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 ctDNA quantification directly from cell-free DNA fragment length profiles. We developed Fragle using low-pass whole genome sequence (lpWGS) data from 8 cancer types and healthy control cohorts, demonstrating high accuracy, and improved lower limit of detection in independent cohorts as compared to existing tumor-naïve methods. Uniquely, Fragle is also compatible with targeted sequencing data, exhibiting high accuracy across both research and commercial targeted gene panels. We used this method to study longitudinal plasma samples from colorectal cancer patients, identifying strong concordance of ctDNA dynamics and treatment response. Furthermore, prediction of minimal residual disease in resected lung cancer patients exhibited significant risk stratification beyond a tumor-naïve targeted panel. Overall, Fragle is a versatile, fast, and accurate method for ctDNA quantification with potential for broad clinical utility.

Introduction

The death of 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), enabling 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 [3-5].

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 [6]. Existing methods 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, 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 [11]. cfDNA fragments from cancer patients tend to be shorter and more variably sized than those from healthy individuals [12-14]. These observations have motivated studies exploring how cfDNA fragment length properties can be used to classify cfDNA samples from cancer patients and healthy individuals [14-19]. 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 2500 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. Finally, to explore the use of Fragle for detection of minimal residual disease (MRD), we analyzed ctDNA levels in a cohort of 162 resected lung cancer patients 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 300 cancer plasma samples and 101 plasma samples from healthy individuals (Fig. 1, Suppl. Data 1). This cohort only comprised cancer samples with detectable ctDNA (≥3%, Suppl. Fig. 1) and ground-truth ctDNA levels were estimated using a consensus across distinct methods (see Methods, Suppl. Data 2). Using a large-scale data augmentation approach, we generated 3897 mixture samples from cancer and healthy samples with variable ctDNA fractions (see Methods, Fig. 1, Suppl. Data 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 (Fig. 1, see Methods). 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. Two parallel deep learning-based sub-models were optimized for quantitative prediction in high (≥3%) and low (<3%) ctDNA burden samples, respectively. These two sub-models adopted the same model architecture and input fragment length features but were trained using different loss functions and model parameters. In the last step, we trained a support vector machine (SVM) model to select between the outputs from the high and low ctDNA burden models as the final model prediction (see Methods). Supporting these model choices, we found that the low and high ctDNA burden sub-models showed more accurate predictions for low and high-ctDNA samples, respectively, and that the final combined model performed best across the full range of samples (Suppl. Fig 2).

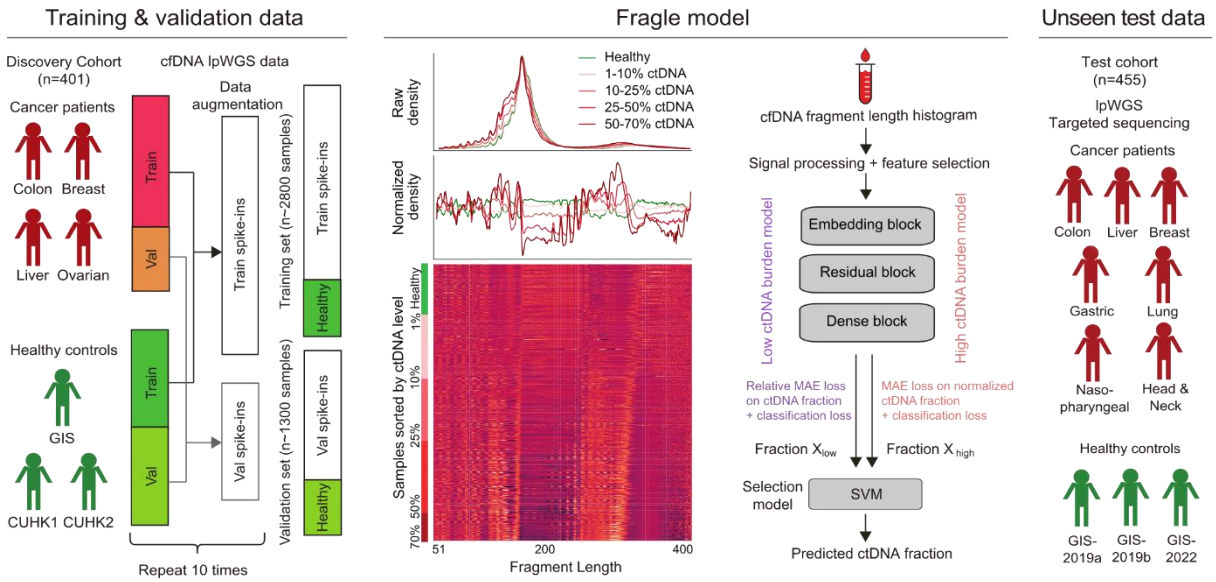


Fig. 1, Overview of Fragle. Fragle is a multi-stage machine learning-based model that estimates the ctDNA level in a blood sample from the cfDNA fragment length density distribution. Fragle was trained using a large-scale data augmentation and cross-validation approach and was further tested using unseen 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-d): Colorectal (mean absolute error (MAE) = 3.27%; Pearson r = 0.92), breast (MAE = 3.56%; r = 0.94), liver (MAE = 3.12%; r = 0.81), and ovarian cancer (MAE = 3.86%; 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 (Suppl. Fig. 3).

We trained the final Fragle model on the full discovery cohort (see Methods) 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 unseen cohorts of colorectal cancer (r = 0.81; P = 6.8e-41; N = 190; Fig. 2e), breast cancer (r = 0.80; P = 3.7e-06; N = 23; Fig. 2f), liver cancer (r = 0.87; P = 5.1e-10; N = 34; Fig. 2g), gastric cancer (r = 0.73; P = 1.9e-09; N = 34; Fig. 2h), and a mixed cohort of cancer types not included in the discovery set (Combined P = 1.3e-4; 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, see Methods). These data demonstrated high concordance between mutation VAFs and Fragle-predicted ctDNA levels (r = 0.88; P = 3.8e-28; Fig. 2i).

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 all 455 samples from the unseen test cohort to render samples with fewer fragments, ranging from 5 million (0.5x) to as low as 10 thousand fragments (0.001x). At 500K fragments (0.05x), Fragle demonstrated excellent concordance (r = 0.97) with the predictions from the original 1x WGS samples (Suppl. Fig. 5). The correlation was maintained when further down-sampling to 250K fragments, but with some discrepancies observed for some low ctDNA fraction samples (Suppl. Fig 5). These results suggest that whole-genome coverage of ~500K (0.05x) 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 demonstrated a fast runtime of ~50 seconds for a 1x coverage WGS sample (runtime increases linearly with sample coverage) and low memory usage (20 MB) independent of the sample sequencing depth (Suppl. Fig 6).

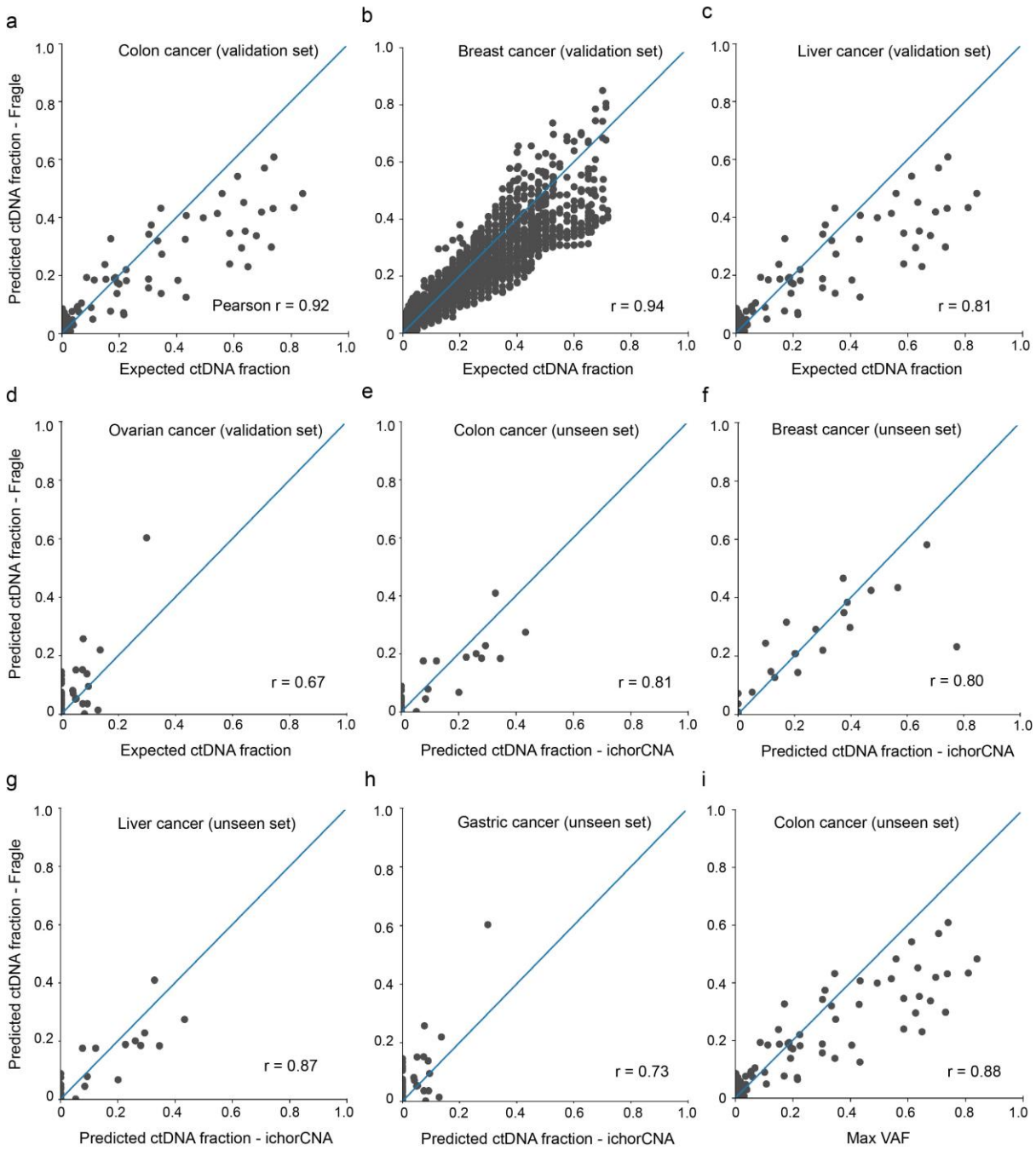


Fig. 2 ctDNA quantification in validation and unseen cohorts. (a-d) Comparison between expected and predicted ctDNA levels for colorectal (CRC), breast (BRCA), liver (HCC), and ovarian (OV) cancer samples in the validation sets. (**e-h)** Comparing ichorCNA and Fragle predicted ctDNA levels in unseen samples from colorectal (N = 190), breast (N = 23), liver (N = 34) and gastric cancer patients (N = 49). **i)** Colorectal cancer plasma samples subjected to both lpWGS and targeted sequencing; comparison of Fragle predicted ctDNA levels (lpWGS) and maximum VAFs (N = 86; samples with detectable somatic mutations).

Determination of the lower limit of detection

To explore the lower limit of detection (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, with 86% of healthy samples predicted to comprise <1% ctDNA. Furthermore, the model could differentiate between healthy and low-ctDNA level samples at the 1% ctDNA level (Wilcoxon rank sum test *P* = 2.5e-24, Fig. 3a), indicating a ~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) applied to the same samples. Expectedly, the AUC increased further when filtering out low-ctDNA burden samples (Suppl. Fig. 7). As an additional comparison, we implemented a model based on fragment length features previously used for the classification of cancer and healthy samples [14]. We trained a random forest model (following the authors of the paper) on the discovery cohort using 4 features derived from the fragment length distribution (see Methods). The 4-feature model demonstrated substantially lower classification accuracy than Fragle in the validation samples, with an AUC of 0.79.

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 all healthy plasma samples from the three different unseen cohorts as negatives (N = 57, Suppl. Data 7-8). Fragle demonstrated an AUC of 0.96 using these samples, outperforming the other models on the same set of samples (ichorCNA = 0.85, 4-feature model = 0.81; Fig. 3c). 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 (see Methods, Suppl. Data 9). In this dilution experiment, Fragle could differentiate healthy from low-ctDNA samples down to the 0.5-1% ctDNA level (*P* = 0.003, healthy vs. 0.625% ctDNA fraction samples, Fig. 3d; Suppl. 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 (see Methods). Across 3 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 94.4% with an LoD of 1%, outperforming ichorCNA with a detection rate of 66.7% (Suppl. Data 10). The second experiment comprised in vitro serial dilutions of 3 high-ctDNA plasma samples from gastric cancer patients (with 3 technical replicates, see 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, Suppl. Data 11). Overall, these results collectively suggest that Fragle can accurately quantify and detect ctDNA with an LoD of ~1%.

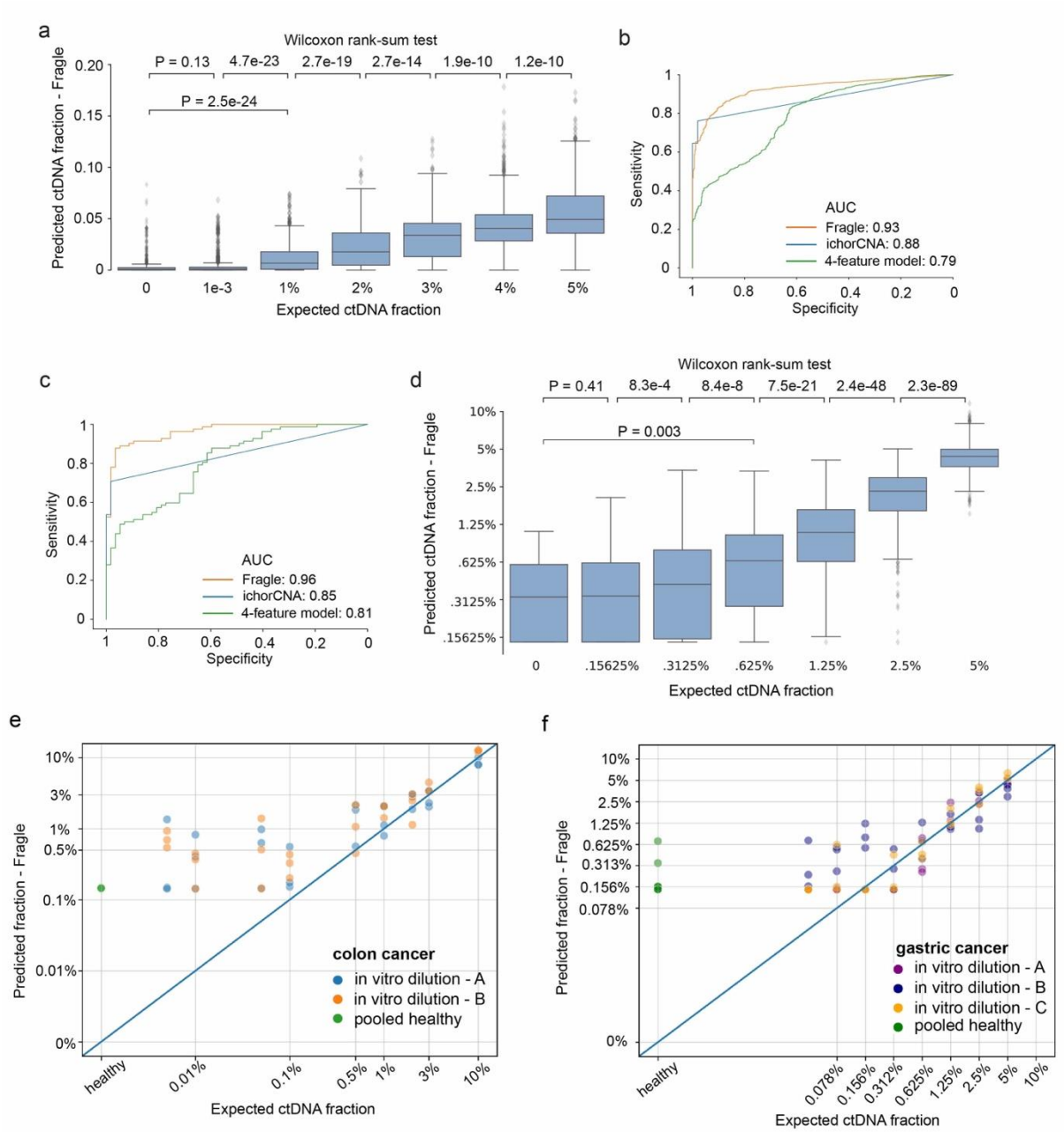


Fig. 3 Lower limit of detection. a) Predicted ctDNA fractions for healthy and low-ctDNA level samples in validation set samples. Boxplots are represented by median and interquartile range (IQR), with +/–1.5 IQR as whiskers. **b)** ROC analyses for classification of healthy control and cancer (≥1% ctDNA) samples (validation samples). **c)** ROC analysis for classification of cancer (N=86) and healthy (N=57, 3 distinct cohorts) plasma samples in the unseen test cohort. **d)** Predicted ctDNA fractions for healthy and low-ctDNA level samples using in silico dilution of 20 cancer samples (unseen cohort). Boxplots are represented by median and interquartile range (IQR), with +/–1.5

IQR as whiskers. **e)** Expected vs. predicted ctDNA fractions using in vitro ctDNA dilution for 2 colorectal cancer samples. **f)** Expected vs. predicted ctDNA fractions using in vitro ctDNA dilution for 3 gastric cancer samples.

Application of Fragle to targeted sequencing data

Targeted gene sequencing of plasma samples is routinely used for tumor 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 hematopoiesis could introduce noise [20]. To explore whether Fragle could quantify ctDNA levels using targeted sequencing data, we analyzed four cfDNA cohorts having both lpWGS and targeted sequencing data (Fig. 4a, see 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. 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 (r = 0.86, *P* = 0.001, N = 10; discovery cohort), colon cancer samples from discovery cohort (r = 0.96, *P* = 1.64e-30, N = 56), colon cancer samples from unseen cohort (r = 0.97, *P* = 3.69e-58, N = 96), and gastric cancer samples from unseen cohort (r = 0.96, *P* = 2.9e-27, N = 49) (Fig. 4c). We found that the targeted sequencing samples contained between 100K to 10M off-target fragments (equivalent to ~0.01-1.0X) across the different samples, with >95% of samples having >250K off-target fragments (~0.025X; Suppl. Fig 9). To further explore if these results generalize to targeted sequencing panels employed by commercial vendors, we evaluated a cohort of 44 plasma samples subjected to a liquid biopsy gene panel from Foundation Medicine [21]. 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 (Suppl. Data 12-13; Methods). In this cohort comprising samples from a variety of cancer types, we observed strong concordance of the reported VAFs and ctDNA levels estimated from off-target reads (*r* = 0.70, *P* = 1.01e-07, N = 44; Figure 4d). Overall, these results support that Fragle can estimate ctDNA levels using both lpWGS and targeted sequencing data.

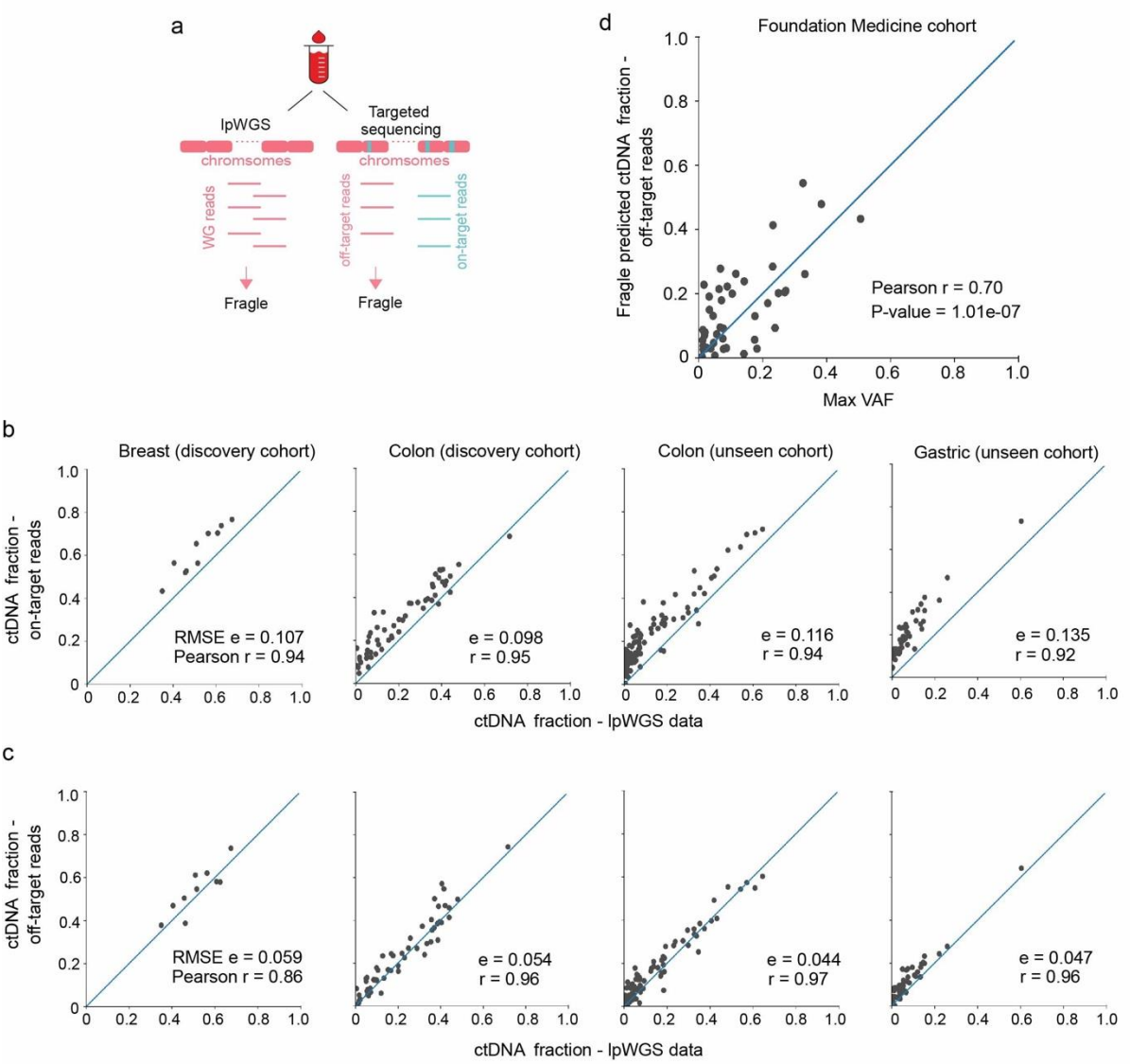


Fig. 4. Application of Fragle to targeted sequencing data. **a)** Application of Fragle to samples having both lpWGS and targeted gene panel sequencing data. **b)** ctDNA levels predicted using lpWGS data and on-target reads from targeted sequencing samples. **c)** ctDNA levels predicted using lpWGS data and off-target reads from targeted sequencing samples. **d)** Targeted sequencing data generated with commercial liquid biopsy assay (Foundation Medicine, N=44). Correlation of maximum VAFs (reported by company, germline variants filtered) and Fragle-predicted ctDNA levels using off-target reads.

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 late-stage colorectal cancer patients. In these samples, we wanted to explore the temporal relationship between Fragle-estimated ctDNA dynamics and disease progression measured by radiographic imaging (RI). Firstly, we observed strong temporal correlations between mutation VAFs and Fragle ctDNA levels across the longitudinal samples from the four patients (Fig 5a-d; Suppl. Data 14). The first patient displayed 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 progressive disease 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% VAF, Suppl. Data 15). We finally 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. 5e; Suppl. Data 14). 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 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 mutation allele frequencies more than 2-fold higher at many time points (e.g. 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 radiographic imaging. Secondly, they outline how Fragle could be used to interpret and resolve heterogeneous and variable mutation VAFs profiled with targeted sequencing assays.

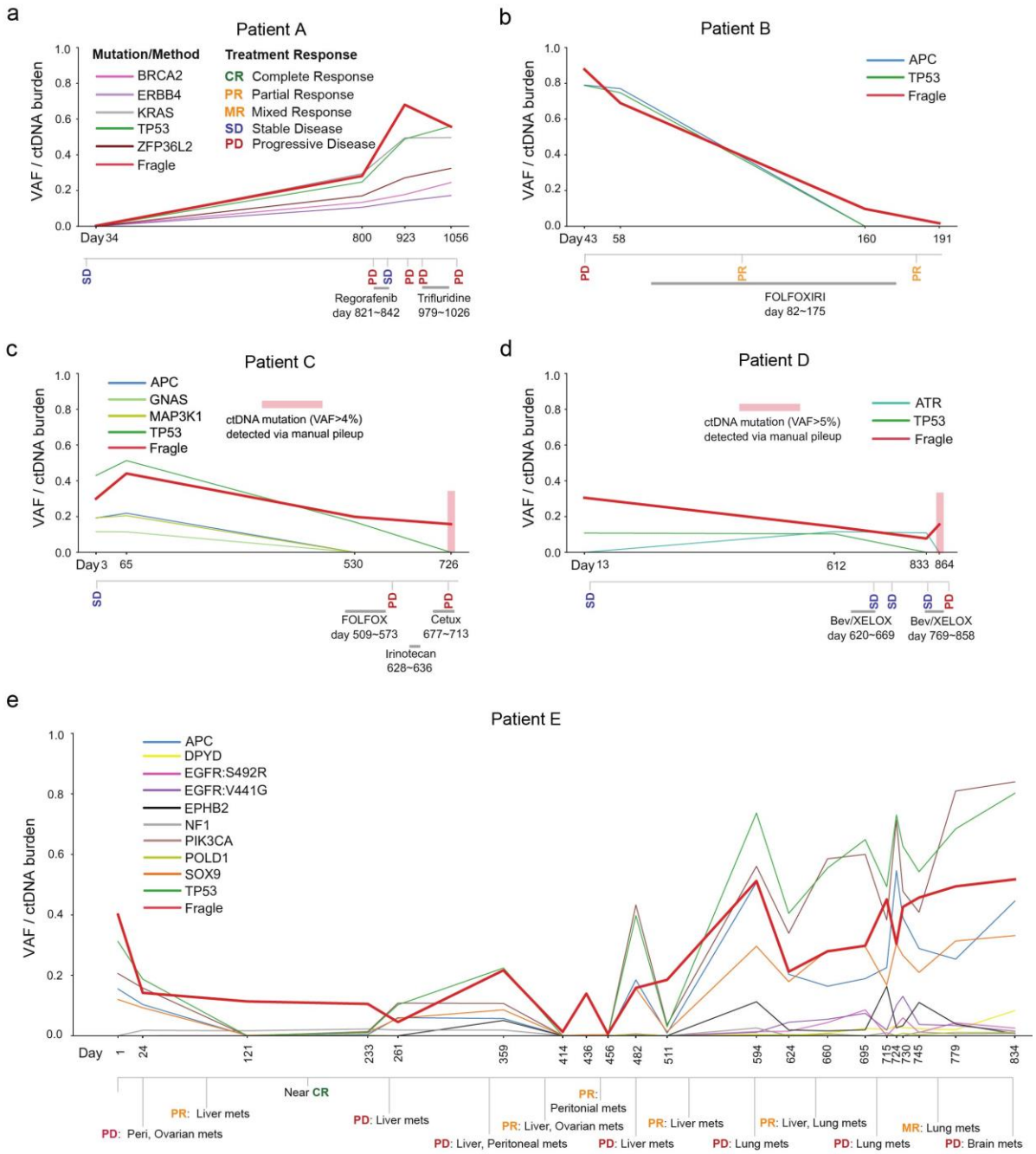


Fig. 5. Monitoring of ctDNA levels and disease progression from targeted sequencing. **a-e**) Simultaneous longitudinal profiling of Fragle ctDNA levels and mutation VAFs in metastatic colorectal cancer patients using plasma targeted gene panel sequencing. Disease progression was captured with radiographic imaging. 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.

Risk stratification for early-stage lung cancer patients

Blood-based detection of minimal residual disease (MRD) following treatment has the potential to improve risk stratification and management strategies for cancer patients [22, 23]. Given the ~1% LoD for Fragle, we explored if the method could be used for tumor-naïve MRD screening, with no requirements for a matching tissue sample. We obtained targeted sequencing data from a published cohort (MEDAL) of 162 early-stage lung cancer patients that had plasma samples collected at the landmark timepoint (~30 days following curative surgery) [24]. In this study, plasma samples were subjected to a commercial tumor-agnostic targeted sequencing assay, and the authors classified samples into ctDNA positive (N=4) and negative (N=158) groups based on mutation VAFs (Fig. 6a, Suppl. Data 16). 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 hazard model) while controlling for known clinical prognostic variables such as tumor type and stage (Fig. 6c). Overall, these data demonstrate the potential clinical utility of Fragle as a supplement to a standard tumor-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.

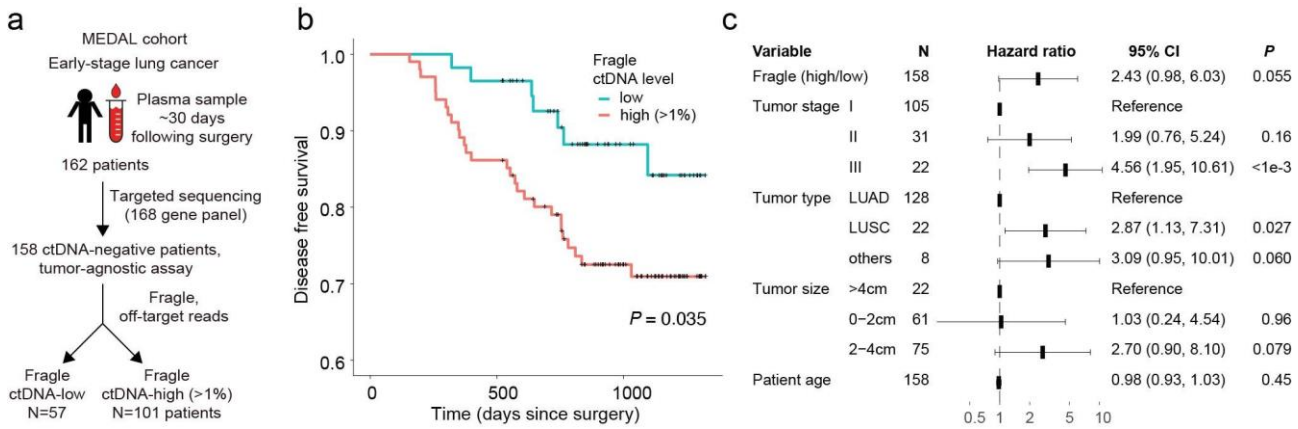


Fig. 6. Risk-stratification of early-stage lung cancer patients: a) Fragle was used to predict ctDNA levels in 158 early-stage lung cancer patients classified as ctDNA-negative with a tumor-agnostic targeted gene panel assay. Plasma samples were obtained at the landmark timepoint (~30 days after surgery) 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 (DFS) 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 DFS while controlling for other clinical variables.

Discussion

While previous studies have explored how cfDNA fragment length signatures can be used to classify plasma samples from cancer patients and healthy individuals [14-19], 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 tumor biopsy or matched normal sample. Using an in-silico data augmentation approach, we trained and evaluated Fragle on nearly 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.

Fragle is the first method to report 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 low-pass whole genome sequencing (lpWGS) data [7], DNA methylation data [8, 9], or modifications to the targeted sequencing panel [10]. Using colorectal, breast, and gastric cancer plasma samples sequenced with both lpWGS and a targeted gene panel, we demonstrate high concordance of Fragle estimates across assays. Furthermore, we demonstrated increased accuracy when input data was 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 [25]. We used this feature to analyze longitudinal targeted sequencing samples from colorectal cancer patients, demonstrating strong concordance of Fragle-inferred ctDNA dynamics and tumor progression measured from radiographic imaging. This analysis also highlighted patients where the dynamic range of mutation VAFs varied extensively across different mutations and time points. Under these conditions, ctDNA quantification using Fragle could provide an orthogonal approach to interpret and resolve heterogeneous mutation VAFs profiled with targeted sequencing.

We also explored the potential for detecting minimal residual disease (MRD) with Fragle. In a cohort of early-stage lung cancer patients with MRD evaluated at the landmark timepoint

following surgery, Fragle ctDNA levels could risk-stratify patients that had otherwise been classified as ctDNA-negative using a commercial tumor-agnostic targeted sequencing assay. This result highlights the potential clinical utility of Fragle for MRD classification in settings where tumor-informed sequencing assays are not feasible or available. While tumor-informed ctDNA detection approaches offer increased MRD detection sensitivity [24], these methods impose additional requirements for tissue sample availability and logistics. In contrast, a tumor-naïve MRD-classification approach could be applied directly to a plasma sample. Indeed, our analysis demonstrates the utility of using Fragle for MRD classification as an add-on to a standard tumor-naïve targeted sequencing panel.

Overall, Fragle showed robust performance across plasma samples from 8 solid tumor types and distinct healthy cohorts. We observed strong concordance with radiographic imaging and tumor VAFs in longitudinal samples from colorectal cancer patients 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. 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 both an in-house as well as commercial targeted panel, future studies are needed to further characterize the performance using distinct sequencing panels, unseen tumor types, 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 [20]. 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 from CHIP and solid tumor 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 tumor 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 [26]. 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.

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 [10, 12, 14, 27]. Similarly, the test cohort was composed of internal samples as well as samples from a previous study [27], 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 2018/2709, 2018/2795, 2018/3046, 2019/2401, and 2012/733/B approved by the Singhealth Centralised Institutional Review Board, as well as for volunteers recruited from National University Health System (NUHS). 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 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 Illumina NovaSeq6000 system. Low-pass WGS (~4x, 2×151 bp) was performed on cfDNA samples from cancer patients and healthy individuals. We used bwa-mem [28] 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 4 cancer types using distinct orthogonal methods. 12 CRC and 10 BRCA plasma samples had ~90x cfDNA and ~30x matched buffy coat WGS data, and their ctDNA fractions were estimated using four tumor tissue-based methods as previously reported [10]. 53 CRC samples had lp-WGS and targeted promoter-region sequencing data, and we inferred ctDNA fractions by averaging ichorCNA and NDRquant estimates [10] in these samples. The remaining 26 CRC and 38 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 Suppl. Data 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 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 validation sets, respectively (see Suppl. Fig. 1).

Overview of Fragle

Fragle quantifies ctDNA levels directly 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 Suppl. Fig. 10 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. In the initial stage, we input the processed fragment length histogram to both models and obtained two ctDNA fraction estimates. In the second stage, a selection model uses these two ctDNA fraction predictions as input features. This model determined whether the sample is a high or low ctDNA burden sample and selected the corresponding ctDNA fraction estimate as the final output. To train a final Fragle model based on the discovery cohort, we first trained a Fragle pre-model on the samples to obtain their intermediate ctDNA burden estimates. We excluded the samples with a large deviation of ctDNA fractions estimated by ichorCNA and Fragle pre-model (i.e. 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%), followed by training a final Fragle model using all the remaining samples.

Fragle model feature extraction

Steps of Fragle feature extraction from fragment length have been graphically shown in Suppl. Fig 10. We computed the fragment length profile for all fragments sized 51-400 bp using the paired-end reads in the sample with Pysam [29]. 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 (column-wise z-score normalization). For each fragment length, we compared the frequency of the fragments with this length between cancer and healthy samples to examine whether cancer and healthy samples show significant differences based on the two-sided Wilcoxon rank sum test. Suppl.

Fig 11 shows the significance plot for all 350 fragment lengths, of which 281 fragment lengths were selected as features based on the P-value threshold of 0.01 (Suppl. Data 17).

High and low ctDNA burden model architecture

Detailed model architecture has been shown in Suppl. Fig 10. The model used a neural network starting with a feature embedding layer. We have 16 fully connected layers with batch normalization [30] and residual connections [31]. Dropout regularization [32] 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 the high ctDNA burden model is the quantification loss function. The high ctDNA burden model directly utilizes mean absolute error (MAE) loss, while the low ctDNA burden model uses relative MAE loss:

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

For relative MAE loss, the same absolute difference for a low-ctDNA sample would result in a higher penalty compared to a high-ctDNA sample. Here, σ is a hyperparameter tuned based on cross-validation data.

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 labeled 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 in the unseen colorectal cancer dataset (N = 109) and in the unseen metastatic gastric cancer (N = 49) 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 [28]. 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 [33] using a minimal VAF threshold of 0.05%, and annotated all variants using Variant Effect Predictor [34]. We removed low-impact variants such as synonymous variants, and low-quality variants such as those that fail to fulfill 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 [35] 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 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. In an additional analysis, we used targeted sequencing data from 44 patients profiled with the Foundation Medicine Liquid CDx assay. Samples belonged to different cancer types such as lung, liver, colon, breast, pancreas, rectum, esophagus, ovarian, uterus, gallbladder, and stomach cancer (Suppl. Data 12). We filtered known germline variants using gnomAD (v4) and analyzed variant allele frequencies 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 the panel (Suppl. Data 13).

Lung cancer survival analysis

Plasma targeted sequencing data from the MEDAL cohort (Project ID: OEP004204) [36] was retrieved from National Omics Data Encyclopedia (NODE). Alignment to the human genome (hg19) was conducted using bwa-mem [28]. Duplicates in the aligned data were marked using

Samblaster [37]. 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 2x 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 samples (n = 162) from the targeted MEDAL cohort samples using samtools.

Unseen in vitro dilution experiments

The first in vitro dilution experiment included high ctDNA burden cfDNA samples from 2 individual CRC patients which were selected to create a starting point for the dilution series. The ctDNA content for each sample was experimentally determined by 2 methods (ichorCNA [7] and NDRquant [10]), 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 a 9-point serial dilution of the ctDNA fraction for each sample (Suppl. Data 10), with 3 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, 8.7% by Fragle, sample 2: 17.2% and 14.7%, sample 3: 13.4% and 19.1%, respectively). 12 control plasma cfDNA samples were purchased from Ripple Biosolutions and were used to set up a 7-point serial dilution of the ctDNA fraction for each cancer sample (Suppl. Data 11), with 3 technical replicates. We randomly selected 3 control samples and pooled them before diluting the plasma of cancer. lp-WGS was performed with a depth of ~ 4-5x.

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 (GIS2019b). A 6-point serial dilution for each sample was set up using these healthy mixtures to dilute the 20 cancer samples (Suppl. Data 9), with 20 technical replicates. A total of 2400 dilution samples were created ranging from 5% to as low as ~0.1% ctDNA fraction, each dilution point containing 400 samples (Suppl. Data 9).

Data availability

Published data used in this study and their access codes are present in Suppl. Data 1. Data generated in this study have been deposited at the European Genome-phenome Archive (EGA; Dataset ID: EGAD50000000167). 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 18 and will be made publicly available via GitHub. The software can be applied to (1) raw lpWGS or targeted sequencing bam files, and (2) fragment length histograms generated from these bam files.

Supplementary materials

Supplementary Fig. 1-11

Supplementary Data 1-18

Disclosure

The authors declare no competing interests.

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 cfDNA cohorts.

References

1 Lui YY, Chik K-W, Chiu RW et al. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clinical chemistry 2002; 48 (3): 421-427.

2 Pantel K, Alix-Panabières C. Liquid biopsy and minimal residual disease—latest advances and implications for cure. Nature Reviews Clinical Oncology 2019; 16 (7): 409-424.

3 Sanz-Garcia E, Zhao E, Bratman SV et al. Monitoring and adapting cancer treatment using circulating tumor DNA kinetics: Current research, opportunities, and challenges. Science Advances 2022; 8 (4): eabi8618.

4 Kilgour E, Rothwell DG, Brady G et al. Liquid biopsy-based biomarkers of treatment response and resistance. Cancer cell 2020; 37 (4): 485-495.

5 Vasan N, Baselga J, Hyman DM. A view on drug resistance in cancer. Nature 2019; 575 (7782): 299-309.

6 Razavi P, Li BT, Brown DN et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nature medicine 2019; 25 (12): 1928-1937.

7 Adalsteinsson VA, Ha G, Freeman SS et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature communications 2017; 8 (1): 1324.

8 Li S, Zeng W, Ni X et al. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proceedings of the National Academy of Sciences 2023; 120 (28): e2305236120.

9 Li W, Li Q, Kang S 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): e89-e89.

10 Zhu G, Guo YA, Ho D et al. Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nature communications 2021; 12 (1): 2229.

11 Lo YMD, Chan KA, Sun H et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science translational medicine 2010; 2 (61): 61ra91-61ra91.

12 Jiang P, Chan CW, Chan KA et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proceedings of the National Academy of Sciences 2015; 112 (11): E1317-E1325.

13 Underhill HR, Kitzman JO, Hellwig S et al. Fragment length of circulating tumor DNA. PLoS genetics 2016; 12 (7): e1006162.

14 Mouliere F, Chandrananda D, Piskorz AM et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Science translational medicine 2018; 10 (466): eaat4921.

15 Cristiano S, Leal A, Phallen J et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019; 570 (7761): 385-389.

16 Foda ZH, Annapragada AV, Boyapati K et al. Detecting liver cancer using cell-free DNA fragmentomes. Cancer discovery 2023; 13 (3): 616-631.

17 Renaud G, Nørgaard M, Lindberg J et al. Unsupervised detection of fragment length signatures of circulating tumor DNA using non-negative matrix factorization. Elife 2022; 11: e71569.

18 Yu SC, Choy LL, Lo YD. ‘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: 1-9.

19 Hudecova I, Smith CG, Hänsel-Hertsch R et al. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. Genome Research 2022; 32 (2): 215-227.

20 Ptashkin RN, Mandelker DL, Coombs CC et al. Prevalence of clonal hematopoiesis mutations in tumor-only clinical genomic profiling of solid tumors. JAMA oncology 2018; 4 (11): 1589-1593.

21 Woodhouse R, Li M, Hughes J 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 2020; 15 (9): e0237802.

22 Audinot B, Drubay D, Gaspar N et al. ctDNA quantification improves estimation of outcomes in patients with high grade osteosarcoma: a translational study from the OS2006 trial. Annals of Oncology 2023.

23 Bratman SV, Yang SC, Iafolla MA et al. Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. Nature Cancer 2020; 1 (9): 873-881.

24 Chen K, Yang F, Shen H et al. Individualized tumor-informed circulating tumor DNA analysis for postoperative monitoring of non-small cell lung cancer. Cancer Cell 2023; 41 (10): 1749-1762. e1746.

25 Talevich E, Shain AH, Botton T et al. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS computational biology 2016; 12 (4): e1004873.

26 Tsui DW, Cheng ML, Shady M et al. Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. Genome medicine 2021; 13 (1): 1-15.

27 Jiang P, Sun K, Peng W et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. Cancer Discovery 2020; 10 (5): 664-673.

28 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 2013.

29 Li H, Handsaker B, Wysoker A et al. The sequence alignment/map format and SAMtools. bioinformatics 2009; 25 (16): 2078-2079.

30 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning.* pmlr; 2015:448-456.

31 He K, Zhang X, Ren S et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770-778.

32 Srivastava N, Hinton G, Krizhevsky A et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 2014; 15 (1): 1929-1958.

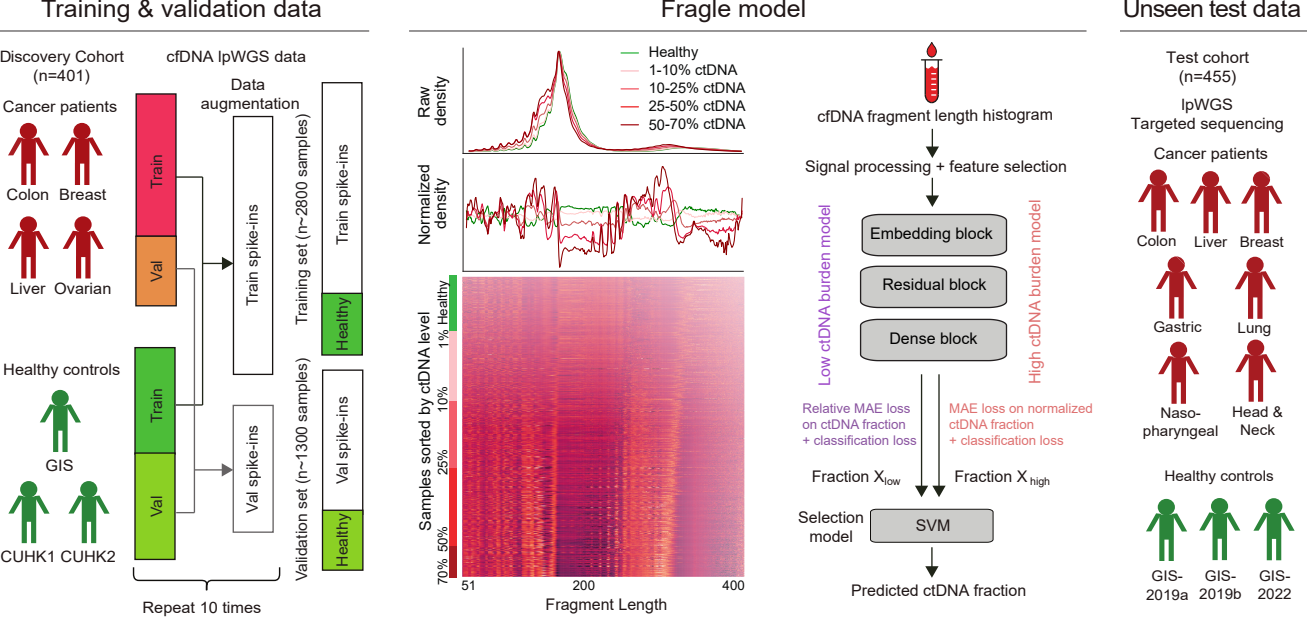
33 Lai Z, Markovets A, Ahdesmaki M et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic acids research 2016; 44 (11): e108-e108.

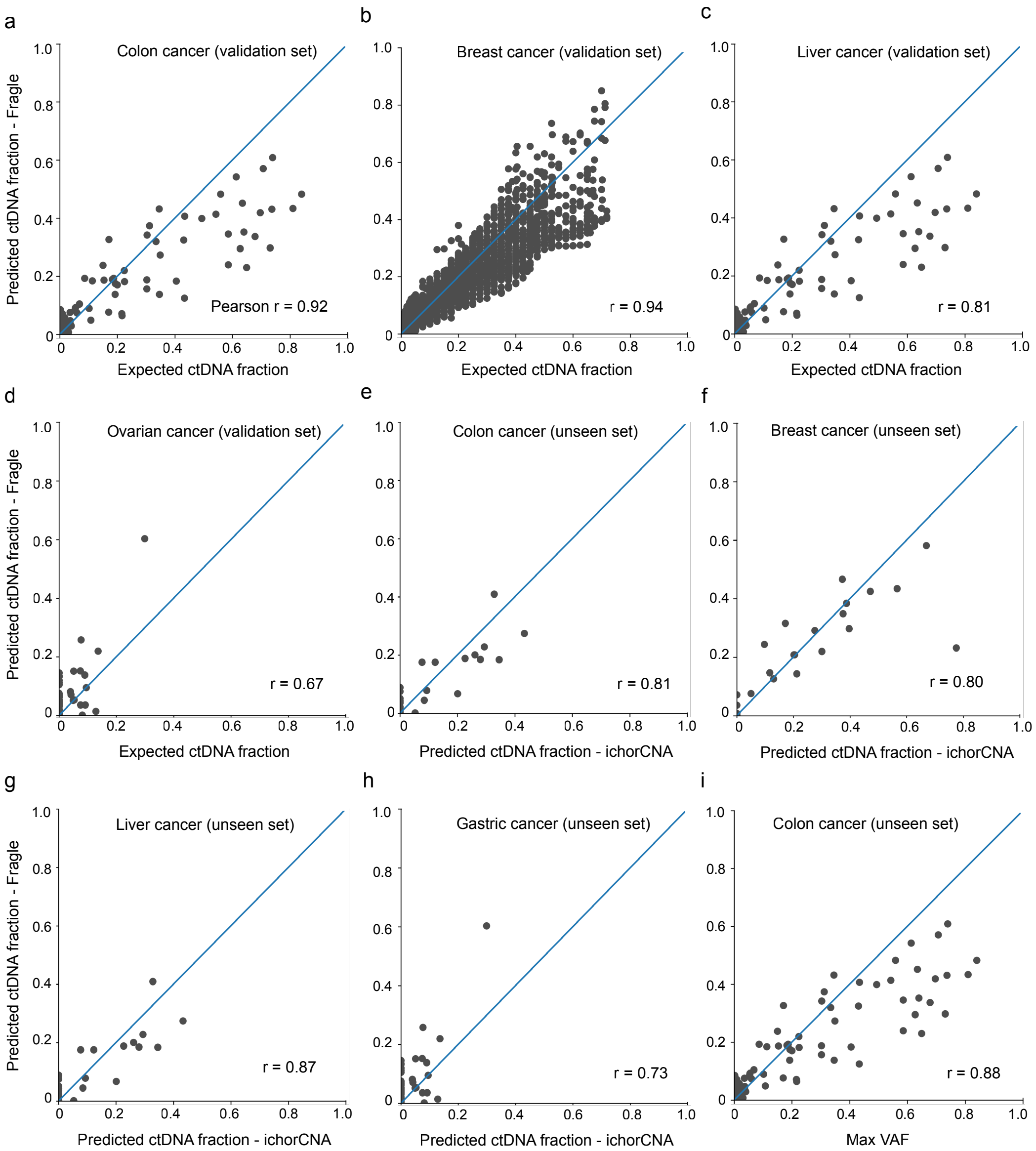
34 McLaren W, Gil L, Hunt SE et al. The ensembl variant effect predictor. Genome biology 2016; 17 (1): 1-14.

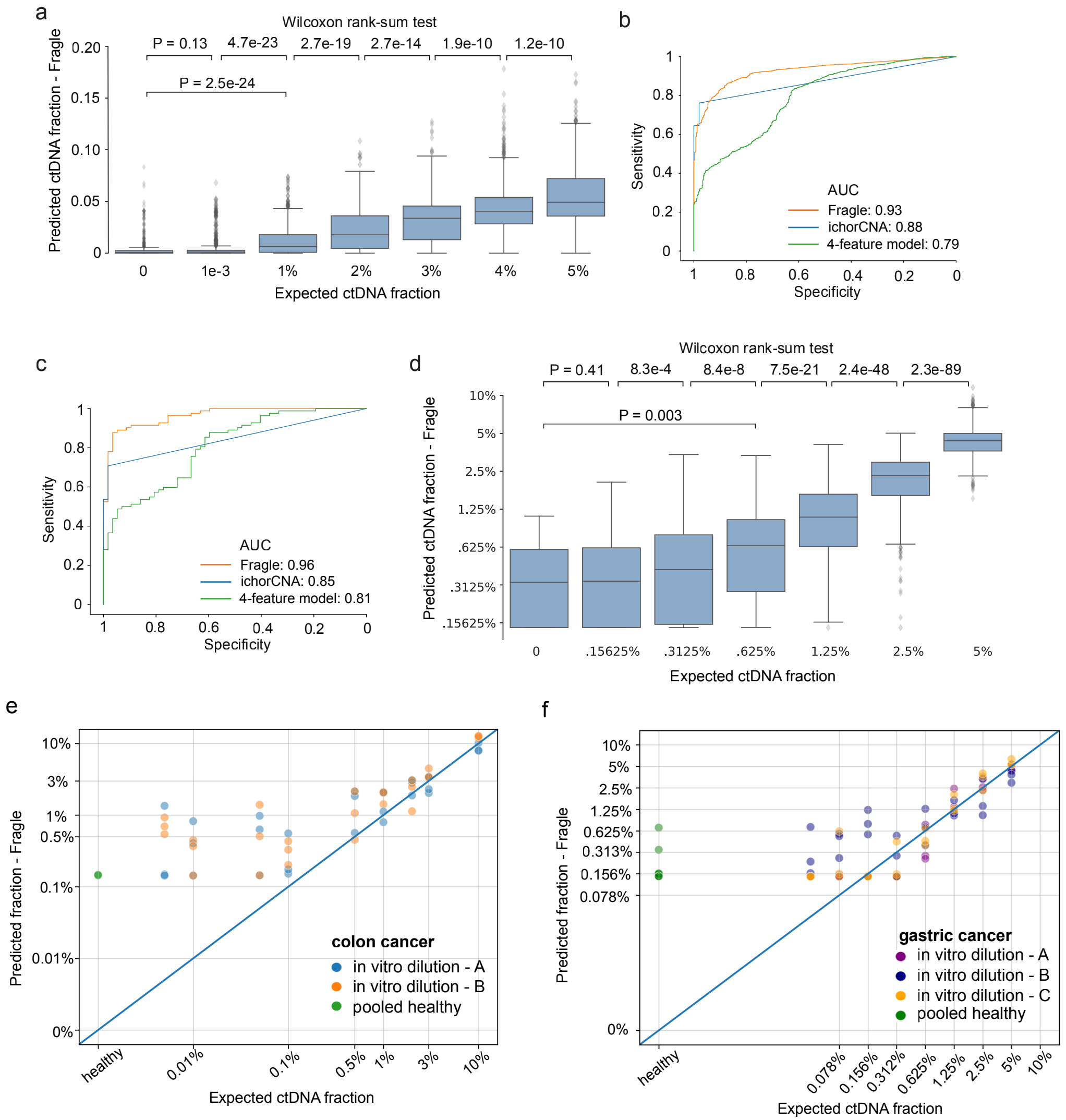
35 Mansukhani S, Barber LJ, Kleftogiannis D et al. Ultra-sensitive mutation detection and genome-wide DNA copy number reconstruction by error-corrected circulating tumor DNA sequencing. Clinical chemistry 2018; 64 (11): 1626-1635.

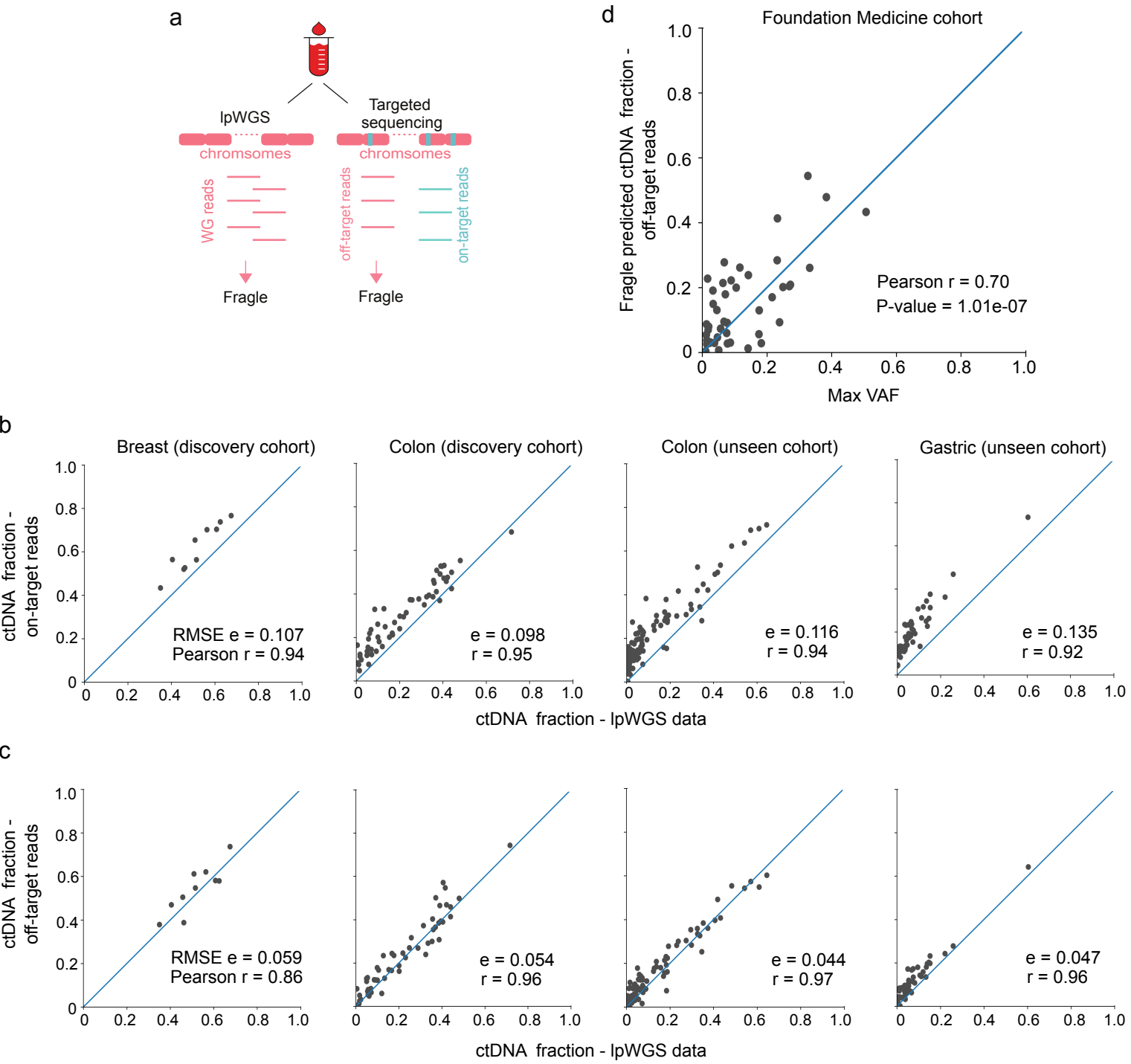
36 Chen K, Kang G, Zhang Z et al. Individualized dynamic methylation-based analysis of cell-free DNA in postoperative monitoring of lung cancer. BMC medicine 2023; 21 (1): 255.

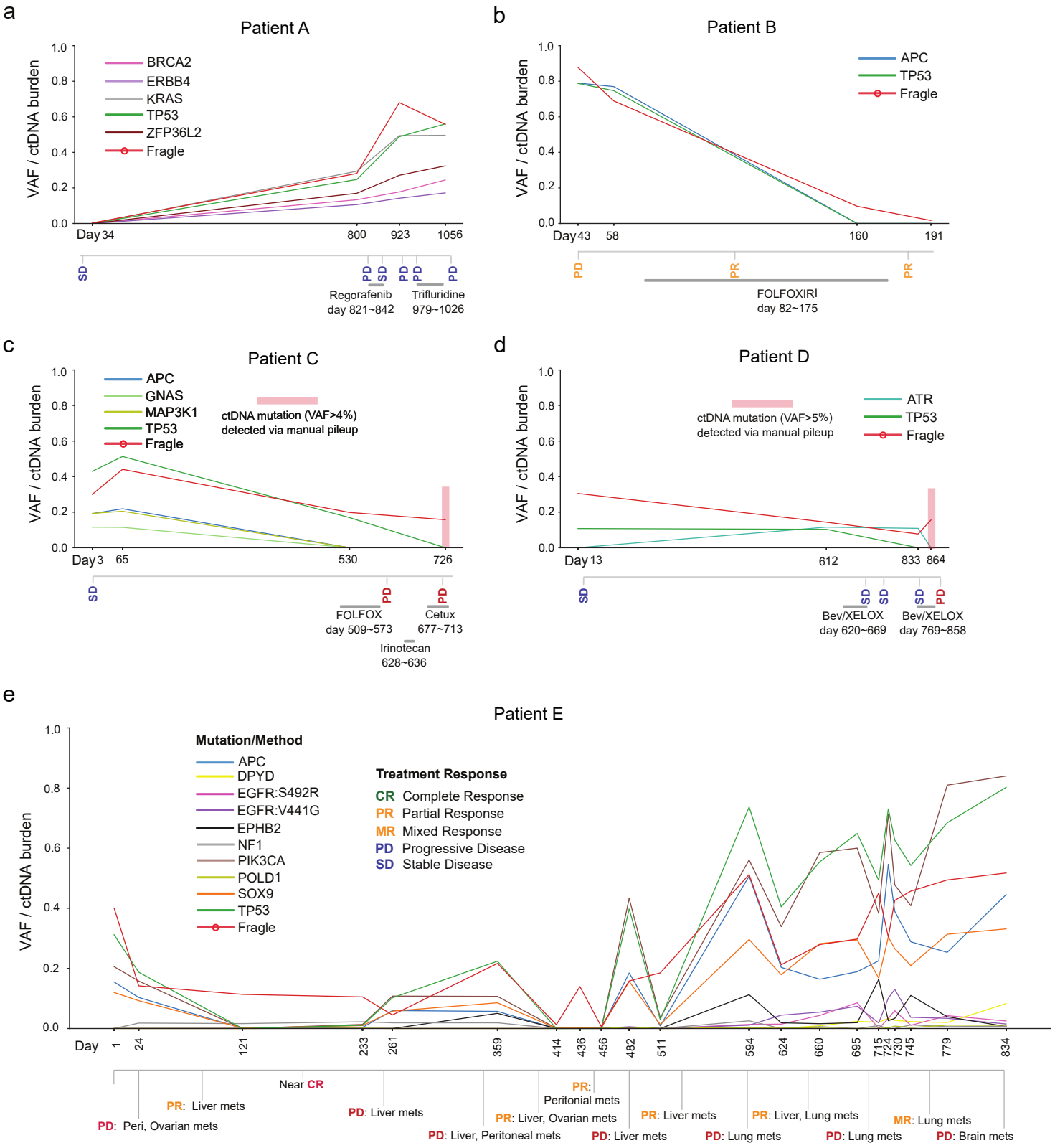
37 Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 2014; 30 (17): 2503-2505.

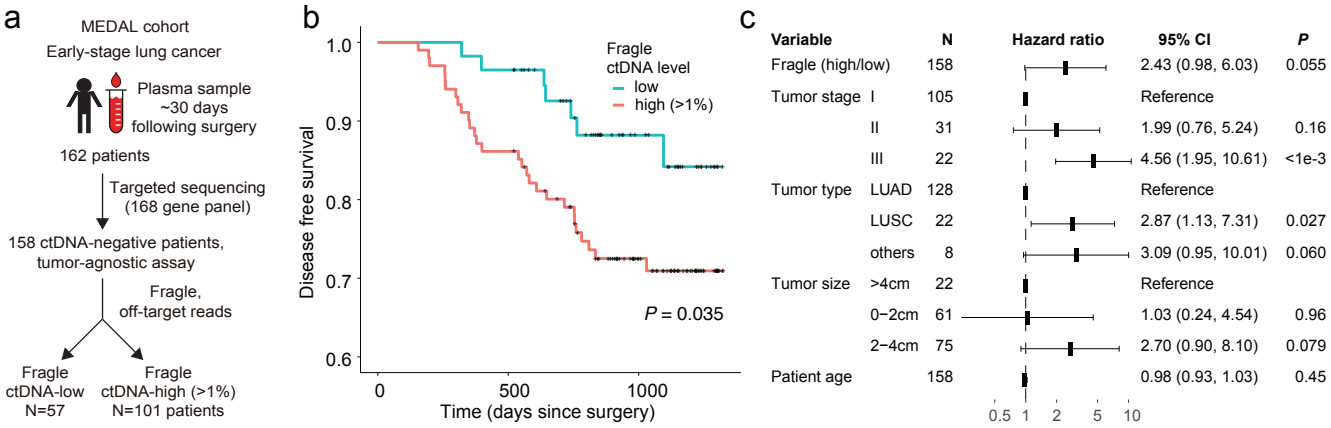

Training & validation data
Fragle model
Unseen test data
Discovery Cohort
(n=401)
cfDNA lpWGS data
Cancer patients
Colon
Breast
Liver
Ovarian
Healthy controls
GIS
CUHK1
CUHK2
Data augmentation
Train
Val
Train
Val
Train spike-ins
Val spike-ins
Training set (n~2800 samples)
Validation set (n~1300 samples)
Train spike-ins
Healthy
Val spike-ins
Healthy
Repeat 10 times
Healthy
1-10% ctDNA
10-25% ctDNA
25-50% ctDNA
50-70% ctDNA
Raw density
Normalized density
Samples sorted by ctDNA level
1% Healthy
10%
25%
50%
70%
51
200
400
Fragment Length
cfDNA fragment length histogram
Signal processing + feature selection
Embedding block
Residual block
Dense block
Low ctDNA burden model
High ctDNA burden model
Relative MAE loss on ctDNA fraction + classification loss
MAE loss on normalized ctDNA fraction + classification loss
Fraction X low
Fraction X high
Selection model
SVM
Predicted ctDNA fraction
Test cohort
(n=455)
lpWGS
Targeted sequencing
Cancer patients
Colon
Liver
Breast
Gastric
Lung
Naso-pharyngeal
Head & Neck
Healthy controls
GIS-2019a
GIS-2019b
GIS-2022


a
Colon cancer (validation set)
Pearson r = 0.92
Predicted ctDNA fraction - Fragle
Expected ctDNA fraction
b
Breast cancer (validation set)
r = 0.94
Expected ctDNA fraction
c
Liver cancer (validation set)
r = 0.81
Expected ctDNA fraction
d
Ovarian cancer (validation set)
r = 0.67
Predicted ctDNA fraction - Fragle
Expected ctDNA fraction
e
Colon cancer (unseen set)
r = 0.81
Predicted ctDNA fraction - ichorCNA
f
Breast cancer (unseen set)
r = 0.80
Predicted ctDNA fraction - ichorCNA
g
Liver cancer (unseen set)
r = 0.87
Predicted ctDNA fraction - Fragle
Predicted ctDNA fraction - ichorCNA
h
Gastric cancer (unseen set)
r = 0.73
Predicted ctDNA fraction - ichorCNA
i
Colon cancer (unseen set)
r = 0.88
Max VAF


a
Wilcoxon rank-sum test
P = 0.13
4.7e-23
2.7e-19
2.7e-14
1.9e-10
1.2e-10
P = 2.5e-24
Predicted ctDNA fraction - Fragle
0.20
0.15
0.10
0.05
0
0
1e-3
1%
2%
3%
4%
5%
Expected ctDNA fraction
b
Sensitivity
Specificity
AUC
Fragle: 0.93
ichorCNA: 0.88
4-feature model: 0.79
c
Sensitivity
Specificity
AUC
Fragle: 0.96
ichorCNA: 0.85
4-feature model: 0.81
d
Wilcoxon rank-sum test
P = 0.41
8.3e-4
8.4e-8
7.5e-21
2.4e-48
2.3e-89
P = 0.003
Predicted ctDNA fraction - Fragle
10%
5%
2.5%
1.25%
.625%
.3125%
.15625%
0
.15625%
.3125%
.625%
1.25%
2.5%
5%
Expected ctDNA fraction
e
Predicted fraction - Fragle
10%
3%
1%
0.5%
0.1%
0.01%
0%
healthy
0.01%
0.1%
0.5%
1%
3%
10%
Expected ctDNA fraction
colon cancer
in vitro dilution - A
in vitro dilution - B
pooled healthy
f
Predicted fraction - Fragle
10%
5%
2.5%
1.25%
0.625%
0.313%
0.156%
0.078%
0%
healthy
0.078%
0.156%
0.312%
0.625%
1.25%
2.5%
5%
10%
Expected ctDNA fraction
gastric cancer
in vitro dilution - A
in vitro dilution - B
in vitro dilution - C
pooled healthy


a
lpWGS
Targeted sequencing
chromsomes
chromsomes
WG reads
off-target reads
on-target reads
Fragle
Fragle
d
Foundation Medicine cohort
Fragle predicted ctDNA fraction - off-target reads
Max VAF
Pearson r = 0.70
P-value = 1.01e-07
b
Breast (discovery cohort)
Colon (discovery cohort)
Colon (unseen cohort)
Gastric (unseen cohort)
ctDNA fraction - on-target reads
RMSE e = 0.107
Pearson r = 0.94
e = 0.098
r = 0.95
e = 0.116
r = 0.94
e = 0.135
r = 0.92
ctDNA fraction - lpWGS data
c
ctDNA fraction - off-target reads
RMSE e = 0.059
Pearson r = 0.86
e = 0.054
r = 0.96
e = 0.044
r = 0.97
e = 0.047
r = 0.96
ctDNA fraction - lpWGS data


a
Patient A
VAF / ctDNA burden
BRCA2
ERBB4
KRAS
TP53
ZFP36L2
Fragle
Day 34
800
923
1056
SD
PD
SD
PD
PD
PD
Regorafenib
day 821~842
Trifluridine
979~1026
b
Patient B
APC
TP53
Fragle
Day 43
58
160
191
PD
PR
PR
FOLFOXIRI
day 82~175
c
Patient C
APC
GNAS
MAP3K1
TP53
Fragle
ctDNA mutation (VAF>4%)
detected via manual pileup
Day 3
65
530
726
SD
PD
PD
FOLFOX
day 509~573
Cetux
677~713
Irinotecan
628~636
d
Patient D
ctDNA mutation (VAF>5%)
detected via manual pileup
ATR
TP53
Fragle
Day 13
612
833
864
SD
SD
SD
SD
PD
Bev/XELOX
day 620~669
Bev/XELOX
day 769~858
e
Patient E
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
1
24
121
233
261
359
414
436
456
482
511
594
624
660
695
715
724
730
745
779
834
PD: Peri, Ovarian mets
PR: Liver mets
Near CR
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


a
MEDAL cohort
Early-stage lung cancer
Plasma sample
~30 days
following surgery
162 patients
Targeted sequencing
(168 gene panel)
158 ctDNA-negative patients,
tumor-agnostic assay
Fragle,
off-target reads
Fragle
ctDNA-low
N=57
Fragle
ctDNA-high (>1%)
N=101 patients
b
Fragle
ctDNA level
low
high (>1%)
P = 0.035
Disease free survival
1.0
0.9
0.8
0.7
0.6
0
500
1000
Time (days since surgery)
c
Variable
N
Hazard ratio
95% CI
P
Fragle (high/low) 158 2.43 (0.98, 6.03) 0.055
Tumor stage I 105 Reference
II 31 1.99 (0.76, 5.24) 0.16
III 22 4.56 (1.95, 10.61) <1e-3
Tumor type LUAD 128 Reference
LUSC 22 2.87 (1.13, 7.31) 0.027
others 8 3.09 (0.95, 10.01) 0.060
Tumor size >4cm 22 Reference
0−2cm 61 1.03 (0.24, 4.54) 0.96
2−4cm 75 2.70 (0.90, 8.10) 0.079
Patient age 158 0.98 (0.93, 1.03) 0.45
0.5
1
2
5
10