# Quantification of circulating tumor DNA using deep learning

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) in blood enables non-invasive surveillance of cancer progression. *Fragle* is a fast deep learning-based method for estimation of ctDNA levels directly from cell-free DNA fragment length profiles. *Fragle* accurately estimates ctDNA levels across different cancer types from both low-pass whole genome and targeted sequencing data, enabling uniform quantification of blood ctDNA levels across sequencing assays and cancer patients.

## 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-cancer 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 [12-14]. Here, we developed *Fragle*, a multi-stage machine learning model that quantifies ctDNA levels directly from a cfDNA fragment length density distribution (Fig. 1a). *Fragle* is based on two parallel deep learning-based sub-models optimized for prediction in high (≥3%) and low (<3%) ctDNA burden samples. The two sub-models differ with respect to loss function, model architecture, as well as regions used to generate the input fragment length features, and the final prediction is based on a support vector machine (SVM) selection model (Methods).

## Results

We assembled a discovery cohort comprising lpWGS data from 213 cancer plasma samples (4 tumor types) and 101 plasma samples from healthy individuals (Fig. 1a, 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 of cancer and healthy samples with variable ctDNA fractions (Methods, Fig. 1a, Suppl. Data 3). The model was trained and evaluated using cross-validation, demonstrating high predictive accuracy on validation samples across all 4 cancer types (Fig. 1b): Colorectal (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 agreement for ovarian cancer could be attributed to two outlier samples, removal of whom increased the correlation to 0.90 (Suppl. Fig. 2). The model predicted a very low ctDNA fraction (median = 0.07%) for the healthy samples in the validation sets, with 94% of healthy samples predicted to comprise <1% ctDNA (Fig. 1c). We also observed that the model could quantitatively differentiate healthy from low-ctDNA samples at around 0.1%-1% ctDNA level (Fig. 1c), 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. 1d), improving on the performance of ichorCNA (AUC=0.88). Expectedly, the AUC increased further after filtering out low-ctDNA burden samples (Suppl. Fig. 3).

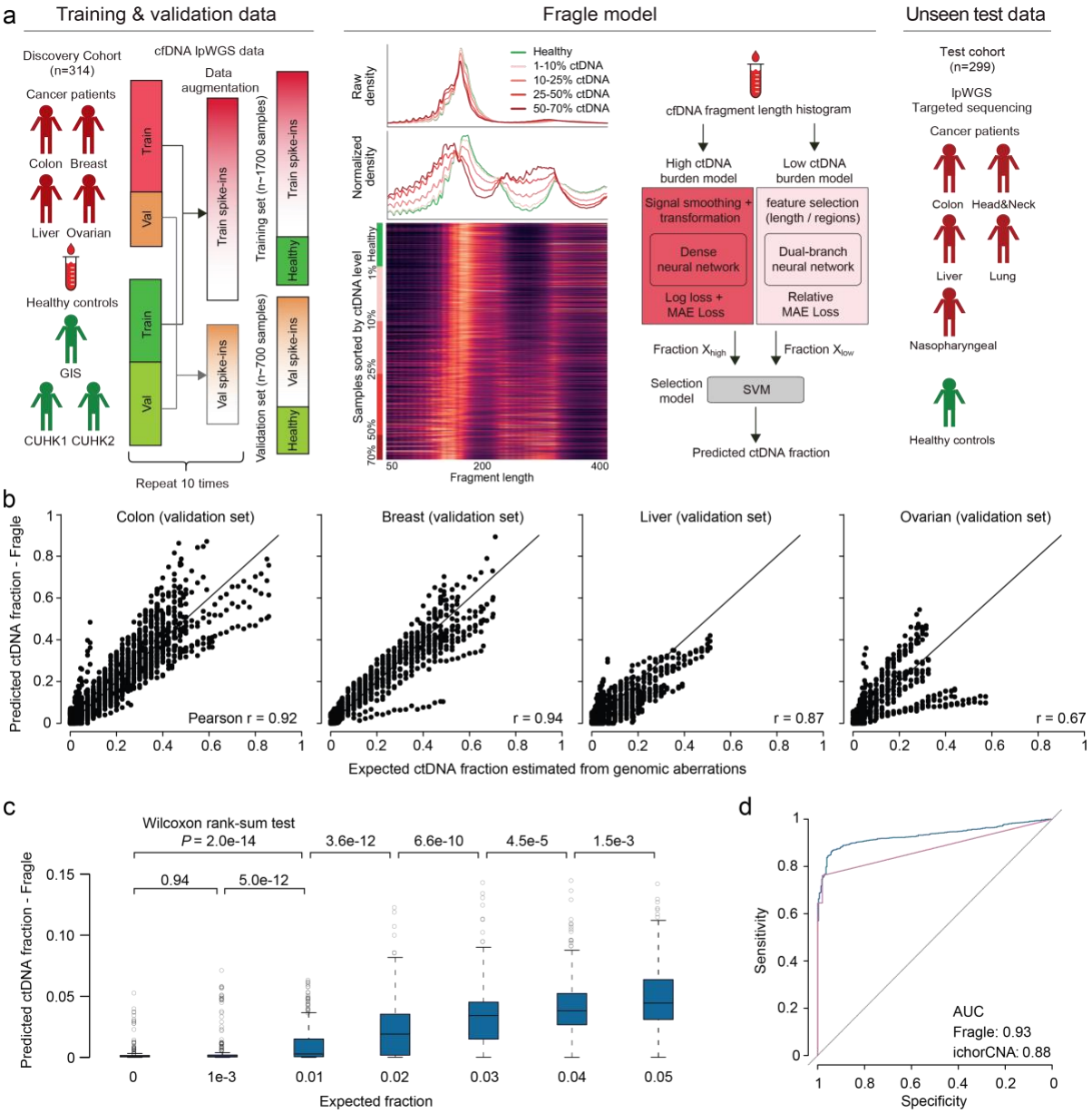



**Fig. 1, Overview of Fragle**. **a**) Fragle is a multi-stage machine learning-based model that estimates the ctDNA level in a blood sample from the cfDNA fragment length density distribution. Fragle is trained using a large-scale data augmentation approach. **b**) Comparison between expected and predicted ctDNA levels for colorectal (CRC), breast (BRCA), liver (HCC), and ovarian (OV) cancer samples in the validation sets. **c**) Predicted ctDNA fractions for healthy and low-ctDNA level samples in the validation sets. Boxplots are represented by median and interquartile range (IQR), with +/–1.5 IQR as whiskers. **d**) Classification of validation healthy control and cancer (≥1% ctDNA) samples.

We trained Fragle 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 ctDNA fraction estimates across an unseen cohort of colorectal cancer ($r$=0.82; $P$=2.6e-48; N=190; Fig. 2a), an unseen cohort of liver cancer ($r$ =0.90; $P$=4.1e-13; N=34; Fig. 2b), and a mixed cohort of cancer types not included in the discovery set (Pearson $r$ =0.54; $P$=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. 2c). To further evaluate the LoD, we used these 86 samples 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. 2d). These results were further confirmed using an in vitro dilution experiment comprising serial dilution of two high-ctDNA burden CRC samples 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. 2e). 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). Fragle was developed using samples each comprising 10 million cfDNA fragments, equivalent to ~1x WGS using 151bp paired-end sequencing. To further evaluate the effect of sequencing depth, down-sampling analysis in the test cohort demonstrated accurate quantification down to 1 million fragments, equivalent to ~0.1x WGS (Pearson $r$ = 0.97; Fig. 2f; Suppl. Fig. 6).

To explore the accuracy of Fragle when applied to targeted gene sequencing data, we evaluated the concordance across three cohorts having both lpWGS and targeted sequencing data (Fig. 2g, Methods). Using standard on-target reads obtained from the targeted sequencing data, Fragle overestimated the ctDNA burden as compared to the lpWGS data (Suppl. Fig. 7). 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 strong concordance between lpWGS and off-target reads from the targeted sequencing across all three cohorts ($r$ = 0.87-0.95; P < 0.0012; Fig. 2g). Next, we explored how Fragle could add value in a clinical setting where targeted sequencing may be preferred to monitor for treatment response, relapse, and disease evolution. We considered a metastatic colorectal cancer patient for whom we had collected 21 serial blood plasma samples over a treatment course of 3 years. Using deep targeted sequencing, we observed strong temporal correlation of Fragle ctDNA levels predicted using off-target reads, mutation VAFs, and treatment response (Fig. 2h). Moreover, while the dynamic range of VAFs varied extensively across different mutations and time points, Fragle provided an orthogonal and independent measure of ctDNA levels.

## Discussion

In summary, Fragle enables accurate quantification of plasma ctDNA levels with a ~1% LoD. Fragle is fast and flexible, estimating ctDNA levels in less than a minute (Suppl. Fig. 8) using aligned paired-end reads from both lpWGS and targeted sequencing. By enabling orthogonal

ctDNA quantification from targeted sequencing data, Fragle could limit the need for running multiple assays for disease monitoring and interpretation of negative results from plasma genotyping [15], facilitating simultaneous discovery of actionable cancer mutations and accurate estimation of ctDNA levels.

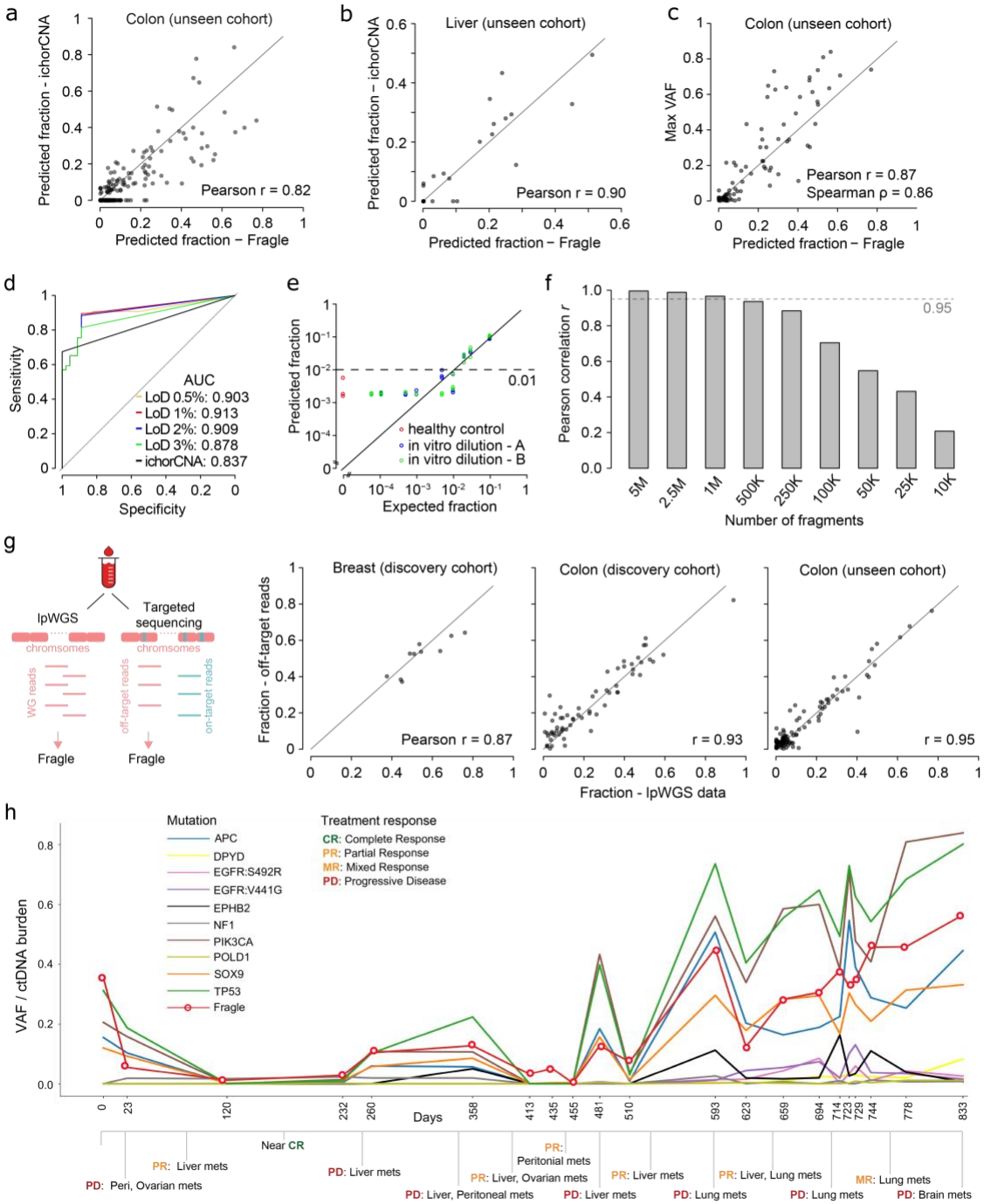



**Fig. 2, Validation of Fragle**. Comparison of ichorCNA and Fragle predicted ctDNA levels in test samples from **a**) colorectal cancer patients (N=190) and **b**) liver cancer patients (N=34). **c**) Comparison of predicted ctDNA levels and maximum VAFs in colorectal cancer patients (N = 86). **d**) ROC analyses for classification of cancer (N=86) and healthy (N=45) plasma samples based on different LoD thresholds. **e**) Expected vs. predicted ctDNA fractions in two in-vitro ctDNA dilution series. **f**) Correlation between predicted ctDNA fractions (10 million fragments) and down-sampled counterparts. **g**)

Comparison of predicted ctDNA levels for Fragle applied to lpWGS and targeted sequencing data (off-target reads) obtained from same samples. **h**) Simultaneous longitudinal profiling of Fragle ctDNA levels and mutation VAFs in a metastatic colorectal cancer patient using plasma targeted sequencing.

## 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, 16]. Similarly, the test cohort was composed of internal samples as well as samples from a previous study [16], 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 [17] 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 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 was normalized using the highest observed count, followed by a moving average normalization ($\log_{10}$ z-score of 32-nt window). The transformed length histogram for a given sample was further standardized relative to the training set. The model uses a neural network consisting of 16 fully connected layers with batch normalization and residual connections, using a composite loss function combining L1 loss (ctDNA fraction) and binary cross entropy loss (cancer vs. healthy classification).

**Low ctDNA burden model**

We computed fragment length histograms (51-400 bp) in 10 Mb genomic bins using pysam. 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 [18], 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, $\sigma$ 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.

**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 [17]. 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 [19] using a minimal VAF threshold of 0.05%, and annotated all variants using Variant Effect Predictor [20]. 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 [21] 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 12 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.

**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. 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.
16. 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.
17. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.
18. Shrikumar, A., P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*. in *International conference on machine learning*. 2017. PMLR.
19. 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.
20. McLaren, W., et al., *The ensembl variant effect predictor.* Genome biology, 2016. **17**(1): p. 1-14.
21. 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.

Fig. 1

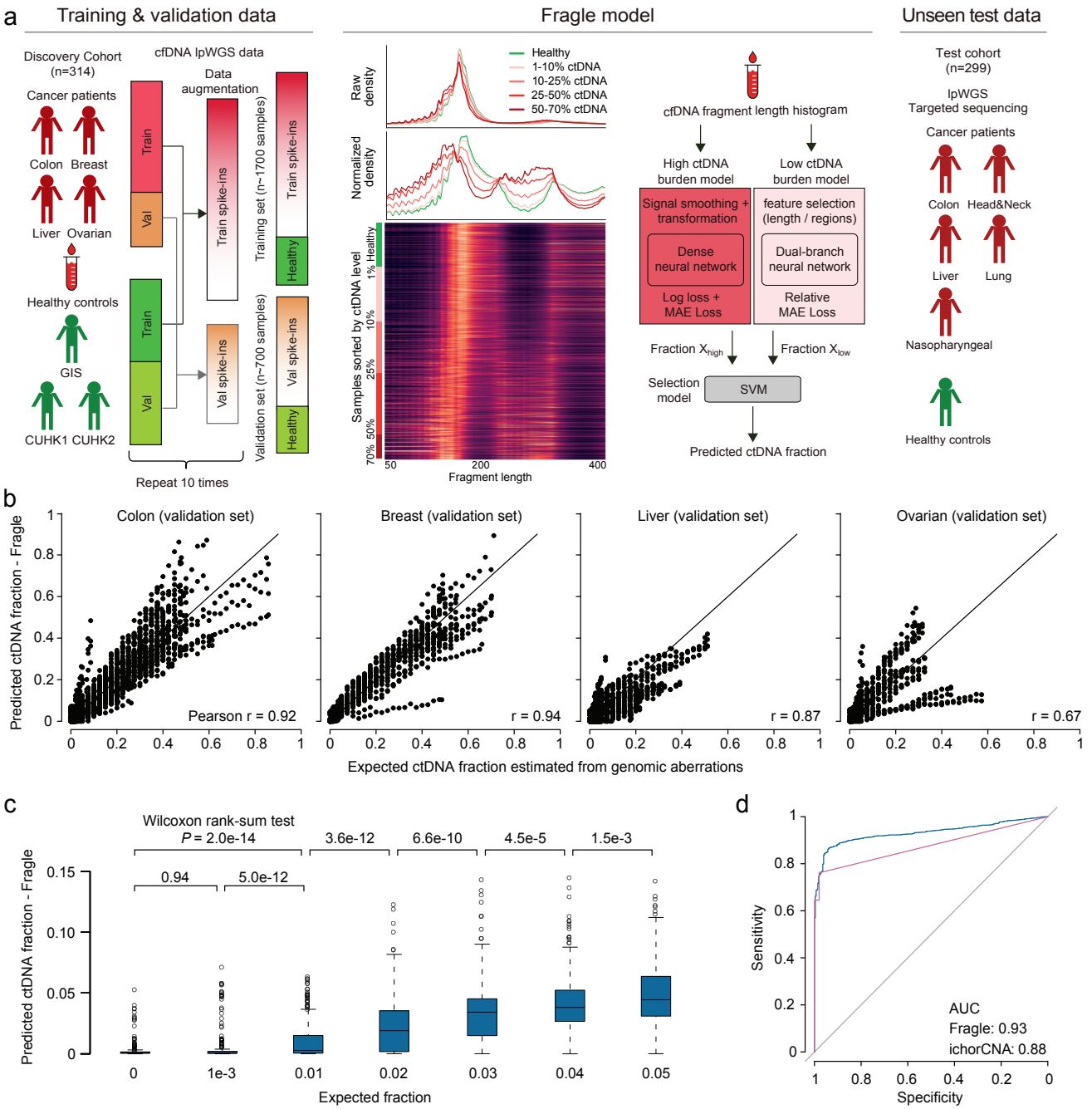
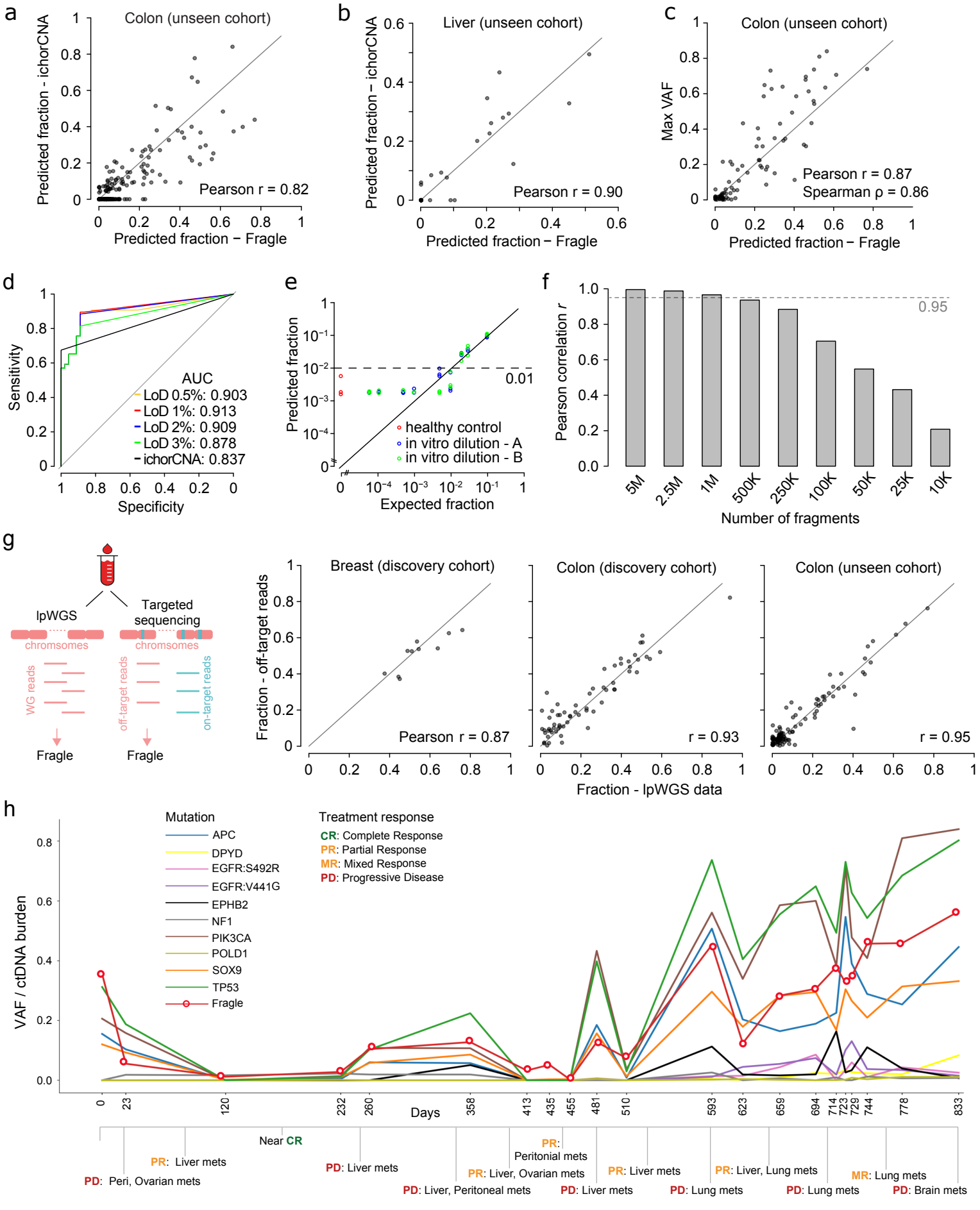

a
Training & validation data
Fragle model
Unseen test data
Discovery Cohort (n=314)
cfDNA lpWGS data
Data augmentation
Cancer patients
Colon
Breast
Liver
Ovarian
Healthy controls
GIS
CUHK1
CUHK2
Train
Val
Train spike-ins
Val spike-ins
Training set (n~1700 samples)
Validation set (n~700 samples)
Healthy
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%
Fragment length
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
b
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
c
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
Expected fraction
d
Sensitivity
Specificity
AUC
Fragle: 0.93
ichorCNA: 0.88

Fig. 2


a
Colon (unseen cohort)
Predicted fraction - ichorCNA
Predicted fraction – Fragle
Pearson r = 0.82
b
Liver (unseen cohort)
Predicted fraction – ichorCNA
Predicted fraction – Fragle
Pearson r = 0.90
c
Colon (unseen cohort)
Max VAF
Predicted fraction – Fragle
Pearson r = 0.87
Spearman ρ = 0.86
d
Sensitivity
Specificity
AUC
LoD 0.5%: 0.903
LoD 1%: 0.913
LoD 2%: 0.909
LoD 3%: 0.878
ichorCNA: 0.837
e
Predicted fraction
Expected fraction
0.01
healthy control
in vitro dilution - A
in vitro dilution - B
f
Pearson correlation r
0.95
5M
2.5M
1M
500K
250K
100K
50K
25K
10K
Number of fragments
g
lpWGS
Targeted sequencing
chromsomes
WG reads
off-target reads
on-target reads
Fragle
Breast (discovery cohort)
Colon (discovery cohort)
Colon (unseen cohort)
Fraction - off-target reads
Pearson r = 0.87
r = 0.93
r = 0.95
Fraction - lpWGS data
h
Mutation
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
VAF / ctDNA burden
Days
Near CR
PR: Liver mets
PD: Peri, Ovarian mets
PD: Liver mets
PR: Liver, Ovarian mets
PD: Liver, Peritoneal 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

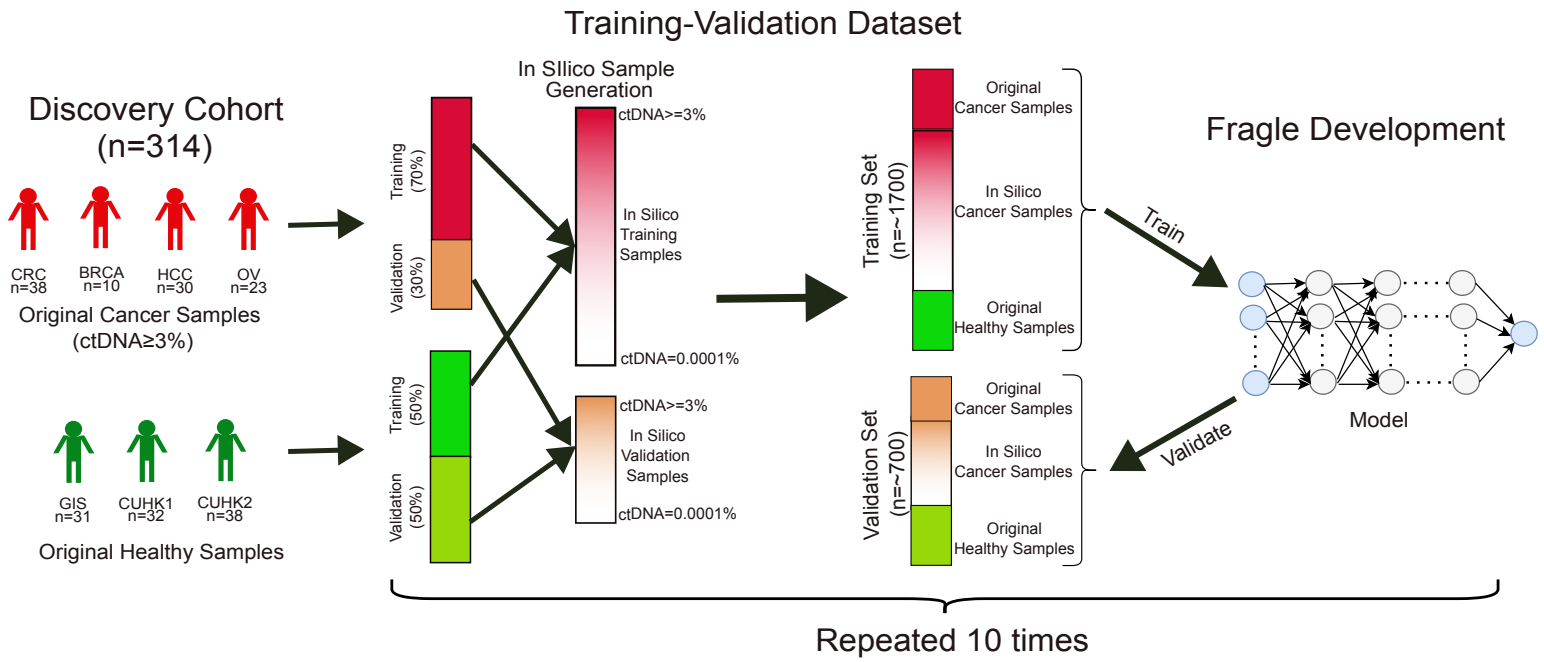



**Fig. S1 Schematic showing how the predictive models were trained and validated using the original cancer samples, healthy control samples and in silico cancer spike-ins in the discovery cohort**

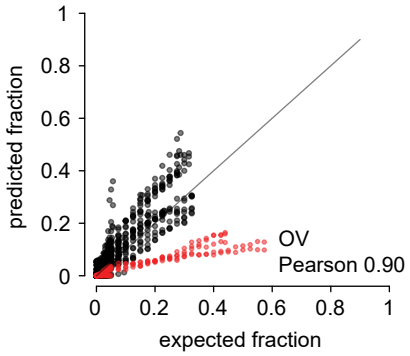



**Fig. S2 Comparison between expected and predicted ctDNA fractions of the plasma samples of OV in the validation set.** The outlier samples (red) were excluded, inporving Pearson correlation to 0.90 and MAE to 0.03

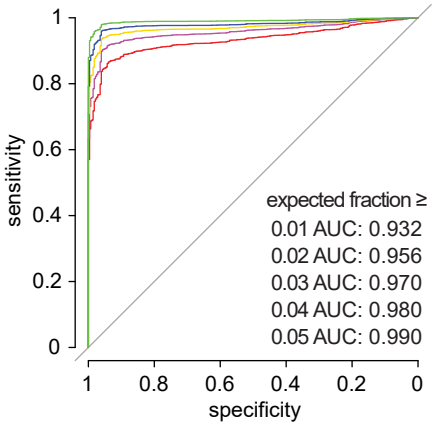



**Fig. S3 ROC analysis to examine the diagnostic performance on the validation set samples with different threshold to exclude low ctDNA samples**

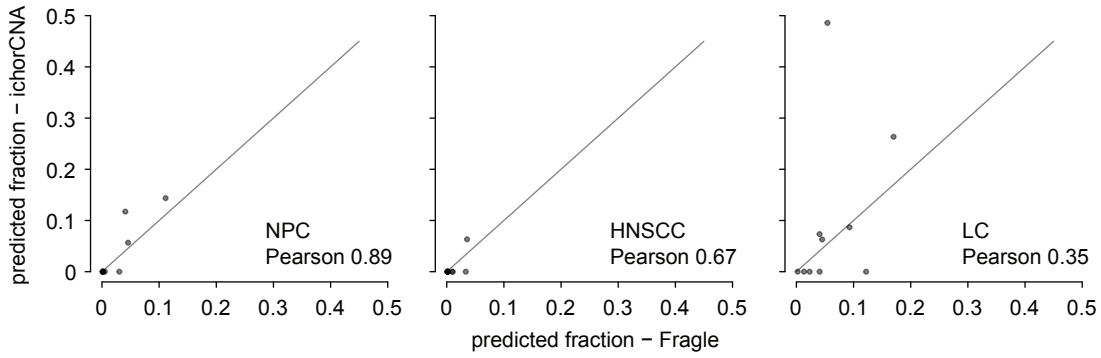



**Fig. S4 Comparison between Fragle and ichorCNA in their predicted ctDNA fractions in unseen samples from LC, NPC, and HNSCC patients.**

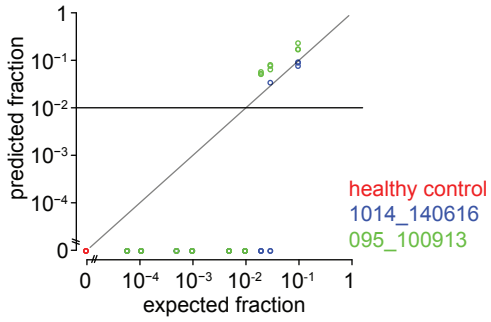



**Fig. S5 ctDNA fractions inferred by ichorCNA in in vitro cfDNA spike-in samples of low ctDNA burden.**

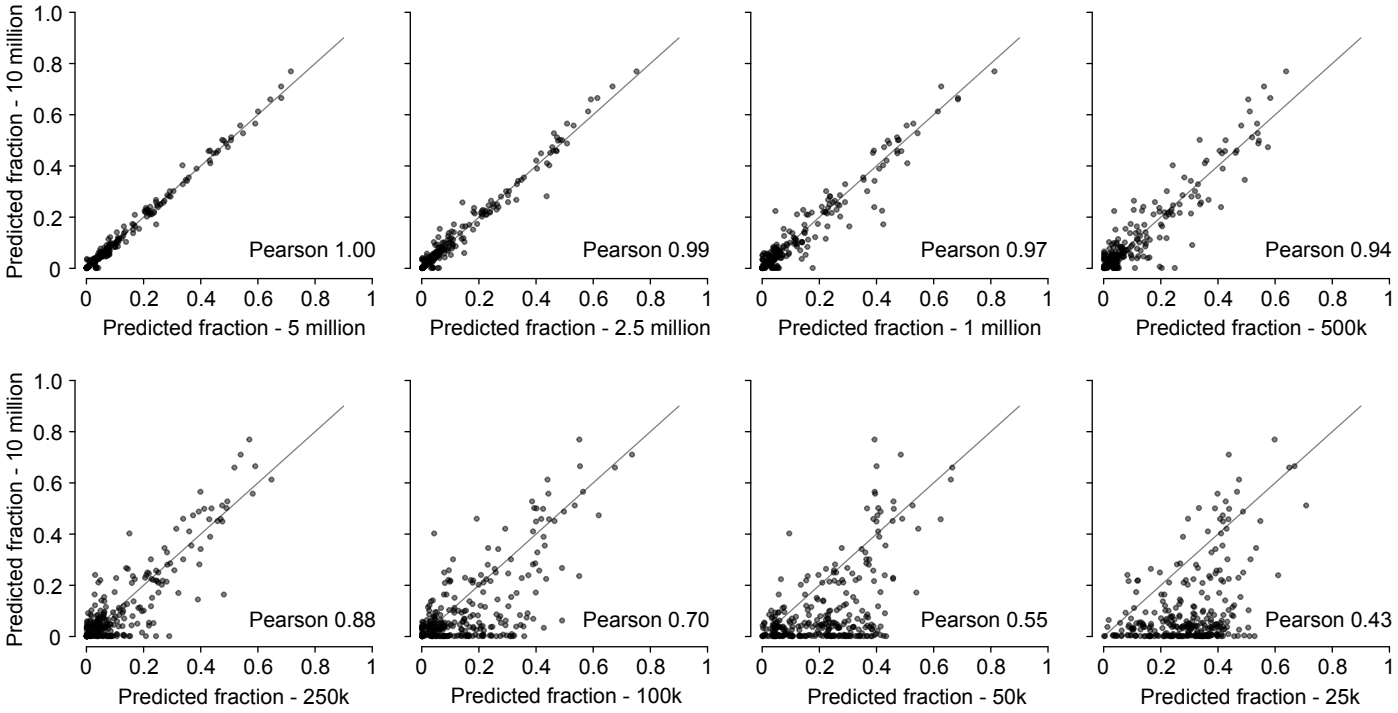



**Fig. S6 Comparison of the predicted ctDNA fractions in the unseen cfDNA WGS samples with 10 million cfDNA fragments and their downsampled counterparts.**

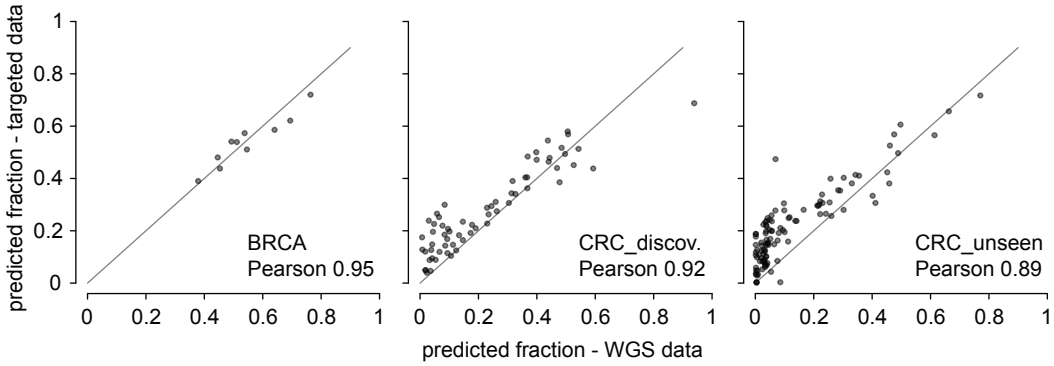



**Fig. S7 Comparison of predicted ctDNA fractions in cfDNA samples between WGS and on-target reads from targeted data for the samples in BRCA_discovery, CRC_discovery, and CRC_unseen.**

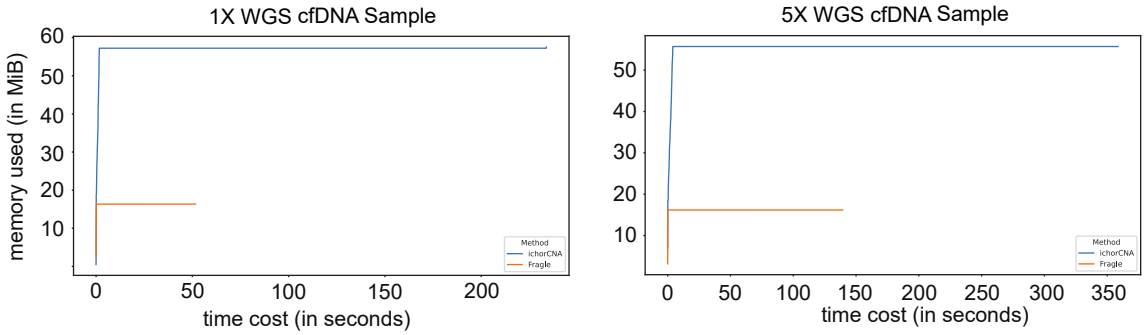



**Fig. S8 Runtime and memory usage for Fragle and ichorCNA to compute ctDNA fraction from input bam files.**