Fragle: Universal ctDNA quantification using deep learning of fragmentomic profiles

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

1 Genome Institute of Singapore (GIS), Agency for Science, Technology and Research, Singapore
2 School of Computing, National University of Singapore, Singapore
3 Institute of Data Science, National University of Singapore, Singapore
4 Integrative Sciences and Engineering Programme, Graduate School, National University of Singapore, Singapore
5 Duke-NUS Medical School, National University of Singapore, Singapore
6 National Cancer Center Singapore, Singapore
7 Singapore General Hospital, Singapore

^ These authors contributed equally to this work.
* Correspondence: Anders Jacobsen Skanderup: skanderupamj@gis.a-star.edu.sg

Abstract

Quantification of circulating tumor DNA (ctDNA) levels in blood enables non-invasive surveillance of cancer progression. Fragle is an ultra-fast deep learning-based method for quantification of ctDNA levels directly from cell-free DNA fragmentomic profiles with no requirement of tumor biopsy or matched normal sample. Fragle was trained on a large collection of low-pass whole-genome sequencing (lpWGS) plasma samples from cancer patients with distinct solid tumor types and healthy cohorts. Separate deep learning models designed for low and high ctDNA burden samples, coordinated by a selection model, enabled robust quantification in both low and high ctDNA burden samples. We tested the accuracy and robustness of Fragle across samples from 7 cancer types and healthy control cohorts from different research groups. Fragle demonstrated a lower limit of detection (~1%) compared to existing tumor-naïve approaches based on lpWGS data. Uniquely, we show that Fragle can also predict ctDNA levels from targeted sequencing data, enabling universal quantification of blood ctDNA levels across sequencing assays and cancer patients. Overall, Fragle is a versatile, fast, and accurate method for ctDNA quantification in blood samples.

Introduction

The death of normal non-malignant cells, primarily of the hematopoietic lineage, releases cell-free DNA (cfDNA) into the blood circulation [1]. In cancer patients, the blood plasma also carries circulating tumor DNA (ctDNA), which can be profiled to enable non-invasive diagnostics and disease surveillance [2]. The ability to monitor tumor 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, 4].

Ultra-deep targeted cfDNA sequencing assays are often preferred in the clinic due to their ability to identify actionable mutations. While mutation variant allele frequencies (VAFs) can be used to approximate ctDNA levels, not all tumors 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, as well as potential confounding noise from clonal hematopoiesis mutations which originate from non-cancerous cells [5]. Existing methods developed for ctDNA quantification are not compatible with targeted sequencing data, requiring either low-pass whole genome sequencing (lpWGS) data [6], DNA methylation data [7], or modifications to the targeted sequencing panel [8]. Thus, there is an unmet need to develop accurate and orthogonal approaches for ctDNA quantification that can generalize across patients, tumor 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 display increased protection from DNA degradation [9]. cfDNA fragments from cancer patients tend to be shorter and more variably sized than those from healthy individuals [10-12]. These observations have motivated studies exploring how cfDNA fragment length properties can be used to classify cfDNA samples from cancer patients and healthy individuals [12-15]. Here, we developed *Fragle*, a multi-stage machine learning model that quantifies ctDNA levels directly from a cfDNA fragment length density distribution. Using an in silico data augmentation approach, we trained and evaluated Fragle on more than 2000 lpWGS samples from four cancer types and three healthy cohorts. We further evaluated the accuracy and the limit of detection (LoD) across independent cohorts and cancer types. We also demonstrated that Fragle can be applied to cfDNA fragmentomic profiles obtained from targeted sequencing panels, making Fragle the first ctDNA quantification method that works across both lpWGS and targeted sequencing assays. Finally, we applied Fragle to longitudinal targeted sequencing cfDNA samples collected across multiple late-stage colorectal cancer patients, demonstrating that Fragle ctDNA quantification is highly correlated with treatment response measured through radiographic imaging.

Results

Quantitative prediction of ctDNA levels from fragmentomic data

We assembled a discovery cohort comprising lpWGS data from 213 cancer plasma samples (4 tumor types) and 101 plasma samples from healthy individuals (Fig. 1, Suppl. Data 1). We only used cancer samples with detectable ctDNA (≥3%, Suppl. Fig. 1), with ground truth ctDNA levels estimated using a consensus across distinct methods (Methods, Suppl. Data 2). Using a large-scale data augmentation approach, we generated 2374 mixture samples from cancer and healthy samples with variable ctDNA fractions (Methods, Fig. 1, Suppl. Data 3). To explore how cfDNA fragment length distributions could predict for 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 (Fig. 1, Methods). The transformed fragment length distributions, in combination with their labels in the form of predicted ctDNA levels, served as input to a multi-stage supervised machine learning approach. We used deep neural networks to enable modelling of position-specific dependencies in the fragment length distributions. Two parallel deep learning-based sub-models were optimized for quantitative prediction in high (≥3%) and low (<3%) ctDNA burden samples, respectively. The two sub-models differed with respect to loss function, model architecture, as well as regions used to generate the input fragment length features. Finally, we trained a support vector machine (SVM) model to select between the output from the high and low ctDNA burden models as the final model prediction (Methods).

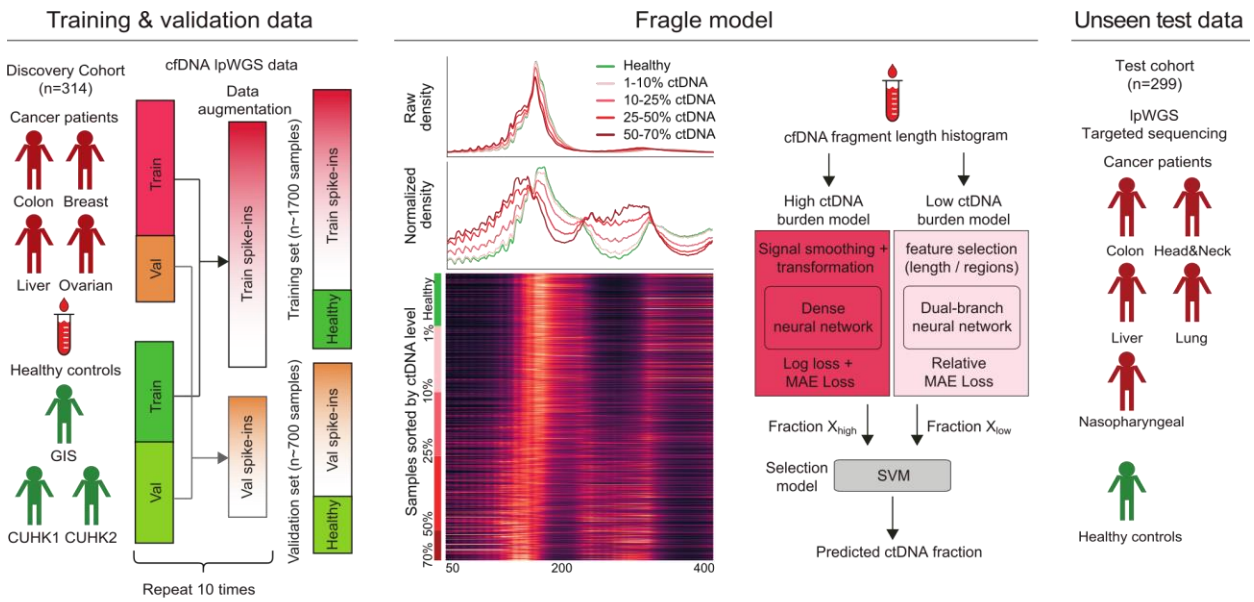


Fig. 1, Overview of Fragle. Fragle is a multi-stage machine learning-based model that estimates the ctDNA level in blood sample from the cfDNA fragment length density distribution. Fragle was trained using a large-scale data

augmentation and cross-validation approach, and was further tested using unseen samples from multiple cancer types and healthy control cohorts.

The model was trained and evaluated using cross-validation, demonstrating high predictive accuracy on validation samples across all 4 cancer types (Fig. 2a): Colorectal (mean absolute error (MAE) = 3.8%; Pearson *r* = 0.92), breast (MAE = 4.3%; *r* = 0.94), liver (MAE = 3.0%; *r* = 0.87), and ovarian cancer (MAE = 4.4%; *r* = 0.67). The lower concordance for ovarian cancer could be attributed to in silico samples from two outlier patients, removal of whom increased the correlation to 0.90 (Suppl. Fig. 2).

We then trained the model on all samples in the discovery cohort and tested its performance on additional cohorts of unseen plasma lpWGS samples. We observed strong correlation between Fragle and ichorCNA-based ctDNA fraction estimates across an unseen cohort of colorectal cancer (*r* = 0.82; *P* = 2.6e-48; N = 190; Fig. 2b), an unseen cohort of liver cancer (*r* = 0.90; *P* = 4.1e-13; N = 34; Fig. 2c), and a mixed cohort of cancer types not included in the discovery set (Pearson *r* = 0.89, 0.67, and 0.35 for nasopharyngeal, head and neck, and lung cancer, respectively; combined *P* = 0.002; N = 30; Suppl. Fig. 4). In the unseen colorectal cancer cohort, we also performed targeted gene sequencing and identified high-confidence somatic mutations in 86 samples (Suppl. Data 4-6), demonstrating high concordance between mutation VAFs and Fragle-predicted ctDNA levels (*r* = 0.87; *P* = 1.1e-27; Fig. 2d).

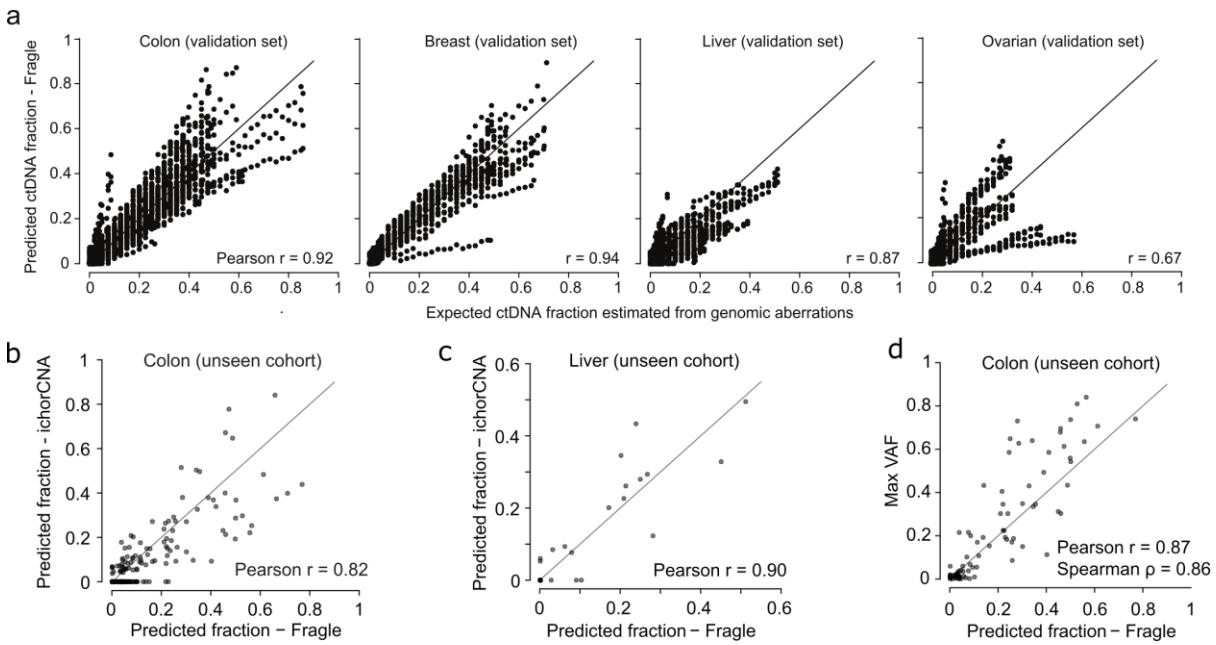


Fig. 2 Evaluation of the quantitative measurement of ctDNA. a) Comparison between expected and predicted ctDNA levels for colorectal (CRC), breast (BRCA), liver (HCC), and ovarian (OV) cancer samples in the validation sets. **b, c**) Validation of Fragle in unseen cohorts, comparing ichorCNA and Fragle predicted ctDNA levels in test

samples from **b**) colorectal cancer (N = 190) and **c**) liver cancer patients (N = 34). **d**) Comparison of Fragle predicted ctDNA levels and maximum VAFs in colorectal cancer patients with detectable somatic mutation (N = 86).

Determination of the lower limit of detection

The model predicted very low ctDNA fractions (median = 0.07%) for the healthy samples in the validation sets, with 94% of healthy samples predicted to comprise <1% ctDNA (Fig. 3a). Furthermore, the model could quantitatively differentiate healthy from low-ctDNA samples at around 0.1%-1% ctDNA level (Fig. 3a), suggesting a ~1% lower limit of detection (LoD). We also examined the performance of Fragle for classification of healthy and cancer samples in the validation sets. Using cancer samples with a diluted ctDNA level ≥1% in the validation sets, Fragle demonstrated an area under the curve (AUC) of 0.93 (Fig. 3b), improving on the performance of ichorCNA (AUC = 0.88) on these samples. Expectedly, the AUC increased further after filtering out low-ctDNA burden samples (Suppl. Fig. 3).

We further evaluated the LoD using unseen test samples. We used 86 cfDNA samples from CRC patients with detectable mutations as positive cancer samples and 45 unseen healthy plasma samples as negatives (Suppl. Data 7-8). Using a previously reported approach [6], Fragle demonstrated an AUC of 0.91 at 1% LoD level, outperforming ichorCNA on the same set of samples (AUC = 0.84, Fig. 3c). These results were further confirmed using an in vitro dilution experiment comprising serial dilution of two high-ctDNA burden plasma samples of CRC patients using pooled cfDNA from healthy individuals (Methods). Across three technical replicates, Fragle accurately predicted ctDNA fractions for both patients at 2-3% ctDNA level and consistently predicted <1% in all <1% ctDNA samples (Fig. 3d). In comparison, ichorCNA failed to detect ctDNA in 3/6 2% ctDNA samples and 2/6 3% ctDNA samples (Suppl. Fig. 5; Suppl. Data 9).

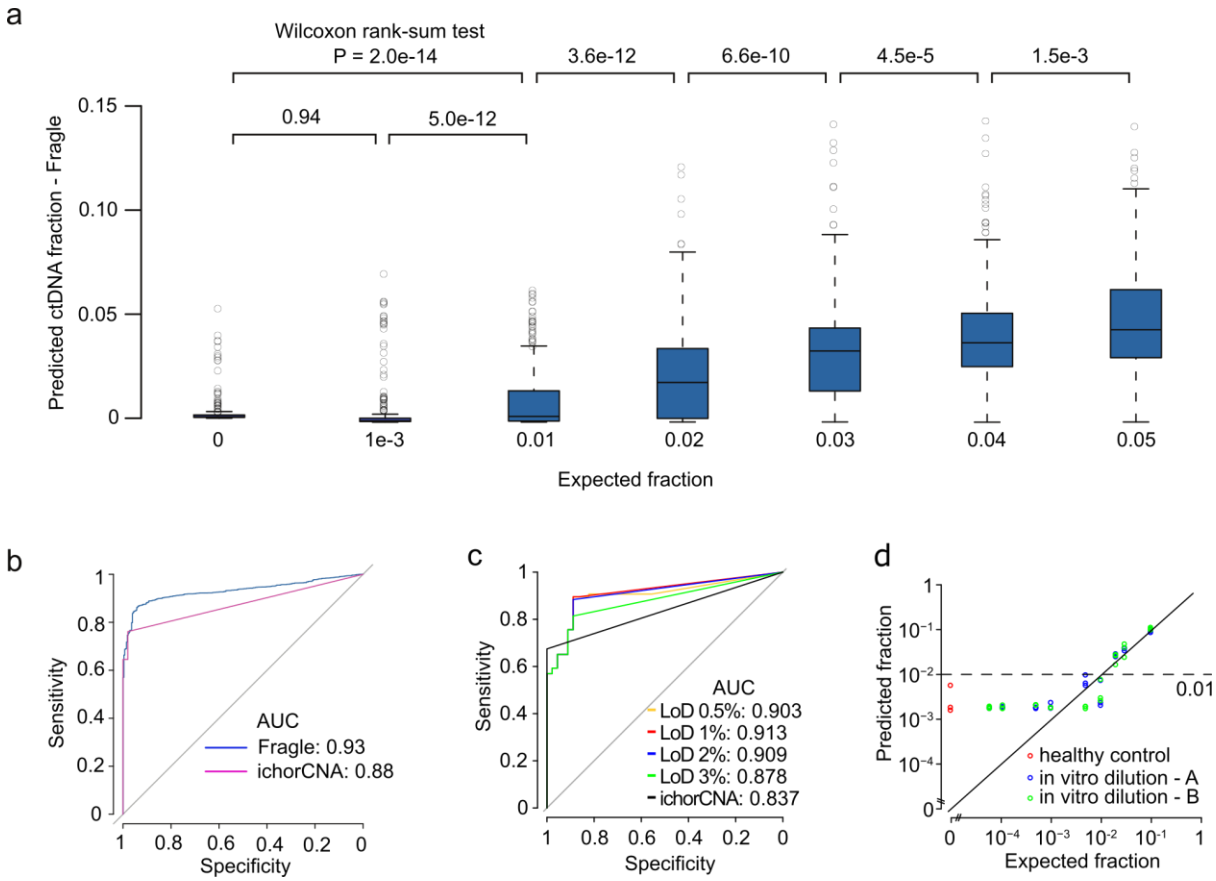


Fig. 3 Determination of lower limit of detection for Fragle. a) Predicted ctDNA fractions for healthy and low-ctDNA level samples in the validation sets in the discovery cohort. Boxplots are represented by median and interquartile range (IQR), with +/–1.5 IQR as whiskers. **b**) ROC analyses for classification of validation healthy control and cancer (⩾1% ctDNA) samples. **c**) ROC analyses for classification of cancer (N=86) and healthy (N=45) plasma samples based on different LoD thresholds using unseen test data. **d**) Expected vs. predicted ctDNA fractions in two in vitro ctDNA dilution series.

Assessment of sequencing depth and computational requirements

Fragle was trained using samples comprising 10 million cfDNA fragments, equivalent to ~1x WGS using 151bp paired-end sequencing. To further evaluate the sequencing coverage requirements for Fragle, we down-sampled samples from the unseen test cohort to render samples with lower fragment counts, ranging from 5 million (0.5x) to as low as 10 thousand fragments (0.001x). At 0.1x coverage (1M fragments), Fragle demonstrated near-perfect concordance (r = 0.97) with the predictions from the original 1x samples (Fig 4a). The performance significantly deteriorated (r < 0.8) at <0.01x coverage, suggesting that whole-genome coverage of ~0.1x provides a good trade-off between prediction accuracy and sequencing cost (Fig 4a, Suppl. Fig 6). We next tested and compared the computational requirements of Fragle as a software tool. While Fragle integrates predictions from multiple machine learning models, the method demonstrated an extremely fast runtime of ~50 seconds on a standard computer for a 1x coverage WGS sample (Fig. 4b). The memory use was also

very low (20 MiB) and independent of the sample sequencing depth. In these benchmark experiments, both compute and memory requirements were significantly lower (3-4x) than another widely used software tool for ctDNA quantification, ichorCNA. Overall, these results demonstrate that Fragle has low requirements for sequencing depth and computational resources.

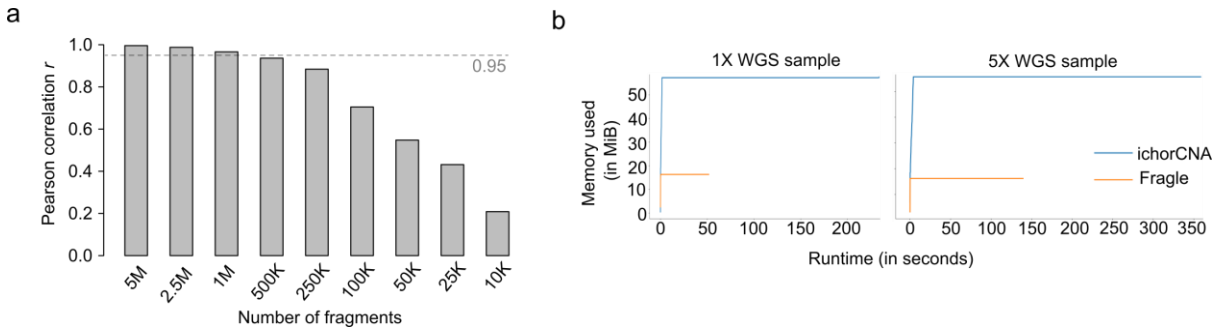


Fig. 4. Fragle software usability analysis. a) Correlation of the ctDNA fractions predicted by Fragle using 10 million (M) fragments and their downsampled counterparts (5M, 2.5M, 1M, 500K, 250K, 100K, 50K, 25K, and 10K, respectively). b) Compute time and memory requirements of Fragle and ichorCNA when applied to 1x and 5x coverage lpWGS samples.

Application of Fragle to targeted sequencing data

Targeted gene sequencing of plasma samples is routinely used in the clinic for tumor genotyping. However, absolute ctDNA quantification using mutation VAFs alone remains challenging, for example due to the lack of clonal mutations in the panel and noise from variants associated with clonal hematopoiesis [42]. Furthermore, existing algorithms for ctDNA quantification using plasma lpWGS data are not compatible with targeted gene panel sequencing data. To explore whether Fragle could quantify ctDNA levels using targeted sequencing data, we applied Fragle to samples from three cohorts having both lpWGS and targeted sequencing data (Fig. 5a, Methods). Using standard on-target reads obtained from the targeted sequencing data, Fragle tended to overestimate the ctDNA burden as compared to the lpWGS data (Fig. 5b), especially for samples with ctDNA levels <40%. We then evaluated the method when applied to off-target reads, which are often filtered and ignored in a targeted sequencing experiment. Remarkably, we observed stronger concordance between lpWGS predictions and off-target read-based predictions from the targeted sequencing across all three cohorts for Fragle (r = 0.87-0.95; $P < 0.0012$; Fig. 5c). While Fragle was trained using lpWGS data, these results support that Fragle can accurately estimate ctDNA levels using both WGS and targeted sequencing data.

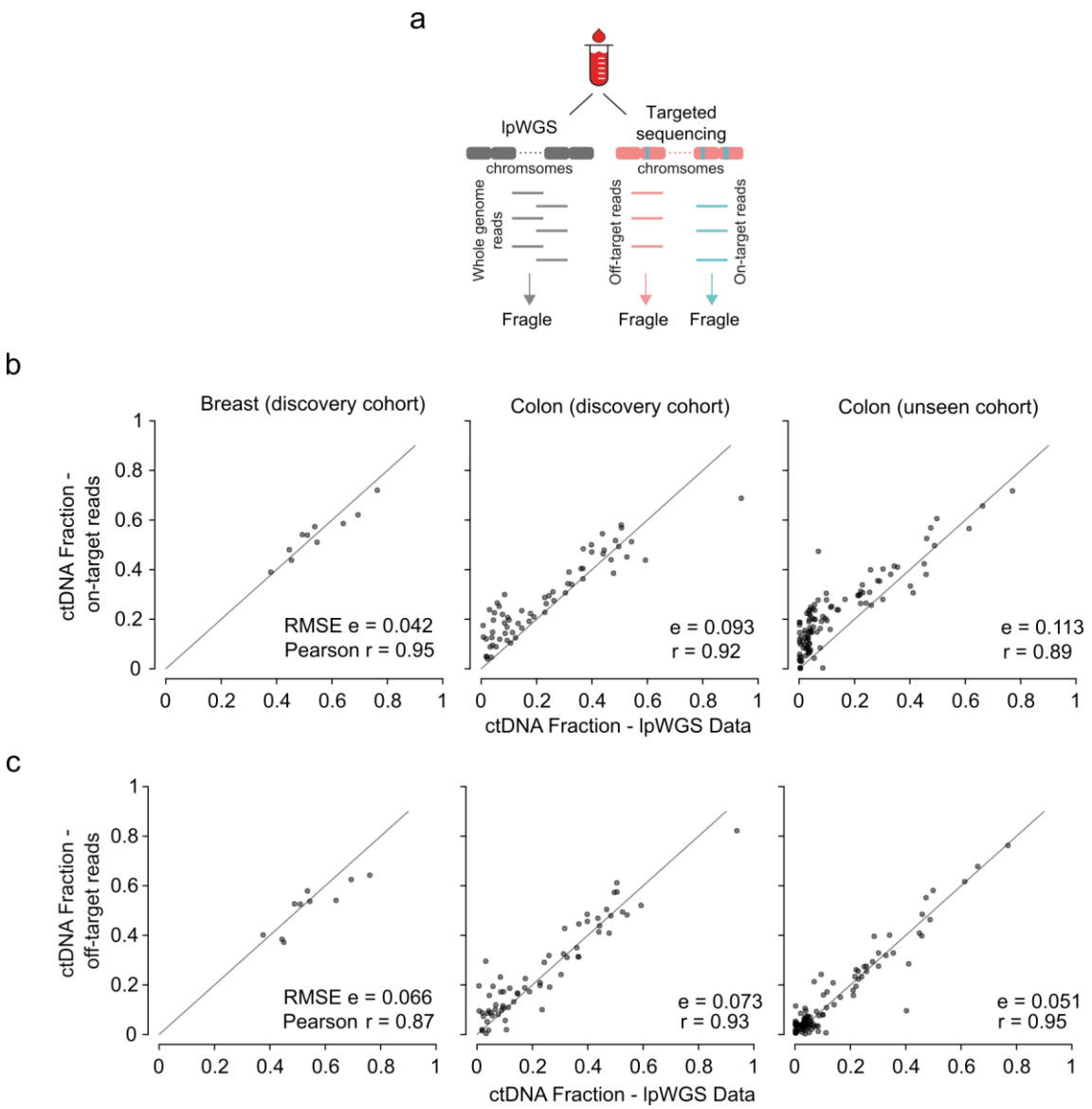


Fig. 5. Application of Fragle to targeted sequencing data. **a**) Application of Fragle to samples having both lpWGS and targeted gene panel sequencing data. **b**) Fragle predicted ctDNA levels using lpWGS data and on-target reads from targeted sequencing data. **c**) Fragle predicted ctDNA levels using lpWGS data and off-target reads from targeted sequencing data.

Monitoring of ctDNA levels 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 to jointly profile actionable mutations, ctDNA levels, and disease progression. We first considered a metastatic colorectal cancer patient for whom we had collected 21 serial blood plasma samples over a cetuximab/chemotherapy treatment course of 3 years (Fig. 6a; Suppl. Data 12). In this patient, we observed strong temporal correlation of Fragle predicted ctDNA levels, mutation VAFs, and treatment response determined from radiographic imaging (Fig. 6a). Both VAFs and Fragle ctDNA levels were low when the patient experienced near

complete response (days 121-233), and they subsequently increased as the patient developed advanced progressive disease (days >500). However, the dynamic range of VAFs varied extensively across different mutations and time points, 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 allele frequencies measured over twice as high at many time points. In these samples, Fragle provided an orthogonal and independent measure of ctDNA levels.

We also observed strong temporal correlations between mutation VAFs and Fragle ctDNA levels in longitudinal samples from four other colorectal cancer patients (Fig 6b-e; Suppl. Data 12). The first of these patients displayed concordant and increasing VAFs and Fragle ctDNA levels, consistent with emergence of progressive disease (PD) at late timepoints (Fig 6b). The second patient developed partial response following FOLFOXIRI treatment, consistent with both reductions in VAFs and Fragle ctDNA burden (Fig 6c). The last two patients both developed PD and displayed high Fragle ctDNA levels (>10%) at late time points (Fig 6d-e). However, the automated variant calling pipeline failed to detect mutations at these timepoints despite the presence of PD. We therefore performed manual inspection of sequencing reads at these positions, which confirmed presence of TP53 and ATR mutations in these samples (4-5% VAF, Suppl. Data 13). Overall, these data demonstrate high concordance of Fragle estimated ctDNA levels and disease progression estimated from radiographic imaging. Secondly, they outline how Fragle could be used to interpret and resolve heterogenous and variable mutation VAFs profiled with targeted sequencing assays.

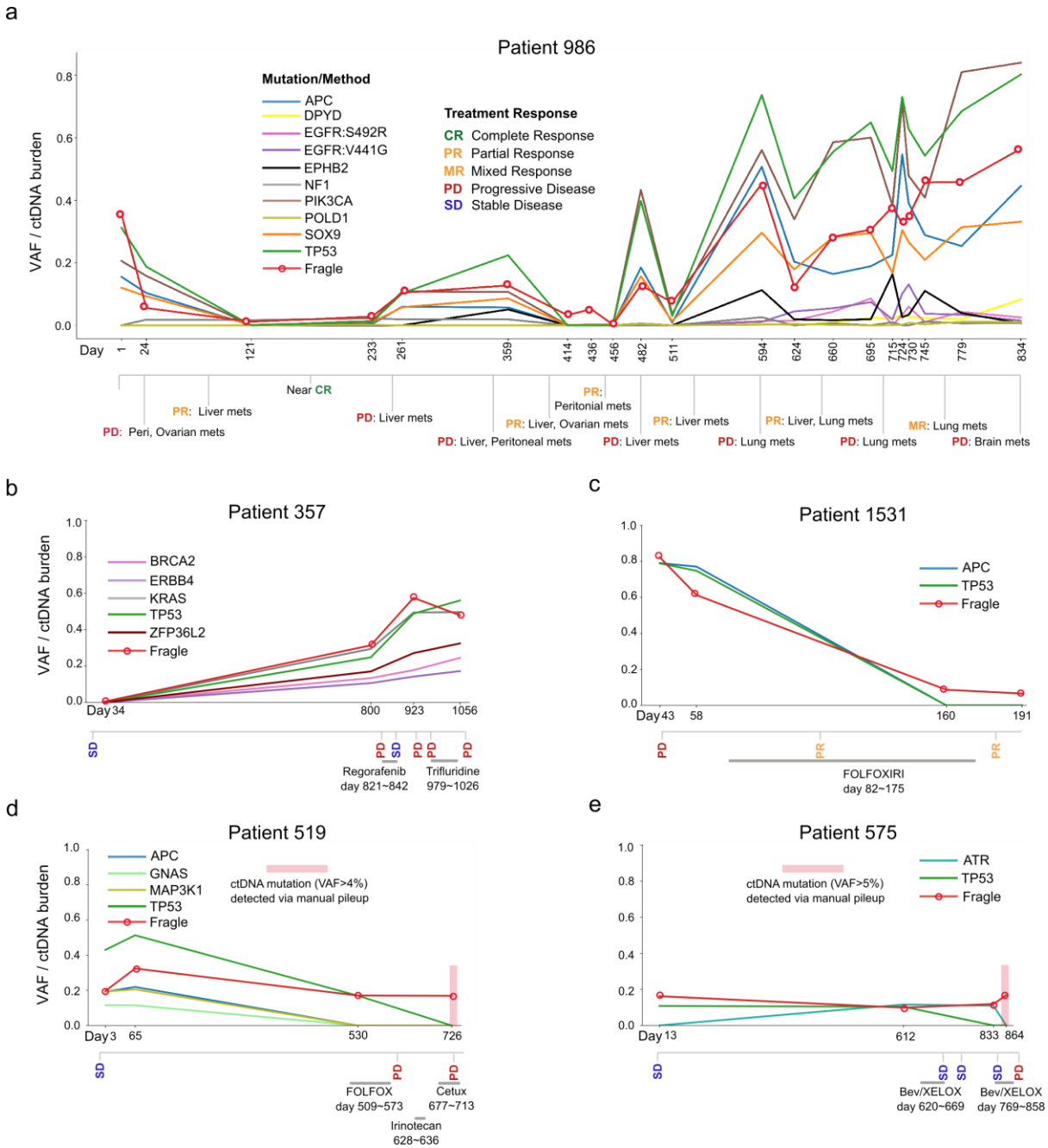


Fig. 6. Monitoring of ctDNA levels and disease progression from targeted sequencing. **a-e**) Simultaneous longitudinal profiling of Fragle ctDNA levels and mutation VAFs in metastatic colorectal cancer patients using plasma targeted gene panel sequencing. Only mutations detected in at least two timepoints were included. Mutation VAFs were profiled using automated pipeline. Manual pileup was performed at highlighted timepoints where mutation detection failed.

Discussion

While previous studies have explored how cfDNA fragment length signatures can be used to classify plasma samples from cancer patients and healthy individuals [12-15], it remained unknown whether such fragmentomic signatures could 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 cfDNA fragment length density distribution with no requirement of tumor biopsy or matched normal sample. Using an in silico data augmentation approach, we trained and evaluated Fragle on more than 2000 lpWGS samples spanning multiple cancer types and healthy cohorts. Using both in vitro and in silico dilution data, Fragle demonstrated accurate quantification of plasma ctDNA levels with a lower LoD than the current state-of-the-art approach, ichorCNA, for ctDNA quantification using lpWGS data.

Fragle is the first method to demonstrate accurate ctDNA quantification directly from a targeted gene panel sequencing assay. Existing methods developed for ctDNA quantification are not directly compatible with targeted sequencing data, requiring either low-pass whole genome sequencing (lpWGS) data [6], DNA methylation data [7], or modifications to the targeted sequencing panel [8]. Using colorectal and breast cancer plasma samples sequenced with both lpWGS and a targeted gene panel, we demonstrate high concordance of Fragle estimates across assays. Furthermore, we demonstrated how accuracy increased further as we limited Fragle to off-target reads from the targeted assay. Interestingly, these reads are usually ignored and filtered from targeted sequencing data. However, compared to on-target reads originating from specific and limited genomic regions, off-target reads are nonspecifically captured and spread across the whole genome [16], thereby likely exhibiting more robust and unbiased fragmentation signatures.

Using longitudinal targeted sequencing samples from colorectal cancer patients, we demonstrated strong concordance of Fragle-inferred ctDNA dynamics and tumor progression measured from radiographic imaging. In a targeted sequencing assay, mutation VAFs may also be used as a proxy for the ctDNA level in the sample. However, this approach is hindered by several factors. First, narrow tumor-naïve gene panels may fail to overlap with the mutations present in the tumor, rendering the assay futile. Moreover, accurate ctDNA quantification from VAFs require identification of clonal mutations and their chromosomal copy number levels, both of which are challenging to infer from narrow tumor-naïve gene panels. Indeed, in our analysis of longitudinal targeted sequencing data from individual colorectal cancer patients, the dynamic range of mutation VAFs varied extensively across different mutations and time points. These results highlight how Fragle ctDNA quantification may also

be used to interpret and resolve heterogenous and variable mutation VAFs profiled with targeted sequencing.

Overall, Fragle showed robust performance across plasma samples from 7 common solid tumor types and distinct healthy plasma cohorts. Furthermore, the method showed strong concordance with radiographic imaging and tumor VAFs in longitudinal samples from colorectal cancer patients undergoing varied 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. Since Fragle uses off-target reads to quantify ctDNA with targeted sequencing, it is likely that the method will also generalize well across distinct targeted sequencing panels. However, testing in additional cohorts is needed to further verify that Fragle generalizes across unseen tumor types, sequencing panels, and therapeutic exposures. Clonal hematopoiesis of indeterminate potential (CHIP) is a known contributor of cfDNA fragments in some patients, with CHIP mutations reported to occur at ~1-2% VAFs [42]. While we demonstrated a ~1% LoD using in vitro and in silico diluted plasma samples, further evaluation in subjects with confirmed high levels of CHIP is needed to determine if Fragle can robustly discriminate between fragmentomic signatures of CHIP and solid tumor cells.

Fragle is fast and flexible, estimating ctDNA levels in less than a minute using paired-end reads from both lpWGS and targeted sequencing assays, without the need for a paired tumor or buffy coat sample. By also enabling orthogonal ctDNA quantification from targeted sequencing data, the method could limit the need for running multiple assays for disease monitoring and interpretation of negative results from plasma genotyping [28], facilitating simultaneous discovery of actionable cancer mutations and accurate estimation of ctDNA levels with a single assay. Overall, Fragle is a low-cost and versatile approach for repeated and cost-effective tracking of ctDNA dynamics, enabling novel studies aiming to monitor for disease relapse and progression during cancer treatment.

Materials and 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 [8, 10, 12, 22]. Similarly, the test cohort was composed of internal samples as well as samples from a previous study [22], all described in Suppl. Data 1. For new samples generated as part of this study, volunteers were recruited at the National Cancer Centre Singapore, under studies 2018/2795 and 2019/2401 (for colorectal cancer (CRC) patients) and 2012/733/B (for healthy individuals) approved by the Singhealth Centralised Institutional Review Board. Written informed consent was obtained from patients. Plasma was separated from blood within 2 hours of venipuncture via centrifugation at 10 min × 300 g and 10 min × 9730 g, and then stored at −80 °C. DNA was extracted from plasma using the QIAamp Circulating Nucleic Acid Kit following manufacturer's instructions. Sequencing libraries were made using the KAPA HyperPrep kit (Kapa Biosystems, now Roche) following manufacturer's instructions and sequenced on Illumina NovaSeq6000 system. Low-pass WGS (~4x, 2×151 bp) was performed on 190 cfDNA samples from CRC patients and 45 cfDNA samples from healthy individuals in the test cohort. We used bwa-mem [29] to align WGS reads to the hg19 human reference genome.

Estimation of ctDNA fractions in the training cohort

We estimated the ctDNA fractions in the plasma samples of 4 cancer types using orthogonal methods. In the colorectal cancer (CRC) dataset (n = 65), 12 samples had ~90x cfDNA and ~30x matched buffy coat WGS data, and their ctDNA fractions were estimated from four tumor tissue-based methods as previously reported [8]. The remaining 53 CRC samples had lp-WGS and targeted promoter-region sequencing data, and we inferred ctDNA fractions by averaging ichorCNA and NDRquant estimates [8]. In the breast cancer (BRCA) dataset (n = 10), in addition to the reported ~90x cfDNA WGS data [8], we also obtained matched buffy coat sequencing data (WGS ~30x), enabling consensus estimation of ctDNA fractions using the same consensus approach as with CRC. The liver (HCC) and ovarian cancer (OV) datasets only had lp-WGS data and ichorCNA [6] was used to quantify ctDNA levels in these samples. The details of the estimation of ctDNA fractions are shown in Suppl. Data 2.

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 = 101) into training and validation sets. We repeated this 10 times, creating 10 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 (Suppl. Data 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 validations sets, respectively (see Suppl. Fig. 1 for details).

Overview of Fragle

Fragle quantifies ctDNA levels directly from a cfDNA fragment length histogram. Using paired-end sequencing data, we compute 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 Suppl. Fig 7 for details): a quantification and a model-selection stage. The quantification stage employs two models: (i) the low ctDNA burden model, and (ii) the high ctDNA burden model. These models were designed and optimized to quantify accurately in low (<3%) and high ctDNA fraction (≥3%) samples, respectively. Although these models differ in terms of architecture and loss function, the input features for both models are extracted from the fragment length histogram of the input cfDNA sequencing data. In the initial quantification stage, we pass the fragment length histogram (after performing appropriate normalization and feature extraction) to both models and obtain two ctDNA fraction estimates. In the second stage, a selection model uses these two ctDNA fraction predictions as input features. This model decides whether the sample is a high or low ctDNA burden sample and selects the corresponding ctDNA fraction estimate as the final output.

High ctDNA burden model

We computed the fragment length histogram for all fragments sized 51-400 bp using the paired-end reads in the sample (reads were filtered as mentioned above). The histogram of each sample was normalized using the highest observed fragment length count found in that sample, 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 as an additional smoothing step. The transformed length histogram features for a given sample was further standardized relative to the training set (column-wise z-score normalization). The model uses a neural network consisting of 16 fully connected layers with batch normalization [39] and residual connections [40]. While batch normalization speeds up training and helps convergence, the residual connections help in alleviating the gradient vanishing problem inherent in the earlier layers of our deep neural network. Dropout regularization [41] was used

in between intermediate layers with a dropout rate of 30% in order to prevent the model from overfitting. A composite loss function was used combining L1 loss and binary cross entropy loss. The L1 loss was used for ctDNA quantification (main objective), while the binary cross entropy loss was used to enable the model in distinguishing the healthy samples from the low tumor burden cancer patients.

Low ctDNA burden model

We computed fragment length histograms (51-400 bp) in 10 Mb genomic bins using pysam library of Python. Motivated by the observation that tissue specific regulatory regions may display differential fragmentation signatures [8, 18, 24], we identified 28 cancer-specific genomic bins containing only genes with high expression in solid tumors (FPKM < 1 in blood cells, and FPKM > 10 in 20 cancer types; Suppl. Data 10), and 59 blood-specific bins that only contain genes with high expression in whole blood cells (FPKM > 10 in blood cells, and FPKM < 1 in 20 cancer types; Suppl. Data 10). We merged the fragment length histograms from all cancer-specific bins to create a cancer fragment length histogram. Similarly, we merged the fragment length histograms from all blood-specific bins to create a blood fragment length histogram. These two histograms were standardized across training samples (relative to maximum value per length feature). We further performed feature selection for informative fragment lengths in histograms using DeepLIFT [30], identifying 127 fragment length features (out of 350) strongly informative of ctDNA burden prediction (Suppl. Data 11). The model uses a two branch (cancer and blood histograms) neural network, with each branch consisting of four fully connected dense layers. The output feature vectors from these two branches are concatenated and passed through a coordination branch (fully connected dense layers) that provides the estimated ctDNA fraction as the output. We use a customized relative MAE (mean absolute error) loss function that guides the model to focus more on the low ctDNA burden cancer samples and the healthy samples:

$$\textit{Relative MAE} = (MAE + \sigma) / (true_fraction + \sigma)$$

Here, σ is a hyperparameter tuned based on cross-validation data. A lower value of $\sigma = 0.01$ implicitly adds more importance to the samples of lower ctDNA burden. But lowering this value too much makes the training process unstable.

Selection model

The low and high ctDNA burden 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.

Targeted sequencing data

Plasma and patient-matched buffy coat samples were isolated from whole blood within two hours 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 (n = 109) in the unseen colorectal 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 pre-processed to append unique molecular identifiers (UMIs) into the fastq headers, followed by read alignment using bwa-mem [29]. 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 [31] using a minimal VAF threshold of 0.05%, and annotated all variants using Variant Effect Predictor [32]. 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 [33] 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.

In vitro dilution experiments

High ctDNA burden cfDNA samples from 2 individual CRC patients were selected to create a starting point for the dilution series. The ctDNA content for each sample was experimentally determined by 2 methods (ichorCNA [6] and NDRquant [8]), with high concordance across methods (sample 1: ctDNA content 38% by ichorCNA, 37% by NDRquant; sample 2: 39% and

38%, respectively). Commercial pooled cfDNA from healthy volunteers (0% ctDNA) was purchased from PlasmaLab (lot numbers 2001011, 210302) and was used to set up an 8 point serial dilution of the ctDNA content for each sample. 3 technical replicates were performed for each dilution point. lp-WGS was performed with a depth of ~4-5x.

Application of Fragle to targeted sequencing data

Duplicates were removed from the targeted sequencing data using Picard MarkDuplicates function (https://broadinstitute.github.io/picard/), and on-target and off-target reads were extracted from the bam files using samtools view function. The resulting reads were used to generate the input fragment length histograms as detailed above.

Data availability

Published data used in this study and their access codes are present in Suppl. Data 1. WGS data generated in this study will be deposited at the European Genome-phenome Archive (EGA). Data are available under restricted access and will be released subject to a data transfer agreement.

Code availability

The Fragle software is attached as Suppl. Data 14 and will be made publicly available via Github. The software can be applied directly on raw bam files with just a single line of command.

Supplementary materials

Supplementary Fig. 1-7

Supplementary Data 1-14

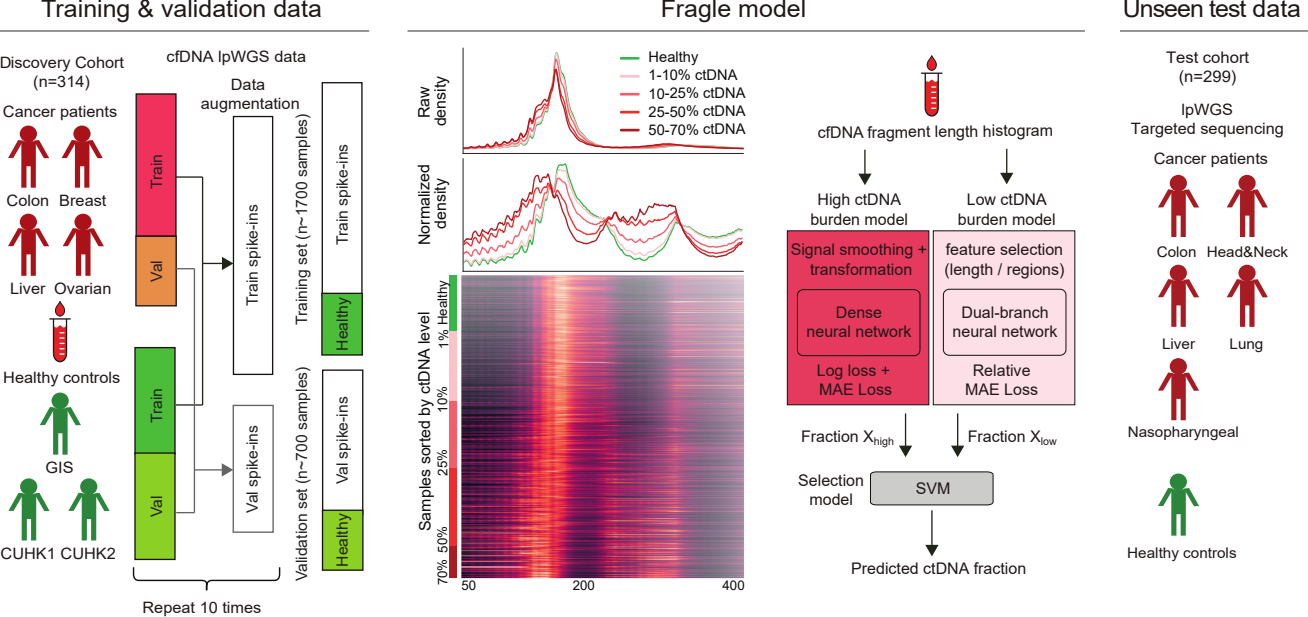
Acknowledgments

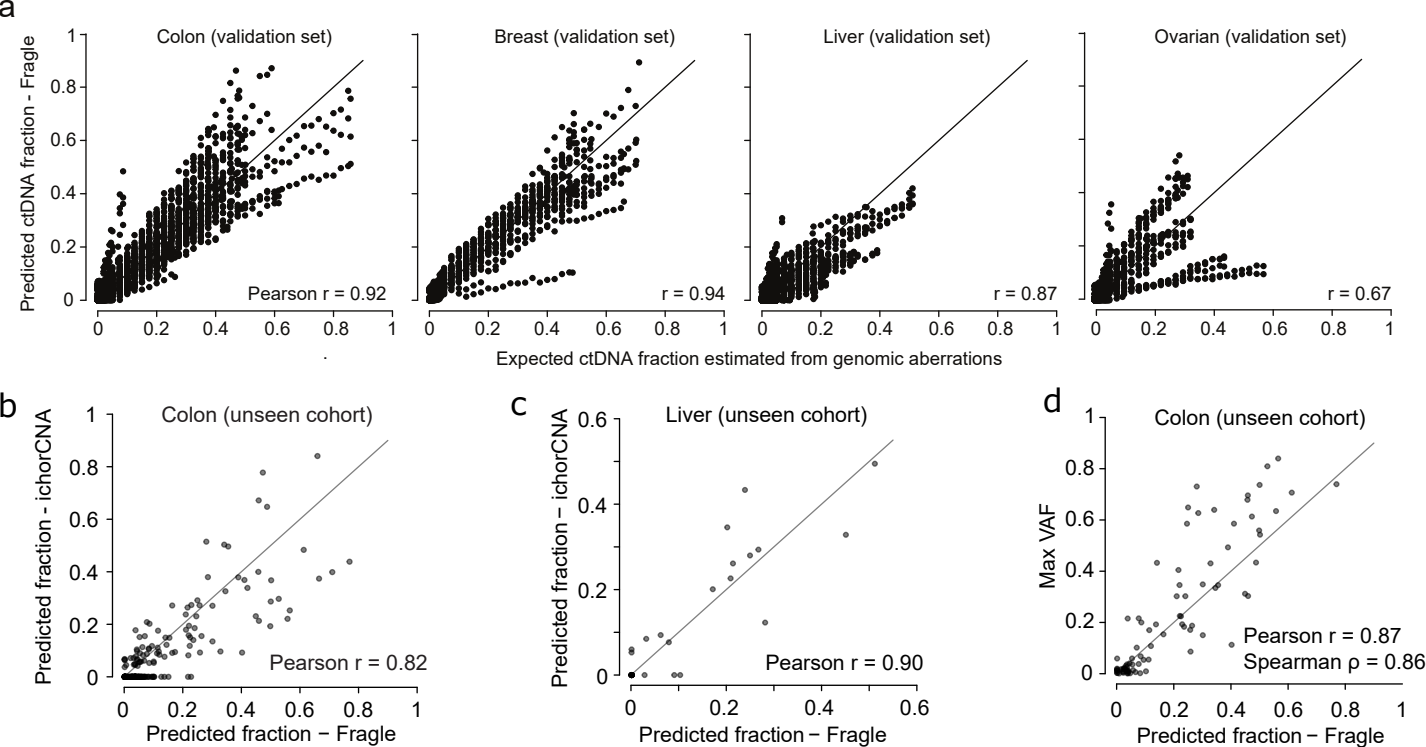
This work was supported by Agency for Science, Technology and Research under its IAF-PP programme (grant ID: H1801a0019) and the Singapore Ministry of Health's National Medical Research Council under its OF-IRG program (OFIRG21nov-0083). This work makes use of data from the Chinese University of Hong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

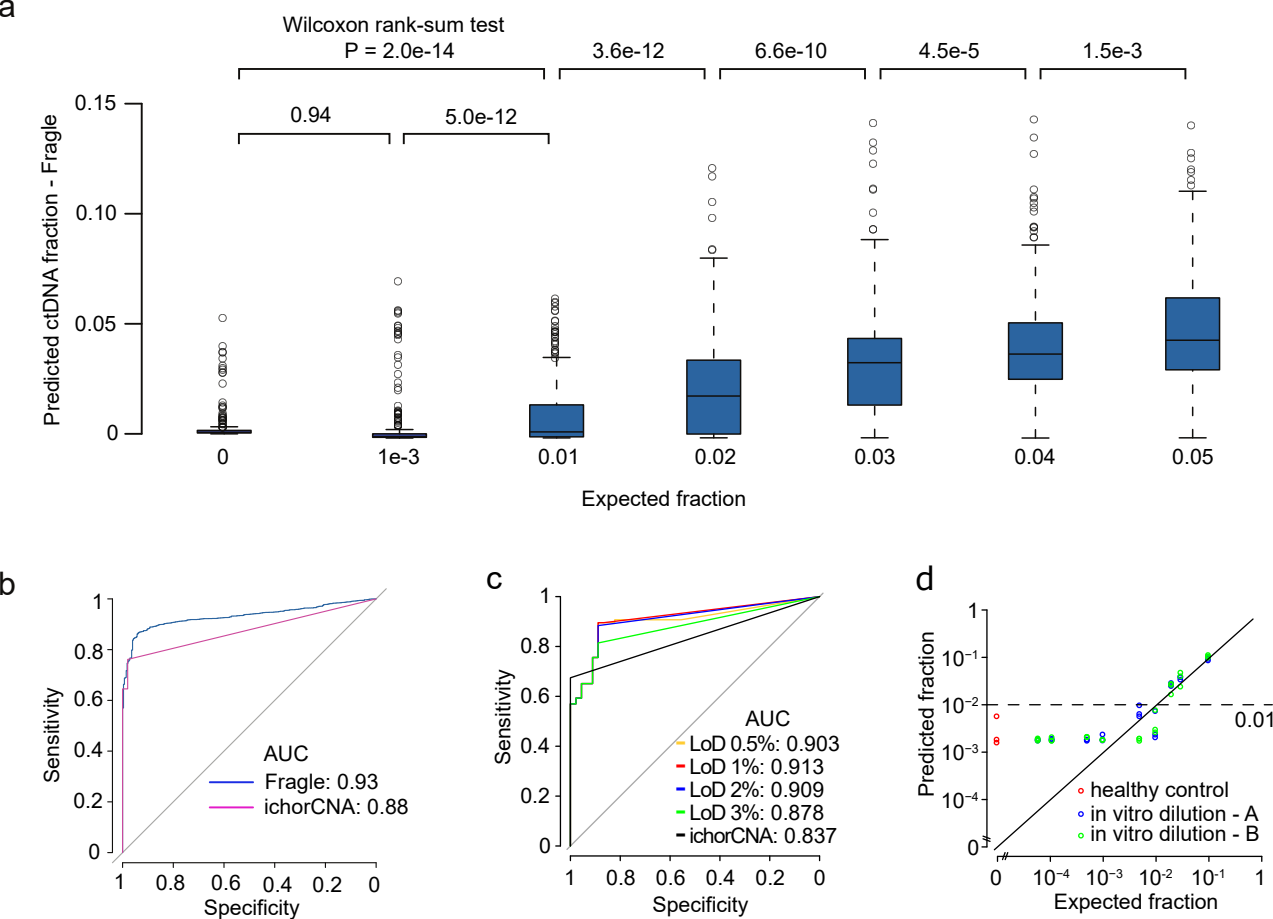
References

1. Lui, Y.Y., et al., *Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation.* Clinical chemistry, 2002. **48**(3): p. 421-427.
2. Pantel, K. and C. Alix-Panabières, *Liquid biopsy and minimal residual disease—latest advances and implications for cure.* Nature Reviews Clinical Oncology, 2019. **16**(7): p. 409-424.
3. Kilgour, E., et al., *Liquid biopsy-based biomarkers of treatment response and resistance.* Cancer cell, 2020. **37**(4): p. 485-495.
4. Vasan, N., J. Baselga, and D.M. Hyman, *A view on drug resistance in cancer.* Nature, 2019. **575**(7782): p. 299-309.
5. Razavi, P., et al., *High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants.* Nature medicine, 2019. **25**(12): p. 1928-1937.
6. Adalsteinsson, V.A., et al., *Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors.* Nature communications, 2017. **8**(1): p. 1324.
7. 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 research, 2018. **46**(15): p. e89-e89.
8. Zhu, G., et al., *Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden.* Nature communications, 2021. **12**(1): p. 2229.
9. Lo, Y.D., et al., *Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus.* Science translational medicine, 2010. **2**(61): p. 61ra91-61ra91.
10. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.* Proceedings of the National Academy of Sciences, 2015. **112**(11): p. E1317-E1325.
11. Underhill, H.R., et al., *Fragment length of circulating tumor DNA.* PLoS genetics, 2016. **12**(7): p. e1006162.
12. Mouliere, F., et al., *Enhanced detection of circulating tumor DNA by fragment size analysis.* Science translational medicine, 2018. **10**(466): p. eaat4921.
13. Foda, Z.H., et al., *Detecting liver cancer using cell-free DNA fragmentomes.* Cancer discovery, 2023. **13**(3): p. 616-631.
14. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer.* Nature, 2019. **570**(7761): p. 385-389.
15. Yu, S.C., L.L. Choy, and Y.D. Lo, *'Longing'for the Next Generation of Liquid Biopsy: The Diagnostic Potential of Long Cell-Free DNA in Oncology and Prenatal Testing.* Molecular Diagnosis & Therapy, 2023: p. 1-9.
16. Talevich, E., et al., *CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing.* PLoS computational biology, 2016. **12**(4): p. e1004873.
17. Lo, Y.D., et al., *Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies.* Science, 2021. **372**(6538): p. eaaw3616.
18. Snyder, M.W., et al., *Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin.* Cell, 2016. **164**(1-2): p. 57-68.
19. Ulz, P., et al., *Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection.* Nature communications, 2019. **10**(1): p. 4666.
20. Sun, K., et al., *Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin.* Genome research, 2019. **29**(3): p. 418-427.
21. Jiang, P., et al., *Detection and characterization of jagged ends of double-stranded DNA in plasma.* Genome research, 2020. **30**(8): p. 1144-1153.

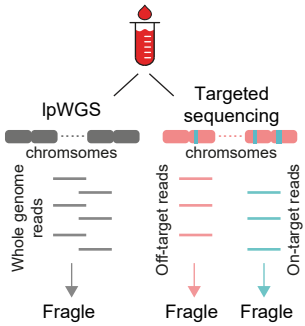
22. Jiang, P., et al., *Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation.* Cancer Discovery, 2020. **10**(5): p. 664-673.
23. Hudecova, I., et al., *Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA.* Genome Research, 2022. **32**(2): p. 215-227.
24. Esfahani, M.S., et al., *Inferring gene expression from cell-free DNA fragmentation profiles.* Nature biotechnology, 2022. **40**(4): p. 585-597.
25. Zhou, Q., et al., *Epigenetic analysis of cell-free DNA by fragmentomic profiling.* Proceedings of the National Academy of Sciences, 2022. **119**(44): p. e2209852119.
26. Doebley, A.-L., et al., *A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA.* Nature Communications, 2022. **13**(1): p. 7475.
27. Budhraja, K.K., et al., *Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer.* Science translational medicine, 2023. **15**(678): p. eabm6863.
28. Tsui, D.W., et al., *Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients.* Genome medicine, 2021. **13**(1): p. 1-15.
29. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.
30. Shrikumar, A., P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*. in *International conference on machine learning*. 2017. PMLR.
31. Lai, Z., et al., *VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.* Nucleic acids research, 2016. **44**(11): p. e108-e108.
32. McLaren, W., et al., *The ensembl variant effect predictor.* Genome biology, 2016. **17**(1): p. 1-14.
33. Mansukhani, S., et al., *Ultra-sensitive mutation detection and genome-wide DNA copy number reconstruction by error-corrected circulating tumor DNA sequencing.* Clinical chemistry, 2018. **64**(11): p. 1626-1635.
34. Pasaniuc, B., et al., *Extremely low-coverage sequencing and imputation increases power for genome-wide association studies*. Nature Genetics, 2012. **44**(6): p. 631-635.
35. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nature biotechnology, 2013. **31**(3): p. 213-219.
36. Koboldt, D., et al., *VarScan: variant detection in massively parallel sequencing of individual and pooled samples*. Bioinformatics, 2009. **25**(17): p. 2283-2285.
37. Koboldt, D. et al., *Best practices for variant calling in clinical sequencing*. Genome Medicine, 2020. **12**(1): p. 1-13.
38. Zviran, E. et al., *Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring*. Nature Medicine, 2020. **26**(7): p. 1114-1124.
39. Ioffe, S. et al., Batch normalization: *Accelerating deep network training by reducing internal covariate shift*. International conference on machine learning, 2015. P. 448-456.
40. He, K. et al., *Deep residual learning for image recognition*. IEEE conference on computer vision and pattern recognition, 2016. P. 770-778.
41. Srivastava, N. et al., *Dropout: a simple way to prevent neural networks from overfitting*. journal of machine learning research, 2014. **15**(1): p. 1929-1958.
42. Ptashkin, R. et al., *Prevalence of clonal hematopoiesis mutations in tumor-only clinical genomic profiling of solid tumors*. JAMA oncology, 2018. **4**(11); p. 1589-1593.


Training & validation data
Fragle model
Unseen test data
Discovery Cohort
(n=314)
Cancer patients
Colon
Breast
Liver
Ovarian
Healthy controls
GIS
CUHK1
CUHK2
cfDNA lpWGS data
Data augmentation
Train
Val
Train
Val
Train spike-ins
Val spike-ins
Training set (n~1700 samples)
Validation set (n~700 samples)
Train spike-ins
Healthy
Val spike-ins
Healthy
Repeat 10 times
Raw density
Normalized density
Healthy
1-10% ctDNA
10-25% ctDNA
25-50% ctDNA
50-70% ctDNA
Samples sorted by ctDNA level
1% Healthy
10%
25%
50%
70%
50
200
400
cfDNA fragment length histogram
High ctDNA burden model
Low ctDNA burden model
Signal smoothing + transformation
Dense neural network
Log loss + MAE Loss
feature selection (length / regions)
Dual-branch neural network
Relative MAE Loss
Fraction X_{high}
Fraction X_{low}
Selection model
SVM
Predicted ctDNA fraction
Test cohort
(n=299)
lpWGS
Targeted sequencing
Cancer patients
Colon
Head&Neck
Liver
Lung
Nasopharyngeal
Healthy controls

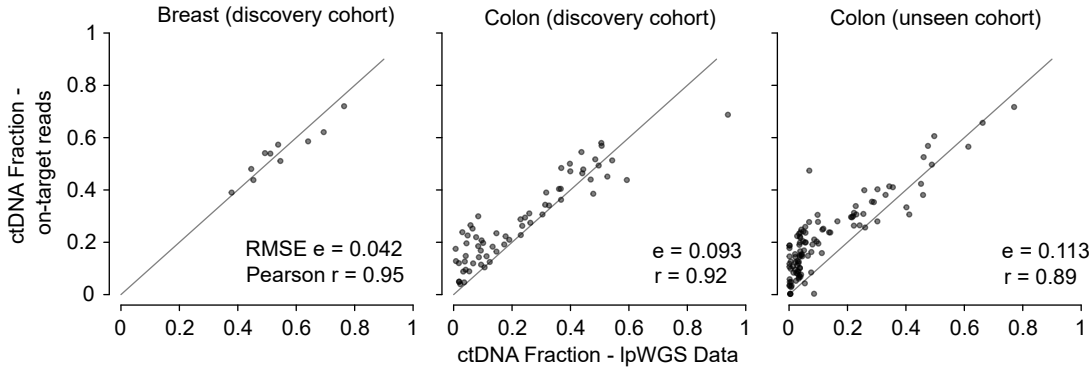

a
Colon (validation set)
Breast (validation set)
Liver (validation set)
Ovarian (validation set)
Predicted ctDNA fraction - Fragle
Pearson r = 0.92
r = 0.94
r = 0.87
r = 0.67
Expected ctDNA fraction estimated from genomic aberrations
b
Colon (unseen cohort)
Predicted fraction - ichorCNA
Pearson r = 0.82
Predicted fraction − Fragle
c
Liver (unseen cohort)
Predicted fraction − ichorCNA
Pearson r = 0.90
Predicted fraction − Fragle
d
Colon (unseen cohort)
Max VAF
Pearson r = 0.87
Spearman ρ = 0.86
Predicted fraction − Fragle


a
Wilcoxon rank-sum test
P = 2.0e-14
3.6e-12
6.6e-10
4.5e-5
1.5e-3
0.94
5.0e-12
Predicted ctDNA fraction - Fragle
0.15
0.10
0.05
0
0
1e-3
0.01
0.02
0.03
0.04
0.05
Expected fraction
b
Sensitivity
Specificity
AUC
Fragle: 0.93
ichorCNA: 0.88
c
Sensitivity
Specificity
AUC
LoD 0.5%: 0.903
LoD 1%: 0.913
LoD 2%: 0.909
LoD 3%: 0.878
ichorCNA: 0.837
d
Predicted fraction
Expected fraction
0.01
healthy control
in vitro dilution - A
in vitro dilution - B

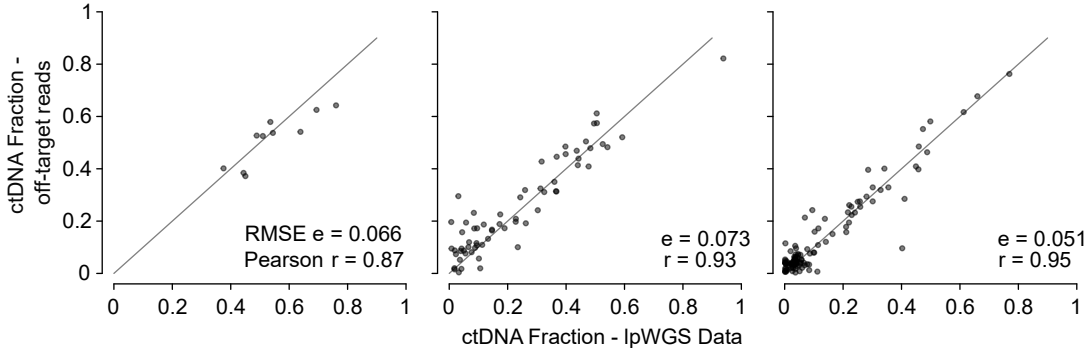
a


lpWGS
Targeted sequencing
chromsomes
chromsomes
Whole genome reads
Off-target reads
On-target reads
Fragle
Fragle
Fragle

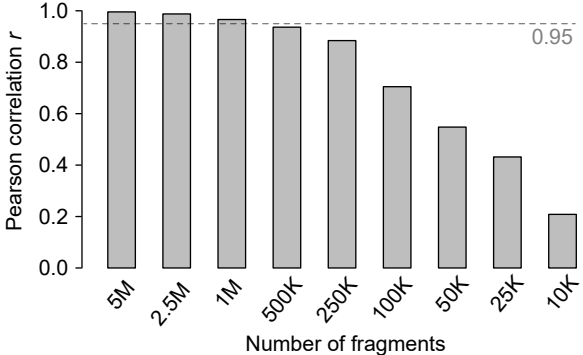
b


Breast (discovery cohort)
Colon (discovery cohort)
Colon (unseen cohort)
ctDNA Fraction - on-target reads
RMSE e = 0.042
Pearson r = 0.95
e = 0.093
r = 0.92
e = 0.113
r = 0.89
ctDNA Fraction - lpWGS Data

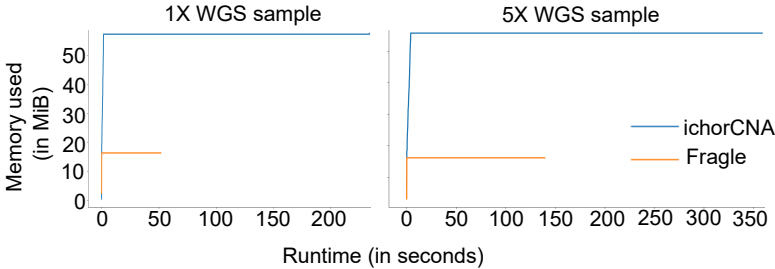
c

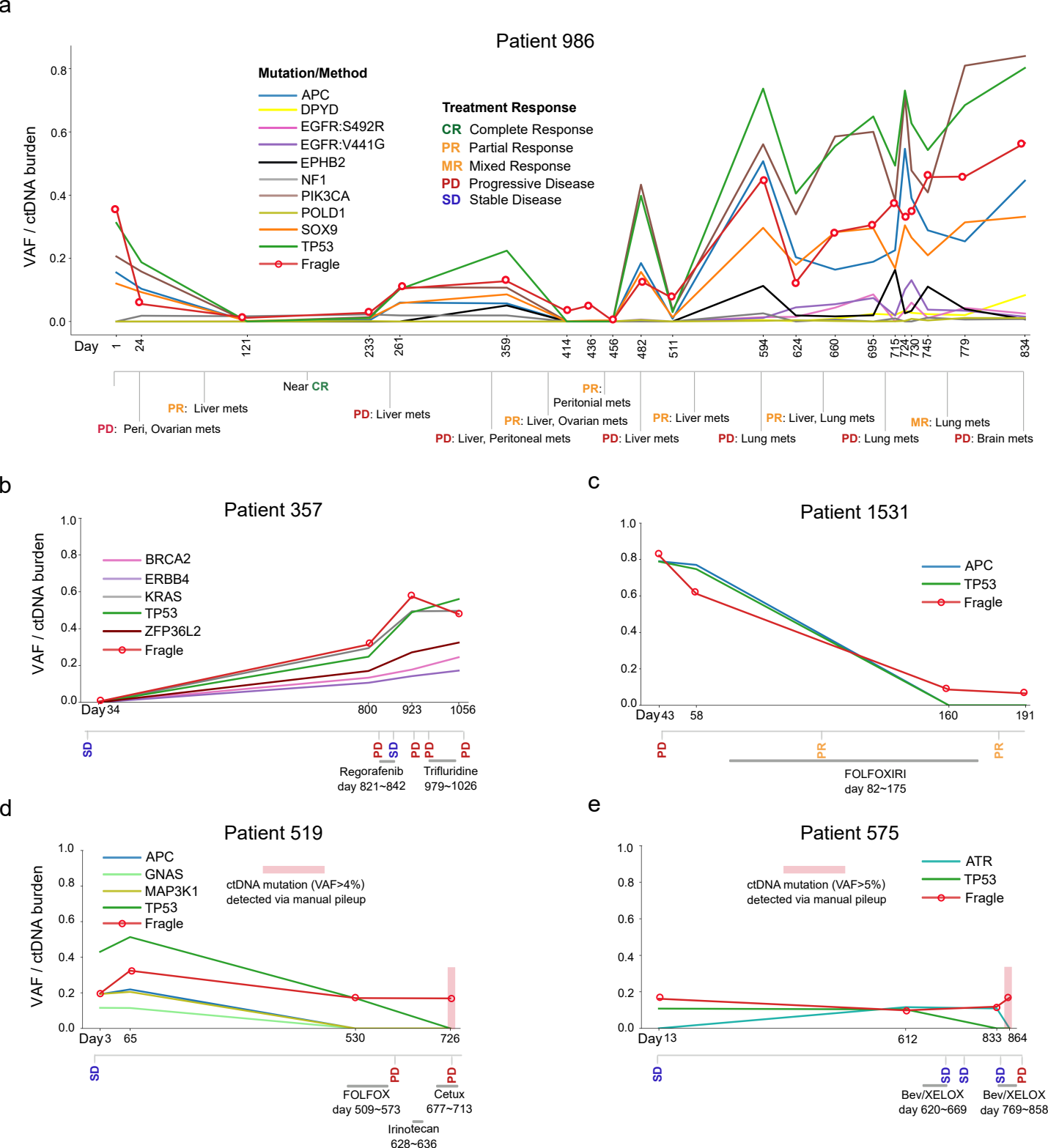

ctDNA Fraction - off-target reads
RMSE e = 0.066
Pearson r = 0.87
e = 0.073
r = 0.93
e = 0.051
r = 0.95
ctDNA Fraction - lpWGS Data

a


Pearson correlation r
1.0
0.8
0.6
0.4
0.2
0.0
0.95
5M
2.5M
1M
500K
250K
100K
50K
25K
10K
Number of fragments

b


1X WGS sample
5X WGS sample
Memory used (in MiB)
50
40
30
20
10
0
0
50
100
150
200
0
50
100
150
200
250
300
350
Runtime (in seconds)
ichorCNA
Fragle


a
Patient 986
VAF / ctDNA burden
Mutation/Method
APC
DPYD
EGFR:S492R
EGFR:V441G
EPHB2
NF1
PIK3CA
POLD1
SOX9
TP53
Fragle
Treatment Response
CR Complete Response
PR Partial Response
MR Mixed Response
PD Progressive Disease
SD Stable Disease
Day
Near CR
PR: Liver mets
PD: Peri, Ovarian mets
PD: Liver mets
PD: Liver, Peritoneal mets
PR: Liver, Ovarian mets
PR: Peritonial mets
PD: Liver mets
PR: Liver mets
PD: Lung mets
PR: Liver, Lung mets
PD: Lung mets
MR: Lung mets
PD: Brain mets
b
Patient 357
VAF / ctDNA burden
BRCA2
ERBB4
KRAS
TP53
ZFP36L2
Fragle
Regorafenib
day 821~842
Trifluridine
979~1026
c
Patient 1531
APC
TP53
Fragle
FOLFOXIRI
day 82~175
d
Patient 519
APC
GNAS
MAP3K1
TP53
Fragle
ctDNA mutation (VAF>4%)
detected via manual pileup
FOLFOX
day 509~573
Cetux
677~713
Irinotecan
628~636
e
Patient 575
ATR
TP53
Fragle
ctDNA mutation (VAF>5%)
detected via manual pileup
Bev/XELOX
day 620~669
Bev/XELOX
day 769~858