Sensitive detection of stage I lung adenocarcinoma using plasma cell-free DNA breakpoint motif profiling

Wei Guo,[a,b] Xin Chen,[c] Rui Liu,[c] Naixin Liang,[d] Qianli Ma,[e] Hua Bao,[c] Xiuxiu Xu,[c] Xue Wu,[c] Shanshan Yang,[c] Yang Shao,[c,f] Fengwei Tan,[a,b] Qi Xue,[a,b] Shugeng Gao,[a,b]* and Jie He[a]

[a]Department of Thoracic Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[b]Key Laboratory of Minimally Invasive Therapy Research for Lung Cancer, Chinese Academy of Medical Sciences, Beijing, China
[c]Geneseeq Research Institute, Nanjing Geneseeq Technology Inc., Nanjing, Jiangsu, China
[d]Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China
[e]Department of Thoracic Surgery, China-Japan Friendship Hospital, Beijing, China
[f]School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China

eBioMedicine 2022;81: 104131
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104131

Summary

Background Early diagnosis benefits lung cancer patients with higher survival, but most patients are diagnosed after metastasis. Although cell-free DNA (cfDNA) analysis holds promise, its sensitivity for detecting early-stage lung cancer is unsatisfying. We leveraged cfDNA fragmentomics to develop a predictive model for invasive stage I lung adenocarcinoma (LUAD).

Methods 292 stage I LUAD patients from three medical centers were included together with 230 healthy controls whose plasma cfDNA samples were profiled by whole-genome sequencing (WGS). Multiple cfDNA fragmentomic motif features and machine learning models were compared in the training cohort to select the best model. Model performance was assessed in the internal and external validation cohorts and an additional dataset.

Findings A logistic regression model using the 6bp-breakpoint-motif feature was selected. It yielded 98·0% sensitivity and 94·7% specificity in the internal validation cohort [Area Under the Curve (AUC): 0·985], while 92·5% sensitivity and 90·0% specificity were achieved in the external validation cohort (AUC: 0·954). It is sensitive for early-stage (100% sensitivity for minimally invasive adenocarcinoma, MIA) and <1 cm (92·9%–97·7% sensitivity) tumors. The predictive power remained high when reducing sequencing depth to 0·5× (AUC: 0·977 and 0·931 for internal and external cohorts).

Interpretation Here we have established a cfDNA breakpoint motif-based model for detecting early-stage LUAD, including MIA and very small-size tumors, shedding light on early cancer diagnosis in clinical practice.

Funding National Key R&D Program of China; National Natural Science Foundation of China; CAMS Initiative for Innovative Medicine; Special Research Fund for Central Universities, Peking Union Medical College; Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences; Beijing Hope Run Special Fund of Cancer Foundation of China.

Keywords: Cell-free DNA; Lung cancer; Early detection; Whole-genome sequencing; Fragmentomics

Introduction

Lung cancer is a leading cause of cancer-related mortality in the world, making up 25% of all cancer deaths. As an important public health issue, the chance that people will develop lung cancer in their lifetime is quite high, about 1 in 15 for a man and 1 in 17 for a woman.[1,2]

*Corresponding author at: Department of Thoracic Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China.
E-mail address: gaoshugeng@cicams.ac.cn (S. Gao).

Research in context

Evidence before this study

The emerging cfDNA analysis is promising for cancer early detection, which can benefit patients with higher survival. However, the cfDNA methodology for lung cancer early screen is often hampered by its unsatisfactory sensitivity. A literature search in PubMed was conducted from the database inception date to October 27, 2021, with the keywords "lung cancer," "stage I," "early detection," and "cfDNA." There was no publication about cfDNA-based early detection primarily targeting stage I lung cancer, which can offer the most benefit to corresponding patients. When previous models built on cfDNA methylation or fragmentation pattern were applied to stage I lung cancer patients, their performance could not suffice the early diagnosis use with sensitivities of merely 40% or even lower. Recently, cfDNA motif sequence profiles have been utilized to identify hepatocellular carcinoma patients with >80% sensitivity, holding potential for other cancer types.

Added value of this study

In this study, we evaluated the combinations of different cfDNA motif features from whole-genome sequencing data and machine learning algorithms for detecting stage I lung adenocarcinoma. A logistic regression model using the 6bp-breakpoint-motif feature demonstrated the detection sensitivity of 96·5% at 93·0% specificity, outperforming other cfDNA-based methods for stage I lung adenocarcinoma detection. In particular, our model is robust for detecting early-stage, small-size tumors, and at sequencing coverage depth down to 0·5×, promoting affordable early-stage cancer screening. We also confirmed its high performance in external validation cohorts.

Implications of all the available evidence

Our study has established a machine learning model using cfDNA breakpoint motifs for sensitive detection of stage I lung adenocarcinoma. In general, noninvasive detection of early-stage lung cancer using plasma cfDNA has attracted increasing attention and is still in progress for leveraging its performance. Epigenetic, fragmentomic, and topological analyses would shed light on its underlying biological mechanisms. Optimal cfDNA features and machine learning algorithms, together with other clinical factors, could improve the prediction power. Furthermore, including high-risk controls would leverage the model's clinical utility for the early screening in the targeted population.

Appropriate therapy, such as surgical resection for lung cancer detected at an early stage, offers the opportunity for a favorable prognosis.[3,4] Indeed, the survival rate of patients diagnosed at the early localized stage is about nine times greater than that at the late distant stage.[5] However, approximately 60% of lung cancer patients are not diagnosed until metastasis has already occurred.[6] Delayed diagnosis can be attributed to reasons including lack of early-stage symptoms, misdiagnosis due to misrepresented early symptoms, unaffordable cost, and limited access to state-of-the-art detection methods.[4] Although radiological approaches such as the Low-dose computed tomography (LDCT) scan have been recommended for early screening purposes, which can contribute to a 20% reduction in cancer-related deaths, their usage has been limited due to radiation exposure, high financial cost, and mediocre detection ability, etc.[7,8] The current situation imposes a growing need for developing accurate, easy-to-access, and cost-effective early detection methods.

The emerging liquid biopsy-based cell-free DNA (cfDNA) analysis has offered a noninvasive approach to the clinical practice of disease detection. cfDNA in the circulation is composed of extracellular DNA fragments released from cell apoptosis and necrosis,[9] therefore bearing the molecular signatures from the cell and tissue of origin.[10] In particular, circulating tumor DNA (ctDNA), as a fraction of the total cfDNA in cancer patients, represents DNA shed from tumor cells.[11] Detection of tumor somatic mutations can serve as biomarkers to distinguish ctDNA and nontumorous cfDNA.[12] However, with sensitivities as low as 40% for early-stage lung cancer, the performance of ctDNA mutation calling-based detection method cannot suffice the requirement of clinical application.[13] Recently, epigenetic modifications and fragmentomic signatures on ctDNA are known to manifest cancer pathobiology, and have opened new avenues for early diagnosis of lung cancer.[14] To date, ctDNA signatures, including methylation and fragment size patterns, have been deployed in lung cancer predictive models.[15,16] However, the current methylation signature-based approach was not ideal for stage I lung cancer diagnosis with a sensitivity of only 25%.[16] On the other hand, the development of the fragment size approach has mainly relied on the later-stage lung cancer patients.[15,17] Therefore, their detection powers exhibited bias against early-stage cases that can receive the greatest clinical benefits.

The profile of cfDNA end motifs represents another class of biomarkers for liquid biopsy in oncology and holds promise in lung cancer detection. Recently, researchers have revealed the tumor-associated cfDNA preferred end coordinates by the comparison between hepatocellular carcinoma (HCC) patients and non-cancer patients.[18] A follow-up study demonstrated that patients with HCC showed a preferential pattern of cfDNA 4mer end motifs compared to non-HCC subjects with a differentiating sensitivity of over 80% at >90% specificity.[19] The preferred end motifs were also observed in other cancers, including lung cancer. These preliminary findings suggested that the motif-based approach could outperform other fragmentomic metrics such as cfDNA fragment size in identifying various cancers.[19] Mechanistically, changes in chromatin

accessibility and nuclease activity could have global impacts on the cfDNA end motifs.[20] Chromatin accessibility is important for the fragmentation of cfDNA and contributes to their preferred ends.[18] The ATAC-seq experiment in different human primary cancers has identified cancer type-specific chromatin accessibility landscape.[21] On the other hand, multiple DNA nucleases have exerted combinatorial effects on the cfDNA end motifs in the mouse model, generating preferential sequences both intracellularly and in circulation.[20] Meanwhile, the downregulation of the DNA nuclease *DNASE1L3* in HCC may contribute to the alterations in cfDNA end motifs.[19] Therefore, the preferred end motifs in cancer can be attributed to tumor-specific nucleosome positioning and nuclease activity, which may implicate cancer classification and origin identification. However, their application for the early detection of lung cancer needs to be systematically verified.

In this study, we established a robust cfDNA fragmentomic machine learning model for early lung adenocarcinoma (LUAD) detection using whole-genome sequencing (WGS) data. The cfDNA genomic breakpoint motifs that profile sequences upstream and downstream of the cleavage sites were extracted from WGS of stage I LUAD patients and healthy controls. We incorporated the breakpoint motifs into the predictive model with the logistic regression algorithm. Our model could sensitively detect stage I LUAD, including small-size (< 1cm) tumors. When the sequencing depth was down to 0·5×, the model performance remained consistent. Furthermore, we verified its performance in an external validation cohort and an additional dataset of benign nodule cases. Hence, this model offers a promising strategy for developing early lung cancer diagnosis and management in clinical practice.

Methods

Patient cohorts and sample collection

This study primarily enrolled 292 stage I LUAD patients from three medical centers in China (Center I: Cancer Hospital, Chinese Academy of Medical Sciences; Center II: Peking Union Medical College Hospital; Center III: China-Japan Friendship Hospital) and 230 healthy volunteers from routine physical checks at Center I (Supplementary Tables 1 and 2). The LUAD patients included invasive adenocarcinoma (ADC) and minimally invasive adenocarcinoma (MIA). In addition, to assess our model performance on benign lung disease, we retrieved plasma samples used in other studies from our plasma biobank for analysis. This additional dataset included 52 noncancer participants with known lung nodule status by computerized tomography (CT) scan (Supplementary Table 3). We performed plasma sample collection, shipping and storage, cfDNA extraction, library preparation, and WGS analysis uniformly as described in Supplementary Materials and Methods. In brief, the blood draw of the participants was performed from 2019 to 2021. The steps of cfDNA extraction, library preparation, and WGS were performed immediately after each other in batches by the Clinical Laboratory Improvement Amendments (CLIA)- certified and College of American Pathologists (CAP)-accredited clinical testing laboratory (Nanjing Geneseeq Technology Inc., China). The study was approved by the ethics committee of the National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College (Approval No. NCC3328) and followed the ICH-GCP Guidelines. Written informed consent was obtained from all participants.

cfDNA extraction and whole-genome sequencing

We performed plasma sample collection, cfDNA extraction followed by WGS as described in Supplementary Materials and Methods. Briefly, the venous blood samples were collected during routine physical checks (healthy volunteers) or preoperatively (cancer patients). All samples were collected, shipped, and processed uniformly. 5-10 ng of plasma cfDNA per sample was subject to PCR-free WGS library construction with the KAPA HyperPrep Kit (Roche). The libraries underwent paired-end sequencing on NovaSeq6000 platform (Illumina) to a mean sequencing depth of 10·92 × (range 5·35×-24·81×). To minimize bias, the sample operating team was blinded to the case/control status of the samples in the whole process.

Due to the varied WGS depths among samples (range 5·35×-24·81×), we down-sampled all of them to 5× for model optimization. This strategy allowed us to eliminate the potential impact of coverage difference on the model while maintaining the maximal sample inclusion. The optimized model was then validated using the WGS data of raw sequencing depth or down-sampling to 4×, 3×, 2×, 1×, and 0·5×.

Bioinformatic analysis and modeling

Raw sequencing data processing was carried out as described in Supplementary Materials and Methods. For sample classification, we built in-house scripts that extracted the features of cfDNA 4bp end motif, 6bp end motif, and 6bp breakpoint motif from the WGS data. In brief, the cfDNA 4bp end motif referred to the 5′ end 4 bp sequences reported by Jiang et al.[19] cfDNA 6bp end motif was the extension of the 4bp end motif to 6bp sequences. cfDNA 6bp breakpoint motif was defined as the 3bp extensions to both directions of the aligned cfDNA 5′ breakpoints in the human reference genome hg19 (Supplementary Figure 1). The frequency of every particular motif over the total motifs was computed for each sample (Supplementary Materials and Methods). Three representative machine learning approaches- logistic regression (with elastic net

regularization), deep learning, and XGBoost, with different advantages (Supplementary Materials and Methods), were individually tested for model construction. Using samples solely derived from the healthy and LUAD participants of the training cohort, we trained the classifiers in the machine learning algorithms with the motif feature frequencies and generated the models for predicting cancer scores of each sample. Notable, all the validation datasets remained untouched during model training. The scale of cancer score ranges from 0 to 1, and a higher score value indicates a higher cancer probability. After we finalized the parameters in the training cohort, the models were subsequently applied to the validation datasets to generate the cancer prediction score for each validation sample and evaluate model performance. For the assessments, we compared the AUC values in the validation cohorts of different models and their sensitivity/specificity with fixed specificity in the internal validation cohort (Supplementary Materials and Methods). After the assessment, we selected the best performance model for downstream analyses.

The *in silico* dilution of tumor cfDNA was conducted by mixing the 5× WGS data of cancer and noncancer samples in the designated ratio while maintaining the average coverage of the resultant sample the same as the original undiluted cancer cfDNA sample. The details are described in Supplementary Materials and Methods.

Statistical analysis

For statistical analysis, the receiver operating characteristic (ROC) curves were generated using the pROC package (v. 1.17.0.1). Based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) of cancer prediction, we calculated the sensitivity [TP/(TP+FN)], specificity [TN/(TN+FP)], positive (PPV) [TP/(TP+FP)] and negative predictive values (NPV) [TN/(TN+FN)], accuracy [(TP+TN)/(TP+FP+TN+FN)], as well as the corresponding 95% confidence intervals, using the epiR package (v 2.0.19) in R (v 4.0.3). Propensity score matching of validation cohorts was performed using the MatchIt package (4 4.2.0) in R (v 4.0.3). The preProcess (caret version 6.0-88) function was used to calculate Z-scores from motif frequencies. The cfDNA tumor fraction was calculated using ichorCNA.[22] The hierarchical clustering analysis was generated using the ComplexHeatmap package (3.13) in R. The Fisher's exact test was performed using GraphPad, and the Wilcoxon test and t-test were performed using R.

Role of funding source

The funding agencies had no role in study design, in the collection, analysis and interpretation of data, in the writing of the report and in the decision to submit the paper for publication.

Results

Cohort characteristics

We included 522 participants with previously untreated stage I LUAD (n = 292) or without cancer (n = 230) (Figure 1a) for model construction and validation. The participants from Center I were randomly assigned to the training [150 cancer patients (115 ADC and 35 MIA; 144 Stage IA and 6 Stage IB; tumor size < 1 cm: 57); 115 healthy controls], and internal validation [102 cancer (78 ADC and 24 MIA; 99 Stage IA and 3 Stage IB; tumor size < 1 cm: 44); 75 healthy] cohorts. The 40 cancer patients (35 ADC and 5 MIA; 36 Stage IA and 4 Stage IB; tumor size < 1 cm: 14) from Centers II and III were assigned to the external validation cohort together with 40 healthy controls. The details of sample characteristics in each cohort are summarized in Supplementary Table 4.

cfDNA fragmentomics feature and machine learning model selection for predictive model construction

We carried out the predictive model construction and selection by testing three cfDNA motif features in combination with three machine learning approaches. The cfDNA fragmentomic features include the 4bp end motif, 6bp end motif, and 6bp breakpoint motif (Supplementary Figure 1). To achieve clinically useful detection power, we started by adapting the recently published cfDNA 5′ end 4bp motif feature and performed further feature adjustment. We speculated the extended cfDNA 5′ end 6bp motif feature (4^6) could confer more information than the 4bp one (4^4). Furthermore, as different nucleases manifested varied preferences for the last one or two nucleotides at the end of cfDNA molecules,[20] we inferred that a broader context around the genomic cleavage sites as a whole would contribute to the nuclease cutting site selection. Therefore, the profile of the 6bp breakpoint motif was also included in the test. The three machine learning algorithms we tested are logistic regression, deep learning, and XGBoost. After being tested in the internal validation cohort, the logistic regression + 6bp breakpoint motif model reached an AUC of 0·985 and a sensitivity of 98·0% at the 94·7% specificity outperformed other combinations (Supplementary Table 5). In the external validation cohort, the logistic regression + 6bp breakpoint motif model also excelled in AUC and sensitivity/specificity (Supplementary Table 5). Based on the assessments, we determined that the logistic regression model of 6bp breakpoint motifs constantly yielded higher detection ability than other models, and thus pursued this one for detailed performance evaluation (Figure 1b). When the two study cohorts were combined, the predictive model has achieved an AUC of 0·977 and a sensitivity of 96·5% at 93·0% specificity for detecting stage I LUAD, exceeding any other

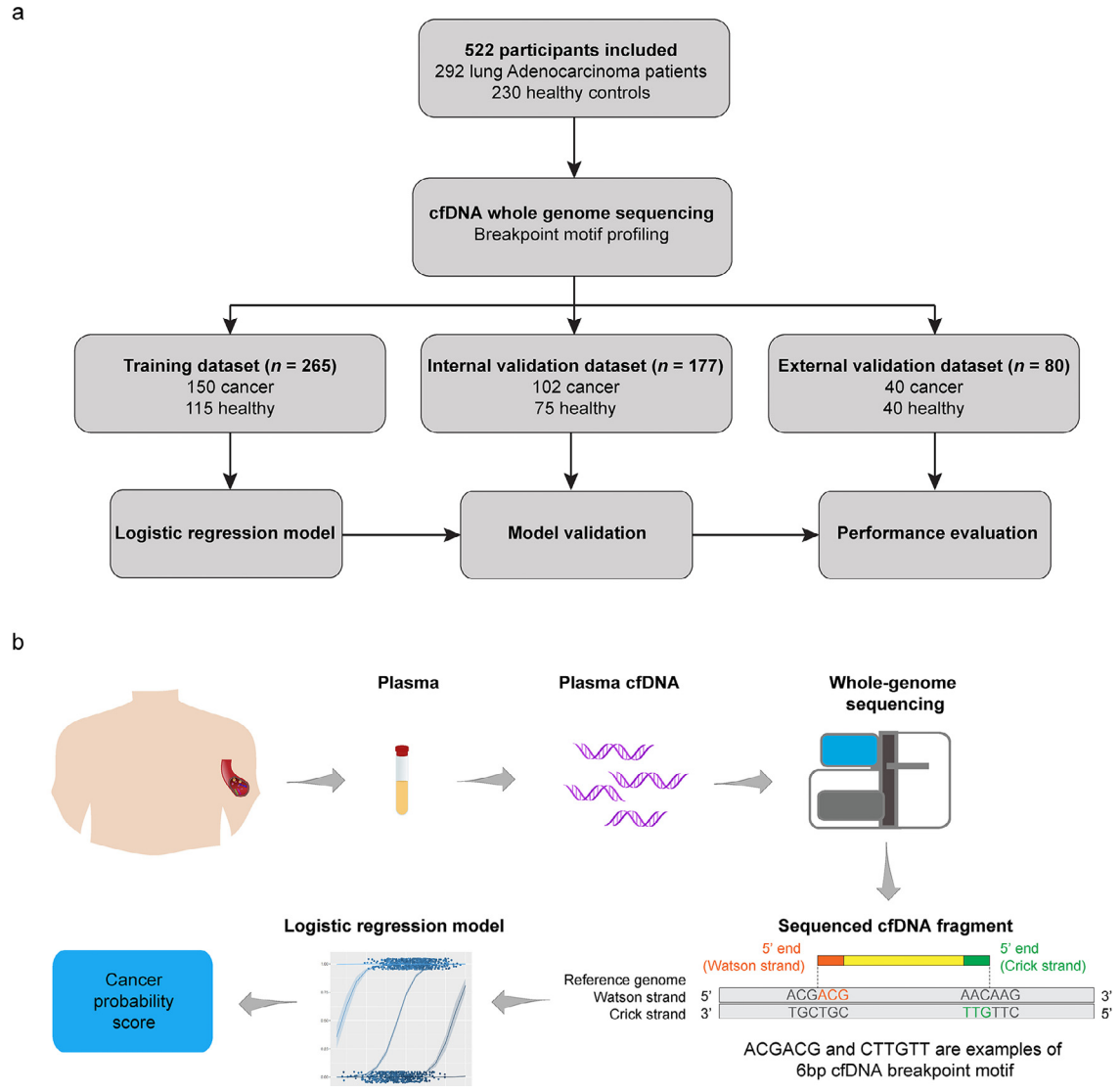

Figure 1. Schematic illustration of the study. (a) Study design. A total of 522 participants (cancer 292, healthy 230) were included in this study. Whole-genome sequencing of plasma cfDNA was performed, and their cfDNA breakpoint motif was profiled. 265 participants (cancer 150, healthy 115) were allocated to training for building the logistic regression algorithm-based machine learning model. 177 participants (cancer 102, healthy 75) were allocated to internal validation for confirming the model performance and determining the cutoff score. 80 participants (cancer 40, healthy 40) were allocated to external validation for evaluating model performance. **(b)** Schematic diagram of cancer probability score determination. Plasma cfDNA was extracted from the participant's plasma sample and subject to whole-genome sequencing. The sequencing reads that were mapped to a human reference genome were used to determine the 6-nucleotide sequence (i.e., a 6bp breakpoint motif) on each 5′ fragment end (Watson and Crick strands) of plasma cfDNA relative to the genome. The genome-wide breakpoint motif profile was then applied in the logistic regression algorithm to calculate the participant's cancer probability score.

combination (Figures 2a and 2b). In addition, our test indicated that different combinations of machine learning algorithms and motif features could render varied detection performance between study cohorts (Figures 2a and 2b, Supplementary Table 5), highlighting the importance of optimizing model parameters.

Furthermore, hierarchical clustering analysis was used to identify the different characteristics of plasma cfDNA 6bp breakpoint motifs between the cancer patients and healthy controls (Figure 2c). In the training model, we identified 65 6bp breakpoint motifs with non-zero coefficients (Supplementary Materials and Methods), and applied these motifs for the hierarchical clustering analysis in the combined internal and external validation datasets. The Z-scores of selected breakpoint motifs were calculated from the motifs'

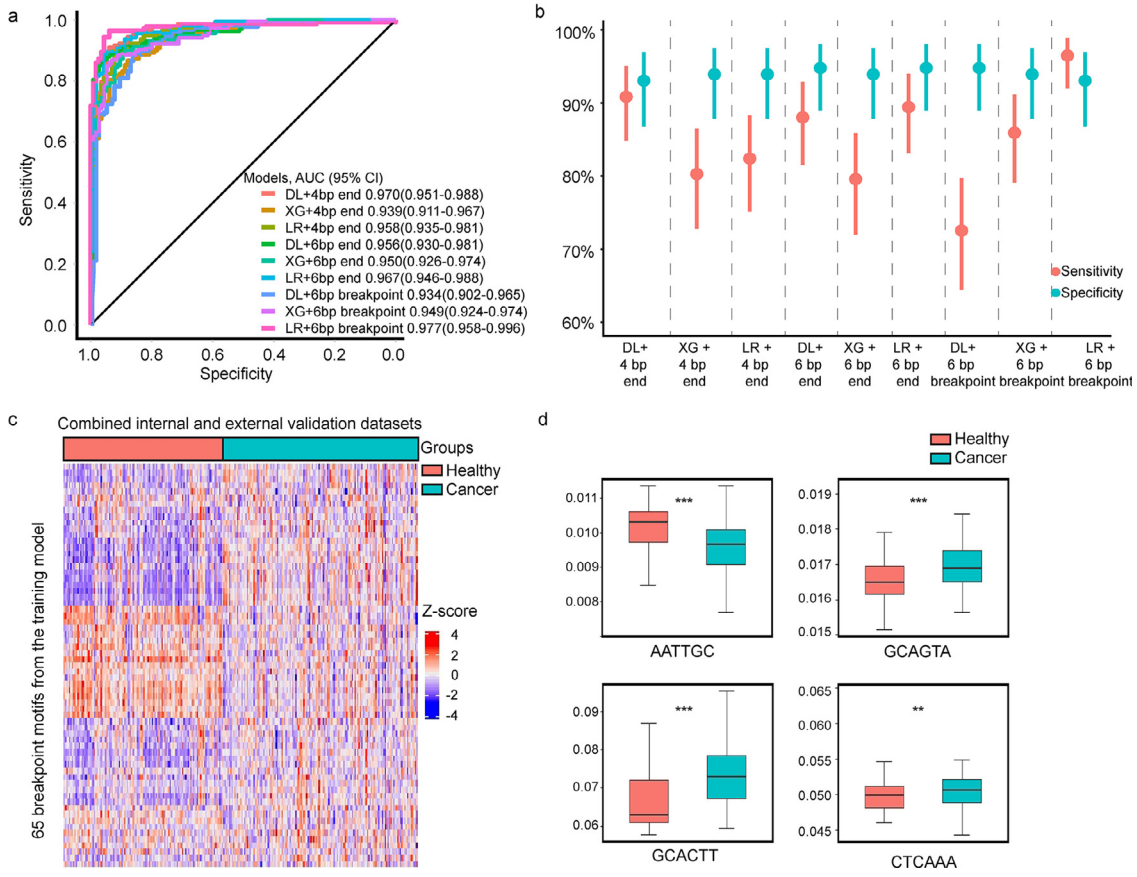

Figure 2. Identification and evaluation of the 6bp breakpoint motif logistic regression model for cancer prediction. (a) ROC curve evaluating the performance of predictive models built on different cfDNA motif features and machine learning algorithms in distinguishing early lung cancer from healthy subjects for the combined validation cohorts (DL: deep learning; XG: XGBoost; RL: logistic regression); **(b)** the sensitivities and specificities of different predictive models from (a) in the combined validation cohorts; **(c)** Heat map analyzing frequencies of the 65 breakpoint motifs in the training model with non-zero coefficients between healthy and cancer subjects in the validation cohorts. The data are row-normalized; **(d)** Box plot showing frequencies between healthy and cancer subjects in the validation cohorts for the four representative motifs contributing most significantly to the model (*: $0{\cdot}01 < p < 0{\cdot}05$; **: $0{\cdot}001 < p < 0{\cdot}01$; ***: $p < 0{\cdot}001$, Wilcoxon rank-sum test).

frequencies in the plasma cfDNA of each participant. Consistently, we found that the motifs identified from the training cohort also tend to form different clusters between lung cancer patients and healthy controls in the combined validation cohorts. Four representative breakpoint motifs that contributed most significantly to the machine learning model in the training cohort were further analyzed. Significant differences in motif frequencies were found for all four motifs between the LUAD and healthy subjects in the validation cohorts (Figure 2d). In detail, the motif AATTGC frequency showed a significant decrease in the patients, while the frequencies of the other three motifs, GCAGTA, GCACTT, and CTCAAA, are significantly higher in the cancer subjects (Wilcoxon rank-sum test).

Built on 5× coverage WGS data, our model exhibited superior detection power, and the model AUC is 0·985 (95% CI: 0·965–1·000) and 0·954 (95% CI: 0·909–1·000) in the internal and external validation cohorts, respectively (Figure 3a). Its prediction sensitivity reached 98·0% (95% CI: 93·1%–99·8%) at the specificity of 94·7% (95% CI: 86·9%–98·5%) in the internal validation cohort (Table 1). When applying the sample cancer score of 0·3275 (Figure 3b) that granted 94·7% specificity in the internal validation cohort to the external validation cohort, the model is consistently sensitive with a 92·5% sensitivity (95% CI: 79·6%–98·4%) at the specificity of 90·0% (95% CI: 76·3%–97·2%) (Table 1). As shown in Figure 3b, the lung cancer subjects were associated with higher cancer scores than the healthy subjects. Specifically, the cancer score values of cancer subjects in the internal (median, 0·8392; 95% CI: 0·3670–0·9760) and external (median, 0·6419; 95% CI: 0·1777–0·9407) validation cohorts were found to be significantly higher than corresponding healthy controls

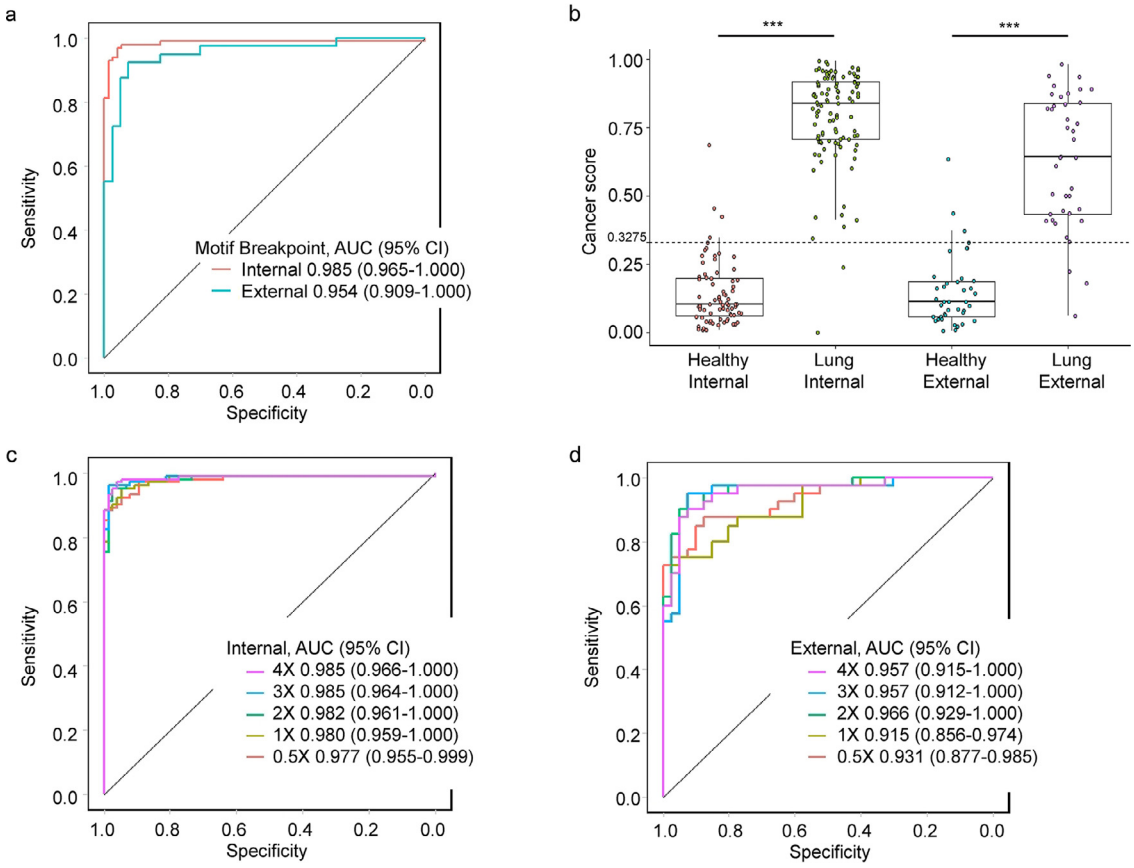

Figure 3. Development and evaluation of the predictive model in internal and external validation cohorts. (a) ROC curve evaluating the overall performance of the predictive model all using 5× coverage WGS data in distinguishing early lung cancer from healthy subjects for the internal and external validation cohorts; **(b)** The boxplot showing the distribution of cancer scores based on the 5× WGS model in the patient and control groups of the validation cohorts. The 95% specificity cutoff score for the internal validation set is 0·3725 (*: $0{\cdot}01 < p < 0{\cdot}05$; **: $0{\cdot}001 < p < 0{\cdot}01$; ***: $p < 0{\cdot}001$, t-test); **(c)** and **(d)** ROC curves evaluating the 5× WGS-based model performance using low-coverage (4×-0·5×) WGS data in internal and external validation cohorts.

(internal: median, 0·1019; 95% CI: 0·0097–0·4290; external: median, 0·1127; 95% CI: 0·0097–0·4410, t-test), respectively. In addition, we performed the principal component analysis (PCA) of the motif features and cancer score analysis of all the participants to ensure the healthy and LUAD samples had no apparent batch effects due to different medical centers or sample preparations.

To evaluate the potential model dependence on the confounding factors of age and sex between the disease and healthy groups, we preprocessed the internal and external validation cohorts for propensity score matching with the R package "MatchIt". The matched internal (45 cancer patients and 42 healthy controls) and external (10 cancer patients and 9 healthy controls) validation datasets were then classified using the same model. The AUC of the matched validation cohorts (0·967, 95% CI: 0·934–1·000) was not statistically different from the unmatched cohorts (t-test), suggesting an age- and sex-independence nature of our predictive model.

Evaluating the robustness and detection sensitivity of the predictive model

With the 6bp breakpoint motif machine learning model, we revisited the raw coverage WGS data. Using raw coverage WGS for model training and disease prediction, the model yielded the AUC of 0·982 (95% CI: 0·961–1·000) and 0·961 (95% CI: 0·920–1·000) in the internal and external validation cohorts, respectively (Supplementary Figure 2a and Supplementary Table 6). We also tested the raw coverage WGS data in the 5× WGS data-trained model, and the resulting models essentially showed consistent detection power no matter whether raw coverage or 5× coverage WGS data were used for modeling and predicting (Supplementary Figure 2b and Supplementary Table 7). A key attribute of affordable NGS-based cancer detection is the ability to differentiate disease from healthy subjects using the low sequencing depth data. Therefore, we pursued the rest of our evaluation with 5× coverage-based WGS data and model unless noted otherwise.

Internal validation cohort		Actual	
		Cancer	Healthy
Predicted	Cancer	100	4
	Healthy	2	71
Sensitivity (95% CI)		98·0% (93·1–99·8%)	
Specificity (95% CI)		94·7% (86·9–98·5%)	
PPV (95% CI)[†]		96·2% (90·4–98·9%)	
NPV (95% CI)[‡]		97·3% (90·5–99·7%)	
Accuracy (95% CI)		96·6% (92·8–98·7%)	
External validation cohort		**Actual**	
		Cancer	**Healthy**
Predicted	**Cancer**	**37**	**4**
	Healthy	**3**	**36**
Sensitivity (95% CI)		92·5% (79·6–98·4%)	
Specificity (95% CI)		90·0% (76·3–97·2%)	
PPV (95% CI)		90·2% (76·9–97·3%)	
NPV (95% CI)		92·3% (79·1–98·4%)	
Accuracy (95% CI)		91·3% (82·8–96·4%)	
Combined validation cohorts		**Actual**	
		Cancer	**Healthy**
Predicted	**Cancer**	**137**	**8**
	Healthy	**5**	**107**
Sensitivity (95% CI)		96·5% (92·0–98·8%)	
Specificity (95% CI)		93·0% (86·8–96·9%)	
PPV (95% CI)		94·5% (89·4–97·6%)	
NPV (95% CI)		95·5% (89·9–98·5%)	
Accuracy (95% CI)		94·9% (91·5–97·3%)	

***Table 1*: The diagnostic performance of the predictive model in the validation cohorts.**
[†] Positive predictive value.
[‡] Negative predictive value.

Next, we sought to identify the robustness of the model and evaluated its performance at even lower sequencing depth by further gradually down-sampling the WGS data to ~ 0·5×. Upon down-sampling WGS coverages to 4×, 3×, 2×, 1× and 0·5×, we found their AUC values remained consistently high in both internal (> 0·97) and external (> 0·91) validation cohorts (Figures 3c and 3d). Under the scrutiny of the model performance, the coverages of 1× and below appeared to cause a slight decrease in the external validation cohort but still yielded usable detection sensitivities (1×: 75·0%, 95% CI: 58·8%–87·3%; 0·5×: 77·5%, 95% CI: 61·6%–89·2%;) at specificities of 95·0% and 92·5%, respectively (Supplementary Table 8).

Our lung cancer patients in the validation cohorts are solely composed of stage I cases with low cfDNA tumor fraction (mean: 1·1%) estimated by the ichorCNA algorithm.[22] To further determine the detection sensitivity of our predictive model, we mixed the WGS data of every cancer patient in the validation cohorts (n = 142) with 20 random noncancer participants' data in the 3:1 and 1:1 ratios, resulting in the *in silico* dilution of the original tumor DNA fraction to 75% and 50%. The virtually diluted cancer samples were analyzed by our predictive model to generate the cancer scores and prediction. This experiment was conducted randomly in triplicates. We observed a downward trend of cancer score distribution and resultant model sensitivity accompanying the decrease of tumoral DNA fraction (Supplementary Figure 3). Thus, our model is more robust with a higher tumor fraction, while it maintained high sensitivity at 83·8% (95% CI: 82·4%–85·2%) with the 1:1 ratio dilution.

Performance of the predictive model in identifying early-stage LUAD

We further examined the model performance in different lung cancer subgroups. As depicted in Figure 4 and Table 2, detection sensitivity is consistent between

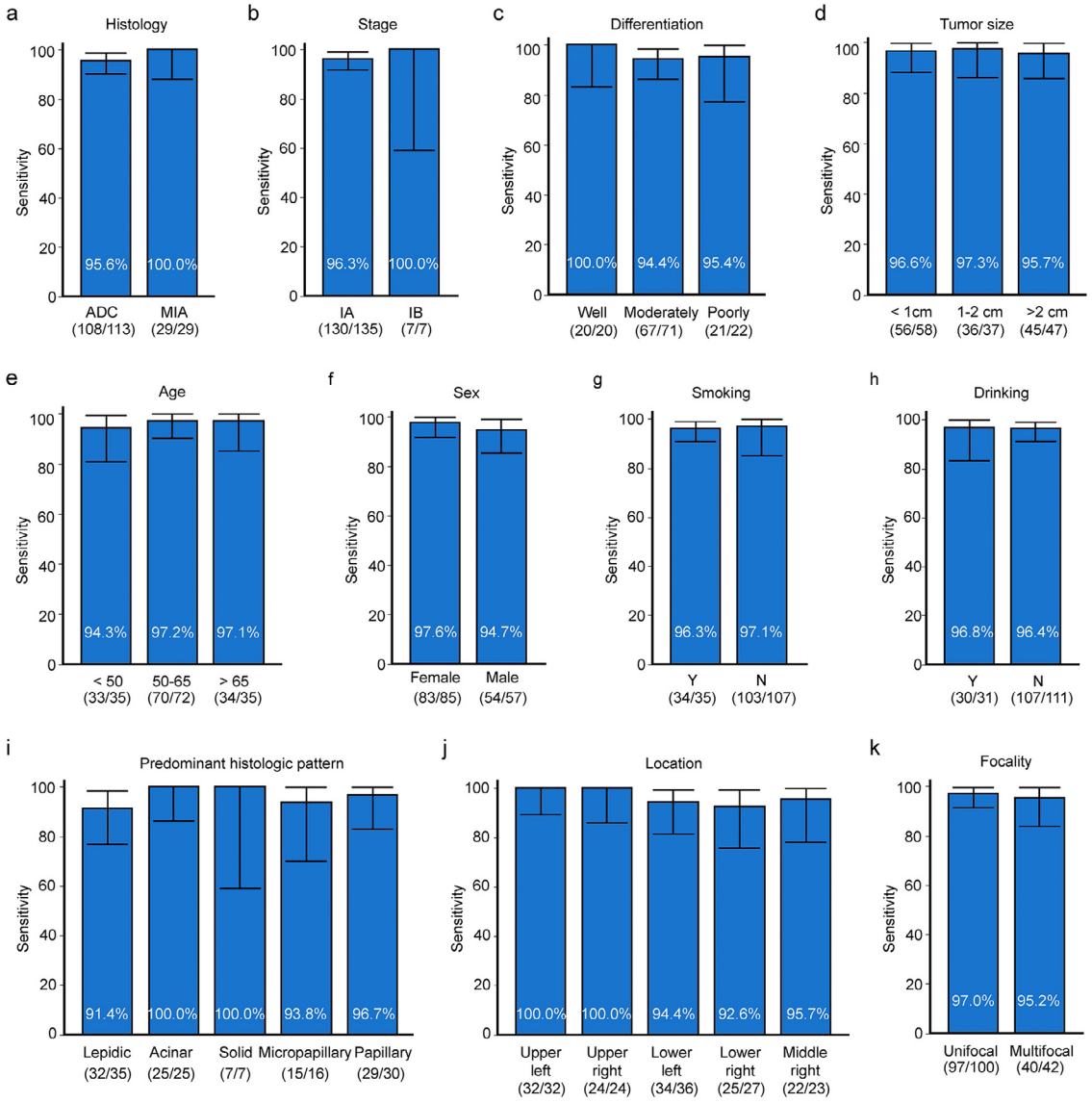

Figure 4. The model's diagnostic sensitivities in different subgroups of the combined validation cohorts at 95% specificity. The sensitivities (%) were calculated with 95% confidence interval as indicated by the bars for subgroups of **(a)** histology, **(b)** stage, **(c)** differentiation level, **(d)** tumor size, **(e)** age, **(f)** sex, **(g)** smoking, **(h)** drinking, **(i)** predominant histologic pattern, **(j)** tumor location and **(k)** focality. The numbers in the parentheses represent the true positive and total cases in each subgroup category.

categories within different subgroups of lung cancer patients, and shows no significant difference between different categories (Fisher's exact test, p values are all > 0·1). Notably, for patients in both internal and external validation cohorts, our model showed high detection sensitivity on early pathological features, including minimally invasive adenocarcinoma (MIA) (100·0%, 95% CI: 88·1%–100·0%), tumors at stage IA (96·3%, 95% CI: 91·6%–98·8%), and of small size (<1 cm) (96·6%, 95% CI: 88·1%–99·6%) at the 95% specificity. With the high performance in identifying stage I LUAD, the output of our predictive model is also consistent with the disease development. By profiling the patients' cancer scores in the validation cohorts grouped by their tumor stages or sizes, we observed an upward trend of score distribution from stage IA to stage IB or from small to bigger tumors, respectively (Supplementary Figures 4a and 4b). Furthermore, our model held consistent detection abilities on both aggressive and less aggressive tumors based on the differentiation grades (well-differentiated: 100·0%, 95% CI: 83·2%–100·0%; moderately-differentiated: 94·4%, 95% CI: 86·2%–98·

Cohort		Internal	External
Histology	ADC[†]	91·4% (76·9–98·2%)	97·4% (91·0–99·7%)
	MIA[‡]	100·0% (47·8–100·0%)	100·0% (85·8–100·0%)
Stage	IA	91·7% (77·5–98·2%)	98·0% (92·9–99·8%)
	IB	100·0% (39·8–100·0%)	100·0% (29·2–100·0%)
Differentiation level	Well	100·0% (66·4–100·0%)	100·0% (71·5–100·0%)
	Moderate	83·3% (58·6–96·4%)	98·1% (89·9–100·0%)
	Poor	100·0% (63·1–100·0%)	92·9% (66·1–99·8%)
Tumor size	< 1 cm	92·9% (66·1–99·8%)	97·7% (88·0–99·9%)
	1 – 2 cm	90·0% (55·5–99·7%)	100·0% (87·2–100·0%)
	> 2 cm	93·8% (69·8–99·8%)	96·8% (83·3–99·9%)
Focality	Unifocal	95·5% (77·2–99·9%)	97·4% (91·0–99·7%)
	Multifocal	88·9% (65·3–98·6%)	100·0% (85·8–100·0%)
Sex	Female	94·4% (72·7–99·9%)	98·5% (92·0–100·0%)
	Male	91·0% (70·8–98·9%)	97·1% (85·1–99·9%)
Age	< 50	90·9% (58·7–99·8%)	95·8% (78·9–99·9%)
	50 – 65	93·8% (69·8–99·8%)	98·2% (90·4–100·0%)
	> 65	92·3% (64·0–99·8%)	100·0% (84·6–100·0%)
Predominant histologic pattern	Lepidic	78·6% (49·2–95·3%)	100·0% (83·9–100·0%)
	Acinar	100·0% (47·8–100·0%)	100·0% (83·2–100·0%)
	Solid	100·0% (29·2–100·0%)	100·0% (39·8–100·0%)
	Micropapillary	100·0% (47·8–100·0%)	90·9% (58·7–99·8%)
	Papillary	100·0% (63·1–100·0%)	95·5% (77·2–99·9%)
Location	Upper left	100·0% (59·0–100·0%)	100·0% (78·2–100·0%)
	Upper right	100·0% (54·1–100·0%)	100·0% (76·8–100·0%)
	Lower left	85·7% (42·1–99·6%)	95·0% (75·1–99·9%)
	Lower right	85·7% (42·1–99·6%)	93·3% (68·1–99·8%)
	Middle right	87·5% (47·3–99·7%)	100·0% (76·8–100·0%)
Smoking	Y	90·0% (55·5–99·7%)	100·0% (86·3–100·0%)
	N	93·3% (77·9–99·2%)	97·4% (90·9–99·7%)
Drinking	Y	88·9% (51·8–99·7%)	100·0% (84·7–100·0%)
	N	93·5% (78·6–99·2%)	97·5% (91·3–99·7%)

***Table 2*: The diagnostic sensitivities of the predictive model in different lung cancer patient subgroups of the validation cohorts at the 95% specificity.**

[†] Invasive lung adenocarcinoma.

[‡] Minimally invasive adenocarcinoma.

4%; poorly-differentiated: 95·5%, 95% CI: 77·2%–99·9% at the specificity of 95%). Finally, the model showed consistent and high sensitivities in identifying LUAD regardless of gender, age, tumor location, focality, histologic pattern, and potentially risky behaviors such as cigarette smoking and alcohol drinking (Table 2).

The additional dataset of noncancer participants with known benign lung nodule status (Supplementary Table 3) allowed us to extend our assessment to subjects with benign lung disease conditions. When applying the same cutoff score of 0·3275 to it (Supplementary Figure 5), the model is consistently credible with a detection specificity of 94·2% (95% CI: 84·1%–98·8%). It is worth noting that the age distribution (mean: 59·5 years; range: 52–75, Supplementary Table 3) of the additional noncancer subjects resembles that of the cancer subjects in the validation cohorts (Supplementary Table 4), and the specificity of our model is not affected by the difference in age. We scrutinized the noncancer participants for their benign lung nodule status and found our model's specificity is not affected by the presence of the benign nodule (p value=0.2723, Fisher's exact test).

Discussion

Here we focused on the LUAD early detection and demonstrated that a machine learning model of plasma cfDNA represents a promising approach to differentiate early-stage patients from noncancer subjects. Our method utilized the cfDNA fragmentomic breakpoint motifs derived from the WGS data. As demonstrated by us and others, the WGS-based approaches theoretically allow us to evaluate the distribution and frequency of motifs and sizes of any naturally

occurring cfDNA fragment. Thus, the findings comprehensively represent a large number of tumor-derived changes.[15,23]

Several existing studies have utilized cfDNA methylation as the classifier for detecting early-stage lung cancer. However, the cfDNA methylation-based approaches often use immunoprecipitation enrichment or targeted enrichment, which only analyze a relatively limited number of loci.[24–26] Furthermore, those models were either not evaluated for their performance on stage I patients that will receive the most clinical benefits from early detection,[24] or may experience overfitting as their performance dropped drastically in independent validation.[25] The study that investigated ctDNA detection in early-stage lung cancer with targeted methylation sequencing reached suboptimal sensitivity below 60% for identifying adenocarcinoma *in situ* (AIS) and MIA at 92.8% specificity, likely due to a lack of detected signals.[27] Recently, Liu et al. applied cfDNA methylation patterns to stage I lung cancer patients but only reported an approximately 25% sensitivity of distinguishing such patients and noncancer controls.[16] Liang et al. have deployed deep methylation sequencing aided by machine learning to improve this classifier. As a result, they reached the sensitivity of 52% and 64% for stage IA and IB lung cancer, respectively, at the 96% specificity,[26] which is still inferior to the performance of our model. Similarly, the signatures of cfDNA fragment size can be used to identify later-stage lung cancer,[15,17] but the performance of this approach on early-stage lung cancer is also suboptimal for the development of early detection.

As cleavage and fragmentation of cfDNA are nonrandom processes, the cleavage site preference can be associated with tissue sources, disease status, chromatin accessibility, and nuclease activities.[14,18,19,28] Thus, we postulated that the feature of cfDNA end motif is promising to achieve clinically usable detection performance as demonstrated by the studies in HCC,[19,29] and set out to explore its application in the predictive model for early-stage LUAD. Leveraging the advantages of cfDNA sequence motifs for tumor markers,[19] our breakpoint motif model outperformed the models mentioned above using other ctDNA features in distinguishing early LUAD and noncancer subjects. Hence, the breakpoint motif feature may confer detection power from the cleavage site preference in the genomic context, which warrants further investigation.

The cfDNA breakpoint motif machine learning model using the logistic regression algorithm has reached the sensitivity of 96·5% at the 93·0% specificity in the validation cohorts, exceeding the detection sensitivities of other reported lung cancer early detection models.[13,16] More importantly, we tested our model with the emphasis on the understudied MIA and very small-size tumors in stage IA LUAD. Remarkably, our training cohort was solely built on stage I LUAD, including 96·0% stage IA and 23·3% MIA samples, to facilitate the detection of the early-stage subjects. Our model exhibited a consistently high-level capacity for distinguishing these patients from noncancer subjects in both internal and external validation cohorts. It also appears robust for correctly identifying relatively old noncancer samples with benign nodules. Studies by others have proposed the detection of plasma cfDNA from AIS and even earlier-stage lung lesions, suggesting the potential of using cfDNA for very early-stage lung cancer,[30,31] but the additional application of cfDNA in the early detection is still to be explored. Furthermore, the profile of the 65 breakpoint motifs selected by our model revealed apparent frequency changes between cancer and noncancer cases in different datasets, which would support its broad application for detecting early-stage LUAD.

We took multiple approaches to validate the performance of our predictive model. The mean ages of cancer subjects are higher than healthy volunteers in the training and internal/external validation cohorts, as there are more older cancer patients (Supplementary Table 4). The impact of age on prediction was ruled out by confirming the model performance in the additional dataset containing older noncancer participants (mean age: 59·5 years; range: 52–75, Supplementary Table 4). Meanwhile, the model accurately differentiated cancer and noncancer cases in multiple cohorts with varied sex ratios. Thus, our model is likely independent of the confounding factors of age and sex. In addition, we performed *in silico* dilution of our cancer samples (Supplementary Figure 3) and profiled the cancer score distribution in subgroups by tumor stage or size (Supplementary Figure 4). The detection robustness increases at higher tumor fraction, as well as later disease stage, and larger tumor size, consistent with the findings of other studies.[16,23,32] Collectively, these validations reassured us of the appropriateness of our model. As early diagnosis is a key to better treatment, while most lung cancer is still diagnosed at the later metastatic stage,[5,6] our predictive model holds great promise for improving treatment outcomes.

We attempted to develop a convenient and reliable way to promote early detection. Several groups sought to leverage model performance by integrating different genomic features from multiple platforms and sequencing methods,[29,30] which would drastically increase the cost and impede their wide application in practice. On the contrary, the method we presented could employ low-depth WGS data to yield clinically useful results even at the coverage of 0·5×, therefore offering a more affordable solution to implementing early detection and eliminating the inequity in lung cancer treatment.

The U.S. Preventive Services Task Force (USPSTF) recommended annual screening for lung cancer with LDCT in the high-risk population for a moderate net benefit.[33] The imaging methods play an irreplaceable role in lung cancer diagnosis by identifying the imaging

features of the lesions and facilitating treatment decisions. However, the relatively high false-positive rates of LDCT screening could lead to the risks such as invasive procedures, overdiagnosis, increases in patient anxiety during follow-up, and radiation exposure.[34] Moreover, a significant portion of lung cancer occurs in lower-risk populations, while the USPSTF does not recommend LDCT for this population due to the potential risk.[23] We expected our noninvasive plasma cfDNA method to complement the imaging method. Further studies are needed to determine how cfDNA-based liquid biopsy tests could facilitate lung cancer screening programs and promote health equity in lung cancer.

There are several limitations of this study. Despite its superior performance for detecting stage I LUAD in this study, the mechanism underlying cfDNA breakpoint motifs has not been fully understood. Chromatin accessibility is important for cfDNA fragmentation and fragment end preference.[18] Studies on different human primary cancers have revealed cell-specific nucleosome positioning landscapes.[21] Thus, the changes in cancer chromatin accessibility could cause a global effect on cfDNA end motifs for cancer detection and classification. Meanwhile, intracellular and extracellular DNA nucleases function synergistically in generating plasma cfDNA end motifs.[20] Among them, the change of *DNASE1L3* could affect cfDNA end motifs in HCC.[19] Therefore, the cfDNA end motifs are likely under combinatorial control, and their mechanism warrants further investigation. We are currently pursuing a comprehensive investigation of the most significant breakpoint motifs to optimize the model's performance. This is a retrospective study. Although model construction only used the training cohort in which all patients and healthy controls were from the same medical center (Center I), we acknowledge that the validation using the external cohort may have confounding results. Due to the sample availability, we made the external validation cohort with cancer patients collected from Center II/III and healthy controls randomly selected from Center I. Samples from the three centers were processed together, but the differences in medical centers may contribute to the patient/control discrimination in the external cohort. Prospective studies from different medical centers with matched cancer patients and healthy controls would further help eliminate the impact of pre-analytic confounding factors and provide independent assessments. In addition, the limited size of our sample population may result in overestimation in certain cancer subgroups. For instance, the 100% sensitivity we observed for MIA detection could be attributed to only 29 MIA patients in the validation cohorts. Expanding the sample size of these subgroups would improve the statistical power for a more accurate estimation of the model's prediction performance. The test of the additional dataset has shown that our model is suitable for identifying older noncancer subjects (mean age: 59·5 years; range: 52–75) with benign nodules at the lung cancer screening age. While the results are promising, we acknowledge that the participants in this current study may not fully represent the lung cancer early screening population. Also, testing samples of lung lesions like AIS and atypical adenomatous hyperplasia (AAH) would further verify the validity of our assay and leverage its application for early detection. Hence, we plan to perform a large prospective study of the screening population to validate the model before clinical utility.

Taken together, we herein reported a machine learning model using the profile of cfDNA breakpoint motifs for stage I LUAD detection. Our model has exhibited superior detection capacity, especially for the tumors at the very early stage and small size, and performed consistently with low-coverage WGS data down to 0·5×. We further validated the model performance in the external cohorts. Hence, this model provides an accurate and cost-effective approach for developing early detection of lung cancer and benefits more patients.

Contributors

WG, YS, and SG designed and supervised the study. WG, NL, and QM contributed to clinical information collection. WG, XC, RL, HB, and XX performed technical development, data analysis, and have directly accessed and verified the underlying data. FT and QX provided technical support. WG, XC, HB, and XW drafted the manuscript. WG, XC, RL, HB, XX, and SY contributed to the revision. YS, SG, and JH contributed to project administration. All authors had full access to all the data in the study, discussed the results, and accepted the responsibility to submit the final manuscript for publication. All authors have read and approved the final version of the manuscript.

Data sharing statement

The raw data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of interests

XC, RL, HB, XX, XW, SY, and YS are employees of Nanjing Geneseeq Technology Inc., China. All other authors have declared no conflicts of interest.

Acknowledgments

The authors thank all the patients, volunteers, and their families, the investigators, and the site personnel who participated in this study. This study was supported by the National Key R&D Program of China (grant number 2021YFC2500900), the National Natural Science Foundation of China (grant number 82002451), the CAMS Initiative for Innovative Medicine (grant number 2021-1-I2M-015), the Special Research Fund for Central

Universities, Peking Union Medical College (grant number 3332020024), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (grant number 2018PT32033), and the Beijing Hope Run Special Fund of Cancer Foundation of China (grant number LC2019B15).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104131.

References

1 Sorber L, Zwaenepoel K, Deschoolmeester V, et al. Circulating cell-free nucleic acids and platelets as a liquid biopsy in the provision of personalized therapy for lung cancer patients. *Lung Cancer*. 2017;107:100–107.
2 Key statistics for lung cancer. American Cancer Society; 2021. https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html. Accessed 9 March 2021.
3 International Early Lung Cancer Action Program I, Henschke CI, Yankelevitz DF, et al. Survival of patients with stage I lung cancer detected on CT screening. *N Engl J Med*. 2006;355(17):1763–1771.
4 Goebel C, Louden CL, McKenna R, Jr., Onugha O, Wachtel A, Long T. Diagnosis of non-small cell lung cancer for early stage asymptomatic patients. *Cancer Genom Proteom*. 2019;16(4):229–244.
5 Lung cancer survival rates. American Cancer Society; 2021. https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html. Accessed 9 March 2021.
6 Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol*. 2017;7(9):170070.
7 Can lung cancer be found early? American Cancer Society; 2021. https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/detection.html. Accessed 9 March 2021.
8 National Lung Screening Trial Research TChurch TR, Black WC, Aberle DR, et al. Results of initial low-dose computed tomographic screening for lung cancer. *N Engl J Med*. 2013;368(21):1980–1991.
9 Stroun M, Maurice P, Vasioukhin V, et al. The origin and mechanism of circulating DNA. *Ann NY Acad Sci*. 2000;906:161–168.
10 Sun K, Jiang P, Chan KC, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A*. 2015;112(40):E5503–E5512.
11 Fece de la Cruz F, Corcoran RB. Methylation in cell-free DNA for early cancer detection. *Ann Oncol*. 2018;29(6):1351–1353.
12 Benesova L, Belsanova B, Suchanek S, et al. Mutation-based detection and monitoring of cell-free tumor DNA in peripheral blood of cancer patients. *Anal Biochem*. 2013;433(2):227–234.
13 Chabon JJ, Hamilton EG, Kurtz DM, et al. Integrating genomic features for noninvasive early lung cancer detection. *Nature*. 2020;580(7802):245–251.
14 Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*. 2021;372(6538):eaaw3616.
15 Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 2019;570(7761):385–389.
16 Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV. Consortium C. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31(6):745–759.
17 Mathios D, Johansen JS, Cristiano S, et al. Early detection of lung cancer using cfDNA fragmentation. *J Clin Oncol*. 2021;39(15_suppl):8519.
18 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 Nat Acad Sci USA*. 2018;115(46):E10925–E10E33.
19 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.
20 Han DSC, Ni M, Chan RWY, et al. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am J Hum Genet*. 2020;106(2):202–214.
21 Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362(6413):eaav1898.
22 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.
23 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.
24 Shen SY, Singhania R, Fehringer G, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563(7732):579.
25 Liang WH, Zhao Y, Huang WZ, et al. Noninvasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). *Theranostics*. 2019;9(7):2056–2070.
26 Liang N, Li B, Jia Z, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nat Biomed Eng*. 2021;5(6):586–599.
27 Liang W, Zhao Y, Huang W, et al. Noninvasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). *Theranostics*. 2019;9(7):2056–2070.
28 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–E8E68.
29 Chen L, Abou-Alfa GK, et al. Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. *Cell Res*. 2021;31:589–592.
30 Chen K, Sun J, Zhao H, et al. Noninvasive lung cancer diagnosis and prognosis based on multi-analyte liquid biopsy. *Mol Cancer*. 2021;20(1):23.
31 Tailor TD, Rao X, Campa MJ, Wang J, Gregory SG, Patz EF, Jr. Whole exome sequencing of cell-free DNA for early lung cancer: a pilot study to differentiate benign from malignant CT-detected pulmonary lesions. *Front Oncol*. 2019;9:317.
32 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.
33 Force USPST, Krist AH, Davidson KW, et al. Screening for lung cancer: US preventive services task force recommendation statement. *JAMA*. 2021;325(10):962–970.
34 Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US preventive services task force. *JAMA*. 2021;325(10):971–987.