Claims

1. A system for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the system comprising:

    at least one processor configured to:

        receive cell free DNA (cfDNA) sequencing data obtained by processing the plasma sample using a sequencing platform; and

        a ctDNA quantification model configured to perform fragment length analysis on the sequencing data using a ctDNA quantification model to estimate a ctDNA fraction of the 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 fraction for low tumour burden samples;

    the HT model being trained to estimate a ctDNA fraction for high tumour burden samples; and

    a selection model for making a determination of whether the plasma sample is a high ctDNA burden sample or a low ctDNA burden sample and, based on the determination, selecting a corresponding one of the estimated ctDNA fraction for low tumour burden samples and the estimated ctDNA fraction 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 fraction 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 the ctDNA fraction 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 method for quantifying circulating tumour DNA (ctDNA) in a plasma sample, the method comprising:

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

performing fragment length analysis of the sequencing data using a ctDNA quantification model to estimate a ctDNA fraction 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.

10. The method of claim 9, 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 fraction using records corresponding to low tumour burden;

the HT model being trained to estimate a ctDNA fraction using records corresponding to high tumour burden; and

a selection model for making a determination of whether the plasma sample is a high ctDNA burden sample or a low ctDNA burden sample and, based on the determination, selecting a corresponding one of the estimated ctDNA fraction for low tumour burden samples and the estimated ctDNA fraction for high tumour burden samples.

11. The method of claim 10, 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.

12. The method of claim 9, 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.

13. The method of claim 12, wherein the method further comprises:

processing the cfDNA sequencing data to generate a global fragment length histogram;

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;

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 the ctDNA fraction of the blood plasma sample.

14. The method of claim 9, 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 the ctDNA fraction of the blood plasma sample.

15. The system of any one of claims 1 to 8 or the method of any one of claims 9 to 14, wherein the cfDNA sequencing data is low-pass cfDNA sequencing data, and optionally wherein the cfDNA sequencing data is whole-genome sequencing cfDNA sequencing data.