LETTER

https://doi.org/10.1038/s41586-018-0703-0

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

Shu Yi Shen[1,12], Rajat Singhania[1,12], Gordon Fehringer[2,12], Ankur Chakravarthy[1,12], 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]*

The use of liquid biopsies for cancer detection and management is rapidly gaining prominence[1]. Current methods for the detection of circulating tumour DNA involve sequencing somatic mutations using cell-free DNA, but the sensitivity of these methods may be low among patients with early-stage cancer given the limited number of recurrent mutations[2–5]. By contrast, large-scale epigenetic alterations—which are tissue- and cancer-type specific—are not similarly constrained[6] and therefore potentially have greater ability to detect and classify cancers in patients with early-stage disease. Here we develop a sensitive, immunoprecipitation-based protocol to analyse the methylome of small quantities of circulating cell-free DNA, and demonstrate the ability to detect large-scale DNA methylation changes that are enriched for tumour-specific patterns. We also demonstrate robust performance in cancer detection and classification across an extensive collection of plasma samples from several tumour types. This work sets the stage to establish biomarkers for the minimally invasive detection, interception and classification of early-stage cancers based on plasma cell-free DNA methylation patterns.

The analysis of circulating tumour DNA (ctDNA) has numerous potential clinical applications. However, certain settings—such as cancer screening and the 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 obstacles to improved sensitivity of these methods include the limited number of recurrent mutations available to distinguish between tumour and normal circulating cell-free DNA (cfDNA) in a cost-effective manner, and technical artefacts (errors) introduced during sequencing. We reasoned that specific enrichment of methylated DNA fragments from cfDNA could overcome both of these issues.

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 that examined the detection probability across varying numbers of differentially methylated regions (DMRs), coverage and ctDNA abundance (Fig. 1a, Extended Data Fig. 1a). We found improved sensitivity as the number of DMRs increased, even at lower sequencing depth and ctDNA abundance, which suggests that the recovery of cancer-specific DNA methylation changes could enable highly sensitive and low-cost detection, classification and monitoring of cancer.

However, this is challenging in practice owing to the low abundance and the fragmented nature of plasma cfDNA[3], which have restricted

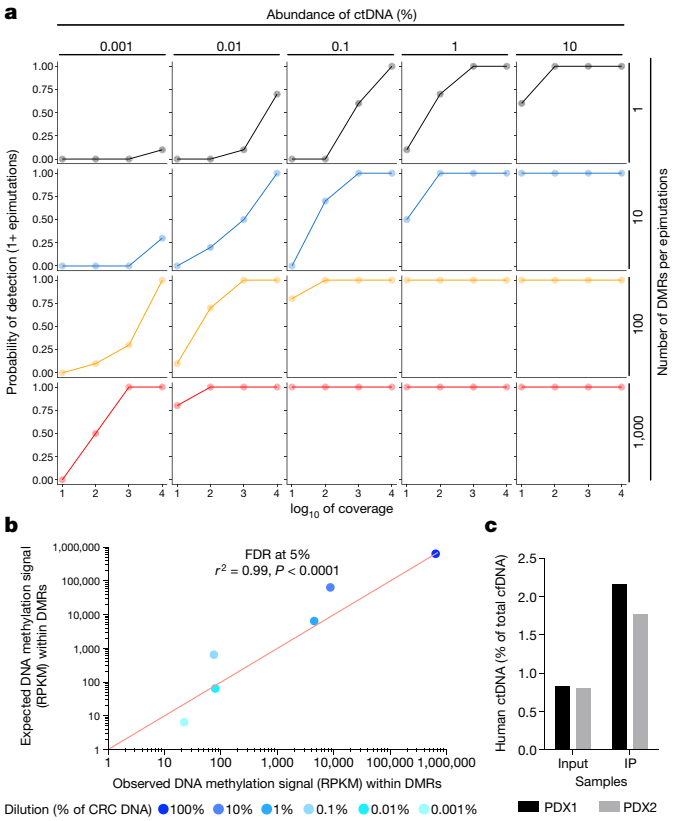


Fig. 1 | The cfDNA methylome as a sensitive approach to detect ctDNA in low levels of input DNA. a, Simulated probability of detecting at least one epimutation as a function of ctDNA concentration (0.001% to 10%; columns), number of DMRs analysed (1 to 10,000; rows) and sequencing depth (10× to 10,000×; *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 DNA, near-perfect correlations are observed between observed and expected methylation signal within DMRs in reads per kilobase of transcript per million mapped reads (RPKM). FDR at 5%, $r^2 = 0.99$; $P < 0.0001$. **c**, Frequency of ctDNA (human) as a percentage of total cfDNA (human + mouse) in the plasma from two colorectal cancer, patient-derived xenografts (PDX) before (input) and after (IP) cfMeDIP–seq.

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada. [3]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [5]Genome Technologies, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [6]UMR_S 1236, Univ Rennes 1, Inserm, Etablissement Français du sang Bretagne, 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, Ontario, Canada. [9]Fred Litwin Centre for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [10]Department of Surgery, Toronto General Hospital, Toronto, Ontario, Canada. [11]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [12]These authors contributed equally: Shu Yi Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy. *e-mail: rayjean.hung@lunenfeld.ca; ddecarv@uhnresearch.ca

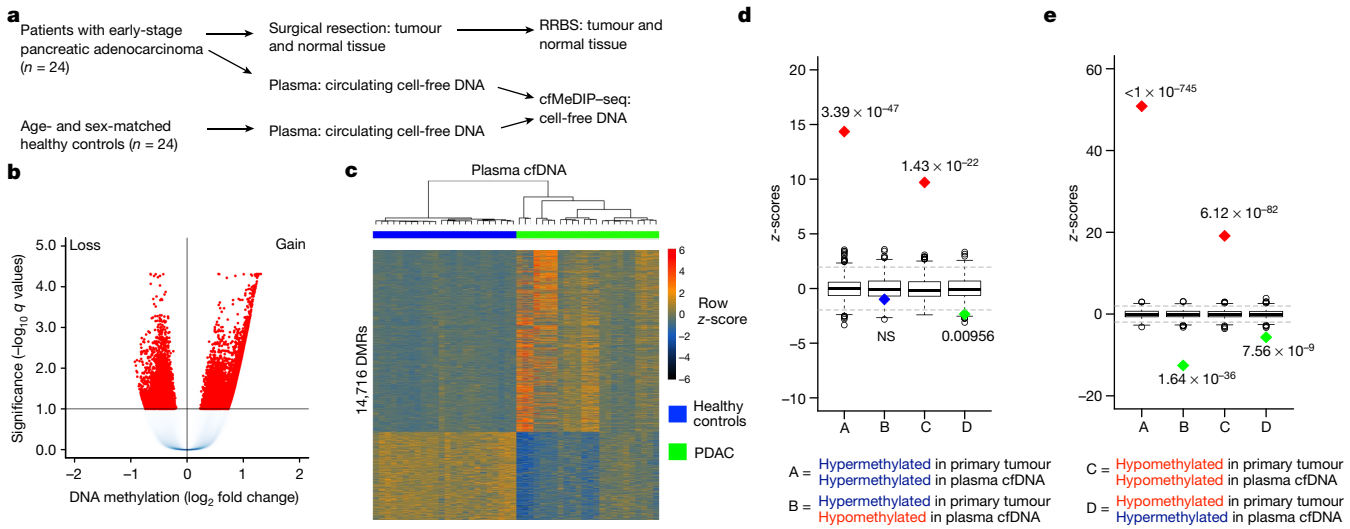


Fig. 2 | The cfMeDIP–seq method can identify thousands of DMRs in circulating cfDNA obtained from patients with pancreatic adenocarcinoma. a, Experimental design. **b**, Volcano plot of DMRs from patients with pancreatic cancer (cases, $n = 24$) versus healthy donors (controls, $n = 24$) using cfMeDIP–seq. Red dots indicate windows significant at BHFDR < 0.1 (negative binomial GLM, two-sided P values). **c**, Heat map of the 14,716 DMRs identified in the plasma cfDNA from cases and controls (Euclidean distance, Ward clustering). Dendrogram shows separation by status (case or control). **d**, **e**, Overlap between case-versus-control plasma-derived DMRs and RRBS tumour-DMR-matched normal tissue (**d**) and PBMCs (**e**). Box plots represent the expected null distribution of overlaps from 1,000 permutations (two-sided, P values computed using standard normal distribution). The extremes of the boxes define the upper and lower quartiles and the centre lines define the median. Whiskers indicate 1.5× 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.

most of the previous plasma methylation profiling to locus-specific PCR-based assays[7–9]. Although whole-genome bisulfite sequencing (WGBS) of cfDNA has been attempted[10,11], this approach is inefficient owing to degradation of around 84–96% of the input DNA during bisulfite conversion[12], high cost, and limited information recovery given the low genome-wide abundance of CpGs. Therefore, we developed cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP–seq) for genome-wide bisulfite-free plasma DNA methylation profiling. This method can enrich CpG-rich, potentially more informative fragments, thus enhancing cost-effectiveness.

In brief, we optimized an existing low-input MeDIP–seq protocol[13] that is robust down to 100 ng of input DNA, using exogenous *Enterobacteria phage* λ DNA (filler DNA) to increase the initial amount (Extended Data Fig. 1b). This is crucial for applications that are based on plasma cfDNA samples, which yield much less than 100 ng of cfDNA. We then performed extensive benchmarking of the optimized protocol. A comparison of low-input cfMeDIP–seq with gold-standard MeDIP–seq using colorectal cancer (CRC) HCT116 DNA that was 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–seq (100 ng), reduced representation bisulfite sequencing (RRBS) (1,000 ng) and WGBS (2,000 ng) (Extended Data Fig. 2e).

Next, cfMeDIP–seq was compared to ultra-deep hybrid capture mutation sequencing based on unique molecular identifiers (UMIs)[14] across a serial dilution of CRC DNA into multiple myeloma MM.1S cell-line DNA (Extended Data Fig. 3a). With cfMeDIP–seq, near-perfect linear associations were found 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, Extended Data Fig. 3b–e). Hybrid capture mutation sequencing, however, detected CRC-specific mutations down to only 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 that carry patient-derived xenografts was used for cfMeDIP–seq, and a twofold enrichment of human-tumour-derived cfDNA was found after 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 patients with primary early-stage pancreatic cancer (pancreatic ductal adenocarcinoma; PDAC) (cases) and 24 age- and sex-matched healthy controls (controls) (Fig. 2a, Extended Data Fig. 4a–f). In addition to plasma cfDNA, the microdissected primary tumours and adjacent normal tissue from the same patients with PDAC were used to generate DNA methylation profiles using RRBS. We identified 14,716 DMRs between the cfDNA of cases and controls (9,931 hypermethylated in cases, 4,785 in controls, based on negative-binomial generalized linear model (GLM) of fragment counts at a significance level of Benjamini–Hochberg FDR (BHFDR) of 0.1) (Fig. 2b, c, Supplementary Table 1).

In comparison, 45,173 differentially methylated CpGs (DMCs) were found between tumour 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 and DMC pairs that are concordantly hypermethylated ($P = 3.39 \times 10^{-47}$) and concordantly hypomethylated ($P = 1.43 \times 10^{-22}$) in the case of cfDNA and tumour tissue. This significant enrichment was not observed in the discordant methylation pattern between cfDNA and tumour DNA (Fig. 2d). Furthermore, signals in overlapping plasma cfDNA and tissue DNA methylation were correlated (Extended Data Fig. 5a). These findings suggest that cfMeDIP–seq of plasma cfDNA can detect tumour-derived DNA methylation events in ctDNA.

As non-tumour-derived cfDNA is mostly released from blood cells, we performed similar permutation-based enrichment testing between case-versus-control cfMeDIP–seq DMRs and the 95,388 RRBS DMCs between PDAC tumour tissue ($n = 24$) and normal peripheral blood mononuclear cells (PBMCs) ($n = 5$) (Supplementary Table 3). Again, we observed significant enrichment for concordant hypermethylated ($P < 1 \times 10^{-745}$) and hypomethylated ($P = 6.12 \times 10^{-82}$) sites

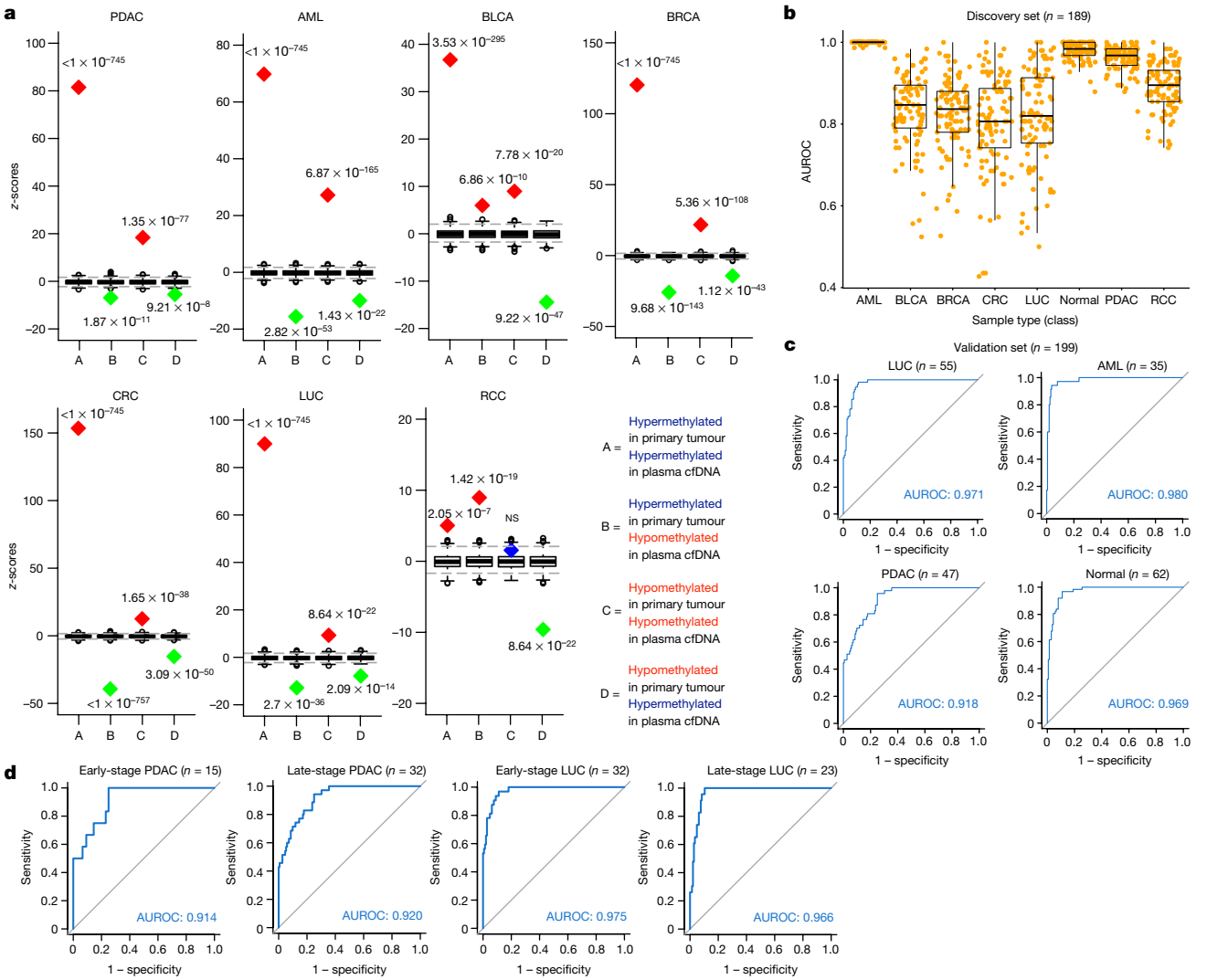


Fig. 3 | Methylome analysis of plasma cfDNA enables tumour classification. a, cfMeDIP–seq carried out on a discovery cohort consisting of 189 samples from seven different tumour types: PDAC, AML, BLCA, BRCA, CRC, LUC and RCC, including early- and late-stage tumours, and healthy controls (normal). For 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-tumour DMRs (tumour tissue versus adjacent normal tissue) for the corresponding cancer type using TCGA data. Box plots represent the expected null distribution of overlaps from 1,000 permutations (two-sided, *P* values computed using standard normal distribution). The extremes of boxes define the upper and lower quartiles and the centre lines define the medians. Whiskers indicate 1.5× 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 versus-other-classes binomial GLMnet classifiers. The *y* axis depicts distributions of AUROC for each held-out test set for each class. Dots represent performance in individual test sets. The extremes of boxes define the upper and lower quartiles and the centre lines define the medians. Whiskers indicate 1.5× IQR. **c**, ROC 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-versus-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.

in cfMeDIP–seq DMRs and tumour compared with PBMC DMCs, whereas discordant calls were underrepresented (Fig. 2e). In addition, signals in overlapping DMRs and DMCs were correlated (Extended Data Fig. 5b), and altogether indicated that DMRs identified using cfMeDIP–seq, between cases and controls, were probably derived from ctDNA (Extended Data Fig. 5c).

On the basis of the enrichment of tumour-derived DMRs and the known methylation-specific variable binding of transcription factors[16], we hypothesized that cfMeDIP–seq methylomes could identify active transcriptional networks in tumours or other tissues using plasma cfDNA. Upon motif enrichment analysis on cfMeDIP–seq DMRs and taking methylation preferences of candidate transcription factors into account[16], we identified 42 transcription factors as binding in healthy controls and 52 as binding in cases of pancreatic cancer (Supplementary Tables 4, 5). As expected, the former included haematopoietic-lineage-specific transcription factors such as PU.1, NFE2 and GATA1, whereas the latter included the pancreas-associated transcription factors PTF1a, Onecut1 (HNF6) and NR5A2 (Extended Data Fig. 6a, c). Compared to random sets of transcription factors, those inferred as active in healthy controls are overexpressed in blood according to data from the Genotype-Tissue Expression (GTEx) project, whereas those inferred as active in cases of pancreatic cancer were found to be over-expressed in pancreatic tissues (according to GTEx data) and PDAC tissue (according to data from The Cancer Genome Atlas (TCGA;

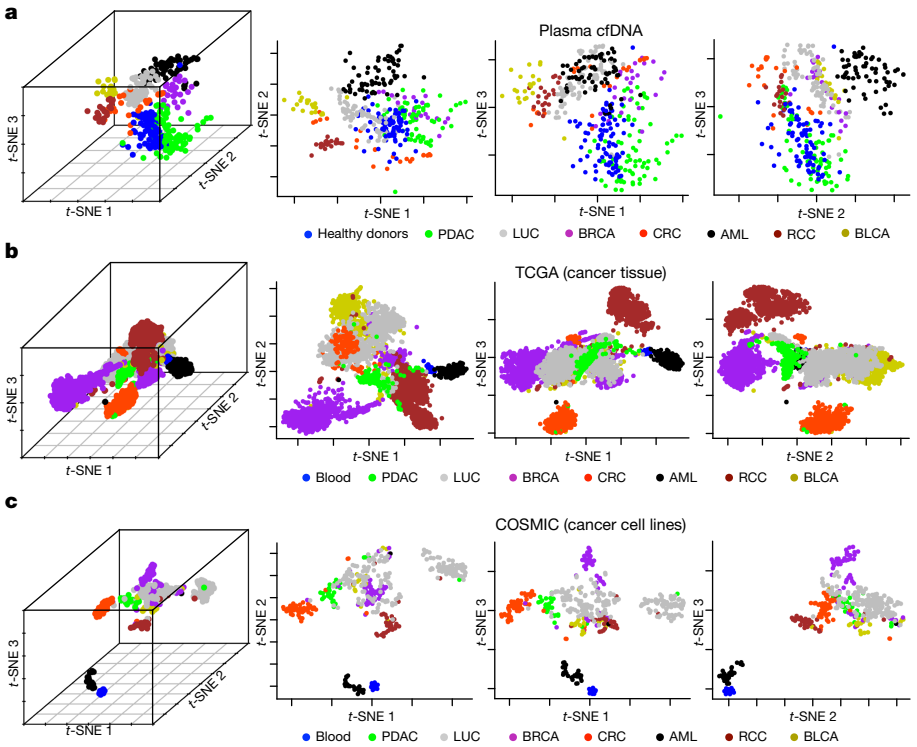


Fig. 4 | Plasma-derived DMRs are informative of cancer type. a, The plasma-derived DMRs identified as informative of cancer type in the discovery cohort of 189 plasma samples were used to generate 3D and 2D *t*-SNE plots for the entire cohort of plasma samples ($n = 388$). **b**, **c**, The DNA methylation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

Extended Data Fig. 6b, d, e)). Collectively, these findings indicate that cfMeDIP–seq might permit non-invasive characterization of active transcription-factor networks in cancer.

Given that we could detect tumour-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 seven different tumour types (PDAC, CRC, breast cancer (BRCA), lung cancer (LUC), renal cancer (RCC), bladder cancer (BLCA) and acute myeloid leukaemia (AML)) and healthy controls (Extended Data Figs. 7a–l, 8a).

We first identified plasma cell-free DMRs for each tumour type relative to healthy controls. We then asked whether these cancer-type-specific DMRs identified on the plasma cfDNA were enriched for the expected tumour DMRs for each cancer type using tumour tissue methylation data from TCGA ($n = 4{,}032$) (Fig. 3a). We observed a marked enrichment of sites that were hypermethylated in the primary tumour 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 tumour 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 into balanced training (80%) and test (20%) sets. Using only training-set samples, we selected the top 300 DMRs by limma-trend test statistic for each class compared with other classes. We then trained a series of one-versus-other-classes regularized binomial GLMs using these features on the training-set data. The training procedure consisted of three rounds of 10-fold cross-validation across a grid of values for alpha and lambda with optimisation for Cohen's kappa. The use of multiple rounds of 10-fold cross-validation was motivated by a desire to leverage additional randomization for more generalizable 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], culminating in a collection of 800 models, with 100 models for each one-versus-all-others comparison (hereafter termed E100). High values of the area under the receiver operator characteristic curve (AUROC) were observed for test-set samples across classes (Fig. 3b, Extended Data Fig. 9a).

Subsequently, we assessed performance across batches by applying the ensemble to a 199-sample validation cohort (35 AML, 47 PDAC, 55 LUC and 62 healthy controls). Averaging the class probabilities output by E100 for each sample yielded high AUROCs for AML versus others (0.980), PDAC versus others (0.918), LUC versus others (0.971) and normal versus others (0.969) (Fig. 3c). Notably, performance was similar between early- and late-stage samples, suggesting applicability to the detection of early-stage cancers (Fig. 3d, Extended Data Fig. 9b).

We then investigated whether the DMRs (non-zero coefficients) selected during the training of E100 were tumour-specific. Visualization using *t*-distributed stochastic neighbour embedding (*t*-SNE) plots showed clear separation by tumour type in the plasma cohort (Fig. 4a). This was notably reproduced in the 450k dataset of 4,032 TCGA cancers and normal blood samples, and 400 cancer cell lines from the Catalogue Of Somatic Mutations In Cancer (COSMIC) and PBMCs (Fig. 4b, c). This suggests that our plasma cfDNA methylation classifiers are mainly driven by tumour-specific DNA methylation patterns rather than by fluctuations in blood cells or cell composition in the tumour microenvironment.

However, these results do not rule out that some plasma cell-free DMRs could originate from changes in the proportions of circulating immune cells[18,19]. To further test our inference, we identified 38,352

cfMeDIP windows that were lowly methylated across a range of leukocyte types in WGBS data from the International Human Epigenome Consortium (IHEC), of which 27,088 overlapped with the TCGA 450k data (Extended Data Fig. 10a). Out of these 27,088 regions, we separated those that were identified as hypermethylated through the comparisons of plasma cfDNA of each cancer type to healthy controls. We then checked the methylation status of these regions in the tumour 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). We found these regions to be hypermethylated in tumour tissue (Extended Data Fig. 10b), reinforcing the hypothesis that these plasma cell-free DMRs are a direct measurement of tumour-derived DNA (that is, 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 tumour detection for cancer interception, and for multi-cancer classification.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0703-0.

Received: 5 December 2016; Accepted: 25 September 2018;
Published online: 14 November 2018

1. Diaz, L. A., Jr & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* **32**, 579–586 (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 (2017).
3. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
4. Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
5. Phallen, J. et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (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 (2014).
7. Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl Acad. Sci. USA* **113**, E1826–E1834 (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.* **35**, 751–758 (2017).
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 (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. USA* **110**, 18761–18768 (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. USA* **112**, E5503–E5512 (2015).
12. Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.* **29**, E65 (2001).
13. Taiwo, O. et al. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* **7**, 617–636 (2012).
14. Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
15. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27–36 (2010).
16. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
17. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–492 (2005).
18. Pedersen, K. S. et al. Leukocyte DNA methylation signature differentiates pancreatic cancer patients from healthy controls. *PLoS ONE* **6**, e18223 (2011).
19. Teschendorff, A. E. et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE* **4**, e8274 (2009).

Acknowledgements This study was conducted with support from the University of Toronto McLaughlin Centre (MC-2015-02), the 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 the Princess Margaret Cancer Foundation to D.D.D.C. as well as Canadian Cancer Society (CCSRI 701717) to R.J.H., CCSRI 704716 to R.J.H. and D.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 G.L. We acknowledge the Princess Margaret Genomics Centre for carrying out the next-generation sequencing and the Bioinformatics and HPC Core, Princess Margaret Cancer Centre for their expertise in generating the next-generation sequencing data.

Reviewer information *Nature* thanks E. Collisson, A. Teschendorff and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.Y.S. and D.D.D.C. designed and developed the cfMeDIP–seq protocol. R.J.H. and G.F. conceived and designed the study related to the pancreatic cancer component. S.Y.S., R.S., A.C. and D.D.D.C. conceived and designed the study related to the other cancer types. S.Y.S., S.V.B., T.J.P. and D.D.D.C. designed the experiments. S.Y.S., D.C., M.H.A.R., P.C.Z., Z.C., T.L., O.K., D.R., I.E., Z.C., S.C., G.M.O., J.L., M.M. and Z.Z. performed the experiments. T.d.S.M., Y.W. and C.O. performed the mouse experiments. R.S., A.C., G.F., T.T.W., A.G., T.J.P., M.M.H. and D.D.D.C. analysed the data with scientific input from R.J.H. G.F., A.B., D.C., A.S., T.M., A.A., N.L., M.H.A.R., J.D.M., P.L.B., N.F., G.L., M.D.M., S.G., T.J.P. and R.J.H. collected the clinical data related to the samples, determined the sample selection criteria and matching scheme, and provided the clinical samples. S.Y.S., R.S., A.C. and D.D.D.C. wrote the paper with feedback from all authors.

Competing interests D.D.D.C., S.Y.S., A.C., S.V.B., R.S. and R.J.H. are listed as inventors/contributors on patents filed related to this work.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0703-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0703-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.J.H. or D.D.D.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. Plasma samples were blinded during the sample preparation and sequencing. Data analysis was performed unblinded on the discovery cohort and blinded on the validation cohort.

Bioinformatic simulation of tumour-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 tenfold increments. Diploid genomes (14,500, the expected copy number in 100 ng cfDNA) were then sampled from these mixtures and further sampled 10–10,000× in tenfold 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, Extended Data Fig. 1a).

cfMeDIP–seq. A schematic representation of the cfMeDIP–seq protocol is shown in Extended Data Fig. 1b. Before cfMeDIP, the samples were subjected to library preparation using Kapa HyperPrep Kit (Kapa Biosystems), following the manufacturer's protocol with minor modifications. In brief, after end-repair and A-tailing, samples were ligated to 0.181 μM of NEBNext adaptor (NEBNext Multiplex Oligos for Illumina kit, New England BioLabs) 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) followed by purification with Qiagen MinElute PCR Purification Kit (MinElute columns) before 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 (C02010021) using a previously published protocol[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 adaptor-ligated cfDNA libraries. Its addition ensures a constant ratio of antibody to input DNA and helps to 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 *Arabidopsis 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 for the sample to be subjected to immunoprecipitation (79 μl). The included 5-mC monoclonal antibody 33D3 (C15200081) from the MagMeDIP kit was diluted 1:15 before generating the diluted antibody mix and was added to the sample. Washed magnetic beads (following the manufacturer's instructions) were also added before incubation at 4 °C for 17 h. The samples were purified using the Diagenode iPure Kit v2 (C03010015) and eluted in 50 μl of buffer C. The success of the reaction (QC1) was validated by qPCR to detect recovery of the spiked-in methylated and unmethylated *A. thaliana* DNA. The percentage recovery of unmethylated spiked-in DNA should be <1% (relative to input control, adjusted for input control being 10% of the overall sample) and the percentage 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), before proceeding to the next step. The optimal number of cycles to amplify each library was determined by 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 adaptor dimers. All the final libraries were submitted for BioAnalyzer analysis before sequencing at the Princess Margaret Genomics Centre on an Illumina HiSeq 2500, SBS V4 chemistry, single read 50 bp, multiplexed as seven samples per lane. After sequencing, the sequenced reads were aligned to λ and hg19 using Bowtie[20] with the default settings. On the basis of virtually no alignment to the λ genome, the filler DNA does not interfere with the generation of sequencing data (Supplementary Tables 7, 8).

The generated SAM files from hg19 alignment were converted to BAM format, ensuring the removal of duplicate reads, and the reads were then sorted and indexed using SAMtools[21] before subsequent analysis with the R package MEDIPS[22]. The 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), STR tested for authentication, mycoplasma free) was extracted using PureLink Genomic DNA Mini Kit (Thermo Fisher 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 were 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 two biological replicates per input. The fold enrichment of a methylated human DNA region (*HIST1H2BA*) over unmethylated human DNA region (*GAPDH* promoter), using primers provided in the MagMeDIP kit, was determined before 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 cell line MM.1S (source: American Type Culture Collection (ATCC), STR tested for authentication, mycoplasma free). A dilution series of CRC into multiple myeloma DNA was carried out following the scheme in Extended Data Fig. 3a. 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 five genes (13 kb) 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 >100,000× read coverage using Illumina HiSeq 2500 instrument, SBS V4 chemistry, paired-end 125 bp, as four samples per lane. Average target coverage of unprocessed reads was 186,312× (range: 154,419× – 216,434×) (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 adaptor ligation and a spacer filter was enforced to remove reads that were 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[27] (Q) of 30. Reads derived from the same strand of a unique fragment were collapsed to form SSCSs, suppressing polymerase and sequencer errors. These condensed reads were subsequently combined with their complementary strand into DCSs. This enables an additional layer of error suppression as double-strand consolidated sequences can correct for asymmetric damage accrued during the first cycle of PCR or induced by oxidation[28].

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

Specimen processing of patient-derived xenograft cfDNA. All mouse work was carried out in compliance with animal use protocol and ethical regulations approved by the Animal Care Committee at University Health Network (UHN). Human colorectal tumour tissue obtained with patient consent and UHN Research Ethics Board approval from the UHN Biobank was digested to single cells using collagenase A. Single cells were subcutaneously injected into 4–6-week-old NOD/SCID male mice. Mice were euthanized by CO_2 inhalation before blood was collected by cardiac puncture and stored in EDTA tubes. From the collected 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 the cfMeDIP–seq protocol as previously mentioned, sequenced and analysed (Supplementary Table 10).

Donor recruitment and sample acquisition. All patients provided written informed consent, and all samples were obtained upon approval of the institutional ethics committees and Research Ethics Boards from UHN and Mount Sinai Hospital, in compliance with all relevant ethical regulations. Pancreatic adenocarcinoma cases 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 Leukaemia Biobank. Bladder and renal cancer plasma samples were obtained from the UHN Genitourinary Biobank from consenting urologic oncology patients, procured before nephrectomy and cystectomy respectively. Healthy controls were recruited through the Family Medicine Centre at Mount Sinai Hospital in Toronto, Canada.

Specimen processing and methylation analysis of purified tumour 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. Laser capture microdissection was performed on the same day as 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 analysed using RRBS following a previously published protocol[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 the manufacturer's protocol, followed by gel size selection for fragments of 160–300 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 before sequencing at the Princess Margaret Genomics Centre on an Illumina HiSeq 2000, using sequencing by synthesis (SBS) V3 chemistry, single read 50 bp and multiplexed as four samples per lane. After sequencing, the raw data for each sample was trimmed with Trim Galore! using the RRBS settings before 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 for patient cfDNA. Plasma samples collected using EDTA and acid citrate dextrose tubes were obtained from the UHN BioBanks and Mount Sinai Hospital 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 before use. The sex, age and pathology stage of the patients from which the samples were collected are available in Supplementary Table 12, and extracted DNA quantities are available in Extended Data Fig. 8a.

Calculation and visualization of differentially methylated regions from cfDNA of patients with pancreatic cancer and healthy donors. DMRs between cfDNA samples from 24 patients with pancreatic cancer (PDAC) 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 visualization.

Enrichment analyses for plasma-derived DMRs in tumour-specific methylation signals in PDAC. Five normal PBMC samples profiled by RRBS were downloaded from the Gene Expression Omnibus (accession number GSE89473) for comparison with the 24 pancreatic cancer tissue RRBS samples. The 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 1,000 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 seven 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 ($n = 3{,}979$) (obtained from SAGE synapse) and PBMCs ($n = 53$, obtained from the Gene Expression Omnibus) 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 supplementary R Markdowns in Zenodo (ID 10.5281/zenodo.1205756) (Supplementary Table 13)—and median expression per tissue was visualized in heat maps. To look for enrichment of transcription factor expression and DMR-associated transcription factor motifs, we selected 1,000 random sets of transcription factors. As part of the analysis, we considered the known sensitivity to the methylation status of each transcription factor[16], yielding 42 transcription factors that are enriched in healthy donors and 52 that are enriched in pancreatic adenocarcinoma cases.

We computed ssGSEA (single-sample gene set enrichment analysis) scores for the expression of these transcription factors per sample, for pancreatic cancer (TCGA), blood (GTEx) and normal pancreas (GTEx) and compared distributions to those from random sets of transcription factors using Wilcoxon's Rank Sum Test. Violin plots were constructed as described in the supplementary 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 tumour 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 $\times$ 7 other classes = 2,100 features) using three iterations of 10-fold cross-validation 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 eight one-class-versus-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 the minimization of optimistic estimates owing 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-versus-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-versus-all classifiers trained to identify the class in question.

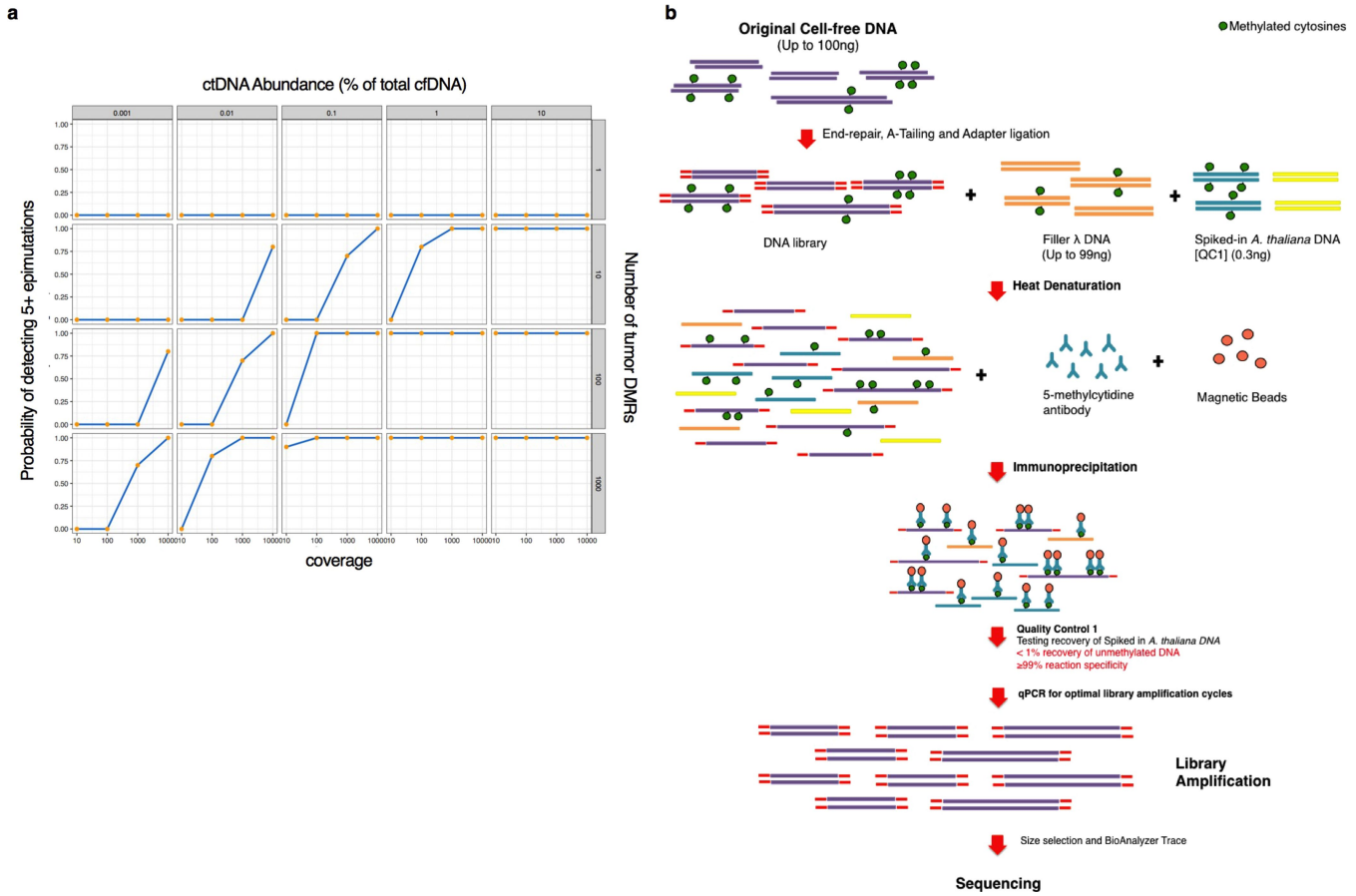
Validation in cell lines. 450K profiles for 1,028 cell lines previously characterized[37] were obtained as IDAT files. The data were then uniformly processed using the ssNoob method in the *minfi* package[38]. We reduced this dataset to tissue types for which cfMeDIP data were available ($n = 400$).

Data availability

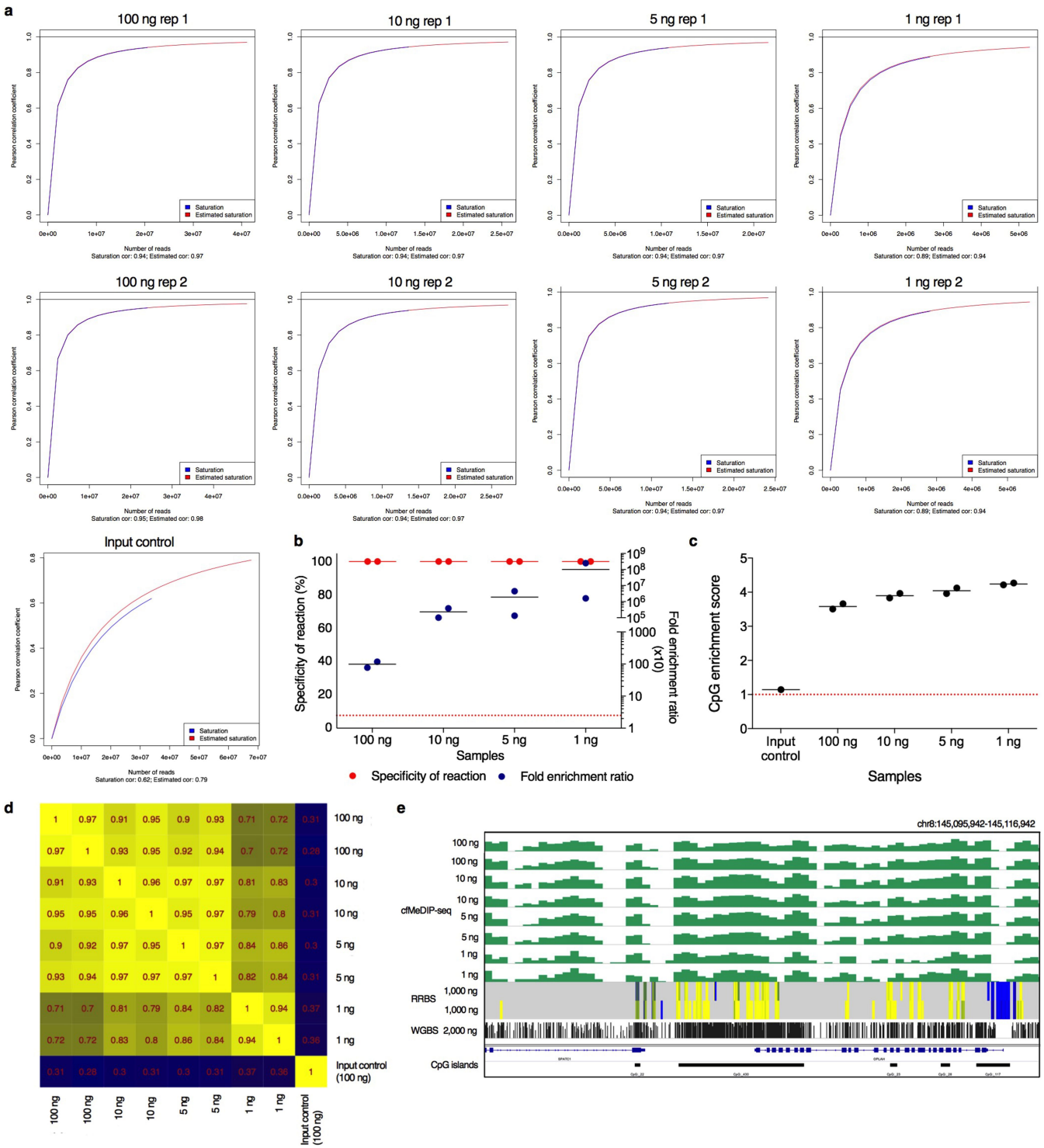
R markdowns (either knit or raw) and scripts used to generate the findings in this study have been deposited on Zenodo (DOIs in Supplementary Table 13). All the cell line datasets generated and/or analysed during the current study are available in the Gene Expression Omnibus repository under accession code GSE79838. The cfMeDIP–seq next-generation sequencing data for patient samples that support the findings of this study are available upon request from the corresponding author to comply with institutional ethics regulation. Source data for Fig. 1b and Extended Data Fig. 3e are provided in Supplementary Table 9, and for Fig. 1c in Supplementary Table 10. Additional source data can be found on Zenodo (Supplementary Table 13).

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 (2009).
21. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (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 (2014).
23. Kis, O. et al. Circulating tumour DNA sequence analysis as an alternative to multiple myeloma bone marrow aspirates. *Nat. Commun.* **8**, 15086 (2017).

24. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (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 (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. USA* **109**, 14508–14513 (2012).
29. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
30. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
31. Gu, H. et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–481 (2011).
32. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (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 (2014).
35. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, (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 (2014).
37. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (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 (2014).

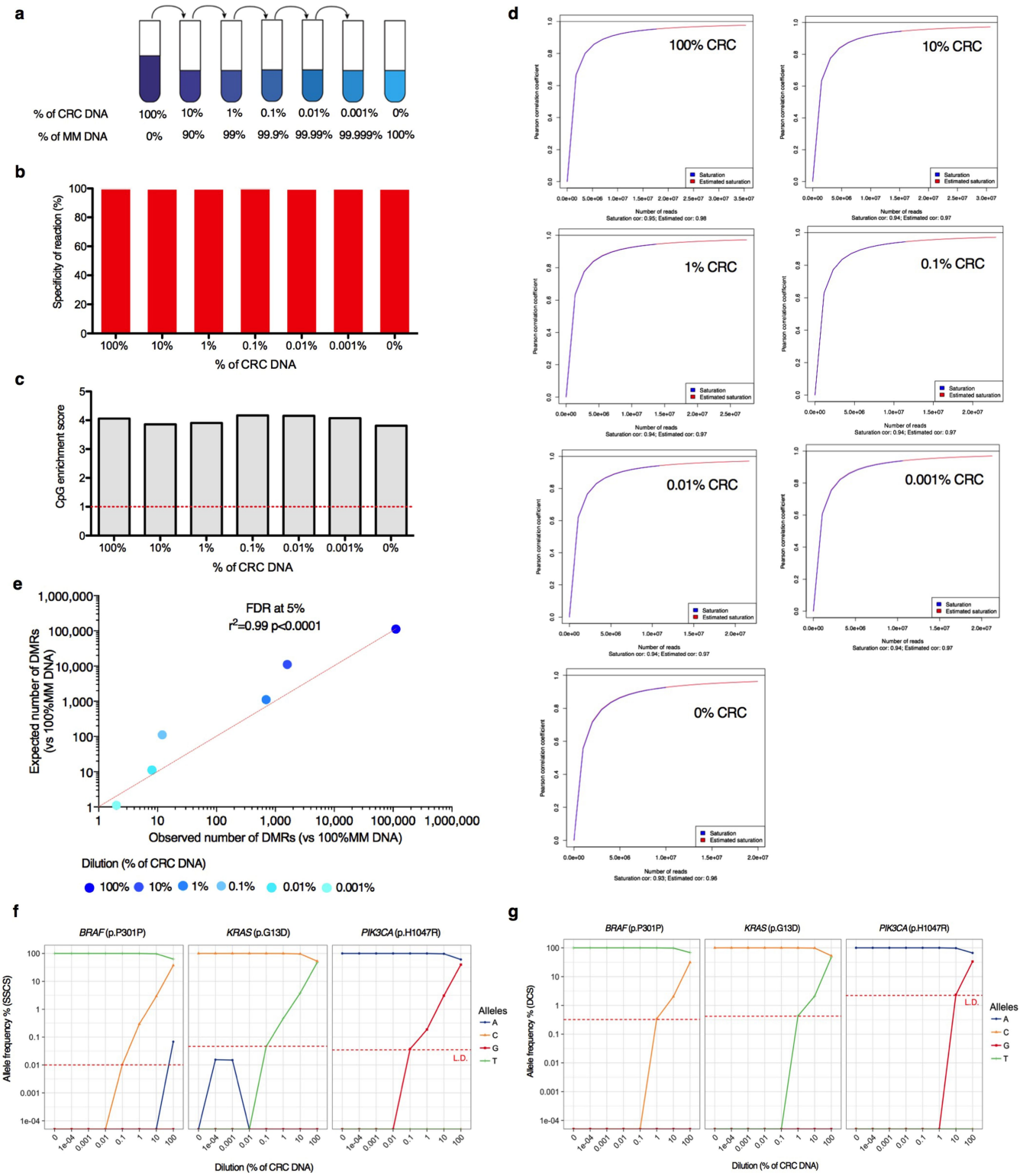


Extended Data Fig. 1 | Simulation of the probability of detecting ctDNA as a function of the number of DMRs, sequencing depth and percentage of ctDNA in plasma cfDNA, and a proposed method to enrich ctDNA. **a**, Bioinformatic simulation of scenarios with different proportions of ctDNA present in the sample (0.001% to 10%, columns), and a range of tumour-specific DMRs—from 1, 10, 100, 1,000 or 10,000—determined through the comparison of ctDNA to normal cfDNA (rows), with reads sampled at varying sequencing depths at each locus (10×, 100×, 1,000× and 10,000×) (*x* axis). The probability of detecting at least five epimutations per 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.



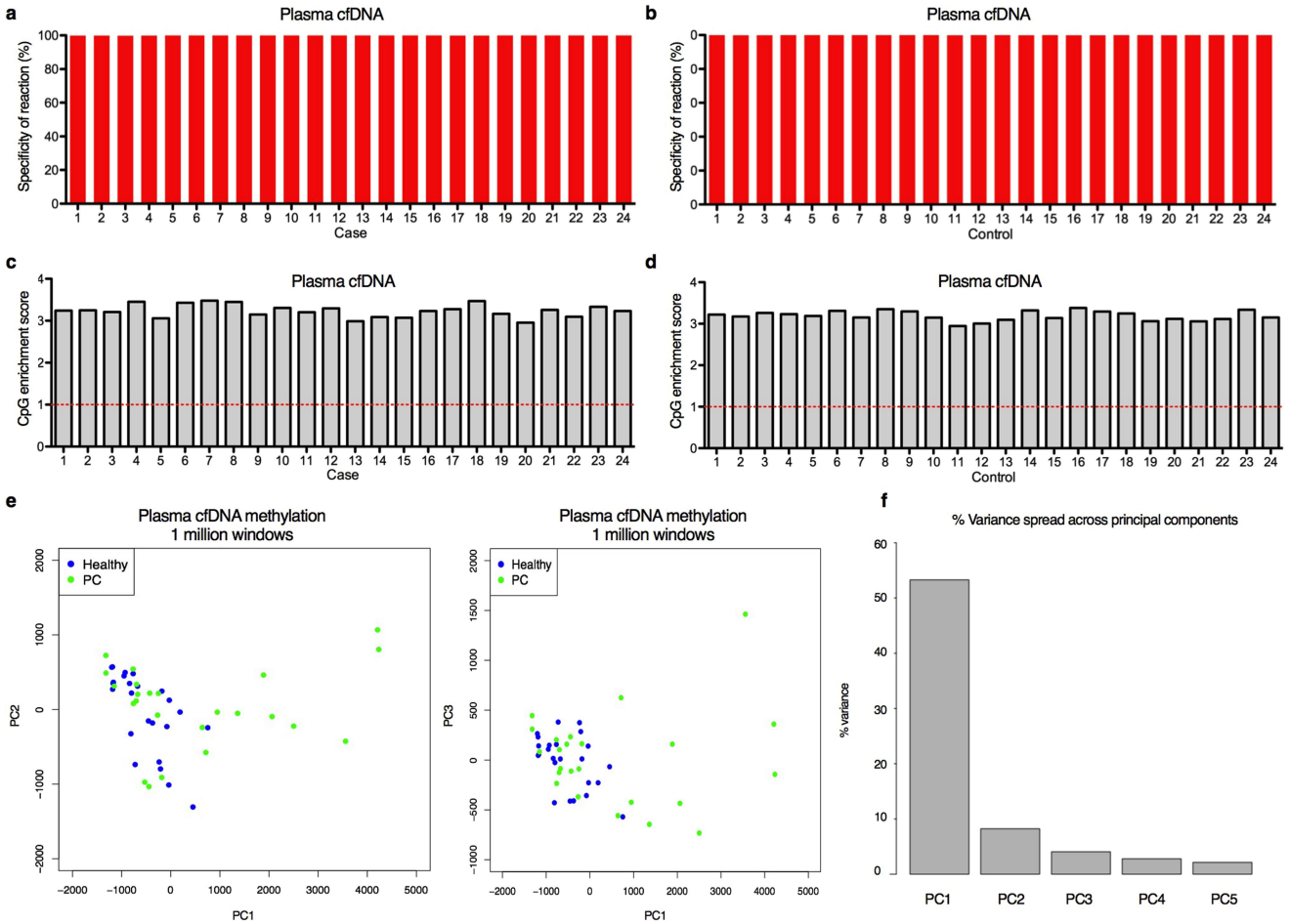
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Sequencing saturation analysis and quality controls of MeDIP–seq and cfMeDIP–seq carried out on varying starting inputs of HCT116 DNA sheared to mimic cfDNA. a, Results of the saturation analysis from the Bioconductor package MEDIPS analysing 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 DNA sheared to mimic cfDNA. The specificity of the reaction was calculated using methylated and unmethylated spiked-in *A. thaliana* DNA. The fold-enrichment ratio was calculated using genomic regions of the fragmented HCT116 DNA (human methylated *HIST1H2BA* and unmethylated *GAPDH*). The horizontal dotted line indicates a fold-enrichment ratio threshold of 25, dots represent biological replicates, with lines representing the 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 lines representing the 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 four examined loci, at different starting DNA inputs (1 to 100 ng, in biological replicates), compared with RRBS (ENCODE: ENCSR000DFS) and WGBS (Gene Expression Omnibus: GSM1465024) data (aligned to hg19). For cfMeDIP–seq, the *y* axis indicates RPKMs; for RRBS, yellow and blue blocks represent hypermethylated and hypomethylated CpGs, respectively. In the WGBS track, peak heights indicate methylation level.

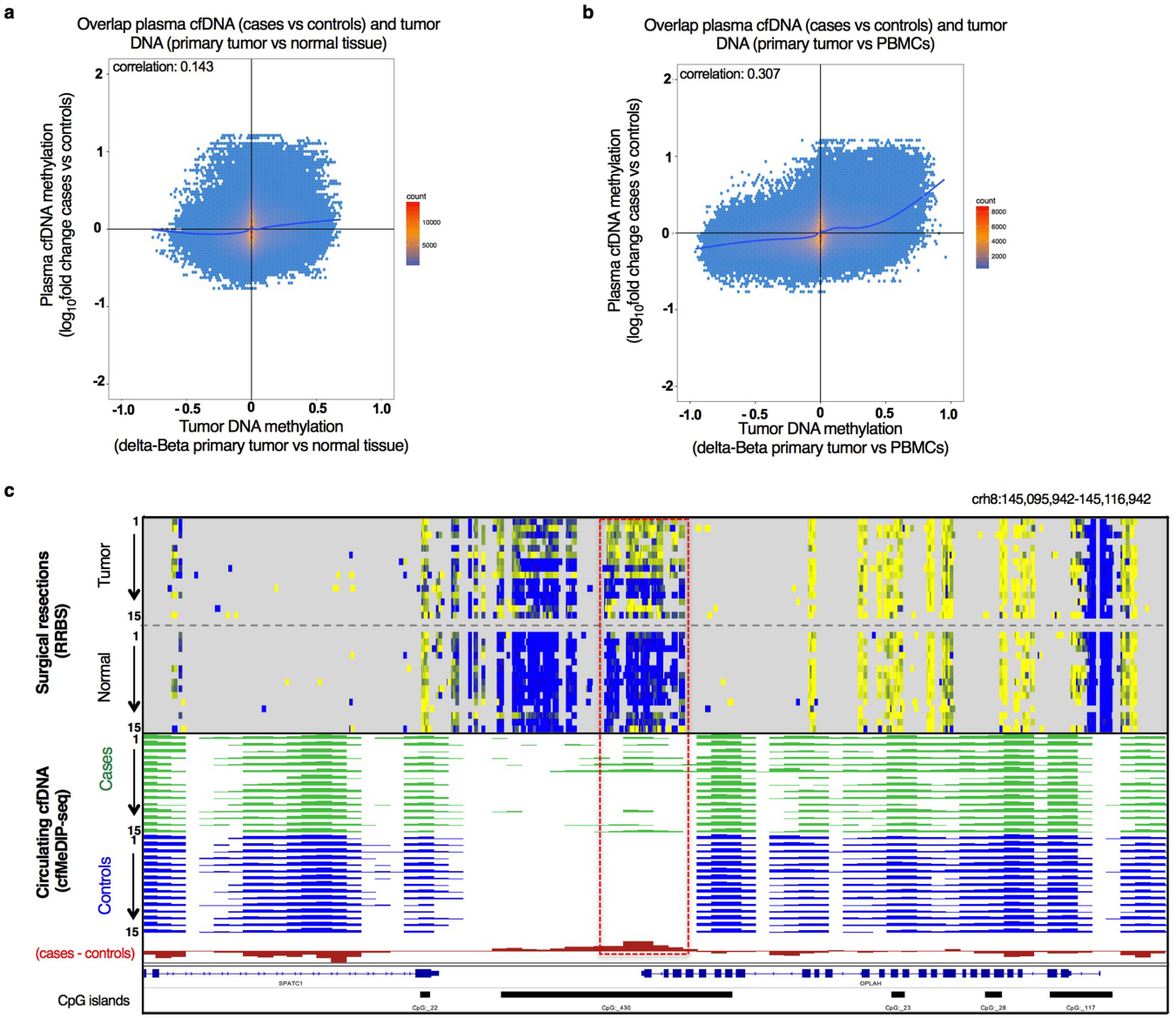


Extended Data Fig. 3 | See next page for caption.

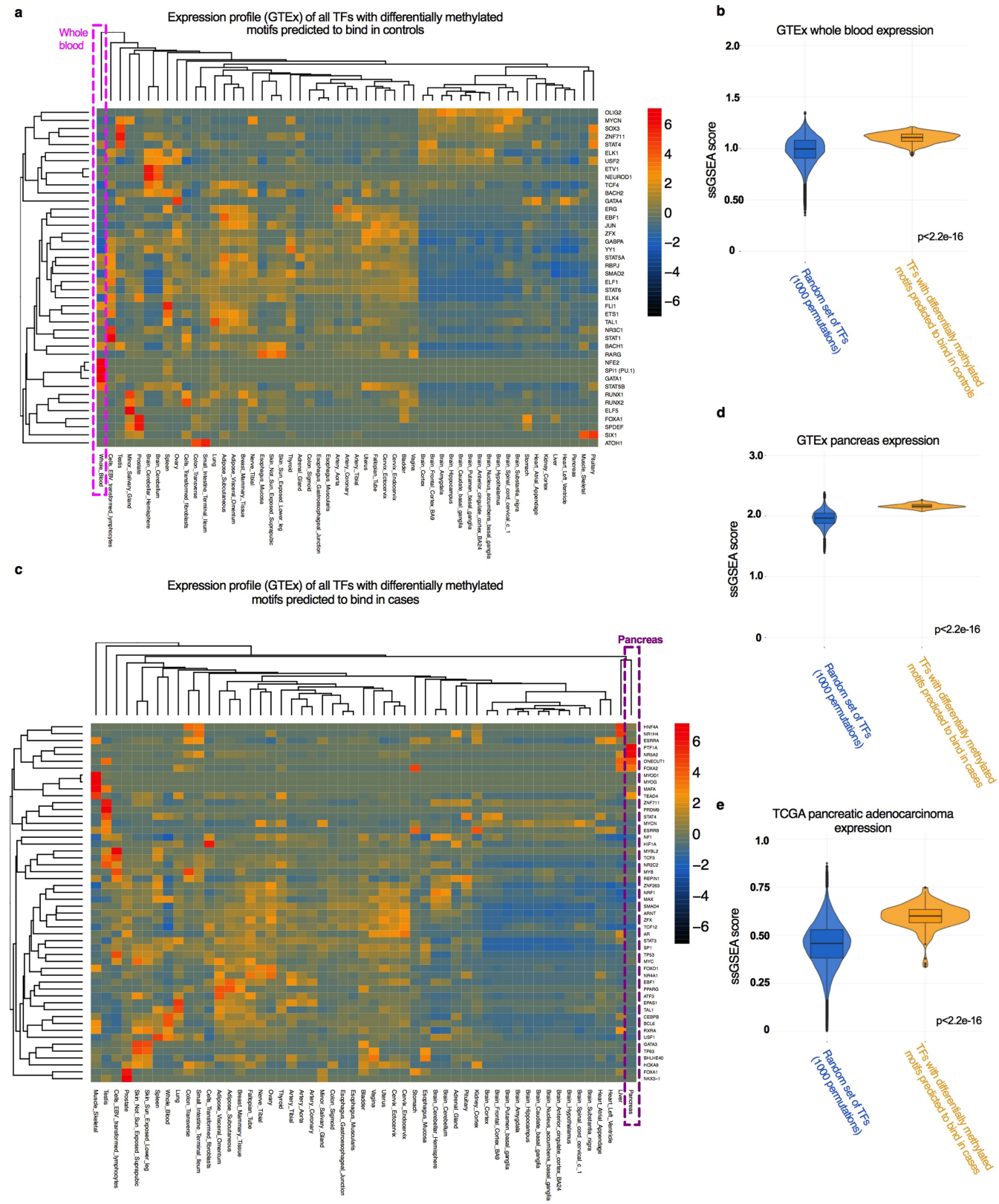
Extended Data Fig. 3 | Sequencing saturation analysis and quality controls of cfMeDIP–seq from serial dilution. a, Schematic representation of the CRC DNA (HCT116) dilution series into multiple myeloma DNA (MM.1S). For both CRC and multiple myeloma DNA, the genomic DNA was sheared to mimic cfDNA fragmentation. The entire dilution series was used to carry out cfMeDIP–seq ($n = 1$) and ultra-deep sequencing for mutation detection ($n = 1$). **b**, The specificity of the reaction for each dilution in the series ($n = 1$) was calculated using methylated and unmethylated spiked-in *A. thaliana* DNA. **c**, CpG enrichment representing the 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 by cfMeDIP–seq. The horizontal dashed line represents a 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 DNA, near-perfect correlations are observed between observed and 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 multiple myeloma DNA) ($n = 1$). UMIs were incorporated into the sequencing adapters and used to create SSCSs (**f**) and DCSs (**g**) for the detection of allele frequency for each mutation at each locus. For each mutation, the reference allele is found at the top. The dashed red line indicates the limit of detection.



Extended Data Fig. 4 | Quality control of cfMeDIP–seq from circulating cfDNA from patients with PDAC (cases) and healthy donors (controls). **a**, **b**, Specificity of reaction calculated using methylated and unmethylated spiked-in *A. thaliana* DNA for each case sample (**a**) and each control sample (**b**). The fold-enrichment ratio was not calculated owing 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**). The horizontal dashed line represents a 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 patients with PDAC, using the 1 million most variable windows by median absolute deviation (300 bp) genome-wide. Left, PC2 against PC1; right, PC3 against PC1. **f**, Percentage of variance explained by each principal component.

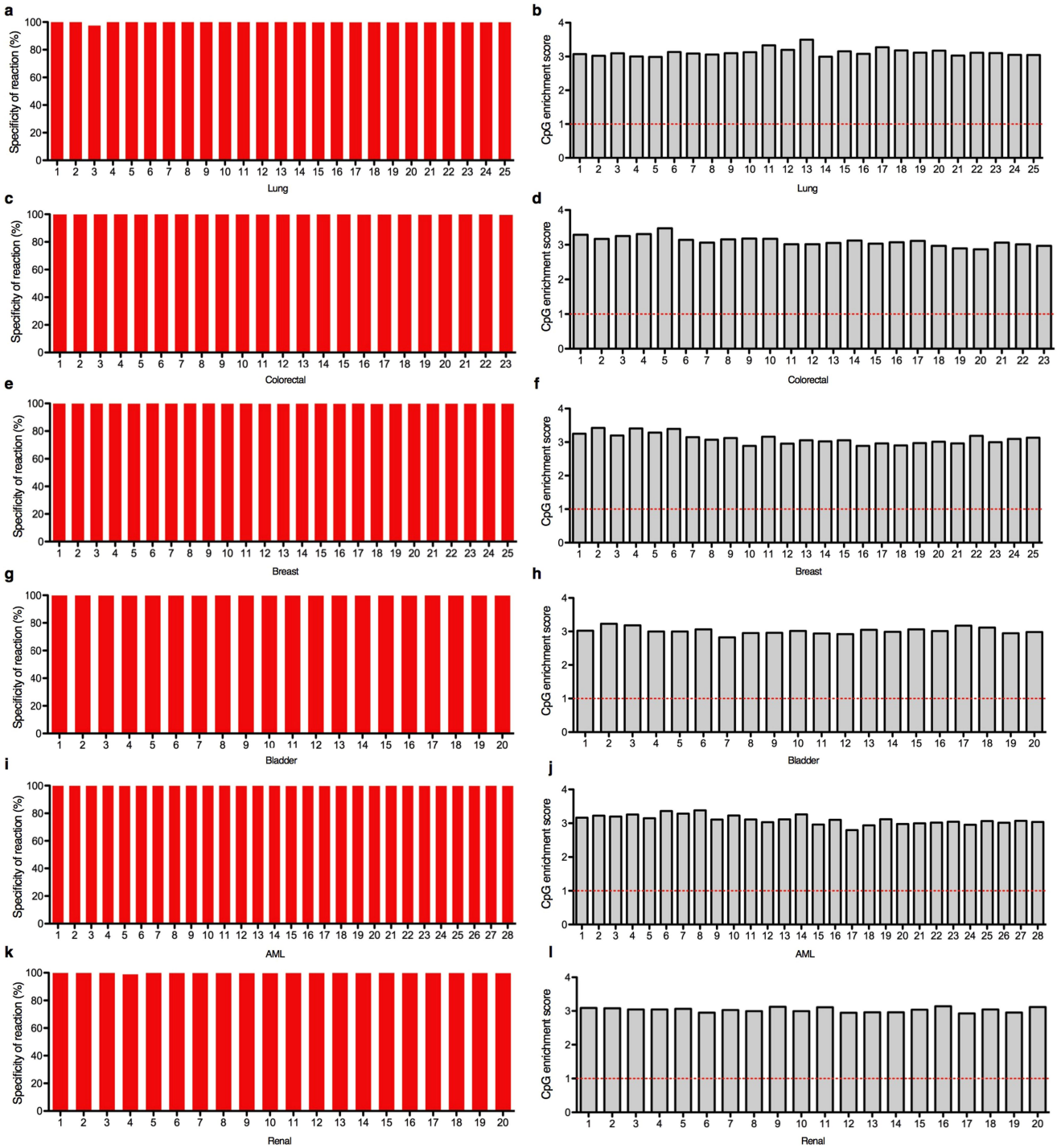


Extended Data Fig. 5 | Methylome analysis of plasma cfDNA distinguishes patients with early-stage PDAC from healthy controls. **a**, The difference in plasma cfDNA methylation plotted against the difference in tumour DNA methylation for each overlapping window ($n = 547{,}887$). The difference in plasma cfDNA methylation between patients with PDAC and healthy controls is $\log_{10}$-fold, as measured by cfMeDIP–seq. Tumour DNA methylation difference is delta beta from primary PDAC tumour to normal tissue, as measured by RRBS. The blue line is a trend line, with the correlation determined by Pearson's correlation. **b**, Scatter plot showing the DNA methylation difference for each overlapping window. The *x* axis shows the DNA methylation difference for the primary PDAC tumour compared with normal PBMCs from the RRBS data. The *y* axis shows the DNA methylation difference for the plasma cfDNA methylation from patients with PDAC compared with 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 laser capture microdissection tissues from PDAC tumour cases and the matching normal tissue, from the same patient, shown in the same order. Each coloured block represents 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 PDAC 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.

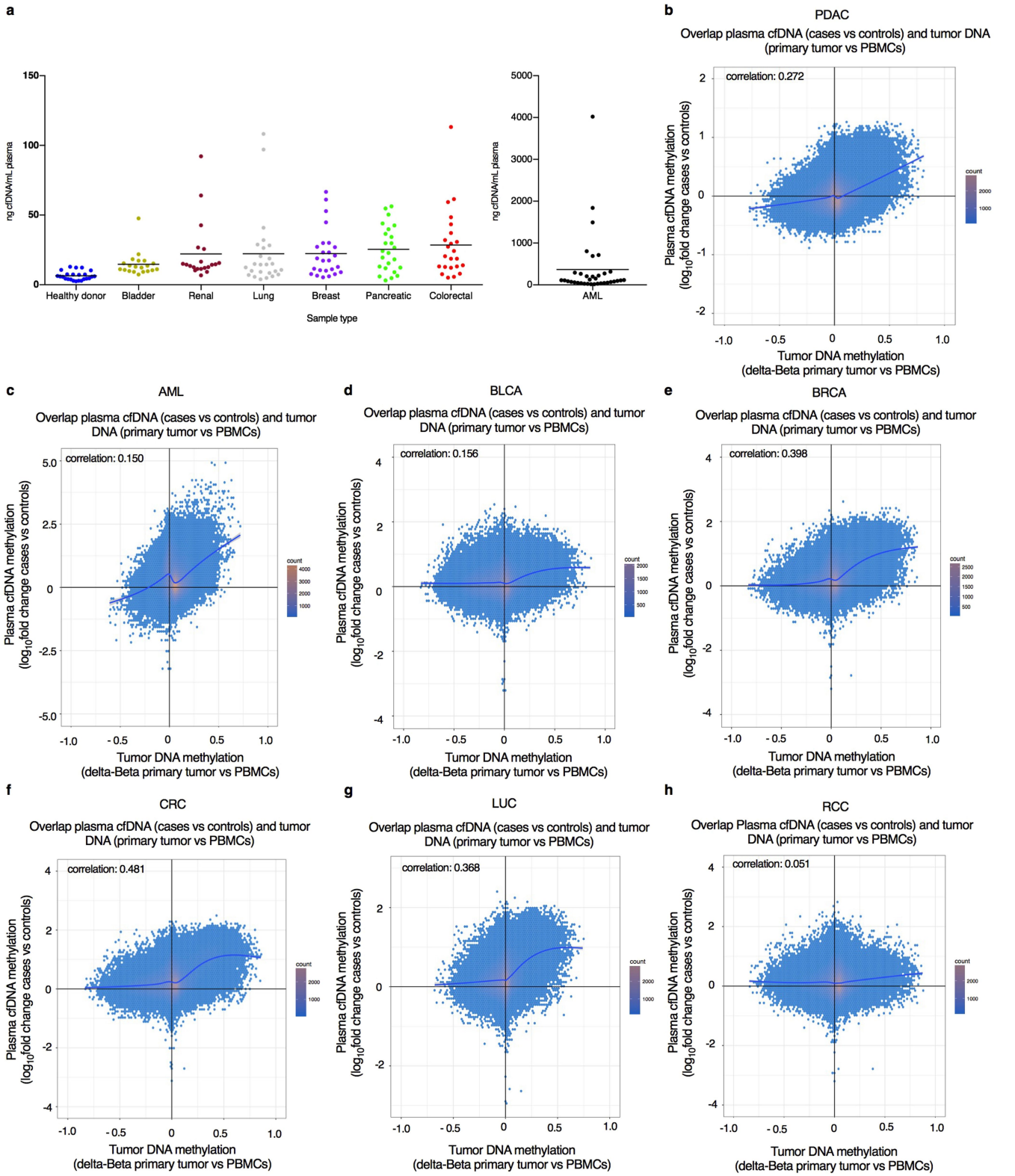


Extended Data Fig. 6 | See next page for caption

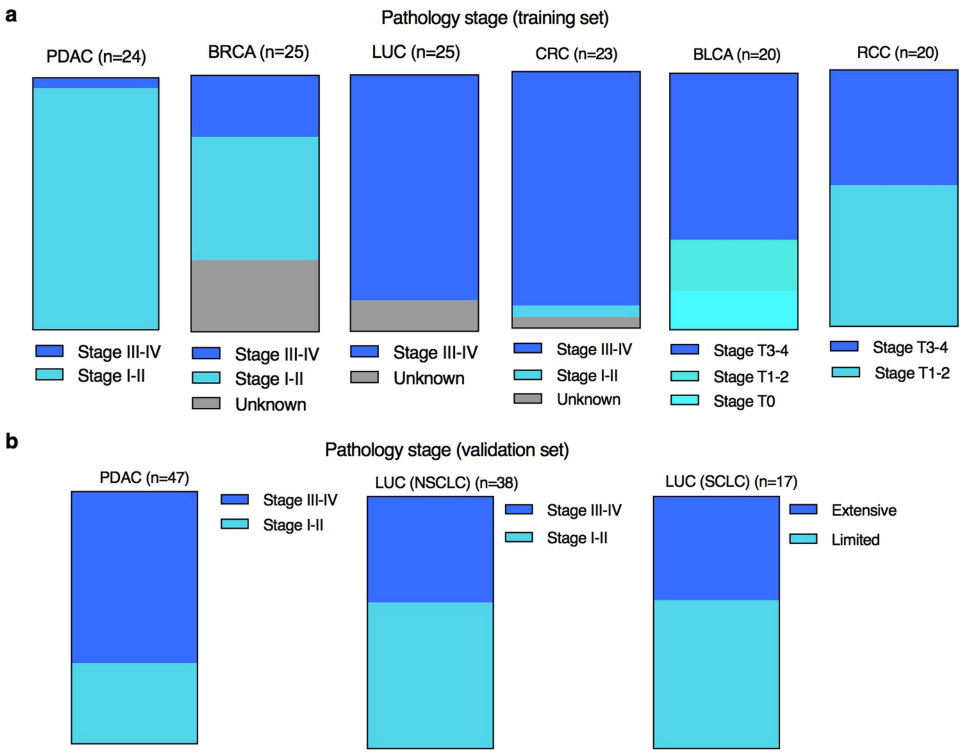
Extended Data Fig. 6 | Circulating cfDNA methylation profiles can identify transcription factor footprints and infer active transcriptional networks in the tissue of origin. a, Expression profile of all transcription factors ($n = 42$) that were characterized as binding in healthy controls across 53 human tissues from the GTEx project. Several transcription factors that are preferentially expressed in the haematopoietic system were identified (PU.1, NFE2 and GATA1). **b**, Expression profiles (ssGSEA scores; single-sample gene set enrichment analysis) of all transcription factors with hypomethylated motifs in controls ($n = 42$) are overexpressed compared with those of 1,000 random sets of 42 transcription factors across GTEx whole-blood data ($P < 2.2 \times 10^{-16}$, Wilcoxon's Rank Sum test, two-sided). **c**, Expression profile of all transcription factors ($n = 52$) characterized as binding in patients with PDAC. Several pancreas-specific or pancreatic-cancer-associated transcription factors were identified. Moreover, hallmark transcription factors that drive molecular subtypes of pancreatic cancer were also identified. **d**, Expression profile (ssGSEA scores) of all transcription factors with hypomethylated motifs in cases ($n = 52$) are overexpressed compared with those of 1,000 random sets of 52 transcription factors in the normal pancreas (GTEx data) (Wilcoxon Rank Sum test, two-sided test, $P < 2.2 \times 10^{-16}$). **e**, Expression profile of all transcription factors with hypomethylated motifs in PDAC cases ($n = 52$) are overexpressed compared those of 1,000 random sets of 52 transcription factors in PDAC tissue (TCGA data) (Wilcoxon Rank Sum test, two-sided test, $P < 2.2 \times 10^{-16}$). For violin plots (**b**, **d**, **e**) the ends of the boxes represent the lower and upper quartiles and the middle line indicates the median. Whiskers represent 1.5× IQR, and outliers are excluded. Rotated kernel densities are also displayed.



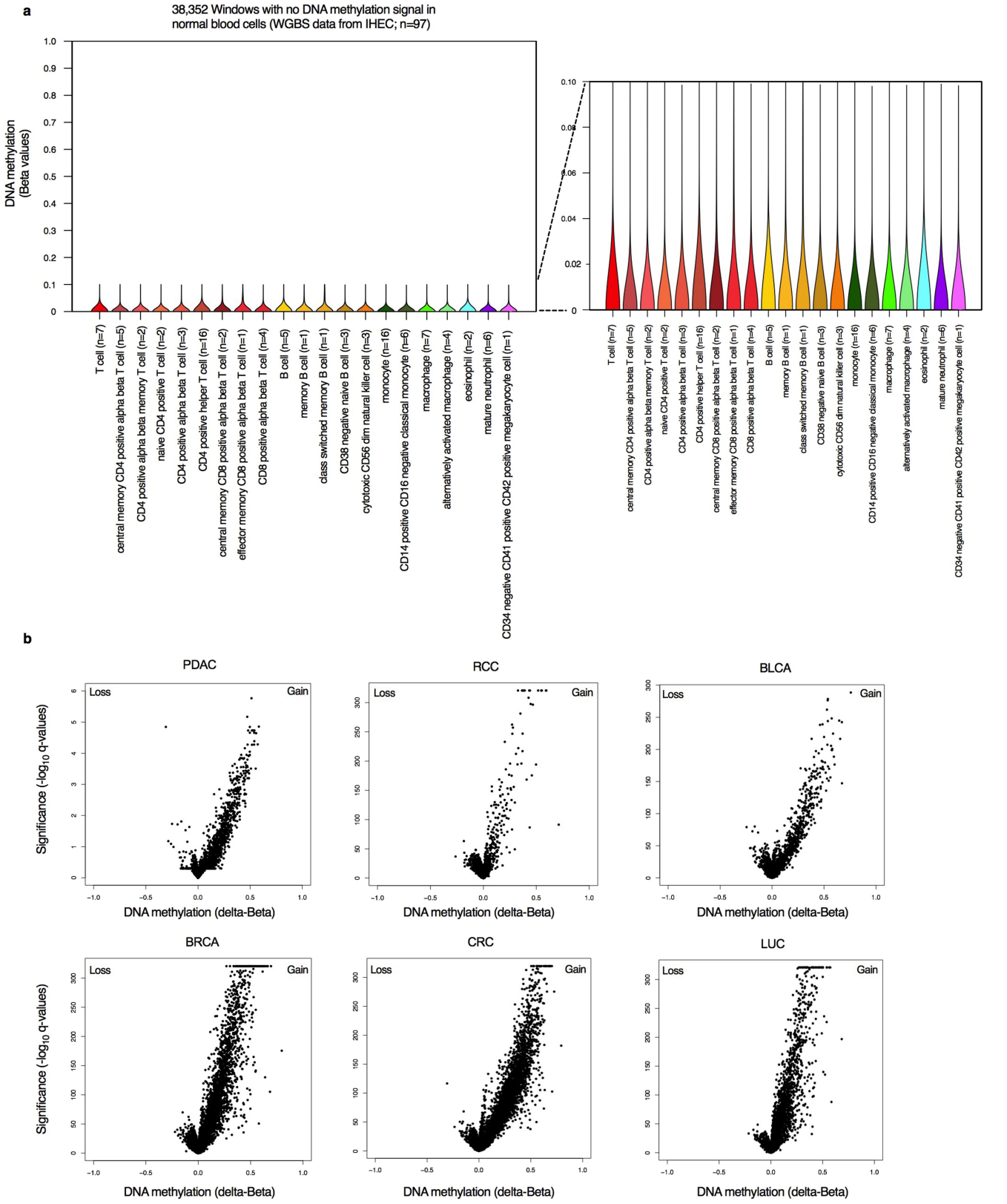
Extended Data Fig. 7 | Quality control of cfMeDIP–seq from circulating cfDNA from multiple cancer types. a, **c**, **e**, **g**, **i**, **k**, Specificity of the reaction; and **b**, **d**, **f**, **h**, **j**, CpG enrichment score for each sample per cancer type. The horizontal dashed lines represent a CpG enrichment of 1.



Extended Data Fig. 8 | Comparison of plasma cfDNA DMRs with tumour 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**, Scatter plots showing the DNA methylation difference for all overlapping windows in PDAC ($n = 245{,}980$ windows) (**b**), AML ($n = 206{,}735$ windows) (**c**), BLCA ($n = 193{,}943$ windows) (**d**), BRCA ($n = 204{,}623$ windows) (**e**), CRC ($n = 210{,}645$ windows) (**f**), LUC ($n = 193{,}043$ windows) (**g**) and RCC ($n = 198{,}390$ windows) (**h**). The *x* axis shows the DNA methylation difference between the primary tumour (TCGA data) and normal PBMCs. The *y* axis shows the DNA methylation difference between the plasma cfDNA methylation for each cancer type and healthy controls from the cfMeDIP–seq data. The blue line is a trend line, with the correlation determined by 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 (according to the AJCC/UICC 7th Edition) breakdown by tumour type for samples in the training set (**a**) and in the validation set (**b**). Non-small-cell lung carcinoma, LUC (NSCLC); 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 (plotted as beta value) of 38,352 regions in normal blood cells selected on the basis of low DNA methylation levels using IHEC whole-genome bisulfite sequencing data. For violin plots, the ends of the boxes represent the lower and upper quartiles and the middle line represents the median. Whiskers represent 1.5× IQR, and outliers are excluded. Rotated kernel densities are also displayed. **b**, Volcano plots 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. The *x* axis represents DNA methylation (plotted as delta beta value), obtained from tumour data from TCGA for cancers other than PDAC and RRBS for PDAC. The *y* axis represents $-\log_{10} q$ values (Benjamini Hochberg false discovery rate, BHFDR).

nature research

Corresponding author(s): Daniel D. De Carvalho and Rayjean J. Hung

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	qPCR data collection was carried out using StepOne Software v2.3 and Bio-Rad CFX Manager 3.1
Data analysis	Shell scripts were used to process MeDIP sequencing data to produce RPKM estimates for downstream analysis. Custom R scripts with a collection of R/Bioconductor packages were used to analyse the resulting dataset. The code used for analysis is available in the form of R markdowns deposited onto Zenodo as detailed in Supplementary Table 13. Open source code/packages used for data analysis: R, SAMtools (v. 1.3.1), Bowtie (v. 0.12.9), MEDIPS (v. 1.22.0), BWA (v. 0.7.1220), Genome Analysis ToolKit (GATK) IndelRealigner, MuTect (v.1.1.5) , Trim Galore! (v. 0.4.4), Bismark (v. 0.10.1), Bowtie2 (v. 2.0.5), MethylKit (v 0.99.2), DESeq2 (v. 1.4.5), caret R package, minfi bioconductor package, Rtsne Commercial software used for data analysis: Microsoft Excel for Mac (v. 16.16.1)

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

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. 1d and e are provided in Supplementary Table 9, and for Figure 1G are provided in Supplementary Table 10. Additional source data can be found on Zenodo (Supplementary Table 13).

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No sample-size calculations were performed. For human samples, sample sizes were determined based on availability. With regards to mouse samples, samples were also chosen based on availability of PDX mice.
Data exclusions	No data were excluded from the analyses.
Replication	During assay development using cell line DNA, cfMeDIP-seq was performed in duplicate (Figure 1b and 1c). For plasma cfDNA from patient samples, the assays were performed only once per sample because of the limited amount of plasma cfDNA available per patient.
Randomization	All human participants belonged to pre-defined groupings based on the cancer type. With regards to the mouse experiments, samples were not randomized, since there was no experimental groups. Plasma cfDNA from mice was immunoprecipitated and compared against the input (no immunoprecipitation) for each mouse. There was no control for other covariates since each animal was compared against itself (IP versus input).
Blinding	Plasma samples were blinded during the sample preparation and sequencing. Data analysis was performed unblinded on the discovery cohort and blinded on the validation cohort.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

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

Antibodies

Antibodies used	5-mC monoclonal antibody 33D3 (Cat#C15200081) from the MagMeDIP kit from Diagenode, diluted 1/15 prior to using 0.16 ug of antibody per reaction. Antibody lots used#: GF-003, D006, RD001
Validation	Antibody validation for against 5-mC DNA shown by Diagenode: https://www.diagenode.com/files/products/antibodies/

Datasheet_5-mC33D3_C15200081.pdf, using MeDIP-seq, Dot blot, immunofluorescence and Surface plasmon resonance (SPR) analysis

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HCT116 and MM.1S obtained from ATCC
Authentication	ATCC standard protocols - Human STR Profiling Cell Authentication
Mycoplasma contamination	HCT116 and MM.1S: mycoplasma free
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	4-6 week old NOD/SCID male mice.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not include field-collected samples from animal or other organisms

Human research participants

Policy information about studies involving human research participants

Population characteristics	Population characteristics for study participants can be found in Supplementary Tables 12 and Extended Data Fig. 9. In brief, the mean age of cohorts (healthy controls and the different cancer samples) ranged from 54 to 68. Most of the cancer groups included approximately equal proportion of male vs female participants per sample availability, with the exception of breast cancer samples, where all patients were female.
Recruitment	Pancreatic cancer patients were identified through the Ontario Cancer Registry. Eligible cases were residents of Ontario with a first primary, pathologically confirmed adenocarcinoma of pancreas or adenocarcinoma metastasis confirmed by treating physicians. The healthy controls were selected randomly from individuals registered in the family medicine clinics databases in the Greater Toronto Area and were frequency matched with pancreatic cancer cases on age and sex. All subjects were interviewed, and information on lifestyle risk factors, occupational history and medical and family history was collected using a standard questionnaire. Lung cancer patients were recruited as part of the Mount Sinai Hospital-Princess Margaret Hospital study, or lung cancer screening programs. The eligible lung cancer patients have histologically or pathologically confirmed lung cancer diagnosis. Samples from breast, colorectal, urological cancers and AML were obtained from the biobanks with no particular selection criteria. All patients have provided written informed consents, and all samples have been obtained upon approval of the institutional ethics committees (University Health Network and Mount Sinai Hospital).