a026187 (2017).
44. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
45. Stackpole, M. L. et al. Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. *Nat. Commun.* **13**, 5566 (2022).
46. Lu, H. et al. Detection of ovarian cancer using plasma cell-free DNA methylomes. *Clin. Epigenetics* **14**, 74 (2022).
47. 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).
48. Cheng, N. et al. Pre-diagnosis plasma cell-free DNA methylome profiling up to seven years prior to clinical detection reveals early signatures of breast cancer. Preprint at *medRxiv* https://doi.org/10.1101/2023.01.30.23285027 (2023).
49. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
50. McCartney, D. L. et al. An epigenome-wide association study of sex-specific chronological ageing. *Genome Med.* **12**, 1 (2019).
51. Deng, Z. et al. DNASE1L3 as a prognostic biomarker associated with immune cell infiltration in cancer. *Onco. Targets. Ther.* **14**, 2003–2017 (2021).
52. 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).
53. 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).
54. Luchini, C. et al. Liquid biopsy as surrogate for tissue for molecular profiling in pancreatic cancer: a meta-analysis towards precision medicine. *Cancers* **11**, 1152 (2019).
55. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
56. 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).
57. 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).
58. Maksimovic, J., Oshlack, A. & Phipson, B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol.* **22**, 173 (2021).
59. Zeng, Y. pan_cancer_cfMeDIP_Figures_source_data [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.15191455 (2025).

Acknowledgements

We 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 the analyses included in this work. This work was supported by the Cancer Genetics and Epigenetic Program at Princess Margaret Cancer Center. H.H.H. is supported by the 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 (OICR) and the Gattuso-Slaight Personalized Cancer Medicine Fund. M.L. is supported by the CIHR (FRN-153234, 158225, 168933 and 191847), the 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. S.V.B. is supported by the Gattuso-Slaight Personalized Cancer Medicine Fund and the Dr. Mariano Antonio Elia Chair in Head & Neck Cancer Research at the UHN and the University of Toronto. We would like to acknowledge the Princess Margaret's Head & Neck Translational Research Program, supported by philanthropic funds from Joe's Team and the Wharton, Elia, Tozer, Riley and Reed Families. Research in the B.H. Lok Laboratory is supported by the Canada Foundation for Innovation, CIHR, National Institutes 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 OICR Genomics Program (http://genomics.oicr.on.ca) and Translational Genomics Laboratory, a joint initiative between the Princess Margaret Cancer Centre and the OICR (D. Torti, K. Marsh, B. Lam, M. Taschuk, L. Heisler and C. 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 no. 32383 and 38401); Ontario Ministry of Research and Innovation, Ontario Research Fund Small Infrastructure Program; and the OICR. E.Y.S.-Z. acknowledges funding from a Cancer Research Institute Irvington Postdoctoral Fellowship, Princess Margaret Global Oncology Fellowship, and a Hold'em for Life Oncology Fellowship. S.C.M. is supported by a Canadian Cancer Society Research Training Award (no. 708002).

Author contributions

Study conception and design: Y.Z., D.D.A., M.L., T.J.P. and H.H.H.; data acquisition and clinical interpretation: Y.Z., S.L.W., N.Z., K.C., E.S.-G., F.N., A.G., P.A., B.H.L., M.M.H., R.H.K., G.Z., D.D.D.C., S.V.B., T.J.P. and H.H.H.; cfMeDIP-seq dataset integration and methylation analysis: Y.Z. (lead) and D.D.A., and fragmentomic and machine-learning analysis: D.D.A. (lead) and Y.Z.; supporting data analysis and interpretation: A.S., N.C., Y.F., S.C.M., E.B., W.Y., P.L., E.Y.S.-Z., D.W. and S.M.; paper first draft: Y.Z., D.D.A., M.L., T.J.P. and H.H.H. All authors revised and approved the paper.

Competing interests

S.V.B. declares stock ownership in Adela; a leadership position in Adela; patents licensed to Roche, Adela; and royalties from Roche. T.J.P. 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.L. reports grants from Pfizer and grants, personal fees and nonfinancial support from AstraZeneca and personal fees from Daiichi-Sankyo outside the submitted work. E.S.G. reports research funding from GSK and Rgenta, as well as consulting roles with GSK and Lutris. H.H.H. reports personal fees from Synth-Med outside the submitted work. S.L.W. and M.M.H. have a patent application licensed to Adela. E.Y.S.-Z. has provided consulting for Adela. N.Z. reports stock ownership and a leadership position in Otaksa. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-026-01116-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-026-01116-3.

Correspondence and requests for materials should be addressed to Yong Zeng, Mathieu Lupien, Trevor J. Pugh or Housheng Hansen He.

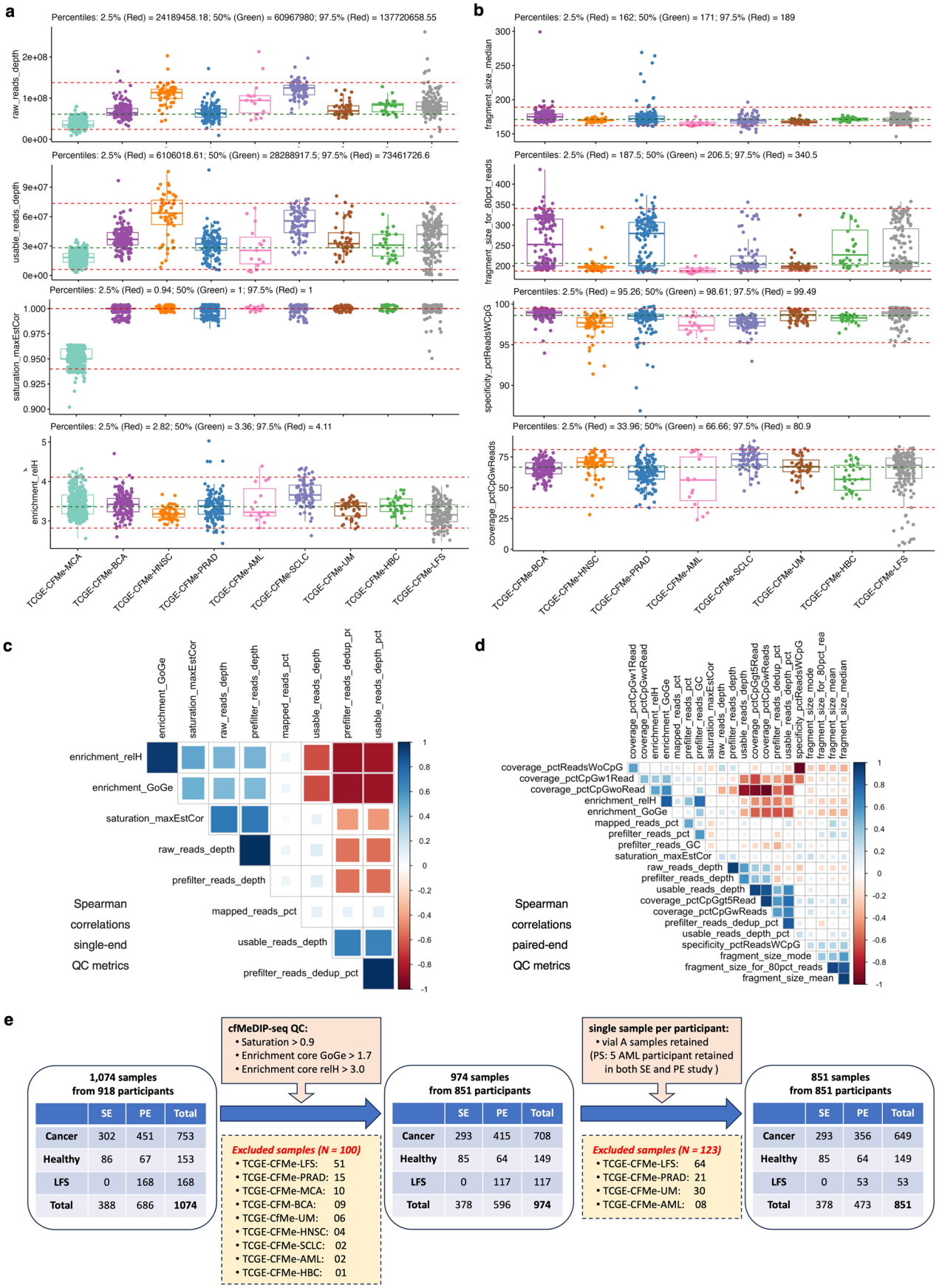
Peer review information *Nature Cancer* thanks Yoshiaki Nakamura and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [3]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Obstetrics and Gynecology, McMaster University, Hamilton, Ontario, Canada. [5]Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [6]Toronto General Hospital Research Institute, University Health Network, Toronto, Ontario, Canada. [7]Division of Medical Oncology, Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [8]Division of General Surgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada. [9]Department of Population Health, Big Data Institute, University of Oxford, Oxford, UK. [10]Vector Institute, Toronto, Ontario, Canada. [11]The Hospital for Sick Children, Toronto, Ontario, Canada. [12]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada. [13]These authors contributed equally: Yong Zeng, Dor D. Abelman. ✉e-mail: Yong.Zeng@uhn.ca; Mathieu.Lupien@uhn.ca; Trevor.Pugh@utoronto.ca; Hansen.He@uhn.ca

https://doi.org/10.1038/s43018-026-01116-3

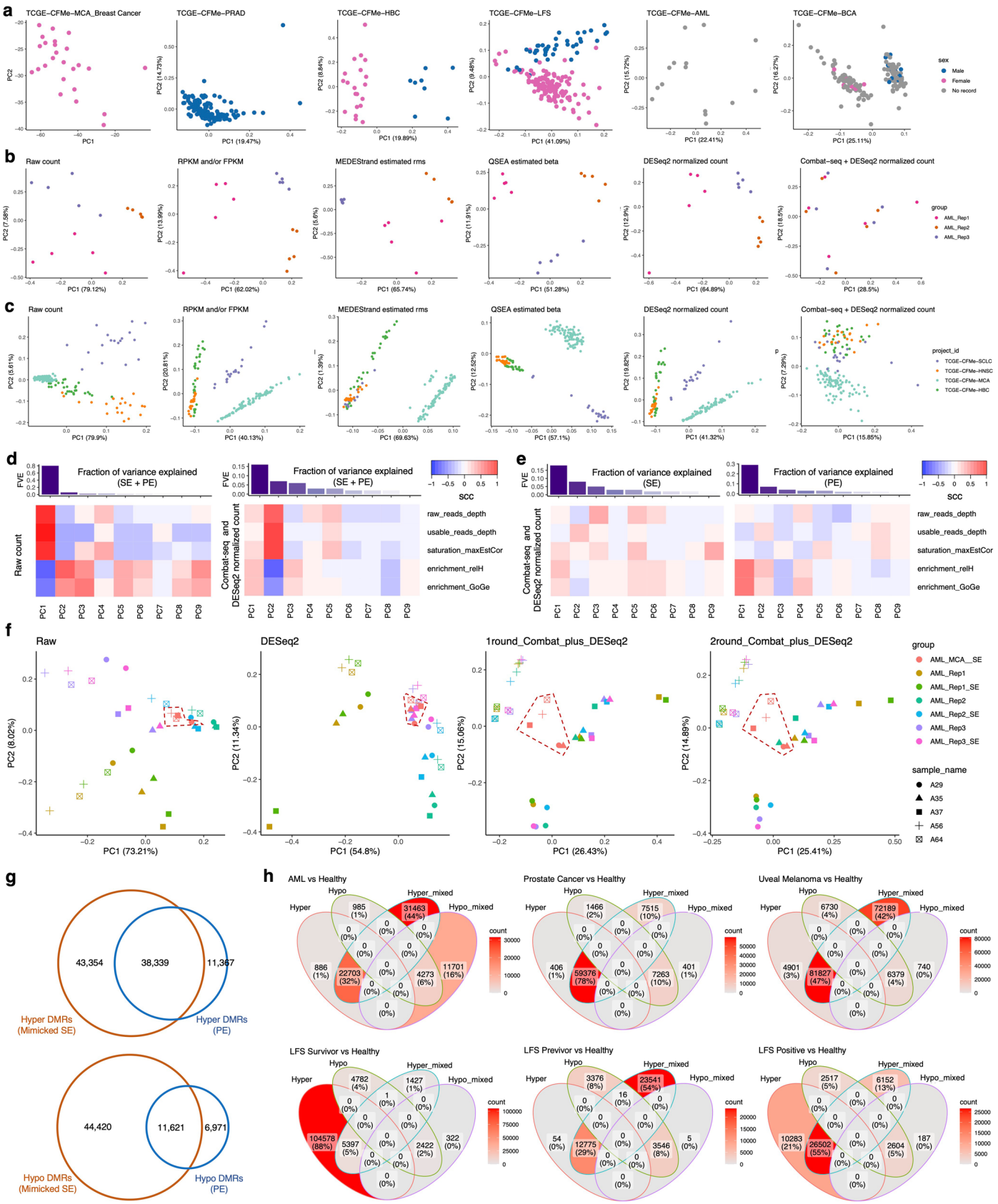


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 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. **b**, The distribution of median fragment size, the fragment range that covers 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 (TCGE-CFMe-MCA) were excluded from this figure. Box plots in a and b represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. The 2.5%, 50%, and 97.5% percentiles are shown with dashed lines accordingly (samples: $n_{TCGE\text{-}CFMe\text{-}MCA} = 388$, $n_{TCGE\text{-}CFMe\text{-}BCA} = 161$, $n_{TCGE\text{-}CFMe\text{-}HNSC} = 56$, $n_{TCGE\text{-}CFMe\text{-}PRAD} = 133$, $n_{TCGE\text{-}CFMe\text{-}AML} = 15$, $n_{TCGE\text{-}CFMe\text{-}SCLC} = 79$, $n_{TCGE\text{-}CFMe\text{-}UM} = 46$, $n_{TCGE\text{-}CFMe\text{-}HBC} = 28$ and $n_{TCGE\text{-}CFMe\text{-}LFS} = 168$). **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. **e**, Flowchart of sample filtering and selection.

https://doi.org/10.1038/s43018-026-01116-3

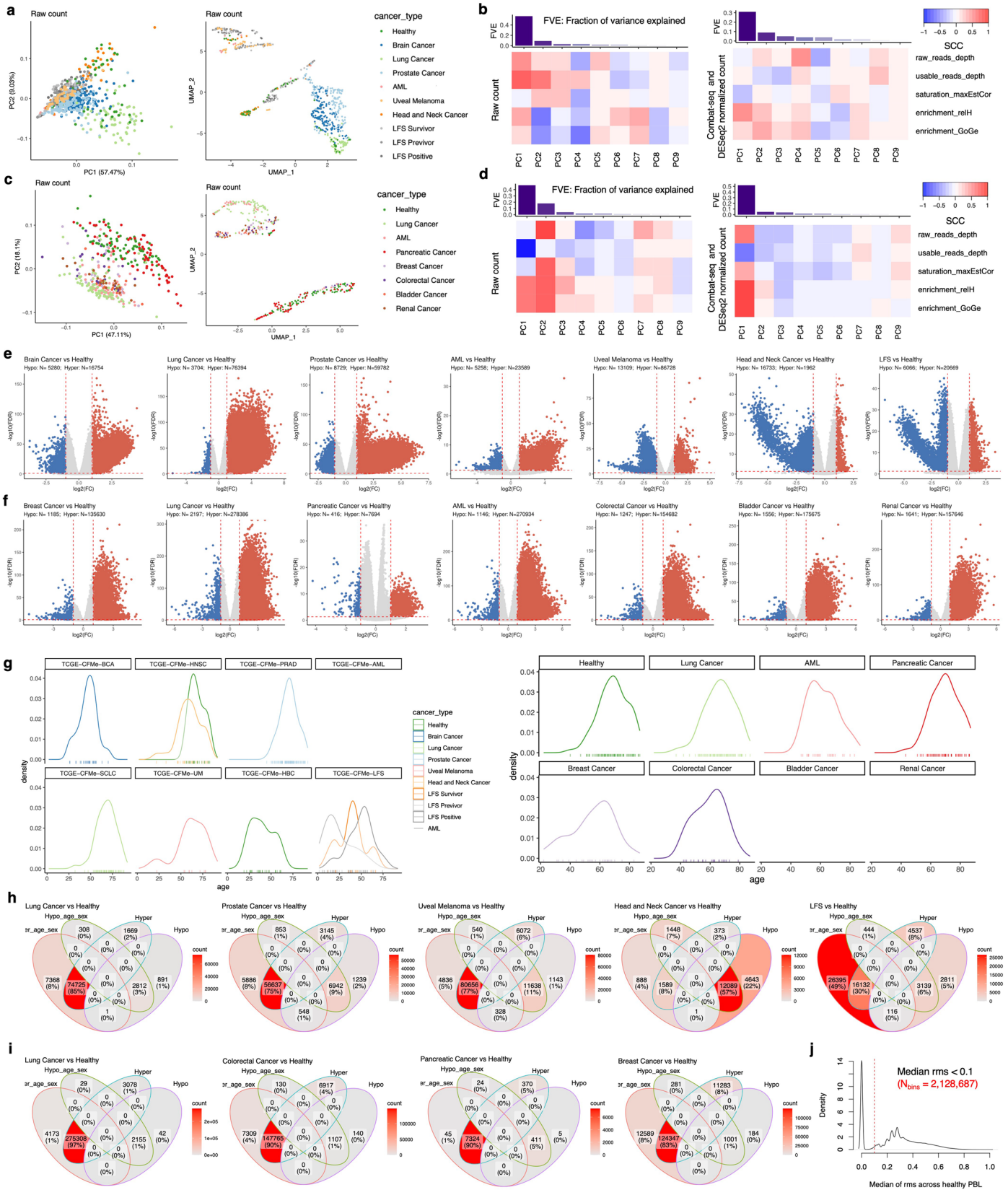


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

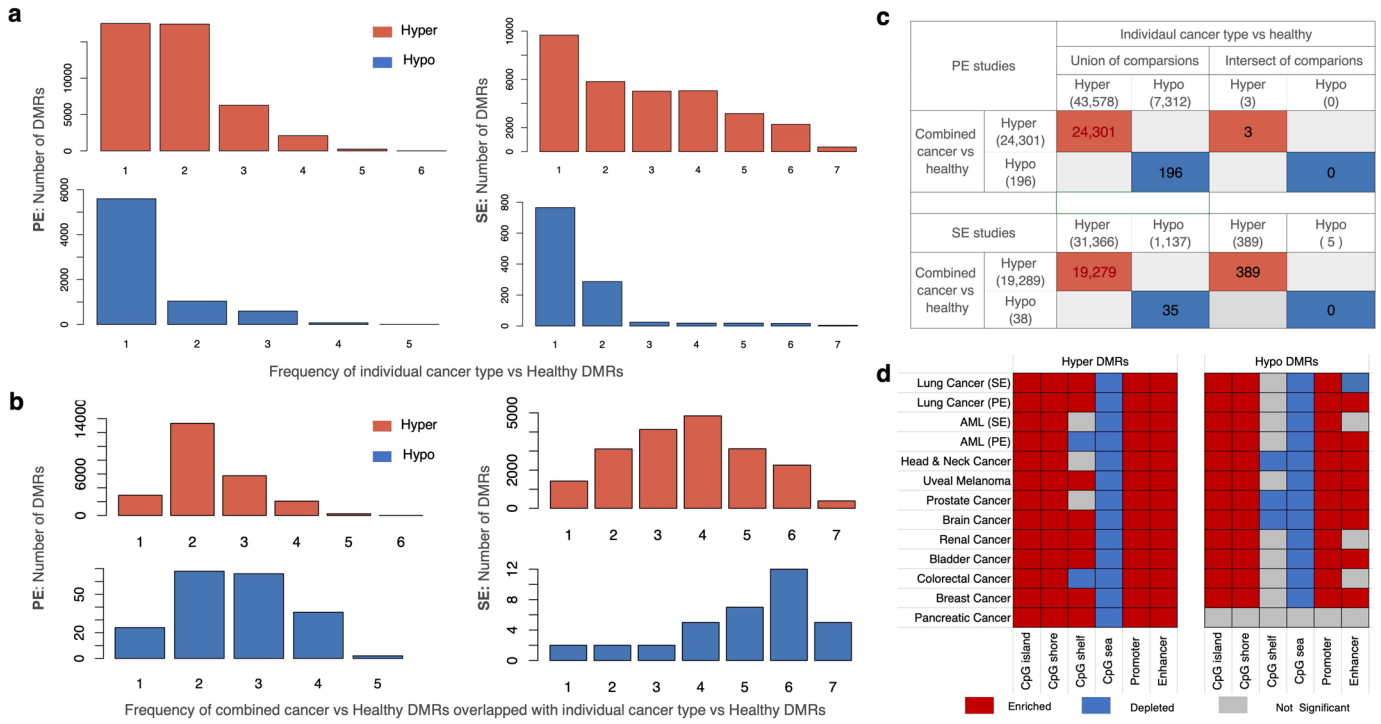
Extended Data Fig. 2 | Comparisons of DNA methylation quantification and normalization methods. a, PCA plots based on the methylation signals on the chromosome X across female-only (TCGE-CFMe-MCA: breast cancer), male-only (TCGE-CFMe-PRAD: prostate cancer), mixed (TCGE-CFMe-HBC and TCGE-CFMe-LFS), and unlabeled-sex samples (TCGE-CFMe-AML and TCGE-CFMe-BCA). **b**, PCA plots for the TCGE-CFMe-AML study, illustrating 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. Colors indicate sample replicates. **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. **f**, PCA plots display data using raw count, DESeq2 normalized count, DESeq2 normalized count with one round ComBat-seq batch correction, and an additional round of batch correction for mimicked SE and PE samples for all AML samples derived from the same 5 participants. **g**, Overlap of (top) hypermethylated and (bottom) hypomethylated DMRs identified from five AML samples compared to five healthy controls using full PE data (blue) and mimicked SE data (orange). **h**, The overlap of DMRs before (Hyper_mixed and Hypo_mixed) and after (Hyper and Hypo) accounting for the mixed effects due to multiple sampling from the same participant in prostate cancer, uveal melanoma, AML and LFS cohort.

https://doi.org/10.1038/s43018-026-01116-3



Extended Data Fig. 3 | Differential methylation analysis between individual cancer types against healthy controls. 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 DMRs before filtering in the comparison of each individual cancer type against the healthy controls for PE studies (**e**) and SE studies (**f**). **g**, Age distribution within each cancer type across PE samples (left) and SE samples (right). Overlap of the DMRs identified with (Hyper_age_sex and Hypo_age_sex) and without (Hyper and Hypo) age and sex as confounding factors across PE samples (**h**) and SE samples (**i**). **j**, The distribution of the median of MEDEstrand estimated rms across 20 healthy PBL samples.

https://doi.org/10.1038/s43018-026-01116-3

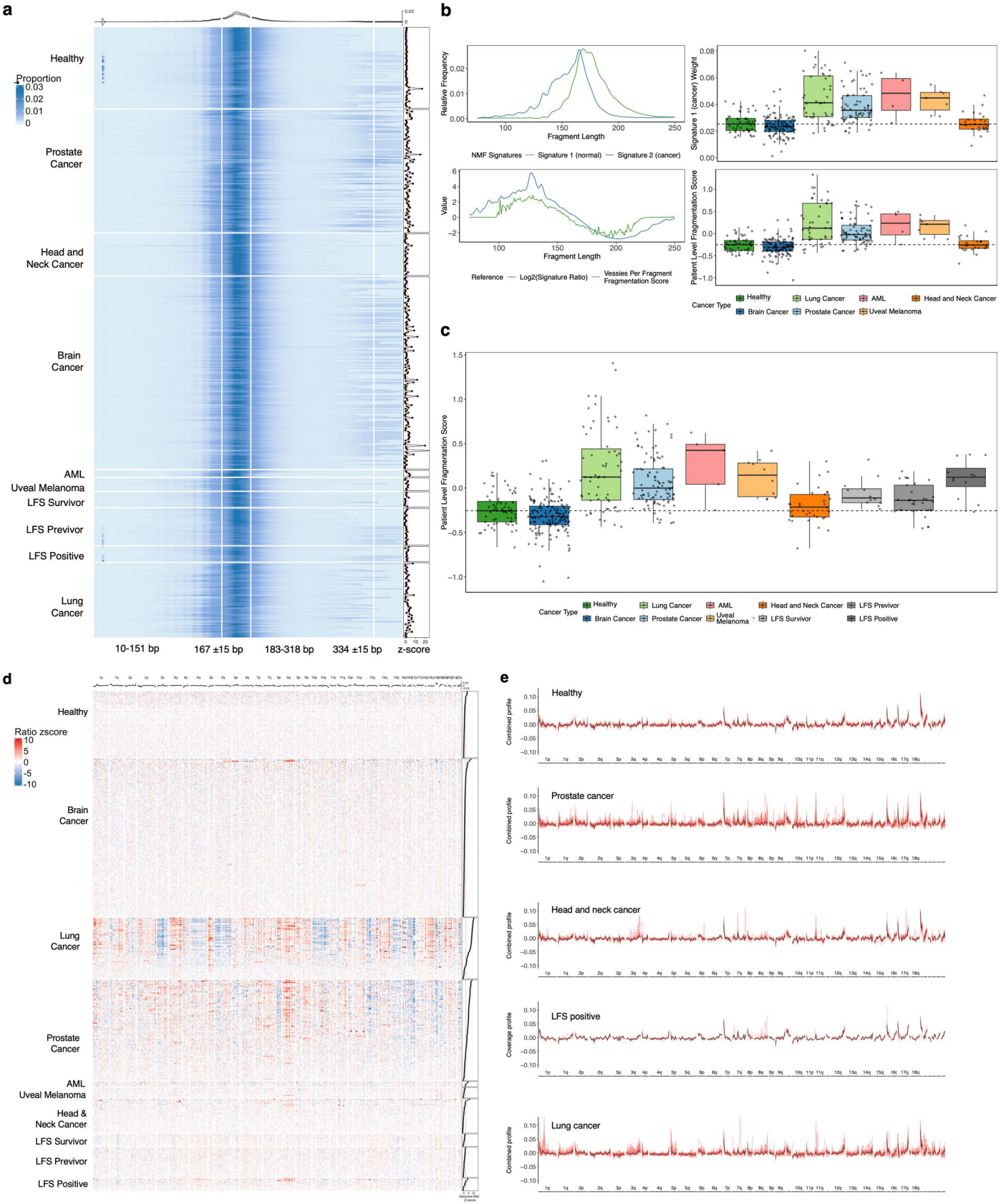


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 4 | Characterization of pancancer and cancer-specific methylated regions. a, Histogram depicting hyper- and hypo-DMRs frequencies in each individual cancer compared to healthy controls for PE (left 2 rows) and SE (right 2 rows) studies. **b**, Histogram of overlapping hyper- and hypo-DMRs between combined cancer vs health control and individual cancer vs healthy controls for PE (left 2 rows) and SE (right 2 rows) studies. **c**, Overlapping of DMRs identified in combined cancer vs health control and the union/common of DMRs across all individual cancer vs healthy controls for PE (top 2 rows) and SE (bottom 2 rows) studies. **d**, Enrichment of hyper- and hypo-DMRs from individual cancer type vs healthy control in annotated CpG, promoter and enhancer regions. **e**, Enrichment of hyper- and hypo-DMRs, identified using combined PE (left 2 columns) and SE (right 2 columns) samples, in annotated regions as d. Top barplots show the number of corresponding DMRs within annotated regions, while the bottom plots depict the permutation test results (n = 1,000), with red, blue, and gray indicating DMRs enriched, depleted, and not significantly different within annotated regions, respectively. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **f**, Histogram of pancancer hyper-DMRs across each individual cancer compared to healthy controls within the SE (TCGE-CFMe-MCA) study. **g**, Enrichment of pancancer hyper-DMRs (as detailed for panel **e**). **h**, Distribution of pancancer hyper-DMRs across annotated genomic regions (top) and around the TSS (bottom). **i**, Histogram of cancer-specific hyper- (top) and hypo-DMRs (bottom) regions across the comparisons. **j**, Distribution of cancer-specific hyper- (top) and hypo-DMRs (bottom) around the TSS for the cancer types with more than 100 DMRs. Top-enriched KEGG terms (hypergeometric test with Benjamini–Hochberg adjusted *P* values) for genes associated with cancer-specific hyper-DMRs (**k**) and hypo-DMRs (**l**) located in promoter regions.

https://doi.org/10.1038/s43018-026-01116-3

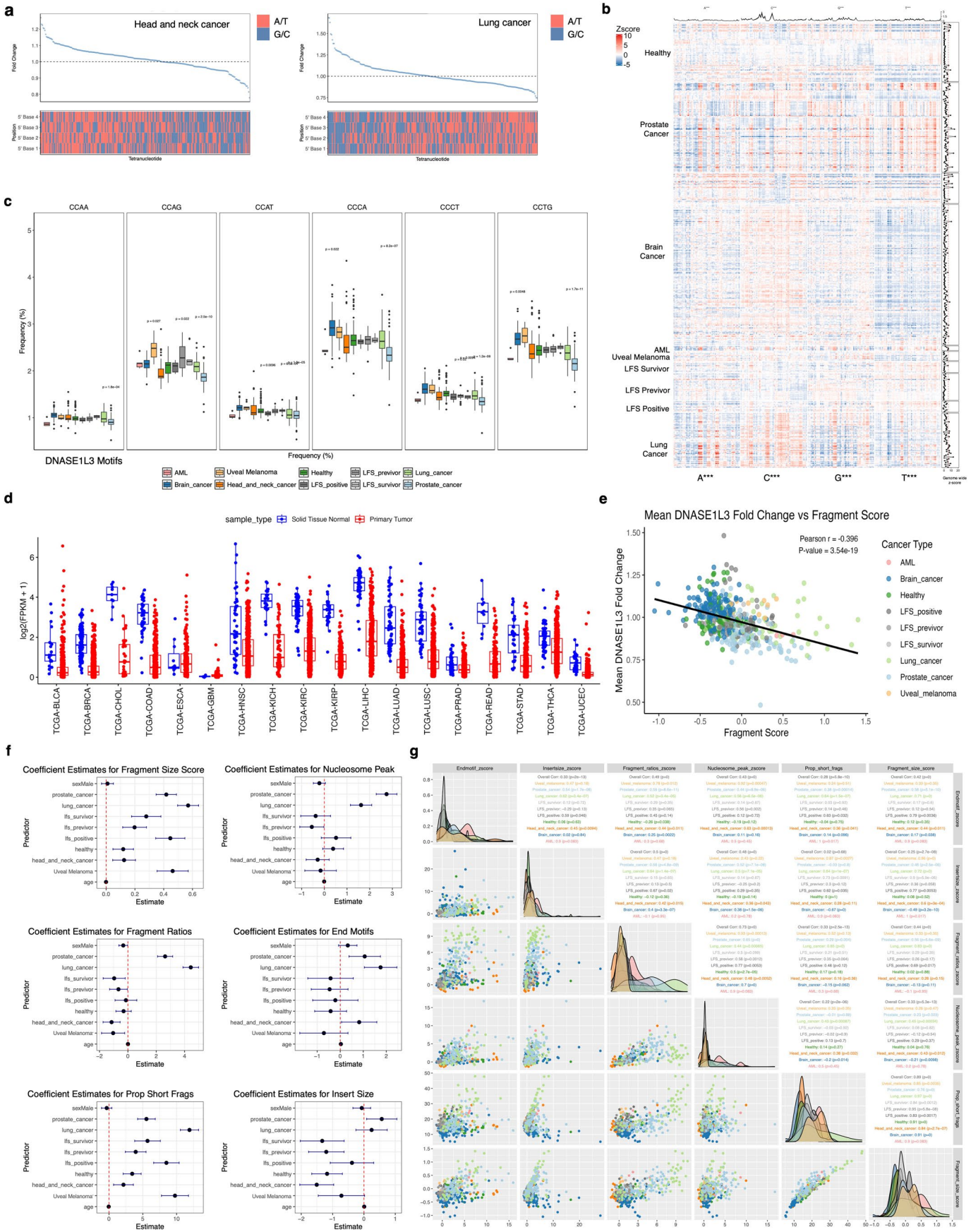


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 5 | Evaluation of fragment insert size and genome-wide short/long fragment ratios. a, Heatmap showing the proportion of fragments of various insert sizes. 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 the 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. Bottom right: the log change of the signature across fragment lengths, indicating the probability of a fragment being cancer or healthy across each length (samples: $n_{Healthy} = 46$, $n_{AML} = 4$, $n_{Lung\ cancer} = 42$, $n_{Uveal\ Melanoma} = 7$, $n_{Prostate\ cancer} = 68$, $n_{Head\ \&\ neck\ cancer} = 24$, $n_{Brain\ cancer} = 107$). 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 (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **c**, Patient-level fragmentation scores across all PE samples (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$, $n_{LFS\ positive} = 12$, $n_{LFS\ survivor} = 12$, $n_{LFS\ previvor} = 29$). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **d**, 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. **e**, Line plot showing differences in short/long fragment profiles across sample groups. The black line represents the median proportion of short/long fragments within each bin. Each red line represents an individual sample (no error bars are plotted), so every red trace is a single sample's proportion of short/long fragments. 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.

https://doi.org/10.1038/s43018-026-01116-3

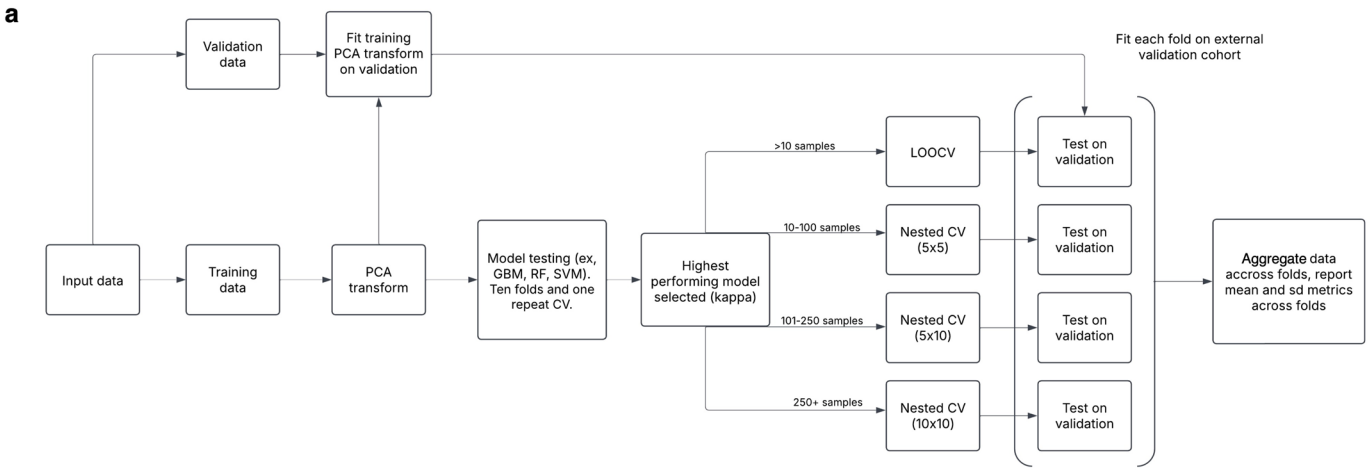


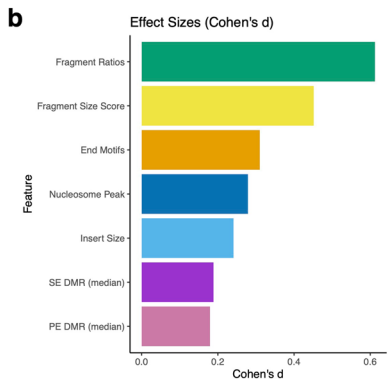
Extended Data Fig. 6 | See next page for caption.

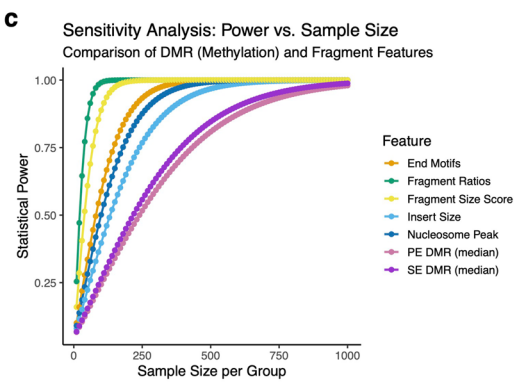
https://doi.org/10.1038/s43018-026-01116-3

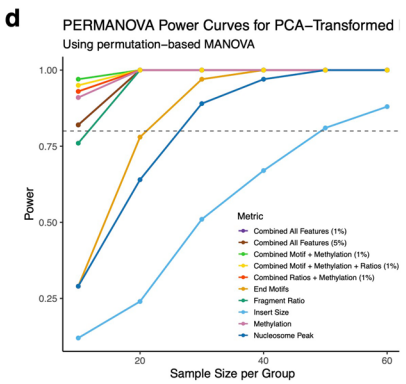
Extended Data Fig. 6 | Characterization of fragment 5′ end motif and relationship between fragmentomic features across cancer types. **a**, Median fold changes between head & neck cancer vs healthy controls and lung cancers vs healthy controls. The specific 5′ end motif being explored is plotted underneath and colored by base as either A/T or C/G. **b**, Heatmaps showing z-score differences in 5′ 4-mer end-motif frequencies between cancer and healthy controls for PE samples. 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 alphabetically from AAAA to TTTT. **c**, 5′ end-motif frequencies between healthy controls and cancer types at motifs associated with DNASE1L3 for PE samples. Box plots show the median (center line), the 25th and 75th percentiles (box edges), and whiskers that extend to the most extreme data points within 1.5 × the interquartile range; points beyond are outliers (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$, $n_{LFS\ positive} = 12$, $n_{LFS\ survivor} = 12$, $n_{LFS\ previvor} = 29$). Significance is based on Welch two-sample t-tests (healthy vs. cancer) with BH-adjusted p values. Exact *P* values: AML: CCAA ($P = 2.0 \times 10^{-19}$, FC = 1.20); CCAT ($P = 4.6 \times 10^{-16}$, FC = 1.15); CCCT ($P = 6.6 \times 10^{-14}$, FC = 1.31); CCAG ($P = 2.1 \times 10^{-2}$, FC = 1.07). Bladder_cancer: CCCT ($P = 1.0 \times 10^{-17}$, FC = 1.37); CCAT ($P = 5.6 \times 10^{-17}$, FC = 1.17); CCAA ($P = 4.3 \times 10^{-15}$, FC = 1.18); CCCA ($P = 1.6 \times 10^{-4}$, FC = 1.16). Breast_cancer: CCCT ($P = 1.3 \times 10^{-12}$, FC = 1.31); CCAA ($P = 3.2 \times 10^{-9}$, FC = 1.14); CCAT ($P = 3.8 \times 10^{-8}$, FC = 1.11); CCCA ($P = 3.3 \times 10^{-2}$, FC = 1.11). Colorectal_cancer: CCCT ($P = 2.7 \times 10^{-12}$, FC = 1.31); CCAA ($P = 9.0 \times 10^{-11}$, FC = 1.15); CCAT ($P = 1.5 \times 10^{-7}$, FC = 1.11). Lung_cancer: CCAA ($P = 3.0 \times 10^{-17}$, FC = 1.17); CCCT ($P = 8.8 \times 10^{-15}$, FC = 1.32); CCAT ($P = 2.0 \times 10^{-14}$, FC = 1.12); CCCA ($P = 1.8 \times 10^{-2}$, FC = 1.11). Pancreatic_cancer: CCAA ($P = 2.6 \times 10^{-2}$, FC = 0.92). Renal_cancer: CCCT ($P = 9.2 \times 10^{-17}$, FC = 1.37); CCAA ($P = 1.1 \times 10^{-13}$, FC = 1.16); CCAT ($P = 2.0 \times 10^{-11}$, FC = 1.14); CCCA ($P = 1.0 \times 10^{-3}$, FC = 1.15). **d**, *DNASE1L3* expression levels across cancer and normal tissues using TCGA data. Expression levels (log2(FPKM + 1)) are shown for various cancer types, with tumor samples in red and normal samples in blue. Sample sizes are: $n_{BLCA.Tumor} = 411$, $n_{BLCA.Normal} = 19$, $n_{BRCA.Tumor} = 1{,}097$, $n_{BRCA.Normal} = 113$, $n_{CHOL.Tumor} = 36$, $n_{CHOL.Normal} = 9$, $n_{COAD.Tumor} = 469$, $n_{COAD.Normal} = 41$, $n_{ESCA.Tumor} = 161$, $n_{ESCA.Normal} = 11$, $n_{GBM.Tumor} = 155$, $n_{GBM.Normal} = 5$, $n_{HNSC.Tumor} = 500$, $n_{HNSC.Normal} = 44$, $n_{KICH.Tumor} = 65$, $n_{KICH.Normal} = 24$, $n_{KIRC.Tumor} = 534$, $n_{KIRC.Normal} = 72$, $n_{KIRP.Tumor} = 288$, $n_{KIRP.Normal} = 32$, $n_{LIHC.Tumor} = 371$, $n_{LIHC.Normal} = 50$, $n_{LUAD.Tumor} = 524$, $n_{LUAD.Normal} = 59$, $n_{LUSC.Tumor} = 501$, $n_{LUSC.Normal} = 49$, $n_{PRAD.Tumor} = 498$, $n_{PRAD.Normal} = 52$, $n_{READ.Tumor} = 166$, $n_{READ.Normal} = 10$, $n_{STAD.Tumor} = 375$, $n_{STAD.Normal} = 32$, $n_{THCA.Tumor} = 502$, $n_{THCA.Normal} = 51$, $n_{UCEC.Tumor} = 547$, $n_{UCEC.Normal} = 35$). The statistically significant differences between tumor and normal tissues with two-sided Wilcoxon rank-sum test are: $P_{TCGA\text{-}BLCA} = 9.25 \times 10^{-8}$, $P_{TCGA\text{-}BRCA} = 8.20 \times 10^{-56}$, $P_{TCGA\text{-}CHOL} = 1.02 \times 10^{-7}$, $P_{TCGA\text{-}COAD} = 1.98 \times 10^{-22}$, $P_{TCGA\text{-}ESCA} = 0.80$, $P_{TCGA\text{-}GBM} = 0.14$, $P_{TCGA\text{-}HNSC} = 1.22 \times 10^{-6}$, $P_{TCGA\text{-}KICH} = 8.32 \times 10^{-11}$, $P_{TCGA\text{-}KIRC} = 6.28 \times 10^{-35}$, $P_{TCGA\text{-}KIRP} = 9.51 \times 10^{-20}$, $P_{TCGA\text{-}LIHC} = 5.63 \times 10^{-25}$, $P_{TCGA\text{-}LUAD} = 8.53 \times 10^{-30}$, $P_{TCGA\text{-}LUSC} = 1.39 \times 10^{-22}$, $P_{TCGA\text{-}PRAD} = 1.14 \times 10^{-5}$, $P_{TCGA\text{-}READ} = 5.52 \times 10^{-7}$, $P_{TCGA\text{-}STAD} = 7.54 \times 10^{-10}$, $P_{TCGA\text{-}THCA} = 1.07 \times 10^{-8}$, $P_{TCGA\text{-}UCEC} = 4.27 \times 10^{-17}$. Box plots show the median (center line), the 25th and 75th percentiles (box edges), and whiskers that extend to the most extreme data points within 1.5 × the interquartile range. **e**, Scatter-plot showing the relationship between the average weighted FS and the average DNASE1L3 fold change across multiple cancer types. Each point represents a sample, colored according to cancer type as labeled in the legend. The black line indicates a linear regression fit, illustrating the overall trend between FS and DNASE1L3 variation. The association was assessed using a Pearson correlation test with a simple linear regression line shown (Pearson $r = -0.396$, $P = 3.54 \times 10^{-19}$, samples $n = 473$). **f**, These six panels show coefficient estimates for age, sex, and multiple cancer-type categories in linear models of six fragmentomic features (weighted FS, fragment ratios, proportion of short fragments, nucleosome peak distances, 5′ end motifs, and insert size, n = 316 samples with complete information). Each point is centered at the model's estimated coefficient, with error bars showing ±1 standard error. Z-scores for the features were calculated relative to healthy controls. The red dashed line at zero marks no effect; coefficients for age, sex, or cancer-type categories that differ from zero indicate an association with the feature beyond the reference (healthy) group. **g**, Correlation matrix showing the Spearman correlation between different fragmentomic features across PE samples. Features include genome-wide z-scores for 5′ end motifs, insert size, fragment ratios, nucleosome peak distances, as well as the proportion of short fragments and weighted FS. 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 corresponding p values. Spearman correlation was used since not all groups achieved the normality conditions required for Pearson correlation to be used.

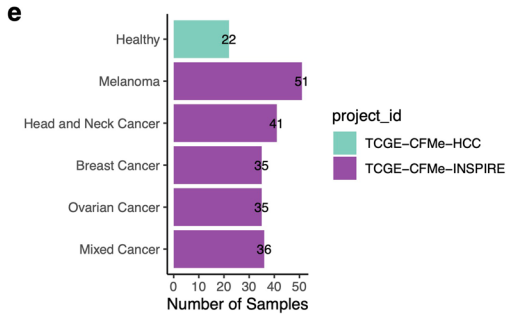
https://doi.org/10.1038/s43018-026-01116-3

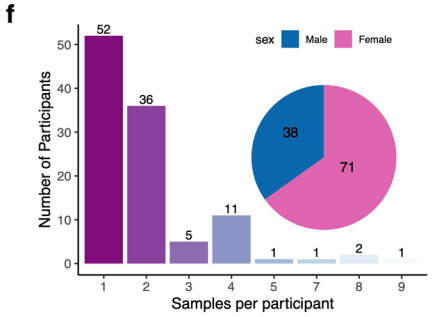


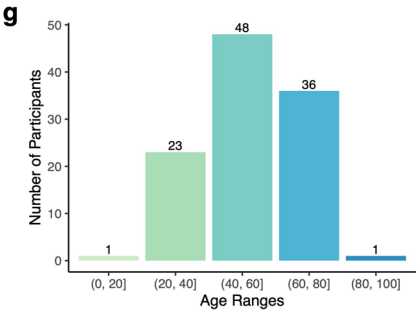


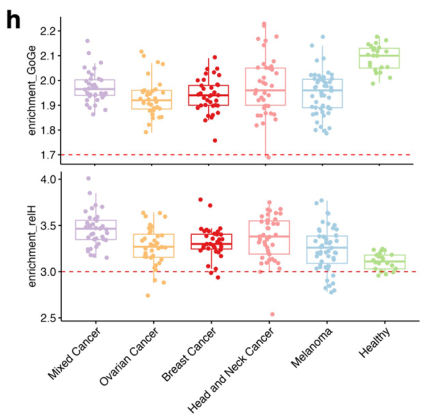


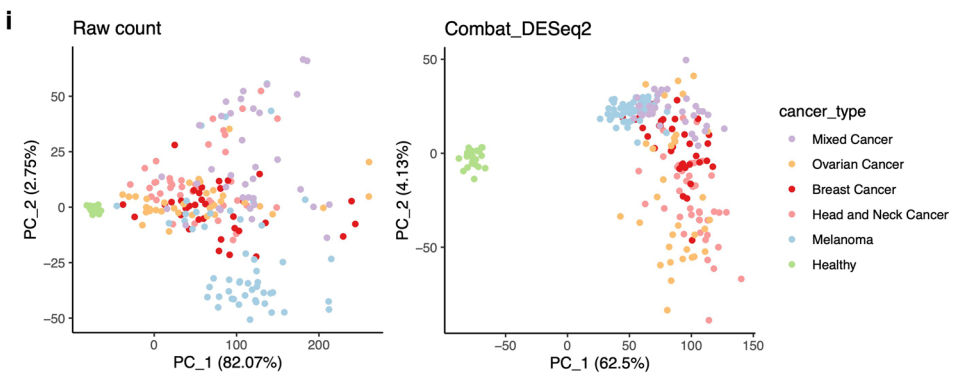










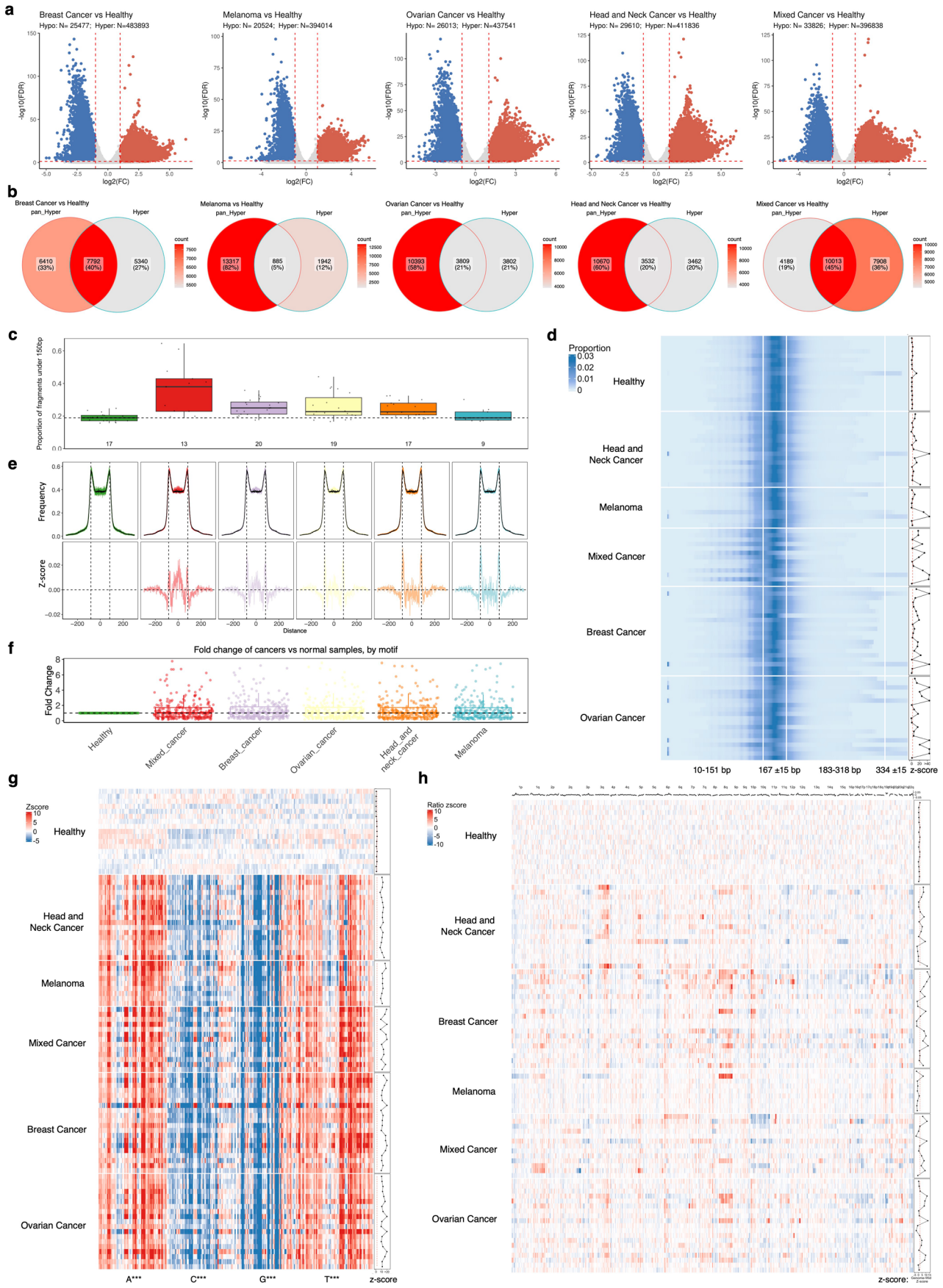


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 7 | Power and variability of methylation and fragmentomic features, classification workflow, and overview of external independent validation dataset. a, Flowchart for the classification models' training, testing, and validation. **b**, Effect sizes for the methylation and fragmentomic features. The bar chart displays the absolute Cohen's *d* values for fragmentomic features, calculated using genome-wide *z*-scores, between cancer samples in the primary PE dataset (total samples n = 356 [n_{AML} = 5, $n_{Lung\ cancer}$ = 59, $n_{Uveal\ Melanoma}$ = 10, $n_{Prostate\ cancer}$ = 97, $n_{Head\ \&\ neck\ cancer}$ = 33, $n_{Brain\ cancer}$ = 152] and healthy controls n = 64). The point for DMRs indicates the median absolute effect size across n = 14,202 differentially methylated regions identified between the same cancer and healthy cohorts. These values quantify the magnitude of differences between cancer and healthy samples for each feature type. **c**, The combined sensitivity plot illustrates statistical power as a function of sample size per group, based on two-sample *t*-test power calculations, using the same sample set as in (**b**). For DMRs, power was estimated using the median effect size across specific differentially methylated regions. For fragmentomic features, the effect size of the genome-wide *z*-score was used. The plot reflects the power to detect a representative DMR or a significant deviation in genome-wide fragmentation patterns between cancer and healthy samples. **d**, PERMANOVA-based power analysis of PCA-transformed fragmentomic and methylation features. We simulated statistical power across varying sample sizes using a permutation-based multivariate analysis of variance (PERMANOVA) applied to PCA-transformed features to define the cumulative power across that feature for classification using the same sample set as in (b). Each curve represents a different feature set or combination, with the y-axis denoting statistical power and the x-axis showing the number of samples per group. **e**, Number of samples categorized by sample type, along with the breakdown of sample numbers from individual validation cohorts ($n_{Mixed\ cancer}$ = 36, $n_{Ovarian\ cancer}$ = 35, $n_{Breast\ cancer}$ = 35, $n_{Head\ \&\ neck\ cancer}$ = 41, $n_{Melanoma}$ = 51, $n_{Healthy}$ = 22). **f**, Distribution of the number of samples per participant (participants: n_1 = 52, n_2 = 36, n_3 = 5, n_4 = 11, n_5 = 1, n_7 = 1, n_8 = 2, n_9 = 1), accompanied with the sex composition for all participants (n_{Male} = 38, n_{Female} = 71). **g**, Number of participants in different age ranges ($n_{(0,20)}$ = 1, $n_{(20,40)}$ = 23, $n_{(40,60)}$ = 48, $n_{(60,80)}$ = 36, $n_{(80,100)}$ = 1). **h**, Distribution cfMeDIP-seq enrichment scores, GoGe (top) and relH (bottom), grouped by cancer type. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × interquartile range (IQR) (samples: $n_{Mixed\ cancer}$ = 36, $n_{Ovarian\ cancer}$ = 35, $n_{Breast\ cancer}$ = 35, $n_{Head\ \&\ neck\ cancer}$ = 41, $n_{Melanoma}$ = 51, $n_{Healthy}$ = 22). **i**, The PCA plots of raw (left) and ComBat-seq + DESeq2 normalized count (right).

https://doi.org/10.1038/s43018-026-01116-3

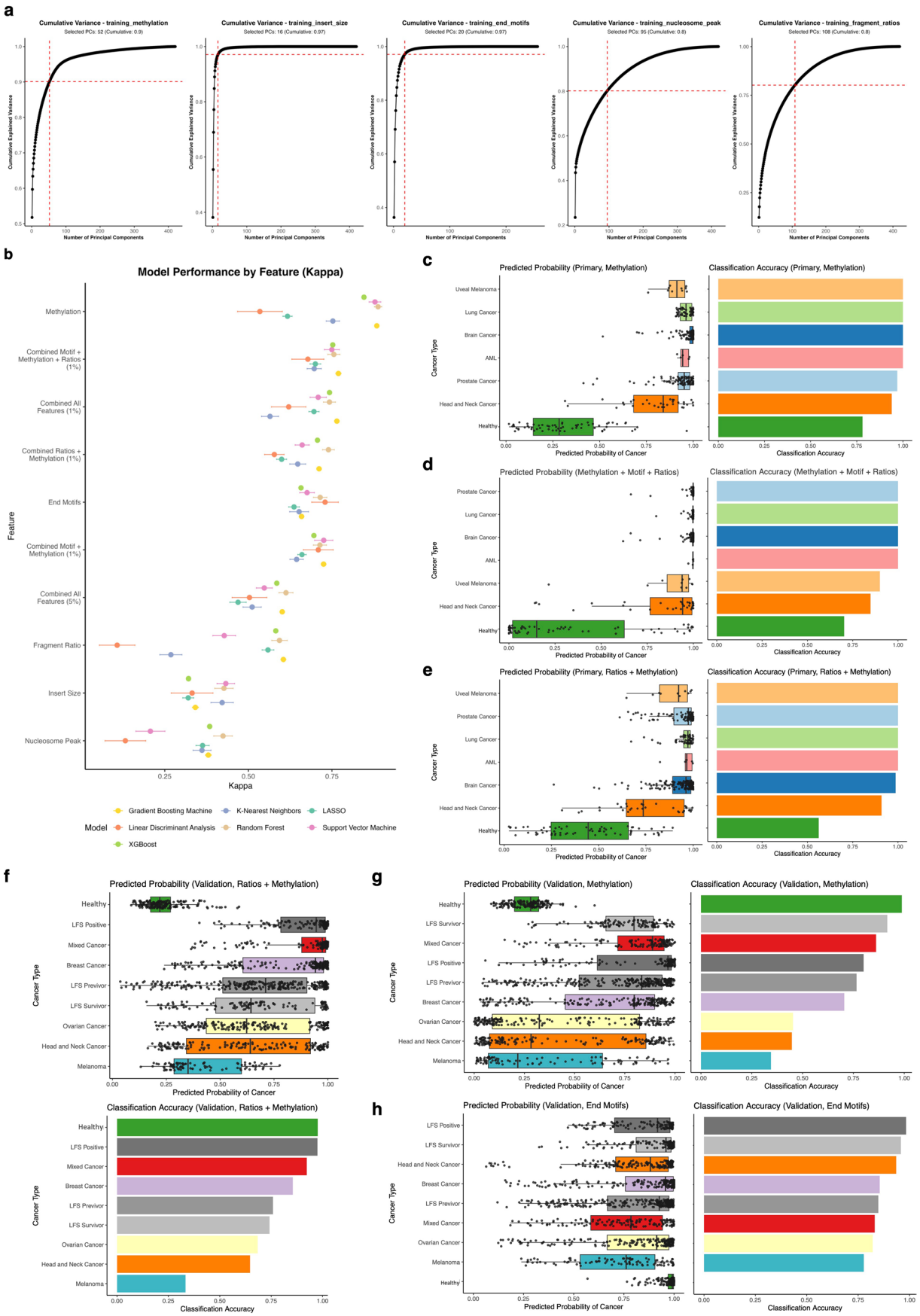


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 8 | Methylation and fragmentomic features of the external independent validation cohorts. a, The volcano plots for DMRs before filtering in the comparison of each cancer type against the healthy control from validation cohorts. **b**, Overlap of the hyper-DMRs identified from validation cohorts and the predefined pancancer hyper-DMRs (pan_Hyper). **c**, The proportion of short fragments between 20:150 bp/20:600 bp in length across cancer types. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR (samples: $n_{Healthy} = 17$, $n_{Mixed\ cancer} = 13$, $n_{Ovarian\ cancer} = 19$, $n_{Breast\ cancer} = 20$, $n_{Head\ \&\ neck\ cancer} = 17$, $n_{Melanoma} = 9$). **d**, 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. **e**, 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. **f**, Median fold changes of 5′ end motifs relative to median frequencies of healthy controls. Each point represents a 4-mer end motif, with 256 possible motifs included per cancer type (derived from n = 95 total patients in the validation cohort). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR. **g**, Heatmap showing z-score differences in 5′ 4-mer end motif frequencies between cancer 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 alphabetically from AAAA to TTTT. **h**, Heatmap showing z-score differences in the proportion of short/long fragments across 5 Mb bins between cancer 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.

https://doi.org/10.1038/s43018-026-01116-3

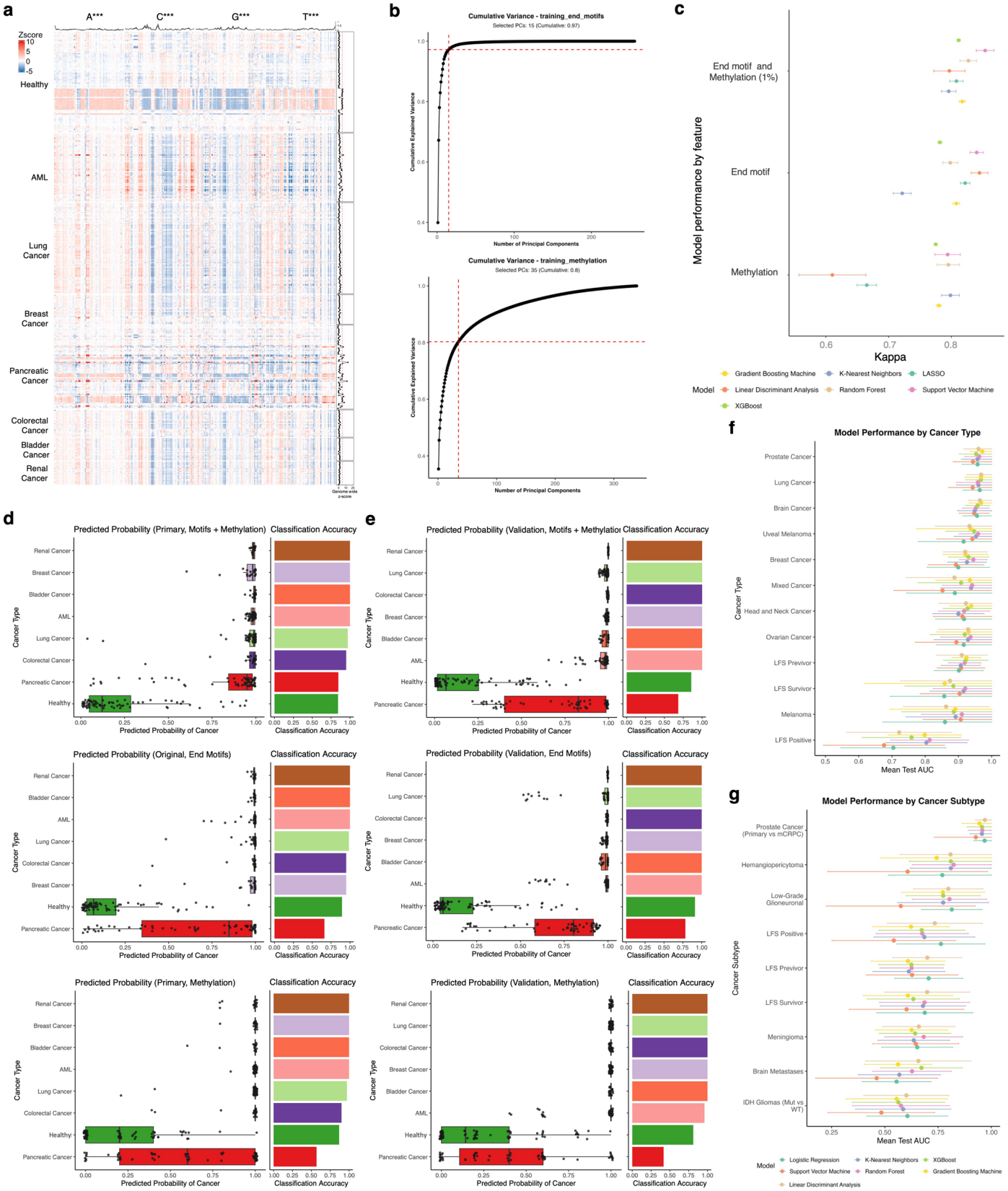


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 9 | Training, testing, and independent validation of cancer vs healthy classification models in PE datasets. a, The plots show the number of PCs to retain for downstream analysis by explaining corresponding variance for methylation signatures and different fragmentomic features, including insert size, 5′ end motifs, nucleosome peak distances, and fragment ratios. 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. **b**, Model performance, measured using Kappa, across seven different algorithms trained on methylation and/or fragmentomic features. Points show the mean Cohen's κ across resamples; horizontal bars indicate 95% CI. Kappa was calculated on all paired-end samples from the primary PE dataset (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$) using a repeated 10-fold cross-validation with class balancing by down-sampling. **c**, Distribution of out-of-fold predicted cancer probabilities (left) and per-cancer-type classification accuracy (right) for the primary PE dataset using the methylation-only model, stratified by cancer type. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$). **d**, Same as (**c**), but using a combined model of 5′ end-motif features, fragment ratios, and methylation (n = 420 samples). **e**, Same as (**c**), but using a combined fragment ratio + methylation model (n = 420 samples). **f**, Predicted probability of cancer (top) and classification accuracy (bottom) on the validation dataset for the combined fragment ratio + methylation model, stratified by cancer type. Predicted cancer probability (right) and accuracy (left) for the methylation-only model are shown for the same validation set (n = 140 samples), with each sample generating one out-of-fold prediction in each of the 10 cross-validation folds (10 predictions per sample). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5 × IQR (samples: $n_{Healthy} = 64$, $n_{AML} = 5$, $n_{Lung\ cancer} = 59$, $n_{Uveal\ Melanoma} = 10$, $n_{Prostate\ cancer} = 97$, $n_{Head\ \&\ neck\ cancer} = 33$, $n_{Brain\ cancer} = 152$). **g**, Same as (**f**) for the methylation-only model. **h**. Same as (**f**) for the 5′ end-motif–only model.

https://doi.org/10.1038/s43018-026-01116-3



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s43018-026-01116-3

Extended Data Fig. 10 | Performance of cancer vs. healthy classifiers using the TCGE-CFMe-MCA study (SE data) and comparison of cancer type and subtype classification models. a, Heatmaps showing z-score differences in 5′ 4-mer end motif frequencies between cancer and healthy controls for SE samples. **b**, PCA for methylation and 5′ end motifs feature selection, showing the cumulative variance explained by each principal component in the training set (red dashed line indicates the number of components selected). **c**, The model performance was measured by Kappa, using seven different algorithms across various methylation and fragmentomic features, either alone or in combination. Points show the mean Cohen's κ across resamples; horizontal bars indicate 95% CI. Kappa was calculated on all paired-end samples from the primary SE dataset (n = 339) with repeated 10-fold cross-validation with class balancing by down-sampling. **d**, Predicted probability of cancer (left) and classification accuracy (right) for the primary SE dataset (total samples: n = 339 [$n_{Healthy}$ = 76, n_{AML} = 53, $n_{Bladder\ cancer}$ = 17, $n_{Breast\ cancer}$ = 20, $n_{Colorectal\ cancer}$ = 20, $n_{Lung\ cancer}$ = 70, $n_{Pancreatic\ cancer}$ = 65, $n_{Renal\ cancer}$ = 18]), stratified by cancer type, using methylation (top), 5′ end motifs (middle), and their combination (bottom) in the primary SE dataset. Each data point corresponds to an out-of-fold prediction in each of the 10 cross-validation folds (10 predictions per sample). Box plots show the median (center line), the 25th and 75th percentiles (box edges), and whiskers that extend to the most extreme data points within 1.5 × the interquartile range; points beyond are outliers. **e**, Predicted probability of cancer (left) and classification accuracy (right) for the validation SE dataset (n = 39 samples [$n_{Healthy}$ = 9, n_{AML} = 5, $n_{Bladder\ cancer}$ = 2, $n_{Breast\ cancer}$ = 5, $n_{Colorectal\ cancer}$ = 3, $n_{Lung\ cancer}$ = 7, $n_{Pancreatic\ cancer}$ = 6, $n_{Renal\ cancer}$ = 2]), with each sample generating one out-of-fold prediction in each of the 10 cross-validation folds (10 predictions per sample), stratified by cancer type. Shown is methylation (top), 5′ end motifs (middle), and the combination of 5′end motifs and methylation (bottom). Box plots show the median (center line), the 25th and 75th percentiles (box edges), and whiskers that extend to the most extreme data points within 1.5 × the interquartile range; points beyond are outliers. **f**, Mean test AUCs for models trained to distinguish each cancer type from other cancer types across various machine-learning classifiers (colors), ordered by decreasing AUC, with error bars representing the standard deviation across test sets (n = 465, 12 cancer types across both primary and validation PE datasets). **g**, Mean test AUCs for models trained to classify each cancer subtype (n = 295 samples across 11 cancer subtypes included [6 brain cancer, 3 LFS, and 2 prostate cancer subtypes]), compared to other similar cancer subtypes using the same classifiers, with error bars representing the standard deviation (±1 SD) across test sets.

nature portfolio

Corresponding author(s): Housheng Hansen He
Trevor Pugh
Mathieu Lupien
Yong Zeng

Last updated by author(s): Jan 7, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specific software was used for data collection in this study.
Data analysis	All data analysis details, including method and detailed parameters are described in Methods. 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. The following software or packages were used: MEDIPIPE (version 1.1.0), ComBat-seq (sva version 3.52.0), DESeq2 (version 1.44.0), MEDEStrand (version 0.0.0.9000), QSEA (Beta, version 1.20.0), methyclock (version 1.4.0), annotatr (version 1.24.0), ClusterProfiler (version 4.6.2), missMethyl (version 1.32.1), ChIPseeker (version 1.34.1), Picard(version 4.0.1.2), pwr.t.test (v.1.3-0), PERMANOVA (adonis2, vegan v.2.6-10), mvnormtest (v0.1.9.3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw cfMeDIP-seq data that support the findings of this study from TCGE-CFMe-MCA, TCGE-CFMe-BCA, TCGE-CFMe-HNSC, and TCGE-CFMe-SCLC were obtained directly from the corresponding authors of the respective cohorts upon request. TCGE-CFMe-MCA and TCGE-CFMe-BCA data can be requested from Dr. Daniel D. De Carvalho (ddecarv@uhnresearch.ca), TCGE-CFMe-HNSC data from Dr. Scott V. Bratman (scott.bratman@rmp.uhn.ca), and TCGE-CFMe-SCLC data from Dr. Benjamin H. Lok (benjamin.lok@rmp.uhn.ca) via submission of a data access application. The other cfMeDIP-seq datasets were deposited in the European Genome-Phenome Archive (EGA): TCGE-CFMe-AML (EGAS00001005069), TCGE-CFMe-PRAD (EGAS00001005522), TCGE-CFMe-UM (EGAD00001008998), TCGE-CFMe-HBC (EGAS00001006539), TCGE-CFMe-LFS (EGAS00001006539), TCGE-CFMe-INSPIRE (EGAD00001011312), and TCGE-CFMe-HCC (EGAD50000000652). Access to all seven UHN-generated datasets is made available upon completion of the required data access agreement, which will be reviewed by the UHN genomics data access committee (dac@uhn.ca). Data access will be granted to qualified investigators for appropriate and compliant use. The source data for DMRs analyses (Fig. 2b; Extended Data Fig. 3e, f; Extended Data Fig. 4d, e, g; Extended Data Fig. 48a) and methylation-based PCA and UMAP plots (Fig. 2a; Extended Data Fig. 2a-c, f; Extended Data Fig. 3a, c; Extended Data Fig. 7i), together with BED files for pan-cancer and cancer-specific DMRs as well as age- and sex-associated regions, are available on Zenodo (https://zenodo.org/records/15191455). All other source data is provided in this paper. The remaining data are available within the article and Supplementary Tables.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	For samples collected in this study, all sex information was self-reported. For the primary dataset, samples came from 390 male donors and 305 female donors. The remaining 223 individuals lacked study-reported sex information. For the validation dataset, sample were from 38 male and 71 female donors. Although potential confounding effects of sex was evaluated, it was not included as covariate in our analyses.
Reporting on race, ethnicity, or other socially relevant groupings	None of this information was included in this study.
Population characteristics	The population characteristics (age and clinical features) and summary for our primary and validation samples were provided in Supplementary Table 1 and Supplementary Table 9, respectively.
Recruitment	Recruitment is not applicable in this case as all datasets analysed here were derived from published studies.
Ethics oversight	All samples obtained in this study complied with the relevant ethical regulations approved by the institutional ethics committee and Research Ethics Board at the University Health Network (UHN).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not predetermine by any statistical methods. Sample size was based on sample availability, and was demonstrated the sufficient power to obtain definitive results.
Data exclusions	All data were uniformly processed, provided as a resource. While the primary analyses in this study were restricted to the high quality and baseline time point samples.
Replication	One replication per sample was obtained
Randomization	The work requires no randomization: Human specimen were allocated into groups according to disease types

Blinding | Blinding is not relevant as no treatment for patient is needed

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks — *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes — *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication — *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*