Genomics 114 (2022) 110502

Contents lists available at ScienceDirect

Genomics

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

Detection and monitoring of HBV-related hepatocellular carcinoma from plasma cfDNA fragmentation profiles

Xinfeng Sun [a,b,1], Wenxing Feng [a,b,1], Pin Cui [c,1], Ruyun Ruan [d,1], Wenfeng Ma [a,b], Zhiyi Han [a,b], Jialing Sun [a,b], Yuanke Pan [d], Jinxin Zhu [d], Xin Zhong [a,b], Jing Li [a,b], Mengqing Ma [a,b], Rui Hu [a,b], Minling Lv [a,b], Qi Huang [a,b], Wei Zhang [a,b], Mingji Feng [c], Xintao Zhuang [c], Bingding Huang [d,**], Xiaozhou Zhou [a,b,*]

[a] Department of Liver Disease, the fourth Clinical Medical School, Guangzhou University of Chinese Medicine, Shenzhen 518033, China
[b] Department of Liver Disease, Shenzhen Traditional Chinese Medicine Hospital, Shenzhen 518033, China
[c] Shenzhen Rapha Biotechnology Incorporate, Shenzhen 518118, China
[d] College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China

ARTICLE INFO

Keywords:
Hepatocellular carcinoma
cfDNA
ctDNA
Machine learning
Fragmentation profile

ABSTRACT

Most hepatocellular carcinomas (HCCs) are associated with hepatitis B virus infection (HBV) in China. Early detection of HCC can significantly improve prognosis but is not yet fully clinically feasible. This study aims to develop methods for detecting HCC and studying the carcinogenesis of HBV using plasma cell-free DNA (cfDNA) whole-genome sequencing (WGS) data. Low coverage WGS was performed for 452 participants, including healthy individuals, hepatitis B patients, cirrhosis patients, and HCC patients. Then the sequencing data were processed using various machine learning models based on cfDNA fragmentation profiles for cancer detection. Our best model achieved a sensitivity of 87.10% and a specificity of 88.37%, and it showed an increased sensitivity with higher BCLC stages of HCC. Overall, this study proves the potential of a non-invasive assay based on cfDNA fragmentation profiles for the detection and prognosis of HCC and provides preliminary data on the carcinogenic mechanism of HBV.

1. Introduction

According to WHO statistics, primary liver cancer (PLC) ranks sixth for incidence rate and third for mortality among all malignant tumors, with approximately half of cases originating in China [1]. Hepatocellular carcinoma (HCC) accounts for ~85% of all PLC cases and is subject to a low five-year survival rate (10–19%) [2]. Common risk factors of HCC include chronic hepatitis B virus (HBV) and hepatitis C virus (HCV) infections, alcohol addiction, metabolic liver disease (particularly nonalcoholic fatty liver disease), exposure to dietary toxins such as aflatoxins and aristolochic acid [3], and cirrhosis [4]. Upon diagnosis, most HCC patients are already in stage C or D (Barcelona clinic liver cancer [BCLC] staging), for which any radical therapeutic intervention, like surgical resection, is no longer feasible. In contrast, the five-year survival rate for HCC at BCLC stage A can be as high as 40%.

Serological tests and imaging examinations are widely used in HCC surveillance. Serum Alpha-fetoprotein (AFP) test in combination with biannual liver ultrasound is currently a major strategy for screening HCC in high-risk patients [5]. Other serum biomarkers used in the clinical setting include AFP lectin fraction (AFP-L3), prothrombin induced by vitamin K absence-II (PIVKA-II), alpha-fucosidase, and glypican-3 [6]. However, AFP suffers from low accuracy with a specificity of 85–90% and sensitivity of 18–60% and is especially inefficient for early-stage liver neoplasm; an abnormal increase of AFP cannot be detected in serum for around 80% of patients with liver nodules [7]. For small tumor lesions (approximately 10–20 mm in size), the sensitivity of

Abbreviations: cfDNA, cell-free DNA; ctDNA, circulating tumor DNA; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; HCV, hepatitis C virus; WGS, whole-genome sequencing.

* Corresponding author at: Department of Liver Disease, the fourth Clinical Medical School, Guangzhou University of Chinese Medicine, Shenzhen 518033, China.
** Corresponding author.
E-mail addresses: huangbingding@sztu.edu.cn (B. Huang), zxz1006@gzucm.edu.cn (X. Zhou).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2022.110502
Received 3 May 2022; Received in revised form 30 September 2022; Accepted 4 October 2022
Available online 8 October 2022

multidetector CT combined with MRI for HCC diagnosis is only 55.1% [8]. The high expense, radiation exposure, and the risk of contrast agent allergy of CT or MRI prevent use in large-scale screening [9]. Hence, accurate and affordable means for early detection of HCC is a clinical necessity, although difficult to achieve. Moreover, the rate of relapse and metastasis remains high for HCC patients, even for BCLC stage A patients (50–70%) [10], and, therefore, a method for accurate prognosis is also urgently required.

Although antiviral therapies have gained success in curing HCV infections, such treatments are only partially effective against and cannot eradicate HBV [11]. Annually there are 292 million global HBV infections, with which more than half of all HCC cases are associated [3], making HBV infection the top factor for the carcinogenesis of HCC [12]. Through clinical observations, clinicians have reached a consensus for the causative role of HBV to a series of liver diseases, including hepatitis, cirrhosis, and eventually cancer [13]. However, the molecular mechanism for this process has not been thoroughly studied and therefore requires clarification [14].

The past decade has witnessed the broad application of cell-free DNA (cfDNA) as a non-invasive means for the diagnosis and prognosis of many cancer types [15], known as tumor liquid biopsy. Plasma cfDNAs are degraded DNA fragments shed into blood circulation resulting from cell death in surrounding tissues, both healthy and tumorous. Circulating tumor DNA (ctDNA) refers to the portion of cfDNAs originating from tumor cells [16]. Two kinds of tumor-specific alterations in ctDNA (present in the mixture of total cfDNA) can be used as biomarkers: genetic and epigenetic [17]. The genetic aspect represents tumorigenic mutations, especially driver or actionable mutations, which have been clinically utilized for companion diagnosis, treatment guide, and prognosis in the form of targeted sequencing panels. However, such applications are limited to middle- and late-stage cancers since the abundance of tumorous genetic biomarkers is dramatically lower in early-stage cancers due to the low proportion of ctDNA in the total cfDNA (0.01%–1%). Routinely, 8–10 mL plasma is used for a mutation detection panel, yielding 20–30 ng of total cfDNA. However, for early-stage cases, it might contain less than 0.3 ng ctDNA, representing only 1–100 genome copies, insufficient for any conventional genetic test. Hence, the genetic approach requires infeasibly high volumes of plasma and is impractical for early cancer detection [18].

Numerous studies have demonstrated the feasibility of cfDNA based epigenetic biomarkers and their advantages over mutation markers for cancer early detection, owing to their early occurrence in the carcinogenic process, wide distribution in the genome, and fingerprinting for tissue of origin [17,19]. The cfDNA fragmentome combines features like cfDNA fragmentation size and end motif, resulting from different nucleosome assembly patterns and copy number variations between healthy and tumorous tissues. A previous study (DELFI) developed an approach for cancer detection by exploring the discrepancies between plasma cfDNA fragmentation profiles of 236 patients across seven cancer types and that of healthy individuals, termed DNA evaluation of fragments for early interception [20]. DELFI incorporated a machine learning algorithm and achieved sensitivities of detection ranging from 57% to more than 99% across seven cancer types at 98% specificity. However, DELFI required relatively large amounts of sequencing data (2× coverage on average, but up to 9× coverage in some cases) and contained no data about HCC. Furthermore, the authenticity and repeatability of DELFI, which relied solely on fragmentation profile, have been questioned since it lacks the support of a large clinical sample size [17].

To systematically investigate the cfDNA fragmentation profile of the HBV-associated transition from chronic hepatitis B to cirrhosis, and finally to HCC, we recruited patients with hepatitis B, cirrhosis, and HBV-related HCC, together with healthy individuals. Patients in all disease cohorts were clinically confirmed to have chronic HBV infections. This study developed a low-coverage whole-genome sequencing (WGS) assay and applied to all plasma cfDNA samples. Then we developed different machine-learning classifiers based on cfDNA WGS fragmentation profiles to discriminate HCC patients from healthy people. Distinguishable patterns in cfDNA fragmentation profiles were observed between different BCLC stages of HCC, between different disease phases of HCC patients with HBV infection and whether the patient had undergone antiviral treatment. Therefore, we present a proof-of-principle study for a cfDNA-based non-invasive test for sensitive and cost-efficient early detection of HCC, and molecular clues for the carcinogenic effect of HBV infection to HCC in this study.

2. Material and methods

2.1. Study design and participant enrollment

In total, 452 participants were recruited for this study, including a healthy cohort (143 individuals) and four patient cohorts (100 hepatitis B patients, 99 cirrhosis patients, 105 HBV-related HCC patients, 5 HCV-related HCC patients; all patients except for the 5 HCV-related ones were clinically confirmed to have chronic HBV infections). The cohorts of our study were defined as follows: (1) the participants without any cancer or chronic viral liver disease were enrolled as a healthy cohort; (2) Hepatitis B patients were participants who have chronic hepatitis B but without advanced fibrosis and cirrhosis based on medical imaging; (3) Cirrhosis participants with HBV or HCV infection were diagnosed based on tissue biopsy or medical imaging such as abdominal ultrasound, magnetic resonance elastography (MRE), Fibroscan, etc., according to the Chinese guidelines on the management of liver cirrhosis [21]. (4) HCC patients were clinically diagnosed in combination with clinical tests (such as AFP) and medical imaging (CT, MRI, etc.) or tissue biopsy according to the Chinese guidelines for diagnosis and treatment of primary liver cancer (2019 edition) [22]. All patients were enrolled at the time of diagnosis in the Shenzhen Traditional Chinese Medicine Hospital from March to December 2021. All protocols of this study were approved by the ethics committee of Shenzhen Traditional Chinese Medicine Hospital with Approval No. 201858 and were performed following international standards of good clinical practice. All participants were acknowledged in this study with informed consent. Detailed information such as age, gender, and related clinical records are listed in Supplementary Table S1.

2.2. Cell-free DNA extraction from plasma samples

Blood samples were collected from all these 452 participants, and 6–10 mL of peripheral blood was drawn from each participant into a cfDNA preservation tube (cat. 20,092,421, Hebei Xinle Medical Instrument Technology Inc., Xinle, China) and shipped at room temperature to the Clinical Laboratory (Shenzhen RAFA Biotechnology Inc., China) for cfDNA extraction within 72 h after blood draw. All participants underwent experiments and data analysis workflow illustrated in Fig. 1.

Plasma was obtained by centrifuging whole blood at 1600*g* for 10 min. The supernatant was transferred to a new tube and 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 ultra-high concentration were identified as contaminated with genomic DNA (mainly from dead white blood cells during logistics).

2.3. Whole-genome sequencing library construction and sequencing

For all 452 participants in this study, whole-genome sequencing

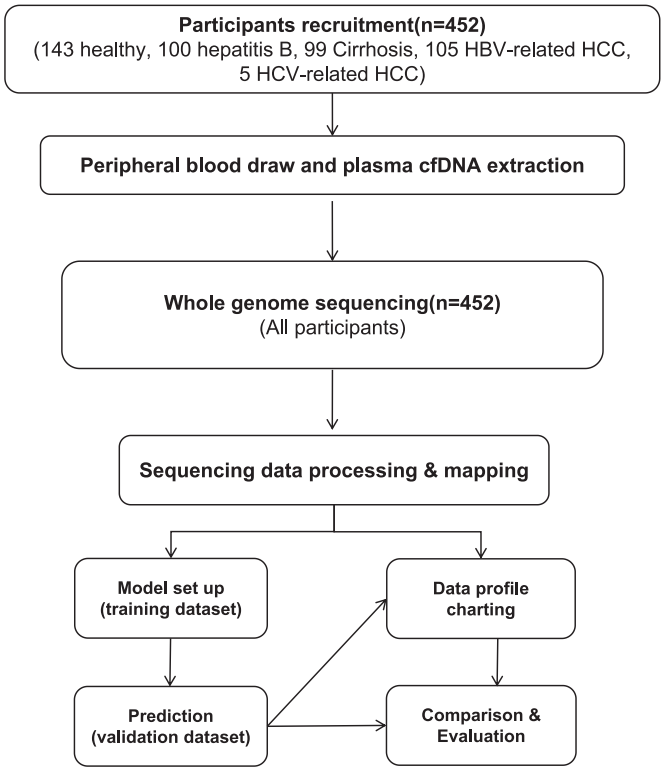

Fig. 1. The workflow of our study.

(WGS) was performed using 10 ng of cfDNA input for each participant. WGS libraries were constructed using 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). Then the libraries were sent for sequencing in batches of 24, each on one lane of MGI-2000 sequencer (BGI Genomics Inc., Wuhan, China) using DNBSEQ™ technology and sequencing mode.

2.4. WGS data processing

A total of 452 WGS data (covering all cohorts of this study) were included in the data analysis. First, raw sequencing data were filtered by fastp [23] 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 [24]. PCR duplicates were then marked by Samtools [25]. The fragment size for every read pair with a mapping quality score below 30 for either read was extracted from every sample by in-house scripts. Short fragments were defined as having lengths between 100 and 150 bp, and long fragments were between 151 and 220 bp. The fragment profile was generated using the short/long fragments ratio, as previously described in the DELFI approach [20]. 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 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/) and the low mappability regions. All these 472 fragment features were then used as the input for building the prediction model for HCC detection.

2.5. Prediction model for HCC detection using WGS data

To build an automatic prediction model for HCC detection using WGS data, several popular machine learning methods (including Random Forest [26], LightGBM [27], XGBoost [28]) and 1 to 15-layer neuron networks were implemented and tested. GridSearchCV [29] was used for searching hyper-parameters such as the number of estimators, max depth, and learning rate. Eventually, parameters were set as 150 estimators, max depth 2, and learning rate 0.1 for training the model. In the end, the best prediction model was selected for further investigation.

In the first step, 143 samples from the healthy cohort and 105 samples from the HBV-related HCC cohort were used to build a model, namely the Healthy/HCC discrimination model. 70% of samples from the healthy cohort, as well as 70% of samples from the HBV-related HCC group were randomly selected as the training set, with the remaining 30% of samples from the healthy cohort and HBV-related HCC group used as the validation set to build this model.

In the second step, considering a significant portion of hepatitis B patients and cirrhosis patients would eventually develop to HCC [2], the Healthy/HCC discrimination model was applied to the data of hepatitis B and cirrhosis patients to distinguish HCC-like Hepatitis B from Healthy-like Hepatitis B, and HCC-like Cirrhosis from Healthy-like Cirrhosis based on cfDNA fragmentation profile, respectively. In either of these cases, all samples of healthy cohorts and HBV-related HCC cohorts were used as a training set, while samples from either the hepatitis B or cirrhosis cohorts were used as a validation set.

In the third step, to build a model to discriminate HCC samples from non-HCC samples, we set up a non-HCC group containing all samples from the healthy cohort, Healthy-like hepatitis B, and Healthy-like Cirrhosis. We randomly selected 70% samples from the non-HCC group and 70% samples from the HBV-related HCC cohort as the training set, and the remaining 30% samples from the non-HCC group and the HBV-related HCC cohort as the validation set to build this model.

To assess the performance of our prediction models, the sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)], and accuracy [(TP + TN)/(TP + FP + TN + FN)] were calculated using Numpy (v 1.18.5) (TP: true-positive; TN: true-negative; FP: false-positive; FN: false-negative). For area under curve (AUC) evaluation, the cut-off value for charting the receiver operating characteristic (ROC) curve was calculated using the Sklearn (v 1.0.2) package of Python 3.8.

3. Results

3.1. Participants characteristics

As listed in Supplementary Table S1, 452 participants were recruited in this study, including 143 healthy individuals, 100 hepatitis B patients, 99 cirrhosis patients, 105 HBV-related HCC patients, and 5 HCV-related HCC patients. All patients except for the HCV-related HCC cohort had been clinically diagnosed with chronic HBV. The average ages of the four major cohorts, healthy, hepatitis B, cirrhosis, and HBV-related HCC, were 32, 45, 54, and 59, respectively.

3.2. Plasma cfDNA fragmentation profiles

The average sequencing depth for each cfDNA WGS data ranged from 1.49× to 4.65× (see Supplementary Table S2).To visualize the cfDNA fragmentation profiles obtained from WGS data for each sample, we plotted fragment feature of the bin minus 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. Significant differences of cfDNA fragmentation profile between four cohorts (healthy individuals, hepatitis B patients, cirrhosis patients, and HBV-related HCC patients) were observed and shown in Fig. 2. In general, the profiles of the three patient

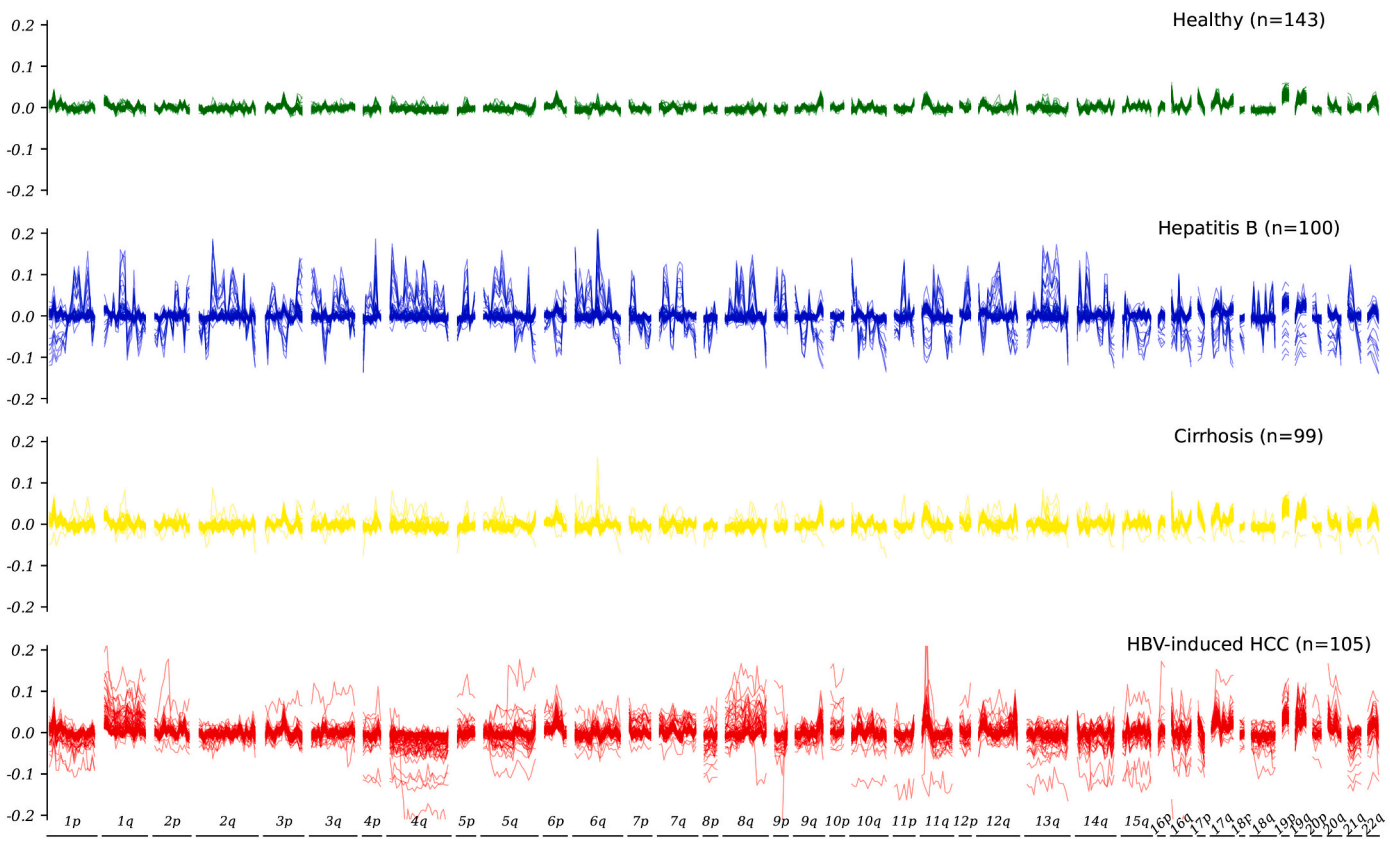

Fig. 2. Plasma cfDNA fragmentation profiles of healthy individuals (curves in green), hepatitis B patients (curves in blue), cirrhosis patients (curves in yellow) and HBV-related HCC patients (curves in red). The X-axis represents 5 Mb bins across the human genome, while Y-axis represents the difference to the average ratio of short to long cfDNA fragments for each bin. The cfDNA fragmentation profile of the healthy cohort shows significant difference from those non-healthy groups (Hepatitis B, Cirrhosis, HCC). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cohorts showed much stronger fluctuations than the healthy cohort, while the fluctuation of the curves of healthy individuals was limited to a very narrow range. In contrast, the profile of hepatitis B patients shows strong and sharp peaks across the whole genome. The cirrhosis patient profile showed less sharp and lower peaks than the hepatitis B patient profile. The HCC patient group also showed significant abnormalities in the cfDNA fragmentation profile, which appeared in a parallel undulation pattern similar to the cirrhosis cohort but with a much broader range of fluctuation. Next, we plotted the cfDNA fragmentation profile of HCC patients colored for each BCLC stage. As shown in Fig. 3 and Supplementary Fig. S1, we concluded that as the stage increased, the fragmentation profile of HCC patients deviated further from the profile range of the healthy cohort, especially in some chromosome arms such as 4q, 8q, and 14q.

3.3. Prediction model for HCC detection using WGS data

To evaluate the performance of our machine learning methods, the healthy cohort and HBV-related HCC cohort were randomly assigned to the training dataset and validation dataset with a 7:3 ratio. The validation result of these methods are shown in Fig. 4A and B. As a result, the XGBoost method outperformed other machine learning and neural-network methods in terms of AUC and accuracy values. Overall, the discrimination model using XGBoost gives a sensitivity of 87.10% and a specificity of 88.37% for discriminating healthy individuals from HBV-related HCC patients.

Secondly, we found significant differences in cfDNA fragmentation profiles among the healthy cohort, hepatitis B cohort, cirrhosis cohort, and different BCLC stages in the HCC cohort. In the meantime, clinical observations showed that some hepatitis B and cirrhosis cases would eventually develop to HCC. Therefore, we investigated whether a machine learning model could be built to predict the prognosis of hepatitis

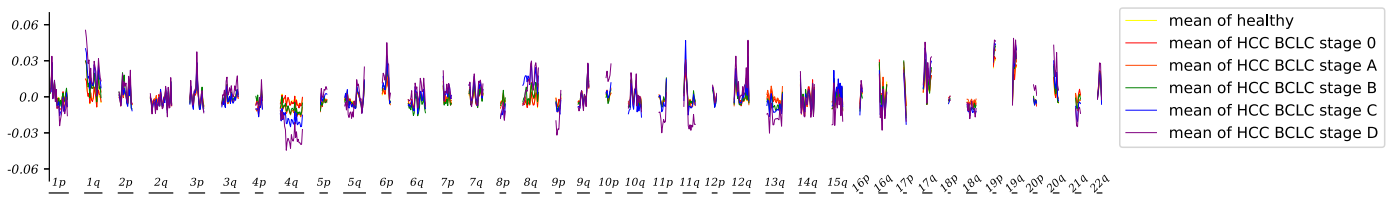

Fig. 3. The plasma cfDNA fragmentation profiles for HCC patients with different BCLC stages. The yellow line represents the average profile of healthy individuals, while the red line represents HBV-related HCC patients at stage 0. The curves in orange, green, blue, and purple represent the average profile of HBV-related HCC patients at BCLC stages A, B, C, and D, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

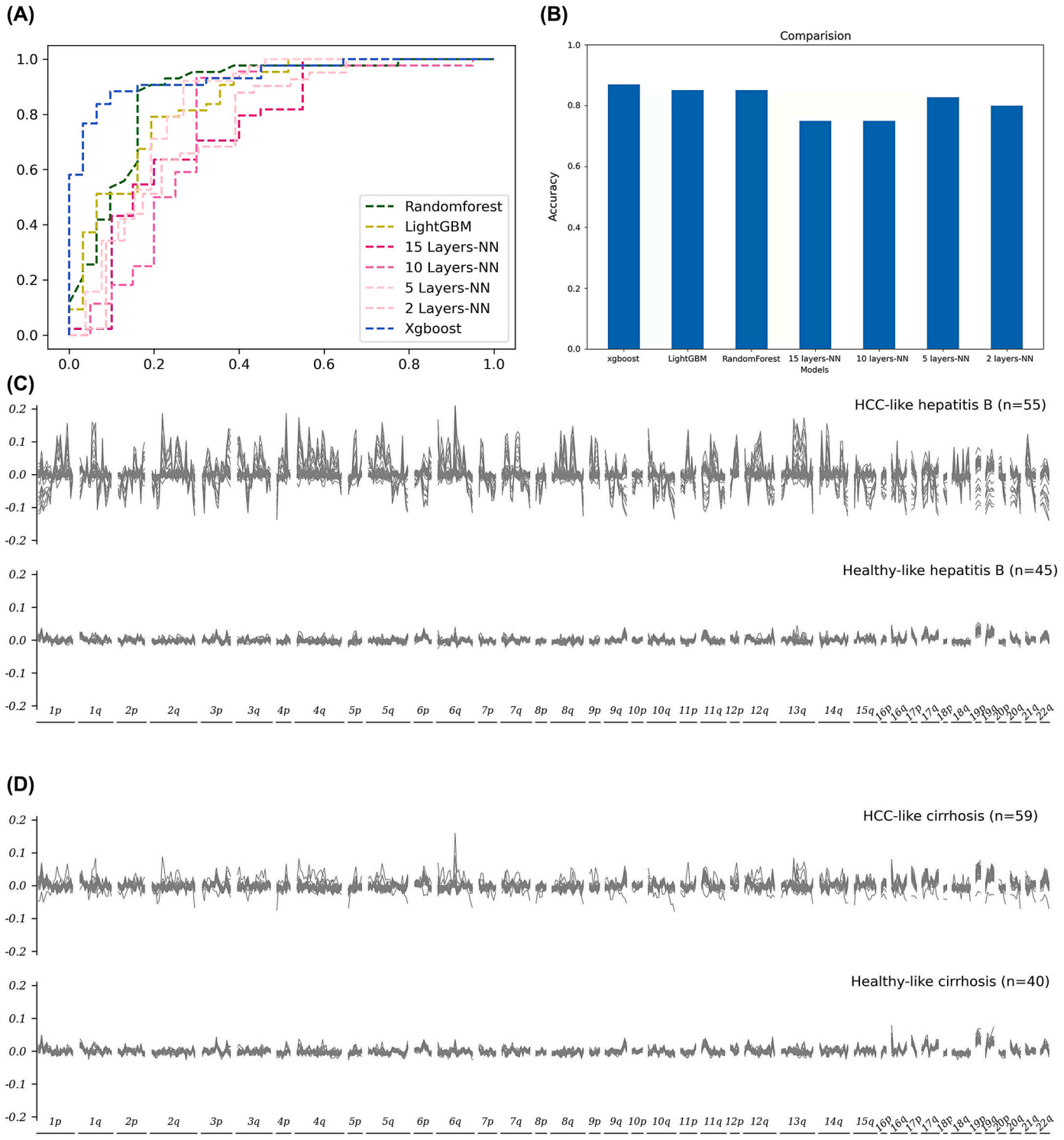

Fig. 4. Our machine learning models distinguish HCC-like hepatitis B from Healthy-like hepatitis, and HCC-like cirrhosis from Healthy-like cirrhosis based on cfDNA fragmentation profiles. (A): the AUC curves of different machine learning models. (B): The prediction accuracy of different models; (C): The cfDNA fragmentation profile of HCC-like (upper, 55 cases) and Healthy-like hepatitis B sub-groups (lower, 45 cases). (D): The cfDNA fragmentation profile of HCC-like (upper, 59 cases) and Healthy-like cirrhosis sub-groups (lower, 40 cases).

B and cirrhosis patients regarding HCC development. Thus, the XGBoost model was employed to stratify the hepatitis B and cirrhosis cohorts further, as shown in Fig. 4C and D. As a result, the XGBoost model separated the hepatitis B cohort into two sub-groups, 55 HCC-like hepatitis B patients and 45 Healthy-like hepatitis,B patients (Fig. 4C). Similarly, it separated the cirrhosis cohort into two sub-groups, 59 HCC-like cirrhosis patients and 40 Healthy-like cirrhosis patients (Fig. 4D). Using this model to discriminate HCC samples from non-HCC samples demonstrated 80.65% sensitivity at 94.12% specificity.

3.4. Treatment assessment and relapse prognosis based on cfDNA fragmentation profile

To evaluate the effectiveness of antiviral treatment, cfDNA fragmentation profiles of patients with and without antiviral treatment were compared (Fig. 5A), and the profile of the group with antiviral treatment generally appeared to have smaller fluctuations, indicating that antiviral treatment attenuated the detrimental effect of HBV to the host genome. However, the shape and location of the peaks of the two groups still looked the same. As the sharp peaks may represent HBV integration sites, this scenario suggests the profile could be used as a potential biomarker for efficacy evaluation of the antiviral treatment.

Finally, we divided the HBV-related HCC patients into three subgroups, incipient, relapsed, and non-relapsed, based on clinical records (see Supplementary Table S1) and compared their fragmentation profiles. As shown in Fig. 5B, the profile of both incipient and relapsed patients are typical of the HCC-like profile, while the profile of the non-relapsed group is typical of the Healthy-like profile.

4. Discussion

4.1. Plasma cfDNA fragmentation profile comparison between different cohorts

The cfDNA fragmentation profiles of three patients with benign liver hyperplasia and chronic HBV infection appeared very similar to those of the 140 healthy individuals (Supplementary Fig. S2) but very different from those of HBV-related HCC patients. Therefore, they were grouped into the healthy cohort. This suggests that the fragmentation profile of benign tumors is healthy-like, while the cancer-like profile only represents the malignant ones. Additionally, HCV-related HCCs also showed an abnormal fragment profile compared to the healthy cohort and a similar pattern to HBV-related HCCs (Supplementary Fig. S3). However, as this study was designed to investigate HBV carcinogenesis, the sampling of HCV-related HCCs is very limited, with only one patient in each of the 5 BCLC stages. Therefore, the relationship between fragmentation profiles of HBV- and HCV-related HCCs requires further investigation

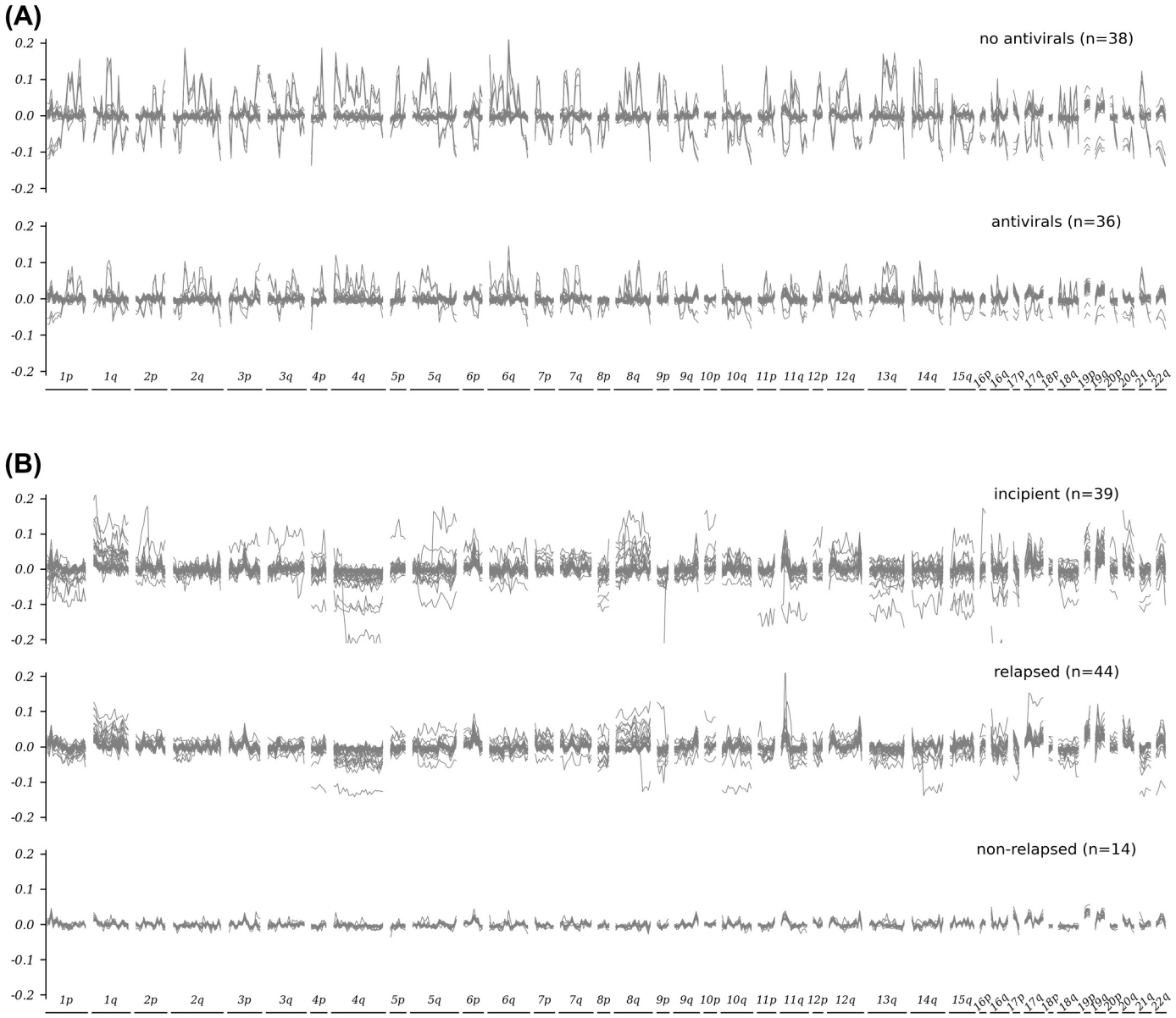

Fig. 5. Treatment assessment and relapse prognosis based on cfDNA fragmentation profiles. (A): Hepatitis B anti-viral treatment assessment (upper: no antiviral treatment, 38 cases; lower: with antivirals treatment, 36 cases) and (B): The cfDNA fragment profiles of patients with different relapse prognosis (upper: incipient, 39 cases, middle: relapse, 44 cases, lower: non-relapse, 14 cases).

with more samples. Furthermore, the comparison between the cfDNA fragmentation profile of HBV-related HCC patients with or without cirrhosis history showed no significant difference (Supplementary Fig. S4), suggesting that HBV might play a more critical role than cirrhosis in the carcinogenesis of HCC.

4.2. Prediction modeling for HCC detection

This study aimed to develop a cost-efficient method of cancer detection. Compared to previous studies [20,30–33], the sequencing data quantity per sample was purposely designed to be very small for low coverage WGS, at an average of 3.74 G (equivalent to 1.2× coverage) across all samples.

For WGS analysis, the model to discriminate HCC samples from non-HCC samples had a sensitivity of 80.65% at 94.12% specificity. However, based on the principle of machine learning methods, the performance of this model could improve as sample size increases. It should be noted that a false positive, C70 (male, 58 years old), was diagnosed with cirrhosis (HBV positive) at the time of blood draw, but was diagnosed to have primary HCC 4 months later. The fragmentation profile of C70 is slightly different from the profile of the healthy group. This suggests our model could predict the occurrence of HCC 4 months in advance, and further trials are needed to prove that with a larger data set.

4.3. Sub-grouping of hepatitis B and cirrhosis patients

We used the machine learning based model for the sub-grouping of hepatitis B and cirrhosis patients and successfully further stratified these two cohorts into two sub-groups, HCC-like and Healthy-like. Interestingly, the fragmentation profiles of the two HCC-like sub-groups are similar to that of the HCC cohort, while the profiles of the two healthy-like sub-groups are indeed like that of the healthy cohort (Fig. 4C and D). Such sub-groups indicate that the fragmentation profile from cfDNA WGS data could be used as a potential prognosis biomarker to determine whether hepatitis B and cirrhosis patients develop HCC. However, further follow-up study with larger sample size need to design and perform to validate this hypothesis.

Furthermore, as we built this model based on sophisticated machine learning methods, we investigated whether we could alternatively use a simpler statistical method, sample variance threshold, for the fragment features of each 5 Mb bin to distinguish HCC-like samples from Healthy-like samples in the hepatitis B and cirrhosis cohorts. As shown in Supplementary Fig. S5 and S6, when designating the sample variance cut-off threshold at 44% for the hepatitis B cohort and 60% for the cirrhosis cohort, the sub-grouping between HCC-like samples and Healthy-like samples by sample variance threshold is similar to the XGBoost model, with the same ratio of HCC-like samples vs. Healthy-like samples, but slightly different in sample contents (Supplementary Fig. S6). Furthermore, the cut-off sample variance for sub-grouping could be more accurate when trained using a larger sample size, indicating that the sample variance threshold method could be a simple and fast replacement to the time-consuming machine learning based methods.

Generally, epidemiology reports only discuss how many HCC patients have hepatitis B or cirrhosis history, which is mainly studied retrospectively following HCC diagnosis and thus called reverse prediction. For example, a recent report claimed that 77% of cirrhosis and 84% of HCC cases in China were caused by HBV infection [34]. In contrast, our study reported, for the first time, a prediction method following the natural cause of this disease and the progression from hepatitis or cirrhosis to HCC, namely forward-prediction. This novel forward prediction requires further study to determine its efficacy in the larger population of HBV, HCV, cirrhosis, and HCC patients. Therefore, further retrospective research with more relevant patients for data modeling, and a larger sample size, should be performed, tracking disease progression in follow-up observational studies.

4.4. Role of HBV in carcinogenesis of HCC

HBV integration has been found in the majority of HBV-associated HCCs, and studies have reported that HBV integration into specific genomic sites may give the host cells a growth advantage for clonal expansion by inducing chromosomal instability or altering the expression of host genes through *cis*-acting mechanisms [35]. Furthermore, the integrated viral DNA may allow the continuous expression of viral oncoproteins such as Hepatitis B virus-encoded X protein (HBx) and truncated preS2/S [36]. Numerous studies have shown that recurrent HBV integration occurred near actively transcribed gene coding chromosomal regions and within, or near, fragile genomic sites or repetitive regions (e.g., Alu sequences and LINEs) [36–39]. Sequence analysis has revealed integration sites in the proximity of a list of genes involved in cell survival, proliferation, metabolism, and cell cycle regulation [36–39]. Ding et al. [39] summarized the number of HBV integrations per chromosome and found the frequencies of integration are relatively higher in chromosomes 2, 3, 5, 7, 8, 10, 12, 17, and 19, with chromosome 17 having the highest. Additionally, a study reported evidence for the association between the frequency of HBV integrations and patient survival [38].

The cfDNA fragmentation profile reflects the nucleosome packaging pattern along with linear DNA molecules originally in the nucleus before cell death. In this study, we found the fragmentation profile of hepatitis B samples harbor genome-wide sharp peaks (Fig. 2), and this pattern is different from those of cirrhosis and HCC patients, possibly indicating that HBV integrations occurred in these genomic regions. However, these sharp peaks occur in all chromosomes rather than clustering in specific chromosomes, suggesting that our approach might reveal a more holistic view of the epigenomic/genomic consequences of HBV integration than previous studies [36–39]. Additionally, it is surprising that fragmentation profiles of multiple hepatitis B patients (Fig. 4C upper) display such high consonance. Therefore, we hypothesize that these shared peaks in the fragmentation profile represent recurrent HBV integration which might be shared among most hepatitis B patients. This hypothesis should be tested in a follow-up study involving a comprehensive investigation of the exact HBV integration sites in a large cohort of hepatitis B patients.

Interestingly, compared with the profile of the non-antiviral therapy group, the profile of the group with antiviral therapy clearly shows generally lower peaks. However, the shape and location of the peaks of the two groups still looked the same (Fig. 5A), indicating that antiviral therapy has weakened the effect of HBV on host genomes, but the HBV integration sites remained the same. This provides proof that our method can be used to assess the effect of antiviral therapy for hepatitis B patients. Moreover, given that the patients in these two groups shared the abnormal pattern, we hypothesize that these HBV integration sites might be recurrent, biologically functional, and able to respond to antiviral therapy. Therefore, the detailed genomic loci of these hypothesized recurrent integration sites are worthy of investigation, and their biological functions should be further studied for possible biomarkers in the development of antiviral therapies.

Finally, although all patients in the hepatitis B cohort had been confirmed to have a chronic HBV infection, five were HBsAg negative, aged 22，83，46，51, and 59, respectively. Their cfDNA fragmentation profiles (Supplementary Fig. S7) were similar to the healthy cohort, except for one, whose profile was typically HCC-like. This HCC-like patient was only 22 years old, while the ages of the four Healthy-like patients ranged from 46 to 83. The duration between HBV clearance and blood draw for this study was less than five years for the HCC-like patient; in contrast, the duration for the four Healthy-like cases was more than 15 years. Therefore, we hypothesize that it takes time for the epigenetic patterns of human genomes to recover from the alterations the HBV has caused. This duration could be longer than 5 years. However, the shortest possible time and the mechanism of patient chromatin architecture recovery after clinically identified HBV clearance requires

further study.

Financial support

This study was supported by the grants from the Shenzhen Science and Technology Project (NO.JCYJ20210324120405015, JCYJ20180302173542393, JCYJ20170817094901026), the Project of Administration of Traditional Chinese Medicine of Guangdong Province of China (20202152).

Authors' contributions

Xinfeng Sun: Conceptualization, Formal analysis, Validation, Writing – original draft. **Wenxing Feng**: Investigation, Methodology, Writing – original draft. **Pin Cui:** Conceptualization, Methodology, Project administration, Writing – original draft. **Ruyun Ruan**: Methodology, Software, Writing – original draft. **Wenfeng Ma**: Data curation, Formal analysis. **Zhiyi Han**: Data curation, Formal analysis. **Jialing Sun**: Data curation, Formal analysis. **Yuanke Pan:** Methodology, Software. **Jinxin Zhu**: Software, Visualization. **Xin Zhong**: Formal analysis, Resources. **Jing Li**: Data curation, Formal analysis. **Mengqing Ma**: Formal analysis, Resources. **Rui Hu**: Data curation, Formal analysis. **Minling Lv:** Data curation, Formal analysis. **Qi Huang**: Data curation, Formal analysis. **Wei Zhang**: Data curation, Validation. **Mingji Feng**: Formal analysis. **Xintao Zhuang**: Formal analysis. **Bingding Huang:** Supervision, Writing – review & editing. **Xiaozhou Zhou:** Funding acquisition. Supervision.

Conflict of interest

Pin Cui is the founder of Shenzhen Rapha Biotechnology Inc., Shenzhen, China. Mingji Feng and Xintao Zhuang are employees of Shenzhen Rapha Biotechnology Inc. The remaining authors declare no conflicts of interest.

Data availability

The WGS data in this study are available from the corresponding authors upon request.

Acknowledgments

We would like to thank the patients who gave their consent to present the data in this study and the research staff involved. We thank Dr. Xin Wang for his helpful discussion and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110502.

References

[1] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (3) (2021) 209–249.
[2] C. Allemani, et al., Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries, Lancet 391 (10125) (2018) 1023–1075.
[3] J.D. Yang, et al., A global view of hepatocellular carcinoma: trends, risk, prevention and management, Nat. Rev. Gastroenterol. Hepatol. 16 (10) (2019) 589–604.
[4] A.G. Singal, et al., Benefits and harms of hepatocellular carcinoma surveillance in a prospective cohort of patients with cirrhosis, Clin. Gastroenterol. Hepatol. 19 (9) (2021) 1925–1932.e1.
[5] EASL, EASL clinical practice guidelines: management of hepatocellular carcinoma, J. Hepatol. 69 (1) (2018) 182–236.
[6] J.C. Nault, A. Villanueva, Biomarkers for hepatobiliary cancers, Hepatology 73 (2021) 115–127.
[7] M. Patel, et al., Hepatocellular carcinoma: diagnostics and screening, J. Eval. Clin. Pract. 18 (2) (2012) 335–342.
[8] C. Aubé, et al., EASL and AASLD recommendations for the diagnosis of HCC to the test of daily practice, Liver Int. 37 (10) (2017) 1515–1525.
[9] V.M. Runge, Safety of the gadolinium-based contrast agents for magnetic resonance imaging, focusing in part on their accumulation in the brain and especially the dentate nucleus, Investig. Radiol. 51 (5) (2016) 273–279.
[10] P. Tabrizian, et al., Recurrence of hepatocellular cancer after resection: patterns, treatments, and prognosis, Ann. Surg. 261 (5) (2015) 947–955.
[11] Y. Xie, Hepatitis B virus-associated hepatocellular carcinoma, Adv. Exp. Med. Biol. 1018 (2017) 11–21.
[12] J. Ding, H. Wang, Multiple interactive factors in hepatocarcinogenesis, Cancer Lett. 346 (1) (2014) 17–23.
[13] L. Liu, Clinical features of hepatocellular carcinoma with hepatitis B virus among patients on Nucleos(t) ide analog therapy, Infect. Agent Cancer 15 (2020) 8.
[14] V.K. Chaturvedi, et al., Molecular mechanistic insight of hepatitis B virus mediated hepatocellular carcinoma, Microb. Pathog. 128 (2019) 184–194.
[15] D.W. Cescon, et al., Circulating tumor DNA and liquid biopsy in oncology, Nat. Can. 1 (3) (2020) 276–290.
[16] D.S.C. Han, et al., The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB, Am. J. Hum. Genet. 106 (2) (2020) 202–214.
[17] R.W.K. Chiu, et al., Cell-free DNA fragmentomics: the new "omics" on the block, Clin. Chem. 66 (12) (2020) 1480–1484.
[18] M. Chen, H. Zhao, Next-generation sequencing in liquid biopsy: cancer screening and early detection, Hum. Genom. 13 (2019).
[19] A.P. Koval, et al., The detection of cancer epigenetic traces in cell-free DNA, Front. Oncol. 11 (2021), 662094.
[20] S. Cristiano, et al., Genome-wide cell-free DNA fragmentation in patients with cancer, Nature 570 (7761) (2019) 385–389.
[21] Chinese Society of Hepatology, C.M.A, Chinese guidelines on the management of liver cirrhosis, Zhonghua Gan Zang Bing Za Zhi 27 (11) (2019) 846–865.
[22] Department of Medical Administration, N.H. and C. Health Commission of the People's Republic of, Guidelines for diagnosis and treatment of primary liver cancer in China (2019 edition), Zhonghua Gan Zang Bing Za Zhi 28 (2) (2020) 112–128.
[23] S. Chen, et al., fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890.
[24] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[25] P. Danecek, et al., Twelve years of SAMtools and BCFtools, Gigascience 10 (2) (2021) giab008.
[26] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[27] G. Ke, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Proces. Syst. 30 (2017) 3146–3154.
[28] 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.
[29] F. Pedregosa, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[30] X. Zhang, et al., Ultra-sensitive and affordable assay for early detection of primary liver cancer using plasma cfDNA fragmentomics, Hepatology 76 (2) (2022) 317–329.
[31] N. Liang, et al., Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning, Nat. Biomed. Eng. 5 (6) (2021) 586–599.
[32] X. Chen, et al., Prognostic significance of blood-based multi-cancer detection in plasma cell-free DNA, Clin. Cancer Res. 27 (15) (2021) 4221–4229.
[33] E.A. Klein, et al., Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set, Ann. Oncol. 32 (9) (2021) 1167–1177.
[34] G. Wang, Z. Duan, Guidelines for prevention and treatment of chronic hepatitis B, J. Clin. Transl. Hepatol. 9 (5) (2021) 769–791.
[35] C. Bréchot, et al., Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC), Semin. Cancer Biol. 10 (3) (2000) 211–231.
[36] Z. Jiang, et al., The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients, Genome Res. 22 (4) (2012) 593–601.
[37] Y. Murakami, et al., Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas, Gut 54 (8) (2005) 1162–1168.
[38] W.K. Sung, et al., Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma, Nat. Genet. 44 (7) (2012) 765–769.
[39] D. Ding, et al., Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach, PLoS Genet. 8 (12) (2012), e1003065.