Sensitive tumour detection and classification using plasma cell-free DNA methylomes

Shu Yi Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael H. A. Roehrl, Dianne Chadwick, Philip C. Zuzarte, Ayelet Borgida, Ting Ting Wang, Tiantian Li, et al.

Sensitive tumor detection and classification using plasma cell-free DNA methylomes

Shu Yi Shen[1*], Rajat Singhania[1*], Gordon Fehringer[2*], Ankur Chakravarthy[1*], Michael H. A. Roehrl[1,3,4], Dianne Chadwick[1], Philip C. Zuzarte[5], Ayelet Borgida[2], Ting Ting Wang[1,4], Tiantian Li[1], Olena Kis[1], Zhen Zhao[1], Anna Spreafico[1], Tiago da Silva Medina[1], Yadon Wang[1], David Roulois[1,6], Ilias Ettayebi[1,4], Zhuo Chen[1], Signy Chow[1], Tracy Murphy[1], Andrea Arruda[1], Grainne M. O'Kane[1], Jessica Liu[4], Mark Mansour[4], John D. McPherson[7], Catherine O'Brien[1], Natasha Leighl[1], Philippe L. Bedard[1], Neil Fleshner[1], Geoffrey Liu[1,4,8], Mark D. Minden[1], Steven Gallinger[9,10], Anna Goldenberg[11], Trevor J. Pugh[1,4], Michael M. Hoffman[1,4,11], Scott V. Bratman[1,4], Rayjean J. Hung[2,8§], Daniel D. De Carvalho[1,4,12§]

1. Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
2. Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada
3. Memorial Sloan Kettering Cancer Center, New York, NY
4. Department of Medical Biophysics, University of Toronto, Toronto, Canada
5. Genome Technologies, Ontario Institute for Cancer Research, Toronto, Canada
6. UMR_S 1236, Univ Rennes 1, Inserm, Etablissement Français du sang Bretagne, F-35000 Rennes, France
7. Department of Biochemistry and Molecular Medicine, UC Davis Comprehensive Cancer Center, Sacramento, CA, USA

8. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
9. Fred Litwin Centre for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada
10. Department of Surgery, Toronto General Hospital, Toronto, Ontario, Canada
11. Department of Computer Science, University of Toronto, Toronto, Canada
12. Lead Contact

*These authors made equal contributions

[§]Corresponding authors:

Daniel D. De Carvalho: ddecarv@uhnresearch.ca

Rayjean J. Hung: rayjean.hung@lunenfeld.ca

The use of liquid biopsies for cancer detection and management is rapidly gaining prominence[1]. Current circulating tumor DNA (ctDNA) detection methods involve sequencing somatic mutations using cell-free DNA (cfDNA), but their sensitivity may be low among early-stage cancer patients given the limited availability of recurrent mutations[2-5]. In contrast, large-scale epigenetic alterations, which are tissue and cancer-type specific are not similarly constrained[6], thus potentially have enhanced ability to detect and classify cancers in early-stage patients. Here, we developed a sensitive, immunoprecipitation-based protocol for methylome analysis of small quantities of circulating cfDNA and demonstrated the ability to detect large-scale DNA methylation changes that are enriched for tumor-specific patterns. We also demonstrated robust performance in cancer detection and classification across an extensive collection of plasma samples from multiple tumor types, setting the stage for the development of minimally invasive biomarkers for early cancer detection, interception, and classification based on plasma cfDNA methylation patterns.

Analysis of ctDNA has numerous potential clinical applications, but certain settings such as cancer screening and detection of minimal residual disease after treatment require a degree of analytical sensitivity that is often beyond current technical limits of mutation-based ctDNA detection methods. The major hurdles to improved sensitivity of these methods include (1) the limited number of recurrent mutations available to distinguish between tumor and normal circulating cfDNA in a cost-effective manner, and (2) technical artefacts (errors) introduced during sequencing. We reasoned that specific enrichment of methylated DNA fragments from cfDNA could overcome both of these hurdles.

To assess whether the higher number of DNA methylation changes in cancers could translate to increased sensitivity at lower sequencing costs, we performed bioinformatic simulations examining detection probability across varying numbers of Differentially Methylated Regions (DMRs), coverage, and ctDNA abundance (Fig. 1a and Extended Data Fig. 1a). We found improved sensitivity as the number of DMRs increased, even at lower sequencing depth and ctDNA abundance, suggesting the recovery of cancer-specific DNA methylation changes could allow for highly sensitive and low-cost detection, classification and monitoring of cancer.

However, this is challenging in practice due to the low-abundance and fragmented nature of plasma cfDNA[3], which has restricted most of the previous plasma methylation profiling to locus-specific PCR-based assays[7-9]. While Whole-Genome Bisulfite Sequencing (WGBS) of cfDNA has been attempted[10,11], this approach is inefficient due to degradation of ~84-96% of the input DNA during bisulfite conversion[12], high-costs, and limited information recovery given the low genome-wide abundance of CpGs. Therefore, we developed cfMeDIP-seq (cell-free Methylated DNA Immunoprecipitation and high-throughput sequencing) for genome-wide bisulfite-free plasma DNA methylation profiling, based on its ability to enrich for CpG rich, potentially more-informative fragments, thus enhancing cost-effectiveness.

Briefly, we optimized an existing low-input MeDIP-seq protocol[13] that is robust down to 100 ng input DNA, using exogenous *Enterobacteria phage* λ DNA (filler DNA) to inflate starting amounts (Extended Data Fig. 1b). This is crucial for applications based on plasma cfDNA samples, which yield much less than 100 ng of cfDNA. We then performed extensive benchmarking of the optimized protocol. First, comparing low-input cfMeDIP-seq versus gold-standard MeDIP-seq using colorectal cancer (CRC) HCT116 DNA sheared to mimic cfDNA

showed robust CpG-enrichment (Extended Data Fig. 2a-c) and inter-replicate correlation (Extended Data Fig. 2d). cfMeDIP-seq (1 to 10 ng input DNA) also recapitulated profiles from gold-standard MeDIP (100 ng), RRBS (Reduced Representation Bisulfite Sequencing) (1,000 ng) and WGBS (2,000 ng) (Extended Data Fig. 2e).

Next, cfMeDIP-seq was compared to ultra-deep, unique molecular identifiers (UMIs) based, hybrid capture mutation sequencing[14] across a serial dilution of CRC DNA into multiple myeloma (MM) MM.1S cell-line DNA (Extended Data Fig. 3a). The former showed near-perfect linear associations between observed and expected numbers of DMRs (5% False Discovery Rate (FDR) threshold) and signals within DMRs, down to 0.001% dilution (both r^2=0.99, $p < 0.0001$) (Fig. 1b and Extended Data Fig. 3b-e). The latter, however, detected CRC specific mutations only down to 0.1% and 1% with single strand consensus sequence (SSCS) and duplex consensus sequence (DCS) respectively (Extended Data Fig. 3f-g). This highlights the excellent analytical sensitivity of cfMeDIP-seq for the detection of cancer-derived DNA. We also evaluated the ability of cfMeDIP-seq to enrich ctDNA through biased sequencing of CpG-rich sequences that are frequently hypermethylated in cancer when compared to normal tissue[15]. Plasma from mice harboring patient-derived xenografts (PDX) was used for cfMeDIP-seq, and a 2-fold enrichment of human tumor-derived cfDNA was found following immunoprecipitation as compared to the input sample (Fig. 1c).

To investigate whether cfMeDIP-seq could detect ctDNA in early-stage cancer, we generated cfMeDIP-seq profiles from pre-surgery plasma cfDNA of 24 primary early-stage pancreatic cancer (PDAC) patients (cases) and 24 age and sex-matched healthy controls (controls) (Fig. 2a and Extended Data Fig. 4a-f). In addition to plasma cfDNA, the microdissected primary tumors

and adjacent normal tissue from the same PDAC patients were used to generate DNA methylation profiles via RRBS. We identified 14,716 DMRs between cases and controls cfDNA (9,931 hypermethylated in cases, 4,785 in controls, based on negative-binomial generalized linear model (GLM) of fragment counts at significance level of Benjamini Hochberg FDR (BHFDR) of 0.1) (Fig. 2b-c and Supplementary Table 1).

In comparison, 45,173 Differentially Methylated CpGs (DMCs) were found between tumor and normal tissue in RRBS data (Supplementary Table 2). Permutation testing to estimate the significance of overlaps between cfMeDIP-seq cell-free DMRs and RRBS tissue DMCs revealed significant enrichment for DMR/DMC pairs concordantly hypermethylated ($p=3.39e-47$), and concordantly hypomethylated in case cfDNA and tumor tissue ($p=1.43e-22$). This significant enrichment was not observed in the discordant methylation pattern between cfDNA and tumor DNA (Fig. 2d). Furthermore, signals in overlapping plasma cfDNA and tissue DNA methylation were correlated (Extended Data Fig. 5a). These findings suggested that cfMeDIP-seq of plasma cfDNA could detect tumor-derived DNA methylation events in ctDNA.

As non-tumor derived cfDNA is mostly released from blood cells, we performed similar permutation-based enrichment testing between case-vs-control cfMeDIP-seq DMRs and the 95,388 RRBS DMCs between PDAC tumor tissue (n=24) and normal peripheral blood mononuclear cell (PBMCs) (n=5) (Supplementary Table 3). Again, we observed significant enrichment for concordant hypermethylated ($p<1e-745$) and hypomethylated ($p=6.12e-82$) sites in cfMeDIP-seq DMRs and tumor vs. PBMC DMCs, while discordant calls were underrepresented (Fig. 2e). In addition, signals in overlapping DMRs/DMCs were correlated

(Extended Data Fig. 5b) and altogether indicated that DMRs identified using cfMeDIP-seq, between cases and controls, were likely derived from ctDNA (Extended Data Fig. 5c).

Based on the enrichment of tumor-derived DMRs and the known methylation-specific variable binding of transcription factors (TFs)[16], we hypothesized that cfMeDIP-seq methylomes could identify active transcriptional networks in tumors or other tissues using plasma cfDNA. Upon motif enrichment analysis on cfMeDIP-seq DMRs and incorporating methylation preferences of candidate transcription factors into account[16], we identified 42 TFs as binding in healthy controls and 52 as binding in pancreatic cancer cases (Supplementary Table 4-5). As expected, the former included hematopoietic-lineage specific TFs such as PU.1, NFE2, and GATA1, while the latter included pancreas-associated TFs, PTF1a, Onecut1 (HNF6), and NR5A2 (Extended Data Fig. 6a and c). Compared to random sets of TFs, TFs inferred as active in healthy controls are overexpressed in blood based on GTEx data, while those inferred as active in pancreatic cancer cases are overexpressed in pancreatic tissues in GTEx and PDAC tissue in TCGA (Extended Data Fig. 6b, d, and e). Collectively, these findings indicated that cfMeDIP-seq might permit non-invasive characterization of active TF-networks in cancer.

Given that we could detect tumor-specific DMRs in the plasma of PDAC cases relative to controls, we then investigated whether cfMeDIP-seq could non-invasively classify multiple cancer types from healthy controls. Consequently, we performed cfMeDIP-seq in a discovery cohort of 189 plasma samples from 7 different tumor types (PDAC, colorectal cancer (CRC), breast cancer (BRCA), lung cancer (LUC), renal cancer (RCC), bladder cancer (BLCA), and AML) and healthy controls (Extended Data Fig. 7a-l and Extended Data Fig. 8a).

We first identified plasma cell-free DMRs for each tumor type relative to healthy controls and asked if these cancer type-specific DMRs identified on the plasma cfDNA were enriched for the expected tumor DMRs for each cancer type using tumor tissue methylation data from TCGA (n=4032) (Fig. 3a). We observed a marked enrichment of sites hypermethylated in the primary tumor tissue (TCGA), within the regions we identified as hypermethylated in the plasma cfDNA for each cancer type, coupled with significantly correlated signals between cfMeDIP-seq plasma methylation and TCGA 450k tumor data (Extended Data Fig. 8b-h). These results indicate the ability to recover ctDNA-associated methylation profiles across a range of cancer types.

Finally, we carried out a set of machine learning analyses on our discovery cohort to rigorously evaluate the utility of cfMeDIP profiles in cancer detection and classification. We initially reduced our dataset to 505,027 windows mapping to CpG islands, shores, shelves and FANTOM5 enhancers for computational efficiency. Unbiased performance estimates, while accounting for training-set biases, were then derived from the reduced dataset. We split the discovery cohort 80%-20% into balanced training and test sets. Using only training set samples, we selected the top 300 DMRs by limma-trend test statistic for each class versus other classes. We then trained a series of one-vs-other-classes regularised binomial GLMs using these features on the training set data. The training procedure consisted of 3 rounds of 10-Fold Cross-Validation (CV) across a grid of values for alpha and lambda with optimisation for Cohen's Kappa. The use of multiple rounds of 10-Fold CV was motivated by a desire to leverage additional randomisation for more generalisable model tuning.

The performance of these classifiers was then evaluated using receiver operating characteristic (ROC) statistics derived from the test-set samples that were not used for either DMR selection or model training. The whole process was repeated 100 times to prevent training-set biases[17],

171 culminating in a collection of 800 models, with 100 models for each one-vs-all-others
172 comparison (hereby termed E100). High area under the receiver operator characteristic curve
173 (AUROC) values were observed for test set samples across classes (Fig. 3b and Extended Data
174 Fig. 9a).
175 Subsequently, we assessed performance across batches by applying the ensemble to a 199-
176 sample validation cohort (35 AMLs, 47 PDACs, 55 LUC and 62 healthy controls). Averaging
177 the class probabilities output by E100 for each sample, yielded high ROCs for AML vs. others
178 (0.980), PDAC vs. others (0.918), LUC vs. others (0.971) and normal vs. others (0.969) (Fig. 3c).
179 Notably, performance was similar between early and late stage samples, suggesting applicability
180 to cancer early detection (Fig. 3d and Extended Data Fig. 9b).
181 We then investigated whether the DMRs (non-zero coefficients) selected during the training of
182 E100 were tumor-specific. Visualization using t-distributed stochastic neighbour embedding
183 (tSNE) plots showed clear separation by tumor type in the plasma cohort (Fig. 4a). This was
184 notably reproduced in the 450k dataset of 4,032 TCGA cancers and normal blood samples and
185 400 COSMIC cancer cell lines and PBMCs (Fig. 4b-c). This suggests that our plasma cfDNA
186 methylation classifiers are mainly driven by tumor-specific DNA methylation patterns rather
187 than fluctuations in blood cells or cell composition on the tumor microenvironment.
188 However, these results do not rule out that some plasma cell-free DMRs could originate from
189 shifts in proportions of circulating immune cells[18,19]. To further test our inference, we identified
190 38,352 cfMeDIP-windows lowly methylated across a range of leukocyte types in IHEC WGBS
191 data, of which 27,088 overlapped with TCGA 450k data (Extended Data Fig. 10a). Out of these
192 27,088 regions, we separated those that were identified as hypermethylated through the
193 comparisons of plasma cfDNA of each cancer type to healthy controls. We then checked the

methylation status of these regions in the tumor tissue compared to PBMCs, using TCGA data for each cancer type. For PDAC, we used *in house* methylation data generated for the matched patients (cfDNA and tissue DNA). Indeed, we found these regions to be hypermethylated in tumor tissue (Extended Data Fig. 10b), reinforcing the hypothesis that these plasma cell-free DMRs are a direct measurement of tumor-derived DNA (i.e., ctDNA).

In summary, we developed a robust, sensitive and bisulfite-free methodology for immunoprecipitation-based profiling of methylation patterns in cfDNA. Our approach awaits further validation in completely independent datasets, but our findings underscore the potential utility of cfDNA methylation profiles as a basis for non-invasive, cost-effective, sensitive and accurate early tumor detection for cancer interception, and for multi-cancer classification.

Reference

1 Diaz, L. A., Jr. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* **32**, 579-586, doi:10.1200/JCO.2012.45.2011 (2014).

2 Aravanis, A. M., Lee, M. & Klausner, R. D. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574, doi:10.1016/j.cell.2017.01.030 (2017).

3 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548-554, doi:10.1038/nm.3519 (2014).

4 Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, doi:10.1126/science.aar3247 (2018).

5 Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aan2415 (2017).

6 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

7 Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-1834, doi:10.1073/pnas.1519286113 (2016).

8 Visvanathan, K. *et al.* Monitoring of Serum DNA Methylation as an Early Independent Marker of Response and Survival in Metastatic Breast Cancer: TBCRC 005 Prospective Biomarker Study. *J Clin Oncol*, JCO2015662080 (2016).

9 Potter, N. T. *et al.* Validation of a real-time PCR-based qualitative assay for the detection of methylated SEPT9 DNA in human plasma. *Clin Chem* **60**, 1183-1191, doi:10.1373/clinchem.2013.221044 (2014).
10 Chan, K. C. *et al.* Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A* **110**, 18761-18768, doi:10.1073/pnas.1313995110 (2013).
11 Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-5512, doi:10.1073/pnas.1508736112 (2015).
12 Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res* **29**, E65-65 (2001).
13 Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **7**, 617-636, doi:10.1038/nprot.2012.012 (2012).
14 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547-555, doi:10.1038/nbt.3520 (2016).
15 Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27-36, doi:10.1093/carcin/bgp220 (2010).
16 Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, doi:10.1126/science.aaj2239 (2017).
17 Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-492, doi:10.1016/S0140-6736(05)17866-0 (2005).
18 Pedersen, K. S. *et al.* Leukocyte DNA methylation signature differentiates pancreatic cancer patients from healthy controls. *PLoS One* **6**, e18223, doi:10.1371/journal.pone.0018223 (2011).
19 Teschendorff, A. E. *et al.* An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* **4**, e8274, doi:10.1371/journal.pone.0008274 (2009).

Supplementary information

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

This study was conducted with support from the University of Toronto McLaughlin Centre (MC-2015-02), Canadian Institutes of Health Research (CIHR FDN 148430 and CIHR New Investigator Salary award 201512MSH-360794-228629), Ontario Institute for Cancer Research (OICR) with funds from the province of Ontario, Canada Research Chair (950-231346), and Princess Margaret Cancer Foundation to D.D.C. as well as Canadian Cancer Society (CCSRI 701717) to R.J.H., CCSRI (704716) to R.J.H. and D.D.C. and CCSRI 703827 to M.M.H. Recruitment of healthy individuals was supported by Cancer Care Ontario Chair of Population Health and CCSRI 020214 awarded to R.J.H. Collection of lung cancer samples was supported by the Alan B Brown chair in molecular genomics and the Lusi Wong Lung Cancer Early Detection Program to GL. We would like to acknowledge the Princess Margaret Genomics Centre for carrying out the NGS sequencing and the Bioinformatics and HPC Core, Princess Margaret Cancer Centre for their expertise in generating the NGS data.

Author Contributions

SYS and DDC designed and developed the cfMeDIP-seq protocol. RJH and GF conceived and designed the study related to the pancreatic cancer component. SYS, RS, AC, and DDC conceived and designed the study related to the other cancer types. SYS, SVB, TJP, and DDC designed the experiments. SYS, DC, MHAR, PCZ, ZC, TL, OK, DR, IE, ZC, SC, GMO, JL,

MM and ZZ performed the experiments. TM, YW, and COB performed the mouse experiments. RS, AC, GF, TTW, AG, TJP, MMH and DDC analyzed the data with scientific input from RJH. GF, AB, DC, AS, TM, AA, NL, MHAR, JDM, PLB, NF, GL, MDM, SG, TJP, and RJH collected the clinical data related to the samples, determined the sample selection criteria and matching scheme, and provided the clinical samples. SYS, RS, AC and DDC wrote the paper with feedback from all authors.

Competing Interests

DDC, SYS, AC, SVB, RS, and RJH are listed as inventors/contributors on patents filed related to this work.

 Correspondence and requests for materials should be addressed to Daniel D. De Carvalho.

Figure Legends

Fig. 1: cfDNA methylome as a sensitive approach to detect ctDNA in low input DNA. a, Simulated probability of detecting at least one epimutation as a function of ctDNA concentration (0.001% to 10%) (columns), number of DMRs analyzed (1 to 10,000) (rows), and sequencing depth (10X to 10,000X) (x-axis). **b,** Across a serial dilution series (n=7 dilution points, two technical replicates, each replicate was used per protocol) of HCT116 DNA spiked into MM.1S multiple myeloma cells, near-perfect correlations are observed between observed vs expected methylation signal within DMRs in RPKMs. **c,** Frequency of ctDNA (human) as a percentage of total cfDNA (human + mice) in the plasma from two colorectal cancer, patient-derived xenografts (PDX) before and after cfMeDIP-seq.

Fig. 2: cfMeDIP-seq method can identify thousands of differentially methylated regions (DMRs) in circulating cfDNA obtained from pancreatic adenocarcinoma patients. a, Experimental design. **b,** Volcano plot of DMRs from pancreatic cancer (cases, n=24) versus healthy donors (controls, n=24) using cfMeDIP-seq. Red dots indicate windows significant at BHFDR < 0.1, (Negative Binomial GLM, two-sided p-values). **c,** Heatmap of the 14,716 DMRs identified in the plasma cfDNA from cases and controls (Euclidean distance, Ward clustering). Dendrogram shows separation by case/control status**. d, e,** Overlap between case-vs-control plasma-derived DMRs and RRBS tumour DMRs matched normal tissue (**d**) and PBMCs (**e**). Boxplots represent expected null distribution of overlaps from 1000 permutations (two sided, p-values computed using standard normal distribution). Extremes of boxes and center-lines define upper and lower quartiles and medians. Whiskers indicate 1.5 x interquartile range (IQR). Diamonds represent observed overlap (red if significantly enriched, green if significantly

depleted, and blue if not significant). Horizontal lines indicate thresholds for statistical significance.

Fig. 3: Methylome analysis of plasma cfDNA allows tumor classification. a, cfMeDIP-seq carried out on a discovery cohort consisting of 189 samples from 7 different tumor types: pancreatic cancer (PDAC), AML, bladder cancer (BLCA), breast cancer (BRCA), colorectal cancer (CRC), lung cancer (LUC) and renal cancer (RCC), including early and late stage tumors, and healthy controls (normal). For each cancer type, DMRs between the cancer type and normal controls were identified. Overlap is shown between plasma-derived DMRs for each cancer type and primary tumor DMRs (tumor tissue versus adjacent normal tissue) for the corresponding cancer type using TCGA data. Boxplots represent expected null distribution of overlaps from 1000 permutations (two sided, p-values computed using standard normal distribution). Extremes of boxes and center-lines define upper and lower quartiles and medians. Whiskers indicate 1.5 x interquartile range (IQR). Diamonds represent observed overlap (red if significantly enriched, green if significantly depleted, and blue if not significant). Horizontal lines indicate thresholds for statistical significance. **b,** Evaluation of classification accuracy on the discovery cohort. The discovery cohort (n=189) was partitioned into 100 independent training and test sets in an 80%-20% manner, consisting of 8 classes (cancer types and healthy controls). Training sets were used for DMR-selection and model training, yielding 100 sets of 8 one-class vs-other-classes binomial GLMnet classifiers. Y-axis depicts distributions of AUROC (area under the receiver operator characteristic curve) for each held-out test set for each class. Dots represent performance in individual test sets. Ends of boxes represent upper and lower-quartiles, line within box represents median, and whiskers represent 1.5 x IQR. **c,** ROC (receiver operating characteristic)

curves constructed using averaged class probabilities for independent validation set samples (n=199, 55 LUC, 35 AML, 47 PDAC and 62 healthy controls) from the 100 models for each one-class-vs-other-classes comparison trained using the discovery cohort. **d,** ROC curves for the PDAC and LUC validation set divided into early and late stage, showing that the ability to discriminate PDAC or LUC samples is similar when considering early and late stage samples of that class separately.

Fig. 4: Plasma-derived DMRs are informative of cancer type. a, The plasma-derived DMRs identified as informative of cancer type in the discovery cohort of 189 plasma samples was used to generate 3D and 2D tSNE (t-distributed stochastic neighbor embedding) plots for the entire cohort of plasma samples (n=388). **b, c,** The DNA methylation Beta value for probes within the plasma-derived DMRs were used to generate 3D and 2D tSNE plots for (**b**) TCGA cancer tissue (n=4,032) and (**c**) COSMIC cancer cell line (n=400 cell lines).

Methods

Bioinformatic simulation of tumor-specific features and probability of detection by sequencing depth

We created 145,000 simulated genomes with 1, 10, 100, 1,000 and 10,000 independent loci with 0.001 – 10% cancer-specific DMRs in 10-fold increments. 14,500 diploid genomes (expected copy number in 100 ng cfDNA) were then sampled from these mixtures and further sampled 10 – 10,000x in 10-fold increments at each locus. The process was repeated 100 times for each combination of parameters. Probability curves were plotted for successful detection of >1 and >5 DMRs (Fig. 1a and Extended Data Fig.1a).

cfMeDIP-seq

A schematic representation of the cfMeDIP-seq protocol is shown in Extended Data Fig. 1b. Prior to cfMeDIP, the samples were subjected to library preparation using Kapa HyperPrep Kit (Kapa Biosystems), following manufacturer's protocol with minor modifications. Briefly, after end-repair and A-tailing, samples were ligated to 0.181 μM of NEBNext adapter (NEBNext Multiplex Oligos for Illumina kit, New England BioLabs, Canada) by incubating at 20˚C for 20 min and purified with AMPure XP beads (Beckman Coulter). The eluted library was digested using the USER enzyme (New England BioLabs, Canada) followed by purification with Qiagen MinElute PCR Purification Kit (MinElute columns) prior to MeDIP.

The prepared libraries were combined with the filler λ DNA (to ensure the total amount of DNA [cfDNA + filler] was 100 ng) and subjected to MeDIP with Diagenode MagMeDIP kit (Cat# C02010021) using the protocol from Taiwo et al. 2012[13] with some modifications. The filler DNA consists of a mixture of unmethylated and *in vitro* methylated λ amplicons of different CpG densities (Supplementary Table 6), similar in size to adapter-ligated cfDNA libraries. Its addition ensures a constant ratio of antibody to input DNA and helps maintain similar immunoprecipitation efficiency across samples regardless of available cfDNA, while minimizing non-specific binding by the antibody and DNA loss due to binding to plasticware. For MeDIP, the prepared library/filler DNA mixture was combined with 0.3 ng of control methylated and 0.3 ng of the control unmethylated *A. thaliana* DNA provided in the kit, and the buffers. The mixture was heated to 95˚C for 10 min, then immediately placed into an ice water bath for 10 min. Each sample was partitioned into two 0.2 mL PCR tubes: one for the 10% input control (7.9 μl) and the other one for the sample to be subjected to immunoprecipitation (79 μl). The included 5-mC

monoclonal antibody 33D3 (Cat#C15200081) from the MagMeDIP kit was diluted 1:15 prior to generating the diluted antibody mix and added to the sample. Washed magnetic beads (following manufacturer instructions) were also added prior to incubation at 4˚C for 17 hours. The samples were purified using the Diagenode iPure Kit v2 (Cat# C03010015) and eluted in 50 µl of Buffer C. The success of the reaction (QC1) was validated through qPCR to detect recovery of the spiked-in methylated and unmethylated *A. thaliana* DNA. The % recovery of unmethylated spiked-in DNA should be <1% (relative to input control, adjusted for input control being 10% of overall sample) and the % specificity of the reaction should be >99% (as calculated by (1 – [recovery of spiked-in unmethylated control DNA over recovery of spiked-in methylated control DNA])*100), prior to proceeding to the next step. The optimal number of cycles to amplify each library was determined through qPCR, after which the samples were amplified using Kapa HiFi Hotstart Mastermix and NEBNext multiplex oligos, added to a final concentration of 0.3 µM. The final libraries were amplified as follows: activation at 95˚C for 3 min, followed by predetermined cycles of 98˚C for 20 s, 65˚C for 15 s and 72˚C for 30 s and a final extension of 72˚C for 1 min. The amplified libraries were purified using MinElute columns, then gel size selected with 3% Nusieve GTG agarose gel to remove any adapter dimers. All the final libraries were submitted for BioAnalyzer analysis prior to sequencing at the Princess Margaret Genomics Centre (PMGC) on an Illumina HiSeq 2500, SBS V4 chemistry, single read 50 bp, multiplexed as 7 samples/lane. After sequencing, the sequenced reads were aligned to λ and hg19 using Bowtie[20] with the default settings. Based on virtually no alignment to λ genome, the filler DNA does not interfere with the generation of sequencing data (Supplementary Table 7 and 8).

The generated SAM files from hg19 alignment were converted to BAM format, ensuring removal of duplicate reads, sorting and indexing of reads using SAMtools[21] prior to subsequent

analysis with R package MEDIPS[22]. CpG Enrichment Score, as a quality control measure for the immunoprecipitation reaction, was calculated as part of the MEDIPS package.

Validation of cfMeDIP-seq against MeDIP-seq

DNA from human colorectal cancer cell (CRC) line HCT116 (American Type Culture Collection (ATCC), mycoplasma free) was extracted using PureLink Genomic DNA Mini Kit (Thermofisher Scientific). HCT116 was chosen because of the availability of public DNA methylation data. Genomic DNA was sheared to mimic cfDNA using a Covaris sonicator, and larger size fragments excluded using AMPure XP beads (Beckman Coulter) to mimic the fragment size of cell-free DNA. cfMeDIP-seq was carried out on 1, 5, 10 and 100 ng of sheared DNA as input, with 100 ng representing the gold-standard MeDIP-seq protocol, with 2 biological replicates per input. The fold enrichment of a methylated human DNA region (*TSH2B*) over unmethylated human DNA region (*GAPDH* promoter), using primers provided in MagMeDIP kit, was determined prior to sequencing libraries to saturation (Extended Data Fig. 2a-c, Supplementary Table 7).

Dilution series of sheared cell line DNA

As with the CRC DNA, the same extraction and shearing protocol was used with multiple myeloma (MM) cell line MM.1S. A dilution series of CRC into MM DNA was carried out following Extended Data Fig. 3a scheme. This dilution series was used for cfMeDIP-seq (Supplementary Table 9) and for ultra-deep targeted sequencing for CRC point mutation detection, using a starting input of 60 ng of DNA. For the mutation detection**,** DNA libraries were prepared using Kapa HyperPrep Kit (Kapa Biosystems) and Illumina compatible molecular

barcoded adapters with 2-bp in-line barcodes (unique molecular identifiers (UMIs)) to ensure optimal analytical sensitivity for mutation detection[14]. A customized biotinylated DNA capture probe panel (xGen Lockdown Custom Probes Mini Pool, Integrated DNA Technologies) targeting exons from 5 genes (13kb) was used[23]. In brief, the barcoded libraries were pooled, and hybrid capture was performed according to the manufacturer's instructions (IDT xGEN Lockdown protocol version 4). The amplified post-capture libraries were sequenced to >100,000X read coverage using Illumina HiSeq 2500 instrument, SBS V4 chemistry, paired-end 125 bp, as 4 samples/lane. Average target coverage of unprocessed reads was 186,312X (range: 154,419X – 216,434X) (Supplementary Table 9).

After sequencing, reads were de-multiplexed using sample specific indices into separate paired-end FASTQ files. A two base pair molecular barcode and a one base pair invariant spacer sequence were removed from each read. A thymine base was encoded in the third position for adapter ligation and a spacer filter was enforced to remove reads incompliant with this design. The extracted barcodes from paired-end reads were grouped and written into the header of each sequence for downstream *in silico* molecular identification[24]. FASTQ files were mapped to the human reference genome hg19 using BWA[25], processed using the Genome Analysis ToolKit (GATK) IndelRealigner[26], and sorted and indexed using SAMtools[21].

Barcodes were used in combination with endogenous sequence features (genome coordinates, mapping alignments, read orientation, and read number in pair) to confer sequences from individual molecules. Consensus sequences were formed from two or more reads supporting the same molecule with 70% agreement amongst bases above Phred quality scores (Q)[27] of 30. Reads derived from the same strand of a unique fragment were collapsed to form single strand

446 consensus sequences (SSCS), suppressing polymerase and sequencer errors. These condensed
447 reads were subsequently combined with their complementary strand into duplex consensus
448 sequences (DCS). This enables an additional layer of error suppression as double strand
449 consolidated sequences can correct for asymmetric damage accrued during the first cycle of PCR
450 or induced by oxidation[28].

451 We selected variants based on annotated SNPs from the Cancer Cell Line Encyclopedia
452 (CCLE)[29] overlapping our target panel. SNVs were called with MuTect[30] using the following
453 parameters: --enable_extended_output --tumor_f_pretest 0.000001f --downsampling_type
454 NONE --force_output --force_alleles --gap_events_threshold 1000 --fraction_contamination
455 0.00f --coverage_file[30]. We force called every base for each variant to assess limit of detection
456 and background noise at each stage of barcode-mediated error correction. Analysis of the UMI-
457 processed error-suppressed reads revealed unique molecule (i.e., SSCS) and DCS average target
458 coverage of 6,276X (4,284X – 8,068X) and 1,043X (654X – 1,602X), respectively
459 (Supplementary Table 9).

460 **Specimen processing – Patient Derived Xenograft (PDX) cfDNA**

461 All mouse work was carried out in compliance with animal use protocol and ethical regulations
462 approved by the Animal Care Committee at University Health Network. Human colorectal tumor
463 tissue obtained with patient consent and UHN Research Ethics Board approval from the UHN
464 Biobank, was digested to single cells using collagenase A. Single cells were subcutaneously
465 injected into 4-6 week old NOD/SCID male mouse. Mice were euthanized by CO_2 inhalation
466 prior to blood collection by cardiac puncture and stored in EDTA tubes. From the collected
467 blood samples, plasma was isolated and stored at -80 ˚C. cfDNA was extracted from 0.3-0.7 ml

of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen). Two biological samples with 10 ng of starting cfDNA were subjected to cfMeDIP-seq protocol as previously mentioned, sequenced and analyzed (Supplementary Table 10).

Donor recruitment and sample acquisition

All patients have provided written informed consents, and all samples have been obtained upon approval of the institutional ethics committees and Research Ethics Boards from University Health Network (UHN) and Mount Sinai Hospital (MSH), in compliance with all relevant ethical regulations. Pancreatic adenocarcinoma cases (PDAC) were obtained from the Ontario Pancreatic Cancer Study and the UHN Biobank. Colorectal and breast cancer plasma samples were obtained from the UHN Biobank. Lung cancer plasma samples were obtained from the UHN Thoracic Biobank. AML samples were obtained from the UHN Leukemia Biobank. Bladder and renal cancer plasma samples were obtained from the UHN Genitourinary (GU) Biobank from consenting urologic oncology patients, procured prior to nephrectomy and cystectomy respectively. Lastly, healthy controls were recruited through the Family Medicine Centre at MSH in Toronto, Canada.

Specimen processing and methylation analysis of purified tumor and normal cells from PDAC samples

For primary PDAC samples, specimens were processed immediately following resection and representative sections were used to confirm the diagnosis. Laser capture microdissection (LCM) of freshly liquid nitrogen-frozen tissue samples was performed on a Leica LMD 7000 instrument. LCM was performed on the same day when sections were cut to minimize nucleic acid degradation. Qiagen Cell Lysis Buffer was used to extract genomic DNA.

Quantified 10 ng of genomic DNA for each sample was analyzed using RRBS following the protocol from Gu et al., 2011[31] with minor modifications. DNA libraries ligated to Illumina TruSeq methylated adapters were subjected to bisulfite conversion using the Zymo EZ DNA methylation kit following manufacturer's protocol, followed by gel size selection for fragments of 160 bp-300 bp in size. After determining the optimal number of cycles to amplify each purified library, samples were amplified using Kapa HiFi Uracil+ Mastermix (Kapa Biosystems) and purified with AMPure beads (Beckman Coulter). The final libraries were submitted for BioAnalyzer analysis prior to sequencing at PMGC on an Illumina HiSeq 2000, using sequencing by synthesis (SBS) V3 chemistry, single read 50bp and multiplexed as 4 samples/lane. After sequencing, the raw data for each sample was trimmed with Trim Galore! using the RRBS settings prior to aligning to hg19 using Bismark[32] with Bowtie2[33] (Supplementary Table 11). The generated SAM files were then converted to BAM format, sorted and indexed using SAMtools.

Specimen Processing – patient cell-free DNA (cfDNA)

Plasma samples collected using EDTA and ACD (acid citrate dextrose) tubes, were obtained from the UHN BioBanks and MSH and were kept frozen until use. cfDNA was extracted from 0.5-3.5 mL of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen) and quantified through Qubit prior to use. Sex, age and pathology stage are available in Supplementary Table 12, while extracted DNA quantities are available in Extended Data Fig. 8a.

Calculation and visualization of differentially methylated regions from cfDNA of pancreatic cancer patients and healthy donors

DMRs between cfDNA samples from 24 pancreatic cancer (PDAC) patients and 24 healthy donors (controls) were calculated using MEDIPS and DESeq2 R packages[22,34]. For each sample, we computed counts per 300 bp non-overlapping windows, filtered out windows with less than 10 counts across all samples and fit a negative binomial model to call DMRs at FDR < 0.1 (Wald test). Z-scores of DMR RPKM values with Euclidean Distance and Ward clustering were used for visualisation.

Enrichment analyses for plasma-derived DMRs in tumor-specific methylation signals in PDAC

Five normal PBMC samples profiled by RRBS were downloaded from GEO (GSE89473) for comparison with the 24 pancreatic cancer tissue RRBS samples. R package MethylKit was used to parse files and autosomal CpGs detected in at least 18 out of the 24 PDACs and 4 out of the 5 PBMCs were retained for further analysis. We obtained DMCs at FDR <0.01, delta-Beta >0.25. A null distribution was then generated, from 1000 resamples, preserving the relationship between the number of CpGs in windows that were seen in the original intersections between RRBS features and cfMeDIP DMRs. Then we computed the frequency of overlap between DMRs hypermethylated in both, hypermethylated in one but not the other, hypomethylated in one but not the other, and finally, hypomethylated in both comparisons. The distributions were then standardized based on z-scores and used to compute Bonferroni-adjusted p-values to determine enrichment. The same procedure was employed for subsequent enrichment tests in the manuscript.

Enrichment analyses for cfMeDIP DMRs in TCGA 450K DMCs relative to normal tissues and PBMCs.

189 cfDNA samples were obtained across 7 cancer types (AML, bladder (BLCA), breast (BRCA), colorectal (CRC), lung (LUC), pancreatic (PDAC) and renal cancer (RCC)) and healthy donors (normal) (Supplementary Table 12). After processing of cfMeDIP-seq data from these samples, DMRs were calculated using DESeq2 between each cancer type and healthy donors as described above. DMCs were also calculated between TCGA 450K methylation array samples from each corresponding cancer type (n = 3979) (obtained from SAGE synapse) and PBMCs (n = 53, obtained from GEO) samples using *limma* (FDR < 0.01, absolute delta-Beta 0.25). Statistical tests for enrichment were performed as described above for PDAC RRBS samples. The same procedure was carried out for DMCs calculated between TCGA 450K methylation array samples from a cancer type and normal samples from the same tissue, for BLCA, BRCA, CRC, LUC and RCC.

Examination of transcription factors associated with differentially methylated motifs in cfMeDIP-seq DMRs.

RNA-seq data obtained as median RPKMs from the GTEx consortium across 53 human tissues – as described in the supplemental R Markdowns in Zenodo (ID 10.5281/zenodo.1205756) (Supplementary Table 13), and median expression per tissue was visualized in heatmaps. To look for enrichment of TF expression and DMR-associated TF motifs, we selected 1000 random sets of TFs. As part of the analysis, we considered the known sensitivity to the methylation status of each TF[16], yielding 42 TFs that are enriched in healthy donors, and 52 TFs that are enriched in pancreatic adenocarcinoma cases.

We computed ssGSEA (single-sample gene set enrichment analysis) scores for the expression of these TFs per sample, for pancreatic cancer (TCGA), blood (GTEx) and normal pancreas (GTEx)

and compared distributions to those from random sets of TFs using Wilcoxon's Rank Sum Test. Violin plots were made as described in the supplemental R Markdown 10.5281/zenodo.1205735 (Supplementary Table 13).

Machine learning analyses for evaluation of classification accuracy.

Model training and evaluation on the discovery cohort.

In order to evaluate the performance of cfMeDIP data in tumor classification without high computational cost, we reduced the initial set of possible candidate features to windows encompassing CpG Islands, shores, shelves and FANTOM5 enhancers ("regulatory features"), yielding a matrix of 189 samples and 505,027 features.

We then used the *caret* R package[35] to partition the discovery cohort data into 100 class-balanced independent training and test sets in an 80%-20% manner. Then, we selected the top 300 DMRs by moderated t-statistic (150 hypermethylated, 150 hypomethylated) on the training data partition using *limma-trend*[36] for each class versus other classes. A binomial GLMnet was then trained using these DMRs (up to 300 DMRs x 7 other classes = 2100 features) using of 3 iterations of 10-Fold Cross-Validation (CV) to optimize values of the mixing parameter (alpha, values = 0, 0.2, 0.5, 0.8 and 1) and the penalty (lambda, values = 0 – 0.05 in increments of 0.01) using Cohen's Kappa as the performance metric. For each training set, this yielded a collection of 8 one-class vs-other-classes binomial classifiers.

We then estimated classification performance on the held-out test set using the AUROC (area under the receiver operating characteristic curve). These estimates represent unbiased measures of classification, as the held-out test set samples were not used for either DMR pre-selection or

GLMnet training and tuning. The 100 independent training and test sets also permitted for minimization of optimistic estimates due to training-set bias.

Model evaluation on the validation cohort.

For each validation cohort cfMeDIP sample, we estimated class probabilities for the AML, PDAC, LUC and normal one-vs-all binomial classifiers trained on the 100 different training sets within the discovery cohort. The probabilities from the 100 models were averaged to produce a single score that was then used for AUROC estimation. We also evaluated if disease stage (applicable to only LUC and PDAC) affected performance by estimating AUROC when either early (Stages I and II) or late stage samples (Stages III and IV) of a particular class were left out for the one-vs-all classifiers trained to identify the class in question.

Validation in cell lines

450K profiles for 1,028 cell lines previously characterised in Iorio et al (2016)[37] were obtained as IDAT files. The data were then uniformly processed using the ssNoob method in the *minfi* package[38]. We reduced this dataset to tissue types for which cfMeDIP data were available (n=400).

Data availability statement

We have deposited R markdowns (either knit or raw) and scripts used to generate the findings in this study on Zenodo (DOIs in Supplementary Table 13). All the cell line datasets generated during and/or analysed during the current study are available in the GEO repository under accession code GSE79838. The cfMeDIP-seq NGS data for patient samples that support the findings of this study are available upon request from the corresponding author (D.D.C) to comply with institutional ethics regulation. Source data for Fig. 1b and Extended Data Fig. 3e

598 are provided in Supplementary Table 9, and for Fig. 1c are provided in Supplementary Table 10.

599 Additional source data can be found on Zenodo (Supplementary Table 13)

600

Methods Reference

20 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

21 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

22 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, doi:10.1093/bioinformatics/btt650 (2014).

23 Kis, O. *et al.* Circulating tumour DNA sequence analysis as an alternative to multiple myeloma bone marrow aspirates. *Nat Commun* **8**, 15086, doi:10.1038/ncomms15086 (2017).

24 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).

25 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

26 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

27 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

28 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513, doi:10.1073/pnas.1208715109 (2012).

29 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

30 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

31 Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**, 468-481, doi:10.1038/nprot.2010.190 (2011).

32 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).

33 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

34 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

35 Kuhn, M. Building Predictive Models inRUsing thecaretPackage. *Journal of Statistical Software* **28**, doi:10.18637/jss.v028.i05 (2008).

36 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).
37 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754, doi:10.1016/j.cell.2016.06.017 (2016).
38 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).

Extended Data Figure Legends

Extended Data Fig. 1 Simulation of the probability of detecting ctDNA as a function of the number of DMRs, sequencing depth and percentage of ctDNA in plasma cfDNA and proposed method to enrich ctDNA. a, Bioinformatic simulation of scenarios with different proportions of ctDNA present in the sample (0.001% to 10%, column facets), and a range of tumor specific Differentially Methylated Regions (DMRs), from 1, 10, 100, 1,000, or 10,000, determined through the comparison of ctDNA to normal cfDNA (row facets), with reads sampled at varying sequencing depths, at each locus (10X, 100X, 1,000X, and 10,000X) (x-axis). The probability of detecting at least 5 epimutations/DMR increases as the number of available features increases, even at shallow coverage per locus (left y-axis). Each panel depicts probability of detection against coverage per candidate DMR for one simulation scenario. **b,** Schematic representation of the cfMeDIP-seq protocol.

Extended Data Fig. 2 Sequencing saturation analysis and quality controls of MeDIP-seq and cfMeDIP-seq carried out on varying starting inputs of HCT116 cell line DNA sheared to mimic cfDNA (HCT116 cfDNA mimic DNA). a, The figure shows the results of the saturation analysis from the Bioconductor package MEDIPS analyzing cfMeDIP-seq data from each replicate, for each starting input amount and including an input control. **b,** The protocol was tested in two biological replicates of four starting DNA inputs (100, 10, 5, and 1 ng) of HCT116 cfDNA mimic DNA. Specificity of the reaction was calculated using methylated and unmethylated spiked-in *A. thaliana* DNA. Fold enrichment ratio was calculated using genomic regions of the fragmented HCT116 DNA (human methylated *TSH2B0* and unmethylated *GAPDH*). The horizontal dotted line indicates a fold-enrichment ratio threshold of 25, dots

represent biological replicate, with line representing mean **c,** CpG Enrichment Scores of the sequenced samples (two biological replicates each of four starting DNA inputs (100, 10, 5, and 1 ng), and one input control) show a robust enrichment of CpGs within the genomic regions from the immunoprecipitated samples compared to the input control. The CpG Enrichment Score was obtained by dividing the relative frequency of CpGs of the regions by the relative frequency of CpGs of the human genome. The horizontal dotted line indicates a CpG Enrichment Score of 1, dots represent biological replicates, with line representing mean. **d,** Genome-wide Pearson correlations of normalized read counts per 300 bp window between cfMeDIP-seq signal for 1 to 100 ng of input HCT116 DNA sheared to mimic cfDNA (2 biological replicates per concentration). **e,** Genome Browser snapshot of HCT116 cfMeDIP-seq signal across a window (chr8:145,095,942-145,116,942) selected out of 4 examined loci, at different starting DNA inputs (1 to 100 ng, in biological replicates), compared with RRBS (ENCODE ENCSR000DFS) and WGBS (GEO GSM1465024) data (aligned to hg19). For cfMeDIP-seq, y-axis indicates RPKMs, for RRBS, yellow and blue blocks represent hypermethylated and hypomethylated CpGs, respectively. In WGBS track, peak heights indicate methylation level.

Extended Data Fig. 3 Sequencing saturation analysis and quality controls of cfMeDIP-seq from serial dilution. a, Schematic representation of the CRC DNA (HCT116) dilution series into MM DNA (MM1.S). For both CRC and MM DNA, the genomic DNA was sheared to mimic cfDNA fragmentation. The entire dilution series was used to carry out cfMeDIP-seq (n=1) and ultra-deep sequencing for mutation detection (n=1). **b**, Specificity of reaction for each dilution in the series (n=1) was calculated using methylated and unmethylated spiked-in *A. thaliana* DNA. **c**, CpG enrichment representing ratio of relative frequency of CpGs in regions to relative frequency of CpGs in the human genome for each dilution in the series (n=1),

determined via cfMeDIP-seq. Horizontal dashed line represents CpG enrichment of 1. **d**, Saturation analysis of cfMeDIP-seq sequenced reads from each dilution point in the series (n=1). **e,** Across a serial dilution series (n=7 dilution points, two technical replicates, each replicate was used per protocol) of HCT116 DNA spiked into MM.1S multiple myeloma cells, near-perfect correlations are observed between observed vs expected numbers of DMRs. **f-g,** Ultra-deep sequencing for mutation detection of three CRC-specific point mutations within *BRAF* (p.P301P), *KRAS* (p.G13D) and *PIK3CA* (p.H1047R) in the same dilution series (of CRC into MM DNA) (n=1). Unique molecular identifiers (UMIs) were incorporated into the sequencing adapters and used to created single strand consensus sequence (SSCS) (**f)** duplex consensus sequences (DCS) (**g)** for the detection of allele frequency for each mutation at each locus. For each mutation, the reference allele is found at the top. Dashed red line: limit of detection.

Extended Data Fig. 4 Quality controls for cfMeDIP-seq from circulating cfDNA from pancreatic adenocarcinoma patients (cases) and healthy donors (controls). a-b, Specificity of reaction calculated using methylated and unmethylated spiked-in *A. thaliana* DNA for **(a)** each case sample and **(b)** each control sample. Fold enrichment ratio was not calculated due to the very limited amount of DNA available after final libraries were generated. **c-d,** CpG enrichment of the sequenced cases (**c**) and controls (**d**), Horizontal dashed line represents CpG enrichment of 1. **e,** Principal component (PC) analysis of cfDNA methylation from 24 plasma cfDNA samples from healthy donors and 24 plasma cfDNA samples from pancreatic adenocarcinoma patients, using the 1 million most variable windows by Median Absolute Deviation (MAD) (300bp) genome-wide. Left: PC2 against PC1; right: PC3 against and PC1. **f**, Percentage of variance explained by each PC.

Extended Data Fig. 5 Methylome analysis of plasma cfDNA distinguishes early stage pancreatic adenocarcinoma patients (PDAC) from healthy controls. a, Difference in plasma cfDNA methylation against difference in tumor DNA methylation for each overlapping window (n=547,887). Plasma cfDNA methylation difference is $\log_{10}$ fold change from pancreatic adenocarcinoma patients to healthy measured by cfMeDIP-seq. Tumor DNA methylation difference is delta-Beta from primary pancreatic adenocarcinoma tumor to normal tissue, measured by RRBS. Blue line: trend line, correlation determined by Pearson's correlation. **b**, Scatter-plot showing the DNA methylation difference for each overlapping window. X-axis shows the DNA methylation difference for the primary pancreatic adenocarcinoma tumor versus normal PBMCs from the RRBS data. Y-axis shows the DNA methylation difference for the plasma cfDNA methylation from pancreatic adenocarcinoma patients versus healthy donors from the cfMeDIP-seq data. Correlation determined by Pearson's correlation. **c,** Genome browser snapshot of RRBS and cfMeDIP-seq signal across a representative chromosomal region selected from four candidate regions (chr8:145,095,942-145,116,942) using reference genome hg19. RRBS tracks show the methylation signal for the LCM tissues from pancreatic adenocarcinoma tumor cases and the matching normal tissue, from the same patient, shown in the same order. Each colored block represents differentially methylated CpGs (DMCs), with yellow representing hypermethylated and blue representing hypomethylated. cfMeDIP-seq tracks show the methylation signal (RPKMs) detected in the cfDNA, with cases representing plasma from the same pancreatic adenocarcinoma cases and controls corresponding to plasma from age and sex matched healthy controls. For the cfMeDIP-seq tracks, green and blue peaks indicate the methylation signal (RPKMs) detected in the cfDNA.

Extended Data Fig. 6 Circulating cfDNA methylation profiles can identify transcription factor (TF) footprints and infer active transcriptional networks in the tissue of origin. a, Expression profile of all TFs (n=42) that were characterized as binding in healthy controls across 53 human tissues from the Genotype-Tissue Expression (GTEx) project. Several TFs preferentially expressed in the hematopoietic system were identified (PU.1, NFE2, and GATA1). **b,** Expression profiles (ssGSEA scores) of all TFs with hypomethylated motifs in controls (n=42) are overexpressed versus those of 1,000 random sets of 42 TFs across GTEx whole blood data ($p<2.2e-16$, Wilcoxon's Rank Sum Test, two-sided). **c,** Expression profile of all TFs (n=52) characterized as binding in pancreatic adenocarcinoma patients. Several pancreas-specific or pancreatic cancer-associated TFs were identified. Moreover, hallmark TFs that drive molecular subtypes of pancreatic cancer were also identified. **d,** Expression profile (ssGSEA scores) of all TFs with hypomethylated motifs in cases (n=52) are overexpressed versus those of 1,000 random sets of 52 TFs in normal pancreas (GTEx data) (Wilcoxon Rank Sum Test, two-sided test, p-value<2.2e-16). **e,** Expression profile of all TFs with hypomethylated motifs in cases (n=52) are overexpressed versus those of 1,000 random sets of 52 TFs in pancreatic adenocarcinoma tissue (TCGA data) (Wilcoxon Rank Sum Test, two-sided test, p-value<2.2e-16). For violin plots (**b, d, e)** the ends of the boxes and middle line represent the lower and upper quartiles and median, respectively. Whiskers represent 1.5 times the interquartile range (IQR). Outliers are excluded. Rotated kernel densities are also displayed.

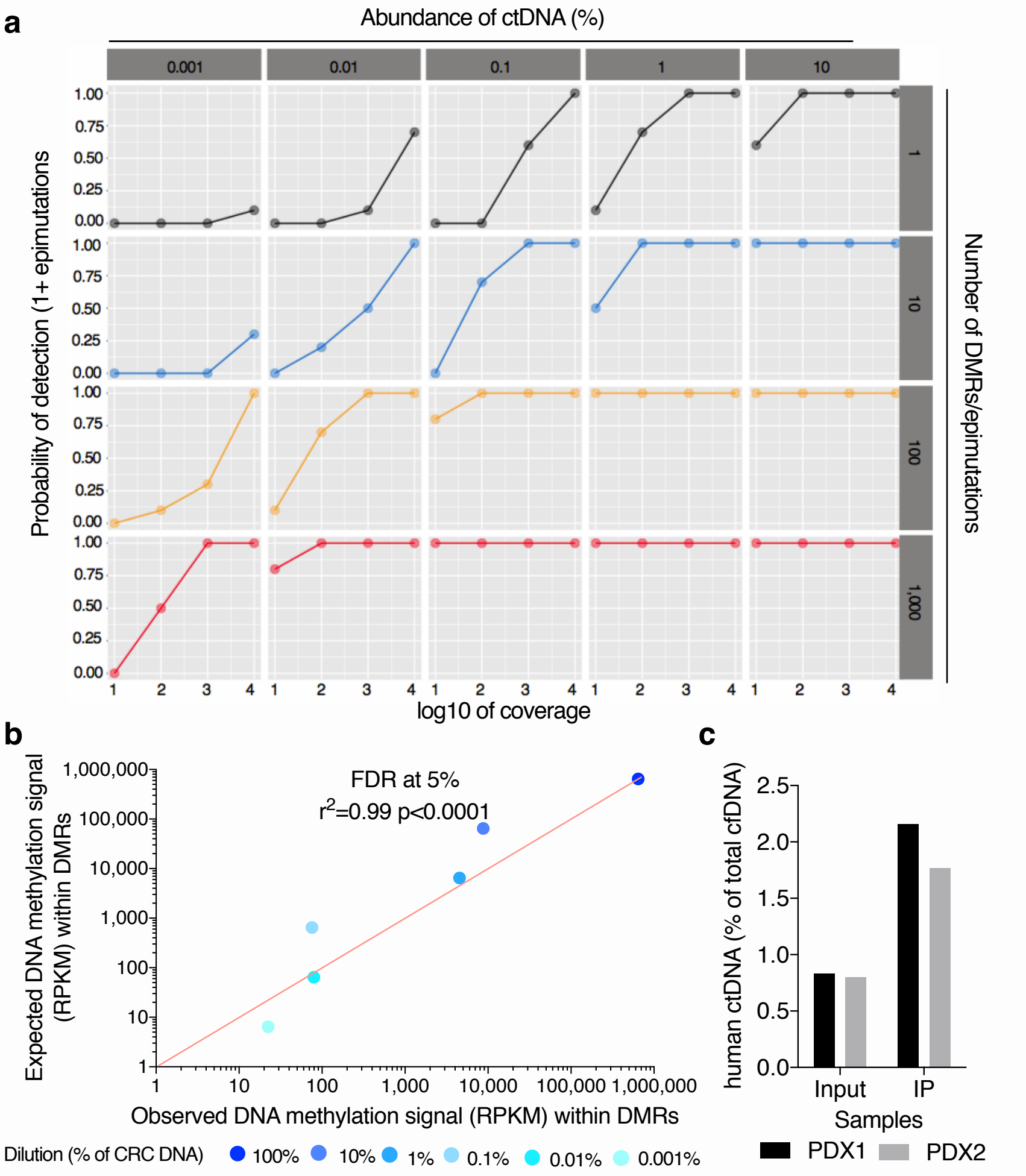
Extended Data Fig. 7 Quality controls for cfMeDIP-seq from circulating cfDNA from multiple cancer types. a, c, e, g, i, k, Specificity of reaction and **b, d, f, h, j,** CpG enrichment score for each sample per cancer type. Horizontal dashed line represents CpG enrichment of 1.

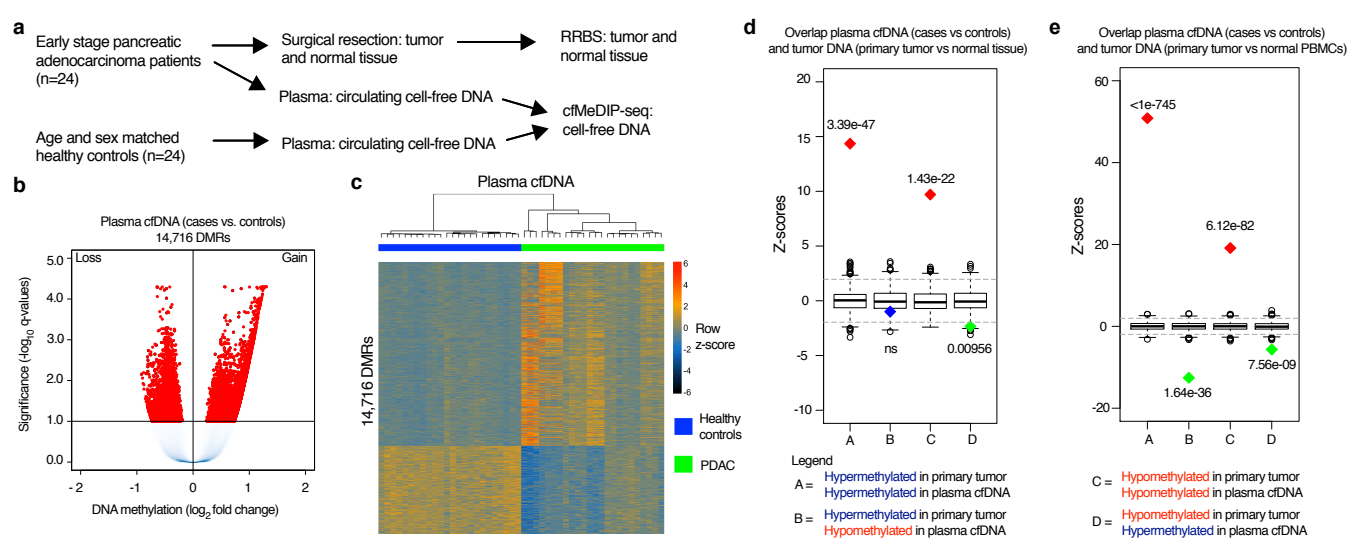
Extended Data Fig. 8 Comparison of plasma cfDNA DMRs with tumor DMCs. a, Yield of cfDNA extracted per mL of plasma from healthy donors (n=24), bladder cancer (n=20), renal cancer (n=20), lung cancer (n=25), breast cancer (n=25), pancreatic cancer (n=24), colorectal cancer (23) and AML (n=28). Horizontal bars represent the mean, with dots representing individual samples. **b-h,** Scatterplot showing the DNA methylation difference for all overlapping windows in pancreatic cancer (PDAC) (n=245,980 windows) (**b**), acute myelogenous leukemia (AML) (n=206,735 windows) (**c**), bladder cancer (BLCA) (n=193,943 windows) (**d**), breast cancer (BRCA) (n= 204,623 windows) (**e**), colorectal cancer (CRC) (n= 210,645 windows) (**f**), lung cancer (LUC) (n= 193,043 windows) (**g**), and renal cancer (RCC) (n= 198,390 windows) (**h**). X-axis shows the DNA methylation difference for the primary tumor (TCGA data) versus normal PBMCs. Y-axis shows the DNA methylation difference for the plasma cfDNA methylation for each cancer type versus healthy controls from the cfMeDIP-seq data. Blue line shows the trend line, correlation determined through Pearson's correlation.

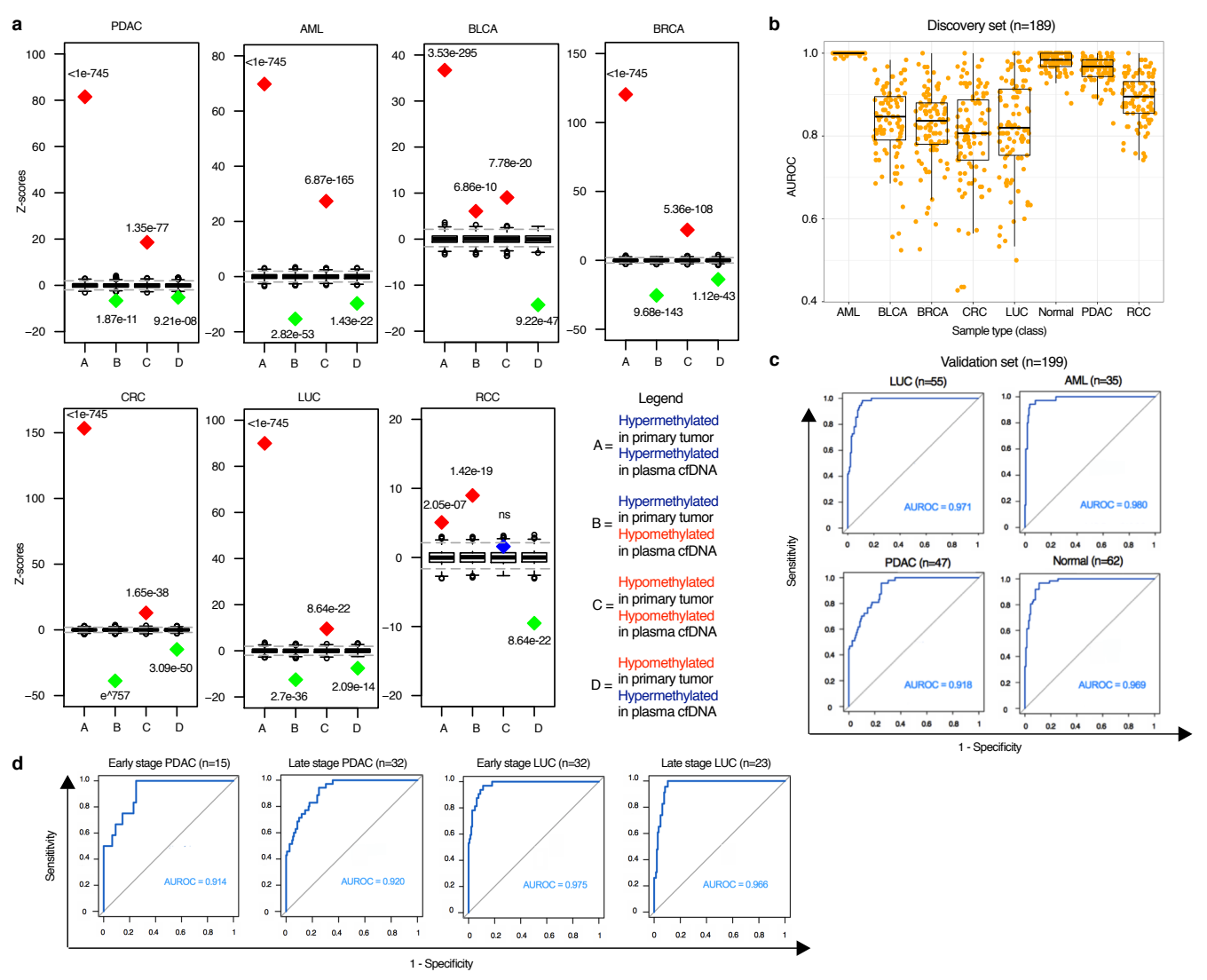
Extended Data Fig. 9 Circulating plasma cfDNA methylation samples used to distinguish between multiple cancer types and healthy donors. a-b, Pathology stage (AJCC/UICC 7^{th} Edition) breakdown by tumor type for samples in the (**a**) training set and in the (**b**) validation set. Pancreatic cancer: PDAC, breast cancer: BRCA, lung cancer: LUC, colorectal cancer: CRC, bladder cancer: BLCA, renal cancer: RCC, Non-small cell lung carcinoma: LUC (NSCLC) and small cell lung cancer: LUC (SCLC).

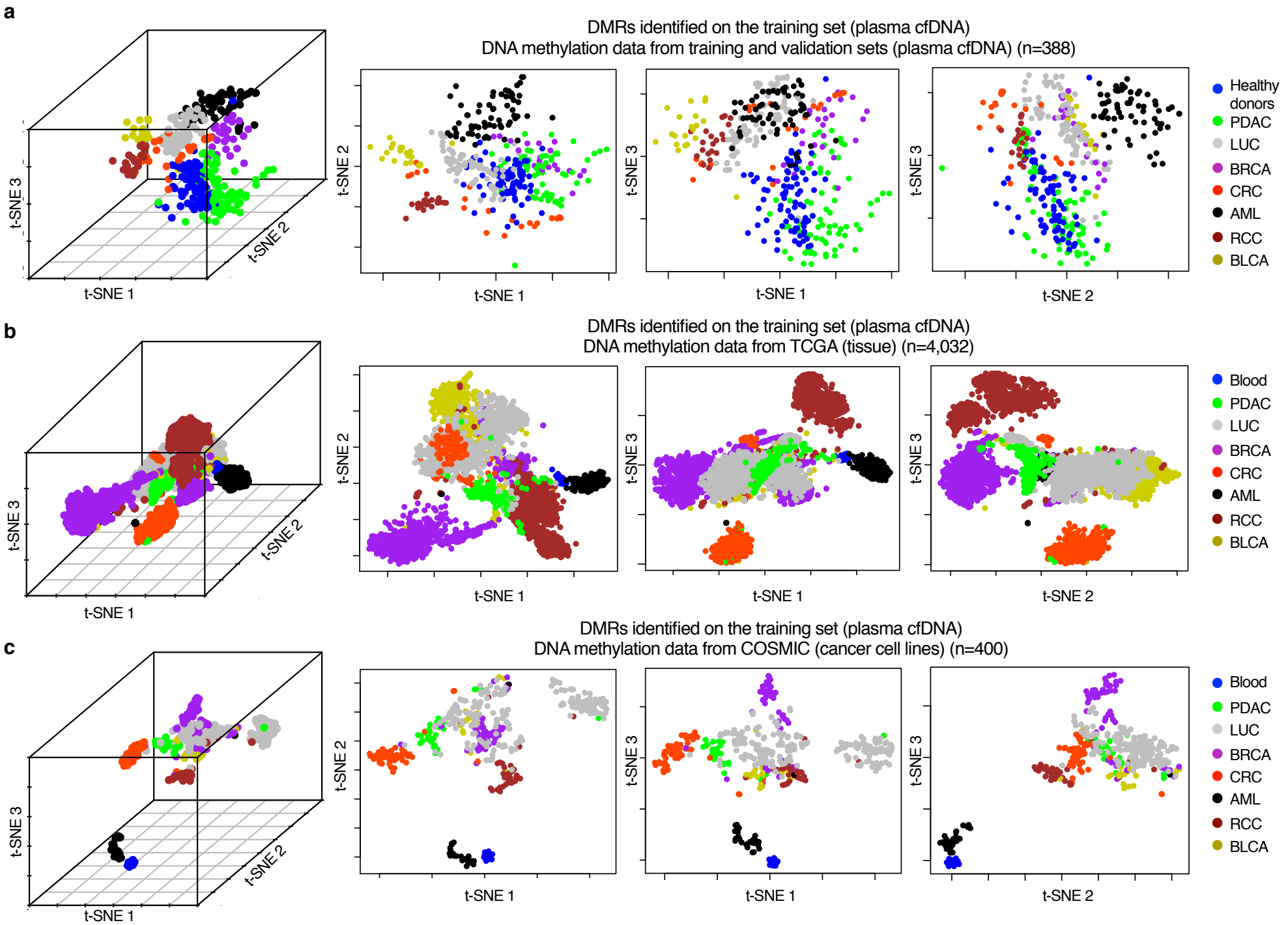
Extended Data Fig. 10 Characterization of hypermethylated regions from cfDNA that are not methylated in leukocytes. **a,** Violin plots for the DNA methylation (Beta value) of 38,352 regions in normal blood cells selected based on low DNA methylation levels using IHEC whole

genome bisulfite sequencing data. For violin plots, the ends of the boxes and the middle line represent the lower and upper quartiles, and medians, respectively. Whiskers represent 1.5 times the interquartile range (IQR). Outliers are excluded. Rotated kernel densities are also displayed. **b,** Volcano plot representing the regions with low DNA methylation levels in normal blood cells that overlap with hypermethylated regions in the plasma cfDNA for PDAC (n=3,146 CpG sites) relative to normal tissue, and RCC (n=2,767 CpG sites), BLCA (n= 3,286 CpG sites), BRCA (n= 6,836 CpG sites), CRC (n= 8,360 CpG sites) and LUC (n= 5,239 CpG sites) relative to PBMCs. X-axis represents the delta-Beta value in methylation in tumor data from TCGA for cancers other than PDAC and RRBS for PDAC. Y-axis represents $-\log_{10}$ q-values (Benjamini Hochberg False Discovery Rate, BHFDR).


a
Abundance of ctDNA (%)
0.001
0.01
0.1
1
10
Probability of detection (1+ epimutations
1.00
0.75
0.50
0.25
0.00
1
10
100
1,000
Number of DMRs/epimutations
1
2
3
4
log10 of coverage
b
Expected DNA methylation signal (RPKM) within DMRs
1,000,000
100,000
10,000
1,000
100
10
1
FDR at 5%
r2=0.99 p<0.0001
Observed DNA methylation signal (RPKM) within DMRs
Dilution (% of CRC DNA)
100%
10%
1%
0.1%
0.01%
0.001%
c
human ctDNA (% of total cfDNA)
2.5
2.0
1.5
1.0
0.5
0.0
Input
IP
Samples
PDX1
PDX2


a
Early stage pancreatic adenocarcinoma patients (n=24)
Surgical resection: tumor and normal tissue
RRBS: tumor and normal tissue
Plasma: circulating cell-free DNA
cfMeDIP-seq: cell-free DNA
Age and sex matched healthy controls (n=24)
Plasma: circulating cell-free DNA
b
Plasma cfDNA (cases vs. controls)
14,716 DMRs
Loss
Gain
Significance (-log10 q-values)
DNA methylation (log2 fold change)
c
Plasma cfDNA
14,716 DMRs
Row z-score
Healthy controls
PDAC
d
Overlap plasma cfDNA (cases vs controls) and tumor DNA (primary tumor vs normal tissue)
Z-scores
3.39e-47
1.43e-22
ns
0.00956
Legend
A = Hypermethylated in primary tumor Hypermethylated in plasma cfDNA
B = Hypermethylated in primary tumor Hypomethylated in plasma cfDNA
e
Overlap plasma cfDNA (cases vs controls) and tumor DNA (primary tumor vs normal PBMCs)
Z-scores
<1e-745
6.12e-82
1.64e-36
7.56e-09
C = Hypomethylated in primary tumor Hypomethylated in plasma cfDNA
D = Hypomethylated in primary tumor Hypermethylated in plasma cfDNA


a
PDAC
AML
BLCA
BRCA
CRC
LUC
RCC
Z-scores
<1e-745
1.35e-77
1.87e-11
9.21e-08
6.87e-165
2.82e-53
1.43e-22
3.53e-295
7.78e-20
6.86e-10
9.22e-47
5.36e-108
9.68e-143
1.12e-43
1.65e-38
3.09e-50
e^757
8.64e-22
2.7e-36
2.09e-14
1.42e-19
2.05e-07
ns
8.64e-22
A
B
C
D
Legend
A = Hypermethylated in primary tumor Hypermethylated in plasma cfDNA
B = Hypermethylated in primary tumor Hypomethylated in plasma cfDNA
C = Hypomethylated in primary tumor Hypomethylated in plasma cfDNA
D = Hypomethylated in primary tumor Hypermethylated in plasma cfDNA
b
Discovery set (n=189)
AUROC
Sample type (class)
AML
BLCA
BRCA
CRC
LUC
Normal
PDAC
RCC
c
Validation set (n=199)
LUC (n=55)
AUROC = 0.971
AML (n=35)
AUROC = 0.980
PDAC (n=47)
AUROC = 0.918
Normal (n=62)
AUROC = 0.969
Sensitivity
1 - Specificity
d
Early stage PDAC (n=15)
AUROC = 0.914
Late stage PDAC (n=32)
AUROC = 0.920
Early stage LUC (n=32)
AUROC = 0.975
Late stage LUC (n=23)
AUROC = 0.966
Sensitivity
1 - Specificity


a
DMRs identified on the training set (plasma cfDNA)
DNA methylation data from training and validation sets (plasma cfDNA) (n=388)
t-SNE 1
t-SNE 2
t-SNE 3
Healthy donors
PDAC
LUC
BRCA
CRC
AML
RCC
BLCA
b
DMRs identified on the training set (plasma cfDNA)
DNA methylation data from TCGA (tissue) (n=4,032)
Blood
c
DMRs identified on the training set (plasma cfDNA)
DNA methylation data from COSMIC (cancer cell lines) (n=400)