nature biomedical engineering

Article

https://doi.org/10.1038/s41551-025-01370-3

A deep-learning model for quantifying circulating tumour DNA from the density distribution of DNA-fragment lengths

Received: 12 August 2023

Accepted: 12 February 2025

Published online: 7 March 2025

Check for updates

Guanhua Zhu[1,2,3,4,11], Chowdhury Rafeed Rahman[1,5,11], Victor Getty[1], Denis Odinokov[1], Probhonjon Baruah[1], Hanaé Carrié[1,5,6,7], Avril Joy Lim[1,5], Yu Amanda Guo[1], Zhong Wee Poh[1,8], Ngak Leng Sim[1], Ahmed Abdelmoneim[1], Yutong Cai[1], Lakshmi Narayanan Lakshmanan[1], Danliang Ho[1], Saranya Thangaraju[1], Polly Poon[1], Yi Ting Lau[1], Anna Gan[1], Sarah Ng[1], Si-Lin Koo[9], Dawn Q. Chong[8,9], Brenda Tay[9], Tira J. Tan[9], Yoon Sim Yap[8,9], Aik Yong Chok[10], Matthew Chau Hsien Ng[8,9], Patrick Tan[1,8], Daniel Tan[1,8,9], Limsoon Wong[5], Pui Mun Wong[1], Iain Beehuat Tan[1,8,9] & Anders Jacobsen Skanderup[1,5,9] ✉

The quantification of circulating tumour DNA (ctDNA) in blood enables non-invasive surveillance of cancer progression. Here we show that a deep-learning model can accurately quantify ctDNA from the density distribution of cell-free DNA-fragment lengths. We validated the model, which we named 'Fragle', by using low-pass whole-genome-sequencing data from multiple cancer types and healthy control cohorts. In independent cohorts, Fragle outperformed tumour-naive methods, achieving higher accuracy and lower detection limits. We also show that Fragle is compatible with targeted sequencing data. In plasma samples from patients with colorectal cancer, longitudinal analysis with Fragle revealed strong concordance between ctDNA dynamics and treatment responses. In patients with resected lung cancer, Fragle outperformed a tumour-naive gene panel in the prediction of minimal residual disease for risk stratification. The method's versatility, speed and accuracy for ctDNA quantification suggest that it may have broad clinical utility.

The death of non-malignant cells, primarily of the haematopoietic lineage, releases cell-free DNA (cfDNA) into the blood circulation[1]. In patients with cancer, the blood plasma also carries circulating tumour DNA (ctDNA), enabling non-invasive diagnostics and disease surveillance[2]. The ability to monitor tumour growth dynamics based on ctDNA levels in the blood provides a promising non-invasive approach to track disease progression during therapy and clinical trials[3–5].

Ultra-deep targeted cfDNA sequencing assays are often preferred in the clinic owing to their ability to identify actionable mutations. While mutation variant allele frequencies (VAFs) can be used to approximate ctDNA levels, not all tumours will have mutations covered by a given targeted sequencing gene panel. Furthermore, the accuracy of this approximation depends on sample-specific and treatment-dynamic properties such as mutation clonality, copy number and potential confounding noise from clonal haematopoiesis[6]. Existing methods

[1]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [2]Centre for Novostics, Hong Kong SAR, China. [3]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China. [4]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China. [5]School of Computing, National University of Singapore, Singapore, Singapore. [6]Institute of Data Science, National University of Singapore, Singapore, Singapore. [7]Integrative Sciences and Engineering Programme, Graduate School, National University of Singapore, Singapore, Singapore. [8]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore. [9]National Cancer Center Singapore, Singapore, Singapore. [10]Singapore General Hospital, Singapore, Singapore. [11]These authors contributed equally: Guanhua Zhu, Chowdhury Rafeed Rahman. ✉e-mail: skanderupamj@gis.a-star.edu.sg

https://doi.org/10.1038/s41551-025-01370-3

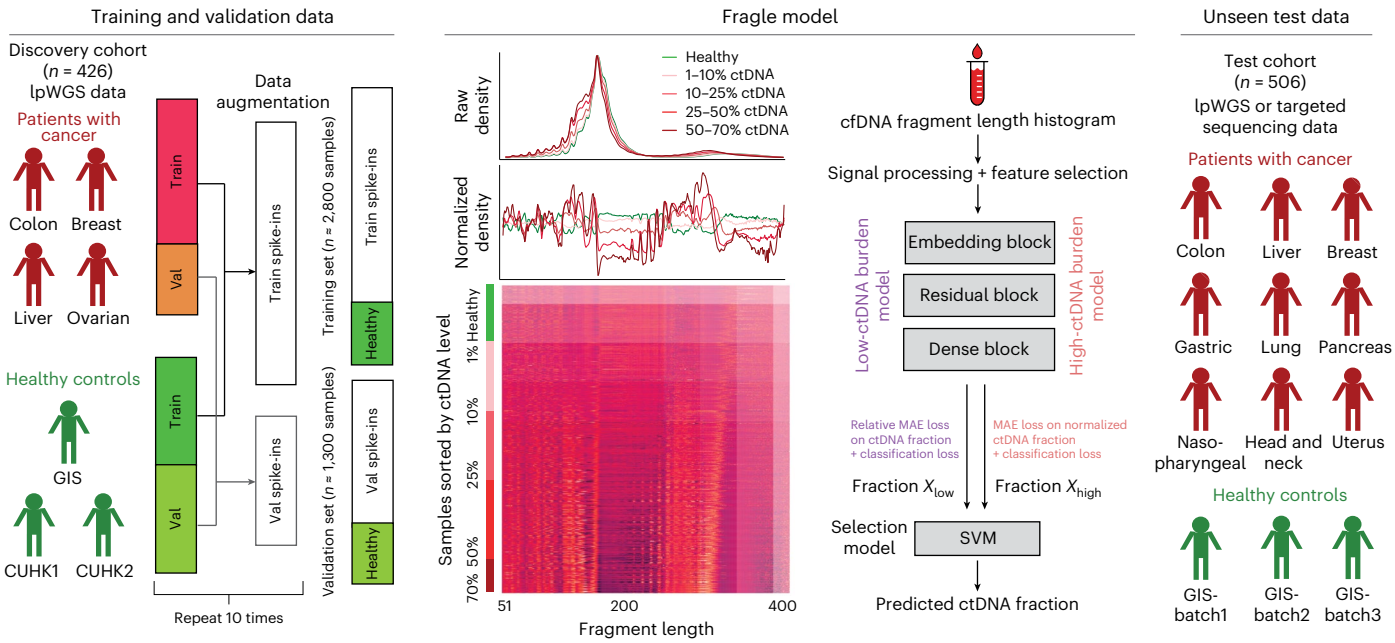


Fig. 1 | Overview of Fragle. Fragle is a multi-stage machine learning-based model that estimates the ctDNA level in a blood sample from the cfDNA fragment length density distribution. Fragle was trained using a large-scale data augmentation and cross-validation approach, with samples divided into training (Train) and validation (Val) sets. Fragle was further tested using unseen samples from multiple cancer types and healthy control cohorts.

developed for ctDNA quantification are not directly compatible with targeted sequencing panels. These methods require either low-pass whole-genome-sequencing (lpWGS) data[7], DNA methylation profiling[8,9] or modifications to the targeted sequencing panel[10]. Thus, there is an unmet need to develop accurate and orthogonal approaches for ctDNA quantification that can generalize across patients, tumour types and sequencing modalities.

The fragment length distribution of cfDNA in plasma has a mode of ~166 base pairs (bp) as nucleosome-bound cfDNA molecules show increased protection from DNA degradation[11]. cfDNA fragments from patients with cancer tend to be shorter than those from healthy individuals, typically with a higher proportion of fragments under 150 bp (refs. 12–14). Shorter cfDNA fragments have also been observed in plasma bisulfite sequencing data from patients with cancer[15]. The size profile of these shorter fragments from patients with cancer also exhibits increased 10-bp oscillation amplitude in the range of 90–145 bp (ref. 16). cfDNA from patients with cancer may also show a higher proportion of fragments longer than 180 bp (refs. 12,16). Other studies have indicated that variation in fragment lengths in patients with cancer could be position dependent within the genome[17]. These observations have motivated studies exploring how cfDNA fragment length properties can be used to classify cfDNA samples from patients with cancer and healthy individuals[12,15–23]. Here we developed Fragle, a multi-stage machine learning model that quantifies ctDNA levels from a cfDNA fragment length density distribution. Using an in silico data augmentation approach, we trained and evaluated Fragle on ~4,000 lpWGS samples across distinct cancer types and healthy cohorts. We evaluated the accuracy and the lower limit of detection (LoD) in independent cohorts and cancer types. Intriguingly, we demonstrate that Fragle can also be applied to cfDNA fragmentomic profiles obtained from targeted sequencing panels. Using this feature, we applied Fragle to longitudinal plasma samples to explore the correlation of ctDNA dynamics and treatment response measured through radiographic imaging (RI). Finally, to explore the use of Fragle for detection of minimal residual disease (MRD), we analysed ctDNA levels in a cohort of 162 patients with resected lung cancer with plasma profiled using a commercial targeted sequencing panel at the landmark timepoint (~30 days following surgery).

Results

Quantitative prediction of ctDNA levels from fragmentomic data

We assembled a discovery cohort comprising lpWGS data from 325 cancer plasma samples from 4 cancer types (colon, breast, liver and ovarian cancer) and 101 plasma samples from healthy individuals (Fig. 1 and Supplementary Dataset 1). In this dataset, we estimated ground-truth ctDNA levels using multiple methods (Methods and Supplementary Dataset 2), and the cancer samples were further selected based on ctDNA levels (>3%, $n = 164$; Extended Data Fig. 1). Using a large-scale data augmentation approach, we performed in silico dilution of these cancer samples and the 101 healthy control samples, generating ~4,000 mixture samples with variable ctDNA fractions for model training (Methods, Fig. 1 and Supplementary Dataset 3). To explore how cfDNA fragment length distributions could predict ctDNA levels in a sample, we derived raw fragment length density distributions using paired-end reads in each sample. Raw density distributions were further normalized and transformed, revealing local differences in the fragment length distributions associated with ctDNA levels in the samples (Methods and Fig. 1). The transformed fragment length distributions, in combination with their labels in the form of ground-truth ctDNA levels, served as input to a multi-stage supervised machine learning approach. We used two parallel sub-models, each designed for either low- or high-ctDNA fraction samples, followed by a model that selects the final predicted ctDNA fraction from the output of the two sub-models (Methods). The two sub-models performed well for the intended low- and high-ctDNA samples, respectively (Supplementary Fig. 1), while the final combined model achieved the lowest overall prediction error (mean absolute error (MAE) = 3.2%) compared with individual sub-models (MAE 4.0% and 3.3% for low- and high-ctDNA sub-models). Notably, although the improvement in overall MAE is modest compared with the high-ctDNA sub-model, the final combined model significantly improved the prediction accuracy for healthy samples (MAE 0.5% versus 1.0%) and specificity at an LoD of 1% (86% versus 68%).

The model was trained and evaluated using cross-validation, demonstrating high predictive accuracy on validation samples across all 4 cancer types (Fig. 2a–d and Supplementary Dataset 4): colorectal

(MAE = 3.3%; Pearson $r = 0.92$), breast (MAE = 3.6%; $r = 0.94$), liver (MAE = 3.1%; $r = 0.81$) and ovarian cancer (MAE = 3.9%; $r = 0.67$). The lower concordance for ovarian cancer could be attributed to samples from one patient; removal of these samples increased the correlation to $r = 0.88$ (Supplementary Fig. 2).

We trained the final Fragle model on the full discovery cohort (Methods) and tested its performance on additional cohorts of unseen plasma lpWGS samples. We observed a strong correlation between Fragle and ichorCNA-based ctDNA fraction estimates across unseen cohorts of colorectal ($r = 0.81$; $P = 6.8 \times 10^{-41}$; $n = 172$; Fig. 2e), breast ($r = 0.80$; $P = 3.7 \times 10^{-6}$; $n = 23$; Fig. 2f), liver ($r = 0.86$; $P = 5.1 \times 10^{-10}$; $n = 34$; Fig. 2g) and gastric cancer ($r = 0.72$; $P = 3.4 \times 10^{-13}$; $n = 74$; Fig. 2h). We also tested Fragle on a mixed cohort of cancer types not included in the discovery set, including lung, nasopharyngeal, and head and neck cancers ($r = 0.63$, 0.75 and 0.23; combined $P = 1.3 \times 10^{-4}$, $n = 10$ for each cancer type; Extended Data Fig. 2). In the unseen colorectal cancer cohort, we also performed targeted gene sequencing and identified high-confidence somatic mutations in 82 samples (Methods and Supplementary Dataset 5). These data demonstrated high concordance between mutation VAFs and Fragle-predicted ctDNA levels ($r = 0.88$; $P = 3.8 \times 10^{-28}$; Fig. 2i). Expectedly, higher ctDNA fractions were generally observed in patients with late-stage tumours (Fig. 2j). Furthermore, we observed a significant difference between ctDNA levels estimated for early-stage cancers (stages 1 and 2; colon, liver and gastric cancer) and healthy controls ($P = 1.3 \times 10^{-9}$, Wilcoxon rank sum test).

We trained Fragle using samples each comprising 10 million cfDNA fragments, equivalent to ~1× WGS using 151 bp paired-end sequencing. To further evaluate the sequencing coverage requirements for Fragle, we down-sampled WGS samples from the unseen test cohort to render samples with fewer fragments, ranging from 5 million (0.5×) to as low as 10 thousand fragments (0.001×). At 500 thousand fragments (0.05×), Fragle demonstrated excellent concordance ($r = 0.97$) with the predictions from the original 1× WGS samples (Extended Data Fig. 3). The correlation was maintained when further down-sampling to 250 thousand fragments, but with some discrepancies observed for some low-ctDNA fraction samples (Extended Data Fig. 3). These results suggest that whole-genome coverage of ~500 thousand (0.05×) fragments provides a good trade-off between prediction accuracy and sequencing cost. In addition, we tested the computational requirements of Fragle as a software tool. Fragle processed a 1×-coverage WGS sample in ~50 s using a single processor and required low memory usage independent of the sample sequencing depth (Supplementary Fig. 3).

Determination of the lower LoD

To explore the LoD for the model, we first observed that Fragle predicted very low ctDNA fractions (median = 0.07%) for the healthy samples in the validation sets. In this healthy cohort, Fragle demonstrated 86% specificity at a 1% LoD level, increasing to 95% at 3% LoD (Supplementary Dataset 6). Furthermore, the model could differentiate between healthy and low-ctDNA level samples at the 1% ctDNA level (Wilcoxon rank sum test, $P = 2.5 \times 10^{-24}$; Fig. 3a), indicating an ~1% LoD in these samples. Similarly, we examined the performance of Fragle for classification of healthy and cancer samples in the validation sets. Using cancer samples with a ctDNA level ≥1% in the validation sets, Fragle demonstrated an area under the curve (AUC) of 0.93 (Fig. 3b), higher than ichorCNA (AUC = 0.88). Notably, after limiting the analysis to the samples in which the ground-truth ctDNA fraction was estimated from a consensus of multiple methods, Fragle further outperformed ichorCNA in classifying cancer and healthy samples (AUC 0.98 versus 0.92; Supplementary Fig. 4). Expectedly, the AUC increased further when filtering out low-ctDNA burden samples (Supplementary Fig. 5). Fragle and ichorCNA achieved AUCs of 0.97 and 0.94, respectively, when excluding cancer samples with ctDNA levels below the LoD of ichorCNA of 3% (Supplementary Fig. 6). We also noticed a trend that Fragle exhibited improved ctDNA detection over ichorCNA in samples with low levels of copy number variation (Supplementary Fig. 7). As an additional comparison, we explored other fragment length features previously used for the classification of cancer and healthy samples[16], and trained a random forest model on the discovery cohort using four features derived from the fragment length distribution (Methods). This four-feature model demonstrated substantially lower classification accuracy (AUC = 0.79) than Fragle in the validation sets.

We further evaluated the LoD using unseen test samples. We used cfDNA samples from patients with CRC with detectable mutations as positive cancer samples ($n = 82$) and all healthy plasma samples from the unseen cohorts as negatives ($n = 57$). Fragle demonstrated an AUC of 0.96 using these samples, outperforming the other models on the same set of samples (Fig. 3c and Supplementary Dataset 7). These results were further confirmed using an in silico dilution experiment. This experiment involved 13 unseen colon cancer and 7 unseen breast cancer samples with high ctDNA burden (>10%), concordantly estimated by Fragle and ichorCNA (Methods). In this dilution experiment, Fragle could differentiate healthy from low-ctDNA samples down to the 0.5–1% ctDNA level ($P = 0.003$, healthy versus 0.625% ctDNA fraction samples; Fig. 3d and Supplementary Fig. 8).

To further examine these results using physical samples, we performed similar dilution experiments in vitro. The first experiment comprised serial dilutions of two high-ctDNA level CRC plasma samples, with samples progressively diluted using pooled cfDNA from healthy individuals (Methods). Across three technical replicates, Fragle accurately predicted ctDNA fractions for both patients down to ~1% ctDNA level, with healthy samples consistently predicted <1% ctDNA (Fig. 3e). For low-ctDNA samples with 1–3% diluted ctDNA fraction, the detection rate was 94% at an LoD of 1%, outperforming ichorCNA with a detection rate of 67%. The second experiment comprised in vitro serial dilutions of three high-ctDNA plasma samples from patients with gastric cancer (each with three technical replicates; Methods). The results from this experiment mirrored our previous observations, with the method accurately quantifying ctDNA down to the 0.5–1% level and predicting <1% ctDNA for healthy samples (Fig. 3f). Overall, these results collectively suggest that Fragle can quantify and detect ctDNA with an LoD of ~1%.

Application of Fragle to targeted sequencing data

Targeted gene sequencing of plasma samples is routinely used for tumour genotyping in the clinic. However, absolute ctDNA quantification based on mutation VAFs remains challenging using targeted sequencing. For example, samples may not have clonal mutations covered by the panel, and non-cancer variants associated with clonal haematopoiesis could introduce noise[24]. To explore whether Fragle could quantify ctDNA levels using targeted sequencing data, we analysed four cfDNA cohorts having both lpWGS and targeted sequencing data (Methods and Fig. 4a). Using standard on-target reads obtained from the targeted sequencing data, Fragle tended to overestimate the ctDNA burden compared with the lpWGS data (Fig. 4b). We then evaluated the method on off-target reads, which are often filtered and ignored in a targeted sequencing experiment. Remarkably, we observed strong concordance of predictions based on lpWGS and off-target reads across all four cohorts: breast cancer samples from the discovery cohort ($r = 0.86$, $P = 0.001$, $n = 10$), colon cancer samples from the discovery cohort ($r = 0.96$, $P = 1.64 \times 10^{-30}$, $n = 56$), colon cancer samples from the unseen cohort ($r = 0.97$, $P = 3.69 \times 10^{-58}$, $n = 96$) and metastatic gastric cancer samples from the unseen cohort ($r = 0.96$, $P = 2.9 \times 10^{-27}$, $n = 49$) (Fig. 4c). We found that the targeted sequencing samples contained between 100 thousand to 10 million off-target fragments (equivalent to ~0.01–1.0× WGS) across the different samples, with >95% of samples having >250 thousand off-target fragments (~0.025×; Supplementary Fig. 9). Expectedly, the off-target coverage levels showed a linear relationship to on-target coverage across samples (Supplementary Fig. 9). To further explore if these results generalize to other targeted

https://doi.org/10.1038/s41551-025-01370-3

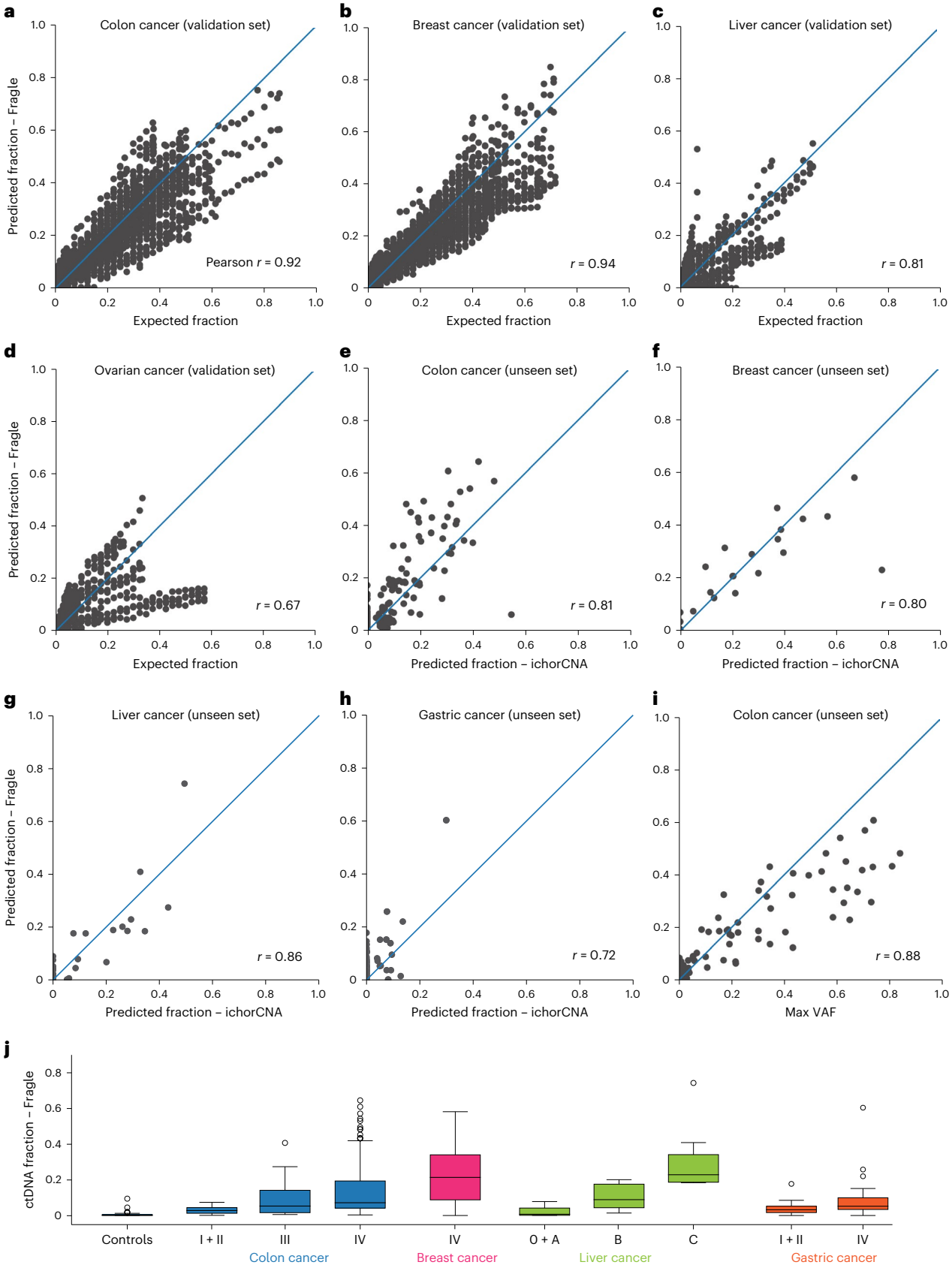


Fig. 2 | ctDNA quantification in validation and unseen cohorts. **a**–**d**, Comparison between expected and predicted ctDNA levels for colorectal (CRC) (**a**), breast (BRCA) (**b**), liver (HCC) (**c**) and ovarian (OV) cancer samples (**d**) in the validation sets. **e**–**h**, Comparing ichorCNA and Fragle predicted ctDNA levels in unseen samples from patients with colorectal ($n = 172$) (**e**), breast ($n = 23$) (**f**), liver ($n = 34$) (**g**) and gastric cancer ($n = 74$) (**h**). **i**, Colorectal cancer plasma samples subjected to both lpWGS and targeted sequencing; comparison of Fragle predicted ctDNA levels (lpWGS) and maximum VAFs ($n = 82$; samples with detectable somatic mutations). **j**, Predicted ctDNA levels in plasma samples from patients with cancer grouped according to tumour stages. Box plots are represented by median and interquartile range (IQR), with ±1.5 IQR as whiskers.

https://doi.org/10.1038/s41551-025-01370-3

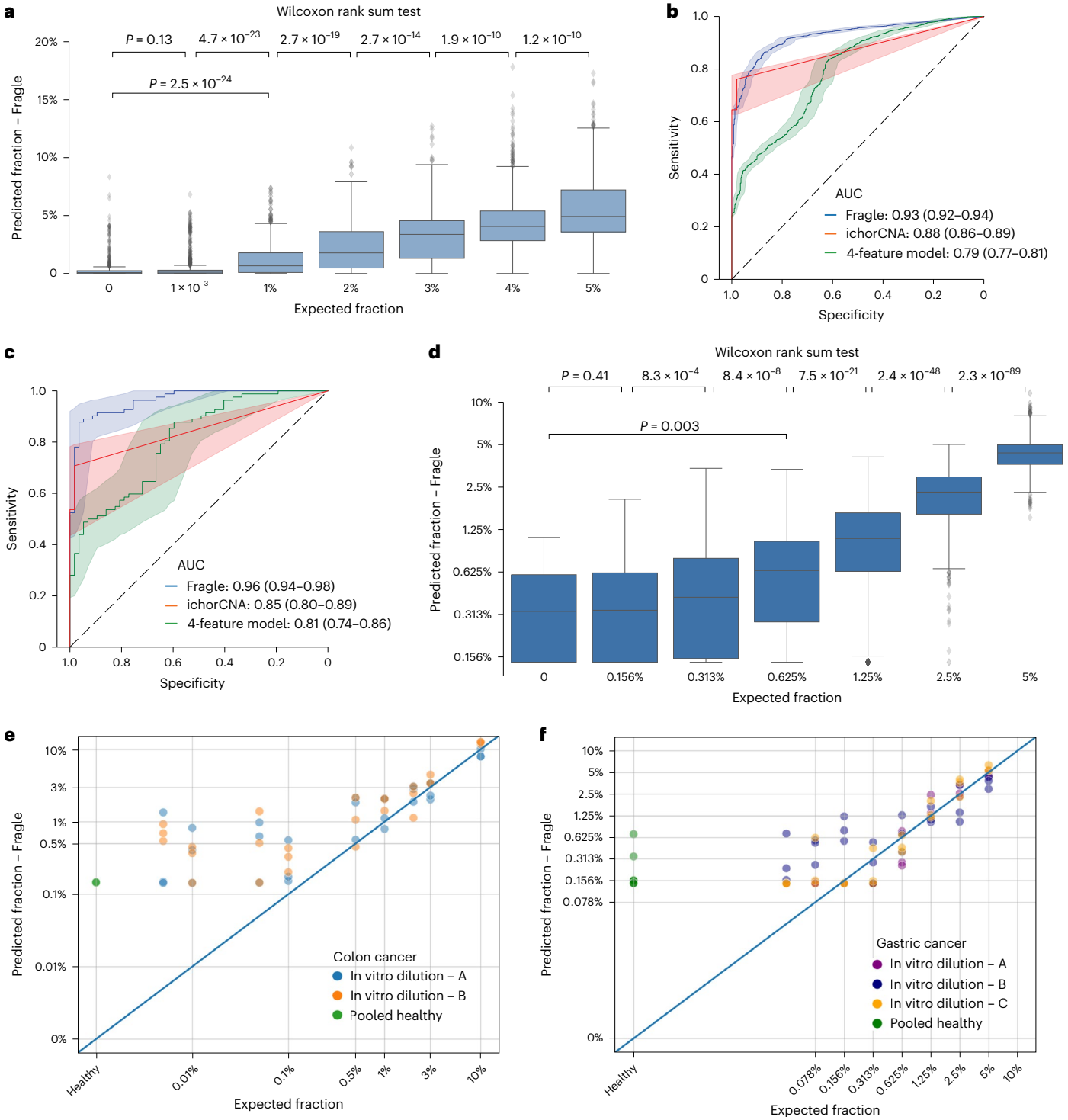


Fig. 3 | Lower LoD. a, Predicted ctDNA fractions for healthy and low-ctDNA level samples in validation set samples. Box plots are represented by median and IQR, with ±1.5 IQR as whiskers. **b**, Receiver operating characteristic analyses for classification of healthy control and cancer (≥1% ctDNA) samples (validation samples). AUC values with 95% confidence intervals are shown. **c**, Receiver operating characteristic analysis for classification of cancer ($n = 82$) and healthy ($n = 57$) plasma samples in the unseen test cohort. AUC values with 95% confidence intervals are shown. **d**, Predicted ctDNA fractions for healthy and low-ctDNA level samples using in silico dilution of 20 cancer samples (unseen cohort). Box plots are represented by median and IQR, with ±1.5 IQR as whiskers. **e**, Expected versus predicted ctDNA fractions using in vitro ctDNA dilution for two colorectal cancer samples. **f**, Expected versus predicted ctDNA fractions using in vitro ctDNA dilution for three gastric cancer samples.

sequencing assays, we evaluated a cohort of 116 plasma samples subjected to a liquid biopsy gene panel from a commercial vendor (Foundation Medicine)[25]. Since these samples did not have matched lpWGS data, we approximated ctDNA levels using the maximum VAFs reported by the company after filtering out germline variants (Methods). In this cohort comprising samples from five different cancer types, we observed that ctDNA levels estimated from off-target reads were generally concordant with the reported VAFs ($r = 0.62$, $P = 1.4 \times 10^{-13}$, $n = 116$; Fig. 4d). Overall, these results support that Fragle can estimate ctDNA levels using both lpWGS and targeted sequencing data.

https://doi.org/10.1038/s41551-025-01370-3

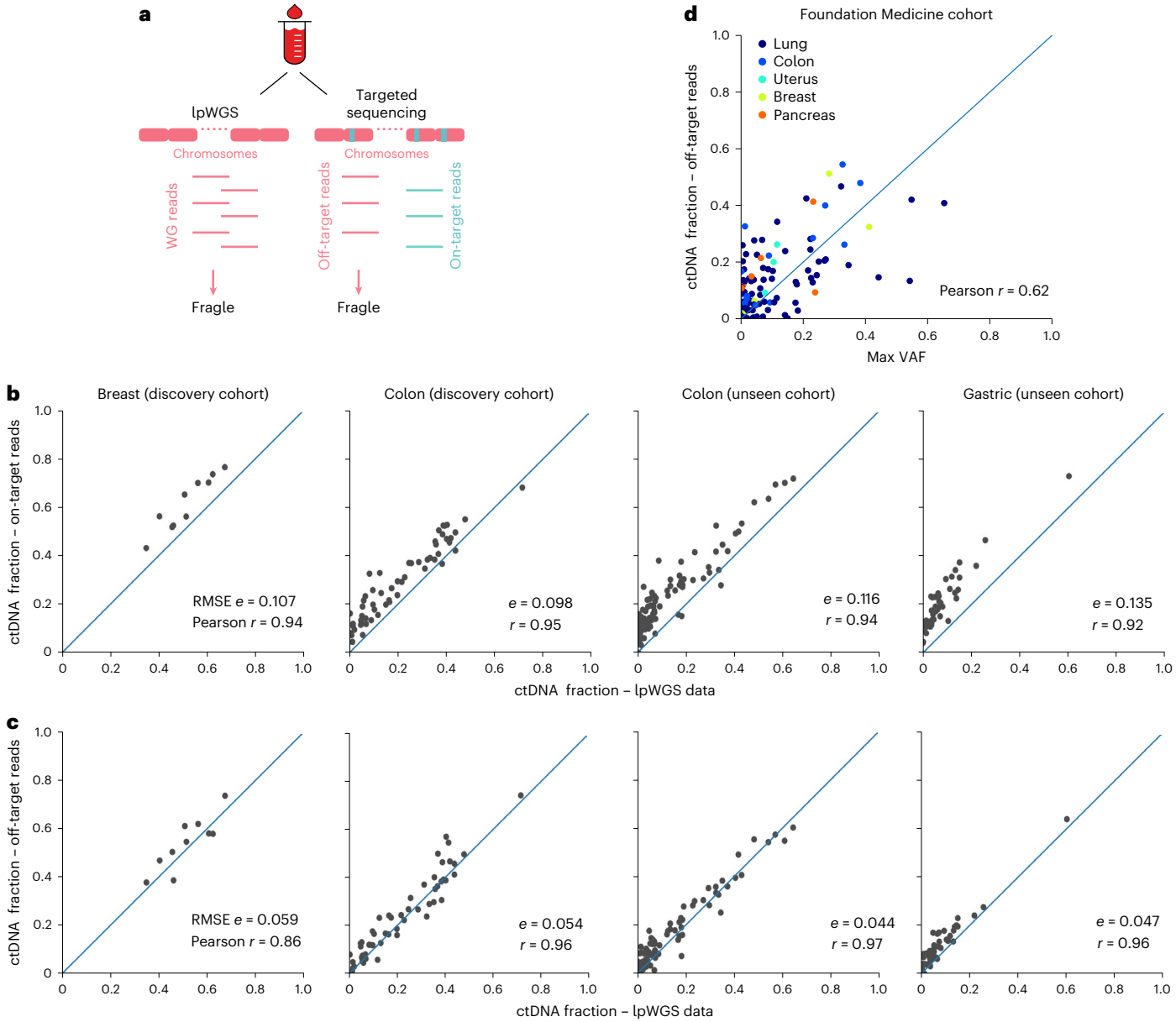


Fig. 4 | Application of Fragle to targeted sequencing data. a, Application of Fragle to samples having both lpWGS and targeted gene panel sequencing data. **b**, ctDNA levels predicted using lpWGS data and on-target reads from targeted sequencing samples. Concordance was assessed using the root mean square error (RMSE) and Pearson correlation coefficient. **c**, ctDNA levels predicted using lpWGS data and off-target reads from targeted sequencing samples. **d**, Targeted sequencing data generated with commercial liquid biopsy assay (Foundation Medicine, $n = 116$). Correlation of maximum VAFs (reported by the company, germline variants filtered) and Fragle-predicted ctDNA levels using off-target reads was shown.

Tracking ctDNA dynamics and disease progression from targeted sequencing

Having demonstrated that Fragle can accurately quantify ctDNA levels with targeted gene panel sequencing, we applied the method to longitudinal targeted sequencing samples from four patients with late-stage colorectal cancer. In these samples, we wanted to explore the temporal relationship between Fragle-estimated ctDNA dynamics and disease progression measured by RI. First, we observed strong temporal correlations between mutation VAFs and Fragle ctDNA levels across the longitudinal samples from the four patients (Fig. 5a–d). The first patient showed concordant and increasing VAFs and Fragle ctDNA levels, consistent with the emergence of progressive disease (PD) via RI at late timepoints (Fig. 5a). The second patient developed a partial response to FOLFOXIRI treatment, consistent with both reductions in VAFs and Fragle ctDNA levels (Fig. 5b). The next two patients showed a similar disease progression trajectory via RI, with initial stable disease evolving into PD following multiple rounds of treatment. ctDNA dynamics inferred by Fragle showed a consistent pattern of disease progression, with ctDNA levels remaining high at all timepoints (>10%; Fig. 5c,d). While the automated variant calling pipeline failed to detect mutations at late timepoints despite the presence of PD, manual inspection of sequencing reads at these positions confirmed the presence of TP53 and ATR mutations in these samples (4–5% VAFs; Supplementary Dataset 8). We finally considered a patient with metastatic colorectal cancer for whom we had collected 21 serial blood plasma samples over a cetuximab/chemotherapy treatment course of 3 years (Fig. 5e). In this patient, we observed an overall temporal correlation of Fragle-based ctDNA levels, mutation VAFs and treatment response determined from RI. However, the dynamic range of VAFs varied

https://doi.org/10.1038/s41551-025-01370-3

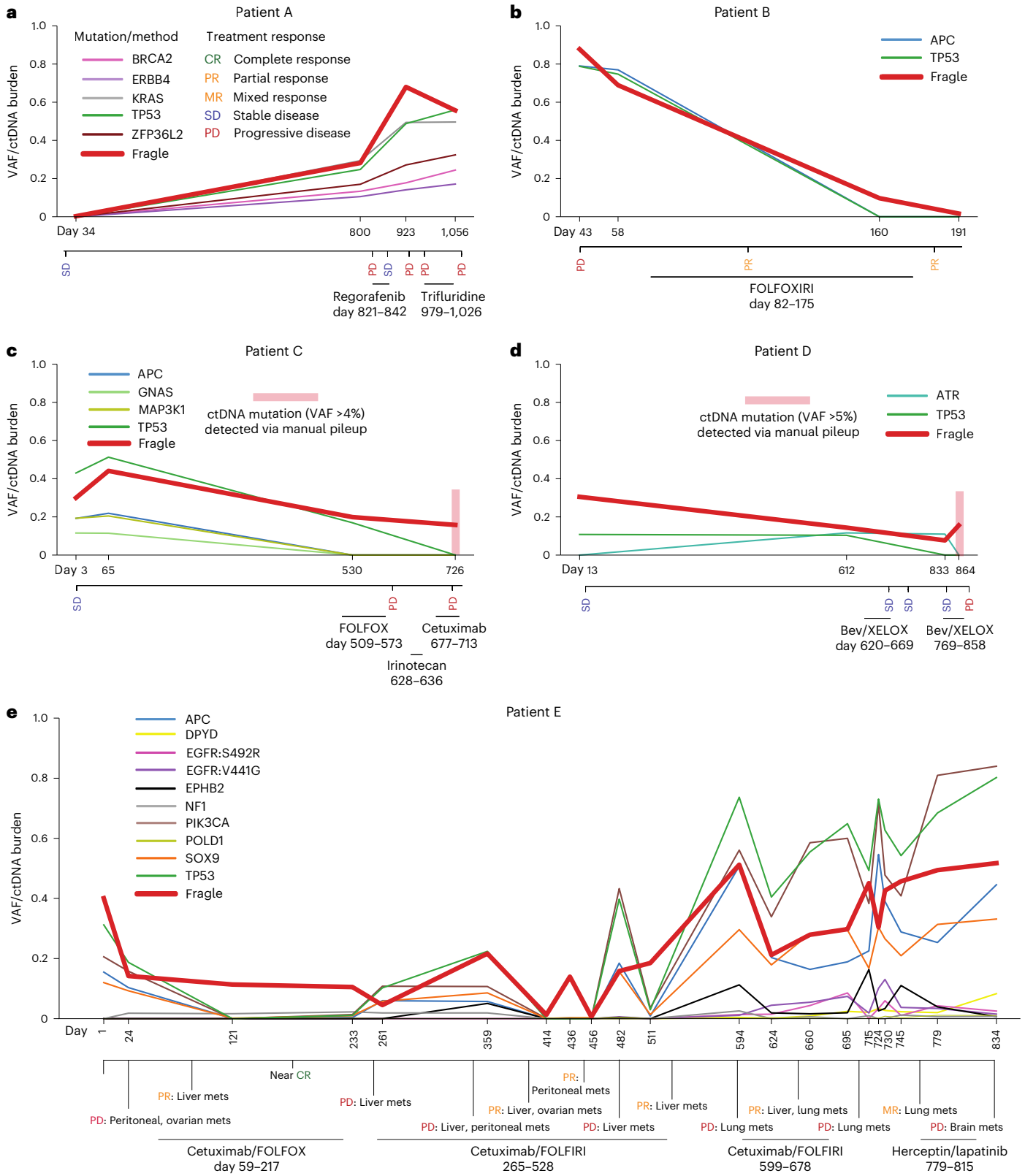


Fig. 5 | Monitoring of ctDNA levels and disease progression from targeted sequencing. a–**e**, Simultaneous longitudinal profiling of Fragle ctDNA levels and mutation VAFs in patients (patient A (**a**), patient B (**b**), patient C (**c**), patient D (**d**) and patient E (**e**)) with metastatic colorectal cancer using plasma targeted gene panel sequencing. Disease progression was captured with RI. Mets, metastasis. Only mutations detected in at least two timepoints for a given patient were included. Mutation VAFs were estimated using an automated pipeline, with manual pileup performed at highlighted timepoints where mutation detection failed.

https://doi.org/10.1038/s41551-025-01370-3

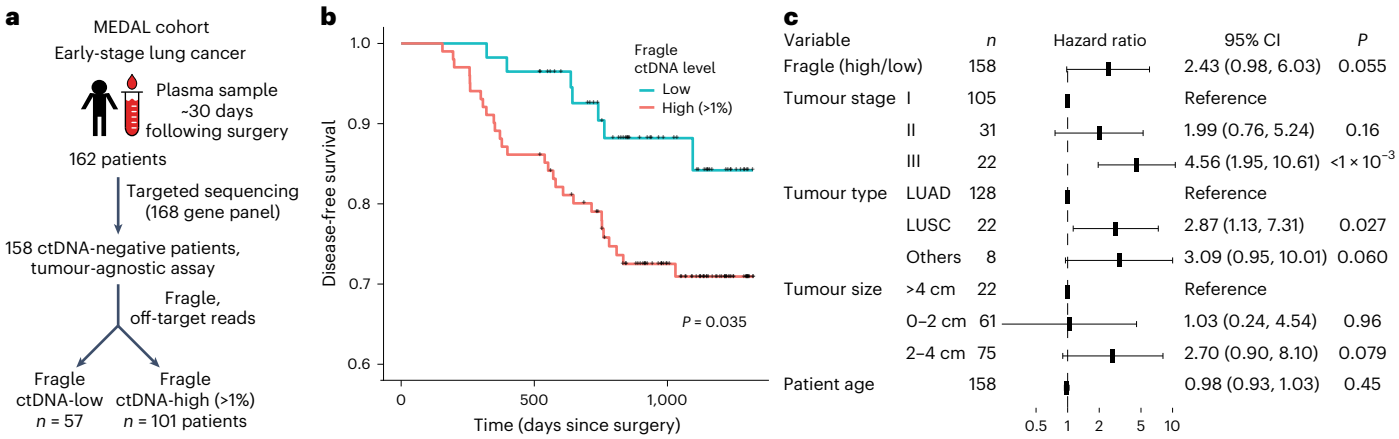


Fig. 6 | Risk stratification of patients with early-stage lung cancer. a, Fragle was used to predict ctDNA levels in 158 patients with early-stage lung cancer classified as ctDNA-negative with a tumour-agnostic targeted gene panel assay. Plasma samples were obtained at the landmark timepoint and Fragle was applied to the off-target reads to infer patients with high (>1%) and low ctDNA levels. **b**, In the 158 ctDNA-negative patients inferred with the targeted gene panel, disease-free survival was evaluated for patients with high and low Fragle ctDNA levels and compared using a log-rank test. **c**, A multivariate Cox proportional hazards model was used to evaluate the association between Fragle ctDNA levels and disease-free survival while controlling for other clinical variables. LUAD and LUSC refer to lung adenocarcinoma and lung squamous cell carcinoma, respectively.

extensively across different mutations and timepoints, highlighting the challenge in estimating absolute ctDNA levels from VAFs. For example, the patient had mutations in APC and TP53, two common clonal driver mutations in colorectal cancer. The VAFs for these two mutations differed markedly, with TP53 mutation allele frequencies more than twofold higher at many timepoints (such as days 779 and 834). In these samples, Fragle provided an orthogonal and independent measure of ctDNA levels. Overall, these data demonstrate high concordance to Fragle-estimated ctDNA levels and disease progression estimated from RI. Second, they outline how Fragle could be used to interpret and resolve heterogeneous and variable mutation VAFs profiled with targeted sequencing assays.

Risk stratification for patients with early-stage lung cancer

Blood-based detection of MRD following treatment has the potential to improve risk stratification and management strategies for patients with cancer[26,27]. Given the ~1% LoD for Fragle, we explored if the method could be used for tumour-naive MRD screening, with no requirements for a matching tissue sample. We obtained targeted sequencing data from a published cohort (MEDAL) of 162 patients with early-stage lung cancer that had plasma samples collected at the landmark timepoint (~30 days following curative surgery)[28]. In this study, plasma samples were subjected to a commercial tumour-agnostic targeted sequencing assay, and the authors classified samples into ctDNA-positive ($n = 4$) and ctDNA-negative ($n = 158$) groups based on mutation VAFs (Fig. 6a). In the ctDNA-negative samples, we used Fragle to further sub-classify the samples into ctDNA-high (>1% ctDNA level, $n = 101$) and low (≤1%, $n = 57$) groups. Intriguingly, despite these samples being classified as ctDNA-negative based on mutation VAFs in the targeted sequencing assay, the Fragle ctDNA-high group demonstrated significantly worse outcomes ($P = 0.035$, log-rank test) (Fig. 6b). Using a multivariate model, the association between Fragle ctDNA levels and outcomes was preserved ($P = 0.055$, Cox proportional hazards model) while controlling for known clinical prognostic variables such as tumour type and stage (Fig. 6c). Overall, these data demonstrate the potential clinical utility of Fragle as a supplement to a standard tumour-agnostic targeted sequencing assay. While Fragle was developed as a ctDNA quantification tool, these results also demonstrate that Fragle could be useful in certain settings where the detection of ctDNA is paramount, such as MRD detection and risk stratification without a matching tissue sample.

Discussion

While previous studies have explored how cfDNA fragment length signatures can be used to classify plasma samples from patients with cancer and healthy individuals[16–21], it remained unknown whether these fragmentomic signatures could also allow for accurate quantification of ctDNA levels in a blood sample. Here we developed Fragle, a multi-stage machine learning model that quantifies ctDNA levels directly from the cfDNA fragment length density distribution, with no requirement for tumour biopsy or matched normal sample. Fragle leveraged fragmentomic features common across multiple cancer types to robustly quantify ctDNA in patients with cancer, and its development and validation involved analysing lpWGS data from eight cancer types and targeted sequencing data from six cancer types. Specifically, using an in silico data augmentation approach, we trained and evaluated Fragle on around four thousand lpWGS samples spanning multiple cancer types and healthy cohorts. Using both in vitro and in silico dilution data from unseen samples, Fragle demonstrated accurate quantification of plasma ctDNA levels with a lower LoD than the current state-of-the-art approaches for ctDNA quantification using lpWGS data. We note that Fragle has been developed and validated exclusively with whole-genome and targeted cfDNA sequencing data; further studies would be needed to evaluate if Fragle could be applied to other sequencing modalities such as bisulfite sequencing data. Moreover, modelling distinct orthogonal fragmentomic features alongside copy number profiles could unlock new opportunities to further enhance quantitative ctDNA profiling methods.

Fragle reports accurate ctDNA quantification directly from a targeted sequencing assay. Existing methods developed for ctDNA quantification are not directly compatible with targeted sequencing data, requiring either lpWGS data[7], DNA methylation data[8,9] or modifications to the targeted sequencing panel[10]. Using colon, breast and gastric cancer plasma samples sequenced with both lpWGS and targeted gene panels, we demonstrate high concordance of Fragle estimates across assays. Furthermore, we demonstrated enhanced accuracy when input data were limited to the off-target reads from the targeted assay. Interestingly, while off-target reads are often filtered and ignored in a targeted sequencing experiment, these reads generally spread across the whole genome potentially mimicking ultra-lpWGS data[29]. We used this feature to analyse longitudinal targeted sequencing samples from patients with colorectal cancer,

demonstrating strong concordance of Fragle-inferred ctDNA dynamics and tumour progression measured from RI. This analysis also highlighted patients where the dynamic range of mutation VAFs varied extensively across different mutations and timepoints. Under these conditions, ctDNA quantification using Fragle could provide an orthogonal approach to interpret heterogeneous mutation VAFs profiled with targeted sequencing.

We also explored the potential for detecting MRD with Fragle. In a cohort of patients with early-stage lung cancer evaluated at the landmark timepoint following surgery, ctDNA levels estimated by Fragle could risk-stratify patients that had otherwise been classified as ctDNA-negative using a commercial tumour-agnostic targeted sequencing assay. This result highlights the potential clinical utility of Fragle for MRD classification in settings where tumour-informed sequencing assays are not feasible or available. While tumour-informed ctDNA detection approaches offer increased MRD detection sensitivity and accuracy[28,30], these methods impose additional requirements for tissue sample availability, sequencing, computing and logistics. By contrast, a tumour-naive MRD classification approach could be applied directly to a plasma sample. Our analysis demonstrates how Fragle has potential to enhance the baseline risk stratification provided by a standard tumour-naive targeted sequencing panel.

Fragle showed robust performance across plasma samples from ten solid tumour types and distinct healthy cohorts. We observed strong concordance with RI and tumour VAFs in longitudinal samples from patients with colorectal cancer undergoing targeted and cytotoxic therapy. These results suggest that the machine learning approach was able to learn properties of ctDNA fragmentation that generalize across cancer types and distinct therapeutic challenges. However, ctDNA release from healthy tissues and cells may vary over time and with different exposures. It will therefore be interesting to further evaluate fragmentomic signatures during distinct types and stages of anti-cancer therapeutic exposures. Since Fragle uses off-target reads to quantify ctDNA with targeted sequencing, we expect the method to generalize across distinct targeted sequencing panels. While we evaluated the method using multiple targeted gene panels, future studies are needed to characterize the performance using additional gene panels, unseen tumour types and therapeutic exposures. Future studies should also explore the extent that model performance could be improved using larger and evenly balanced training datasets comprising additional tumour types and cohorts. Clonal haematopoiesis of indeterminate potential (CHIP) is a known contributor of cfDNA fragments in some patients, with CHIP mutations reported to occur at ~1–2% VAFs[24]. While we demonstrated an ~1% LoD using in vitro and in silico diluted plasma samples, further evaluation in participants with confirmed high levels of CHIP is needed to determine if Fragle can robustly discriminate between fragmentomic signatures from CHIP and solid tumour cells.

Fragle is fast and flexible, estimating ctDNA levels in less than a minute using paired-end cfDNA profiling and without the need for a matching tumour or buffy coat sample. By also enabling orthogonal ctDNA quantification from targeted sequencing assays, the method could limit the need for running multiple assays for disease monitoring and interpretation of negative results from plasma genotyping[31]. This could enable simultaneous discovery of actionable cancer mutations and accurate estimation of ctDNA levels with a single assay. Overall, Fragle is a versatile and accurate method for profiling of ctDNA dynamics with potential for broad clinical utility.

Methods

Plasma sample collection and processing

The discovery cohort was composed of WGS plasma samples obtained from internal cohorts as well as from previous studies[10,12,16,32]. Similarly, the test cohort was composed of internal samples as well as samples from previous studies[32,33], all described in Supplementary Dataset 1. For new samples generated as part of this study, volunteers were recruited at the National Cancer Centre Singapore, under studies 2018/2709, 2018/2795, 2018/3046, 2019/2401 and 2012/733/B. Volunteers were also recruited from the National University Health System (NUHS). The studies were approved by the Singhealth Centralised Institutional Review Board. Written informed consent was obtained from all volunteers. Clinical data for the patients included in this study have been listed in Supplementary Dataset 9. Plasma was separated from blood within 2 h of venipuncture via centrifugation at 300 × *g* for 10 min and 9,730 × *g* for 10 min, and then stored at −80 °C. DNA was extracted from plasma using the QIAamp Circulating Nucleic Acid Kit following the manufacturer's instructions. Sequencing libraries were made using the KAPA HyperPrep kit (Kapa Biosystems, now Roche) following the manufacturer's instructions and sequenced on an Illumina NovaSeq6000 system. Low-pass WGS (~4×, 2 × 151 bp) was performed on cfDNA samples from patients with cancer and healthy individuals. We used BWA-MEM[34] to align WGS reads to the hg19 human reference genome.

Estimation of ctDNA fractions in the discovery cohort

We estimated the ctDNA fractions in the plasma samples of four cancer types using distinct orthogonal methods. Twelve CRC and 10 BRCA plasma samples had ~90× cfDNA and ~30× matched buffy coat WGS data, and their ctDNA fractions were estimated using 4 tumour tissue-based methods[35–38] as previously reported[10]. Fifty-three CRC samples had lpWGS and targeted nucleosome-depleted-region sequencing data, and we inferred ctDNA fractions by averaging ichorCNA and NDR-quant estimates[10] in these samples. The remaining 55 CRC and 57 BRCA samples only had lpWGS data and their ctDNA fractions were inferred using ichorCNA. The liver (HCC) and the ovarian cancer (OV) datasets only had lpWGS data and ichorCNA[7] was used to quantify ctDNA levels in these samples. The details of the estimation of ctDNA fractions are provided in Supplementary Dataset 2.

Discovery cohort data augmentation approach

After identifying the plasma samples with ctDNA level >3% and with at least 10 million fragments in the discovery cohort, we split the cancer samples (n = 164) into training and validation sets. We repeated this ten times, creating ten training–validation set pairs. We then diluted each cancer plasma sample with reads from a random control plasma sample to generate in silico spike-ins, followed by down-sampling to 10 million cfDNA fragments per sample. We generated in silico samples with variable ctDNA fractions ranging from 10^{-6} up to the undiluted fractions (Supplementary Dataset 3). To minimize information leakage to the validation set, we evenly split the healthy control samples (n = 101) into two sets. These two control sets were then used to dilute cancer samples in the training and validation sets, respectively (Extended Data Fig. 1).

Overview of Fragle

Fragle quantifies ctDNA levels from a cfDNA fragment length histogram. Using paired-end sequencing data, we computed the length of each sequenced cfDNA fragment, excluding duplicates and supplementary alignments and only keeping paired reads mapping to the same chromosome with a minimal mapping quality of 30. The machine learning model consists of two stages (see Fig. 1 and Supplementary Fig. 10 for details): a quantification and a model-selection stage. The quantification stage uses two sub-models: (i) the low-ctDNA burden sub-model and (ii) the high-ctDNA burden sub-model. These sub-models were designed and optimized to quantify accurately in low- (<3%) and high-ctDNA fraction (≥3%) samples, respectively. In the initial stage, for any given cfDNA sample, we individually input its processed fragment size profile into the two parallel sub-models. These two parallel sub-models, each with distinct loss functions,

focus on ctDNA quantification for low- and high-ctDNA samples, respectively. The two parallel sub-models independently output their estimated ctDNA fractions. In the second stage, a support vector machine (SVM) model selects the final predicted fraction from these two independent estimates. To train a final Fragle model based on the discovery cohort, we first trained a Fragle model on the training samples to obtain their ctDNA burden estimates. Among samples with ichorCNA-only ground-truth ctDNA estimates, we excluded samples with a large deviation of ctDNA fractions estimated by ichorCNA and Fragle (that is, relative difference >50% for samples with ctDNA fraction >20% by ichorCNA, >40% for samples with ctDNA fraction of 10–20%, and >30% for samples with ctDNA fraction of 3–10%). The final Fragle model was subsequently trained using all remaining samples in the discovery cohort. Notably, no samples were filtered out or selected from the unseen cohorts used to validate the final Fragle. As a result, Fragle remains entirely independent from these unseen cohorts (Fig. 1).

Fragle model feature extraction

The feature extraction steps have been illustrated in Supplementary Fig. 10. We computed the fragment length profile for all fragments sized 51–400 bp using paired-end reads with Pysam[39]. The length profile of each sample was normalized using the highest observed fragment length count, followed by $\log_{10}$ scaling of these sample-wise normalized counts. Next, a moving average normalization (z-score of 32-nt window) was performed sample-wise for smoothing. The transformed length features for a given sample were further standardized relative to the training set. To explore the fragmentomic feature space predictive of cancer, we identified fragment length intervals that differed between cancer samples and healthy individuals in the discovery cohort (Supplementary Fig. 11 and Supplementary Dataset 10). The most predictive length intervals comprised both short and long cfDNA fragments, including 125–140 bp, 170–208 bp and 246–306 bp ($P < 10^{-20}$, Wilcoxon rank sum test). Out of 350 fragment lengths, 281 showed significant differences between cancer and healthy samples ($P < 0.01$); these were selected as candidate length features for model development.

High- and low-ctDNA burden sub-model architecture

The model was implemented as a neural network with a feature embedding layer, 16 fully connected layers with batch normalization[40] and residual connections[41] (Supplementary Fig. 10). Dropout regularization[42] was used in between intermediate layers with a dropout rate of 30% to minimize model overfitting. A composite loss function was used combining a quantification loss and a binary cross-entropy loss. The main differentiating factor between the low- and high-ctDNA burden sub-models was the loss function. Here the high-ctDNA sub-model utilizes MAE loss, while the low-ctDNA burden model uses relative MAE loss:

$$\text{Relative MAE} = (\text{MAE} + \sigma)/(\text{true\_fraction} + \sigma) \quad (1)$$

We used relative MAE in the low-ctDNA fraction model to emphasize proportional errors rather than absolute errors. Relative MAE compares the error to the true value, making the loss less sensitive to larger targets and more focused on percentage-based differences. Relative MAE is thus more sensitive to prediction errors in low-ctDNA samples because these samples will tend to have larger percentage-based differences. The model heuristically determined the hyperparameter σ based on cross-validation data.

Selection model

The low- and high-ctDNA sub-models individually predict ctDNA fractions for each sample. These two predictions are used as input features for the selection model. The training sample ground truth is labelled as 0 or 1 when the expected ctDNA fraction is <3% or ≥3%, respectively. The selection model is a binary support vector machine classifier with radial basis kernel function.

ichorCNA and four-feature model benchmarking

We utilized ichorCNA according to its usage guidelines, using the default parameters to compute read count coverage with the HMMcopy Suite, followed by deducing tumour fractions with the ichorCNA R package. For the four-feature model, we extracted four features from the fragment length profile according to a previously published study[16]: 10-bp amplitude and proportions of fragments sized 160–180 bp, 180–220 bp and 250–320 bp. We used these four features to develop a random forest regression model for estimating ctDNA fractions.

Targeted sequencing assay

Plasma and patient-matched buffy coat samples were isolated from whole blood within 2 h from collection and were stored at −80 °C. DNA was extracted with the QIAamp Circulating Nucleic Acid Kit, followed by library preparation using the KAPA HyperPrep kit. All libraries were tagged with custom dual indexes containing a random 8-mer unique molecular identifier. Targeted capture was performed on the plasma samples in the unseen colorectal cancer dataset and in the unseen metastatic gastric cancer dataset, using an xGen custom panel (Integrated DNA Technologies) of 225 cancer driver genes. We also performed targeted sequencing of six plasma samples from healthy individuals to identify and blacklist unreliable variants likely attributed to sub-optimal probe design. Paired-end sequencing (2 × 151 bp) was done on an Illumina NovaSeq6000 system.

Variant calling

FASTQ files generated from targeted sequencing were preprocessed to append unique molecular identifiers (UMIs) into the fastq headers, followed by read alignment using BWA-MEM[34]. We then performed UMI-aware deduplication using the fgbio package (https://github.com/fulcrumgenomics/fgbio). We grouped reads with the same UMI, allowing for one base mismatch between UMIs, and generated consensus sequences by discarding groups of reads with single members. To identify single-nucleotide variants and small insertions/deletions in the cfDNA samples, we first performed variant screening using VarDict[43] using a minimal VAF threshold of 0.05%, and annotated all variants using Variant Effect Predictor[44]. We removed low-impact variants such as synonymous variants and low-quality variants such as those that fail to fulfil the minimum requirements of variant coverage, signal-to-noise ratio and number of reads supporting alternative alleles. Finally, we removed population SNPs found in Genome Aggregation Database (gnomAD) and 1000 Genomes. To further minimize false-positive variants, we used duplexCaller[45] to identify variants with double-strand support and discarded blacklisted variants that were recurrently found in the plasma of two or more healthy individuals. Finally, when available, we identified high-confidence variants by taking advantage of serial plasma samples collected from the same patient, keeping only variants that were detectable in at least two serial samples, with VAF more than 3% in at least one sample.

Application of Fragle to targeted sequencing data

Duplicates were removed from the targeted sequencing data using the Picard MarkDuplicates function (https://broadinstitute.github.io/picard/), and on-target and off-target reads were extracted from the BAM files using the samtools view function. The resulting reads were used to generate the input fragment length histograms as detailed above. We obtained targeted sequencing data of the plasma samples in the unseen colorectal cancer dataset ($n = 96$) and in the unseen metastatic gastric cancer ($n = 49$) dataset, based on a panel of 225 cancer driver genes, as described above. The targeted sequencing data for the colorectal and breast cancer datasets in the discovery cohort have been reported in the previous studies[10,46], based on a panel

of 100 genes of colorectal cancer mutation and a panel of 77 genes of breast cancer mutation, respectively. Summary statistics for all gene panels such as gene count, genomic coverage, target regions and on-target coverage ratio have been provided in the supplementary material (Supplementary Datasets 11–13). In an additional analysis, we used targeted sequencing data from 116 patients profiled with the Foundation Medicine Liquid CDx assay. Samples belonged to different cancer types, including lung, colon, breast, pancreas and uterus cancer. We filtered known germline variants using gnomAD (v4) and analysed VAFs using all remaining variants reported by the company. Since Fragle requires off-target BAM files for prediction, we constructed a targeted sequencing BED file using the 311 genes reported to comprise this panel (Supplementary Dataset 12).

Lung cancer survival analysis

Plasma targeted sequencing data from the MEDAL cohort (project ID OEP004204)[47] was retrieved from National Omics Data Encyclopedia (NODE). Alignment to the human genome (hg19) was conducted using BWA-MEM[34]. Duplicates in the aligned data were marked using Samblaster[48]. Putative target regions were identified by calculating the median coverage per base from a subset of randomly selected BAM files ($n = 38$). Coverage of regions without any reads was reported as zero. Next, the resulting consensus bedgraph file was segmented into 100 bp bins. Bins with median coverage exceeding 2× were selected and merged if they were within 100 bp of each other, to form contiguous regions. The resulting BED file was used for obtaining the off-target BAM file for each sample using samtools.

Unseen in vitro dilution experiments

The first in vitro dilution experiment included high-ctDNA burden cfDNA samples from two individual patients with CRC, which were selected to create a starting point for the dilution series. The ctDNA fraction for each sample was determined by two methods (ichorCNA[7] and NDRquant[10]), with high concordance across methods (sample 1, ctDNA content 38% by ichorCNA and 37% by NDRquant; sample 2, 39% and 38%, respectively). Commercial pooled cfDNA from healthy volunteers (0% ctDNA) was purchased from PlasmaLab (lot numbers 2001011 and 210302) and was used to set up a nine-point serial dilution of the ctDNA fraction for each sample, with three technical replicates per dilution point. The second in vitro dilution experiment started from 3 plasma samples of gastric cancer that had concordant ctDNA estimates between methods (sample 1, ctDNA content 7.6% by ichorCNA and 8.7% by Fragle; sample 2, 17.2% and 14.7%; sample 3, 13.4% and 19.1%, respectively). Twelve control plasma cfDNA samples were purchased from Ripple Biosolutions and were used to set up a seven-point serial dilution of the ctDNA fraction for each cancer sample, with three technical replicates. We randomly selected three control samples and pooled them before diluting the plasma of cancer. lpWGS was performed with a depth of ~4–5×.

Unseen in silico dilution experiments

Our unseen in silico dilution experiment included 7 unseen breast and 13 unseen colon cancer samples each containing high and concordant ctDNA fraction estimates based on Fragle and ichorCNA (>10% ctDNA based on both methods with a relative difference <5%). We prepared 20 healthy mixtures, each created by pooling 3 random samples from an unseen control cohort. A 6-point serial dilution for each sample was set up using these healthy mixtures to dilute the 20 cancer samples, with 20 technical replicates. A total of 2,400 dilution samples were created ranging from 5% to as low as ~0.1% ctDNA fraction, each dilution point containing 400 samples.

Reporting summary

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

Data availability

The published data used in this study and their access codes are provided in Supplementary Dataset 1. Data generated in this study are available from the European Genome-Phenome Archive (EGA; dataset ID EGAD50000000167). The data are available under restricted access and will be released subject to a data-transfer agreement. Source data are provided with this paper.

Code availability

The Fragle software is available at https://github.com/skandlab/FRAGLE. The software can be directly applied to lpWGS/off-target BAM files aligned to hg19/GRCh37/hg38 reference genomes without any preprocessing.

References

1. Lui, Y. Y. et al. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin. Chem.* **48**, 421–427 (2002).
2. Pantel, K. & Alix-Panabières, C. Liquid biopsy and minimal residual disease—latest advances and implications for cure. *Nat. Rev. Clin. Oncol.* **16**, 409–424 (2019).
3. Sanz-Garcia, E., Zhao, E., Bratman, S. V. & Siu, L. L. Monitoring and adapting cancer treatment using circulating tumor DNA kinetics: current research, opportunities, and challenges. *Sci. Adv.* **8**, eabi8618 (2022).
4. Kilgour, E., Rothwell, D. G., Brady, G. & Dive, C. Liquid biopsy-based biomarkers of treatment response and resistance. *Cancer Cell* **37**, 485–495 (2020).
5. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. *Nature* **575**, 299–309 (2019).
6. Razavi, P. et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* **25**, 1928–1937 (2019).
7. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
8. Li, S. et al. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proc. Natl Acad. Sci. USA* **120**, e2305236120 (2023).
9. Li, W. et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* **46**, e89 (2018).
10. Zhu, G. et al. Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. *Nat. Commun.* **12**, 2229 (2021).
11. Lo, Y. M. D. et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
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. Underhill, H. R. et al. Fragment length of circulating tumor DNA. *PLoS Genet.* **12**, e1006162 (2016).
14. Mouliere, F. et al. High fragmentation characterizes tumour-derived circulating DNA. *PLoS ONE* **6**, e23418 (2011).
15. Nguyen, T. H. et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. *eLife* **12**, RP89083 (2023).
16. Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
17. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
18. Foda, Z. H. et al. Detecting liver cancer using cell-free DNA fragmentomes. *Cancer Discov.* **13**, 616–631 (2023).

19. Renaud, G. et al. Unsupervised detection of fragment length signatures of circulating tumor DNA using non-negative matrix factorization. *eLife* **11**, e71569 (2022).
20. Yu, S. C., Choy, L. L. & Lo, Y. D. 'Longing' for the next generation of liquid biopsy: the diagnostic potential of long cell-free DNA in oncology and prenatal testing. *Mol. Diagn. Ther.* **27**, 563–571 (2023).
21. Hudecova, I. et al. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. *Genome Res.* **32**, 215–227 (2022).
22. Mathios, D. et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat. Commun.* **12**, 5060 (2021).
23. Esfahani, M. S. et al. Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597 (2022).
24. Ptashkin, R. N. et al. Prevalence of clonal hematopoiesis mutations in tumor-only clinical genomic profiling of solid tumors. *JAMA Oncol.* **4**, 1589–1593 (2018).
25. Woodhouse, R. et al. Clinical and analytical validation of FoundationOne Liquid CDx, a novel 324-Gene cfDNA-based comprehensive genomic profiling assay for cancers of solid tumor origin. *PLoS ONE* **15**, e0237802 (2020).
26. Audinot, B. et al. ctDNA quantification improves estimation of outcomes in patients with high grade osteosarcoma: a translational study from the OS2006 trial. *Ann. Oncol.* **35**, 559–568 (2024).
27. Bratman, S. V. et al. Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. *Nat. Cancer* **1**, 873–881 (2020).
28. Chen, K. et al. Individualized tumor-informed circulating tumor DNA analysis for postoperative monitoring of non-small cell lung cancer. *Cancer Cell* **41**, 1749–1762. e1746 (2023).
29. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
30. Zviran, A. et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* **26**, 1114–1124 (2020).
31. Tsui, D. W. et al. Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. *Genome Med.* **13**, 1–15 (2021).
32. Jiang, P. et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
33. Yu, P. et al. Multi-dimensional cell-free DNA-based liquid biopsy for sensitive early detection of gastric cancer. *Genome Med.* **16**, 79 (2024).
34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
35. Bao, L., Pu, M. & Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* **30**, 1056–1063 (2014).
36. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
37. Larson, N. B. & Fridley, B. L. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29**, 1888–1889 (2013).
38. Oesper, L., Satas, G. & Raphael, B. J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**, 3532–3540 (2014).
39. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* 448–456 (PMLR, 2015).
41. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (CVPR, 2026).
42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
43. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
44. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).
45. Mansukhani, S. et al. Ultra-sensitive mutation detection and genome-wide DNA copy number reconstruction by error-corrected circulating tumor DNA sequencing. *Clin. Chem.* **64**, 1626–1635 (2018).
46. Kleftogiannis, D. et al. Detection of genomic alterations in breast cancer with circulating tumour DNA sequencing. *Sci. Rep.* **10**, 16774 (2020).
47. Chen, K. et al. Individualized dynamic methylation-based analysis of cell-free DNA in postoperative monitoring of lung cancer. *BMC Med.* **21**, 255 (2023).
48. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

Acknowledgements

This work was supported by the Agency for Science, Technology and Research under its IAF-PP programme (grant ID H1801a0019) and the Singapore Ministry of Health's National Medical Research Council under its OF-IRG programme (OFIRG21nov-0083). This work makes use of data from the Chinese University of Hong Kong (CUHK) Circulating Nucleic Acids Research Group as reported previously (https://doi.org/10.1073/pnas.1500076112 and https://doi.org/10.1158/2159-8290.CD-19-0622) and data from CRUK_CI, University of Cambridge, Rosenfeld Lab, as reported previously (https://doi.org/10.1126/scitranslmed.aat4921). We gratefully acknowledge D. Lo and his research group at CUHK, as well as N. Rosenfeld and his research group at University of Cambridge, for providing access to cfDNA cohorts.

Author contributions

A.J.S. supervised the project. A.J.S. and G.Z. conceived the project. A.J.S., G.Z. and C.R.R. performed most of data analysis and model development. V.G., D.O., P.B., H.C., A.J.L., Y.A.G., Z.W.P., N.L.S., A.A., Y.C., L.N.L., D.H., S.T. and L.W. assisted in data analysis. V.G., P.B. and A.A. assisted in model development. P.P., Y.T.L., A.G. and S.N. performed the experiments. S.-L.K., D.Q.C., B.T., T.J.T., Y.S.Y., A.Y.C., M.C.H.N., P.T., D.T., P.M.W. and I.B.T. provided samples and clinical information. A.J.S., G.Z. and C.R.R. wrote the paper. All authors reviewed and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

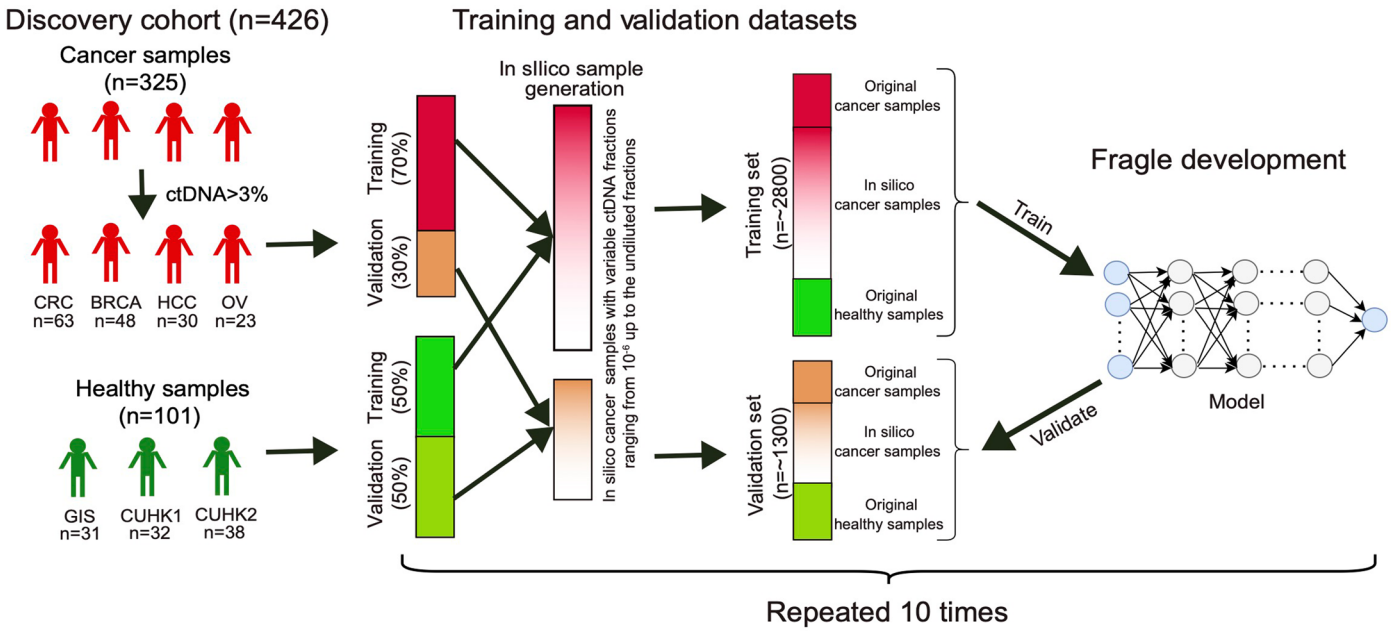
Extended data is available for this paper at https://doi.org/10.1038/s41551-025-01370-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-025-01370-3.

Correspondence and requests for materials should be addressed to Anders Jacobsen Skanderup.

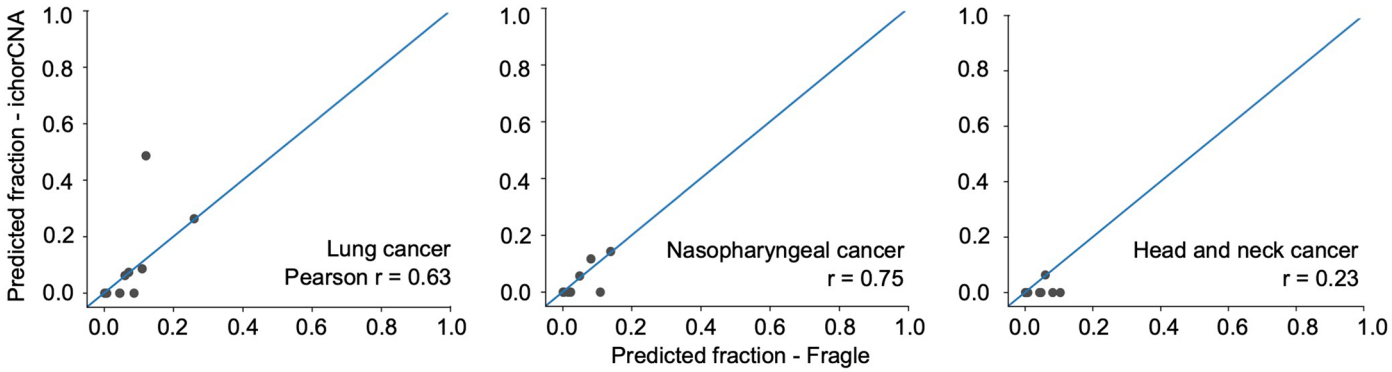
Peer review information *Nature Biomedical Engineering* thanks Le Son Tran and Zhihong Zhang for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41551-025-01370-3



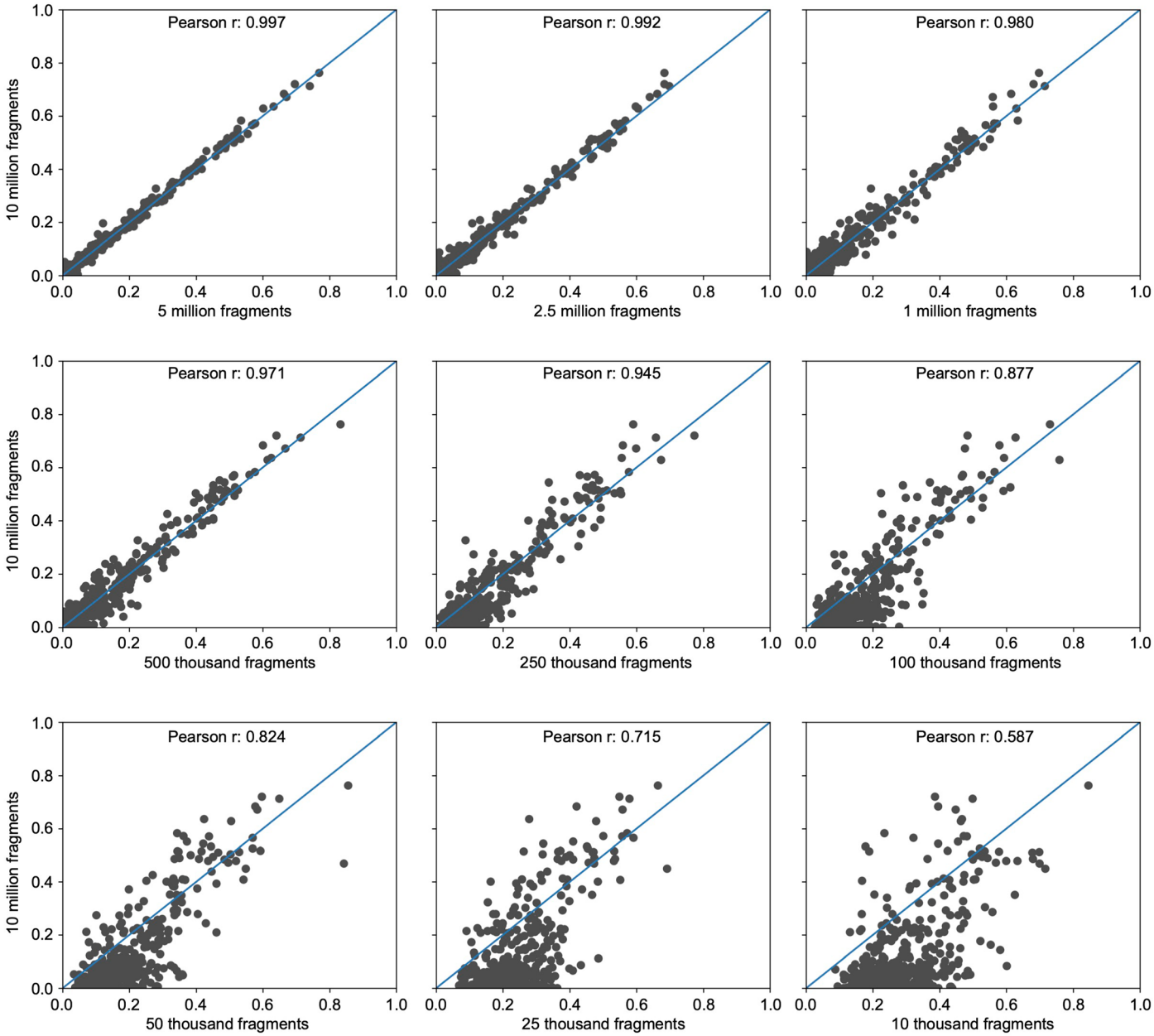
Extended Data Fig. 1 | Schematic showing how the predictive models were trained and validated using the healthy control samples, original cancer samples, and in silico cancer spike-ins in the discovery cohort.

https://doi.org/10.1038/s41551-025-01370-3



Extended Data Fig. 2 | Comparison between Fragle and ichorCNA in their predicted ctDNA fractions in unseen test samples from patients with lung, nasopharyngeal, as well as head and neck cancers.

https://doi.org/10.1038/s41551-025-01370-3



Extended Data Fig. 3 | Comparison of the Fragle-predicted ctDNA fractions in the unseen cfDNA WGS samples with 10 million cfDNA fragments and their downsampled counterparts.

nature portfolio

Corresponding author(s): Anders Skanderup

Last updated by author(s): Feb 2, 2025

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Sequencing data were obtained in fastq format using Illumina instruments.
Data analysis	Machine-learning modelling: Pytorch Read filtering and feature generation: Samtools, Picard, Samblaster, and Pysam Alignment: bwa Estimation of ctDNA fraction: theta2, titanCNA, abscnseq, PurBayes, ichorCNA, and NDRquant Plotting and significance analysis: Seaborn, Matplotlib, and R Variant-calling analysis: fgbio package, VarDict, Variant Effect Predictor, duplexCaller, and gnomAD

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

Data

Policy information about availability of data

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

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

The published data used in this study and their access codes are provided in Supplementary dataset 1. Data generated in this study are available from the European Genome-phenome Archive (EGA; Dataset ID: EGAD50000000167). The data are available under restricted access and will be released subject to a data-transfer agreement.

Research involving human participants, their data, or biological material

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

Reporting on sex and gender	The study did not involve sex-based nor gender-based analyses.
Reporting on race, ethnicity, or other socially relevant groupings	The study did not involve analyses relevant to race, ethnicity or other socially relevant groupings.
Population characteristics	Information for the patients (such as gender and cancer type) included in this study is available as Supplementary dataset 9.
Recruitment	Volunteers were recruited at the National Cancer Centre Singapore.
Ethics oversight	Volunteers were recruited at the National Cancer Centre Singapore, under studies 2018/2709, 2018/2795, 2018/3046, 2019/2401 and 2012/733/B. Volunteers were also recruited from the National University Health System (NUHS). The studies were approved by the Singhealth Centralised Institutional Review Board. Written informed consent was obtained from all volunteers.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We analysed 658 cfDNA lpWGS samples of 8 cancer types and 158 cfDNA lpWGS samples of healthy controls, as listed in Supplementary dataset 1.
Data exclusions	There were no data exclusions.
Replication	The training/validation splitting and in-silico-mixture-generation experiment were performed 10 times to ensure the robustness of the reported results.
Randomization	The study was not a randomized study; hence, no randomization was performed.
Blinding	Blinding was not applicable to the study, because plasma samples from all cancer patients were included for model development and validation regardless of treatment protocol.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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