Genomics 116 (2024) 110876

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Early detection and stratification of colorectal cancer using plasma cell-free DNA fragmentomic profiling

Jiyuan Zhou [a,1], Yuanke Pan [b,1], Shubing Wang [c,1], Guoqiang Wang [d,1], Chengxin Gu [a], Jinxin Zhu [b], Zhenlin Tan [a], Qixian Wu [a], Weihuang He [e], Xiaohui Lin [f], Shu Xu [g], Kehua Yuan [h], Ziwen Zheng [a], Xiaoqing Gong [a], Chenhao JiangHe [a], Zhoujian Han [a], Bingding Huang [b], Ruyun Ruan [b], Mingji Feng [e], Pin Cui [e,**], Hui Yang [a,*]

[a] Department of Gastroenterology, the Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, China
[b] College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[c] Department of Oncology, Shenzhen Key Laboratory of Gastrointestinal Cancer Translational Research, Cancer Institute, Peking University Shenzhen Hospital, Shenzhen-Peking University-Hong Kong University of Science and Technology Medical Center, Shenzhen, China
[d] Department of Gastrointestinal Surgery, the Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, China
[e] Shenzhen Rapha Biotechnology Incorporate, Shenzhen, China
[f] Department of Oncology, People's Hospital of Shenzhen Baoan District, The Second Affiliated Hospital of Shenzhen University, Shenzhen, China
[g] Department of Oncology, Shenzhen Hospital, University of Chinese Academy of Sciences, Shenzhen, Guangdong, China
[h] Department of Oncology, Yantian Hospital, South University of Science and Technology, Shenzhen, Guangdong, China

ARTICLE INFO

Keywords:
Colorectal cancer
Cell-free DNA
Machine learning
Fragmentation profile

ABSTRACT

Timely accurate and cost-efficient detection of colorectal cancer (CRC) is of great clinical importance. This study aims to establish prediction models for detecting CRC using plasma cell-free DNA (cfDNA) fragmentomic features. Whole-genome sequencing (WGS) was performed on cfDNA from 620 participants, including healthy individuals, patients with benign colorectal diseases and CRC patients. Using WGS data, three machine learning methods were compared to build prediction models for the stratification of CRC patients. The optimal model to discriminate CRC patients of all stages from healthy individuals achieved a sensitivity of 92.31% and a specificity of 91.14%, while the model to separate early-stage CRC patients (stage 0-II) from healthy individuals achieved a sensitivity of 88.8% and a specificity of 96.2%. Additionally, the cfDNA fragmentation profiles reflected disease-specific genomic alterations in CRC. Overall, this study suggests that cfDNA fragmentation profiles may potentially become a noninvasive approach for the detection and stratification of CRC.

1. Introduction

Colorectal cancer (CRC) is the third most common malignant neoplasm and the second most deadly cancer. Undoubtedly, a large number of CRC cases substantially contribute to the global burden of public health [1]. Despite advances in therapies for the treatment of colorectal cancer, including local ablative therapies for metastases, radiotherapy, immunotherapy, palliative chemotherapy, targeted therapy, and endoscopic and surgical excision, the morbidity and mortality of CRC are still increasing [2,3]. Reportedly, CRC accounts for 10% of global cancer incidence and 9.4% of cancer deaths in 2020, and the global number of new CRC cases is expected to reach 3.2 million in 2040 [4]. However, the symptoms of CRC appear only in the advanced stages of the disease, which are invasive, malignant, and metastatic. These constraints make the problem difficult to solve. Therefore, early detection of CRC is particularly important.

Recently, liquid biopsy, particularly plasma cell-free DNA (cfDNA), has been developed as a potential tool for diagnosis and monitoring in clinical oncology [5]. Compared to tumor tissues, plasma cfDNA is noninvasive and can provide timely insight into tumor progression,

* Corresponding author at: Department of Gastroenterology, the Second Affiliated Hospital of Guangzhou Medical University, No.250, Changgang East Street, Haizhu District, Guangzhou 440105, China
** Corresponding author at: Shenzhen Rapha Biotechnology Incorporate, Shenzhen 518118, China.
E-mail addresses: cuipin@rafabio.com (P. Cui), yanghui@gzhmu.edu.cn (H. Yang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2024.110876
Received 28 October 2023; Received in revised form 17 April 2024; Accepted 27 April 2024
Available online 5 June 2024

which can be sampled repeatedly with less harm to patients [6]. cfDNA can be detected in the plasma of blood from either normal or tumor tissue, normally at a low content [7,8]. Several studies have shown that the non-random fragmentation patterns of cfDNA can reflect epigenetic regulation [9,10] and can be used as a marker for cancer screening [11,12], so the distribution of cfDNA fragments is related to histopathological status. Notably, the fragment length of cfDNA derived from malignant sources exhibits greater variability compared to its non-neoplastic counterparts [13]. Furthermore, the cfDNA end motif (the first 4 bp nucleotides of molecular ends) of healthy individuals or nonmalignant cancer shows different nucleotide preferences compared to that of cancer patients [14]. Both of these molecular properties have been categorized as cfDNA fragmentomics, which has shown potential for cancer detection in several studies [15–17]. Despite these preliminary studies for the detection of multiple cancer types, CRC has not been studied in detail.

To investigate the potential of cfDNA fragmentomics for CRC detection, we performed whole genome sequencing (WGS) on three cohorts to distinguish healthy individuals, patients with benign colorectal diseases and CRC patients from each other. Hence, this study proved the potential of plasma cfDNA fragmentomic features for sensitive detection and stratification of CRC.

2. Materials and methods

2.1. Study design and participant enrolment

In total, 620 participants were recruited for this study, including a healthy cohort (395 individuals), a benign patient cohort (97 patients with benign colorectal diseases) and a cancer cohort (128 patients with CRC). All benign colorectal and CRC patients were enrolled at the time of diagnosis in the Second Affiliated Hospital of Guangzhou Medical University, the Second Affiliated Hospital of Shenzhen University, and Shenzhen Hospital, University of Chinese Academy of Sciences, while healthy individuals were enrolled in the Peking University Shenzhen Hospital. Detailed information such as age, gender, and related clinical records are listed in Supplementary Table S1. All procedures involving human participants were performed following the Declaration of Helsinki. The protocol obtained review and approval from the Ethics and Scientific Committee of the Second Affiliated Hospital of Guangzhou Medical University (NO. 2022-YJS-ks-22), Peking University Shenzhen Hospital (NO. 2021–075), the Second Affiliated Hospital of Shenzhen University (BYL20230310), and Shenzhen Hospital, University of Chinese Academy of Sciences (LL-KT-2022053).

Fragment size and motif features are extracted from the sequencing data, and all features are used to train a model through a back-propagation algorithm based on the Adam (Adaptive Moment Estimation) optimizer. The model learns and fits the data features of the training set and can make decisions based on the data features of new samples. The resulting score was used to predict the clinical diagnosis of the subject corresponding to the sample. All participants underwent experiments and data analysis workflows, as illustrated in Fig. 1.

2.2. Cell-free DNA extraction from plasma samples

Peripheral blood was collected from all 620 participants, each 6–10 mL into a cfDNA preservation tube (cat. 20,092,421, Hebei Xinle Medical Instrument Technology Inc., Xinle, China) and shipped at room temperature to the Molecular Genetics Laboratory (Shenzhen RAFA Biotechnology Inc., China) for cfDNA extraction within 72 h after blood draw. Plasma was obtained by centrifugation of whole blood at 1600*g* for 10 min. The supernatant was transferred to a new tube and further centrifuged at 10,000*g* for 15 min to remove cell debris from the plasma. For each participant, cfDNA was isolated and purified from 3 mL plasma using the HiPure Circulating DNA Midi Spin Kit S (Magen Biotech Inc., Guangzhou, China) into a final elution volume of 50 μL. Quality control was performed on these libraries using Qsep100 (Bio-optic. Inc., Taiwan, China) for fragment size distribution and Qubit 4.0 (Thermo Fisher Inc., MA, USA) for concentration, and cfDNA samples with abnormal fragment size distribution (showing distribution outside the normal cfDNA peak) and ultrahigh concentration were identified as contaminated with genomic DNA (mainly from dead white blood cells during logistics). None of the 620 participants had genomic DNA contamination, and the data was used for downstream analysis.

2.3. Whole-genome sequencing library construction and sequencing

For all 620 participants in this study, WGS was performed using 10 ng cfDNA input for each participant. WGS libraries were constructed using the RainbowOne Universal DNA Library Prep Kit for MGI (Rapha Biotechnology Inc., China) following the fundamental principles for WGS library preparation, including molecular end repair, sequencing adaptor ligation, and library clean up. The libraries were then amplified using VAHTS HiFi Amplification Mix (cat. N616–01) and purified using VAHTS DNA Clean Beads (cat. N411–02), both purchased from Vazyme Biotech Co., Ltd., Nanjing, China. Quality control was performed on these libraries using Qsep100 (Bio-optic. Inc., Taiwan, China) and Qubit 4.0 (ThermoFisher. Inc., MA, USA), and libraries were sent for

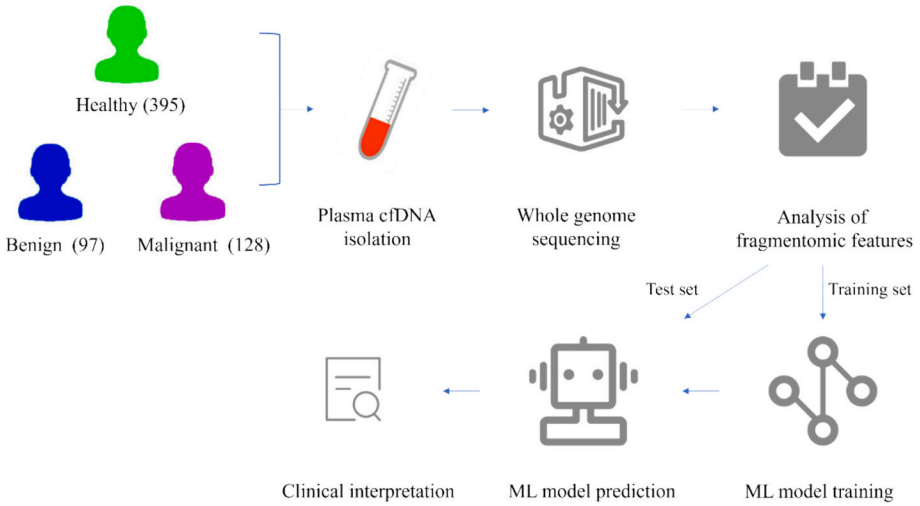

Fig. 1. General workflow of this study.

sequencing in batches of 24, each on one lane of an MGI-2000 sequencer (BGI Genomics Inc., Wuhan, China) using DNBSEQ™ technology and sequencing mode. Following such design of pooling for sequencing, the average sequencing depth for each library is around 1.2×, which is within the recommended range (0.5×-5×) for low depth WGS based cancer detection using cfDNA fragmentation [18].

2.4. Fragmentation profile and end motif feature

A total of 620 WGS data samples were included in the data analysis. First, raw sequencing data were filtered by fastp [19] as part of the quality control protocol. The qualified reads were then mapped onto the human reference genome (GRCh37/UCSC hg19) using the sequence aligner BWA [20]. PCR duplicates were then marked by SAMtools [21]. The fragment size for every read pair with a mapping quality score above 30 for either read was extracted from every sample by in-house scripts. According to the relevant studies of cfDNA fragmentation patterns, short fragments were defined as having lengths between 100 and 150 bp, and long fragments were defined as having lengths between 151 and 220 bp [22,23]. The fragment profile was generated using the short/long fragment ratio, as previously described in the DELFI approach [13]. The ratios of the short/long fragments for each sample were examined in 5 Mb bins, resulting in a total of 472 features from the 472 bins genome-wide after excluding the Duke blacklisted regions and the low mappability regions. All 472 fragment features were then used as the input to build the prediction model for colorectal cancer detection.

Furthermore, the first four bases of the R1 end of each cfDNA fragment were used as motif codes, and the proportion of 4^4 motif codes in each sample was counted as motif features. A total of 728 dimensions of features were used to train our model.

2.5. Prediction models

To build an automatic prediction model using WGS data, three machine learning methods were implemented (including random forest [24], LightGBM [25], XGBoost [26] and 1- to 15-layer neuron networks were tested. GridSearchCV [27] was used to search hyperparameters such as the number of estimators, max depth, and learning rate. Eventually, the parameters were set as 150 estimators, max depth 2, and learning rate 0.1 for training the model. After comparing the results from three machine learning methods, LightGBM was chosen as the optimal approach to build prediction models for multiple clinical stratifications. The participants in each cohort were randomly assigned into a training set and validation set at an 8:2 ratio to build these prediction models according to the designed pipeline (Fig. 1). This ratio of 8:2 was determined by convention.

2.6. Criteria

Typically, the performance of a classifier is evaluated according to four basic statistics: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP is an example of a correctly labelled positive; TN is an example correctly marked as negative; FP is a negative example incorrectly labelled as positive; FN is a positive example that has been incorrectly marked as negative.

Based on the four basic statistics of cancer prediction, the sensitivity, specificity, accuracy, true positive rate (TPR) and false positive rate (FPR) were further calculated.

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

According to the TPR and FPR, the receiver operating characteristic (ROC) curve can be drawn to evaluate the performance of the model. ROC curves were created by plotting TPR against FPR at different threshold settings. The area under the ROC curve (AUC) is used to evaluate the prediction performance, and the higher the AUC value is, the better the prediction performance.

2.7. Transcription factor analyses from CRC

To study transcription factors (TFs) related to CRC, RNA gene expression data for 740 healthy individuals, 524 TCGA colon cancer (COAD) patients, and 177 TCGA rectal cancer (READ) patients were downloaded from the TCGA database. The count values of all samples were concatenated into one file for differential gene expression analysis using edgeR (R package). According to the results of this analysis, only the genes with a fold change >1.5 and a *P*-value <0.05 were identified as significantly differentially expressed genes (SDEGs). Furthermore, these SDEGs were annotated using Animal TFDB (a transcription factor database) and then subjected to Gene Ontology (GO) enrichment analysis for biological interpretation.

We utilized the Gene Transcription Regulation Database (GTRD) to obtain detailed transcription factor binding site (TFBS) information. Due to the potentially high number of TFBSs for each TF, TFs with fewer than 1000 defined sites in the GTRD were omitted. We recalculated the position of each TF by focusing on the peak ChIP-seq signals in the GTRD database, and extracting the top 1000 sites supported by the majority of analysed samples (1000-msTFBS). Coverage data for TFBS within ±1000 bp regions were computed, and the coverage data for each site were standardized using copy number variation and average coverage. For each position surrounding the TFBS, the mean coverage is shown. By comparing the TF coverage across cfDNA data from healthy individuals, patients with colorectal benign diseases, and CRC patients in this study, we identified the final CRC-specific TFs.

3. Results

3.1. cfDNA fragmentation profiles of three cohorts

All 620 participants recruited in this study underwent WGS and data analysis, including QC and mapping, yielding an average sequencing depth of 1.2× and an average map-rate of 91.99% (Supplementary Table S2). To visualize the cfDNA fragmentation profiles obtained from WGS data for each sample from the three cohorts, we plotted the fragment feature of the bin minus the average fragment feature of all 472 bins on the Y-axis, while the X-axis marks the order and location of all 472 bins by chromosome (Fig. 2A). In general, the profiles of the CRC cohort showed much stronger fluctuations than those of the healthy cohort, while the fluctuation of the curves of both the healthy cohort and the colorectal benign disease cohort was limited to a narrow range.

Next, we plotted the cfDNA fragmentation profile of CRC patients by different stages according to clinical records (Fig. 2B). As the stage increased, the fragmentation profile of CRC patients deviated further from the profile range of the healthy cohort, especially in some chromosome arms, such as 4q, 5q, 7p and 18q. To simplify this scenario, we integrated all sample profiles of the same stage into one curve, as shown in Fig. 2C.

3.2. Plasma cfDNA end motifs of three cohorts

From the same WGS data, molecular end motif features were

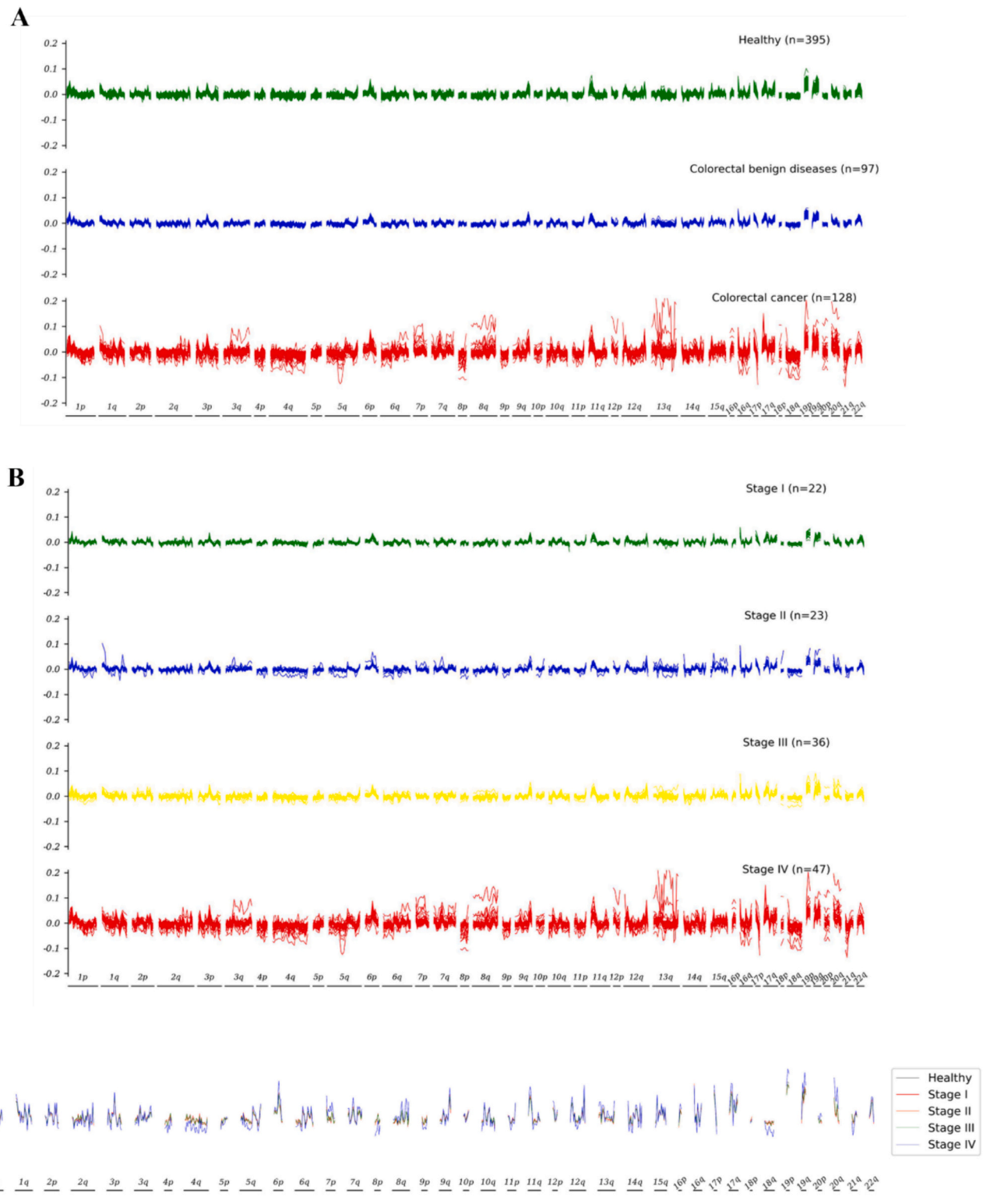

Fig. 2. Fragmentation profiles of cfDNA samples from different cohort. (A) Fragmentation profiles of cfDNA samples from the healthy cohort (curves in green), colorectal benign disease cohort (curves in blue) and CRC cohort (curves in red). (B) Fragmentation profiles of cfDNA samples from CRC patients with different stages. The X-axis represents 5 Mb bins across the human genome, while Y-axis represents the difference to the average ratio of short to long cfDNA fragments for each bin. (C) Simplified scenario of fragmentation profiles of CRC patients at different stages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

extracted for all 620 participants and sorted into three cohorts (Supplementary Table S3). The overall end motif features of the three cohorts were compared, and we found that the motif feature values of the three groups of data were different and had certain regularities on the top 10 selected motifs by proportion (Fig. 3). The motif feature values of healthy individuals were the largest and the smallest of CRC patients. Therefore, 256 motif eigenvalues can be identified (Supplementary Fig. S1), which can be used as markers of CRC prediction model.

3.3. Prediction model for CRC patient detection using cfDNA fragmentation profile

To discriminate between healthy individuals and CRC patients, 523 samples were used for modelling, including a healthy cohort marked as negative (395 samples) and a cancer patient cohort marked as positive (128 samples). Healthy individuals and patient samples were randomly split at a ratio of 8:2 for training and validation, respectively. This ratio of 8:2 is determined by convention. The results on the test set show that our method can accurately detect whether subjects have CRC based on cfDNA information. The sensitivity and specificity were 96.2% and 88.8%, respectively. The optimal ROC curve was plotted as shown in Fig. 4A with an AUC of 0.9494.

To discriminate between colorectal benign patients and CRC, 225 samples were used to build the prediction model, including the colorectal benign diseases cohort (97 samples) representing negative samples and the CRC cohort (128 samples) representing positive samples. Benign and malignant samples were randomly divided in an 8:2 ratio for training and validation, respectively. The validation set gave a sensitivity and specificity of 96.15% and 89.47%, respectively. The optimal ROC curve was plotted as shown in Fig. 4B with an AUC of 0.9575.

Furthermore, we also evaluated the performance of our machine learning methods for discriminating noncancer participants (including healthy individuals and patients with colorectal benign diseases) from CRC patients. As a result, the validation set gave a sensitivity and specificity of 84.62% and 91.84%, respectively. The optimal ROC curve was plotted as shown in Fig. 4C with an AUC of 0.9305. Overall, these results suggests that cfDNA fragmentation profiles may potentially become a noninvasive assay for the detection and stratification of CRC.

3.4. Prediction model for early-stage CRC patient detection using the cfDNA fragmentation profile

To discriminate between healthy individuals and early-stage CRC patients, 440 samples were used for modelling, including a healthy cohort marked as negative (395 samples) and CRC patients at stage I-II marked as positive (45 samples). Healthy individuals and patient samples were randomly split at a ratio of 8:2 for training and validation, respectively. This ratio of 8:2 is determined by convention. The results on the test set show that our method can accurately detect whether subjects have CRC based on cfDNA information. The sensitivity and specificity were 96.2% and 88.8%, respectively. The optimal ROC curve was plotted as shown in Fig. 5A with an AUC of 0.9480.

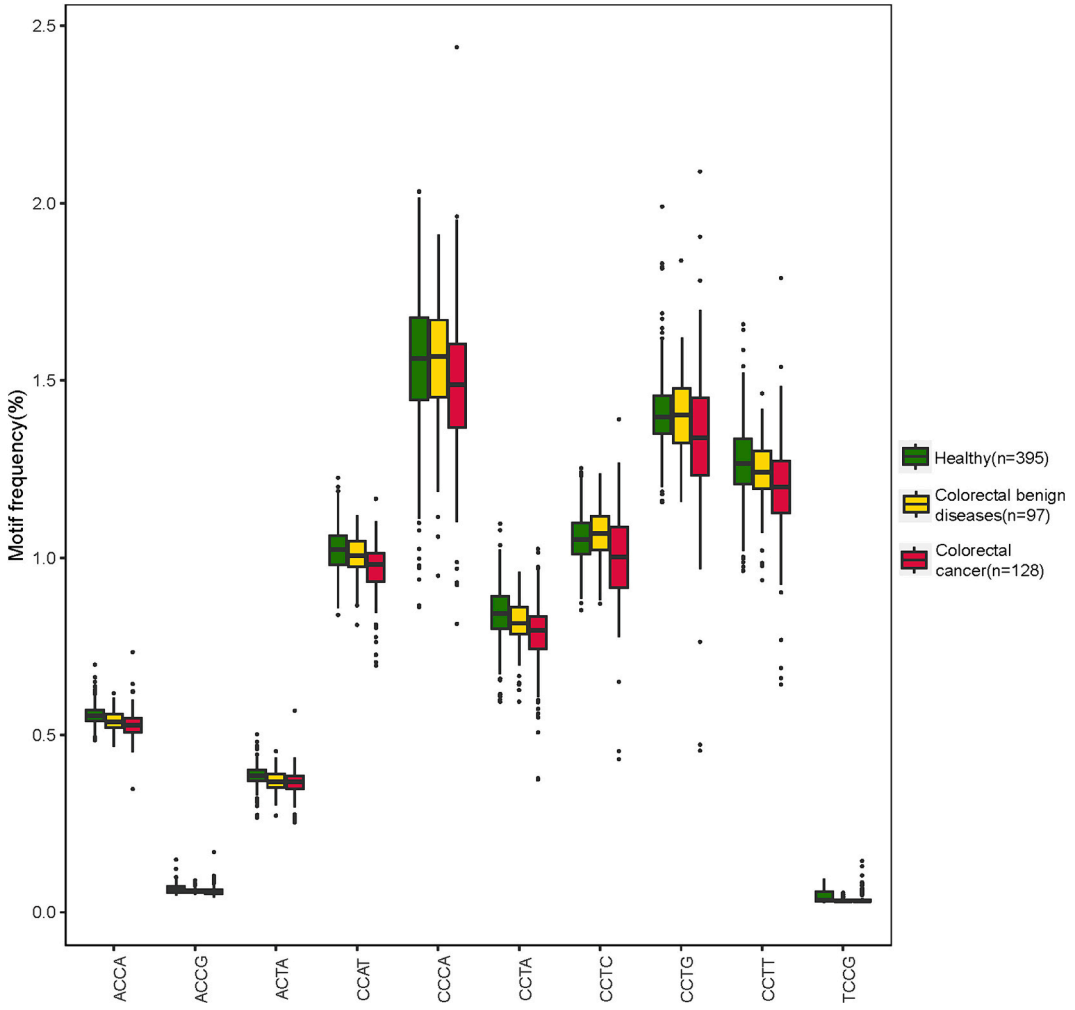

Fig. 3. Motif feature of three cohorts on the top 10 selected motifs by proportion. The abscissa represents the motif sequence, and the ordinate represents the proportion of each of the 10 motifs.

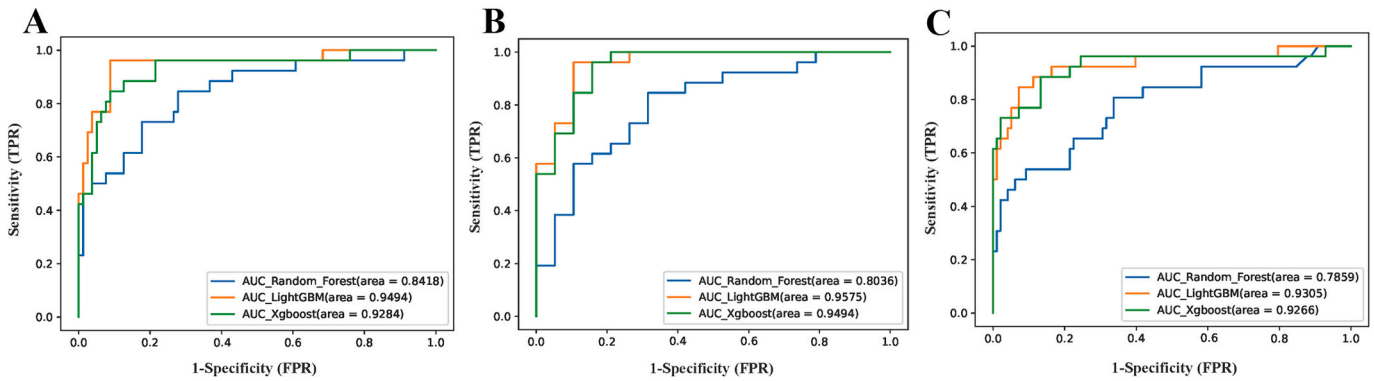

Fig. 4. Machine learning models detect CRC patients with high sensitivity and specificity based on cfDNA fragmentation profiles. (A) ROC curves of the prediction model to discriminate CRC patients from healthy individuals. (B) ROC curves of the prediction model to discriminate CRC patients from patients with colorectal benign diseases. (C) ROC curves of the prediction model to discriminate the malignant patients from noncancer participants.

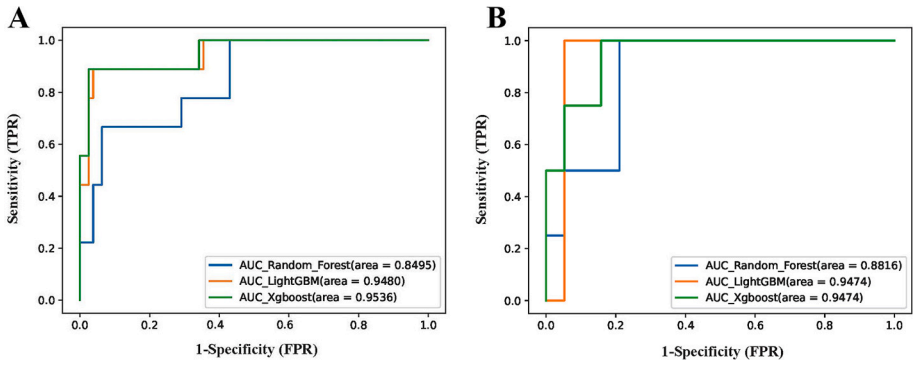

Fig. 5. Machine learning models detect early-stage CRC patients using cfDNA fragmentation profile. (A) ROC curves of the prediction model to discriminate early- stage CRC patients from healthy individuals. (B) ROC curves of the prediction model to discriminate early- stage CRC patients from patients with colorectal benign diseases.

To discriminate between patients with colorectal benign patients and early-stage CRC, 142 samples were used to build the prediction model, including the colorectal benign diseases cohort (97 samples) representing negative and CRC patients at stage I-II (45 samples) representing positive. Benign and malignant samples were randomly divided in an 8:2 ratio for training and validation, respectively. The validation set gave a sensitivity and specificity of 77% and 89.4%, respectively. The optimal ROC curve was plotted as shown in Fig. 5B with an AUC of 0.9474. Altogether, these results indicated that cfDNA fragmentation profiles exhibit a better diagnostic value for the detection of early-stage CRC.

3.5. *CRC-associated transcription factors*

Abnormal gene expression is usually associated with disease [28], especially for the key component for the regulation of gene expression, e.g. transcription factors (TFs). Numerous studies have reported the altered expression of TFs in cancer patients [29]. TFs can be used to predict cancer risk [30] and to differentiate cancer subtypes (such as lung cancer [31]), making the study of TFs related to CRC particularly important.

Gene expression differential analysis was conducted on the three TCGA datasets, with the differential analysis results for the three comparison groups shown in Fig. 6A. Compared to healthy individuals, CRC patients exhibited significant differential expression of 3208 genes (Fig. 6B). Among these differentially expressed genes, several TFs genes, such as ARID3A, BCL6, FOXM1, FOXD2, and ZIC2, were identified, and previous literature has reported that the aberrant expression of the TFs ARID3A, BCL6, FOXM1, FOXD2, ZIC2, etc., is associated with the occurrence and development of CRC [32–36]. Through enrichment analysis via Gene Ontology (GO), the aforementioned genes were identified to be involved in carcinogenic signalling pathways, including pathways related to transcription regulation, cell fate decision and regulation of oncogenes and tumor suppressor genes (Fig. 6C). Furthermore, by analysing the coverage of the TFBS of these TFs in cfDNA data from healthy individuals, patients with colorectal benign diseases, and CRC patients, compared to that in the healthy individual group, there was a significant increase in the depth of HLF binding sites in the colorectal benign diseases group, and an even greater depth in the CRC patient group (Fig. 6D). TFs exhibiting significant inter-group differences in coverage were selected as CRC-specific TFs (Supplementary Table S4). These analyses suggested that the clinicopathological status of CRC patients was closely related to the specific TFs. The aforementioned differential genes of TFs can be used as candidate CRC-specific TFs that will be validated by genome-wide cfDNA fragments in the future studies.

4. Discussion

Over the past decade, multiple invasive and noninvasive screening modalities of CRC have been explored worldwide, including colonoscopy, stool DNA tests, and many more aided by artificial intelligence. However, the sensitivities of these methods are generally insufficient for accurate early detection of CRC [37]. Due to rapid advances in technology, cfDNA detected in patient blood samples is considered a valid indicator of disease progression from tumor occurrence to recurrence [38]. In this study, to obtain the data source to extract cfDNA fragmentomic features, we performed WGS on the plasma cfDNA of 620 participants, including healthy individuals, benign colorectal patients

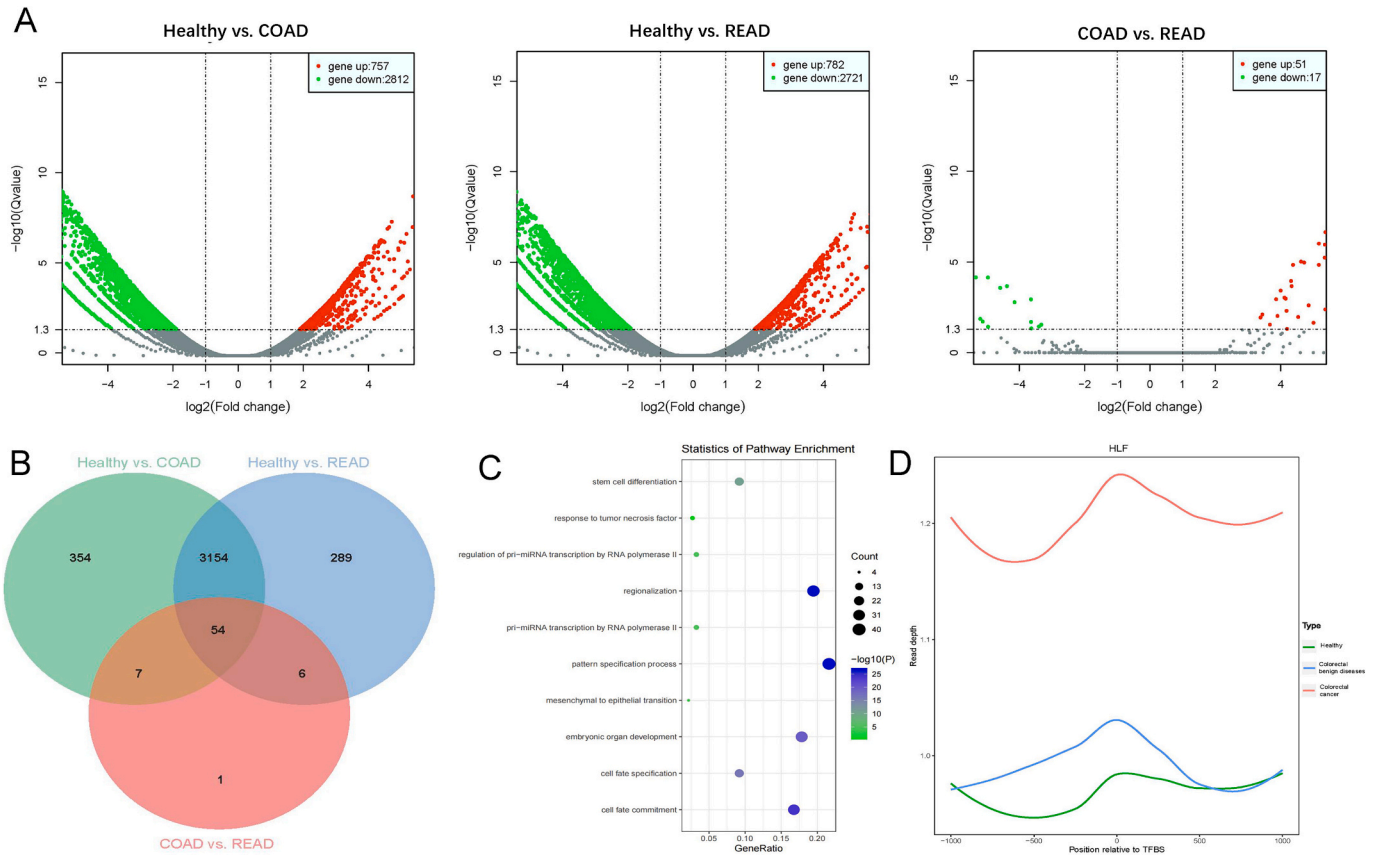

Fig. 6. CRC Candidate TFs Analysis. (A) The results of differential expression analysis among the three comparison groups. (B) The intersection of significantly differentially expressed genes among groups. (C) Enrichment analysis of gene ontology for differential TFs between healthy individuals and CRC patients. The abscissa represents the ratio of number of genes annotated against the total number of genes in each of the pathways. In addition, the ordinates on the left are the names of the GO pathways, while the ordinate on the right are the -log10 (P) (minus log10 of the *P* value obtained using Fisher's exact test and representing the significance of enrichment) and count (number of genes annotated from the data of this study) values. (D) The read depth of HLF binding sites in the healthy cohort (curves in green), colorectal benign disease cohort (curves in blue) and CRC cohort (curves in red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and CRC patients. Then, we built automatic prediction models for early cancer detection using three machine learning methods. The prediction model to discriminate CRC patients from healthy controls and to discriminate CRC patients from benign patients both achieved sensitivity over 90% (92.31% and 96.15%, respectively) and specificity of approximately 90% (91.14% and 89.47%, respectively), which are higher than those of the ctDNA methylation haplotype patterns reported by Mo et al [39]. In technical aspects, all data were obtained only through low pass WGS (approximately 1.2×) in our study, which maintains low cost and labor consumption compared to DNA methylation-based tests. Additionally, the sampling approach of this study was blood draw rather than stool collection, which greatly increases compliance in real-world practice, especially for screening CRC in a large population.

In CRC screening, high sensitivity and specificity in detecting early-stage disease are critical to improve patient outcomes [39]. In this study, we further tested the potential of our method for the early detection of CRC by setting up prediction models to discriminate early-stage CRC patients (stage 0-II) from healthy individuals or colorectal benign patients and these two models showed high diagnostic accuracy for early-stage CRC (AUC = 0.9480 and 0.9474, respectively). However, the sensitivity of the model for the detection of early-stage CRC from colorectal benign patients is only 77%, which is probably due to the small number of negative samples input in the validation set of this model. Therefore, we admit that the precision of our method depends highly on the sampling size and the balance between the positive vs. negative groups. In this sense, larger and balanced sampling of colorectal benign patients and stage 0-II CRC patients should increase the precision of prediction models based on machine learning algorithms, which can be validated further in a larger cohort study.

Furthermore, to test if our method could discriminate all noncancer participants from CRC patients, we combined a healthy cohort and a patient cohort with colorectal benign patients together to form a noncancer group to be marked as negative against the CRC patient cohort marked as positive. This prediction model also achieved remarkable performance with a sensitivity of 84.62% and specificity of 91.84%, which provide enough precision for clinical utility.

Of note, by analysing CRC transcriptome data from the TCGA database and the coverage of TFBS in cfDNA data from each group. we identified specific TFs that are closely associated with clinical pathological status. The fragmentation patterns of plasma cfDNA can reflect the in vivo gene-regulation status across multiple molecular layers, such as nucleosome positioning and gene expression [40,41]. Recent studies have reported specific TFs used for distinguishing small cell lung cancers from non–small cell lung cancers [31] or detecting of liver cancer [17]. Thus, the analysis of disease-specific TFs using genome-wide cfDNA fragments may improve the detection and identification of tissues of origin in cancer patients [29]. Furthermore, future prospective studies are needed to improve machine learning algorithms to detect CRC in combination with CRC-specific TFs.

In summary, this study has identified clearly different profiles of cfDNA fragmentation of the CRC cohort compared to the healthy cohort or colorectal benign cohort, including both the cfDNA fragmentation profile and end motifs. Based on these molecular traits, the detection

and staging of CRC achieved considerable performance. Therefore, this is a proof of principle study for a minimally invasive, accurate, cost-efficient and convenient approach for the early detection of CRC.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2024.110876.

Funding

This study was supported by the Guangdong Basic and Applied Basic Research Foundation (grant number 2019A1515110060), Guangzhou Science and Technology Department-School Joint Project (grant number 2023A03J0417), Guangzhou Medical University Student Innovation Enhancement Project, Plan on Enhancing Scientific Research in GMU (Grant Number 02-410-2302035XM), and Shenzhen Science and Technology Innovation Commission Project (KCXFZ20200201101050887, ZDSYS20190902092855097, and GJHZ20200731095207023), Shenzhen San-Ming Project of Medicine (No.SZSM202211036).

Ethics approval and consent to participate

The protocol obtained review and approval from the Ethics and Scientific Committee of the Second Affiliated Hospital of Guangzhou Medical University (NO. 2022-YJS-ks-22), Peking University Shenzhen Hospital (NO. 2021–075), the Second Affiliated Hospital of Shenzhen University (BYL20230310), and Shenzhen Hospital, University of Chinese Academy of Sciences (LL-KT-2022053). All participants provided written informed consent for the scientific use of their clinical data and sample.

Consent for publication

Not applicable.

Author statement

J.Z analyzed data, and contributed to writing and revising the manuscript; Y.P performed data processing and analysed data; S.W, G. W, C.G, Z.T, C J-H, Z.H, X.L, S.X, and K.Y recruited patients; J.Z, W.H and Z.Z performed data processing and analysed data; Q.W and X.G contributed to revise the manuscript and discussion; R.R and B.H contributed to the discussion and critical evaluation of the manuscript; M.F performed WGS experiments; P.C contributed to the discussion and critical evaluation of the manuscript; H.Y designed experiments, analysed data and provided overall direction. All authors reviewed the manuscript.

CRediT authorship contribution statement

Jiyuan Zhou: Writing – review & editing, Writing – original draft, Funding acquisition. **Shubing Wang:** Resources, Funding acquisition, Data curation. **Guoqiang Wang:** Resources, Data curation. **Chengxin Gu:** Resources, Investigation, Data curation. **Jinxin Zhu:** Methodology, Formal analysis. **Zhenlin Tan:** Project administration, Investigation, Data curation. **Qixian Wu:** Validation, Software, Methodology. **Weihuang He:** Methodology, Formal analysis, Data curation. **Xiaohui Lin:** Formal analysis, Conceptualization. **Shu Xu:** Data curation. **Kehua Yuan:** Supervision, Software, Data curation. **Ziwen Zheng:** Supervision, Data curation. **Xiaoqing Gong:** Data curation. **Chenhao JiangHe:** Resources, Data curation. **Zhoujian Han:** Data curation. **Bingding Huang:** Formal analysis, Data curation. **Ruyun Ruan:** Methodology, Data curation, Conceptualization. **Mingji Feng:** Supervision, Software, Project administration, Data curation. **Pin Cui:** Writing – review & editing, Visualization, Supervision, Project administration. **Hui Yang:** Writing – review & editing, Visualization, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We would like to thank the patients who gave their consent to present the data in this study.

References

[1] N. Keum, E. Giovannucci, Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies, Nat. Rev. Gastroenterol. Hepatol. 16 (12) (2019) 713–732.
[2] E. Dekker, P.J. Tanis, J.L.A. Vleugels, P.M. Kasi, M.B. Wallace, Colorectal cancer, Lancet 394 (10207) (2019) 1467–1480.
[3] M.G. Guren, The global challenge of colorectal cancer, Lancet Gastroenterol. Hepatol. 4 (12) (2019) 894–895.
[4] Y. Xi, P. Xu, Global colorectal cancer burden in 2020 and projections to 2040, Transl. Oncol. 14 (10) (2021) 101174.
[5] K.H. Khan, D. Cunningham, B. Werner, G. Vlachogiannis, I. Spiteri, T. Heide, J. F. Mateos, A. Vatsiou, A. Lampis, M.D. Damavandi, et al., Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal Cancer clinical trial, Cancer Discov. 8 (10) (2018) 1270–1285.
[6] D.M. Hyman, B.S. Taylor, J. Baselga, Implementing genome-driven oncology, Cell 168 (4) (2017) 584–599.
[7] J.H. Strickler, J.M. Loree, L.G. Ahronian, A.R. Parikh, D. Niedzwiecki, A.A. L. Pereira, M. McKinney, W.M. Korn, C.E. Atreya, K.C. Banks, et al., Genomic landscape of cell-free DNA in patients with colorectal Cancer, Cancer Discov. 8 (2) (2018) 164–173.
[8] M. Fleischhacker, B. Schmidt, Circulating nucleic acids (CNAs) and cancer–a survey, Biochim. Biophys. Acta 1775 (1) (2007) 181–232.
[9] M. Ivanov, A. Baranova, T. Butler, P. Spellman, V. Mileyko, Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation, BMC Genomics (2015) S1.
[10] D. Chandrananda, N. Thorne, M. Bahlo, High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA, BMC Med. Genet. 8 (2015) 29.
[11] C. Sanchez, B. Roch, T. Mazard, P. Blache, Z. Dache, B. Pastor, E. Pisareva, R. Tanos, A. Thierry, Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics, JCI Insight 6 (7) (2021).
[12] A. Thierry, Circulating DNA fragmentomics and cancer screening, Cell Genom. 3 (1) (2023) 100242.
[13] S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D.C. Bruhm, S.O. Jensen, J. E. Medina, C. Hruban, J.R. White, et al., Genome-wide cell-free DNA fragmentation in patients with cancer, Nature 570 (7761) (2019) 385–389.
[14] P. Jiang, K. Sun, W. Peng, S.H. Cheng, M. Ni, P.C. Yeung, M.M.S. Heung, T. Xie, H. Shang, Z. Zhou, et al., Plasma DNA end-motif profiling as a Fragmentomic marker in Cancer, pregnancy, and transplantation, Cancer Discov. 10 (5) (2020) 664–673.
[15] Y.V. Zhitnyuk, A.P. Koval, A.A. Alferov, Y.A. Shtykova, I.Z. Mamedov, N. E. Kushlinskii, D.M. Chudakov, D.S. Shcherbo, Deep cfDNA fragment end profiling enables cancer detection, Mol. Cancer 21 (1) (2022) 26.
[16] Y.M.D. Lo, C. Gianni, Cell-free DNA Fragmentomics: a promising biomarker for diagnosis, prognosis and prediction of response in breast Cancer, PLoS Genet. 23 (22) (2022).
[17] Z.H. Foda, A.V. Annapragada, K. Boyapati, D.C. Bruhm, N.A. Vulpescu, J. E. Medina, D. Mathios, S. Cristiano, N. Niknafs, H.T. Luu, et al., Detecting liver Cancer using cell-free DNA Fragmentomes, Cancer Discov. 13 (3) (2023) 616–631.
[18] W. Guo, X. Chen, R. Liu, N. Liang, Q. Ma, H. Bao, X. Xu, X. Wu, S. Yang, Y. Shao, et al., Sensitive detection of stage I lung adenocarcinoma using plasma cell-free DNA breakpoint motif profiling, EBioMedicine 81 (2022) 104131.
[19] S. Chen, Y. Zhou, Y. Chen, J. Gu, Fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890.
[20] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[21] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, et al., Twelve years of SAMtools and BCFtools, Gigascience 10 (2) (2021).
[22] M. Snyder, M. Kircher, A. Hill, R. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin, Cell 164 (2016) 57–68.
[23] F. Mouliere, B. Robert, E. Arnau Peyrotte, M. Del Rio, M. Ychou, F. Molina, C. Gongora, A. Thierry, High fragmentation characterizes tumour-derived circulating DNA, PLoS One 6 (9) (2011) e23418.

[24] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Proces. Syst. 30 (2017) 3146–3154.
[26] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: 2016, 2016, pp. 785–794.
[27] F. Pedregosa, Scikit-learn: machine learning in Python, J. Mach. Learn. 12 (2011) 2825–2830.
[28] S. Kravitz, E. Ferris, M. Love, A. Thomas, A. Quinlan, C. Gregg, Random allelic expression in the adult human body, Cell Rep. 42 (1) (2023) 111945.
[29] P. Ulz, S. Perakis, Q. Zhou, T. Moser, J. Belic, I. Lazzeri, A. Wölfler, A. Zebisch, A. Gerger, G. Pristauz, et al., Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection, Nat. Commun. 10 (1) (2019) 4666.
[30] V. Klepsch, R.R. Gerner, Nuclear orphan receptor NR2F6 as a safeguard against experimental murine colitis, Gut 67 (8) (2018) 1434–1444.
[31] D. Mathios, J. Johansen, S. Cristiano, J. Medina, J. Phallen, K. Larsen, D. Bruhm, N. Niknafs, L. Ferreira, V. Adleff, et al., Detection and characterization of lung cancer using cell-free DNA fragmentomes, Nat. Commun. 12 (1) (2021) 5060.
[32] J. Tang, L. Yang, Y. Li, X. Ning, A. Chaulagain, T. Wang, D. Wang, ARID3A promotes the development of colorectal cancer by upregulating AURKA, Carcinogenesis 42 (4) (2021) 578–586.
[33] X. Meng, J. Peng, X. Xie, F. Yu, W. Wang, Q. Pan, H. Jin, X. Huang, H. Yu, S. Li, et al., Roles of lncRNA LVBU in regulating urea cycle/polyamine synthesis axis to promote colorectal carcinoma progression, Oncogene 41 (36) (2022) 4231–4243.
[34] L. Jun, L. Jipeng, W. Ke, L. Haiming, S. Jianyong, Z. Xinhui, Y. Yanping, Q. Yihuan, W. Ye, Z. Xiaofang, et al., Aberrantly high activation of a FoxM1-STMN1 axis contributes to progression and tumorigenesis in FoxM1-driven cancers, Signal Transduct. Target. Ther. 6 (2021).
[35] K. Hyo-Min, K. Byunghee, P. Sohyun, P. Hyorim, K. Chan Johng, L. Hyeonji, Y. Mijoung, K. Mi-Na, I. Sin-Hyeog, K. Tae Il, et al., Forkhead box protein D2 suppresses colorectal cancer by reprogramming enhancer interactions, Nucleic Acids Res. 51 (2023).
[36] L. Fangting, S. Zhehao, B. Wenming, Z. Jiuyi, C. Kaiyu, L. Zhihui, S. Hao-Nan, L. Xin, D. Qiantong, J. Lei, et al., ZIC2 promotes colorectal cancer growth and metastasis through the TGF-β signaling pathway, Exp. Cell Res. 415 (2022).
[37] P. Kanth, J.M. Inadomi, Screening and prevention of colorectal cancer, BMJ 374 (2021) n1855.
[38] X. Zhou, Z. Cheng, M. Dong, Q. Liu, W. Yang, M. Liu, J. Tian, W. Cheng, Tumor fractions deciphered from circulating cell-free DNA methylation for cancer early diagnosis, Nat. Commun. 13 (1) (2022) 7694.
[39] S. Mo, W. Dai, H. Wang, X. Lan, C. Ma, Z. Su, W. Xiang, L. Han, W. Luo, L. Zhang, et al., Early detection and prognosis prediction for colorectal cancer by circulating tumour DNA methylation haplotypes: a multicentre cohort study, EClinicalMedicine 55 (2023) 101717.
[40] Y. Liu, At the dawn: cell-free DNA fragmentomics and gene regulation, Br. J. Cancer 126 (3) (2022) 379–390.
[41] M. Esfahani, E. Hamilton, M. Mehrmohamadi, B. Nabet, S. Alig, D. King, C. Steen, C. Macaulay, A. Schultz, M. Nesselbush, et al., Inferring gene expression from cell-free DNA fragmentation profiles, Nat. Biotechnol. 40 (4) (2022) 585–597.