Multi-omics integrated circulating cell-free DNA genomic signatures enhanced the diagnostic performance of early-stage lung cancer and postoperative minimal residual disease

Yun Li,[a,g] Guanchao Jiang,[a,g] Wendy Wu,[b,c,g] Hao Yang,[b,g] Yichen Jin,[d,g] Manqi Wu,[a] Wenjie Liu,[b] Airong Yang,[b] Olga Chervova,[e] Sujie Zhang,[a] Lu Zheng,[b] Xueying Zhang,[b] Fengxia Du,[b] Nnennaya Kanu,[e,f] Lin Wu,[b,c,**] Fan Yang,[a,***] Jun Wang,[a,h] and Kezhong Chen[a,i,*]

[a]Thoracic Oncology Institute/Department of Thoracic Surgery, Peking University People's Hospital, Peking University, Beijing, China
[b]Berry Oncology Corporation, Beijing, China
[c]Fujian Key Laboratory of Advanced Technology for Cancer Screening and Early Diagnosis, Fuzhou, China
[d]Department of Clinical Sciences, Peking University Health Science Center, Beijing, China
[e]University College London Cancer Institute, University College London, 72 Huntley St, London, WC1E 6DD, UK
[f]Cancer Research UK Lung Cancer Centre of Excellence, University College London, 72 Huntley St, London, WC1E 6DD, UK

eBioMedicine
2023;91: 104553

Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104553

Summary

Background Liquid biopsy is a promising non-invasive alternative for cancer screening and minimal residual disease (MRD) detection, although there are some concerns regarding its clinical applications. We aimed to develop an accurate detection platform based on liquid biopsy for both cancer screening and MRD detection in patients with lung cancer (LC), which is also applicable to clinical use.

Methods We applied a modified whole-genome sequencing (WGS) -based High-performance Infrastructure For MultIomics (HIFI) method for LC screening and postoperative MRD detection by combining the hyper-co-methylated read approach and the circulating single-molecule amplification and resequencing technology (cSMART2.0).

Findings For early screening of LC, the LC score model was constructed using the support vector machine, which showed sensitivity (51.8%) at high specificity (96.3%) and achieved an AUC of 0.912 in the validation set prospectively enrolled from multiple centers. The screening model achieved detection efficiency with an AUC of 0.906 in patients with lung adenocarcinoma and outperformed other clinical models in solid nodule cohort. When applied the HIFI model to real social population, a negative predictive value (NPV) of 99.92% was achieved in Chinese population. Additionally, the MRD detection rate improved significantly by combining results from WGS and cSMART2.0, with sensitivity of 73.7% at specificity of 97.3%.

Interpretation In conclusion, the HIFI method is promising for diagnosis and postoperative monitoring of LC.

Funding This study was supported by CAMS Innovation Fund for Medical Sciences, Chinese Academy of Medical Sciences, National Natural Science Foundation of China, Beijing Natural Science Foundation and Peking University People's Hospital.

Abbreviations: AUC, Area under the curve; cfDNA, Cell-free DNA; CHIP, Clonal hematopoiesis of indeterminate potential; cSMART, Circulating single-molecule amplification and resequencing technology; ctDNA, Circulating tumor DNA; CTR, Consolidation tumor ratio; DMR, Differentially methylated region; HIFI, High-performance infrastructure for multiomics; LDCT, Low-dose computed tomography; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; mGGO, Mixed ground-glass opacity; MHD, Methylated haplotype diversity; MRD, Minimal residual disease; NPV, Negative predictive value; NSCLC, Non-small cell lung cancer; pGGO, Pure ground-class opacity; RFS, Recurrence-free survival; WGS, Whole genome sequencing

*Corresponding author.
**Corresponding author.
***Corresponding author.

E-mail addresses: chenkezhong@pkuph.edu.cn (K. Chen), wulin@berryoncology.com (L. Wu), yangfan@pkuph.edu.cn (F. Yang).

[g]These authors contributed equally to this work.
[h]Senior author.
[i]Lead contact.

Keywords: Liquid biopsy; Whole-genome sequencing; Lung cancer; Diagnostic model; Prognostic model

Research in context

Evidence before this study

Lung cancer is a severe disease jeopardizing people's health. Early diagnosis and postoperative MRD monitoring are crucial procedures to elevate the overall survival for patients with lung cancer. Liquid biopsy, a non-invasive and sensitive detection method, has shown great potential in both fields of lung cancer. However, a low density of circulating tumor DNA in peripheral blood and interference from normal cells block its way into clinical utility. Economic expense and turnaround time are also critical issues to consider.

Added value of this study

We developed a low-pass WGS-based HIFI platform for both early screening and MRD monitoring, which combined multi-dimensional features with machine learning methods. HIFI method exhibited higher accuracy than anterior classic methods in both fields. The design of HIFI platform also simplified detection procedure for convenience in real applications.

Implications of all the available evidence

HIFI method was built for whole diagnosis-therapy process, in pursuit of convenience and low expense when still keeping its performance at a relative high level. In early screening, it can perform as an auxiliary method for LDCT to eliminate false positive rate. While in MRD detection, HIFI showed great sensitivity in discerning patients with high risk of recurrence.

Introduction

Lung cancer (LC) is the leading cause of cancer-related mortality worldwide,[1] among which non-small cell lung cancer (NSCLC) accounts for ~85% of the cases.[2] The prognosis of LC is significantly correlated with the stage at which it is diagnosed.[3] Therefore, early screening is an effective strategy to reduce the mortality of LC patients. Low-dose computed tomography (LDCT) is the most extensively recommended LC screening method, which can reduce mortality by 20% among high-risk patients.[4] Although LDCT plays a vital role in lung cancer early screening, an unsatisfactory false-positive rate of 96.4% limits its potential in early screening. Surgery is the major therapy for patients in early stages. Although radical surgery cured most early-stage NSCLC (stages I-IIIA) patients, approximately 10–50% of them experienced relapse after surgery,[3] which was probably caused by indiscernible residual cancer cells. Therefore, postoperative surveillance is essential for early identification of patients with high risk of recurrence and is indispensable for the administration of necessary adjuvant therapy. Thus, methods with high accuracy are required to distinguish malignant tumors from benign pulmonary nodules (PNs) and to identify patients with high recurrence risk at an early timepoint.

In recent years, circulating tumor DNA (ctDNA), which refers to the fraction of the cell-free DNA in the peripheral blood that is shed by tumor cells, has emerged as a promising liquid biopsy biomarker for non-invasive cancer screening and post-treatment surveillance.[5,6] We had previously reported that ctDNA was significantly more sensitive to the early diagnosis of LC than serum protein markers[7] and have shown that ctDNA detection could test postoperative MRD which is significantly associated with the poor prognosis of LC patients.[8] With continuous improvements in technology, the detection accuracy of ctDNA keeps improving.[9,10]

However, there are still several concerns regarding clinical applications of ctDNA. Due to the cell-free DNA (cfDNA) abundance of non-cancerous origin coupled with its rapid metabolic rate, the concentration of ctDNA is extremely low, especially in operable early-stage cancers.[10] Very few methods meet the required sensitivity currently; thus, the detection rate of stage I LC, especially for adenocarcinoma, is unsatisfactory even with higher sequencing depth.[11] Similar issues exist in postoperative MRD detection. Moreover, the clonal hematopoiesis of indeterminate potential (CHIP), which increases with age, is another common confounding factor interfering with ctDNA mutation detection.[12] To overcome these technical barriers, many studies enhance sequencing depth to increase the detection rate of ctDNA mutation but the improvement is just passable.[13] On the other hand, WGS focuses on the genomics breadth instead of depth, which captures abnormal signals from another dimension.[14] Our previous retrospective study suggested that incorporating multi-omics features might provide a more comprehensive view of the patients, systematically reducing investigational bias and facilitating more accurate cancer screening and MRD detection strategies.[15] Therefore, a WGS-based approach integrating multiple cancer genomic features may provide a solution to the current challenges.

In this study, we investigated a low-pass WGS technology to acquire diverse genetic variation signatures of cfDNA and developed a hyper-accurate method for NSCLC detection, characterized by cfDNA genomic end motifs,[16,17] fragmentation,[18] and Bincount.[19,20] This HIFI method was further applied to both LC screening and postoperative MRD detection, combined with a hyper-co-methylated read approach[21] and modified cSMART

2.0.[22] Consequently, the HIFI method showed favorable results for the diagnosis and treatment of LC.

Methods

Study design and recruited participants

We performed a prospective observational study involving multiple centers from Beijing and Fuzhou. A total of 670 participants limited to NSCLC were enrolled for constructing the lung cancer model and performing validation tests during 2019–2022 (Fig. 1a). A discovery cohort consisting of 145 NSCLC patients and 163 healthy control (HC) or participants with benign PNs was used for lung cancer early screening model construction. Then, the validation effect of this model was then tested on an independent validation cohort (N = 306) (Table S1). Sex was self-reported by participants and we assigned similar ratios of sex in both cohorts. Detailed clinical information of these participants is summarized in Table S2. For early screening tests of lung cancer, 358 participants with malignant or benign pulmonary nodules were taken from Peking University People's Hospital and Beijing HAIDIAN Hospital with pathological analysis as gold standard and 297 individuals from different medical examination centers in Beijing and Fuzhou as healthy control cases verified by chest CT scan and a one-year follow-up. Sixteen samples in the healthy control set were excluded due to recent surgical resection and prior diagnosis of cancer. Twenty-five individuals were excluded from the cohort due to the poor sample quality or failed raw data QC (Fig. 1a). The modified HIFI (**H**igh-performance **I**nfrastructure **F**or Mult**I**omics) method[21] was used to distinguish lung cancer patients from healthy/benign nodule individuals using machine learning methods (Fig. 1b). To test the monitoring ability of the lung cancer screening model in the postoperative treatment effect of the patients, patients (N = 58) who underwent surgery and conducted MRD detection experiments were enrolled. Postoperative blood sample was drawn on the third day after surgery before applying any adjuvant therapy. The low-pass WGS sequencing method at a depth of 2.5× and cSMART2.0 (the captured single-molecule amplification and resequencing technology) assay were performed for MRD detection[21,22] (Fig. 1b right).

DNA isolation and quality inspection

Peripheral blood was collected from each participant and stored in Streck Cell-Free DNA BCT tubes (STRECK, USA). Plasma was purified by centrifugation at 800g for 10 min at 4 °C. The plasma was centrifuged again at 18,000 g for 10 min at room temperature to remove any remaining cellular debris and stored at −80 °C until DNA extraction. DNA was isolated from the plasma using the Qiagen Circulating Nucleic Acid Kit (Qiagen GmbH) and eluted in LoBind tubes (Eppendorf AG). The concentration and quality of cfDNA were assessed using the Bioanalyzer 2100 (Agilent Technologies). Isolated cfDNA was stored at −20 °C until library preparation. The genomic DNA of fresh frozen tissue was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions and was purified by AxyPrep™ Mag Tissue-Blood genomic DNA Kit (Axygen, America). The quantification of DNA was measured by the Qubit dsDNA HS Assay Kit (Life Technologies, Eugene, OR).

Low-pass WGS library construction

DNA (5 ng) of each sample was prepared for constructing the WGS library. DNA samples were subjected to end repair/dA-tailing (5× ER/A-Tailing Enzyme Mix). The dTTP-tailed adapters were ligated to both ends of the repaired/dA-tailed DNA fragments using WGS ligase and amplified by PCR. The PCR products from each library were subsequently purified with an Agencourt AMPure XP PCR Purification Kit (Beckman Coulter, Brea, CA, USA), and each DNA library was quantified by the KAPA Library Quantification Kit (Kapa Biosystems, USA); the size was confirmed using a Bioanalyzer 2100 (Agilent, USA). Equal amounts of sequencing libraries were pooled and analyzed using the Illumina NovaSeq 6000 platform. The FASTQ files were processed using the Cutadapt software (https://github.com/marcelm/cutadapt/) to remove the adaptor and end sequence along with sequences below 50 bp. The clean data were aligned to the human reference genome GRCh37 using bwa-mem (https://github.com/lh3/bwa). Duplicate reads were marked using Sambamba (https://github.com/biod/sambamba/). The mapping rate, duplicate rate, and genome coverage were calculated using samtools. Reads with a mapping rate above 90%, a duplicate rate below 25%, and coverage above 50% passed the quality control. Low-quality reads, marked duplicates, and sequences with no perfect match between reads 1 and 2 were subsequently removed after being filtered further by samtools.

Genome characteristics

Three genomic features (Fragment, Motif, and Bincount) were determined through low-pass WGS detection to distinguish lung cancer patients from healthy/benign individuals. The detailed filtering process of different genome characteristics is described below.

Motif

The motif mode was built using Pysam (https://pysam.readthedocs.io/en/latest/) to calculate the percentage of 4-mer 5′end. We identified 256 different types of 4-mer 5′end motifs and calculated their percentages without considering chromosome Y and unidentifiable bases.[16] The motif types were filtered as follows: the Wilcoxon rank-sum test was performed to filter out the features of the groups that were significantly different ($p \leq 0.05$); Random Forest (RF) was used for reducing the

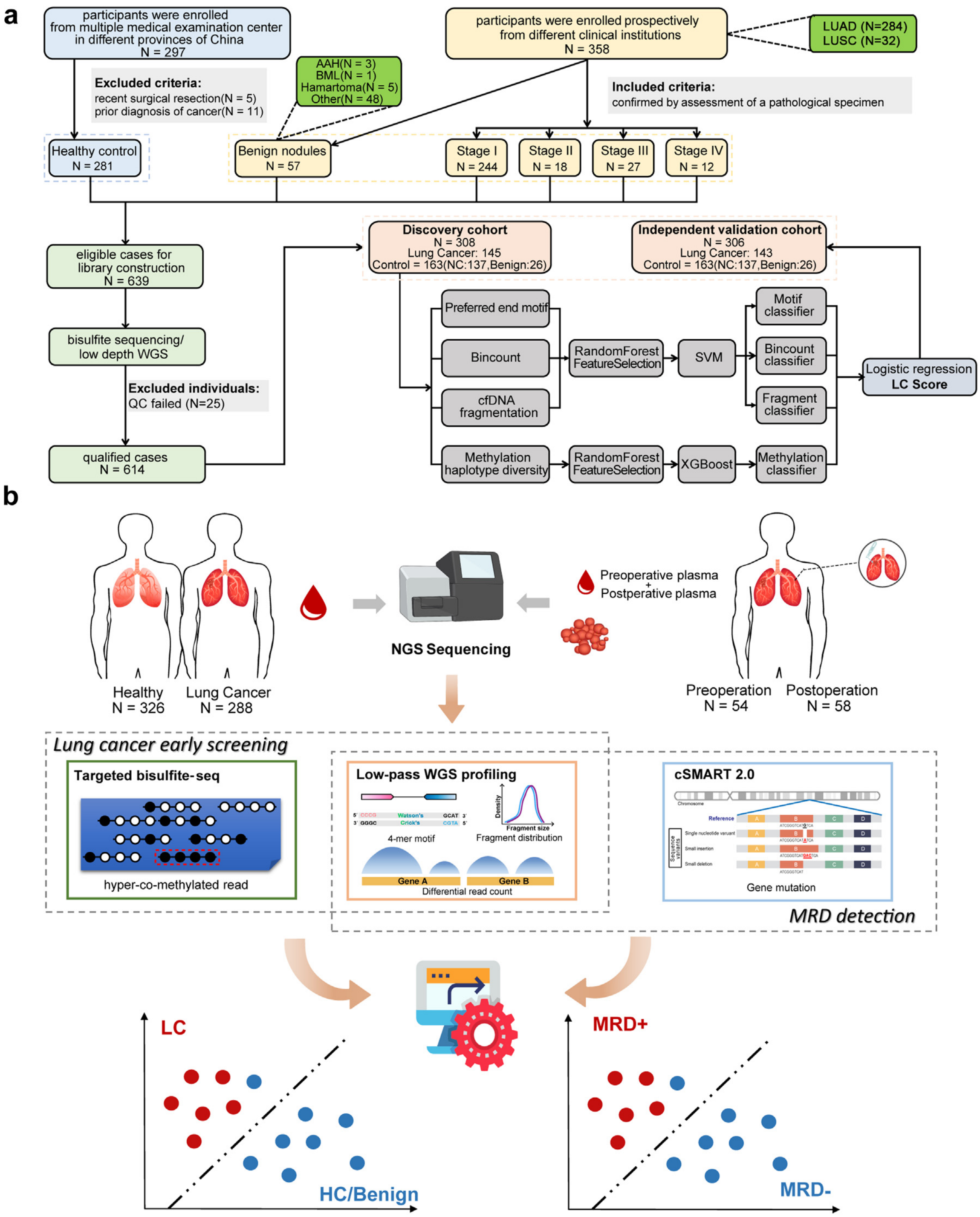

Fig. 1: **Overview of the study design and the cohort of the participants. (a)** The participants of the lung cancer early screening cohort. A discovery cohort consisting of 145 lung cancer patients and 163 healthy control (HC) or participants with benign PNs was used for lung cancer early screening model construction. The validation effect of this model was tested on an independent validation cohort (N = 306). **(b)** Study design for the early detection and MRD monitoring of lung cancer. For the early screening of lung cancer, hyper-co-methylation signatures and the genomic features detected by low-pass WGS were used for lung cancer screening model construction. And genomic features detected by low-pass WGS features were combined with the ctDNA mutation results detected by cSMART2.0 methodology for MRD monitoring. AAH, atypical adenomatous hyperplasia; BML, benign metastasizing leiomyoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

dimensionality of the 133 motifs left in the previous step; finally, 73 motif types were left for further analysis.

Fragment

The fragmentation model was built by Pysam to calculate the length of the insertion fragments and the ratio of short/long fragments in different regions.[23] The entire genome was first excluded from the X, Y, and MT chromosomes and then divided into the continuous windows of 2M bases. The proportion of fragments in each window and the ratio of shorter versus longer fragments were calculated and produced 11,616 features. 2× XP beads were used to recover fragments above 100 bp during library construction. Wilcoxon rank-sum test was performed to filter out the features of the groups that were significantly different ($p \leq 0.05$). Then, Random Forest was used for reducing the dimensions with 1382 features left in the previous step, based on the importance of the features, and finally, 357 areas were retained.

Bincount

The entire genome was first excluded from the X, Y, and MT chromosomes and then divided into the consecutive windows of 2M bases. The Feature Counts software (http://subread.sourceforge.net/) was used to count the number of fragments in each window.[19] To eliminate the bias caused by the size of the library, the formula used for normalization was:

$$BinCount_i = log\left(\frac{Fragment_i * 10^9}{TotalMappedFragments * WindowLength_i} + 1, 10\right)$$

Here, “i” represents the ith window; “Fragment$_i$” represents the number of fragments compared to the ith window; “TotalMappedFragments” represents the total number of fragments of the sample excluding X, Y, and MT chromosomes; “WindowLength$_i$” represents the length of the ith window (2×10^5 in this case). The Wilcoxon rank-sum test was performed to filter out the features of the groups that were significantly different ($p \leq 0.05$). Random Forest was used for reducing the dimensionality of 603 features left in the previous step; finally, 336 areas were retained.

Bisulfite-sequencing library construction

Approximately 5–20 ng of cfDNA was taken from each participant for constructing the bisulfite-sequencing library. The DNA samples were subjected to bisulfite conversion using EZ-96 DNA methylation-lightning MagPrep (Zymo Research, D5047) following the manufacturer’s protocol. Then, the converted cfDNA was ligated with adaptors and amplified using the Hieff NGS® Methyl-seq ssDNA Library Prep Kit for Illumina (Yeasen) following the manufacturer’s protocol. The libraries were hybridized and captured using the Twist Fast Hybridization and Wash Kit and a Twist panel with a customized design. After being purified by Agencourt AMPure XP beads (Beckman Coulter, USA), the libraries were quantified by the KAPA Library Quantification Kit (Kapa Biosystems, USA); finally, the size was confirmed using a Bioanalyzer 2100 (Agilent, USA). Equal amounts of sequencing libraries were pooled. Then, the libraries with an average coverage of 200× were analyzed using the Illumina NovaSeq 6000 platform by performing 2 × 150 bp paired-end sequencing.

Characterizing bisulfite markers

The FASTQ files were processed using Cutadapt (https://github.com/marcelm/cutadapt/) to acquire clean data by removing the adaptor or poly-tail and end sequence, along with sequences below 50 bp. The raw data of methylation were aligned to the human reference hg19 from the 1000 Genomes Phase 3 resources with decoy and patch sequences using the BSMAP (https://code.google.com/archive/p/bsmap/) alignment module. The mapped reads were split by top/bottom strand using bamtools (https://github.com/pezmaster31/bamtools) to remove duplicates, and then these strands were re-merged. The BSMAP aligner included a tag in the BAM file (ZS), which indicated the top/bottom strand and the forward/reverse read. This tag was used to split the BAM file containing the mapped reads. Several paired-end reads overlapped when 150 bp paired-end reads were used. As there was a chance of incorrect counting of the converted/non-converted Cs in the overlapping regions twice while evaluating percent methylation, bamUtil clipOverlap (https://github.com/statgen/bamUtil) was used to prevent such a bias, and BisMark (https://github.com/FelixKrueger/Bismark) was used to determine read-level cytosine methylation states. Each captured area was considered as a unit to count the methylation status of each CpG site of all reads in the area.[24]

We defined methylated haplotype diversity (MHD) for each candidate region which was used to calculate the fraction of haplotype of all possible lengths that were fully methylated in each read.

$$MHD = \frac{1}{N} x \sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{m} h_{ij}}{\sum_{k=1}^{L} k_i}\right)$$

Here, “N” indicates the total number of reads in the candidate region; “m” indicates the total number of haplotypes in the *i*th read, which contains continuous methylated CpGs; “h_{ij}” indicates the *j*th haplotype in the *i*th read; “L” indicates the total number of CpGs (including methylated and un-methylated loci) in the *i*th read.

Co-methylation model construction

The reference data of genomic methylation characteristics of lung cancer were obtained from TCGA database, in-house MeDIP cfDNA, GEO RRBS/WGBS, and other published data.[25–27] It generated 757 regions of interest (ROIs) containing ∼30K CpG sites for lung cancer-specific methylation characteristics of cfDNA (Fig. S1). Then, these methylation CpG sites were clustered in the development set (145 LC, 163 benign/HC individuals) for differentially methylated region (DMR) identification, and 378 were verified. The reads with at least three methylated sites within a sliding window of five CpGs or at least two methylated sites within a sliding window of three or four CpGs were extracted for co-methylated reads.

This step was repeated continuously along all the reads to ensure that consistently enhanced methylation signals were obtained. All obtained CpG clusters were regarded as the methylated haplotypes of the whole training cohort. The lung cancer-specific methylated haplotypes were selected using Fisher's exact test and were significantly enriched in LC individuals compared to HC/benign individuals (p < 5.00E-02). Each of the development sets generated a vector according to the specific methylated haplotypes, and a co-methylation classifier was constructed when applying these vectors to the extreme gradient boosting (XGBoost) classifier, after comparison among adaptive boosting (AdaBoost), Gaussian process, K-nearest neighbor (KNN), RF, support vector machine (SVM) and XGBoost (Fig. S2a).

The modified HIFI method for lung cancer screening

In our previous investigation, the HIFI method showed a strong diagnostic value in differentiating hepatocellular carcinoma (HCC) from liver cirrhosis.[21] Here, we modified the HIFI method for the early detection of lung cancer, which integrated four genomic features: motif, fragment, Bincount, and co-methylation model. A co-methylation model was initially constructed by machine learning to distinguish LC individuals from HC/benign individuals in the training cohort. Three distinctive genomic features (Motif/Fragment/Bincount) were detected through low-pass WGS technology and selected for constructing the lung cancer detection model. To obtain the best diagnostic performance, we constructed logistic regression models using the predictive score of the four individual models as input features to integrate the outcome of each model based on training and validation datasets to establish the signal for detecting cancer.

The LC score model construction

For lung cancer detection, a machine learning method of SVM was selected as classifier with the highest area under the curve (AUC) of 0.945 among AdaBoost, Gaussian process, KNN, RF, SVM and XGBoost (Fig. S2b). SVM was implemented to construct individual genomic feature-based models[28] based on three parameters: (1) C: Penalty coefficient; (2) Kernel function; and (3) Gamma. For the training dataset, 10-fold cross-validation was used to determine the best combination of parameters. The cutoff value was set at the point with the best diagnostic accuracy in the validation set. To obtain the best diagnostic model,[29] we constructed logistic regression models using the predictive score of the four individual models as input features to integrate the outcome of each model based on the discovery and validation datasets. The probability of lung cancer in a patient was calculated as:

$$Pr\,(LC) = exp\,(Z) \big/ \big(1 + exp\,(Z)\big)$$

$$where,\ Z = (2.98 \times PBincount) + \big(1.59 \times PMotif\big) + (3.27 \times P5mC) + \big(4.78 \times PFragment\big) - 5.75$$

MRD detection and postoperative monitoring

Genomic DNA was isolated from frozen fresh tumor samples and quantified using the Bioanalyzer 2100 (Agilent Technologies). WGS was performed using the purified DNA. The ctDNA of the post-treatment plasma samples was prepared as previously described, and plasma DNA was prepared for constructing the WGS library. We conducted a series of experiments on the input amount and sequencing depth of ctDNA, the results showed that better detection sensitivity and accuracy can be obtained when the input amount was 30 ng and the sequencing depth was 35,000×. DNA sequencing was performed by a sensitive assay termed cSMART2.0 (the captured single-molecule amplification and resequencing technology) (Berry Oncology, Fuzhou, China) with a 218 hotspot genes panel (Table S6). The detailed steps of cSMART2.0 refer to the published paper.[22,30,31]

Machine learning

Each classifier (SVM/Random Forest/Logistic Regression) was implemented in python with the scikit-learn (1.0.1, https://scikit-learn.org/stable/) library. Using the processed input data to quantify the group data from two sets, the machine learning classification model was constructed by random forest, and the optimal threshold was evaluated by the receiver operator characteristic (ROC) curve.

Sample size estimation

The required sample size for each non-diseased and diseased group was defined by:

$$N = \frac{Z_{\frac{\alpha}{2}}^{2} V(\widehat{AUC})}{d^{2}}$$

Here, α is calculated as follows, with ϕ^{-1} considered as the inverse of standard cumulative normal distribution (assuming the pre-determined value of AUC = 0.912)[32]:

$$\alpha = \phi^{-1}(0.912)\times 1.414$$

$V(\widehat{AUC})$ can be evaluated as follows:

$$V(\widehat{AUC}) = \left(0.0099 \times e^{\frac{-\alpha^2}{2}}\right)\times\left(6\alpha^2 + 16\right)$$

To estimate AUC with 95% confidence, the degree of precision of estimate was 0.05, and the required sample size was obtained by inserting $V(\widehat{AUC})$ and d = 0.05 in Eq (1). Thus, the required sample size for each group in the validation cohort is 95.

AUC confidence interval

To confirm the obtained results of the LC score performance obtained via bootstrapping, we explicitly calculated the margin of error for AUC, using Hanley and McNeil (1982) method.[33]

In brief, $(1-\alpha)\cdot 100\%$ confidence interval for AUC could be derived from standard normal distribution as $AUC \pm z_{\alpha/2}\cdot SE(AUC)$, where $z_{\alpha/2}$ denotes a quantile of normal distribution, $SE(AUC)$ is standard error, which could be calculated as:

$$SE(AUC) =$$

$$\sqrt{\frac{AUC(1-AUC)+(n_1-1)\left(q_1-AUC^2\right)+(n_2-1)\left(q_2-AUC^2\right)}{n_1 n_2}}cr,$$

where n_1 and n_2 are simple sizes for control and case groups respectively, and q_1 and q_2 are defined as:

$$q_1 = \frac{AUC}{2-AUC},\ q_1 = \frac{2\cdot AUC^2}{1+AUC}.$$

Substituting our data into theses formulae results in $SE(AUC) = 0.0168969$, then for $\alpha = 0.05$ $z_{\alpha/2} \approx 1.96$ and the confidence interval would be:

$$[AUC - z_{\alpha/2}\cdot SE(AUC), AUC + z_{\alpha/2}\cdot SE(AUC)]$$

$$= [0.8788821, 0.9451179],$$

which is very close to the results obtained by the bootstrapping i.e. [0.880, 0.942].

Evaluating the benefit to the population

It is a good practice to optimize the discovered biomarkers for sensitivity or specificity based on their intended clinical application. For tests where a positive biomarker result leads to an action, while a negative biomarker result is associated with the standard of care for the population, the formula reads:

$$\frac{sensitivity}{1-specificity} \geq \frac{1-prevalence}{prevalence}\cdot\frac{harm}{benefit}$$

Here, harm/benefit indicates the ratio of the net harm of a false-positive test result to the net benefit of a true-positive test result.[34] In our case, with 0.15% prevalence, sensitivity 0.518 and specificity of 0.963, we found that $\frac{harm}{benefit} \leq 0.02103155$, or about 1:48, which could be interpreted as unnecessary clinical actions (e.g. lung X-ray or CT scan) on 48 of control subjects testing positive should be tolerated to benefit one case subject testing positive.

Statistics

After model construction, the optimal threshold was evaluated by the ROC curve. The Youden index was used to select the optimal threshold for improving the sensitivity and specificity of tumor detection.

Ethics statement

All procedures were approved by the Medical Ethics Committee (2019PHB058-02) of the Peking University People's Hospital and registered on Clinical ClinicalTrails.gov (NCT04558255). Informed consent was obtained from all enrolled participants prior to participation.

Data deposition and materials sharing

The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0003151.

Role of the funding source

The funder took no role in study design, data collection, data analyses, interpretation, or writing of report.

Results

The overview of the design and the participants

In this study, the HIFI method[21] was used to distinguish lung cancer patients from healthy/benign nodule individuals and filter out postoperative patients with high risk of recurrence, mainly based on three genomic features (motif/fragment/Bincount) which had shown their privilege in previous studies.[17–20] In order to achieve better detection efficiency, HIFI score was further integrated with co-methylation features and ctDNA mutations respectively, in order to construct the LC screening model and MRD monitoring model (Fig. 1b). Patient demographics and clinical information are shown in Table S1. A total of 614 individuals from multiple centers were included according to our inclusion and exclusion criteria. Initially, we performed detection on the discovery cohort (training set, N = 308) with 145 NSCLC patients and 163 healthy control (HC) or participants with benign PNs (Fig. 1a) (Table S2). The LC score, lung cancer screening model, was constructed by co-methylation patterns based on the HIFI method platform with machine learning classifiers. The validation cohort (test set, N = 306) was obtained from the subsequent prospective samples, which comprised 143 NSCLC patients and 163 HC/Benign. Specifically, stage I cancer individuals comprised 80.7% in the discovery set, and 81.1% in a single-blind cohort of validation set (Table S1). The HIFI method was also used to monitor patients after surgery and evaluate their postoperative effects status using WGS combined with cSMART2.0 methodology in a subset cohort containing 58 postoperative individuals. In sum, we applied the HIFI method in the whole diagnosis-treatment procedure of LC.

Genomic alteration features of lung cancer patients

Ultra-low-pass whole-genome sequencing (average sequence depth is 2.5×) of plasma cfDNA was used to distinguish genomic features of lung cancer patients from healthy control, which covers plasma DNA end motifs, fragment, and Bincount. We selected these three vectors to construct models for the following reasons. In the generation of plasma DNA, the cleavage of DNA was not random, but with preferred tendencies. Such preferred ends of several nucleotides called motif were observed in several cancers, including hepatocellular carcinoma and lung cancer.[17] It was believed genome alterations had an influence on global aberrations on DNA endonucleases and type-specific chromatin accessibility also contributed to DNA fragmentation, both of which explained the distinctiveness of the landscape of plasma DNA motifs.[17,35] Fragmentation was a sensitive vector that depicted the dimension of fracture size in plasma DNA. Genomic profiles in patients with cancers were more cluttered than healthy people, while the same scenario applied to fragmentation profiles, for the fragment's strong correlation to nucleosomal DNA and nucleosome distances. Multi-centered results had validated the practicability of fragmentation profiles in pan-cancers including lung cancer.[18] Bincount was another indicator to present the characters of plasma DNA fragments which focused on quantifying tumor genomic instability, exhibiting great detection ability in colorectal cancer and bladder cancer.[19,20]

To determine the differences in fragment size and coverage in a position-dependent manner across the genome, we mapped the fragments to their genomic origin and evaluated fragment lengths covering the whole genome in windows of 2 Mb, which performed first-class sensitivity in low-coverage sequencing.[36] The proportion of fragments 20–150 bp long was counted in each window, and particular fragments proportion of different sizes was calculated to optimize the criterion of fragmentation, which was reported as an optimal classification standard.[23] Fragments of cfDNA in plasma are typically 166 bp in size, which was marked with a dotted line in Fig. 2a. We observed an enrichment shift of cfDNA fragment sizes in this line between lung cancer patients and healthy control (p < 0.05, Wilcoxon rank sum test). The lung cancer ctDNA were more fragmented than non-mutant cfDNA, with the maximum enrichment in 100–150 bp, as well as enrichment in the size ratio of 20–150 bp vs 160–180 bp (Fig. 2b). Genomic instability occurs in most cancers and can be quantified by higher normalized read counts in specific regions.[37] We divided the entire genome into windows of 2M bases and calculated the logarithm of the FPKM (Fragments per Kilobase Million) value of the read count in each window. The Bincount differences between LC and HC/benign were evaluated by the Mann–Whitney U test, and the scores of LC patients were higher than that of HC/benign individuals in certain regions (Fig. 2c). The plasma DNA end motifs were identified using the first four nucleotide sequences (i.e., 4-mer) on each 5′ end of the plasma DNA fragment after alignment to the reference genome. Each end motif had 4^4 or 256 features.[16] The Mann–Whitney U test was performed to cluster the different DNA end motif features between the LC and HC/benign individuals. Among these, the top eight motifs were significantly different between the two groups (Fig. 2d). When applying the 5 mC classifier model to retrieve the validation cohort, the classification could roughly distinguish LC individuals from HC/benign individuals. The heat map of the methylation model might be applied to different clinical subgroups, including consolidation tumor ratio (CTR), nodule type, and histological diagnosis (Fig. 2e).

The evaluation performance of the HIFI method in lung cancer early screening model

To evaluate the diagnostic performance for lung cancer identification, a weighted diagnostic model was constructed using a machine learning method that took all individual genomic features (hypo-methylation, motif, fragment, and Bincount) into account. Methylation had great potential in early screening and were proved to be of the priorities among cfDNA characters in cancer early detection, so we also brought methylation into early screening model as an indicator to improve model's stability.[38] Support vector machine (SVM) was used to classify the best vectors of HIFI platform. When applying this LC score model to retrieve the validation cohort (LC = 143, HC/benign = 163), the principal component analysis (PCA) classification could also roughly distinguish LC patients from HC/benign individuals (Fig. S3). The privilege of SVM was further validated with the highest AUC of 0.945 among adaptive boosting (AdaBoost), Gaussian process, K nearest

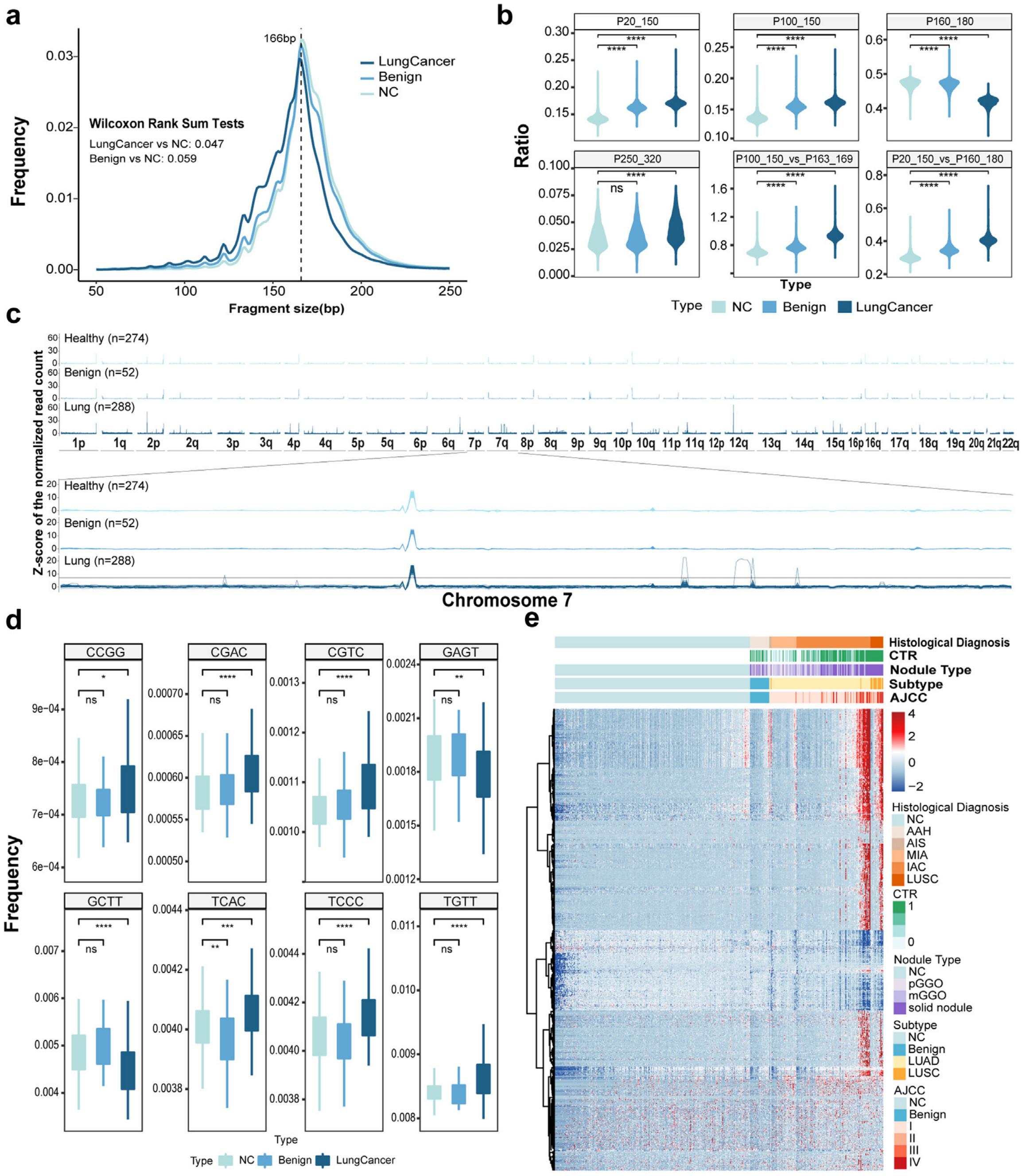

Fig. 2: **Multiple genomic feature alteration and methylation clustering. (a)** Fragment size distributions between the LC patients from HC/benign ones. **(b)** The density distribution of six specific fragment lengths between the LC patients from HC/benign ones. **(c)** The evaluation of genome-wide cfDNA stability characteristics by Bincount, and analyses of healthy cfDNA (top), benign (middle), and lung cancer (bottom) Bincount profiles based on chromosome 7. **(d)** Statistical results of cfDNA genomic end motif differences between the LC patients from HC/benign ones. The top eight end motifs were significantly different between the LC patients from HC/benign ones. **(e)** Heat map classification of LC patients and HC/benign participants using co-methylation features. To all data in Fig. 2, n = 614, *p*-value were computed with Wilcoxon rank sum test. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$; $p > 0.05$ was considered not significant (ns). NC, negative control. AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; LUSC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; CTR, consolidation tumor ratio, representing the value from 0% to 100%. pGGO, pure ground-class opacity; mGGO, mixed ground-glass opacity.

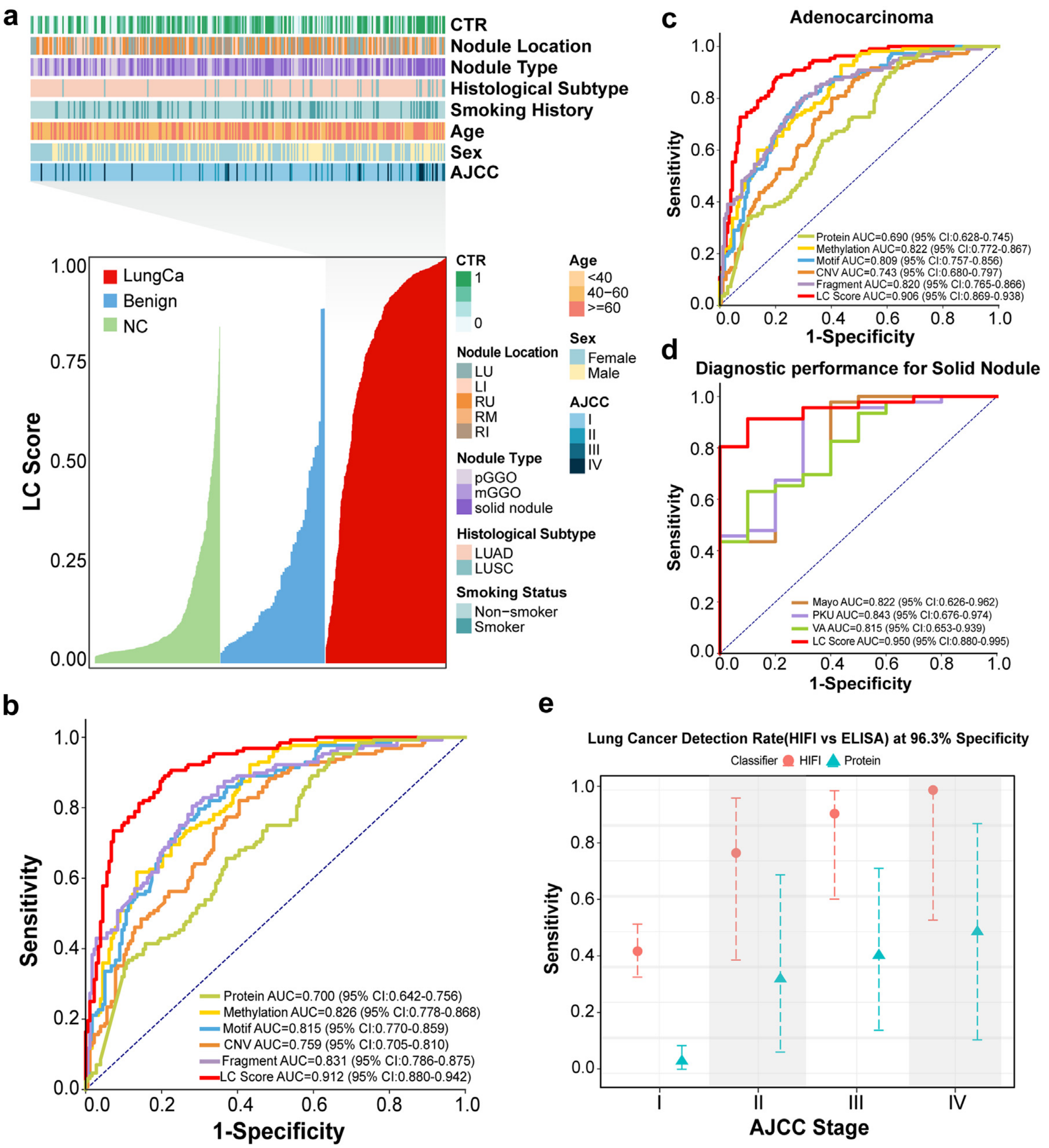

Fig. 3: **The diagnostic performance of the LC score model in lung cancer patients. (a)** The obtained scores of all participants in the validation cohort (n = 306) when evaluated by LC score model. The cutoff value of the LC score was set as 0.38. Upper: clinical characteristics of lung cancer patients. **(b)** ROC curves and associated AUC values in the validation cohort (n = 306) when evaluated by different genomic feature models. **(c)** ROC curves and associated AUC values for lung adenocarcinoma patients (n = 288). **(d)** ROC curves and associated AUC values for solid nodule patients (n = 57) when evaluated by LC, VA, Mayo, and PKU models. **(e)** The obtained sensitivity of different clinical stages in the validation cohort when evaluated by LC score model and protein biomarker model at the specificity of 96.3%.

neighbor (KNN), random forest (RF), SVM and extreme gradient boosting (XGBoost) (Fig. S2b). This model was used to assess the genomic features for each one in the validation cohort and provide an independent score (LC score). The score ranged from 0 to 1. A large proportion of the lung cancer patients had a high score, and almost all healthy people had lower scores (Fig. 3a and Fig. S4). The detection performance of these classifier models was measured by the ROC curve based on the independent validation dataset. The fragmentation model achieved an AUC value of 0.831 (95% CI: 0.765–0.866) (Fig. 3b). When all Bincount features were taken for machine learning using the weight generated by the SVM, the ROC analyses for the detection of patients with lung cancer showed an AUC value of 0.759 (95% CI: 0.705–0.810). To efficiently and accurately identify lung cancer patients, a motif classifier was constructed using SVM to retrieve these motifs, where the magnitude of each end motif was considered. The resulting motif classifier model showed a detection ability with an AUC of 0.815 (95% CI: 0.770–0.859) (Fig. 3b). The methylation model achieved an AUC of 0.826 (95% CI: 0.778–0.868) (Fig. 3b). In this study, five serum protein biomarkers (CEA, CYFRA21-1, NSE, CA-199, CA-125) were also measured for clinical cancer detection (Table S3), and a multivariate predictive model based on SVM was constructed and tested on the validation cohort with bootstrapping AUC (95% CI) of 0.700 (0.642–0.756) (Fig. 3b). Collectively, all four genomic features (hypo-methylation, end motif, fragment, and Bincount) showed better diagnostic characteristics than clinically used protein biomarkers for distinguishing LC individuals from HC/benign ones.

The LC score model, which integrated all the above four genomic features, has an excellent performance on the training set (Fig. S5a), and also shown an AUC value of 0.912 (95% CI: 0.880–0.942) on the validation set (Fig. 3b). Similar detection capability was seen for lung adenocarcinoma and squamous patients, which is the most common type of lung cancer (Figs. 3c and S4). As to patients with different clinical stages (stage I, II, III, and IV), the LC score model still showed the best screening performance than other single feature models (Fig. S5). When concerning the diagnostic performance for solid nodules, the LC score model showed superior detection efficiency than Mayo,[39] PKU,[40] and VA models,[41] and achieved an AUC of 0.950 (95% CI: 0.880–0.995) (Fig. 3d). Protein biomarkers were known as the important biochemical indicators for early screening of lung cancer. From the comparison of the detected sensitivity for lung cancer patients in the validation set, the LC model showed excellent performance in different clinical stages than protein biomarkers, especially in the early clinical stages (Fig. 3e), manifesting its practical value in clinical use. The performance of the screening model was also assessed in different clinical subgroups of lung cancer patients. The LC scores did not show difference for different nodule location subtypes (LU, LI, RU, RM, RI) and sex (Fig. S6a and g). Lung solid nodules are more severity types than GGO, and these patients got much higher LC scores (Fig. S6b, $p < 0.01$, Wilcoxon rank sum test). The LC scores significantly increased with the development of the clinical stage (Fig. S6c, $p < 0.0001$, Wilcoxon rank sum test). Different immune infiltration situations and volume represent different developmental stages of the tumor, and the patients with advanced tumor volume and invasive adenocarcinoma of lung cancer showed significantly higher LC scores (Fig. S6d, e and f, $p < 0.05$, Wilcoxon rank sum test).

In order to verify whether the coverage was enough for HIFI methods and further explore the limit of low-pass WGS, we checked the outcome of 2.5×, 2×, 1.5×, 1×, 0.5× genome coverage in training set and validation set independently (Fig. S7). The AUC was relatively stable even at 1× coverage, proving the veracity of 2.5× coverage we conducted in this research and providing a possibility to reduce the coverage with non-inferior outcome.

Evaluation of the HIFI method in patients with GGO

Ground-glass Opacity (GGO) was a kind of radiologic appearance, which always linked to pre-invasive and indolent tumors.[42] It shed lower level of ctDNA and caused ambiguous result in lung cancer screening.[10] The efficiency of the LC score model was also evaluated in 49 patients who were identified as ground-glass opacity by clinical imaging diagnosis and clinical characters of these cases were present in Table S4. Mutations of these patients were also captured from blood samples and resected tissue based on cSMART2.0 assay. Mutation profiles of tumor tissue and plasma samples in these patients were shown in Fig. 4a. In tumor-naïve methods, we used blood test results as the only criteria, while the tumor-informed methods also took sequences of tissue samples into consideration to make a malignant differentiation. The majority of patients were detected with tumor tissue mutations positive while plasma samples were negative (TpBn). Only a few tissue-positive patients were judged as plasma mutations positive (TpBp). Different detection sensitivity was achieved among these three methods, it was 14.3% by the tumor-naïve method and 16.3% by the tumor-informed plasma mutation detection method, while it was 75.5% when these patients were evaluated by the LC score model (Fig. 4b). These results confirmed that the LC score model also demonstrated superior diagnostic efficiency in the very early lung cancer stage. The LC score model showed different detection performances among different subtypes of GGO samples, with mixed GGO (mGGO) patients showing significantly higher LC scores than pure GGO (pGGO) patients (Fig. 4c, $p < 0.0005$, Wilcoxon rank sum test). The acquired LC score showed a positive correlation with cfDNA concentration and CTR of solid nodules (Fig. 4d and e).

Furthermore, when adjusted to a population (Chinese population, aged 20–80 years) with a low incidence of 0.15%,[43] the LC score model, with sensitivity of 51.8% and the specificity of 96.3%, acquired a very high NPV of 99.9%. Hence, it might be used as a

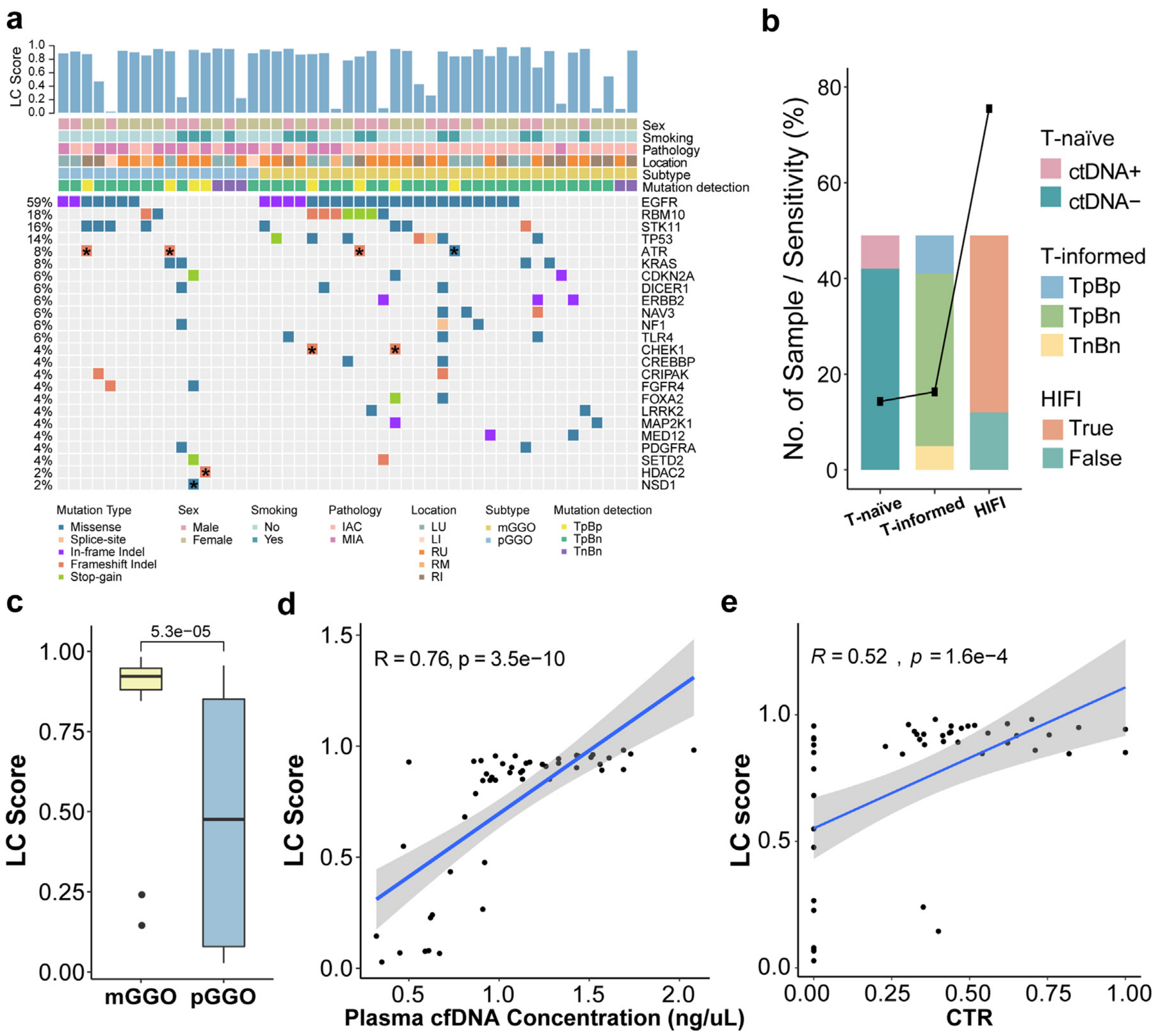

Fig. 4: **The evaluation performance of the LC model in patients with lung ground-glass opacity (GGO). (a)** Mutation profiles of tumor and plasma. Upper column chart: LC score of each patient; Middle block diagram: clinical information of patients; Lower block diagram: the gene list with mutant frequency ≥4% in these patients. The mutant frequency of *HDAC2* and *NSD1* was 2%, but they were detected in both tumor tissue and plasma. Asterisk indicates mutations detected in the tumor tissue and plasma. **(b)** The detected positive sensitivity by different methods (n = 49). TpBp, both tissue and plasma were mutation-positive; TpBn, tissue was positive while blood was negative; TnBn, both tissue and blood were mutation-negative. **(c)** The sample LC score differences between mGGO from pGGO (n = 49, Wilcoxon rank sum test). **(d)** Correlation between the LC score and the cfDNA concentration (n = 49, Spearman's rank correlation rho). **(e)** Correlation between the LC score and the proportion of solid nodules (n = 49, Spearman's rank correlation rho).

population-based LC preliminary screening tool to minimize the risk of radiation exposure and reduce patients' unnecessary concerns. Additionally, a trade-off value was calculated that showed a harm/benefit value below 0.02103155, approximately 1:48 (the formula is presented in the section on methods).[34] These results indicated that with this model we can recognize one lung cancer patients at a maximum cost of 48 participants' healthy or with benign nodules falsely categorized into positive. The results above predicted a beneficial clinical application of the HIFI method in lung cancer early screening.

Multi-omics integration in postoperative residual disease monitoring

The detection of trace tumor mutation information in plasma remains a major technical challenge in assessing the risk of recurrence in patients after surgery. The HIFI method was thus used to evaluate the postoperative MRD status of lung cancer patients combined with mutation panel detection. The clinical information of these patients is listed in Table S5. An advanced targeted next-generation sequencing assay cSMART2.0, which specialized in capturing genomic characters under low ctDNA density, was applied for ctDNA mutation's detection.[22,44]

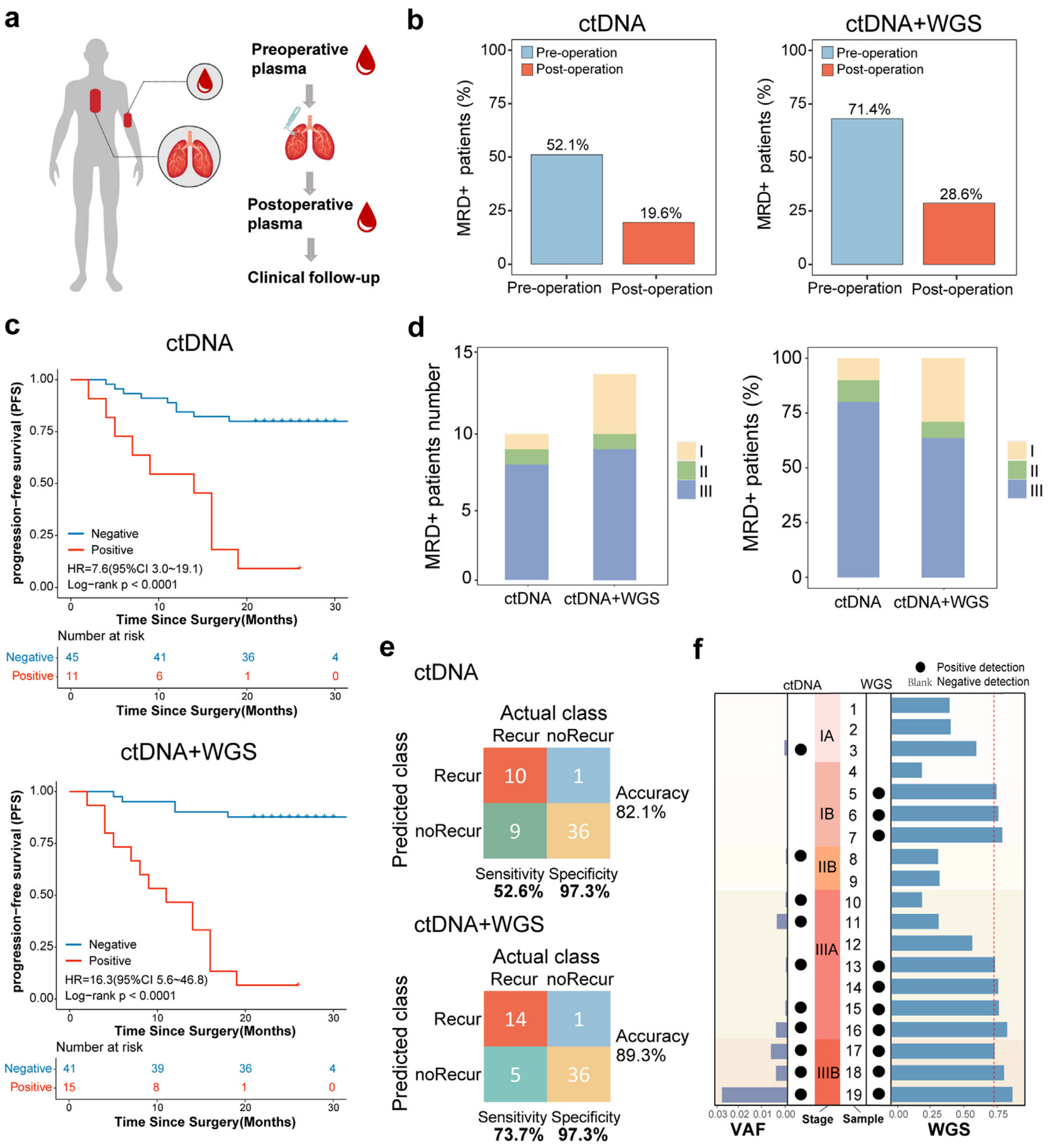

Fig. 5: **Comparisons of ctDNA and ctDNA combined with WGS detection in lung cancer postoperative MRD monitoring. (a)** Study design of the postoperative MRD monitoring. **(b)** The positive rate of ctDNA detection in patients before and after the operation. The left shows only ctDNA signals and the right shows ctDNA detection combined with WGS (ctDNA + WGS) results. Left: 52.1% (24/46) and 19.6% (11/56). Right: 71.4% (30/42) and 28.6% (16/56). **(c)** The statistical of patients' recurrence-free survival (RFS). The above figure showed samples detected only by ctDNA mutation and the bottom was evaluated result by ctDNA + WGS (n = 56, log-rank test). **(d)** The statistical result for the staging of MRD-positive samples after the operation. Left: the number of MRD-positive patients. ctDNA (n = 10), ctDNA + WGS (n = 14). Right: the ratio of MRD-positive patients. **(e)** Confusion matrices showing the ctDNA and ctDNA + WGS detection for recurrence prediction in the postoperative population. **(f)** The mutation detection on tumor recurrence patients after operative (n = 19 patients), using ctDNA (gray) and WGS (blue) methods. Bar plots were used to calculate mutation signature of in the postoperative patient plasma. The bar graph on the left was the VAF of the patient's postoperative plasma ctDNA detection, the right was the LC score when detected by WGS. Black dot indicating it was mutation positive.

A 218-hotspot-gene panel (Table S6) was used for postoperative plasma monitoring (Fig. 5a right) after examination of specific tumor mutational sequencing of these patients from resected tissue. For all patients, when only ctDNA molecules were detected using cSMART2.0 assay, the mutation-positive rate of the preoperative and postoperative specimens was 52.1% (24/46) and 19.6% (11/56), respectively (Table S7). When it was evaluated by single-nucleotide variants detection combined with WGS-based HIFI method (ctDNA + WGS) in most of these patients, the mutation-positive rate of pre-operation and post-operation patients increased to 71.4% (30/42) and 28.6% (16/56), respectively (Fig. 5b). For 21 patients, who were detected as VAF-positive before the surgery, their mutation profiles were compared between ctDNA and HIFI detection (Fig. S8). The results of the recurrence-free survival (RFS) were evaluated in the postoperative population. The recurrence ratio for MRD-positive patients was 7.60 folds (95% CI: 1.3–8.8, $p = 9.00E\text{-}03$, log-rank test) of ctDNA-negative patients when only detected by the ctDNA mutation panel, and increased to 16.25 folds (95% CI, 4.4–35.3, $p < 1.00E\text{-}04$, log-rank test) of MRD-negative samples when detected by ctDNA + WGS (Fig. 5c). There were also some sensitivity differences when the clinical stage of these patients was taken into account. For all MRD-positive patients, the ratio of stage I patients was 10.0% (1 patient) when detected only by ctDNA mutation, while it was 21.0% (4 patients) when WGS-based HIFI method was added (Fig. 5d). The recurrence prediction of these patients was also evaluated by these two methods. When the detection was only performed by ctDNA mutation, the accuracy of the predicted recurrence was 82.1% with a sensitivity of 52.6% at the specificity of 97.3% (10 of 19 patients recurred and 1 patient was false-positive) (Fig. 5e). When it was detected by ctDNA + WGS, the accuracy of the predicted recurrence was 89.3% with a sensitivity of 73.7% at the specificity of 97.3% (14 of 19 patients recurred and 1 patient was false-positive). Among stage I patients, ctDNA mutation only detected 1 in 7 patients who met recurrence in the following 30 months, while ctDNA + WGS strategy had more advantages with an elevated sensitivity of 57.14%. For mutation-positive detection of patients with clinical recurrence, WGS method could improve the accuracy of MRD detection in postoperative risk assessment of patients, especially for very early-stage lung cancer (patients No. 5, 6 and 7. Fig. 5f).

Discussion

Liquid biopsy manifests application prospects in many fields of lung cancer as previously reported,[11] whereas most studies that perform liquid biopsy focus either on cancer screening or MRD detection. MRDetect is the limited methods that run through the whole process of diagnosis-therapy pattern, based on single-nucleotide variation and copy number variation under a coverage depth of 35×.[14] In order to probe the characters of ctDNA in other dimensions and make it more suitable for clinical use, we constructed the WGS-based HIFI platform based on several genomic fragment features, accompanied with epigenetic signatures and other mutation information to detect early-stage cancers and discern postoperative patients with high risk of recurrence. We selected one top-class indicator in each field adding to HIFI method (motif, fragment and Bincount).[38,45,46] For LC screening, the LC score model was constructed using the SVM and XGBoost method based on the HIFI method and the hyper-co-methylated read approach. The LC score model had excellent screening performance in the validation set and achieved an AUC of 0.912. Subsequently, we incorporated the HIFI method with the cSMART 2.0 detection platform to improve its effectiveness in lung cancer MRD detection, which showed high accuracy in MRD monitoring.[8,47] Consequently, the MRD detection achieved a specificity of 97.3% with the accuracy of 89.3%, and the detection rate had better detection efficiency over traditional ctDNA mutation methods.

Our technical approach has overcome challenges that are commonly encountered in the applications of liquid biopsy. Low fraction of ctDNA is a major limiting factor for LC screening, and it is also challenging for MRD detection as a result of low tumor load.[5,48,49] CHIP is another common confounding factor that influences the analysis of ctDNA mutation and contributes to false-positive results.[12] WGS, which enables effective integration across orthogonal data dimensions, can overcome these problems and perform accurate and sensitive ctDNA detection, allowing clinical application in various types of tumors.[14,21] Several studies have elucidated the molecular characteristics of ctDNA, including fragment sizes, nucleosome protection and accessibility, sequence motif of end-points, and genome instability.[17,50,51] Whereupon, we developed HIFI method for whole diagnosis-therapy procedure based on fragmentation, motif and Bincount. cfDNA fragmentation is a non-random process. The most abundant plasma DNA is 166 bp long, and fragmentation end-points show relationships with nucleosome organization.[21,52] As cfDNA fragmentation can comprehensively represent both genomic and chromatin characteristics, it can be used as a genomic feature for many tumor-derived changes.[18,53] Plasma DNA preferred ends contain information on the tissue of origin, which is because these genomic locations are preferentially cleaved when plasma DNA is generated.[16,17,54] Genome instability might cause genomic abnormalities, which can be used as a prognostic predictor of lung cancer.[37,55] Data on reads of different sizes from cfDNA can reflect alterations in the physiological state between tumor and non-tumor derived cell populations.[17,52,56] So, specific genomic features of cfDNA detected by WGS can provide more precise information about tumor cell populations.[52,57]

In clinical practice, it is difficult to confirm the presence of lung cancer before an invasive procedure, especially in early stages.[58] The opinion of the examining physician plays a crucial role in distinguishing malignant lung nodules from benign ones.[59] To reduce subjective bias, some of the clinical guidelines favor objective data-based math models for replacing equivocal subjective judgments for the diagnosis of lung cancer. Examples of these models include the Mayo model,[39] which focuses on out-patients, and the Brock model[60] for wide screening. Liquid biopsy is a sensitive, non-invasive methods to improve accuracy from another aspect. Metabolomics is also a sensitive biomarker for cancer early detection,[61] but significant batch effects are revealed in former studies, which has no unified improvement standard.[62] As a result, we only take features from ctDNA into account during LC model construction. The HIFI method is encouraging for the diagnosis of patients with suspicious lung nodules due to the following reasons: a) Compared to the existing mainstream early-detection math models, the LC score model exhibited higher accuracy, especially in the solid nodule cohort (with an AUC of 0.950, Fig. 3d) that outperformed other models. Its performance in early detection even had comparable efficiency with a new PKU-M machine-learning model.[63] b) The result of liquid biopsy is likewise an objective reference for physicians to consider, avoiding erratic personal biases. c) Many models require multitudinous clinical parameters like smoking history and family history.[39,60,63] This information might be hard to obtain under certain circumstances, while the LC score model does not need information on such parameters, and thus, is more convenient for clinical application. In modulation of Chinese population, LC score reached a preferable NPV, which freed plenty healthy individuals from radiation exposure and liberated them from anxiety during the follow-up procedure. Acting as an assistant means to LDCT, it was potential to make up for the high false positive rate from traditional LDCT, bringing real benefits.

Traditional ctDNA mutation assays and methylation detection methods that cover more loci cannot efficiently detect LC at a very early stage[64] owing to low levels of ctDNA and few available DNA fragments in a typical plasma sample.[13] In recent years, the pathological composition spectrum of LC has changed, and the proportion of adenocarcinoma cases has increased. Detecting patients with stage I lung adenocarcinoma has always remained a major challenge during early screening.[5,10,11] Although recent studies had developed a targeted-methylation method with favorable efficiency for multi-cancer early detection, its availability on early stage lung cancer still needed further exploration.[65] The patients in our study mainly had lung adenocarcinoma (89.2%), mostly in stage I (80.9%). The LC score model achieved significant improvement in sensitivity at the same level of specificity, compared with traditional protein biomarkers, especially in stage I patients of the validation set, which allowed more patients with early-stage LC to be identified. Additionally, most sub-solid nodules, particularly pure GGOs, were very early-stage LC with a low tumor burden,[66,67] which is difficult for ctDNA mutation detection. To further test our model in very early-stage cancer, we compared the LC score with a traditional tumor-informed approach in two independent GGO cohorts. Along with other necessary clinical data, the LC score model achieved a sensitivity of about 80.0%, while mutation detection only had a sensitivity of ≤20%, outperforming tumor-naïve and tumor-informed methods. These results suggested that the LC score model exhibited a robust performance in GGO, and is more sensitive than traditional methods, indicating a breakthrough in the diagnosis of very early-stage LC.

LC score is a convenient and non-invasive detection method, suitable for large population screening. Although liquid biopsy dramatically improves the sensitivity of early detection, the huge economic financial outlay is one of the major impediments on its way to clinical application. On the basis of ensuring model performance, we adopted relatively cheap and convenient detection method for model construction. The average coverage of WGS in HIFI method is as low as 2.5×, which realized authentic low coverage sequencing without extra time cost on both models. By adding the WGS section, the expense added by a few dozen dollars and the total cost price was close to that of a clinical comprehensive tumor marker detection. Furthermore, our research checked the detection efficiency of lower coverage and the result indicated the potential of 1× coverage in HIFI model, which might further shrink the budget by approximately 60%. The requirement for universal population screening is not only money-and-time-saving but also harmless. Low-pass WGS only needs a low volume of plasma (2 mL) to complete the test, much less than other liquid biopsy programs, which reduces the harm to a minimum.

Our MRD detection model also has improvements in clinical applications. The effect of adjuvant chemotherapy is unsatisfactory, with only a 4% survival benefit at 5 years.[68] A major clinical significance of MRD detection is that it can identify patients at high risk of recurrence and guide precision therapy. MRD detection strategies can be roughly classified into tumor-informed and tumor-agnostic. Personalized whole exome sequencing-based tumor-informed panel owns better efficiency in distinguishing low-concentrated ctDNA in plasma, but cost more time and expense for individualized panel. While, the tumor-agnostic one has convenient detection testing cycle, independent of tumor tissue, at the cost of sensitivity and specificity.[69,70] Our MRD model is constructed on the basis of WGS and cSMART2.0, which combines the best of both strategies. First, the use of WGS and tumor-agnostic panel simplifies workflow of current personalized tumor-informed MRD detection method, by eliminating the need for designing panels, spacing a large amount of time and money from panel design. Second, the combination of HIFI and mutation detection on cSMART2.0 significantly improves detection ability. MRD monitoring is also confronting the challenge on low concentration of ctDNA. cSMART2.0 is state of the art in ctDNA mutation detection with its unique inverse primer pairs and specially optimized capture process especially in ultra-low ctDNA frequency which has shown its preponderance in previous studies.[22,71] But the result is still ungratified as its sensitivity in MRD detection is merely 52.6%. When WGS-based HIFI result is added for MRD detection, the sensitivity got significant improvement without prominent declination on specificity and found the proportion of patients with stage I LC who were MRD-positive increased considerably accompanied with corresponding shorter RFS. Our MRD model not only economizes time and expense from intricate panel designing, but also possesses high MRD detection efficiency.

However, the limitations of our study must be recognized. Firstly, the HIFI method constructed a multi-dimensional LC screening model by integrating cfDNA end motifs, fragment, and Bincount information, while the application of various machine learning methods may make the model highly complex and thus lead to over-fitting. Therefore, its wider range of applications performance needs to be evaluated in further independent test sets. Secondly, for pGGO, we have not done a detailed performance evaluation to verify whether the HIFI model has sufficient efficacy in the differential diagnosis of benign and malignant pGGOs. Finally, the enrolled patients inevitably involved some selection bias, the accuracy of this model need to be validated in a cohort that involving larger and more diverse populations.

In conclusion, the application of non-invasive liquid biopsy will increase with improvement of sequencing technology, and multi-omics analysis could improve the detection accuracy of cancer signals. Furthermore, as a simple and effective strategy, low-pass WGS and tumor-agnostic cSMART 2.0 have considerable clinical prospects than precedent studies. Our WGS-based integrative multi-dimensional analytical models and the HIFI method demonstrated the effectiveness for LC screening and postoperative surveillance. The HIFI method could serve as an auxiliary method to significantly reduce the false-positive rate and improve the accuracy of LDCT-based LC screening due to its sufficiently high specificity. Additionally, MRD detection based on liquid biopsy might be routinely used for longitudinal monitoring to reduce radiation exposure and allow timely treatment. Our HIFI method opens a new avenue for whole diagnosis-therapy process.

Contributors

Each author's individual contributions to this work were depicted with CRediT Taxonomy. Yun Li, Guanchao Jiang, Wendy Wu, Hao Yang, and Yichen Jin contributed equally to this work. All authors were involved into conceptualization of the study. Hao Yang, Manqi Wu, and Kezhong Chen were responsible for methodology. Hao Yang, Manqi Wu, Wenjie Liu, and Airong Yang contributed to software, data curation, formal analysis and visualization. Kezhong Chen contributed to validation and funding acquisition. Hao Yang, Yichen Jin, Manqi Wu, Wenjie Liu, Airong Yang, Sujie Zhang, Lu Zheng, Xueying Zhang, and Fengxia Du contributed to resources and investigation. Yun Li, Guanchao Jiang, Wendy Wu, Lin Wu, Fan Yang, Kezhong Chen, and Jun Wang contributed to project administration and supervision of the work. Hao Yang, Yichen Jin, Manqi Wu, Wenjie Liu, Airong Yang, Sujie Zhang, Lu Zheng, Xueying Zhang, and Fengxia Du contributed to writing-original draft preparation. Yun Li, Guanchao Jiang, Wendy Wu, Hao Yang, Yichen Jin, Manqi Wu, Olga Chervova, Nnennaya Kanu, Lin Wu, Fan Yang, Jun Wang, and Kezhong Chen contributed to writing-review & editing part. The order of the five co-first authors was decided according to the workload and total contribution to this research. All authors read and approved the final manuscript. Hao Yang and Yichen Jin have accessed and verified the data in the manuscript.

Data sharing statement

All the data appeared in this manuscript is available to others, including Supplementary Files. Access to these data is present in the part of Data deposition and materials sharing.

Declaration of interests

Wendy Wu, Wenjie Liu, Hao Yang, Airong Yang, and Xueying Zhang are the employee of Berry Oncology Corporation. Two patents are planning to implement on China National Intellectual Property Administration and we declare no relevant interest conflict exists against this article. The other authors declare that they have no competing interests.

Acknowledgements

The authors wish to acknowledge the contribution of Yun Wang and Yanyan Hou in the laboratory of Peking University People's Hospital for their technical assistance.

This work was supported by CAMS Innovation Fund for Medical Sciences (CIFMS) 2022-I2M-C&T-B-120, Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, Chinese Academy of Medical Sciences (2021RU002), National Natural Science Foundation of China (No. 82072566 and No. 92059203), Beijing Natural Science Foundation (L222021) and Peking University People's Hospital Research and Development Funds (RZ2022-03).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104553.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249.

2 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;70(1):7–30.

3 Chansky K, Detterbeck FC, Nicholson AG, et al. The IASLC lung cancer staging project: external validation of the revision of the TNM stage groupings in the eighth edition of the TNM classification of lung cancer. *J Thorac Oncol*. 2017;12(7):1109–1121.

4 Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395–409.

5 Abbosh C, Birkbak NJ, Wilson GA, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*. 2017; 545(7655):446–451.

6 Chaudhuri AA, Chabon JJ, Lovejoy AF, et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov*. 2017;7(12):1394–1403.

7 Chen K, Zhang J, Guan T, et al. Comparison of plasma to tissue DNA mutations in surgical patients with non-small cell lung cancer. *J Thorac Cardiovasc Surg*. 2017;154(3):1123–1131.e2.

8 Chen K, Zhao H, Shi Y, et al. Perioperative dynamic changes in circulating tumor DNA in patients with lung cancer (DYNAMIC). *Clin Cancer Res*. 2019;25(23):7058–7067.

9 Newman AM, Lovejoy AF, Klass DM, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*. 2016;34(5):547–555.

10 Chabon JJ, Hamilton EG, Kurtz DM, et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature*. 2020;580(7802):245–251.

11 Abbosh C, Birkbak NJ, Swanton C. Early stage NSCLC - challenges to implementing ctDNA-based screening and MRD detection. *Nat Rev Clin Oncol*. 2018;15(9):577–586.

12 Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*. 2014;371(26):2488–2498.

13 Avanzini S, Kurtz DM, Chabon JJ, et al. A mathematical model of ctDNA shedding predicts tumor detection size. *Sci Adv*. 2020;6(50):eabc4308.

14 Zviran A, Schulman RC, Shah M, et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med*. 2020;26(7):1114–1124.

15 Chen K, Sun J, Zhao H, et al. Non-invasive lung cancer diagnosis and prognosis based on multi-analyte liquid biopsy. *Mol Cancer*. 2021;20(1):23.

16 Jiang P, Sun K, Peng W, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov*. 2020;10(5):664–673.

17 Jiang P, Sun K, Tong YK, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc Natl Acad Sci U S A*. 2018;115(46):E10925–E10933.

18 Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 2019;570(7761):385–389.

19 Wan N, Weinberg D, Liu TY, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer*. 2019;19(1):832.

20 Liu H, He W, Wang B, et al. MALBAC-based chromosomal imbalance analysis: a novel technique enabling effective non-invasive diagnosis and monitoring of bladder cancer. *BMC Cancer*. 2018;18(1):659.

21 Chen L, Abou-Alfa GK, Zheng B, et al. Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. *Cell Res*. 2021;31(5):589–592.

22 Liu J, Yang Y, Liu Z, et al. Multicenter, single-arm, phase II trial of camrelizumab and chemotherapy as neoadjuvant treatment for locally advanced esophageal squamous cell carcinoma. *J Immunother Cancer*. 2022;10(3):e004291.

23 Mouliere F, Chandrananda D, Piskorz AM, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*. 2018;10(466):eaat4921.

24 Li W, Li Q, Kang S, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res*. 2018;46(15):e89.

25 Hsu HS, Chen TP, Hung CH, et al. Characterization of a multiple epigenetic marker panel for lung cancer detection and risk assessment in plasma. *Cancer*. 2007;110(9):2019–2026.

26 Begum S, Brait M, Dasgupta S, et al. An epigenetic marker panel for detection of lung cancer using cell-free serum DNA. *Clin Cancer Res*. 2011;17(13):4494–4503.

27 Dietrich D, Kneip C, Raji O, et al. Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates. *Int J Oncol*. 2012;40(3):825–832.

28 Heikamp K, Bajorath J. Support vector machines for drug discovery. *Expert Opin Drug Discov*. 2014;9(1):93–104.

29 Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika*. 1967;54(1):167–179.

30 Huang L, Jiang XL, Liang HB, et al. Genetic profiling of primary and secondary tumors from patients with lung adenocarcinoma and bone metastases reveals targeted therapy options. *Mol Med*. 2020;26(1):88.

31 Zhang Q, Jia H, Wang Z, et al. Intertumoural heterogeneity and branch evolution of synchronous multiple primary lung adenocarcinomas by next-generation sequencing analysis. *Front Oncol*. 2021;11:760715.

32 Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform*. 2014;48:193–204.

33 Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29.

34 Mazzone PJ, Sears CR, Arenberg DA, et al. Evaluating molecular biomarkers for the early detection of lung cancer: when is a biomarker ready for clinical use? An Official American Thoracic Society Policy Statement. *Am J Respir Crit Care Med*. 2017;196(7):e15–e29.

35 Serpas L, Chan RWY, Jiang P, et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc Natl Acad Sci U S A*. 2019;116(2):641–649.

36 Gusnanto A, Taylor CC, Nafisah I, Wood HM, Rabbitts P, Berri S. Estimating optimal window size for analysis of low-coverage next-generation sequence data. *Bioinformatics*. 2014;30(13):1823–1829.

37 Pollard JW. Tumour-educated macrophages promote tumour progression and metastasis. *Nat Rev Cancer*. 2004;4(1):71–78.

38 Jamshidi A, Liu MC, Klein EA, et al. Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell*. 2022;40(12):1537–1549.e12.

39 Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med*. 1997;157(8):849–855.

40 Li Y, Chen KZ, Wang J. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in Chinese people. *Clin Lung Cancer*. 2011;12(5):313–319.

41 Gould MK, Ananth L, Barnett PG. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest*. 2007;131(2):383–388.

42 Handa Y, Tsutani Y, Okada M. Transition of treatment for ground glass opacity-dominant non-small cell lung cancer. *Front Oncol*. 2021;11:655651.

43 Li X, Gao S. Trend analysis of the incidence, morbidity and mortality of lung cancer in China from 1990 to 2019. *Chin J Prev Contr Chron Dis*. 2021;29:821–826.

44 Han M, Li Z, Wang W, et al. A quantitative cSMART assay for noninvasive prenatal screening of autosomal recessive non-syndromic hearing loss caused by GJB2 and SLC26A4 mutations. *Genet Med*. 2017;19(12):1309–1316.

45 Xia L, Mei J, Kang R, et al. Perioperative ctDNA-based molecular residual disease detection for non-small cell lung cancer: a prospective multicenter cohort study (LUNGCA-1). *Clin Cancer Res*. 2022;28(15):3308–3317.

46 Gale D, Heider K, Ruiz-Valdepenas A, et al. Residual ctDNA after treatment predicts early relapse in patients with early-stage non-small cell lung cancer. *Ann Oncol*. 2022;33(5):500–510.

47 Lv W, Wei X, Guo R, et al. Noninvasive prenatal testing for Wilson disease by use of circulating single-molecule amplification and resequencing technology (cSMART). *Clin Chem*. 2015;61(1):172–181.

48 Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med*. 2017;9(403):eaan2415.

49 Newman AM, Bratman SV, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*. 2014;20(5):548–554.

50 Chan KC, Zhang J, Hui AB, et al. Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem*. 2004;50(1):88–92.

51 Lo YM, Chan KC, Sun H, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med*. 2010;2(61):61ra91.

52 Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*. 2016;164(1–2):57–68.

53 Mathios D, Johansen JS, Cristiano S, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun*. 2021;12(1):5060.
54 Chan KC, Jiang P, Sun K, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci U S A*. 2016;113(50):E8159–E8168.
55 Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of genomic instability in cancer. *Cell*. 2017;168(4):644–656.
56 Chan KC, Jiang P, Chan CW, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A*. 2013;110(47):18761–18768.
57 Ulz P, Thallinger GG, Auer M, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet*. 2016; 48(10):1273–1278.
58 Balata H, Fong KM, Hendriks LE, et al. Prevention and early detection for NSCLC: Advances in Thoracic Oncology 2018. *J Thorac Oncol*. 2019;14(9):1513–1527.
59 Raghunath S, Maldonado F, Rajagopalan S, et al. Noninvasive risk stratification of lung adenocarcinoma using quantitative computed tomography. *J Thorac Oncol*. 2014;9(11):1698–1703.
60 McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med*. 2013;369(10):910–919.
61 Bratulic S, Limeta A, Dabestani S, et al. Noninvasive detection of any-stage cancer using free glycosaminoglycans. *Proc Natl Acad Sci U S A*. 2022;119(50):e2115328119.
62 Han W, Li L. Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom Rev*. 2022;41(3):421–442.
63 Chen K, Nie Y, Park S, et al. Development and validation of machine learning-based model for the prediction of malignancy in multiple pulmonary nodules: analysis from multicentric cohorts. *Clin Cancer Res*. 2021;27(8):2255–2265.
64 Castro-Giner F, Gkountela S, Donato C, et al. Cancer diagnosis using a liquid biopsy: challenges and expectations. *Diagnostics*. 2018;8(2):31.
65 Klein EA, Richards D, Cohn A, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol*. 2021;32(9): 1167–1177.
66 Kim YW, Kwon BS, Lim SY, et al. Lung cancer probability and clinical outcomes of baseline and new subsolid nodules detected on low-dose CT screening. *Thorax*. 2021;76(10):980–988.
67 Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: a review. *JAMA*. 2022;327(3):264–273.
68 Burdett S, Pignon JP, Tierney J, et al. Adjuvant chemotherapy for resected early-stage non-small cell lung cancer. *Cochrane Database Syst Rev*. 2015;(3):Cd011430.
69 Pellini B, Chaudhuri AA. Circulating tumor DNA minimal residual disease detection of non-small-cell lung cancer treated with curative intent. *J Clin Oncol*. 2022;40(6):567–575.
70 Pascual J, Attard G, Bidard FC, et al. ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group. *Ann Oncol*. 2022;33(8):750–768.
71 Peng M, Huang Q, Yin W, et al. Circulating tumor DNA as a prognostic biomarker in localized non-small cell lung cancer. *Front Oncol*. 2020;10:561598.