LETTER

https://doi.org/10.1038/s41586-019-1272-6

# Genome-wide cell-free DNA fragmentation in patients with cancer

Stephen Cristiano[1,2,15], Alessandro Leal[1,15], Jillian Phallen[1,15], Jacob Fiksel[1,2,15], Vilmos Adleff[1], Daniel C. Bruhm[1], Sarah Østrup Jensen[3], Jamie E. Medina[1], Carolyn Hruban[1], James R. White[1], Doreen N. Palsgrove[1], Noushin Niknafs[1], Valsamo Anagnostou[1], Patrick Forde[1], Jarushka Naidoo[1], Kristen Marrone[1], Julie Brahmer[1], Brian D. Woodward[4], Hatim Husain[4], Karlijn L. van Rooijen[5], Mai-Britt Worm Ørntoft[3], Anders Husted Madsen[6], Cornelis J. H. van de Velde[7], Marcel Verheij[8], Annemieke Cats[9], Cornelis J. A. Punt[10], Geraldine R. Vink[5], Nicole C. T. van Grieken[11], Miriam Koopman[5], Remond J. A. Fijneman[12], Julia S. Johansen[13], Hans Jørgen Nielsen[14], Gerrit A. Meijer[12], Claus Lindbjerg Andersen[3], Robert B. Scharpf[1,2]* & Victor E. Velculescu[1]*

**Cell-free DNA in the blood provides a non-invasive diagnostic avenue for patients with cancer[1]. However, characteristics of the origins and molecular features of cell-free DNA are poorly understood. Here we developed an approach to evaluate fragmentation patterns of cell-free DNA across the genome, and found that profiles of healthy individuals reflected nucleosomal patterns of white blood cells, whereas patients with cancer had altered fragmentation profiles. We used this method to analyse the fragmentation profiles of 236 patients with breast, colorectal, lung, ovarian, pancreatic, gastric or bile duct cancer and 245 healthy individuals. A machine learning model that incorporated genome-wide fragmentation features had sensitivities of detection ranging from 57% to more than 99% among the seven cancer types at 98% specificity, with an overall area under the curve value of 0.94. Fragmentation profiles could be used to identify the tissue of origin of the cancers to a limited number of sites in 75% of cases. Combining our approach with mutation-based cell-free DNA analyses detected 91% of patients with cancer. The results of these analyses highlight important properties of cell-free DNA and provide a proof-of-principle approach for the screening, early detection and monitoring of human cancer.**

Much of the morbidity and mortality of human cancers worldwide results from late diagnosis when therapeutic intervention is less effective[2,3]. Unfortunately, clinically proven biomarkers that can be used to broadly diagnose and treat patients are not widely available[4]. Recent analyses of circulating cell-free DNA (cfDNA) suggest that approaches using tumour-specific alterations may provide new opportunities for early diagnosis, but not all patients have detectable changes[5–8]. Whole-genome sequencing (WGS) of cfDNA can identify chromosomal abnormalities in patients with cancer but detecting such alterations may be challenging owing to the small number of abnormal chromosomal changes[9–12]. Analyses of the size of fragments of cfDNA have been contradictory, indicating both increases[13–15] and decreases in the overall distribution of cfDNA[12,16,17–19]. Recent studies have suggested that size selection of small cfDNA can increase enrichment of circulating tumour DNA in patients with late-stage cancer[17]. Nucleosome positions[18,20], patterns near transcription start sites[20,21], and the end positions of cfDNA[22] may be altered in cancer, but the sequencing needed to identify nucleosomes is impractical for routine analyses.

Conceptually, the sensitivity of any cfDNA approach depends on the number of alterations examined as well as the technical and biological limitations of detecting such changes. As a typical blood sample contains approximately 2,000 genome equivalents of cfDNA per millilitre of plasma[5], the theoretical limit of detection of a single alteration can be no better than one in a few thousand mutant to wild-type molecules. We hypothesized that the detection of a larger number of alterations in the genome may be more sensitive for detecting cancer in

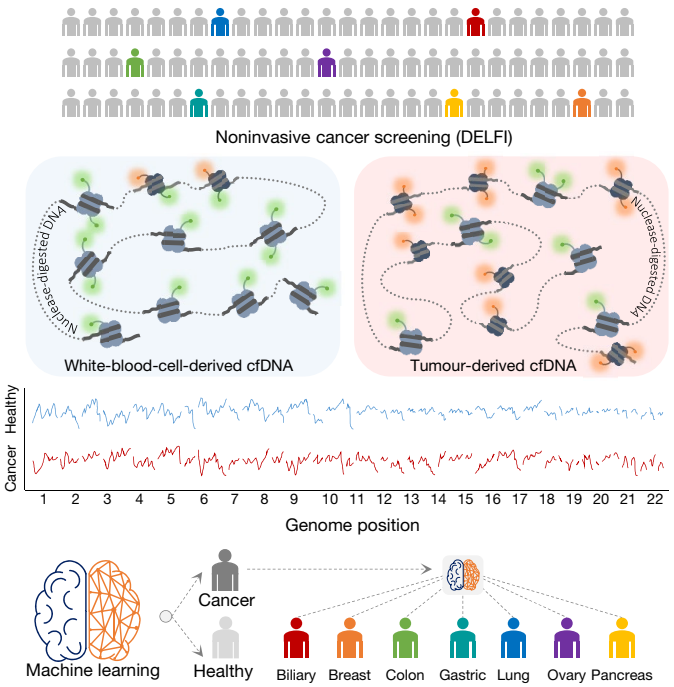



**Fig. 1 | Schematic of DELFI approach.** Blood is collected from healthy individuals and patients with cancer. cfDNA is extracted from plasma, processed into sequencing libraries, examined by WGS, mapped to the genome, and analysed to determine cfDNA fragmentation profiles across the genome. Machine learning is used to categorize whether individuals have cancer and identify the tumour tissue of origin.

[1]The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [4]Division of Hematology and Oncology, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. [5]Department of Medical Oncology, University Medical Center, Utrecht University, Utrecht, The Netherlands. [6]Department of Surgery, Herning Regional Hospital, Herning, Denmark. [7]Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands. [8]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [9]Department of Gastrointestinal Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [10]Department of Medical Oncology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [11]Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands. [12]Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [13]Department of Oncology, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [14]Department of Surgical Gastroenterology 360, Hvidovre Hospital, Hvidovre, Denmark. [15]These authors contributed equally: Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel. *e-mail: rscharpf@jhu.edu; velculescu@jhmi.edu

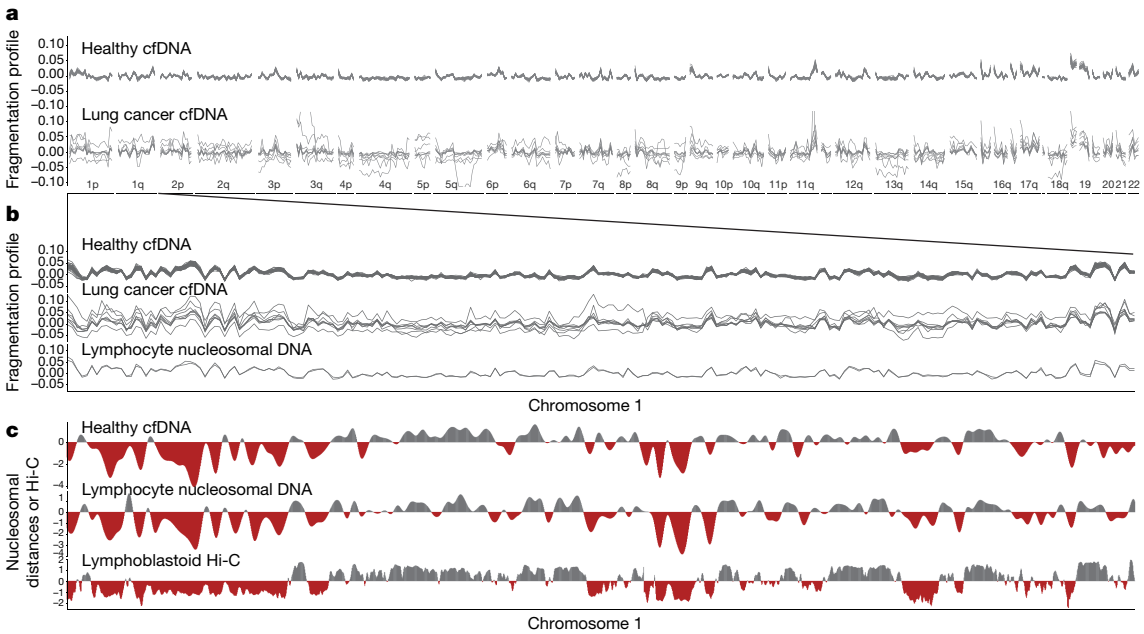



**Fig. 2 | Aberrant cfDNA fragmentation profiles in patients with cancer. a**, Genome-wide cfDNA fragmentation profiles (defined as the ratio of short to long fragments) from approximately 9× WGS are shown in 5-Mb bins for 30 healthy individuals (top) and 8 patients with lung cancer (bottom). **b**, Analyses of healthy cfDNA (top), lung cancer cfDNA (middle), and healthy lymphocyte (bottom) fragmentation profiles from chromosome 1 at 1-Mb resolution. Healthy lymphocyte profiles were scaled with a standard deviation equal to that of the median healthy cfDNA profiles. **c**, Smoothed median distances between adjacent nucleosomes centred at zero using 100 kb bins from healthy cfDNA (top) and nuclease-digested healthy lymphocytes (middle) are depicted together with the first eigenvector for the genome contact matrix from Hi-C analyses of lymphoblastoid cells[27] (bottom).

the circulation. Monte Carlo simulations showed that increasing the number of abnormalities detected from a few to tens or hundreds can improve the limit of detection, similar to recent analyses of methylation changes in cfDNA[23] (Extended Data Fig. 1a).

We developed an approach called 'DNA evaluation of fragments for early interception' (DELFI) (Fig. 1) to detect a large number of abnormalities in cfDNA by genome-wide analysis of fragmentation patterns. The method is based on low-coverage WGS of isolated cfDNA. Mapped sequences are analysed in non-overlapping windows that cover the genome. Conceptually, windows may range in size from thousands to millions of bases, resulting in hundreds to thousands of windows in the genome. We used 5-megabase (Mb) windows to evaluate cfDNA fragmentation patterns as this provided more than 20,000 reads per window at 1–2× genome coverage. Within each window, we examined the coverage and size distribution of cfDNA fragments in healthy and cancer populations (Supplementary Table 1). The genome-wide pattern from an individual can be compared to reference populations to determine whether the pattern is likely to be healthy or cancer-derived. As genome-wide profiles may reveal differences associated with specific tissues, these patterns may also indicate the tissue source of cfDNA.

We focused on fragmentation size of cfDNA as we found that cancer-derived cfDNA may be more variable in length than cfDNA from non-cancer cells. We initially examined cfDNA from targeted regions captured and sequenced at high coverage from patients with breast, colorectal, lung or ovarian cancer[5] (Supplementary Tables 1–3). Analyses of loci containing 165 tumour-specific alterations from 81 patients revealed an average absolute difference of 6.5 base pairs (bp; 95% confidence interval (CI), 5.4–7.6 bp) between the lengths of median mutant and wild-type cfDNA fragments, with mutant cfDNA fragments ranging from 30 bases smaller to 47 bases larger (Extended Data Fig. 1b, Supplementary Table 3). The GC content was similar for mutated and non-mutated fragments, with no correlation between GC content and fragment length (Extended Data Fig. 1c, d). Analyses of 44 germline alterations from 38 patients identified median cfDNA size differences of less than 1 bp between different alleles (Extended Data Fig. 2a, Supplementary Table 3). For 41 alterations related to clonal haematopoiesis[5], there were no significant differences between cfDNA fragments containing such alterations and wild-type fragments (Extended Data Fig. 2b, Supplementary Table 3). Overall, the lengths of cancer-derived cfDNA fragments were more variable than non-cancer cfDNA ($P < 0.001$, variance ratio test). We hypothesized that these differences may reflect changes in chromatin structure as well as other genomic and epigenomic abnormalities in cancer[24,25], and that cfDNA fragmentation in a position-specific manner could serve as a biomarker for cancer detection.

As targeted sequencing analyses a limited number of loci, we investigated whether genome-wide analyses would detect additional abnormalities from cfDNA fragmentation. In a pilot analysis, we isolated cfDNA from around 4 ml of plasma from 8 patients with stage I–III lung cancer and 30 healthy individuals (Supplementary Tables 1, 4, 5), and performed WGS at approximately 9× coverage (Supplementary Table 4). As expected[12,18,19], the median overall lengths of fragments of cfDNA from healthy individuals were larger than those from patients with cancer (167.3 bp and 163.8, respectively, $P < 0.01$, Welch's $t$-test) (Supplementary Table 5). To examine differences in fragment size and coverage in a position-dependent manner across the genome, we mapped fragments to their genomic origin and evaluated fragment lengths in 504 windows of 5 Mb, covering approximately 2.6 Gb of the genome. For each window, we determined the fraction of small cfDNA fragments (100–150 bp) to larger cfDNA fragments (151–220 bp) and overall coverage to obtain genome-wide fragmentation profiles for each sample.

We found that healthy individuals had similar genome-wide fragmentation profiles (Fig. 2a, b, Extended Data Fig. 3a). To examine the origins of cfDNA fragmentation patterns, we isolated and nuclease-treated nuclei from lymphocytes of two healthy individuals to obtain nucleosomal DNA fragments. Healthy cfDNA patterns were highly correlated to lymphocyte nucleosomal DNA fragmentation profiles and nucleosome distances (Fig. 2b, c, Extended Data Fig. 3b, c). Median distances between nucleosomes in lymphocytes were correlated to high-throughput sequencing chromosome conformation capture (Hi-C) open (A) and closed (B) compartments of lymphoblastoid cells[26,27]

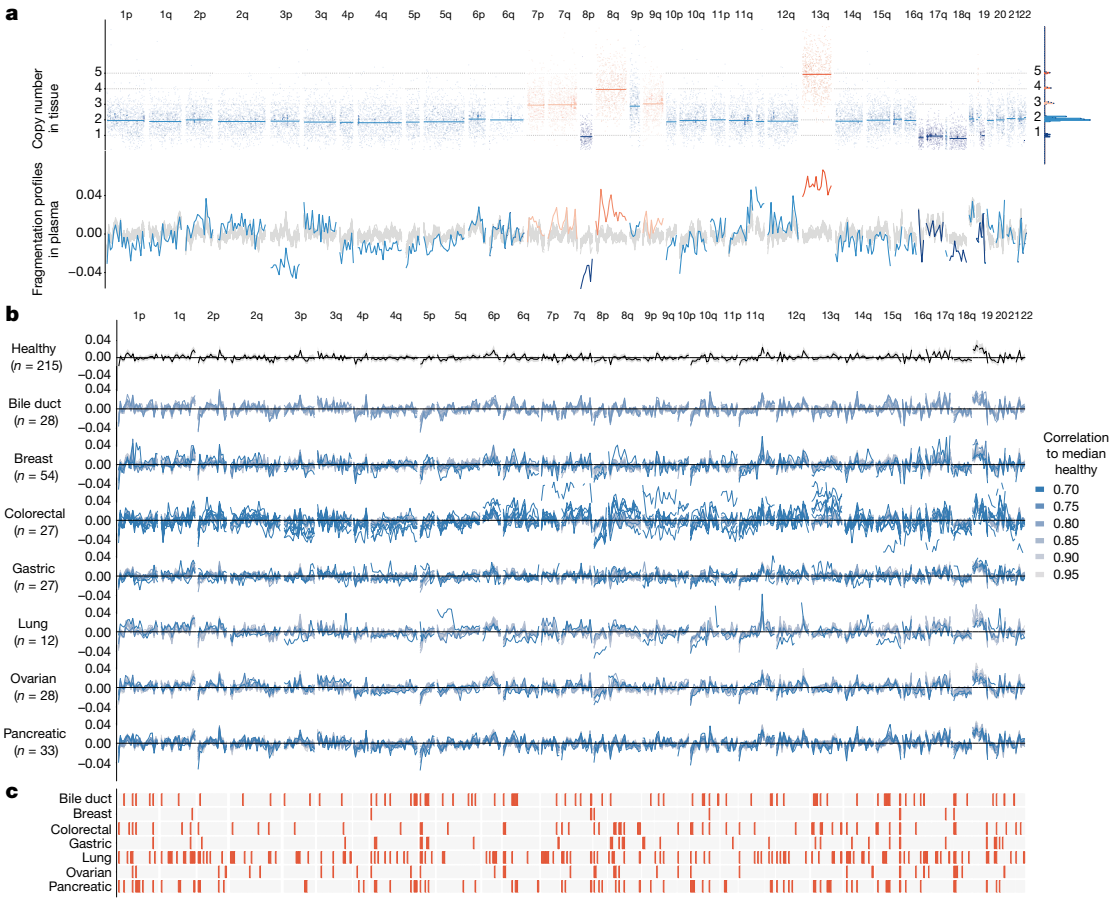



**Fig. 3 | cfDNA fragmentation profiles in healthy individuals and patients with cancer. a**, Fragmentation profiles (bottom) in the context of tumour copy number changes (top) in a patient with colorectal cancer. The distribution of segment means and integer copy numbers are shown at top right. **b**, GC-adjusted fragmentation profiles from WGS at 1–2× coverage for healthy individuals and patients with cancer are depicted per cancer type using 5-Mb windows. The median healthy profile is indicated in black and the 98% confidence band is shown in grey. For patients with cancer, individual profiles are coloured based on their Pearson correlation to the healthy median. **c**, Windows are indicated in orange if more than 10% of the cancer samples had a fragment ratio more than three standard deviations from the median healthy fragment ratio.

(Fig. 2c). These analyses suggest that fragmentation patterns of normal cfDNA are the result of nucleosomal DNA patterns that reflect the chromatin structure of normal blood cells.

In contrast to healthy cfDNA, patients with cancer had several distinct genomic differences with increases and decreases in fragment sizes at different regions (Fig. 2a, b). We performed genome-wide correlation analyses of the fraction of short to long cfDNA fragments for each sample compared to the median fragment length profile of healthy individuals, and found that—although cfDNA fragment profiles were consistent among healthy individuals (median correlation of 0.99)—the median correlation of fragment ratios among patients with cancer was 0.84 ($P < 0.001$, Wilcoxon rank-sum test; Fig. 2a, b, Extended Data Fig. 3d, Supplementary Table 5). Similar differences were observed when comparing cfDNA fragmentation profiles of patients with cancer to fragmentation profiles of healthy lymphocytes (Fig. 2c, Extended Data Fig. 3b, c). To account for potential biases attributable to GC content, we applied a locally weighted smoother and found that differences in fragmentation profiles between healthy individuals and patients with cancer remained after this adjustment (median correlation of patients with cancer to healthy = 0.83, Supplementary Table 5).

We subsampled WGS data at 9× coverage to approximately 2×, 1×, 0.5×, 0.2× and 0.1× genome coverage, and determined that altered fragmentation profiles from patients with cancer were identified even at 0.5× coverage (Extended Data Fig. 3e, f). On the basis of these observations, we performed WGS at 1–2× coverage to evaluate whether fragmentation profiles may change during the course of therapy[28,29]. We evaluated cfDNA from 19 patients with non-small-cell lung cancer during therapy with anti-EGFR or anti-ERBB2 agents (Supplementary Table 6). The degree of abnormality in the fragmentation profiles during therapy closely matched levels of EGFR or ERBB2 mutant allele fractions[29] (Extended Data Fig. 4, Spearman correlation of mutant allele fractions to fragmentation profiles = 0.74). These results demonstrate that fragmentation analyses may be useful for detecting tumour-derived cfDNA and monitoring patients during treatment.

As cfDNA fragmentation profiles would be expected to reflect both epigenomic and genomic alterations, we examined these in a patient with known tumour copy number changes. Altered fragmentation profiles were present in regions of the genome that were copy-neutral and were further affected in regions with copy number changes (Fig. 3a, Extended Data Fig. 5a). Position-dependent differences in fragmentation patterns distinguished cancer-derived cfDNA from healthy cfDNA, whereas analyses of overall fragment sizes of cfDNA would have missed such differences (Extended Data Fig. 5a, b).

We performed WGS at 1–2× coverage of cfDNA from 208 patients with cancer, including breast ($n = 54$), colorectal ($n = 27$), lung ($n = 12$), ovarian ($n = 28$), pancreatic ($n = 34$), gastric ($n = 27$) or bile duct cancer ($n = 26$), as well as 215 healthy individuals (Supplementary Tables 1, 4). All patients with cancer had not undergone previous treatment and most had resectable disease ($n = 183$). After GC adjustment of short and long cfDNA fragment coverage (Extended Data Fig. 6a, b), we examined coverage and size characteristics of fragments in windows throughout the genome (Fig. 3b, Supplementary Tables 4, 7). Healthy individuals had concordant fragmentation profiles whereas patients with cancer had highly variable profiles with decreased correlation to the median healthy profile (Supplementary Table 7). An

**Table 1 | DELFI performance for cancer detection**

| Individuals analysed | | | 95% specificity | | | 98% specificity | | |
|---|---|---|---|---|---|---|---|---|
| | | | Individuals detected | Sensitivity (%) | 95% CI (%) | Individuals detected | Sensitivity (%) | 95% CI (%) |
| Healthy | | 215 | 10 | – | – | 4 | – | – |
| Cancer | | 208 | 166 | 80 | 74–85 | 152 | 73 | 67–79 |
| Type | Breast | 54 | 38 | 70 | 56–82 | 31 | 57 | 43–71 |
| | Bile duct | 26 | 23 | 88 | 70–98 | 21 | 81 | 61–93 |
| | Colorectal | 27 | 22 | 81 | 62–94 | 19 | 70 | 50–86 |
| | Gastric | 27 | 22 | 81 | 62–94 | 22 | 81 | 62–94 |
| | Lung | 12 | 12 | 100 | 74–100 | 12 | 100 | 74–100 |
| | Ovarian | 28 | 25 | 89 | 72–98 | 25 | 89 | 72–98 |
| | Pancreatic | 34 | 24 | 71 | 53–85 | 22 | 65 | 46–80 |
| Stage | I | 41 | 30 | 73 | 53–86 | 28 | 68 | 52–82 |
| | II | 109 | 85 | 78 | 69–85 | 78 | 72 | 62–80 |
| | III | 33 | 30 | 91 | 76–98 | 26 | 79 | 61–91 |
| | IV | 22 | 18 | 82 | 60–95 | 17 | 77 | 55–92 |
| | X | 3 | 3 | 100 | 29–100 | 3 | 100 | 29–100 |

analysis of commonly altered genomic windows revealed a median of 60 affected windows across the cancer types analysed, which highlights position-dependent alterations in fragmentation of cfDNA (Fig. 3c).

We implemented a gradient tree boosting machine learning model to examine whether cfDNA has characteristics of a patient with cancer or healthy individual, and estimated performance characteristics of this approach by tenfold cross-validation repeated ten times (Extended Data Fig. 7a, b). The machine learning model included GC-adjusted short and long fragment coverage characteristics in windows throughout the genome. We also developed a machine learning classifier for copy number changes from chromosomal arm features[10,11] (Extended Data Fig. 8a, Supplementary Table 8) and included mitochondrial copy number changes[12] (Extended Data Fig. 8b). Using this implementation of DELFI, we obtained a score that could be used to classify patients as being healthy or having cancer. We detected 152 out of 208 cancer patients (73% sensitivity, 95% confidence interval 67–79%), and misclassified 4 out of 215 healthy individuals (98% specificity) (Table 1). At a threshold of 95% specificity, we detected 80% of patients with cancer (95% confidence interval 74–85%), including 79% of patients with resectable (stage I–III) disease (145 out of 183) and 82% of patients with stage IV disease (18 of 22) (Table 1). Receiver operator characteristic analyses for the detection of patients with cancer had an area under the curve (AUC) value of 0.94 (95% confidence interval 0.92–0.96), ranging from 0.86 for pancreatic cancer to at least 0.93 for breast, bile duct, colorectal, gastric, lung and ovarian cancers (Fig. 4, Extended Data Fig. 9a), with AUC values of at least 0.92 for each stage (Extended Data Fig. 9b). To assess the contribution of fragment size and coverage across the genome, chromosome arm copy number or mitochondrial copy number to the predictive accuracy of the model, we implemented the cross-validation procedure to assess performance characteristics of these features in isolation. Fragment coverage features alone (AUC = 0.94) were nearly identical to the classifier that combined all features (AUC = 0.94). By contrast, machine learning analyses of changes in chromosomal copy number had lower performance (AUC = 0.88) but were still more predictive than copy number using individual scores (AUC = 0.78) or mitochondrial copy number (AUC = 0.72) (Fig. 4). These results suggest that fragment coverage is the major contributor to our classifier, but we have included all features in our prediction model as they can be obtained from the same WGS data and may contribute in a complementary fashion for cancer detection.

As fragmentation profiles reveal regional differences between tissues, we used machine learning to identify the tissue of origin of circulating tumour DNA. These analyses had a 61% accuracy (95% confidence interval 53–67%) that increased to 75% (95% confidence interval 69–81%) when assigning circulating tumour DNA to one of two sites of origin (Extended Data Fig. 9c, d). For all tumour types, the classification of tissue of origin by DELFI was higher than that by random assignment ($P < 0.01$, binomial test, Extended Data Fig. 9d).

We evaluated whether combining DELFI with mutation detection in cfDNA[5] could increase the sensitivity of cancer detection (Extended Data Fig. 10). An evaluation of cases analysed using both approaches revealed that 82% (103 out of 126) of patients were detected using DELFI, and 66% (83 out of 126) had sequence alterations. For cases with mutant allele fractions of less than 1%, DELFI detected 80% of cases—including those that were undetectable using targeted sequencing (Supplementary Table 7). When these approaches were used together, the combined sensitivity increased to 91% (115 out of 126 patients) with a specificity of 98% (Extended Data Fig. 10).

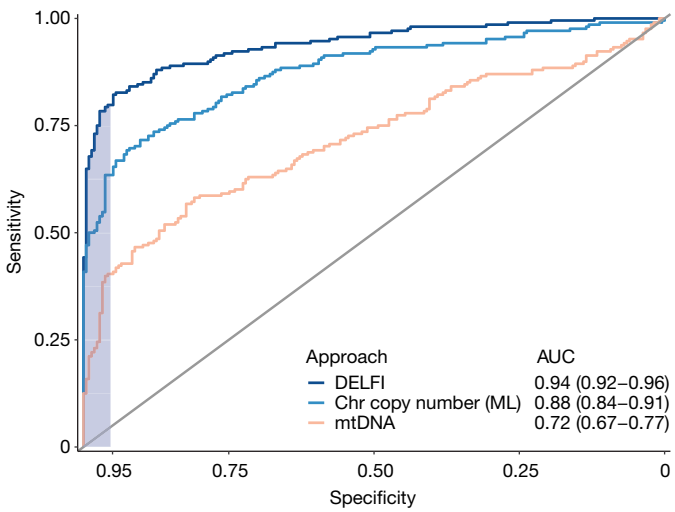



**Fig. 4 | Detection of cancer using DELFI.** Receiver operator characteristics for the detection of cancer using cfDNA fragmentation profiles and other genome-wide features in a machine learning approach are depicted for a cohort of 215 healthy individuals and 208 patients with cancer (DELFI, AUC = 0.94), with ≥95% specificity shaded in blue. Machine learning analyses of chromosomal arm copy number (Chr copy number (ML)), and mitochondrial genome copy number analyses (mtDNA) are shown.

Overall, we have determined that genome-wide fragmentation profiles of cfDNA are different between patients with cancer and healthy individuals. In patients with cancer, fragmentation patterns in cfDNA appear to result from mixtures of nucleosomal DNA from both blood and neoplastic cells. Our approach could be further improved through recovery of smaller fragments[17,30], evaluation of single-stranded libraries[18,30,31] or use of alternative technologies. Additionally, PCR-free libraries could reduce GC bias and sequencing artefacts[18,30,31].

These observations have important implications for non-invasive detection of human cancer. DELFI simultaneously analyses tens to hundreds of tumour-specific abnormalities from minute amounts of cfDNA, overcoming a limitation that has precluded the possibility of more-sensitive analyses of cfDNA. These analyses detected a higher fraction of patients with cancer than previous methods[5–7,12,17], and combining DELFI with the detection of sequence alterations in cfDNA further increased the sensitivity of detection. As fragmentation profiles seem to be related to nucleosomal patterns, DELFI may be useful for determining the source of tumour-derived cfDNA, an aspect that could be further improved using clinical characteristics, methylation changes[23] and other diagnostic approaches[6]. DELFI requires only a small amount of whole-genome sequencing, which suggests that this approach could be broadly applied for the screening and management of patients with cancer.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-019-1272-6.

Received: 19 November 2018; Accepted: 10 May 2019;
Published online: 29 May 2019

1. Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
2. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
3. World Health Organization. *Guide to Cancer Early Diagnosis* https://www.who.int/cancer/publications/cancer_early_diagnosis/en/ (WHO, 2017).
4. National Comprehensive Cancer Network. *NCCN Clinical Practice Guidelines in Oncology* https://www.nccn.org/professionals/physician_gls/default.aspx (accessed 16 April 2019).
5. Phallen, J. et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (2017).
6. Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
7. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
8. Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
9. Leary, R. J. et al. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
10. Leary, R. J. et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4**, 162ra154 (2012).
11. 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).
12. Jiang, P. et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl Acad. Sci. USA* **112**, E1317–E1325 (2015).
13. Wang, B. G. et al. Increased plasma DNA integrity in cancer patients. *Cancer Res.* **63**, 3966–3968 (2003).
14. Umetani, N. et al. Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J. Clin. Oncol.* **24**, 4270–4276 (2006).
15. Chan, K. C., Leung, S. F., Yeung, S. W., Chan, A. T. & Lo, Y. M. Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **14**, 4141–4145 (2008).
16. Mouliere, F. et al. High fragmentation characterizes tumour-derived circulating DNA. *PLoS ONE* **6**, e23418 (2011).
17. Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
18. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
19. Underhill, H. R. et al. Fragment length of circulating tumor DNA. *PLoS Genet.* **12**, e1006162 (2016).
20. Ulz, P. et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* **48**, 1273–1278 (2016).
21. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16** (Suppl. 13), S1 (2015).
22. Jiang, P. et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl Acad. Sci. USA* **115**, E10925–E10933 (2018).
23. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
24. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
25. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
26. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
27. Fortin, J. P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
28. Diehl, F. et al. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).
29. Phallen, J. et al. Early noninvasive detection of response to targeted therapy in non-small cell lung cancer. *Cancer Res.* **79**, 1204–1213 (2019).
30. Burnham, P. et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).
31. Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J. & Thierry, A. R. New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. *NPJ Genom. Med.* 3, 31 (2018).

**Acknowledgements** We thank members of our laboratories for critical review of the manuscript. This work was supported, in part, by the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up to Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Commonwealth Foundation, the Cigarette Restitution Fund, the Burroughs Wellcome Fund and the Maryland Genetics, Epidemiology and Medicine Training Program, the AACR-Janssen Cancer Interception Research Fellowship, the Mark Foundation for Cancer Research, US NIH (grants CA121113, CA006973, and CA180950), the Danish Council for Independent Research (11-105240), the Danish Council for Strategic Research (1309-00006B), the Novo Nordisk Foundation (NNF14OC0012747 and NNF17OC0025052), and the Danish Cancer Society (R133-A8520-00-S41 and R146-A9466-16-S2). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research.

**Reviewer information** *Nature* thanks Daniel De Carvalho, Ellen Heitzer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** S.C., A.L., J.P., J.F., V. Adleff, R.B.S. and V.E.V. designed and planned the study, and developed and optimized experimental protocols. A.L., J.P., V. Adleff, J.E.M. and D.N.P. performed experiments. S.Ø.J., V. Anagnostou, P.F., J.N., K.M., J.B., B.D.W., H.H., K.L.v.R., M.-B.W.Ø., A.H.M., C.J.H.v.d.V., M.V., A.C., C.J.A.P., G.R.V., N.C.T.v.G., M.K., R.J.A.F., J.S.J., H.J.N., G.A.M. and C.L.A. organized patient enrolment, sample collection, and clinical data curation. S.C., A.L., J.P., J.F., V. Adleff, D.C.B., J.E.M., J.R.W., N.N., G.A.M., C.L.A., R.B.S. and V.E.V. analysed and interpreted data. S.C., A.L., J.P., J.F., R.B.S. and V.E.V. wrote the manuscript and incorporated feedback from all authors. S.C., A.L., J.P. and J.F. contributed equally to this study.

**Competing interests** S.C., A.L., J.P., J.F., V. Adleff, R.B.S. and V.E.V. are inventors on patent applications (62/673,516 and 62/795,900) submitted by Johns Hopkins University related to cell-free DNA for cancer detection. V.E.V. is a founder of Delfi Diagnostics and Personal Genome Diagnostics, a member of their Scientific Advisory Boards and Boards of Directors, and owns Delfi Diagnostics and Personal Genome Diagnostics stock, which are subject to certain restrictions under university policy. Within the last five years, V.E.V. has been an advisor to Daiichi Sankyo, Janssen Diagnostics, Ignyta, and Takeda Pharmaceuticals. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies.

**Additional information**
**Extended data** is available for this paper at https://doi.org/10.1038/s41586-019-1272-6.
**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-019-1272-6.
**Reprints and permissions information** is available at http://www.nature.com/reprints.
**Correspondence and requests for materials** should be addressed to R.B.S. or V.E.V.
**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Patient and sample characteristics.** Plasma samples from healthy individuals and plasma and tissue samples from patients with breast, lung, ovarian, colorectal, bile duct or gastric cancer were obtained from ILSBio/Bioreclamation, Aarhus University, Herlev Hospital of the University of Copenhagen, Hvidovre Hospital, the University Medical Center of the University of Utrecht, the Academic Medical Center of the University of Amsterdam, the Netherlands Cancer Institute and the University of California, San Diego. All samples were obtained under Institutional Review Board approved protocols with informed consent from all participants for research use at participating institutions. Plasma samples from healthy individuals were obtained at the time of routine screening, including for colonoscopies or Pap smears. Individuals were considered healthy if they had no previous history of cancer and negative screening results.

Plasma samples from individuals with breast, colorectal, gastric, lung, ovarian, pancreatic and bile duct cancer were obtained at the time of diagnosis, before tumour resection or therapy. Nineteen patients with lung cancer analysed for changes in cfDNA fragmentation profiles across several time points were undergoing treatment with anti-EGFR or anti-ERBB2 therapy[29]. Clinical data for all patients included in this study are listed in Supplementary Table 1. Sex was confirmed by genomic analyses of X and Y chromosome representation. Pathological staging of patients with gastric cancer was performed after neoadjuvant therapy. Samples for which the tumour stage was unknown were indicated as stage X.

**Nucleosomal DNA purification.** Viably frozen lymphocytes were elutriated from leukocytes obtained from a healthy male (C0618) and female (D0808-L) (Advanced Biotechnologies). Aliquots of $1 \times 10^6$ cells were used for nucleosomal DNA purification using EZ Nucleosomal DNA Prep Kit (Zymo Research). Cells were initially treated with 100 μl of Nuclei Prep Buffer and incubated on ice for 5 min. After centrifugation at 200*g* for 5 min, supernatant was discarded and pelleted nuclei were treated twice with 100 μl of Atlantis Digestion Buffer. Finally, cellular nucleic DNA was fragmented with 0.5 U of Atlantis dsDNase at 42 °C for 20 min. Reactions were stopped using 5× Stop Buffer and DNA was purified using Zymo-Spin IIC Columns. Concentration and quality of eluted cellular nucleic DNA were analysed using the Bioanalyzer 2100 (Agilent Technologies).

**Sample preparation and sequencing of cfDNA.** Whole blood was collected in EDTA tubes and processed immediately or within one day after storage at 4 °C, or was collected in Streck tubes and processed within two days of collection for three patients with cancer who were part of the monitoring analysis. Plasma and cellular components were separated by centrifugation at 800*g* for 10 min at 4 °C. Plasma was centrifuged a second time at 18,000*g* at room temperature to remove any remaining cellular debris and stored at −80 °C until the time of DNA extraction. DNA was isolated from plasma using the Qiagen Circulating Nucleic Acids Kit (Qiagen GmbH) and eluted in LoBind tubes (Eppendorf AG). Concentration and quality of cfDNA were assessed using the Bioanalyzer 2100 (Agilent Technologies).

Next-generation sequencing (NGS) cfDNA libraries were prepared for WGS and targeted sequencing using 5–250 ng of cfDNA as previously described[5]. In brief, genomic libraries were prepared using the NEBNext DNA Library Prep Kit for Illumina (New England Biolabs (NEB)) with four main modifications to the manufacturer's guidelines: (i) the library purification steps used the on-bead AMPure XP approach to minimize sample loss during elution and tube transfer steps[32]; (ii) NEBNext End Repair, A-tailing and adaptor ligation enzyme and buffer volumes were adjusted as appropriate to accommodate the on-bead AMPure XP purification strategy; (iii) a pool of eight unique Illumina dual index adaptors with 8-bp barcodes was used in the ligation reaction; and (iv) cfDNA libraries were amplified with Phusion Hot Start Polymerase. Whole-genome libraries were sequenced using 100-bp paired-end runs on the Illumina HiSeq 2000/2500 (Illumina).

**Analyses of targeted sequencing data from cfDNA.** Analyses of targeted NGS data for cfDNA samples were performed as previously described[5]. In brief, primary processing was completed using Illumina Consensus Assessment of Sequence and Variation (CASAVA) software (v.1.8), including demultiplexing and masking of dual-index adaptor sequences. Sequence reads were aligned against the human reference genome (version hg18 or hg19) using NovoAlign with additional realignment of select regions using the Needleman–Wunsch method[33]. The positions of sequence alterations we identified have not been affected by the different genome builds. Candidate mutations, consisting of point mutations, small insertions and deletions, were identified using VariantDx[33] (Personal Genome Diagnostics) across the targeted regions of interest.

To analyse the fragment lengths of cfDNA molecules, we required that each read pair from a cfDNA molecule had a Phred quality score ≥30. We removed all duplicate DNA fragments, defined as having the same start, end and index barcode. For each mutation, we included only fragments for which one or both of the read pairs contained the mutated (or wild-type) base at the given position. This analysis was done using the R packages Rsamtools and GenomicAlignments.

For each genomic locus in which a somatic mutation was identified, we compared the lengths of fragments containing the mutant allele to the lengths of fragments with the wild-type allele. If more than 100 mutant fragments were identified, we used Welch's two-sample *t*-test to compare the mean fragment lengths. For loci with fewer than 100 mutant fragments, we implemented a bootstrap procedure. Specifically, we sampled with replacement *N* fragments containing the wild-type allele, in which *N* denotes the number of fragments with the mutation. For each bootstrap replicate of wild-type fragments, we computed their median length. The *P* value was estimated as the fraction of bootstrap replicates with a median wild-type fragment length as long as, or more extreme than, the observed median mutant fragment length.

**Analyses of WGS data from cfDNA.** Primary processing of whole-genome NGS data for cfDNA samples was performed using Illumina CASAVA (Consensus Assessment of Sequence and Variation) software (v.1.8.2), including demultiplexing and masking of dual-index adaptor sequences. Sequence reads were aligned against the human reference genome (version hg19) using ELAND.

Read pairs with a MAPQ score below 30 for either read and PCR duplicates were removed. We tiled the hg19 autosomes into 26,236 adjacent, non-overlapping 100-kb bins. We excluded regions of low mappability based on previous work[27] in which 10% of bins with the lowest coverage were removed, and excluded reads that fell in the Duke blacklisted regions (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/). Using this approach, we excluded 361 Mb (13%) of the hg19 reference genome, including centromeric and telomeric regions. Short fragments were defined as having lengths between 100 and 150 bp and long fragments as having lengths between 151 and 220 bp.

To account for biases in coverage attributable to GC content of the genome, we applied locally weighted scatterplot smoothing (LOWESS, also known as LOESS) regression analysis with a span setting of 0.75 to the scatterplot of average fragment GC versus coverage calculated for each 100-kb bin. This LOESS regression was performed separately for short and long fragments to account for possible differences in GC effects on coverage in plasma by fragment length—an approach loosely motivated by a previous study[34]. We subtracted the predictions for short and long coverage explained by GC from the LOESS model, obtaining residuals for short and long that were uncorrelated with GC. We returned the residuals to the original scale by adding back the genome-wide median short and long estimates of coverage. This procedure was repeated for each sample to account for possible differences in GC effects on coverage between samples. To reduce the feature space and noise further, we calculated the total GC-adjusted coverage in 5-Mb bins.

To compare the variability of fragment lengths from healthy subjects to fragments in patients with cancer, we calculated the standard deviation of the short to long fragmentation profiles for each individual. We compared the median of the standard deviations in the two groups by a Wilcoxon rank-sum test.

**Analyses of changes in chromosome-arm copy number.** To develop arm-level statistics for copy number changes, we adapted a previously described approach for aneuploidy detection in plasma, which used both chromosome-arm-specific *Z*-scores as well as plasma aneuploidy (PA) scores to summarize arm-level data[10]. This adapted approach divides the genome into non-overlapping 50-kb bins for which GC-corrected $\log_2$-transformed read depth was obtained after correction by LOESS with span setting of 0.75. This LOESS-based correction is comparable to the approach outlined above, but is evaluated on a $\log_2$ scale to increase robustness to outliers in the smaller bins and does not stratify by fragment length. To obtain an arm-specific *Z*-score for changes in copy number, the mean GC-adjusted read depth for each arm was centred and scaled by the mean and standard deviation, respectively, of read depth scores obtained from an independent set of 50 healthy samples.

**Analyses of mitochondrial-aligned reads from cfDNA.** Whole-genome sequence reads that initially mapped to the mitochondrial genome were extracted from .bam files and realigned to the hg19 reference genome in end-to-end mode with Bowtie2 as previously described[35]. The resulting aligned reads were filtered such that both mates aligned to the mitochondrial genome with MAPQ ≥ 30. The number of fragments mapping to the mitochondrial genome was counted and converted to a percentage of the total number of fragments in the original .bam files.

**Prediction model for cancer detection.** To distinguish healthy individuals from patients with cancer using fragmentation profiles, we used a stochastic gradient boosting model (gbm)[36,37]. GC-corrected total and short fragment coverage for all 504 bins were centred and scaled for each sample to have mean zero and unit standard deviation. Additional features included *Z*-scores for each of the 39 autosomal arms and mitochondrial representation ($\log_{10}$-transformed proportion of reads mapped to the mitochondria). To estimate the prediction error of this approach, we used tenfold cross-validation[38]. Feature selection, performed only on the training data in each cross-validation run, removed bins that were highly correlated (correlation > 0.9) or had near-zero variance. Stochastic gradient boosted machine learning was implemented using the R package gbm with parameters: n.trees = 150, interaction.depth = 3, shrinkage = 0.1, and n.minobsinside = 10. To average over the prediction error from the randomization of individuals to folds, we repeated the tenfold cross-validation procedure ten times. Confidence intervals

for sensitivity were obtained from 2,000 bootstrap replicates with specificity fixed at 98% and 95%.

**Prediction model for tumour tissue of origin classification.** For samples correctly identified from patients with cancer at 90% specificity ($n = 174$), a separate stochastic gradient boosting model was trained to classify the tissue of origin. To account for the small number of lung samples used for prediction, we included 18 cfDNA baseline samples from patients with late-stage lung cancer from the monitoring analyses of our study. Performance characteristics of the model were evaluated using tenfold cross-validation repeated ten times. This gbm model was trained using the same features as in the cancer classification model. Features that displayed correlation above 0.9 to each other or had near zero variance were removed within each training dataset during cross-validation. The tissue class probabilities were averaged across the ten replicates for each patient and the class with the highest probability was used as the predicted tissue.

**Analyses of nucleosomal DNA from human lymphocytes and cfDNA.** From the nuclease-treated lymphocytes, fragment sizes were analysed in 5-Mb bins as described for whole-genome cfDNA analyses. A genome-wide map of nucleosome positions was constructed from the nuclease-treated lymphocyte cell lines. This approach identified local biases in the coverage of circulating fragments, indicating a region protected from degradation. A 'window positioning score' (WPS) was used to score each base pair in the genome[18]. Using a sliding window of 60-bp centred around each base, the WPS was calculated as the number of fragments completely spanning the window minus the number of fragments with only one end in the window. Because fragments arising from nucleosomes have a median length of 167 bp, a high WPS indicated a possible nucleosomic position. WPS values were centred at zero using a running median and smoothed using a Kolmogorov–Zurbenko filter[39]. For spans of positive WPS between 50 and 450 bp, a nucleosome peak was defined as the set of base pairs with a WPS above the median in that window. The calculation of nucleosome positions for cfDNA from 30 healthy individuals with sequence coverage of 9× was determined in the same manner as for lymphocyte DNA. To ensure that nucleosomes in healthy cfDNA were representative, we defined a consensus track of nucleosomes consisting only of nucleosomes identified in two or more individuals. Median distances between adjacent nucleosomes were calculated from the consensus track.

**Monte Carlo simulation of detection sensitivity.** We used Monte Carlo simulation to estimate the probability of detecting a molecule with a tumour-derived alteration. In brief, we generated one million molecules from a multinomial distribution. For a simulation with $m$ alterations, wild-type molecules were simulated with probability $p$ and each of the $m$ tumour alterations were simulated with probability $(1 - p)/m$. Next, we sampled $g \times m$ molecules randomly with replacement, in which $g$ denotes the number of genome equivalents in 1 ml of plasma. If a tumour alteration was sampled ($s$) or more times, we classified the sample as cancer-derived. We repeated the simulation 1,000 times, estimating the probability that the in silico sample would be correctly classified as cancer by the mean of the cancer indicator. Setting $g = 2{,}000$ and $s = 5$, we varied the number of tumour alterations by powers of 2 from 1 to 256 and the fraction of tumour-derived molecules from 0.0001% to 1%.

**Statistical analyses.** All statistical analyses were performed using R version 3.4.3. The R packages caret (v.6.0-79) and gbm (v.2.1-4) were used to implement the classification of healthy versus cancer and tissue of origin. Confidence intervals from the model output were obtained with the pROC (v.1.13) R package[40]. Assuming the prevalence of undiagnosed cancer cases in this population is high (1 or 2 cases per 100 healthy), a genomic assay with a specificity of 0.95 and sensitivity of 0.8 would have useful operating characteristics (positive predictive value of 0.25 and negative predictive value near 1). Power calculations suggest that an analysis of more than 200 patients with cancer and an approximately equal number of healthy controls, enable an estimation of the sensitivity with a margin of error of 0.06 at the desired specificity of 0.95 or greater. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

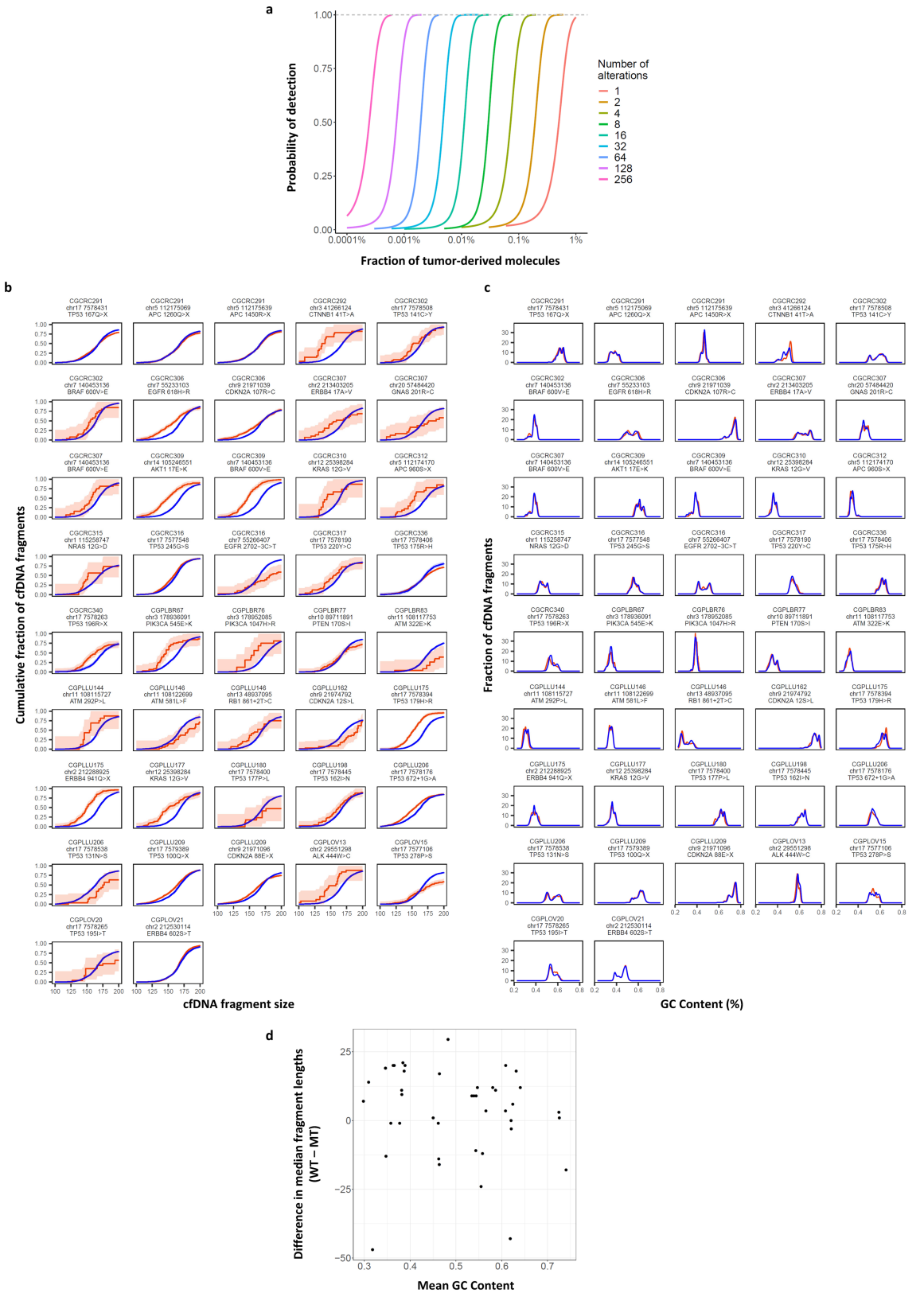
## Data availability

Sequence data used in this study have been deposited at the database of Genotypes and Phenotypes (dbGaP, study ID 34536).

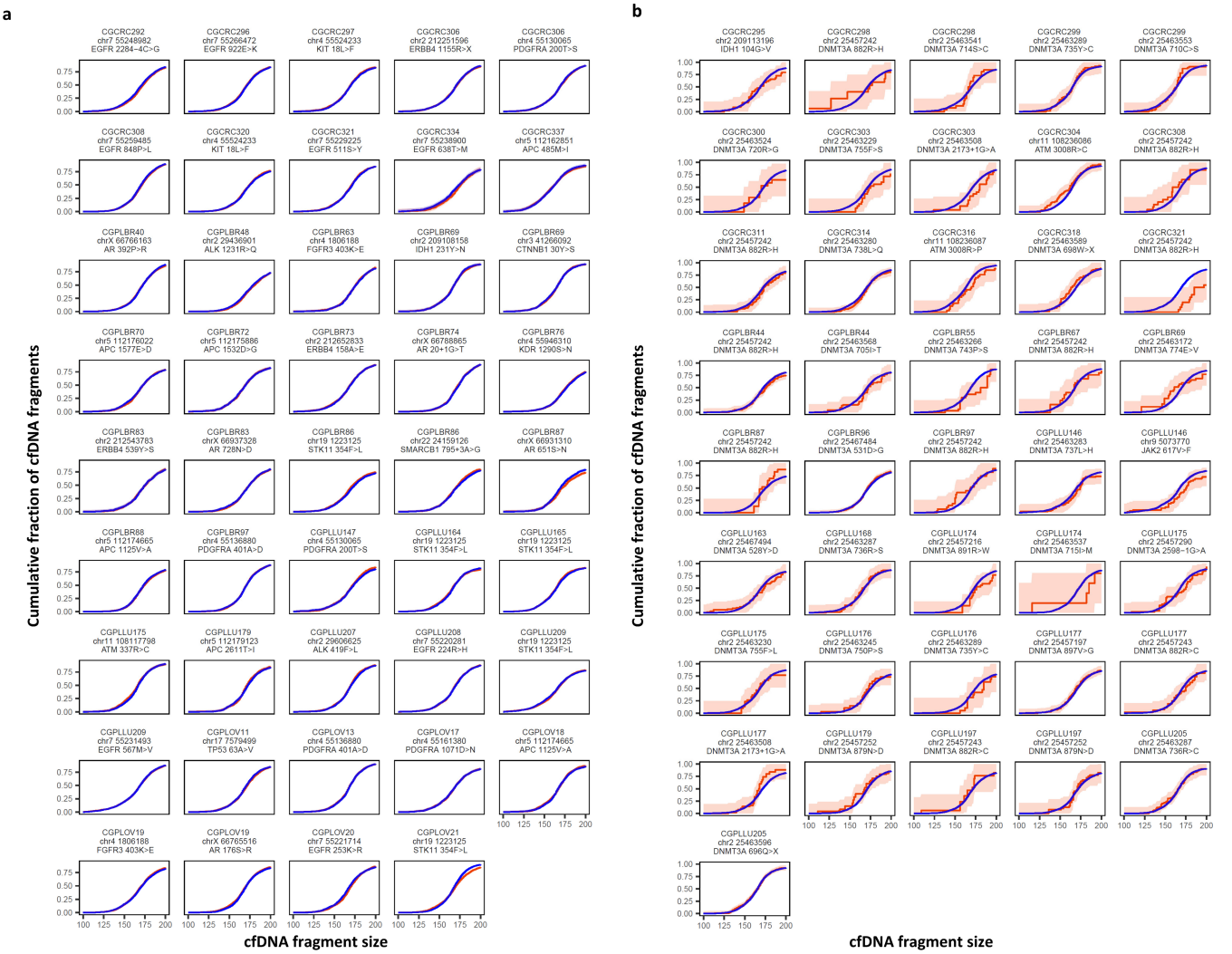
## Code availability

Code for analyses is available at http:github.com/Cancer-Genomics/delfi_scripts.

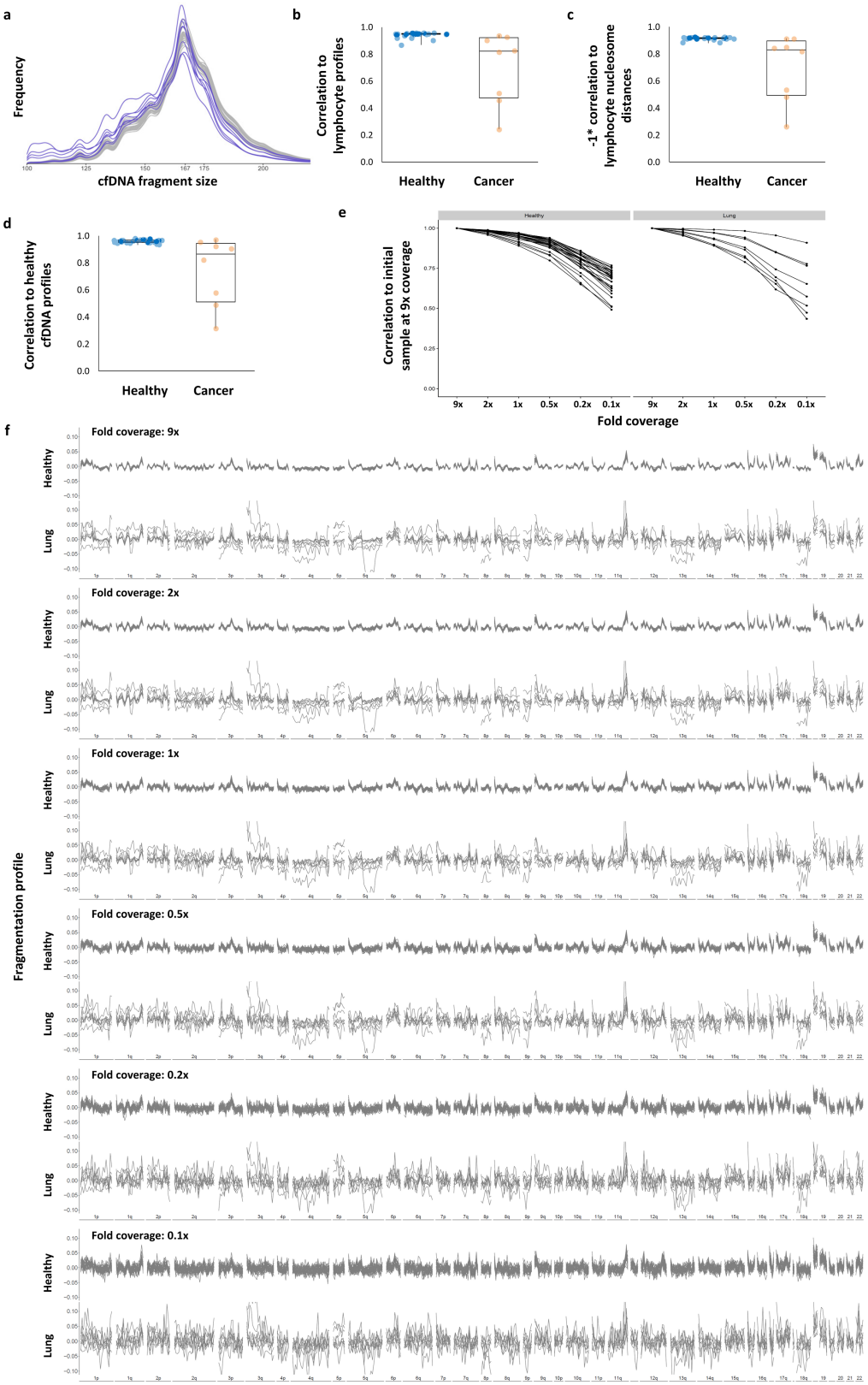

32. Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
33. Jones, S. et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7**, 283ra53 (2015).
34. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
36. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
37. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
38. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997).
39. Zurbenko, I. G. *The Spectral Analysis of Time Series* (Elsevier, 1986).
40. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).



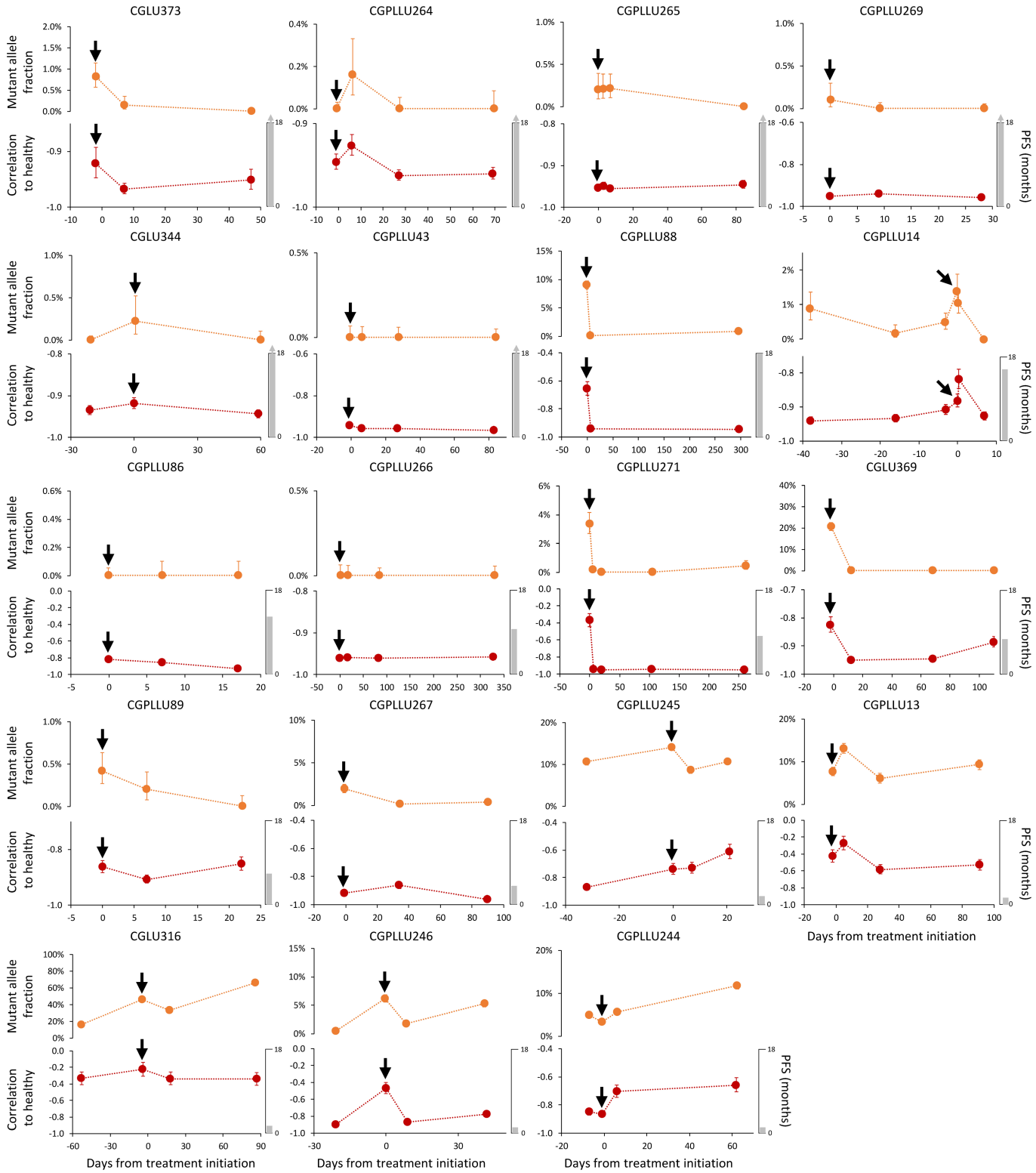

**Extended Data Fig. 1 | Simulations of non-invasive cancer detection based on number of alterations analysed and tumour-derived cfDNA fragment distributions. a**, Monte Carlo simulations were performed using different numbers of tumour-specific alterations to evaluate the probability of detecting cancer alterations in cfDNA at the indicated fraction of tumour-derived molecules. The simulations were performed assuming an average of 2,000 genome equivalents of cfDNA and the requirement of five or more observations of any alteration. These analyses indicate that increasing the number of tumour-specific alterations improves the sensitivity of detection of circulating tumour DNA. **b**, Cumulative density functions of cfDNA fragment lengths of 42 loci containing tumour-specific alterations from 30 patients with breast, colorectal, lung, or ovarian cancer are shown with 95% confidence bands (orange). Lengths of mutant cfDNA fragments were significantly different in size from wild-type cfDNA fragments (blue) at these loci. **c**, GC content was similar for mutated and non-mutated fragments. **d**, GC content was not correlated to fragment length.



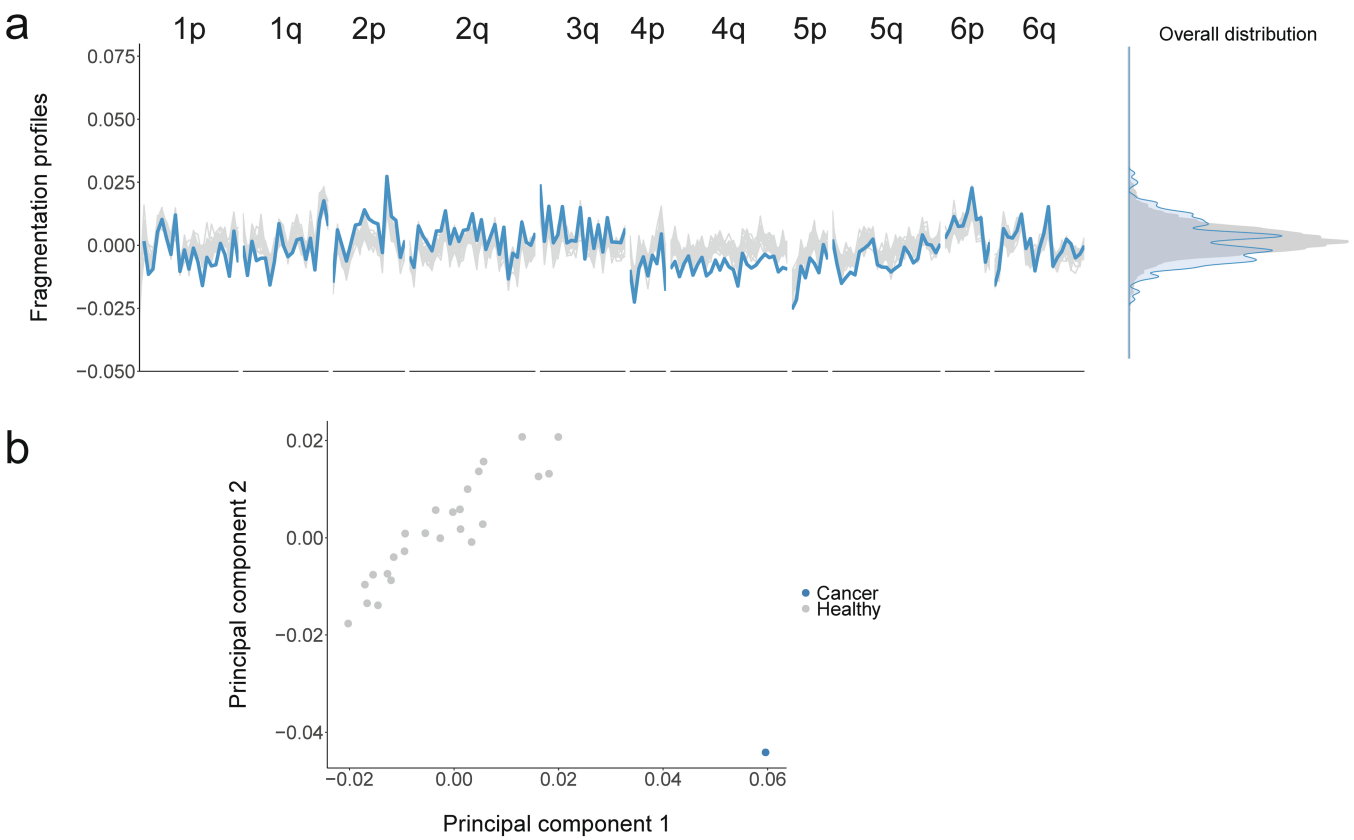

**Extended Data Fig. 2 | Germline and haematopoietic cfDNA fragment distributions. a**, Cumulative density functions of fragment lengths at 44 loci containing germline alterations (non-tumour derived) from 38 patients with breast, colorectal, lung or ovarian cancer are shown with 95% confidence bands. Fragments with germline mutations (orange) were comparable in length to wild-type cfDNA fragment lengths (blue). **b**, Cumulative density functions of fragment lengths at 41 loci containing haematopoietic alterations (non-tumour derived) from 28 patients with breast, colorectal, lung or ovarian cancer are shown with 95% confidence bands. After correction for multiple testing, there were no significant differences ($\alpha = 0.05$) in the size distributions of mutated haematopoietic cfDNA fragments (orange) and wild-type cfDNA fragments (blue).



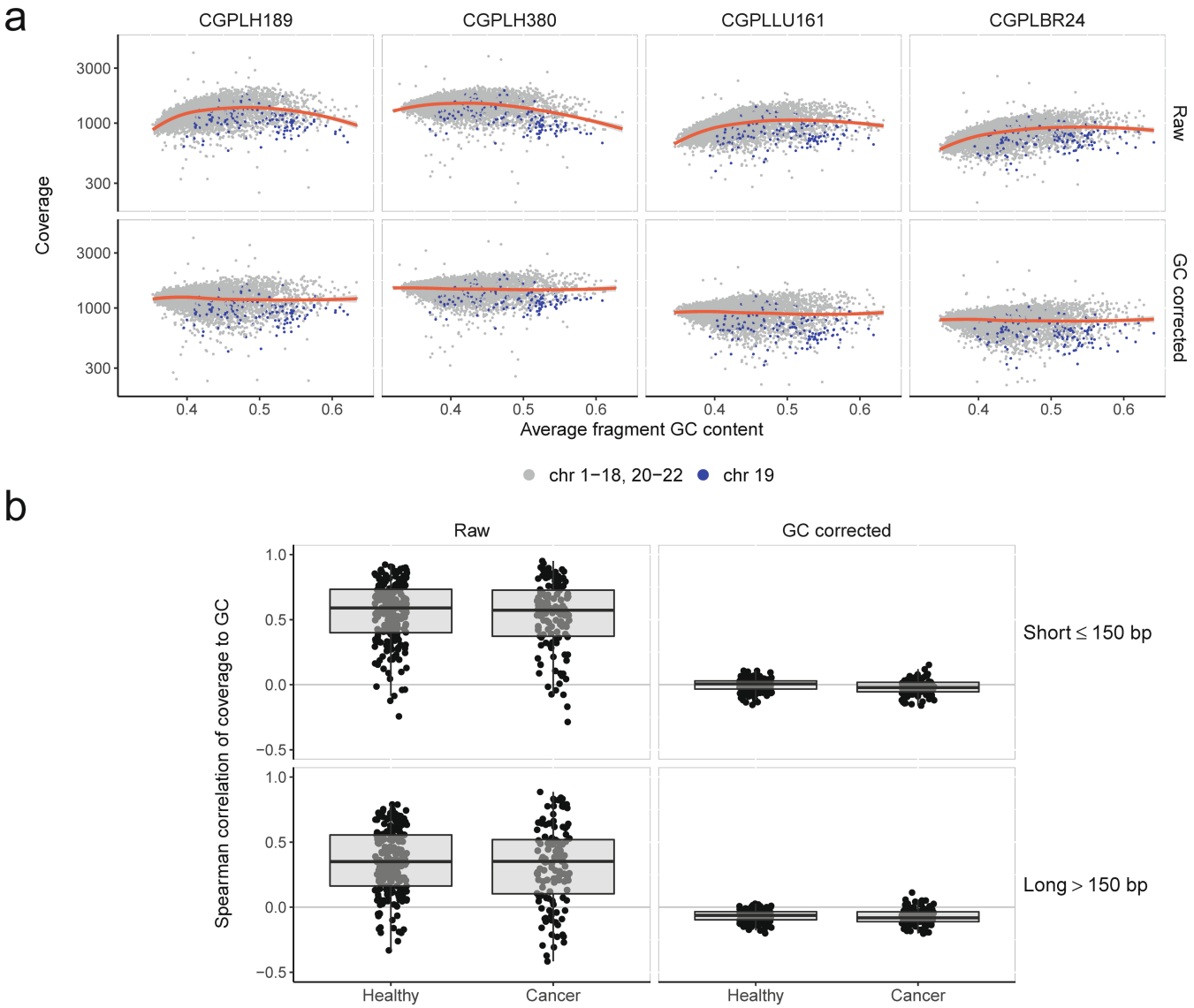

**Extended Data Fig. 3 | cfDNA fragmentation in healthy individuals and patients with lung cancer. a**, cfDNA fragment lengths are shown for healthy individuals ($n = 30$, grey) and patients with lung cancer ($n = 8$, blue). **b**–**d**, cfDNA fragmentation profiles from healthy individuals ($n = 30$) had high correlations, whereas patients with lung cancer ($n = 8$) had lower correlations to median fragmentation profiles of lymphocytes (**b**), lymphocyte nucleosome distances (**c**) and healthy cfDNA (**d**). Pearson correlations are shown with box plots depicting minimum, 25th percentile, median, 75th percentile, and maximum values. **e**, High coverage (9×) WGS data were subsampled to 2×, 1×, 0.5×, 0.2× and 0.1×-fold coverage. Mean centred genome-wide fragmentation profiles in 5-Mb bins for 30 healthy individuals and 8 patients with lung cancer are depicted for each subsampled fold coverage with median profiles shown in blue. **f**, Pearson correlation of subsampled profiles to initial profile at 9× coverage for healthy individuals and patients with lung cancer.



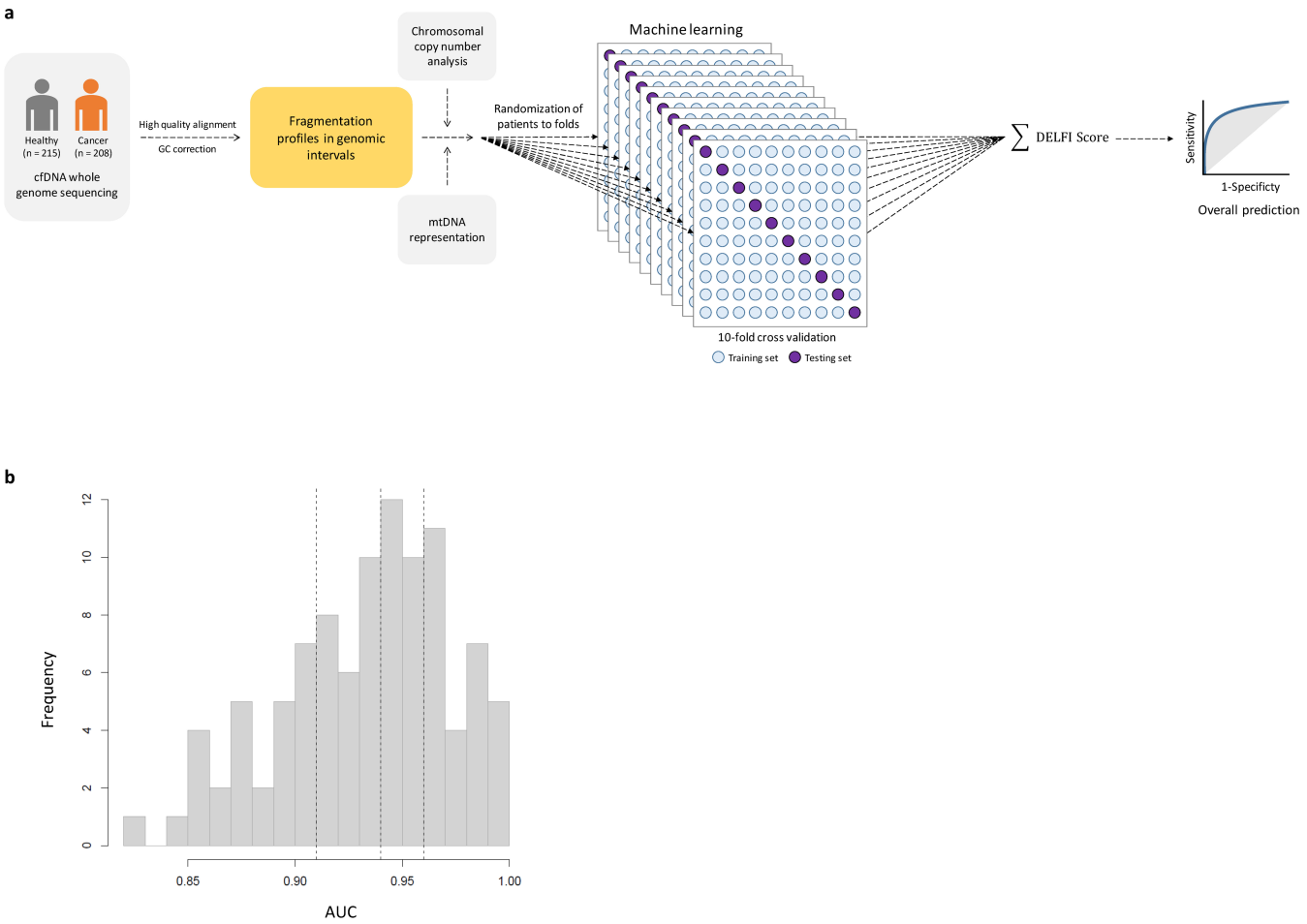

**Extended Data Fig. 4 | cfDNA fragmentation profiles and sequence alterations during therapy.** Detection and monitoring of cancer in serial blood draws from patients with non-small cell lung cancer ($n = 19$) undergoing treatment with targeted tyrosine kinase inhibitors (black arrows) was performed using targeted sequencing (top) as previously reported[29], and genome-wide fragmentation profiles (bottom). For each case, the vertical axis of the bottom panel displays −1 times the Pearson correlation of each sample to the median healthy cfDNA fragmentation profile. Error bars depict confidence intervals from binomial tests for mutant allele fractions, and confidence intervals calculated using Fisher transformation for genome-wide fragmentation profiles. Although the approaches analyse different aspects of cfDNA (whole genome compared with specific alterations), the targeted sequencing and fragmentation profiles were similar for patients responding to therapy as well as those with stable or progressive disease. As fragmentation profiles reflect both genomic and epigenomic alterations (whereas mutant allele fractions only reflect individual mutations), mutant allele fractions alone may not reflect the absolute level of correlation of fragmentation profiles to healthy individuals.



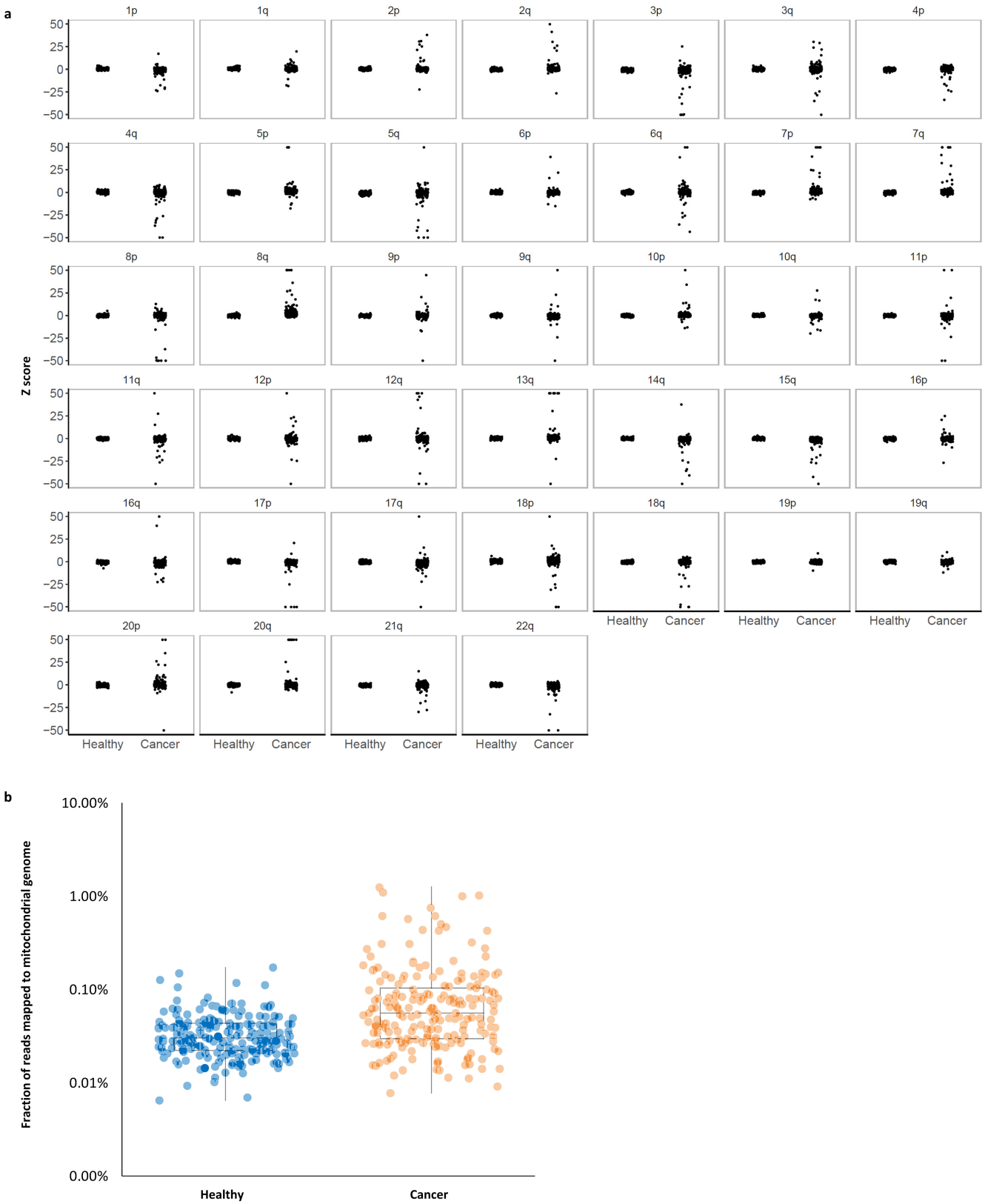

**Extended Data Fig. 5 | Profiles of cfDNA fragment lengths in copy neutral regions in healthy individuals and one patient with colorectal cancer. a**, The fragmentation profiles in 211 copy neutral windows in chromosomes 1–6 are shown for 25 randomly selected healthy individuals (grey). For a patient with colorectal cancer (CGCRC291) with an estimated mutant allele fraction of 20%, we diluted the cancer fragment length profile to an approximate 10% tumour contribution (blue). **a**, **b**, Although the marginal densities of the fragment profiles for the healthy samples and patient with cancer show substantial overlap (**a**, right), the fragmentation profiles are different as can be seen through visualization of the fragmentation profiles (**a**, left) and by the separation of the patient with colorectal cancer from the healthy samples ($n = 25$) in a principal component analysis (**b**).




**Extended Data Fig. 6 | Genome-wide GC correction of cfDNA fragments.** To estimate and control for the effects of GC content on sequencing coverage, we calculated coverage in non-overlapping 100-kb genomic windows across the autosomes. For each window, we calculated the average GC of the aligned fragments. **a**, LOESS smoothing of raw coverage (top row) for two randomly selected healthy subjects (CGPLH189 and CGPLH380) and two patients with cancer (CGPLLU161 and CGPLBR24) with undetectable aneuploidy (PA score < 2.35). After subtracting the average coverage predicted by the LOESS model, the residuals were rescaled to the median autosomal coverage (bottom row). As fragment length may also result in coverage biases, we performed this GC correction procedure separately for short (≤150 bp) and long (>150 bp) fragments. Although the 100-kb bins on chromosome 19 (blue points) consistently have less coverage than predicted by the LOESS model, we did not implement a chromosome-specific correction as such an approach would remove the effects of chromosomal copy number on coverage. **b**, Overall, we found a limited correlation between short or long fragment coverage and GC content after correction among healthy individuals ($n = 211$, interquartile range: −0.03–0.03) and patients with cancer ($n = 128$, interquartile range: −0.06–0.02) with a PA score < 3. Box plots depict 25th percentile, median, and 75th percentile values.

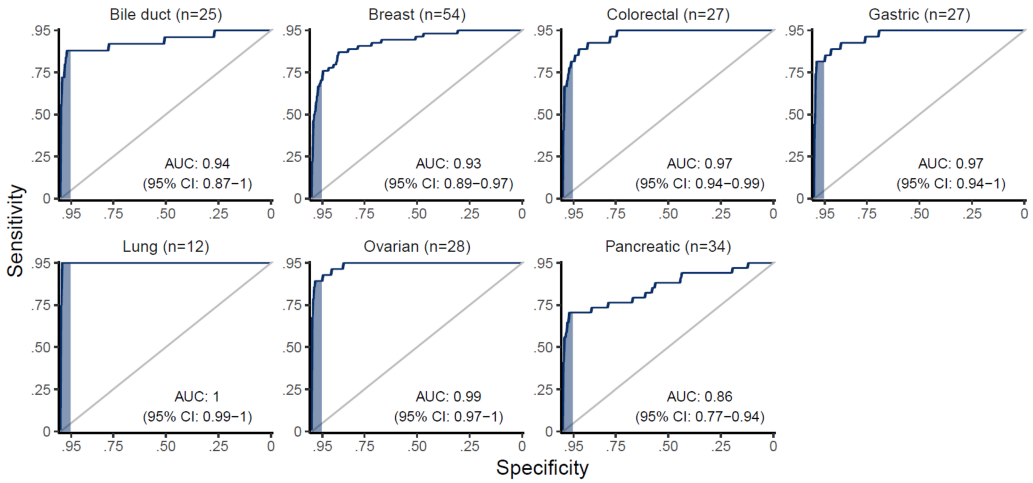



**Extended Data Fig. 7 | Machine learning model. a**, We used gradient tree boosting machine learning to examine whether cfDNA can be categorized as having characteristics of a patient with cancer or a healthy individual. The machine learning model included fragmentation size and coverage characteristics in windows throughout the genome, as well as chromosomal arm and mitochondrial DNA copy numbers. We used a tenfold cross-validation approach in which each sample is randomly assigned to a fold, and nine of the folds (90% of the data) are used for training and one fold (10% of the data) is used for testing. The prediction accuracy from a single cross-validation is an average over the ten possible combinations of test and training sets. As this prediction accuracy can reflect bias from the initial randomization of patients, we repeat the entire procedure, including the randomization of patients to folds, ten times. For all cases, feature selection and model estimation were performed on training data and were validated on test data, and the test data were never used for feature selection. Ultimately, we obtained a DELFI score that could be used to classify individuals as likely to be healthy or having cancer. **b**, Distribution of AUCs across the repeated tenfold cross-validation. The 25th, 50th and 75th percentiles of the 100 AUCs for the cohort of 215 healthy individuals and 208 patients with cancer are indicated by dashed lines.

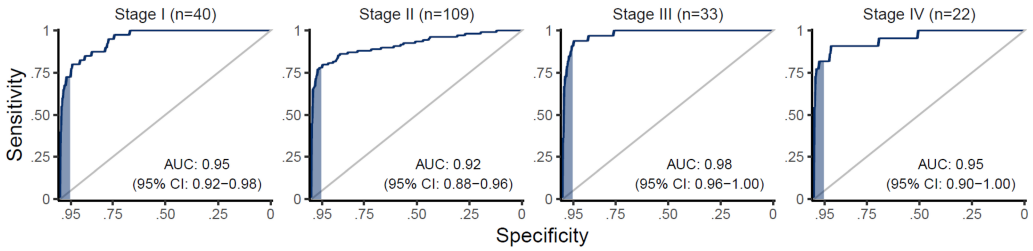



**Extended Data Fig. 8 | Whole-genome analyses of chromosomal arm copy number changes and mitochondrial genome representation.** **a**, *Z*-scores for each autosome arm are depicted for healthy individuals ($n = 215$) and patients with cancer ($n = 208$). The vertical axis depicts normal copy at zero with positive and negative values indicating arm gains and losses, respectively. *Z*-scores greater than 50 or less than −50 are thresholded at the indicated values. **b**, The fraction of reads mapping to the mitochondrial genome is depicted for healthy individuals ($n = 215$) and patients with cancer ($n = 208$). Box plots depict the minimum, 25th percentile, median, 75th percentile, and maximum values.

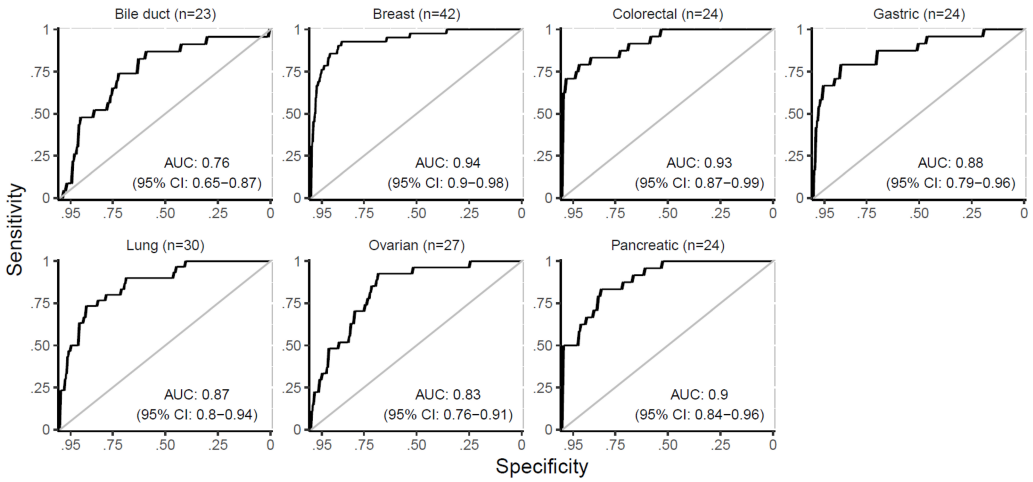
a




b




c




d

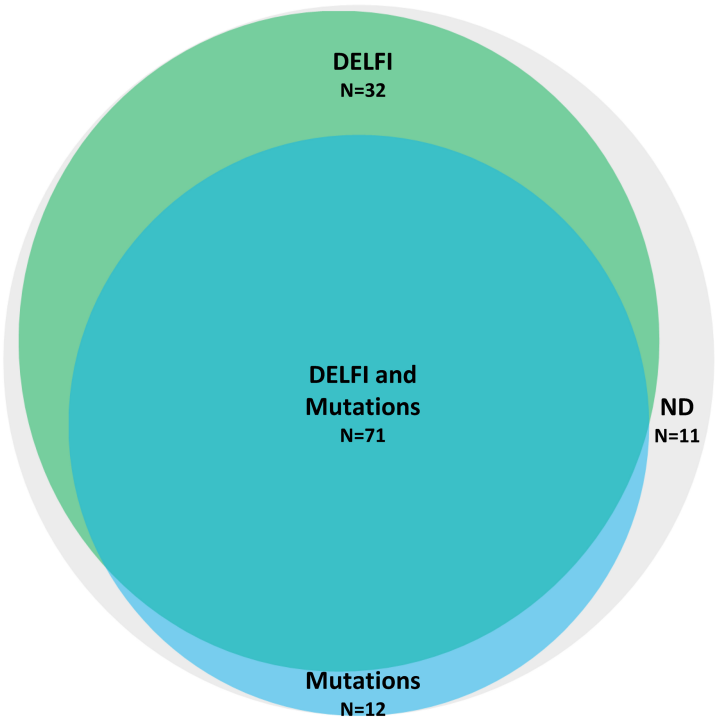
| Cancer Type | Patients Detected* | Top Prediction | | Top Two Predictions | | Random Assignment | |
|---|---|---|---|---|---|---|---|
| | | Patients | Accuracy (95% CI) | Patients | Accuracy (95% CI) | Patients | Accuracy |
| Breast | 42 | 32 | 76% (61%-88%) | 38 | 91% (77%-97%) | 9 | 22% |
| Bile Duct | 23 | 10 | 44% (23%-66%) | 15 | 65% (43%-84%) | 3 | 12% |
| Colorectal | 24 | 17 | 71% (49%-87%) | 19 | 79% (58%-93%) | 3 | 12% |
| Gastric | 24 | 16 | 67% (45%-84%) | 19 | 79% (58%-93%) | 3 | 12% |
| Lung | 30 | 16 | 53% (34%-72%) | 23 | 77% (58%-90%) | 2 | 6% |
| Ovarian | 27 | 13 | 48% (29%-68%) | 16 | 59% (38%-78%) | 4 | 14% |
| Pancreatic | 24 | 12 | 50% (29%-71%) | 16 | 67% (45%-84%) | 3 | 12% |
| Total | 194 | 116 | 61% (53%-67%) | 146 | 75% (69%-81%) | 26 | 13% |

*Patients detected are based on DELFI detection at 90% specificity. Lung cohort includes additional lung cancer patients with prior therapy.

**Extended Data Fig. 9 | DELFI detection of cancer and tissue of origin prediction. a**, Analyses of individual cancer types using DELFI had AUCs ranging from 0.86 to >0.99. **b**, Receiver operator characteristics for detection of cancer using cfDNA fragmentation profiles and other genome-wide features in a machine learning approach are depicted for a cohort of 215 healthy individuals and each stage of 208 patients with cancer with ≥95% specificity shaded in blue. **c**, Receiver operator characteristics for DELFI tissue prediction of bile duct, breast, colorectal, gastric, lung, ovarian or pancreatic cancer are depicted. To increase sample sizes within cancer type classes, we included cases detected with a 90% specificity, and the lung cancer cohort was supplemented with the addition of baseline cfDNA data from 18 patients with lung cancer with prior treatment[36]. **d**, DELFI tissue of origin prediction.

| Detection approach* | | Patients analyzed | Patients detected | Fraction of patients detected | 95% CI |
|---|---|---|---|---|---|
| DELFI | | 126 | 103 | 82% | 74%-88% |
| Mutations | | 126 | 83 | 66% | 57%-74% |
| DELFI and Mutations | | 126 | 115 | 91% | 85%-96% |
| Stage | I | 32 | 27 | 84% | 67%-95% |
| | II | 52 | 48 | 92% | 81%-98% |
| | III | 25 | 23 | 92% | 74%-99% |
| | IV | 16 | 16 | 100% | 79%-100% |

*Cancer detection using DELFI, sequence mutations, and the combination of DELFI and mutations was performed at specificities of 98%, >99%, and 98%, respectively. Per stage sensitivities are included for all cases except for one patient with stage X.




**Extended Data Fig. 10 | Detection of cancer using DELFI and mutation-based cfDNA approaches.** DELFI (green) and targeted sequencing[10] for mutation identification (blue) were performed independently in a cohort of 126 patients with breast, bile duct, colorectal, gastric, lung or ovarian cancer. The number of individuals detected by each approach and in combination are indicated for DELFI detection with a specificity of 98%, targeted sequencing specificity at >99%, and a combined specificity of 98%. ND, not detected.

nature research

Corresponding author(s): Victor E. Velculescu, Robert B. Scharpf

# 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<br>*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<br>*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<br>*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 | Primary processing of whole genome NGS data for cfDNA samples was performed using Illumina CASAVA (version 1.8.2) with alignment using ELAND. Re-alignment of mitochondrial mapped reads was done using Bowtie-2 (version 2.3.4) in end-to-end mode. Sequence reads in the analysis of targeted data were aligned using NovoAlign (version 3.02.12) and variant calling was performed using VariantDx. |
| Data analysis | All analyses were performed using R (version 3.4.3). Binning of the human genome was done using the R package GenomicRanges (version 1.30.3). Prediction was implemented using R package gbm (version 2.1.4), caret (version 6.0.79) and pROC (version 1.13). Bam file processing and analyses of targeted data were performed using Rsamtools (version 1.30.0) and GenomicAlignments (version 1.14.2). |

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

Sequence data utilized in this study have been deposited at the database of Genotypes and Phenotypes (dbGaP, study ID 34536). Code for analyses is available at http:github.com/Cancer-Genomics/delfi_scripts.

# 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 | Assuming the prevalence of undiagnosed cancer cases in this population is high (1 or 2 cases per 100 healthy), a genomic assay with a specificity of 0.95 and sensitivity of 0.8 would have useful operating characteristics (positive predictive value of 0.25 and negative predictive value near 1). An analysis of more than 200 cancer patients and an approximately equal number of healthy controls would provide an estimation of the sensitivity with a margin of error of 0.06 at the desired specificity of 0.95. |
| Data exclusions | Plasma samples that were not collected according the described protocol were not used in the analysis. |
| Replication | We successfully performed internal replication in our study, including replication of the initial observations of our pilot study in a larger analysis of healthy individuals and patients with cancer. |
| Randomization | Not applicable |
| Blinding | Not applicable |

# 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 |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | Plasma samples from healthy individuals and plasma and tissue samples from patients with breast, lung, ovarian, colorectal, bile duct, and gastric cancers were obtained from ILSBio/Bioreclamation, Aarhus University, Herlev Hospital of the University of Copenhagen, Hvidovre Hospital, the University Medical Center of the University of Utrecht, the Academic Medical Center of the University of Amsterdam, the Netherlands Cancer Institute, and the University of California, San Diego. All samples were obtained under Institutional Review Board approved protocols with informed consent for research use at participating institutions. Plasma samples from healthy individuals were obtained at the time of routine screening, including for |

| | |
|---|---|
| | colonoscopies or Pap smears. Individuals were considered healthy if they had no previous history of cancer and negative screening results. |
| Recruitment | Participants were recruited through screening trials, observational trials, or through formal biospecimen collection at University center hospitals. Potential self-selection bias or other biases were not identified. |