



**HAL**  
open science

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

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

► **To cite this version:**

Shu Yi Shen, Rajat Singhanian, Gordon Fehring, Ankur Chakravarthy, Michael H. A. Roehrl, et al.. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*, 2018, 563 (7732), pp.579-583. 10.1038/s41586-018-0703-0 . hal-01974928

**HAL Id: hal-01974928**

**<https://univ-rennes.hal.science/hal-01974928>**

Submitted on 14 Jan 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

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

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

29

30

\*These authors made equal contributions

31

32

**§Corresponding authors:**

33

Daniel D. De Carvalho: [ddecarv@uhnresearch.ca](mailto:ddecarv@uhnresearch.ca)

34

Rayjean J. Hung: [rayjean.hung@lunenfeld.ca](mailto:rayjean.hung@lunenfeld.ca)

35

36

37

38

39

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

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

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

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

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

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

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

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

106 and adjacent normal tissue from the same PDAC patients were used to generate DNA  
107 methylation profiles via RRBS. We identified 14,716 DMRs between cases and controls cfDNA  
108 (9,931 hypermethylated in cases, 4,785 in controls, based on negative-binomial generalized  
109 linear model (GLM) of fragment counts at significance level of Benjamini Hochberg FDR  
110 (BHFDR) of 0.1) (Fig. 2b-c and Supplementary Table 1).

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

120 As non-tumor derived cfDNA is mostly released from blood cells, we performed similar  
121 permutation-based enrichment testing between case-vs-control cfMeDIP-seq DMRs and the  
122 95,388 RRBS DMCs between PDAC tumor tissue ( $n=24$ ) and normal peripheral blood  
123 mononuclear cell (PBMCs) ( $n=5$ ) (Supplementary Table 3). Again, we observed significant  
124 enrichment for concordant hypermethylated ( $p<1e-745$ ) and hypomethylated ( $p=6.12e-82$ ) sites  
125 in cfMeDIP-seq DMRs and tumor vs. PBMC DMCs, while discordant calls were  
126 underrepresented (Fig. 2e). In addition, signals in overlapping DMRs/DMCs were correlated

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

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

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



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

156 Finally, we carried out a set of machine learning analyses on our discovery cohort to rigorously  
157 evaluate the utility of cfMeDIP profiles in cancer detection and classification. We initially  
158 reduced our dataset to 505,027 windows mapping to CpG islands, shores, shelves and  
159 FANTOM5 enhancers for computational efficiency. Unbiased performance estimates, while  
160 accounting for training-set biases, were then derived from the reduced dataset. We split the  
161 discovery cohort 80%-20% into balanced training and test sets. Using only training set samples,  
162 we selected the top 300 DMRs by limma-trend test statistic for each class versus other classes.

163 We then trained a series of one-vs-other-classes regularised binomial GLMs using these features  
164 on the training set data. The training procedure consisted of 3 rounds of 10-Fold Cross-  
165 Validation (CV) across a grid of values for alpha and lambda with optimisation for Cohen's  
166 Kappa. The use of multiple rounds of 10-Fold CV was motivated by a desire to leverage  
167 additional randomisation for more generalisable model tuning.

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

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<sup>18,19</sup>. 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

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

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

#### 204 Reference

- 205 1 Diaz, L. A., Jr. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J Clin*  
206 *Oncol* **32**, 579-586, doi:10.1200/JCO.2012.45.2011 (2014).
- 207 2 Aravanis, A. M., Lee, M. & Klausner, R. D. Next-Generation Sequencing of Circulating  
208 Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574,  
209 doi:10.1016/j.cell.2017.01.030 (2017).
- 210 3 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor  
211 DNA with broad patient coverage. *Nat Med* **20**, 548-554, doi:10.1038/nm.3519  
212 (2014).
- 213 4 Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a  
214 multi-analyte blood test. *Science*, doi:10.1126/science.aar3247 (2018).
- 215 5 Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA.  
216 *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aan2415 (2017).
- 217 6 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular  
218 classification within and across tissues of origin. *Cell* **158**, 929-944,  
219 doi:10.1016/j.cell.2014.06.049 (2014).
- 220 7 Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using  
221 methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-1834,  
222 doi:10.1073/pnas.1519286113 (2016).
- 223 8 Visvanathan, K. *et al.* Monitoring of Serum DNA Methylation as an Early Independent  
224 Marker of Response and Survival in Metastatic Breast Cancer: TBCRC 005  
225 Prospective Biomarker Study. *J Clin Oncol*, JCO2015662080 (2016).

226 9 Potter, N. T. *et al.* Validation of a real-time PCR-based qualitative assay for the  
227 detection of methylated SEPT9 DNA in human plasma. *Clin Chem* **60**, 1183-1191,  
228 doi:10.1373/clinchem.2013.221044 (2014).

229 10 Chan, K. C. *et al.* Noninvasive detection of cancer-associated genome-wide  
230 hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.  
231 *Proc Natl Acad Sci U S A* **110**, 18761-18768, doi:10.1073/pnas.1313995110 (2013).

232 11 Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing  
233 for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad*  
234 *Sci U S A* **112**, E5503-5512, doi:10.1073/pnas.1508736112 (2015).

235 12 Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic  
236 investigation of critical experimental parameters. *Nucleic Acids Res* **29**, E65-65  
237 (2001).

238 13 Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations.  
239 *Nat Protoc* **7**, 617-636, doi:10.1038/nprot.2012.012 (2012).

240 14 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of  
241 circulating tumor DNA. *Nat Biotechnol* **34**, 547-555, doi:10.1038/nbt.3520 (2016).

242 15 Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27-36,  
243 doi:10.1093/carcin/bgp220 (2010).

244 16 Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human  
245 transcription factors. *Science* **356**, doi:10.1126/science.aaj2239 (2017).

246 17 Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a  
247 multiple random validation strategy. *Lancet* **365**, 488-492, doi:10.1016/S0140-  
248 6736(05)17866-0 (2005).

249 18 Pedersen, K. S. *et al.* Leukocyte DNA methylation signature differentiates pancreatic  
250 cancer patients from healthy controls. *PLoS One* **6**, e18223,  
251 doi:10.1371/journal.pone.0018223 (2011).

252 19 Teschendorff, A. E. *et al.* An epigenetic signature in peripheral blood predicts active  
253 ovarian cancer. *PLoS One* **4**, e8274, doi:10.1371/journal.pone.0008274 (2009).

254

255

256 **Supplementary information**

257 Supplementary Information is linked to the online version of the paper at  
258 [www.nature.com/nature](http://www.nature.com/nature).

259 **Acknowledgements**

260 This study was conducted with support from the University of Toronto McLaughlin Centre (MC-  
261 2015-02), Canadian Institutes of Health Research (CIHR FDN 148430 and CIHR New  
262 Investigator Salary award 201512MSH-360794-228629), Ontario Institute for Cancer Research  
263 (OICR) with funds from the province of Ontario, Canada Research Chair (950-231346), and  
264 Princess Margaret Cancer Foundation to D.D.C. as well as Canadian Cancer Society (CCSRI  
265 701717) to R.J.H., CCSRI (704716) to R.J.H. and D.D.C. and CCSRI 703827 to M.M.H.  
266 Recruitment of healthy individuals was supported by Cancer Care Ontario Chair of Population  
267 Health and CCSRI 020214 awarded to R.J.H. Collection of lung cancer samples was supported  
268 by the Alan B Brown chair in molecular genomics and the Lusi Wong Lung Cancer Early  
269 Detection Program to GL. We would like to acknowledge the Princess Margaret Genomics  
270 Centre for carrying out the NGS sequencing and the Bioinformatics and HPC Core, Princess  
271 Margaret Cancer Centre for their expertise in generating the NGS data.

272  
273 **Author Contributions**

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

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

#### 284 Competing Interests

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

287 Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence  
288 and requests for materials should be addressed to Daniel D. De Carvalho.

289

290 **Figure Legends**

291 **Fig. 1: cfDNA methylome as a sensitive approach to detect ctDNA in low input DNA. a,**  
292 Simulated probability of detecting at least one epimutation as a function of ctDNA concentration  
293 (0.001% to 10%) (columns), number of DMRs analyzed (1 to 10,000) (rows), and sequencing  
294 depth (10X to 10,000X) (x-axis). **b,** Across a serial dilution series (n=7 dilution points, two  
295 technical replicates, each replicate was used per protocol) of HCT116 DNA spiked into MM.1S  
296 multiple myeloma cells, near-perfect correlations are observed between observed vs expected  
297 methylation signal within DMRs in RPKMs. **c,** Frequency of ctDNA (human) as a percentage of  
298 total cfDNA (human + mice) in the plasma from two colorectal cancer, patient-derived  
299 xenografts (PDX) before and after cfMeDIP-seq.

300

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

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

315

316 **Fig. 3: Methylome analysis of plasma cfDNA allows tumor classification. a,** cfMeDIP-seq

317 carried out on a discovery cohort consisting of 189 samples from 7 different tumor types:

318 pancreatic cancer (PDAC), AML, bladder cancer (BLCA), breast cancer (BRCA), colorectal

319 cancer (CRC), lung cancer (LUC) and renal cancer (RCC), including early and late stage tumors,

320 and healthy controls (normal). For each cancer type, DMRs between the cancer type and normal

321 controls were identified. Overlap is shown between plasma-derived DMRs for each cancer type

322 and primary tumor DMRs (tumor tissue versus adjacent normal tissue) for the corresponding

323 cancer type using TCGA data. Boxplots represent expected null distribution of overlaps from

324 1000 permutations (two sided, p-values computed using standard normal distribution). Extremes

325 of boxes and center-lines define upper and lower quartiles and medians. Whiskers indicate 1.5 x

326 interquartile range (IQR). Diamonds represent observed overlap (red if significantly enriched,

327 green if significantly depleted, and blue if not significant). Horizontal lines indicate thresholds

328 for statistical significance. **b,** Evaluation of classification accuracy on the discovery cohort. The

329 discovery cohort (n=189) was partitioned into 100 independent training and test sets in an 80%-

330 20% manner, consisting of 8 classes (cancer types and healthy controls). Training sets were used

331 for DMR-selection and model training, yielding 100 sets of 8 one-class vs-other-classes binomial

332 GLMnet classifiers. Y-axis depicts distributions of AUROC (area under the receiver operator

333 characteristic curve) for each held-out test set for each class. Dots represent performance in

334 individual test sets. Ends of boxes represent upper and lower-quartiles, line within box

335 represents median, and whiskers represent 1.5 x IQR. **c,** ROC (receiver operating characteristic)



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

342

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

## 349 **Methods**

### 350 **Bioinformatic simulation of tumor-specific features and probability of detection by** 351 **sequencing depth**

352 We created 145,000 simulated genomes with 1, 10, 100, 1,000 and 10,000 independent loci with  
353 0.001 – 10% cancer-specific DMRs in 10-fold increments. 14,500 diploid genomes (expected  
354 copy number in 100 ng cfDNA) were then sampled from these mixtures and further sampled 10  
355 – 10,000x in 10-fold increments at each locus. The process was repeated 100 times for each  
356 combination of parameters. Probability curves were plotted for successful detection of >1 and >5  
357 DMRs (Fig. 1a and Extended Data Fig.1a).

358 **cfMeDIP-seq**

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

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

380 monoclonal antibody 33D3 (Cat#C15200081) from the MagMeDIP kit was diluted 1:15 prior to  
381 generating the diluted antibody mix and added to the sample. Washed magnetic beads (following  
382 manufacturer instructions) were also added prior to incubation at 4°C for 17 hours. The samples  
383 were purified using the Diagenode iPure Kit v2 (Cat# C03010015) and eluted in 50 µl of Buffer  
384 C. The success of the reaction (QC1) was validated through qPCR to detect recovery of the  
385 spiked-in methylated and unmethylated *A. thaliana* DNA. The % recovery of unmethylated  
386 spiked-in DNA should be <1% (relative to input control, adjusted for input control being 10% of  
387 overall sample) and the % specificity of the reaction should be >99% (as calculated by  $(1 -$   
388 [recovery of spiked-in unmethylated control DNA over recovery of spiked-in methylated control  
389 DNA])\*100), prior to proceeding to the next step. The optimal number of cycles to amplify each  
390 library was determined through qPCR, after which the samples were amplified using Kapa HiFi  
391 Hotstart Mastermix and NEBNext multiplex oligos, added to a final concentration of 0.3 µM.  
392 The final libraries were amplified as follows: activation at 95°C for 3 min, followed by  
393 predetermined cycles of 98°C for 20 s, 65°C for 15 s and 72°C for 30 s and a final extension of  
394 72°C for 1 min. The amplified libraries were purified using MinElute columns, then gel size  
395 selected with 3% Nusieve GTG agarose gel to remove any adapter dimers. All the final libraries  
396 were submitted for BioAnalyzer analysis prior to sequencing at the Princess Margaret Genomics  
397 Centre (PMGC) on an Illumina HiSeq 2500, SBS V4 chemistry, single read 50 bp, multiplexed  
398 as 7 samples/lane. After sequencing, the sequenced reads were aligned to  $\lambda$  and hg19 using  
399 Bowtie<sup>20</sup> with the default settings. Based on virtually no alignment to  $\lambda$  genome, the filler DNA  
400 does not interfere with the generation of sequencing data (Supplementary Table 7 and 8).

401 The generated SAM files from hg19 alignment were converted to BAM format, ensuring  
402 removal of duplicate reads, sorting and indexing of reads using SAMtools<sup>21</sup> prior to subsequent

403 analysis with R package MEDIPS<sup>22</sup>. CpG Enrichment Score, as a quality control measure for the  
404 immunoprecipitation reaction, was calculated as part of the MEDIPS package.

#### 405 **Validation of cfMeDIP-seq against MeDIP-seq**

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

#### 417 **Dilution series of sheared cell line DNA**

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

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

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

441 Barcodes were used in combination with endogenous sequence features (genome coordinates,  
442 mapping alignments, read orientation, and read number in pair) to confer sequences from  
443 individual molecules. Consensus sequences were formed from two or more reads supporting the  
444 same molecule with 70% agreement amongst bases above Phred quality scores (Q)<sup>27</sup> of 30.  
445 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<sup>28</sup>.

451 We selected variants based on annotated SNPs from the Cancer Cell Line Encyclopedia  
452 (CCLE)<sup>29</sup> overlapping our target panel. SNVs were called with MuTect<sup>30</sup> 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_file30`. 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<sub>2</sub> 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

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

#### 471 **Donor recruitment and sample acquisition**

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

#### 483 **Specimen processing and methylation analysis of purified tumor and normal cells from** 484 **PDAC samples**

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

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

### 503 **Specimen Processing – patient cell-free DNA (cfDNA)**

504 Plasma samples collected using EDTA and ACD (acid citrate dextrose) tubes, were obtained  
505 from the UHN BioBanks and MSH and were kept frozen until use. cfDNA was extracted from  
506 0.5-3.5 mL of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen) and quantified  
507 through Qubit prior to use. Sex, age and pathology stage are available in Supplementary Table  
508 12, while extracted DNA quantities are available in Extended Data Fig. 8a.

### 509 **Calculation and visualization of differentially methylated regions from cfDNA of** 510 **pancreatic cancer patients and healthy donors**



511 DMRs between cfDNA samples from 24 pancreatic cancer (PDAC) patients and 24 healthy  
512 donors (controls) were calculated using MEDIPS and DESeq2 R packages<sup>22,34</sup>. For each sample,  
513 we computed counts per 300 bp non-overlapping windows, filtered out windows with less than  
514 10 counts across all samples and fit a negative binomial model to call DMRs at FDR < 0.1 (Wald  
515 test). Z-scores of DMR RPKM values with Euclidean Distance and Ward clustering were used  
516 for visualisation.

### 517 **Enrichment analyses for plasma-derived DMRs in tumor-specific methylation signals in** 518 **PDAC**

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

### 531 **Enrichment analyses for cfMeDIP DMRs in TCGA 450K DMCs relative to normal tissues** 532 **and PBMCs.**

533 189 cfDNA samples were obtained across 7 cancer types (AML, bladder (BLCA), breast  
534 (BRCA), colorectal (CRC), lung (LUC), pancreatic (PDAC) and renal cancer (RCC)) and  
535 healthy donors (normal) (Supplementary Table 12). After processing of cfMeDIP-seq data from  
536 these samples, DMRs were calculated using DESeq2 between each cancer type and healthy  
537 donors as described above. DMCs were also calculated between TCGA 450K methylation array  
538 samples from each corresponding cancer type (n = 3979) (obtained from SAGE synapse) and  
539 PBMCs (n = 53, obtained from GEO) samples using *limma* (FDR < 0.01, absolute delta-Beta  
540 0.25). Statistical tests for enrichment were performed as described above for PDAC RRBS  
541 samples. The same procedure was carried out for DMCs calculated between TCGA 450K  
542 methylation array samples from a cancer type and normal samples from the same tissue, for  
543 BLCA, BRCA, CRC, LUC and RCC.

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

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

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

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

## 558 **Machine learning analyses for evaluation of classification accuracy.**

### 559 *Model training and evaluation on the discovery cohort.*

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

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

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

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

578 *Model evaluation on the validation cohort.*

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

586 *Validation in cell lines*

587 450K profiles for 1,028 cell lines previously characterised in Iorio et al (2016)<sup>37</sup> were obtained as  
588 IDAT files. The data were then uniformly processed using the ssNoob method in the *minfi*  
589 package<sup>38</sup>. We reduced this dataset to tissue types for which cfMeDIP data were available  
590 (n=400).

### 591 **Data availability statement**

592 We have deposited R markdowns (either knit or raw) and scripts used to generate the findings in  
593 this study on Zenodo (DOIs in Supplementary Table 13). All the cell line datasets generated  
594 during and/or analysed during the current study are available in the GEO repository under  
595 accession code GSE79838. The cfMeDIP-seq NGS data for patient samples that support the  
596 findings of this study are available upon request from the corresponding author (D.D.C)  
597 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

- 602  
603 20 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient  
604 alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25,  
605 doi:10.1186/gb-2009-10-3-r25 (2009).
- 606 21 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,  
607 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 608 22 Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide  
609 differential coverage analysis of sequencing data derived from DNA enrichment  
610 experiments. *Bioinformatics* **30**, 284-286, doi:10.1093/bioinformatics/btt650  
611 (2014).
- 612 23 Kis, O. *et al.* Circulating tumour DNA sequence analysis as an alternative to multiple  
613 myeloma bone marrow aspirates. *Nat Commun* **8**, 15086,  
614 doi:10.1038/ncomms15086 (2017).
- 615 24 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing.  
616 *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).
- 617 25 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler  
618 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324  
619 (2009).
- 620 26 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-  
621 generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806  
622 (2011).
- 623 27 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II.  
624 Error probabilities. *Genome Res* **8**, 186-194 (1998).
- 625 28 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation  
626 sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513,  
627 doi:10.1073/pnas.1208715109 (2012).
- 628 29 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of  
629 anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).
- 630 30 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and  
631 heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514  
632 (2013).
- 633 31 Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for  
634 genome-scale DNA methylation profiling. *Nat Protoc* **6**, 468-481,  
635 doi:10.1038/nprot.2010.190 (2011).
- 636 32 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for  
637 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572,  
638 doi:10.1093/bioinformatics/btr167 (2011).
- 639 33 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*  
640 *Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 641 34 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and  
642 dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550,  
643 doi:10.1186/s13059-014-0550-8 (2014).
- 644 35 Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical*  
645 *Software* **28**, doi:10.18637/jss.v028.i05 (2008).

646 36 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear  
647 model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-  
648 2014-15-2-r29 (2014).

649 37 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**,  
650 740-754, doi:10.1016/j.cell.2016.06.017 (2016).

651 38 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the  
652 analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369,  
653 doi:10.1093/bioinformatics/btu049 (2014).

654

655

656

657 **Extended Data Figure Legends**

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

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



679 represent biological replicate, with line representing mean **c**, CpG Enrichment Scores of the  
680 sequenced samples (two biological replicates each of four starting DNA inputs (100, 10, 5, and 1  
681 ng), and one input control) show a robust enrichment of CpGs within the genomic regions from  
682 the immunoprecipitated samples compared to the input control. The CpG Enrichment Score was  
683 obtained by dividing the relative frequency of CpGs of the regions by the relative frequency of  
684 CpGs of the human genome. The horizontal dotted line indicates a CpG Enrichment Score of 1,  
685 dots represent biological replicates, with line representing mean. **d**, Genome-wide Pearson  
686 correlations of normalized read counts per 300 bp window between cfMeDIP-seq signal for 1 to  
687 100 ng of input HCT116 DNA sheared to mimic cfDNA (2 biological replicates per  
688 concentration). **e**, Genome Browser snapshot of HCT116 cfMeDIP-seq signal across a window  
689 (chr8:145,095,942-145,116,942) selected out of 4 examined loci, at different starting DNA  
690 inputs (1 to 100 ng, in biological replicates), compared with RRBS (ENCODE ENCSR000DFS)  
691 and WGBS (GEO GSM1465024) data (aligned to hg19). For cfMeDIP-seq, y-axis indicates  
692 RPKMs, for RRBS, yellow and blue blocks represent hypermethylated and hypomethylated  
693 CpGs, respectively. In WGBS track, peak heights indicate methylation level.

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

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

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

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

746 **Extended Data Fig. 6 Circulating cfDNA methylation profiles can identify transcription**  
747 **factor (TF) footprints and infer active transcriptional networks in the tissue of origin. a,**  
748 Expression profile of all TFs (n=42) that were characterized as binding in healthy controls across  
749 53 human tissues from the Genotype-Tissue Expression (GTEx) project. Several TFs  
750 preferentially expressed in the hematopoietic system were identified (PU.1, NFE2, and GATA1).  
751 **b,** Expression profiles (ssGSEA scores) of all TFs with hypomethylated motifs in controls (n=42)  
752 are overexpressed versus those of 1,000 random sets of 42 TFs across GTEx whole blood data  
753 ( $p < 2.2e-16$ , Wilcoxon's Rank Sum Test, two-sided). **c,** Expression profile of all TFs (n=52)  
754 characterized as binding in pancreatic adenocarcinoma patients. Several pancreas-specific or  
755 pancreatic cancer-associated TFs were identified. Moreover, hallmark TFs that drive molecular  
756 subtypes of pancreatic cancer were also identified. **d,** Expression profile (ssGSEA scores) of all  
757 TFs with hypomethylated motifs in cases (n=52) are overexpressed versus those of 1,000 random  
758 sets of 52 TFs in normal pancreas (GTEx data) (Wilcoxon Rank Sum Test, two-sided test,  $p$ -  
759 value  $< 2.2e-16$ ). **e,** Expression profile of all TFs with hypomethylated motifs in cases (n=52) are  
760 overexpressed versus those of 1,000 random sets of 52 TFs in pancreatic adenocarcinoma tissue  
761 (TCGA data) (Wilcoxon Rank Sum Test, two-sided test,  $p$ -value  $< 2.2e-16$ ). For violin plots (**b, d,**  
762 **e**) the ends of the boxes and middle line represent the lower and upper quartiles and median,  
763 respectively. Whiskers represent 1.5 times the interquartile range (IQR). Outliers are excluded.  
764 Rotated kernel densities are also displayed.

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

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

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

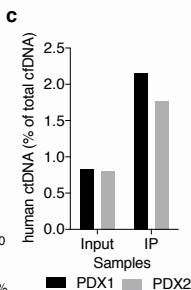
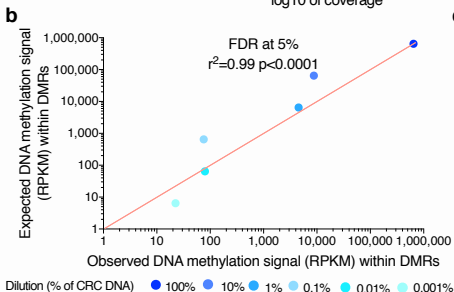
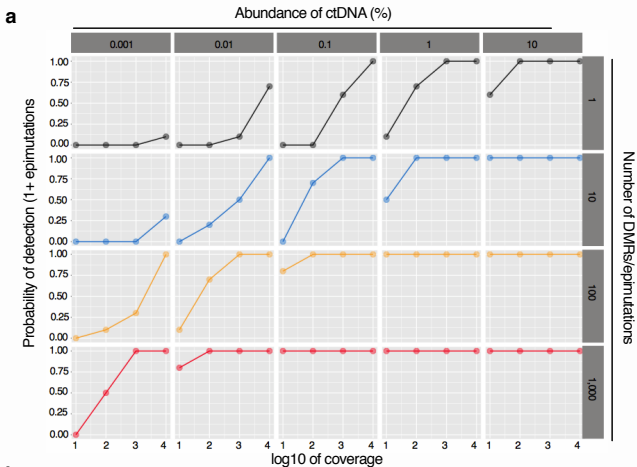
787 **Extended Data Fig. 10 Characterization of hypermethylated regions from cfDNA that are**  
788 **not methylated in leukocytes. a**, Violin plots for the DNA methylation (Beta value) of 38,352  
789 regions in normal blood cells selected based on low DNA methylation levels using IHEC whole

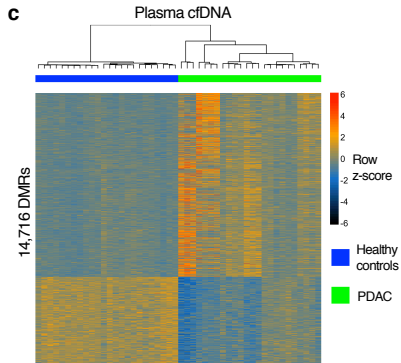
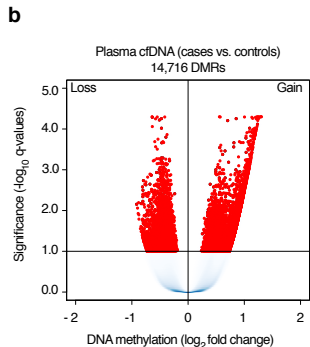
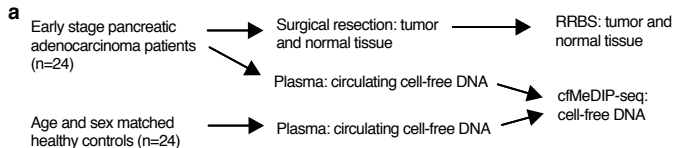
790 genome bisulfite sequencing data. For violin plots, the ends of the boxes and the middle line  
791 represent the lower and upper quartiles, and medians, respectively. Whiskers represent 1.5 times  
792 the interquartile range (IQR). Outliers are excluded. Rotated kernel densities are also displayed.  
793 **b**, Volcano plot representing the regions with low DNA methylation levels in normal blood cells  
794 that overlap with hypermethylated regions in the plasma cfDNA for PDAC (n=3,146 CpG sites)  
795 relative to normal tissue, and RCC (n=2,767 CpG sites), BLCA (n= 3,286 CpG sites), BRCA (n=  
796 6,836 CpG sites), CRC (n= 8,360 CpG sites) and LUC (n= 5,239 CpG sites) relative to PBMCs.  
797 X-axis represents the delta-Beta value in methylation in tumor data from TCGA for cancers other  
798 than PDAC and RRBS for PDAC. Y-axis represents  $-\log_{10}$  q-values (Benjamini Hochberg False  
799 Discovery Rate, BHFDR).

800

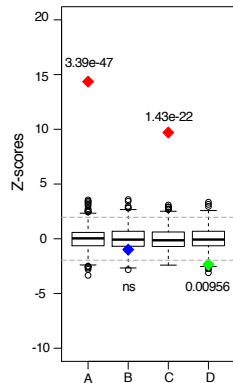
801

802





**d** Overlap plasma cfDNA (cases vs controls) and tumor DNA (primary tumor vs normal tissue)

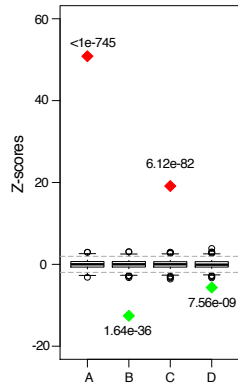


Legend

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)



C = **Hypomethylated** in primary tumor  
**Hypomethylated** in plasma cfDNA

D = **Hypomethylated** in primary tumor  
**Hypermethylated** in plasma cfDNA



