

ROBUST QUANTIFICATION OF CIRCULATING TUMOUR DNA THROUGH FRAGMENT LENGTH ANALYSIS

Technical Field

[0001] This disclosure generally relates to methods and systems for analysis of cell free DNA sequencing data from blood plasma.

Background

[0002] This background description is provided for the purpose of generally presenting the context of the disclosure. Contents of this background section are neither expressly nor impliedly admitted as prior art against the present disclosure.

[0003] The death of non-malignant cells, primarily of the hematopoietic lineage, releases cell-free DNA (cfDNA) into the blood circulation. In cancer patients, the blood plasma also carries circulating tumour DNA (ctDNA), enabling non-invasive diagnostics and disease surveillance. The ability to monitor tumour growth dynamics based on ctDNA levels in the blood provides a promising non-invasive approach to track disease progression during therapy and clinical trials.

[0004] 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 tumours will have mutations covered by a given targeted sequencing gene panel. Furthermore, the accuracy of this approximation depends on sample-specific and treatment-dynamic properties such as mutation clonality, copy number, as well as potential confounding noise from clonal hematopoiesis.

[0005] 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, DNA methylation profiling, or modifications to the targeted sequencing panel. Thus, there is an unmet need to develop accurate and orthogonal approaches for ctDNA quantification that can generalize across patients, tumour types, and sequencing modalities.

[0006] 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. cfDNA fragments from cancer patients tend to be shorter and more variably sized than those from healthy individuals. These observations have motivated studies exploring how cfDNA fragment length properties can be used to classify cfDNA samples from cancer patients and healthy individuals.

[0007] It is challenging to quantify circulating tumour DNA (ctDNA) in blood plasma using cell-free DNA (cfDNA) sequencing. Currently there are limited solutions tailored for measuring ctDNA fraction. Thus, there is an unmet need to develop approaches for quantification of ctDNA.

Summary

[0008] Disclosed is a system for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the system comprising at least one processor configured to:

- receive low-pass cell free DNA (cfDNA) sequencing data obtained by processing the plasma sample using a sequencing platform; and
- perform fragment length analysis on the sequencing data using a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample; wherein the ctDNA quantification model comprises a machine learning integrative model trained using a training dataset generated using cfDNA data augmented with cfDNA data obtained from known cancer and healthy samples to generate artificial training dataset records with a known ctDNA metric.

[0009] Also disclosed is a system for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the system comprising at least one processor configured to:

- receive targeted sequencing data obtained by processing the plasma sample by a sequencing platform, the targeted sequencing data comprising a fragment length histogram feature of an off-target portion of the targeted sequencing data; and
- feed the fragment length histogram feature to a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample; wherein the ctDNA quantification model comprises a neural network integrative model trained using a training dataset comprising labelled sequencing data of off-

target read associated fragments originating from healthy and cancer afflicted individuals.

[0010] As used herein, "off-target read" refers to those reads that are obtained during high-throughput targeted sequencing (such as next-generation sequencing) and are not aligned with the intended target region or gene. They spread throughout the whole genome and provides a similar effect as shallow whole genome sequencing

[0011] Also disclosed is a method for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the method comprising:

- processing the plasma sample by a sequencing platform to obtain low-pass cell free DNA (cfDNA) sequencing data; and

- performing fragment length analysis of the sequencing data using a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample; wherein the ctDNA quantification model comprises a machine learning integrative model trained using a training dataset generated using cfDNA data augmented with ctDNA data obtained from known cancer samples to generate artificial training dataset records with a known ctDNA metric.

[0012] Also disclosed is a method for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the method comprising:

- receiving targeted sequencing data obtained by processing the plasma sample by a sequencing platform, the targeted sequencing data comprising a fragment length histogram feature of an off-target portion of the targeted sequencing data; and

- feeding the fragment length histogram feature to a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample;

wherein the ctDNA quantification model comprises a neural network integrative model trained using a training dataset comprising labelled sequencing data of off-target read associated fragments originating from healthy and cancer afflicted individuals.

Brief Description of the Drawings

[0013] Some embodiments of systems and methods for quantifying circulating tumour DNA (ctDNA) in a plasma sample, in accordance with present disclosure, will

now be described, by way of non-limiting example only, with reference to the accompanying drawings in which:

[0014] Fig. 1 illustrates a pipeline overview of systems and methods for quantifying circulating tumour DNA (ctDNA) in a plasma sample (Fragle). 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.

[0015] Fig. 2 illustrates an overview of systems and methods for quantifying circulating tumour DNA (ctDNA) in a plasma sample (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 is trained on a large collection of cancer/healthy mixture cfDNA samples.

[0016] Fig. 3 illustrates a series of mean absolute error (MAE) histograms showing the low and high ctDNA burden models help the final model achieve a low MAE for both low and high ctDNA burden samples as well as for healthy samples.

[0017] Fig. 4 illustrates correlation analysis between expected and Fragle predicted ctDNA levels for colorectal (CRC) and breast (BRCA) cancer samples in the 10 split validation set (4a – 4b); comparisons between ichorCNA (<https://www.nature.com/articles/s41467-017-00965-y>) and Fragle predicted ctDNA levels in unseen samples from colorectal (N = 190), breast (N = 23), liver (N = 34) and gastric cancer patients (N = 49) (4c – 4f); 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) (4g).

[0018] Fig. 5 shows correlation analysis between ichorCNA and Fragle predicted ctDNA burden for three cancer types – lung (pearson correlation R: 0.632), nasopharyngeal (R: 0.748) and head & neck cancer (R: 0.227). The correlation for head & neck cancer is low, because 90% of the samples of this cancer type have been predicted as 0 ctDNA burden sample by ichorCNA.

[0019] Fig. 6 shows predicted ctDNA fractions for healthy and low-ctDNA level samples in the validation set samples. Boxplots are represented by median and interquartile range (IQR), with ± 1.5 IQR as whiskers (6a). ROC analyses for

classification of healthy control and cancer ($\geq 1\%$ ctDNA) samples (validation samples) for Fragle, ichorCNA and 4-feature model (<https://www.science.org/doi/10.1126/scitranslmed.aat4921>) (6b), ROC analysis for classification of cancer (N=86) and healthy (N=57, 3 distinct cohorts) plasma samples in the unseen test cohort (6c), 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 (6d), expected vs. predicted ctDNA fractions using in vitro ctDNA dilution for 2 colorectal cancer samples (6e), expected vs. predicted ctDNA fractions using in vitro ctDNA dilution for 3 gastric cancer samples (6f).

[0020] Fig. 7 shows ROC analysis to examine the diagnostic performance on the validation set samples with different ctDNA burden thresholds to exclude low ctDNA samples.

[0021] Fig. 8 illustrates Fragle predicted ctDNA burden median at different unseen in silico dilution points and healthy mixture series - the median predictions are on the diagonal line up until expected ctDNA fraction drops to 0.625%.

[0022] Fig. 9 illustrates application of Fragle to targeted sequencing data. Application of Fragle to samples having both lpWGS and targeted gene panel sequencing data (9a). ctDNA levels predicted using lpWGS data and on-target reads from targeted sequencing samples (9b). ctDNA levels predicted using lpWGS data and off-target reads from targeted sequencing samples (9c). 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 (9d).

[0023] Fig. 10 shows correlation between on-target vs off-target prediction coverage (extracted from the same targeted bam file) for the different targeted samples of four different cohorts consisting of three different cancer types.

[0024] Fig. 11a to 11e illustrates monitoring of ctDNA levels and disease progression from targeted sequencing. 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 time points for a given patient were

included. Mutation VAFs were estimated using an automated pipeline, with manual pileup performed at highlighted time points where mutation detection failed.

[0025] Figure 12 illustrates 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.

[0026] Fig. 13 illustrates Fragle bin based modelling design incorporating local and global fragment length histogram feature.

[0027] Fig. 14 shows feature significance matrix of the bin based model input feature for cancer samples ($\geq 1\%$ TF) vs healthy reference: rows represent the genome bins, while columns represent the fragment lengths. Each element of the matrix represents normalized fragment length density in a specific genome bin.

[0028] Fig. 15 provides comparison between normalized cancer (left) and blood (right) TSS region fragment length density feature curves among four samples – two cancer samples (red one is high tumour fraction (TF) sample, while pink one is low TF sample) and two healthy samples (dark and light green).

[0029] Fig. 16 illustrates tumour burden modelling using two fragment length histograms: one for cancer specific gene region and the other for blood specific gene region.

[0030] Fig. 17 evaluates the sequencing coverage requirements for Fragle. 455 samples from the unseen test cohort were downsampled to render samples with lower fragment counts, ranging from 5 million (0.5x) to as low as 10 thousand fragments (0.001x). At 0.05x coverage (500K fragments), Fragle demonstrated near-perfect concordance ($r = 0.971$) with the predictions from the original 1x WGS samples. Even at 0.025X coverage (250K fragments), the concordance was ~ 0.95 , although some discrepancies were there in low ctDNA fraction samples.

[0031] Fig. 18 demonstrates an extremely fast runtime of ~50 seconds on a standard computer for a 1x coverage WGS sample for Fragle. The memory usage is also very low (20 MiB) and independent of the sample sequencing depth. In these benchmark experiments, both compute and memory requirements are significantly lower (3-4x) than another widely used software tool for ctDNA quantification, ichorCNA.

Detailed Description

[0032] Embodiments relate to systems and methods for quantifying circulating tumour DNA (ctDNA) in a plasma sample. While the description will generally be provided with reference to methods for ctDNA quantification, the described embodiments will be implemented by a system. That system requires at least one processor, and is configured to perform the computational steps of the methods of quantifying ctDNA. The processor(s) may be configured to receive low-pass cell free DNA (cfDNA) sequencing data obtained by processing the plasma sample using a sequencing platform. The system then performs fragment length analysis on the sequencing data using a cfDNA quantification model to estimate a cfDNA metric of the blood plasma sample.

[0033] The methods and systems of some embodiments may also be referred to as Fragle. Fragle may comprise a ctDNA quantification model for both whole genome sequencing and targeted sequencing data of cell free DNA (cfDNA).

[0034] Figure 1 illustrates a pipeline overview of systems and methods for quantifying circulating tumour DNA (ctDNA) in a plasma sample, also known as Fragle. A discovery cohort was assembled comprising low-pass whole genome sequence (lpWGS) data from 202 cancer plasma samples (2 cancer types) and 38 plasma samples from healthy individuals. This cohort only comprised cancer samples with detectable ctDNA ($\geq 3\%$) and ground-truth ctDNA levels were estimated using a consensus across distinct methods. Using a large-scale data augmentation approach, around 2500 mixture samples were generated from cancer and healthy samples with variable ctDNA fractions. The samples were split between train, test and validation datasets. Raw fragment length density distributions were derived using paired-end reads in each sample, to provide information on how cfDNA fragment length distributions may be used to predict ctDNA levels in a sample. The raw density distributions may be further normalized and transformed (e.g., log transformed), to

reveal local differences in the fragment length distributions associated with ctDNA levels in the samples.

[0035] The fragment length distributions are parsed to a ctDNA quantification model, discussed with reference to Figure 2, that outputs a Fragle predicted ctDNA fraction from a cfDNA fragment length distribution. This model can then be applied to the test dataset to assess accuracy of the prediction. In addition, the model can then be applied on other cancer types of any external cohort. Ideally, for use on new cancer types, BAM files will be generated for the cancer type.

[0036] In some embodiments, the ctDNA quantification model comprises a machine learning integrative model trained using a training dataset generated using artificial cfDNA data augmented with cfDNA data obtained from known cancer and healthy samples to generate artificial training dataset records with a known ctDNA metric. Figure 2 provides an illustration on the global fragment length histogram feature generation and model details, for embodiments incorporating machine learning based approaches to estimate the ctDNA level in a blood sample directly from the cfDNA fragment length distribution.

[0037] The transformed fragment length distributions, in combination with their labels in the form of ground-truth ctDNA levels, may serve as input to a multi-stage supervised machine learning approach. In some embodiments, the ctDNA quantification model comprises two parallel deep learning-based sub-models, optimized for quantitative prediction in high and low ctDNA burden samples. A high ctDNA burden model (HT) may be trained to estimate a ctDNA metric for high tumour burden samples. A low ctDNA burden model (LT) may be trained to estimate a ctDNA metric for low tumour burden samples. There may be a greater number of models, and thus of divisions or ranges of ctDNA burden, with each model being trained to predict ctDNA fraction for a specific one of the ranges. The HT and LT models may be trained to adopt the same model architecture and input fragment length features, but using different loss functions and model parameters.

[0038] In some embodiments, a high ctDNA burden sample may correspond to a ctDNA level of more than or equal to 3%. A low ctDNA burden sample may correspond to a ctDNA level of less than 3%.

[0039] In some embodiments, the LT and HT models individually predict ctDNA fractions of each training sample. The two predictions may become the input features

for the selection model. The ground truth may be labelled as 0 or 1, when the expected ctDNA fraction is $<3\%$ or $\geq 3\%$, respectively. The selection model of some embodiments is a binary support vector machine (SVM) classifier with radial basis kernel function. In other embodiments, a regression model or other type of model may be used. The hyperparameters of the selection model were tuned and determined using a withheld validation set to maximize the selection accuracy. The LT and HT models may be found to provide more accurate predictions for low ctDNA burden samples and high ctDNA burden samples, respectively. The final combined model may perform the best across the full range of samples.

[0040] As shown in Figure 2, the processed fragment length histogram is inputted to both models and two ctDNA fraction estimates are obtained in the first stage. In the second stage, the selection model (e.g., SVM or regression model) uses these two ctDNA fraction predictions as input features. This model determines whether the sample is a high or low ctDNA burden sample and selects the corresponding ctDNA fraction estimate as the final output. To train a final Fragle model based on the discovery cohort, a Fragle pre-model is trained on the samples to obtain their intermediate ctDNA burden estimates. The samples with a large deviation of ctDNA fractions estimated by ichorCNA and Fragle pre-model are excluded (e.g. 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.

[0041] In some embodiments, there are many samples in the discovery cohort for which only an ichorCNA prediction is available as ground truth. This may be because there is only access to low pass WGS data. The Fragle pre-model is trained on these samples + deep WGS and/or targeted samples. These samples may be accompanied by predictions made using one or more known methods.

[0042] For training the final Fragle model, both the Fragle pre-model and ichorCNA predictions are considered for those samples with only low pass WGS data. Data that don't have good a correlation between Fragle pre-model and ichorCNA are not used since their ground truth ctDNA burden data is low confidence.

[0043] Steps of Fragle feature extraction from fragment length have been graphically shown in Figure 2 according to a preferred embodiment. In experiments, the fragment length profile for all fragments sized 51-400 bp was computed using

the paired-end reads in the sample with Pysam. The length profile of each sample was normalized, presently using the highest observed fragment length count, followed by transformation of the normalized length profile. Transformation was performed by $\log_{10}$ scaling of the sample-wise normalized counts. Next, the transformed length profiles were smoothed – this can employ any mechanism and, presently, involved moving average normalization (z-score of 32-nt window) for sample-wise 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, the frequency of the fragments with this length is compared between cancer and healthy samples to examine whether cancer and healthy samples show significant differences based on the two-sided Wilcoxon rank sum test. Figure 19 shows significance plot for the difference of fragment length histogram count between cancer and healthy samples. It illustrates 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. Fragment size counts that have significance score above the red line have been deemed as significant. All samples were downsampled such that they all have similar number of total fragments in their global histogram before the significance analysis.

[0044] Figure 3 provides a series of mean absolute error (MAE) histograms showing the low and high ctDNA burden models contributing to the final model achieving a low MAE for both low and high ctDNA burden samples as well as for healthy samples. Supporting these model choices, it can be 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 as shown in Figure 3.

[0045] In some embodiments, the processor(s) is or are configured to allocate the cfDNA sequencing data into a plurality of whole genome data bins (per Figure 13), generate a histogram for each of the plurality of data bins, and process the generated histograms by a local branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample. The processor(s) may process the cfDNA sequencing data to generate a global fragment length histogram and process the global fragment length histogram using a second branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.

[0046] In some embodiments, the output of the first branch and the second branch of the ctDNA quantification model can be concatenated to obtain a concatenated intermediate output, and process the concatenated intermediate output by a block of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.

[0047] In some embodiments, the processor(s) is or are further configured to process a subset of the cfDNA sequencing data relating to cancer specific transcription start sites (TSS) and surrounding regions of the genome to generate a cancer specific fragment length histogram. The processor then processes a subset of the cfDNA sequencing data relating to blood specific TSS and surrounding regions of the genome to generate a blood specific fragment length histogram. Finally, it processes the cancer fragment length histogram and the blood fragment length histogram by the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.

[0048] In some embodiments, the received cfDNA sequencing data is obtained by sequencing the plasma sample to a depth of coverage of at least 0.025X. Preferably, the received cfDNA sequencing data is obtained by sequencing the plasma sample to a depth of coverage of 0.05X or above. The ctDNA quantification model was trained and evaluated using cross-validation. Figure 4 provides a demonstration of the high predictive accuracy on validation samples across two cancer types: Colorectal (mean absolute error (MAE) = 3.27%; Pearson $r = 0.92$) and breast (MAE = 3.56%; $r = 0.94$) cancer.

[0049] The final Fragle model was trained on the full discovery cohort and its performance tested on additional cohorts of unseen plasma lpWGS samples. A strong correlation between Fragle and ichorCNA-based ctDNA fraction estimates across unseen cohorts of colorectal cancer ($r = 0.81$; $P = 6.8e-41$; $N = 190$; Figure 4c), breast cancer ($r = 0.80$; $P = 3.7e-06$; $N = 23$; Figure 4d), liver cancer ($r = 0.87$; $P = 5.1e-10$; $N = 34$; Figure 4e), gastric cancer ($r = 0.73$; $P = 1.9e-09$; $N = 34$; Figure 4f), and a mixed cohort of cancer types not included in the discovery set (Combined $P = 1.3e-4$; $N = 30$; Figure 5). In the unseen colorectal cancer cohort, targeted gene sequencing was also performed and high-confidence somatic mutations identified in 86 samples. These data demonstrated high concordance between mutation VAFs and Fragle-predicted ctDNA levels ($r = 0.88$; $P = 3.8e-28$; Figure 4g).

[0050] Some embodiments of the present invention relate to systems for quantifying circulating tumour DNA (ctDNA) in a plasma sample. These systems comprise at least one processor which is configured to receive targeted sequencing data obtained by processing the plasma sample by a sequencing platform. The system then performs fragment length analysis on the off-target fragment or portion derived histogram from targeted sequencing data using a ctDNA quantification model, such as that described above for predicting ctDNA fraction, to estimate a ctDNA metric of the blood plasma sample. This is done irrespective of the on-target regions. The ctDNA quantification model comprises a neural network integrative model trained using a training dataset, comprising labelled sequencing data off-target data originating from both healthy and cancer afflicted individual targeted sequencing data. The targeted ctDNA quantification model may be the same.

[0051] Previous studies showed plasma cfDNA has a length distribution with a dominant mode at ~166 bp, approximately the size of DNA wrapped around a nucleosome plus its linker, suggesting nucleosome-protected DNA molecules shed in blood during DNA degradation. cfDNA fragments of cancer patients are generally shorter and more variably sized than those of healthy individuals. Differences in cfDNA fragment lengths between healthy individuals and cancer patients are exploited to quantify the ctDNA burden. Embodiments provide methods for robust quantification of ctDNA by fragment length analysis with deep learning, which thereby assists in the systemic monitoring of the tumour dynamics in cancer patients. Advantages of some embodiments include:

- a. **Lower Cost:** The method is applicable to low-pass WGS-based methods (~0.05x). Such low pass WGS incurs lower sequencing cost compared to a targeted panel used for sequencing at 10,000x coverage (usual target for coding mutation panels, typically upwards of 100kb-1000kb genomic regions).
- b. **Applicability:** The method is applicable to both WGS and targeted sequencing data.
- c. **Flexibility:** The approach utilizes lp-WGS data allowing for the simultaneous discovery of copy number aberrations. The approach can be directly applied to existing lp-WGS data for ctDNA quantification. Besides, some embodiments provide robust application to targeted sequencing data, which

simultaneously enables the discovery of actionable cancer mutations and the estimation of ctDNA fraction in one assay.

- d. **Accuracy:** Some embodiments can very accurately quantify ctDNA burden in all patients with colorectal, breast, liver, gastric, head & neck, nasopharyngeal and lung cancer.
- e. **Low Computation Cost:** Some embodiments require little to no hard disk space (less than 100 mega bytes) for internal computation, and also require no graphical processing unit (GPU) and low RAM (less than 20 mega bytes). The approach has extremely fast computation speed (around 50 seconds per 1X WGS sample). As shown by Fig. 18, the computation time increases linearly with the increase of sequencing depth.
- f. **Usability:** No parameter of the proposed approach needs to be tuned by the user irrespective of the cancer types and tumour fraction of the samples. The software can be installed and used via simple single line commands.
- g. **Feasibility in Cancer Screening:** Some embodiments do not require any tumour biopsy information and can predict tumour fraction based on data derived from blood plasma only.

Determination of the lower limit of detection

[0052] To explore the lower limit of detection (LoD) for the model, it is 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 as shown in Figure 6a, the model could differentiate between healthy and low-ctDNA level samples at the 1% ctDNA level (Wilcoxon rank sum test $P = 2.5e-24$), indicating a ~1% LoD in these samples. Similarly, the performance of Fragle for classification of healthy and cancer samples is examined in the validation sets. Using cancer samples with a ctDNA level $\geq 1\%$ in the validation sets as shown in Figure 6b, Fragle demonstrated an area under the curve (AUC) of 0.93, higher than ichorCNA (AUC = 0.88) applied to the same samples. Expectedly, the AUC increased further when filtering out low-ctDNA burden samples as shown in Figure 7. As an additional comparison, a model is implemented based on fragment length features previously used for the classification of cancer and healthy samples. A random forest model is trained on the discovery cohort using 4

features derived from the fragment length distribution (<https://www.science.org/doi/full/10.1126/scitranslmed.aat4921>). The 4-feature model demonstrated substantially lower classification accuracy than Fragle in the validation samples, with an AUC of 0.79.

[0053] The disclosure provides that the LoD is further evaluated using unseen test samples. 86 cfDNA samples from CRC patients are used with detectable mutations as positive cancer samples and all healthy plasma samples from the three different unseen cohorts as negatives ($N = 57$). As shown in Figure 6c, 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). 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. 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) as shown in Figure 6d and Figure 8.

[0054] To further examine these results using physical samples, similar dilution experiments were performed 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. Across 3 technical replicates, Fragle accurately predicted ctDNA fractions for both patients down to $\sim 1\%$ ctDNA level, with healthy samples consistently predicted $<1\%$ ctDNA as shown in Figure 6e. 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%. The second experiment comprised in vitro serial dilutions of 3 high-ctDNA plasma samples from gastric cancer patients (with 3 technical replicates). The results from this experiment mirrored previous observations, with the method accurately quantifying ctDNA down to the 0.5-1% level and predicting $<1\%$ ctDNA for healthy samples as shown in Figure 6f. Overall, these results collectively suggest that Fragle can accurately quantify and detect ctDNA with an LoD of $\sim 1\%$.

Application of Fragle to targeted sequencing data

[0055] 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. To explore whether Fragle could quantify ctDNA levels using targeted sequencing data, four cfDNA cohorts having both lpWGS were analyzed and sequencing data targeted as shown in Figure 9a. Using standard on-target reads obtained from the targeted sequencing data, Fragle tended to overestimate the ctDNA burden as compared to the lpWGS data as shown in Figure 9b. In an embodiment, the method on off-target reads is evaluated, which are often filtered and ignored in a targeted sequencing experiment. Strong concordance of predictions was observed 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 ($r = 0.96$, $P = 1.64\text{e-}30$, $N = 56$), colon cancer samples from unseen cohort ($r = 0.97$, $P = 3.69\text{e-}58$, $N = 96$), and gastric cancer samples from unseen cohort ($r = 0.96$, $P = 2.9\text{e-}27$, $N = 49$) as shown in Figure 9c.

[0056] The targeted sequencing samples contained between 100K to 10M off-target fragments (equivalent to ~ 0.01 - 1.0X) across the different cohorts, with $>95\%$ of samples having $>250\text{K}$ off-target fragments ($\sim 0.025\text{X}$) as shown in Figure 10. To further explore if these results generalize to targeted sequencing panels employed by commercial vendors, a cohort of 44 plasma samples were subjected to a liquid biopsy gene panel from Foundation Medicine. Since these samples did not have matched lpWGS data, ctDNA levels were approximated using the maximum VAFs reported by the company after filtering out germline variants. This cohort comprising samples from a variety of cancer types shows strong concordance of the reported VAFs and ctDNA levels estimated from off-target reads ($r = 0.70$, $P = 1.01\text{e-}07$, $N = 44$; Figure 9d). Overall, these results support that Fragle can estimate ctDNA levels using both lpWGS and targeted sequencing data.

Tracking ctDNA dynamics and disease progression from targeted sequencing

[0057] Having demonstrated that Fragle can accurately quantify ctDNA levels with targeted gene panel sequencing, the method was applied to longitudinal targeted sequencing samples from four late-stage colorectal cancer patients. In these samples, the temporal relationship between Fragle-estimated ctDNA dynamics and disease progression was explored by radiographic imaging (RI). Firstly, this showed

strong temporal correlations between mutation VAFs and Fragile ctDNA levels across the longitudinal samples from the four patients (Fig 11a-d). The first patient displayed concordant and increasing VAFs and Fragile ctDNA levels, consistent with the emergence of progressive disease (PD) via RI at late time points (Fig 9a). The second patient developed a partial response to FOLFOXIRI treatment, consistent with both reductions in VAFs and Fragile ctDNA levels (Fig 9b). 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 Fragile showed a consistent pattern of disease progression, with ctDNA levels remaining high at all time points ($>10\%$; Fig. 9c-d). While the automated variant calling pipeline failed to detect mutations at late time points 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). A metastatic colorectal cancer patient was considered, for whom 21 serial blood plasma samples had been collected over a cetuximab/chemotherapy treatment course of 3 years (Fig. 9e). In this patient, results showed an overall temporal correlation of Fragile-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, Fragile provided an orthogonal and independent measure of ctDNA levels. Overall, these data demonstrate high concordance to Fragile estimated ctDNA levels and disease progression estimated from radiographic imaging. Secondly, they outline how Fragile could be used to interpret and resolve heterogeneous and variable mutation VAFs profiled with targeted sequencing assays.

Risk stratification for early-stage lung cancer patients

[0058] Blood-based detection of minimal residual disease (MRD) following treatment has the potential to improve risk stratification and management strategies for cancer patients. Given the $\sim 1\%$ LoD for Fragile, the method was investigated to assess if it could be used for tumor-naïve MRD screening, with no requirements for a matching tissue sample. The targeted sequencing data was obtained 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). 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. 12a). In the ctDNA-negative samples, Fragle was used 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. 12b). 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. 12c). 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.

Plasma sample collection and processing

[0059] Besides the newly generated CRC, BRCA and healthy samples, WGS data of plasma samples in the discovery cohort were obtained from other internal cohorts.

[0060] Similarly, the test cohort was composed of internal cohorts as well as samples from previous studies. 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). The written informed consent was obtained from the 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 paired-end sequenced (2 × 151 bp) on Illumina NovaSeq6000 system.

Whole-genome sequencing of cfDNA

[0061] Low-pass WGS (~4x, 2 x 151bp) was performed on cfDNA samples from cancer patients and from healthy individuals. Deep WGS (~125X) was performed on the two BRCA samples used for unseen in-silico dilution series creation. Some embodiments used bwa-mem to align the WGS reads to the hg19/ GRCh37 human genome.

Estimation of ctDNA fractions in the discovery cohort

[0062] The ctDNA fractions in the plasma samples of 4 cancer types in the training cohort were estimated 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 from four tissue-based methods (<https://www.nature.com/articles/s41467-021-22463-y>). 53 plasma samples of CRC had lp-WGS and targeted NDR sequencing data, and their ctDNA fractions were inferred from averaging the reported ichorCNA and NDRquant estimates in these samples. The remaining CRC and BRCA samples only had lpWGS data and their ctDNA fractions were inferred using ichorCNA.

Discovery cohort data augmentation approach

[0063] After identifying the plasma samples with ctDNA level ≥ 0.03 and with at least 10 million fragments in the discovery cohort, the cancer samples ($n = 103$) were split into training and validation set, independently repeated for 10 times creating 10 training-validation pairs – this follows the methodology schematically reflected in Figure 1. One plasma sample of cancer was in silico diluted with reads from one control plasma sample to generate in silico spike-ins, followed by downsampling to 10 million cfDNA fragments per spike-in sample. In silico samples were generated with variable ctDNA fractions ranging from 10^{-6} up to the undiluted fractions. To minimize information leakage to the cross-validation set, the healthy control samples ($n = 38$) were evenly split into two sets. These two control sets were then used to dilute cancer samples in the training and validation sets, respectively.

Overview of Fragle

[0064] Fragle quantifies ctDNA levels directly from a cfDNA fragment length histogram. Using paired-end sequencing data, we compute the length of each

sequenced cfDNA fragment, excluding duplicates and supplementary alignments and only keeping paired reads mapping to the same chromosome with a minimal mapping quality of 30. The machine learning model consists of two stages (see Fig. 1 and Fig 2 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 ($\geq 3\%$) samples, respectively. Although these models differ in terms of loss function, the model architecture and input feature (extracted from the global fragment length histogram of the input cfDNA sequencing data) for both models are the same. 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.

Fragile model feature extraction

[0065] Each step of Fragile feature extraction process from the global histogram has been graphically shown in Fig. 2. 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) utilizing Pysam library of Python. The histogram of each sample was normalized using the highest observed fragment length count found in that sample, followed by log10 scaling of these sample-wise normalized counts. Next, a moving average normalization (z-score of 32-nt window) was performed sample-wise as an additional smoothing step. The transformed length histogram features for a given sample was further standardized relative to the training set (column-wise z-score normalization). Finally, the feature values related to the 281 significant fragment lengths (significant in differentiating cancer and healthy samples) have been used as input for both the high and the low ctDNA burden model. In order to find these significant fragment lengths, we first downsampled all our discovery cohort healthy and cancer samples such that they all have 10 million fragments in their histograms. Then for each fragment length, we counted the number of that particular fragment for the different cancer and healthy samples; and compared these two groups of count values using the two-sides Wilcoxon rank sum test to see if these counts are significantly different between cancer and healthy

group. Fig 19 shows the negative log P-value plot of the significance values for all 350 fragment lengths. The 281 fragment lengths that have negative log P-value above the red line in the figure have been selected.

High and low ctDNA burden model architecture

[0066] Detailed model architecture of some embodiments has been shown in Fig 2. Some embodiments used a neural network starting with a feature embedding layer. We have 16 fully connected layers with batch normalization and residual connections. Dropout regularization 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 of some embodiments 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:

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

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.

σ was fixed at 0.0005 after tuning.

Selection model

[0067] 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 $\geq 3\%$, respectively. The selection model is a binary support vector machine (SVM) classifier with radial basis kernel function (RBF). For each sample, the selection model uses only two features - low and high ctDNA model predictions for that sample. Based on these 2 features, the selection model predicts whether this sample is low ctDNA burden or high ctDNA burden. If the prediction is low, then we select the low ctDNA burden model prediction as the final prediction and vice versa. The RBF kernel theoretically takes the 2-dimensional input feature to infinite dimension, enabling linear separation of the two classes to be predicted.

Targeted sequencing experiment

[0068] In some experiments, plasma and patient-matched buffy coat samples were isolated from whole blood within two hours from collection and 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 CRC dataset and in the unseen metastatic gastric cancer ($N = 49$) dataset, using xGen custom panel (Integrated DNA Technologies) of 225 genes that are known cancer driver genes. Some experiments included the targeted assay on six plasma samples from healthy individuals to identify a blacklist of frequently occurring and unreliable variants that are likely attributed to sub-optimal probe design. Paired-end sequencing (2×151 bp) was done on an Illumina NovaSeq6000 system.

Variant calling

[0069] During pre-processing, 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. Experiments involved performing UMI-aware deduplication using the fgbio package (<https://github.com/fulcrumgenomics/fgbio>). Experiments 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.

[0070] To identify single-nucleotide variants and small insertions/deletions in the cfDNA samples, a bioinformatics pipeline was developed. Variant screening was performed using VarDict (<https://academic.oup.com/nar/article/44/11/e103/2468391>) using a minimal VAF threshold of 0.05% and annotated all variants using Variant Effect Predictor (<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4>). Low-impact variants such as synonymous or splice region variants and low-quality variants such as those that fail to fulfil the minimum requirements of variant coverage, signal-to-noise ratio, and the number of reads supporting alternative alleles were removed. Finally, we removed population SNPs found in Genome Aggregation Database (gnomAD) and 1000 Genomes.

[0071] To further minimize false positive variants, we used duplexCaller (<https://academic.oup.com/clinchem/article/64/11/1625/5608546>) to identify the variants with double-strand support (<https://www.nature.com/articles/s41598-020-72818-6>), and discarded blacklisted variants that were recurrently found in the plasma of two or more healthy individuals. Finally, 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.

Unseen in vitro dilution experiments

[0072] High ctDNA burden cfDNA samples from 2 individual CRC patients were selected as a benchmark to create a starting point for the dilution series. The ctDNA content for each sample was experimentally determined by 2 methods (ichorCNA and NDR estimated), with good concordance from both methods (sample 1: ctDNA content 38% by ichorCNA, 37% by NDR; sample 2: 39%, 38% respectively). Commercial pooled cfDNA from healthy volunteers (0% ctDNA) was purchased from PlasmaLab (lot numbers 2001011, 210302) and used to set up an 9 point serial dilution of the ctDNA content for each sample. 3 technical replicates were performed for each 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 Fragile, 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, with 3 technical replicates. We randomly selected 3 control samples and pooled them before diluting the plasma of cancer. Ip-WGS was performed with a target depth of ~4-5x.

Unseen in silico dilution experiments

[0073] 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 Fragile and ichorCNA (>10% ctDNA based on both methods with a relative difference < 5%). 20 healthy mixtures are prepared, 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, 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.

Application of Fragle to targeted data

[0074] 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 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, targeted sequencing data from 44 patients profiled with the Foundation Medicine Liquid CDx assay is used. Samples belonged to different cancer types such as lung, liver, colon, breast, pancreas, rectum, esophagus, ovarian, uterus, gallbladder, and stomach cancer. Known germline variants are filtered 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, a targeted sequencing bed file is constructed using the 311 genes reported to comprise the panel.

Lung cancer survival analysis

[0075] Plasma targeted sequencing data from the MEDAL cohort (Project ID: OEP004204) was retrieved from National Omics Data Encyclopedia (NODE). Alignment to the human genome (hg19) was conducted using bwa-mem. Duplicates in the aligned data were marked using Samblaster. 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.

Data availability

[0076] 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.

[0077] The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

[0078] Throughout this specification and the statements or claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

[0079] The scope of this disclosure encompasses all changes, substitutions, variations, alterations, and modifications to the example embodiments described or illustrated herein that a person having ordinary skill in the art would comprehend. The scope of this disclosure is not limited to the example embodiments described or illustrated herein. Moreover, although this disclosure describes and illustrates respective embodiments herein as including particular components, elements, feature, functions, operations, or steps, any of these embodiments may include any combination or permutation of any of the components, elements, features, functions, operations, or steps described or illustrated anywhere herein that a person having ordinary skill in the art would comprehend. Although this disclosure describes or illustrates particular embodiments as providing particular advantages, particular embodiments may provide none, some, or all of these advantages.

Claims

1. A system for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the system comprising at least one processor configured to:
 - receive low-pass cell free DNA (cfDNA) sequencing data obtained by processing the plasma sample using a sequencing platform; and
 - perform fragment length analysis on the sequencing data using a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample;wherein the ctDNA quantification model comprises a machine learning integrative model trained using a training dataset generated using cfDNA data augmented with cfDNA data obtained from known cancer and healthy samples to generate artificial training dataset records with a known ctDNA metric.
2. The system of claim 1, wherein the ctDNA quantification model comprises a low ctDNA burden (LT) model and a high ctDNA burden (HT) model;
 - the LT model being trained to estimate a ctDNA metric for low tumour burden samples; and
 - the HT model being trained to estimate a ctDNA metric for high tumour burden samples.
3. The system of claim 2, wherein the records with a tumour burden greater than or equal to 3% are considered as records with a high tumour burden; and wherein the records with a tumour burden less than 3% are considered as records with a low tumour burden.
4. The system of claim 1, wherein the one processor is further configured to:
 - allocate the cfDNA sequencing data into a plurality of whole genome data bins;
 - generate a histogram for each of the plurality of data bins; and
 - process the generated histograms by a local branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
5. The system of claim 4, wherein the at least one processor is further configured to:

process the cfDNA sequencing data to generate a global fragment length histogram; and
process the global fragment length histogram using a second branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.

6. The system of claim 5, wherein the at least one processor is further configured to:
 - concatenate output of the first branch and second branch of the ctDNA quantification model to obtain a concatenated intermediate output;
 - and
 - process the concatenated intermediate output by a block of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
7. The system of claim 1, wherein at least one processor is further configured to:
 - process a subset of the cfDNA sequencing data relating to cancer specific transcription start sites (TSS) and surrounding regions of the genome to generate a cancer fragment length histogram;
 - process a subset of the cfDNA sequencing data relating to blood specific transcription start sites (TSS) and surrounding regions of the genome to generate a blood fragment length histogram; and
 - process the cancer fragment length histogram and the blood fragment length histogram by the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
8. The system of claim 1, wherein the received cfDNA sequencing data is obtained by sequencing the plasma sample to a depth of coverage of 0.05x or greater.
9. A system for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the system comprising at least one processor configured to:
 - receive targeted sequencing data obtained by processing the plasma sample by a sequencing platform, the targeted sequencing data comprising a fragment length histogram feature of an off-target portion of the targeted sequencing data; and feeding the fragment

length histogram feature to a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample;
wherein the ctDNA quantification model comprises a neural network integrative model trained using a training dataset comprising labelled sequencing data of off-target read associated fragments originating from healthy and cancer afflicted individuals.

10. The system of claim 9, wherein the received targeted sequencing data is obtained by sequencing the plasma sample to a depth of coverage of 60x or greater.
11. The system of claim 9 or 10, wherein the targeted sequencing data comprises off-target reads.
12. A method for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the method comprising:
 - processing the plasma sample by a sequencing platform to obtain low-pass cell free DNA (cfDNA) sequencing data; and
 - performing fragment length analysis of the sequencing data using a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample;wherein the ctDNA quantification model comprises a machine learning integrative model trained using a training dataset generated using cfDNA data augmented with ctDNA data obtained from known cancer samples to generate artificial training dataset records with a known ctDNA metric.
13. The method of claim 12, wherein the ctDNA quantification model comprises a low ctDNA (LT) model and a high ctDNA (HT) model;
 - the LT model being trained to estimate a ctDNA metric using records corresponding to low tumour burden; and
 - the HT model being trained to estimate a ctDNA metric using records corresponding to high tumour burden.
14. The method of claim 13, wherein the records with a tumour burden greater than or equal to 3% are considered as records with a high tumour burden; and

wherein the records with a tumour burden less than 3% are considered as records with a low tumour burden.

15. The method of claim 12, wherein the method further comprises:
 - allocating the cfDNA sequencing data into a plurality of whole genome data bins;
 - generating a histogram of each of the plurality of data bins; and
 - processing the generated histograms by a first branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
16. The method of claim 15, wherein the method further comprises:
 - processing the cfDNA sequencing data to generate a global fragment length histogram; and
 - processing the global fragment length histogram using a second branch of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
17. The method of claim 15, wherein the method further comprises:
 - concatenating output of the first branch and second branch of the ctDNA quantification model to obtain a concatenated intermediate output; and
 - processing the concatenated intermediate output by a block of the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.
18. The method of claim 12, wherein the method further comprises:
 - processing a subset of the cfDNA sequencing data relating to cancer specific transcription start sites (TSS) and surrounding regions of the genome to generate a cancer fragment length histogram;
 - processing a subset of the cfDNA sequencing data relating to blood specific transcription start sites (TSS) and surrounding regions of the genome to generate a blood fragment length histogram; and
 - processing the cancer fragment length histogram and the blood fragment length histogram by the ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample.

19. The method of claim 12, wherein the plasma sample is sequenced to a depth of coverage of 0.05X or greater.
20. A method for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the method comprising:
 - receiving targeted sequencing data obtained by processing the plasma sample by a sequencing platform, the targeted sequencing data comprising a fragment length histogram feature of an off-target portion of the targeted sequencing data; and
 - feeding the fragment length histogram feature to a ctDNA quantification model to estimate a ctDNA metric of the blood plasma sample;wherein the ctDNA quantification model comprises a neural network integrative model trained using a training dataset comprising labelled sequencing data of off-target read associated fragments originating from healthy and cancer afflicted individuals.
21. The method of claim 20, wherein the plasma sample is sequenced to a depth of coverage of 60x or greater.
22. The method of claim 20 or 21, wherein the targeted sequencing data comprises off-target reads.