FUTURE ONCOLOGY
2025, VOL. 21, NO. 11, 1391–1402
https://doi.org/10.1080/14796694.2025.2483154

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Cost-effective shallow genome-wide sequencing for profiling plasma cfDNA signatures to enhance lung cancer detection

Van Thien Chi Nguyen*, Dac Ho Vo*, Thi Trang Tran, Thanh Truong Tran, Thi Hue Hanh Nguyen, Truong Dang Huy Vo, Thi Tuong Vi Van, Thi Luyen Vu, Minh Quang Lam, Giang Thi Huong Nguyen, Trung Hieu Tran, Ngoc Tan Pham, Quang Thinh Trac, Trong Hieu Nguyen, Thi Van Phan, Thi Huyen Dao, Huu Tam Phuc Nguyen, Luu Hong Dang Nguyen, Duy Sinh Nguyen, Hung Sang Tang, Hoa Giang, Minh Duy Phan, Hoai-Nghia Nguyen and Le Son Tran

Research and Development Department, Medical Genetics Institute, Ho Chi Minh, Vietnam

ABSTRACT

Background: Lung cancer (LC) screening via low-dose computed tomography (LDCT) faces challenges including high false-positive rates and low patient compliance. Circulating tumor DNA (ctDNA)-based tests offer a minimally invasive alternative but are limited by high costs and low sensitivity, particularly in early-stage detection. This study introduces a cost-effective, shallow genome-wide sequencing approach for LC detection by profiling multiple cell-free DNA (cfDNA) signatures.

Methods: We developed a multimodal cfDNA assay with shallow sequencing coverage (0.5×) that integrates fragmentomic, nucleosome, end-motif, and copy number alteration analyses. A machine-learning model trained on a discovery cohort (99 LC patients, 168 healthy controls) and validated on an independent cohort (58 LC patients, 71 controls) demonstrated robust performance.

Results: The ensemble model exhibited outstanding performance, achieving an AUC of 0.97 and a specificity of 92% in both the discovery and validation cohorts, with sensitivities of 94% and 90%, respectively. Notably, it outperformed hotspot mutation-based assays and the multi-cancer SPOT-MAS assay in sensitivity across all LC stages.

Conclusions: This assay provides a cost-effective, accurate, and minimally invasive method for LC detection, addressing the limitations of current screening methods. It represents a promising complementary tool to improve early detection and patient outcomes in LC.

PLAIN LANGUAGE SUMMARY

Lung cancer screening can be expensive and sometimes gives false positives, leading to unnecessary worry and extra tests. Many people also skip screening because it is inconvenient. A simple blood test could be a better option, but current ones are costly and not always accurate for early detection. In this study, researchers developed an affordable blood test that looks at tiny fragments of DNA in the blood. They used machine learning to train the test on samples from lung cancer patients and healthy people, then its accuracy was checked on a separate group. The test correctly detected lung cancer in 90% of cases and ruled out healthy individuals with 92% accuracy. It outperformed existing tests and could become an easy and effective way to catch lung cancer early, helping more people get treated in time.

ARTICLE HISTORY
Received 11 January 2025
Accepted 19 March 2025

KEYWORDS
Lung cancer; shallow genome-wide sequencing; cfDNA; fragmentomic; nucleosome; end-motif; copy number alteration

1. Introduction

Lung cancer (LC) is a prevalent and highly lethal malignancy, ranking as the second-leading cause of cancer-related mortality worldwide [1]. In 2022, approximately 2.5 million new cases of LC were reported, along with an estimated 1.8 million deaths, highlighting its profound impact on global health [1]. The prognosis for LC remains dismal, especially for patients diagnosed at advanced stages. Late-stage detection is associated with increased psychological distress, higher healthcare costs, diminished treatment efficacy, and significantly reduced survival rates [1]. Therefore, the development of effective screening methods for early detection is crucial to improving survival outcomes and reducing the economic burden associated with LC.

Low-dose computed tomography (LDCT) is the current gold-standard method for lung cancer (LC) screening, offering significant benefits for high-risk populations. LDCT demonstrates a sensitivity of 93.7% in individuals aged 55–80 years with a smoking history exceeding 30 pack-years [2]. The National Lung Screening Trial (NLST) reported a 20% reduction in LC mortality through LDCT screening [3]. Despite these achievements, the overall five-year survival rate for LC remains below 60%, considerably lower than that of many other cancers [4]. However, LDCT has notable limitations, including an excessive false-positive rate of up to 96.4% [3], leading to

CONTACT Le Son Tran leson1808@gmail.com Research and Development Department, Medical Genetics Institute, Ho Chi Minh 700000, Vietnam
*Van Thien Chi Nguyen and Dac Ho Vo contributed equally to this study.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14796694.2025.2483154.

Article highlights

- Lung cancer screening via low-dose CT faces challenges such as high false-positive rates and low patient compliance.
- Circulating tumor DNA-based tests provide a minimally invasive alternative but are limited by high costs and low sensitivity, especially in early-stage detection.
- This study introduces a cost-effective cfDNA-based assay using shallow genome-wide sequencing for lung cancer detection.
- The assay integrates multiple cfDNA features, including fragmentomics, nucleosome positioning, end-motif patterns, and copy number alterations.
- A machine-learning model was trained on 99 lung cancer patients and 168 healthy controls and validated on an independent cohort of 58 lung cancer patients and 71 controls.
- The model demonstrated high accuracy, achieving an AUC of 0.97 and a specificity of 92%, with sensitivities of 94% and 90% in the discovery and validation cohorts, respectively.
- The cfDNA-based approach outperformed traditional hotspot mutation assays and the multi-cancer SPOT-MAS assay in sensitivity across all lung cancer stages.
- The minimally invasive nature of this test could improve patient compliance and enable broader adoption in routine screening.

unnecessary follow-up tests, invasive procedures, patient anxiety, and concerns regarding cumulative radiation exposure. Additional challenges, including low uptake of LDCT among eligible individuals and the exclusion of high-risk populations not meeting current screening criteria, further hinder its effectiveness [2]. A study analyzing data from over 1 million patients who underwent baseline LDCT screening between 2015 and 2019 found that only 22.3% adhered to subsequent annual screenings [5]. Additionally, approximately 70% of lung cases are diagnosed at an advanced stage, when treatment options are limited, and prognosis is poor [6]. Even when screening is performed, early-stage tumors (<4 mm) may be too small to detect by LDCT, leading to missed diagnoses and delayed intervention [7]. These factors underscore the urgent need for improved screening strategies to enhance early detection and improve patient outcomes.

Liquid biopsy has emerged as a promising minimally invasive alternative for cancer detection, including LC. This approach involves analyzing tumor-derived components in body fluids, such as circulating tumor cells (CTCs), circulating cell-free DNA (cfDNA), and other biomarkers [8,9]. Among these, cfDNA is the most extensively studied biomarker. Distinct differences in cfDNA levels, methylation patterns, and fragmentomic profiles have been observed in early-stage LC patients, enabling the differentiation of cancerous conditions from healthy controls [10,11]. Chabon et al. [12] conducted a lung cancer-specific study to develop and prospectively validate a machine-learning-based ctDNA test, termed "Lung Cancer Likelihood in Plasma" (Lung-CLiP), with a focus on patients with non-small cell lung cancer (NSCLC). By integrating genetic characteristics, Lung-CLiP achieved sensitivities of 41%, 54%, and 67% for stages I, II, and III, respectively, at 98% specificity [12]. Another investigation by Mathios et al. [10] introduced DELFI (DNA Evaluation of Fragments for Early Interception), leveraging cfDNA fragmentation patterns to assess the "fragmentome" across the genome. Using an independent cohort of 385 non-cancer individuals and 46 LC patients, DELFI detected 94% of cancer cases across stages and subtypes at a specificity of 80% by combining fragmentation profiles, clinical risk factors, CEA levels, and CT imaging [10].

Moreover, many studies examining cfDNA as a screening tool for LC have utilized "pan-cancer" assays, where LC was analyzed alongside several other cancer types [13]. Recently, Liu et al. [9] employed high-depth targeted sequencing (139× coverage) to profile the methylation status of over 1 million CpG sites, enabling the detection of more than 50 cancer types, including LC. Similarly, we developed SPOT-MAS (Screening for the Presence of Tumor by DNA Methylation and Size), a multimodal assay combining deep targeted sequencing and shallow genome-wide sequencing to analyze cfDNA methylomics, fragmentomics, copy number variations, and end motifs for detecting breast, colorectal, gastric, liver, and lung cancer [14]. These multi-cancer early detection (MCED) assays rely on high-depth targeted sequencing with extensive panels or multiple library preparation protocols to capture a wide range of cfDNA signatures across various cancer types. This complexity results in high costs, posing challenges to their practicality for single-cancer screening applications. For population-level LC screening, single-cancer detection methods must strike a balance between high performance and cost-effectiveness to ensure widespread adoption.

To address this need, we streamlined the SPOT-MAS protocol by focusing on shallow genome-wide sequencing and performing feature selection to identify the most relevant cfDNA markers for LC. Using these features, we developed an ensemble machine-learning model, validated in a cohort of 157 non-metastatic LC patients and 239 healthy controls [15]. Our findings demonstrate that this simplified, cost-effective approach achieves accurate LC detection, holding significant promise for integration into LC screening programs.

2. Materials and methods

2.1. Patient recruitment

A total of 157 non-metastatic lung cancer patients and 239 healthy controls were recruited at the Medical Genetics Institute (Ho Chi Minh City) between September 2020 and December 2024. Cancer diagnoses were confirmed via tissue biopsy and disease stages were classified according to the TNM staging system established by the American Joint Committee on Cancer (Version VIII). Blood samples were collected before the initiation of any treatment. Healthy control participants were confirmed to have no history or symptoms indicative of cancer at the time of blood collection. They were followed up at six months and one year after enrollment to ensure that they had not developed cancer. Written informed consent was obtained from all participants, and the study was conducted in compliance with the principles outlined in the Declaration of Helsinki. Ethical approval for the study was granted by the Ethics and Scientific Committee of the University of Medicine and Pharmacy and the Medical Genetics Institute, Ho Chi Minh City, Vietnam (Ethics number 164/HDDD).

For machine learning model development, the cohort was divided into a discovery cohort (99 LC patients and 168 healthy controls) and a validation cohort (58 LC patients and 71 healthy controls). The discovery and validation cohorts were randomly split in a 7:3 ratio from the same case-control cohort to minimize deviations in clinical characteristics, such as gender, age, and tumor stage. By using only the discovery cohort for feature selection and model training, we effectively prevented data leakage and overfitting, ensuring an unbiased model evaluation. This approach aligns with best practices for data splitting and follows recommendations to preprocess the entire dataset before model evaluation [16].

2.2. Blood collection and cfDNA isolation

A 10 mL blood sample was collected from each participant using Cell-Free DNA BCT® tubes (Streck). Based on our previous analysis, blood sample hemolysis should remain below 500 mg/dL, as higher hemoglobin levels can interfere with cfDNA-based cancer screening [17]. Therefore, samples with a hemolysis rate exceeding 500 mg/dL were excluded from the study. Plasma was separated through a two-step centrifugation process: an initial spin at 1,600–2,000×g for 10 minutes, followed by a second centrifugation at 16,000×g for 10 minutes. The plasma was then aliquoted and stored at −80°C.

cfDNA was extracted from plasma samples using the MagMAX Cell-Free DNA Isolation Kit (Thermo Fisher Scientific) following the manufacturer's protocol. The cfDNA yield was quantified using the QuantiFluor® dsDNA System (Promega), and the purified cfDNA samples were subsequently stored at −20°C.

2.3. Library preparation and whole genome shallow sequencing

UDI cfDNA libraries were prepared from a minimum of 0.5 ng of cfDNA using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (New England BioLabs) in combination with NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) and quantified by QuantiFluor dsDNA system (Promega).

The cfDNA libraries were sequenced on DNBSEQ-G400 sequencing system (MGI) using 100 bp paired-end reads at the assigned depth of 15 million reads per sample. Data obtained from sequencers were demultiplexed by bcl2fastq (Illumina, CA, USA) to get FASTQ files. All FASTQ files were then examined by FastQC v. 0.11.9 and MultiQC v. 1.12. The paired-end reads were trimmed by Trimmomatic v 0.32 with the option HEADCROP and were mapped onto hg38 – human reference genome [18]. Deduplication and sorting of BAM files were performed with SAM-tools v. 1.11 [19].

Additionally, to evaluate the impact of sequencing depth on test performance, we compared results from samples with coverage $< 0.5X$ and $\geq 0.5X$. We found that shallow-depth sequencing (<0.5X) reduced the model's AUC from 0.98 to 0.92, highlighting the importance of adequate sequencing depth for maintaining test accuracy (Figure S1). Therefore, samples with a sequencing depth below 0.5X were excluded from the analysis.

2.4. Nucleosome footprinting analysis

To determine nucleosome positions in the reference genome, we utilized a published dataset of nucleosome locations [20]. For each nucleosome central point in the reference genome, a 601 bp window was defined, extending 300 bp upstream and 300 bp downstream [21]. The nucleosome footprinting features ($n = 601$) were calculated by dividing the number of reads with endpoints mapped to each position within the 601 bp window (centered at position 0) by the total number of reads.

2.5. Fragment length analysis

From the total pool of reads with lengths ranging from 50 to 350 bp (351 distinct fragment lengths), we calculated the frequency (%) of each fragment length. These frequency values were utilized as fragment length features.

The genome was divided into 2,875 non-overlapping bins, each spanning 1 million bases. Read counts were determined for each bin. Fragment lengths between 50 and 150 bp were classified as short fragments. For each bin, the ratio of short fragments to the total number of fragments was calculated and incorporated as features for data visualization.

2.6. End motifs analysis

The first four consecutive nucleotides (4-mer motifs) at each cleavage site of a sequencing read were identified based on the reference genome. The frequency of each of the 256 possible 4-mer motifs was calculated by dividing the number of reads containing a specific 4-mer motif by the total number of reads in a sample. These frequencies were used as end motif features.

2.7. Copy number aberration analysis

Copy number aberrations (CNA) were assessed for 1-million-base-long genomic bins. After excluding bins with low mappability and those within Duke blacklist regions, a total of 1942 bins were analyzed [22]. For each bin, the total mapped reads were normalized across all bins using z-score normalization, yielding 1942 CNA features per sample.

2.8. Statistical analysis

The Chi-Square test assesses whether a significant association exists between two categorical variables. The Wilcoxon rank-sum (Mann-Whitney) test was used to calculate p-values, which were subsequently adjusted using the Benjamini-Hochberg correction to control false discovery rates. The Kolmogorov-Smirnov test was applied to evaluate whether the two cohorts followed the same statistical distribution. Significant differences were defined when p-value ≤ 0.05.

The statistical analyses were conducted using Python v3.8.10 [23]. The Wilcoxon rank sum test was performed using the stats.ranksums() function with alternative="two-sided" from scipy v1.10.1 [24]. To correct p-values, we used the stats.multitest.fdrcorrection() function from statsmodels

v0.14.1 [25] with parameters set as alpha = 0.05, method="indep," is_sorted = False.

2.9. Construction of machine learning models

A multi-feature classification model was developed to distinguish LC patients from healthy individuals in the discovery cohort. Four feature datasets were constructed: Copy Number Aberration (CNA), Fragment Length Distribution (FLEN), End Motif (EM), and Nucleosome Footprint (NFT). Each dataset underwent 100 bootstrap iterations with replacement enabled. Features with importance scores greater than 0, as identified by the *RandomForestClassifier* algorithm (with "criterion" set to "gini"), were retained if they appeared in at least 50 iterations.

The selected feature sets were used for hyperparameter tuning across multiple machine learning algorithms, including XGBoost (XGB), Logistic Regression (LG), Random Forest (RF), and Support Vector Machines (SVM). A stratified 10-fold cross-validation was performed to identify optimal models for each dataset, with the area under the receiver operating characteristic curve (ROC-AUC) used as the scoring metric.

The probability scores generated from the four datasets were integrated using a simple Logistic Regression model to produce a final aggregated score. This ensemble model was validated using the validation cohort to evaluate its performance

2.10. Hotspot mutation assay

From the COSMIC database and a cohort of 1,100 Vietnamese patients diagnosed with breast, colorectal, gastric, liver, and lung cancers, we identified and selected a panel of 700 hotspot mutations. These mutations were reported to recur in the five cancer types within the COSMIC database, included actionable mutations, and were confirmed in tumor tissue samples from the Vietnamese cancer cohort.

A multiplex PCR approach was employed for the targeted enrichment of 700 hotspot mutations. Library preparation for the enriched target products included an additional indexing PCR step, as detailed in a previous study [26]. To minimize the influence of clonal hematopoiesis of indeterminate potential (CHIP)-related mutations, matched genomic DNA (gDNA) was extracted from the white blood cell fraction of hotspot-positive samples using the GeneJET Whole Blood Genomic DNA Purification Mini Kit (Thermo Fisher). The gDNA was subsequently sequenced, and variant calling was conducted using SAMtools v1.11 [26].

2.11. SPOT-MAS assay

The SPOT-MAS assay is a MCED test designed to detect five cancer types, including breast, colorectal, gastric, liver, and lung cancer [14]. Briefly, circulating cfDNA samples were bisulfite-converted using the EZ DNA Methylation-Gold Kit (Zymo Research), and cfDNA libraries were prepared using the xGen Methyl-Seq DNA Library Prep Kit (Integrated DNA Technologies). Methylomic and fragmentomic features from 450 target regions, as well as genome-wide fractions, were obtained through probe-based hybridization and sequenced on the DNBSEQ-G400 sequencing system (MGI). Advanced machine learning models were developed using these features to differentiate cancer cases from healthy controls [14].

3. Results

3.1. Clinical characteristics of LC patients and healthy participants

A total of 396 participants (Table S1) were recruited for the study, divided into two cohorts: a discovery cohort comprising 267 individuals (99 LC patients and 168 healthy controls) and a validation cohort consisting of 129 individuals (58 LC patients and 71 healthy controls) (Figure 1). In the discovery cohort, the majority of LC patients were male (60.6%, Table 1), slightly higher than the proportion of males among healthy controls (58.9%, Table 1). However, this difference was not statistically significant (Chi-Square test, $p = 0.787$, Table 1). Similarly, in the validation cohort, males accounted for 53.4% of LC patients and 57.7% of healthy controls, with no significant difference in gender distribution (Chi-Square test, $p = 0.625$, Table 1). The median age of LC patients in the discovery cohort was 66 years, which was comparable to the median age of 63 years among healthy controls (Mann-Whitney test, $p = 0.119$, Table 1). A similar trend was observed in the validation cohort, where the median age of LC patients was 62 years, compared to 61 years for healthy controls (Mann-Whitney test, $p = 0.362$, Table 1). Smoking status differed significantly between healthy controls and LC patients in both cohorts. In the discovery cohort, 48.5% of LC patients were smokers, vs. 28% of controls (Chi-Square test, $p = 7.223\times10^{-4}$, Table 1). In the validation cohort, 50% of LC patients smoked, compared to 12.7% of controls (Chi-Square test, $p = 3.727\times10^{-6}$, Table 1).

Regarding disease staging, early-stage and nonmetastatic cases (Stages I, II – IIIA) constituted 23.2% of the discovery cohort (3% in Stage I, 3% in Stage II, and 17.2% in Stage IIIA; Table 1) and 25.9% of the validation cohort (3.4% in Stage I, 3.4% in Stage II, and 19% in Stage IIIA; Table 1). Late-stage metastatic cases (Stages IIIB – IV) accounted for 18.2% of LC patients in the discovery cohort and 27.6% in the validation cohort. Staging information was unavailable for 58.6% and 46.6% of LC patients in the discovery and validation cohorts, respectively. All healthy controls underwent LDCT or X-ray imaging and were monitored for 12 months to confirm their cancer-free status.

3.2. Shallow genome-wide sequencing reveals distinct fragment length, nucleosome footprinting, and 4-mer end motif patterns in LC patients

Previous studies have demonstrated that apoptosis-mediated caspase activity plays a pivotal role in shaping the nonrandom fragmentation patterns of cfDNA, with ctDNA fragments typically being shorter than those of non-cancer cfDNA [27]. In this study, we compared DNA fragment lengths between LC patients and healthy controls. As anticipated, LC patients displayed an enrichment of shorter fragments (<150 bp) (Figure 2(a)). The mean ratio of short fragments (length 50 to

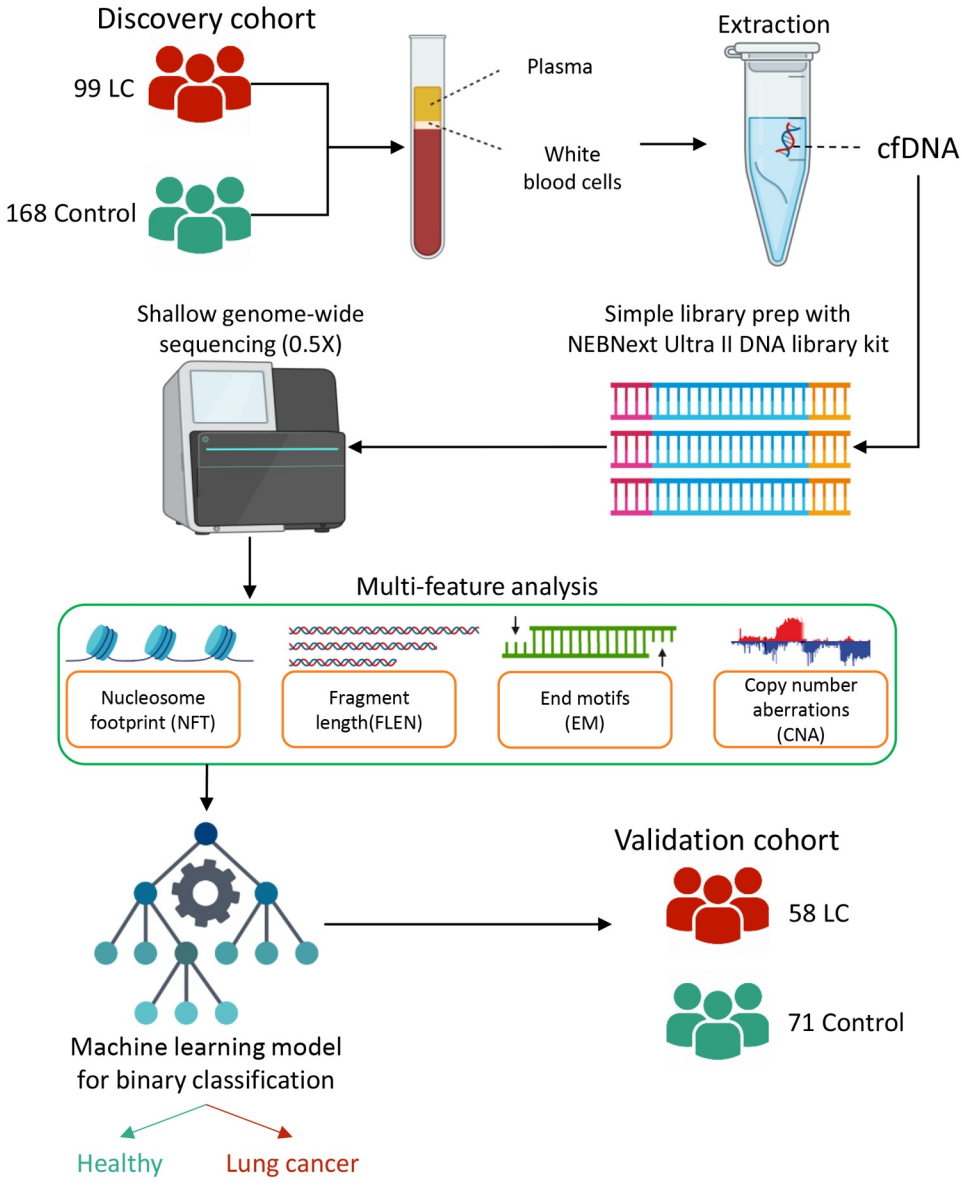

Figure 1. Study design.

A multimodal analysis of cfDNA using low-depth (0.5X) genome-wide sequencing was developed to profile multiple features, including nucleosome footprint (NFT), fragment length (FLEN), end motifs (EM), and copy number aberrations (CNA), in the plasma cfDNA of 99 lung cancer patients and 168 healthy individuals in the discovery cohort. Significant signatures identified from the analysis were used to construct a machine learning model to classify lung cancer patients from healthy individuals. The model's performance was externally validated in a separate validation cohort consisting of 58 lung cancer patients and 71 healthy subjects.

150 bp) to total fragments across 5 Mb bins for all 22 chromosomes was higher in LC compared to healthy controls (Figure 2(b)). These findings suggest that cfDNA fragment length profiles may serve as distinguishing features for differentiating LC patients from healthy subjects.

The fragmentation patterns of plasma cfDNA are closely associated with nucleosome positioning and chromosome structure, offering insights into the biological and pathological processes occurring in the source tissue [21]. Deviations from these patterns across the genome may act as potential markers of cancer-associated signaling. However, this phenomenon has not yet been reported in LC. Using published nucleosome positions in healthy samples as a reference [28], we analyzed cfDNA reads for their distances to nucleosome cores, extending 300 bp upstream and downstream. This analysis enabled us to assess nucleosome footprint distribution patterns in both LC patients and healthy controls (Figure 2(c)). Our findings reveal that LC samples exhibit an M-shaped nucleosome footprint distribution, characterized by a depletion of read starts at nucleosome edges and enrichment at nucleosome centers compared to controls (Figure 2(d)). This pattern suggests a shorter DNA wrapping length around nucleosomes in LC samples. These unique nucleosome footprint patterns in plasma cfDNA underscore the potential of nucleosome distance profiling as a biomarker for LC detection.

In addition to fragment lengths and nucleosome positioning, variations in cfDNA fragment ends may reflect altered cleavage patterns in cancer cells compared to normal cells during apoptosis [29]. To investigate this, we analyzed 256

Table 1. Summary of clinical information of 157 cancer patients and 239 healthy subjects (total $N = 396$).

		Discovery cohort ($N = 267$)					Validation cohort ($N = 129$)				
		Lung cancer ($N = 99$)		Healthy control ($N = 168$)		p-value	Lung cancer ($N = 58$)		Healthy control ($N = 71$)		p-value
Clinical features		N	%	N	%	(All cancer vs Healthy control)	N	%	N	%	(All cancer vs Healthy control)
Gender	Female	39	39.4	69	41.1	Chi-Square test	27	46.6	30	42.3	Chi-Square test
	Male	60	60.6	99	58.9	p-value = 0.787	31	53.4	41	57.7	p-value = 0.625
Age	Median	66		63		Mann-Whitney test	62		61		Mann-Whitney test
	Min	31		31		p-value = 0.119	36		36		p-value = 0.362
	Max	91		91			95		96		
Smoking status	Yes	48	48.5	47	28.0	Chi-Square test	29	50	9	12.7	Chi-Square test
	No	51	51.5	121	72.0	p-value = $7.223e^{-4}$	29	50	62	87.3	p-value = $3.727e^{-6}$
Non-metastasis	IA, IB	3	3.0				2	3.4			
	IIA, IIB	3	3.0				2	3.4			
	IIIA	17	17.2				11	19.0			
Metastasis	IIIB, IV	18	18.2				16	27.6			
Unknown staging information	NA	58	58.6				27	46.6			

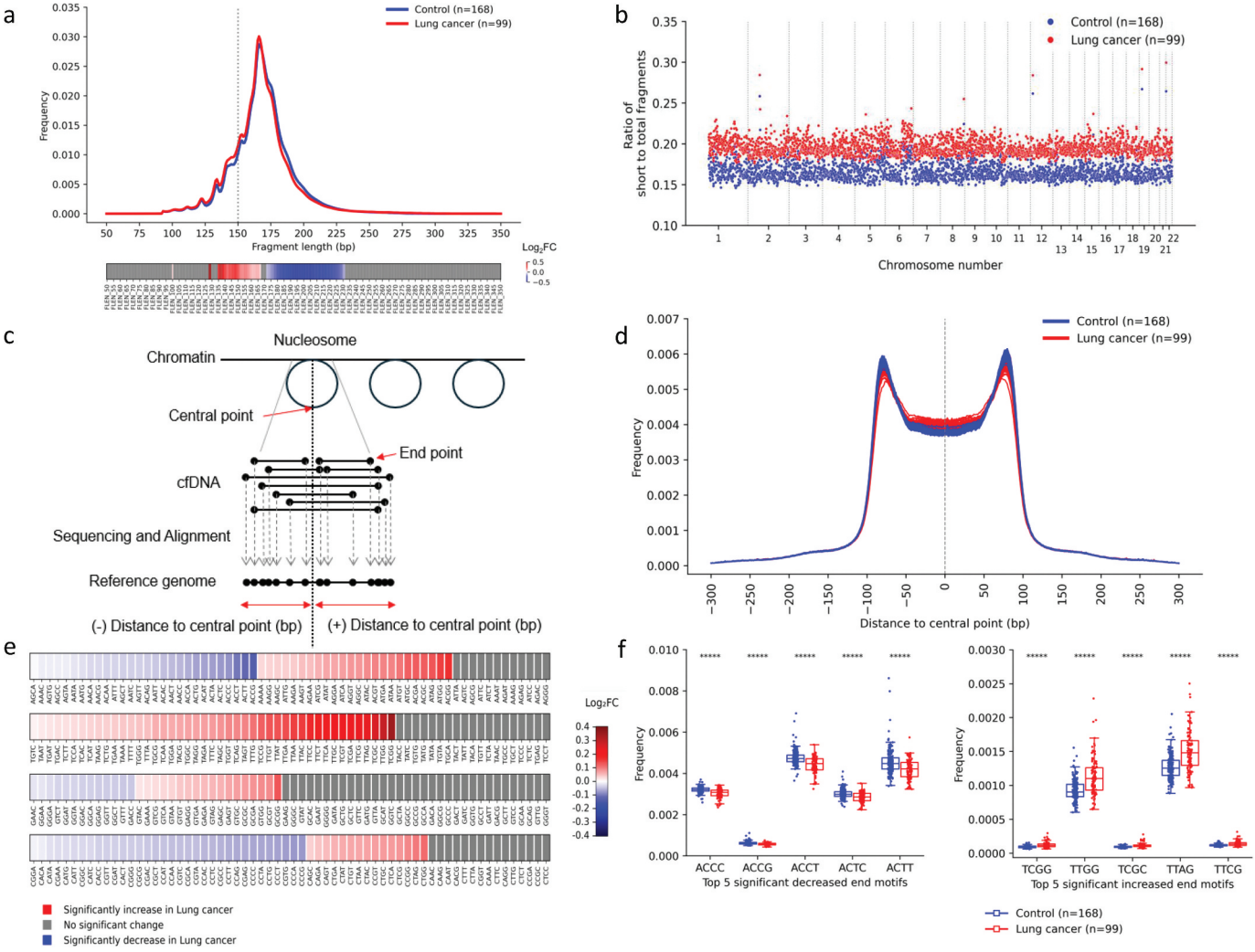

Figure 2. Lung cancer patients exhibit distinct patterns in fragment length, nucleosome footprinting, and 4-mer end motifs.

(a) The fragment length distribution in healthy control individuals (blue) and lung cancer patients (red). The color bar indicates the $\log_2$ fold change ($\log_2$FC) of fragment length features, with red denoting significant increases, blue denoting significant decreases, and gray indicating no significant change in lung cancer compared to the healthy control group. (b) The ratio of short fragments (<150 bp) to total fragments across 22 autosomes in healthy controls (blue) and lung cancer patients (red). (c) The diagram depicts the principle of generating nucleosome footprinting features. (d) The "M-shape" nucleosome footprinting features in healthy control individuals (blue) and lung cancer patients (red). (e) A heatmap displaying the $\log_2$FC of 4-mer end motif frequencies in the lung cancer group relative to the healthy control group. End motifs with significantly higher frequencies in lung cancer patients are highlighted in red, while those with significantly lower frequencies are highlighted in blue. (f) The top 5 most significantly increased and decreased end motifs in the lung cancer group compared to the healthy control group. Data in F are Mean ± SEM. Significance in E and F was assessed with the Mann-Whitney test, $p < 0.00001$.

distinct 4-mer EMs and compared their frequencies between LC patients and controls. Our analysis identified 75 EMs with significantly lower frequencies and 102 EMs with significantly higher frequencies in LC patients (Figure 2(e)). Interestingly, cfDNA from LC patients exhibited significantly lower frequencies of motifs starting with adenine and significantly higher frequencies of motifs starting with thymine (Figure 2(e)). Among the most significant EMs, the top five motifs with increased frequencies in LC patients were TCGG, TTGG, TCGC, TTAG, and TTCG, while those with the greatest decreases were ACCC, ACCG, ACCT, ACTC, and ACTT (Mann-Whitney test, $p < 0.05$) (Figure 2(f)). These results demonstrate that cfDNA from LC patients displays distinct end motif patterns, highlighting their potential utility in distinguishing LC patients from healthy individuals.

3.3. Distinct genome-wide DNA CNAs in plasma cfDNA of lung cancer patients

CNAs, often associated with chromosomal instability, involve the gain or loss of chromosomal segments or even entire chromosomes and represent a critical hallmark of cancer, alongside fragmentomic profiles [30]. In this study, we compared CNA profiles in cfDNA between LC patients and healthy controls. We identified 368 bins (19%) with significant CNA gains and 254 bins (13%) with significant CNA losses (Mann-Whitney test, Benjamini-Hochberg adjusted $p < 0.05$) (Figure 3(a)). Notably, CNA gains were predominantly observed on chromosomes 1, 7, and 20, whereas CNA losses were more frequently detected on chromosomes 4, 9, 13, and 22 (Figure 3(b)). These findings underscore the potential of CNAs as biomarkers for early detection and provide new insights into the chromosomal alterations present in the cfDNA of LC patients.

3.4. Multimodal analysis of cfDNA features could enhance LC detection

The identification of multiple distinct signatures in the cfDNA of LC patients prompted the development of a multi-feature classification model. We constructed four feature datasets: CNA, FLEN, EM, and NFT (Figure 4(a)). Features selected through bootstrapping were utilized for hyperparameter tuning using algorithms such as XGB, LR, RF, and SVM. Stratified 10-fold cross-validation was performed to identify the best-performing models for each dataset, with the "roc_auc" metric used for evaluation. The probability scores generated for each dataset were subsequently combined using a simple LR model, resulting in an ensemble score set.

We compared the classification performance of individual feature models with that of the ensemble model, which integrated all four features. The single-feature models demonstrated comparable performance, with the EM-, CNA-, and FLEN-based models achieving AUC values of 0.92, 0.91, and 0.90, respectively, in the discovery cohort. The NFT-based model exhibited slightly lower performance, with an AUC of 0.84 (Figure 4(b)). By integrating the four best-performing single-feature models using LR, the ensemble model

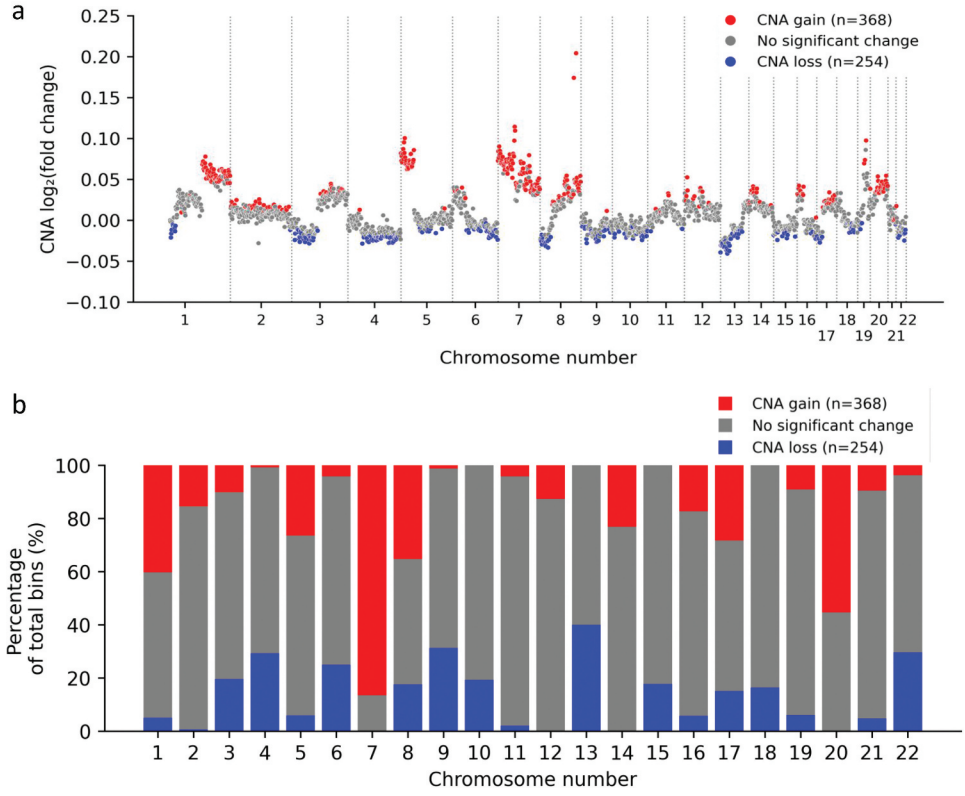

Figure 3. Genome-wide DNA copy number aberrations in lung cancer patients.

(a) Log_2 fold-change (Log_2FC) between LC patients and healthy control groups for all CNA features across 22 autosomes. (b) The proportion of genomic bins categorized as CNA gain (red), CNA loss (blue), or no significant CNA change (gray) across 22 autosomes. Statistical significance for features in panels a and b was determined using the Mann-Whitney test.

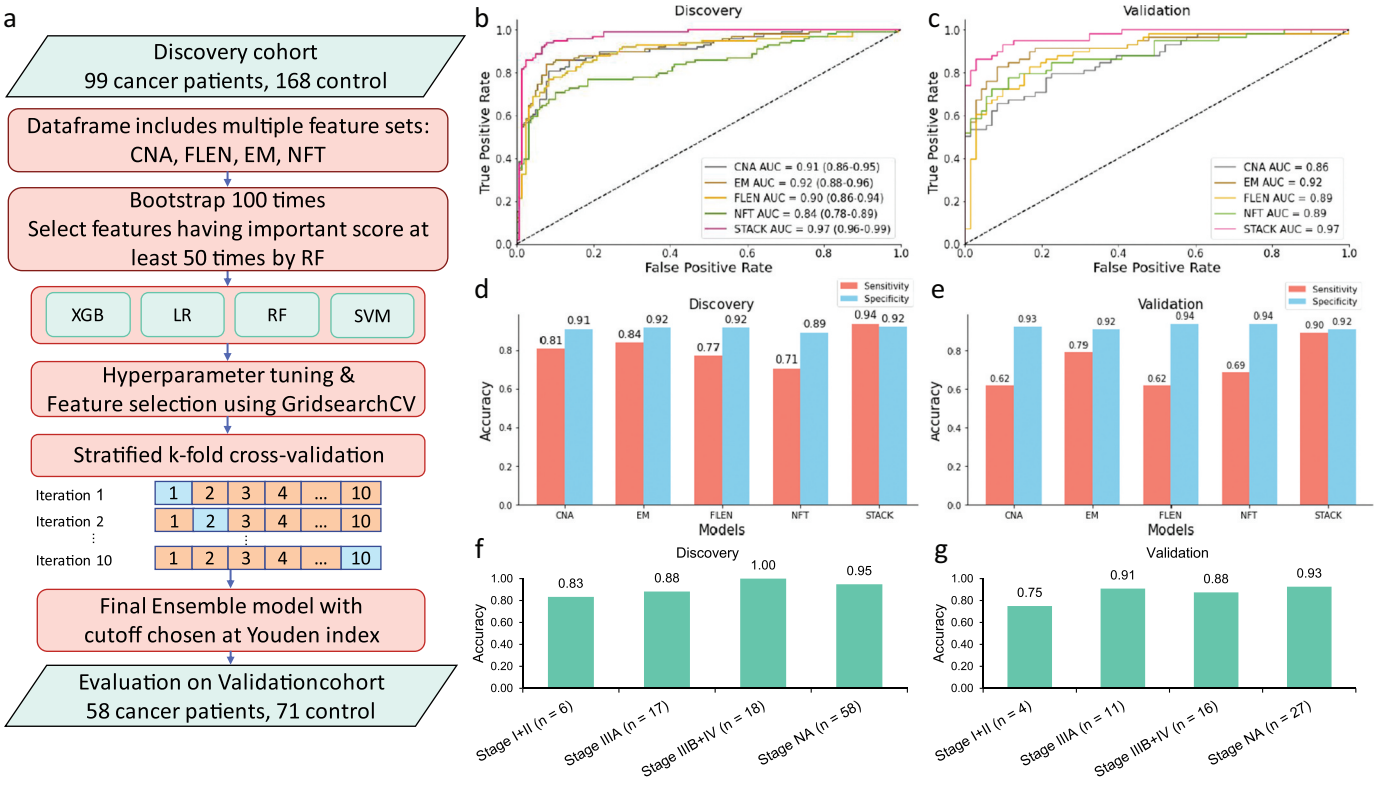

Figure 4. Model construction and performance validation.

(a) Model construction strategy. (b, c) Receiver operating characteristic (ROC) curves comparing the performance of single-feature models and the ensemble model in the discovery cohort (b) and validation cohort (c). (d, e) Bar charts illustrating the specificity and sensitivity of single-feature models versus the ensemble model in the discovery cohort (d) and validation cohort (e). (f, g) Bar charts depicting the sensitivity of the ensemble model across different lung cancer stages in the discovery cohort (f) and validation cohort (g).

significantly improved classification performance, achieving an AUC of 0.97 (95% CI: 0.96–0.99, Figure 4(b)).

To optimize overall accuracy by balancing sensitivity and specificity, we established a cutoff value using the Youden index [31]. The ensemble model consistently outperformed the single-feature models, achieving a sensitivity of 0.94 at a specificity of 0.92 (Figure 4(d)). To further assess its robustness, we tested the ensemble model on an independent validation cohort comprising 58 LC patients and 71 healthy individuals. As observed in the discovery cohort, the ensemble model demonstrated the highest accuracy in detecting LC, achieving an AUC of 0.97 (Figure 4(c)), sensitivity of 0.90, and specificity of 0.92 (Figure 4(e)).

3.5. Evaluating the impact of clinical features on model performance

To assess the robustness of our model, we evaluated its performance stratified by gender, age, and smoking status. Across both cohorts, cancer probability scores were significantly lower in control samples compared to LC patients, regardless of gender (Figure S2(a,b); $p < 0.0001$, Mann-Whitney test), with comparable classification performance observed between males and females (Figure S2(c,d)). Similarly, no significant differences in model scores (Figure S3A-B) or performance (Figure S3(c,d)) were found between older (>63 years) and younger (≤63 years) participants. Stratification by smoking status also showed significantly lower scores in controls across both cohorts (Figure S4A-B; $p < 0.0001$, Mann-Whitney test), with no notable differences in classification performance between smokers and nonsmokers (Figure S4C-D). These findings demonstrate that our model reliably predicts LC while minimizing potential confounding effects of gender, age, and smoking status.

Next, we evaluated the model's sensitivity when stratifying LC samples by stage. In the discovery cohort, sensitivity increased progressively from early stages I and II (0.83) to stage IIIA (0.88) and late-stage IIIB and IV (1.00) (Figure 4(f)). Similarly, in the validation cohort, the model showed the lowest sensitivity (0.75) for detecting stages I+II, compared to stage IIIA (0.91) and later stages IIIB and IV (0.88) (Figure 4(g)). These findings indicate that the multi-feature model maintained consistently high classification performance across both early and late-stage tumors in the validation cohort.

3.6. Performance comparison between LC multi-feature model and hotspot-based or MCED assays

LC is characterized by a high tumor mutation burden, and mutations in plasma cfDNA have been utilized not only for selecting LC patients for targeted therapy but also for early detection [32]. A previous study by Cohen et al. [33] developed the CancerSeek assay, using a targeted panel of 60 mutations across 16 cancer-associated genes, aimed at early detection of eight different cancer types, including LC. Our group also previously developed the SPOT-MAS assay, which

combines deep targeted sequencing and genome-wide sequencing to profile fragmentomic and methylomic signatures for detecting five common cancer types, including LC [14]. In this study, we compared the performance of our LC multi-feature model with a 700-hotspot mutation-based assay and SPOT-MAS in differentiating LC from healthy individuals in the validation cohort. The hotspot mutation assay demonstrated 100% specificity but only 48% sensitivity, while the SPOT-MAS assay showed a lower specificity of 92% but a higher sensitivity of 74% (Figure 5(a)). Notably, the multi-feature LC model exhibited specificity comparable to SPOT-MAS but higher overall sensitivity than both the hotspot and SPOT-MAS assays across all tumor stages (Figure 5(a)).

In particular, the LC multi-feature model showed greater sensitivity for detecting non-metastatic stages (I-II and IIIA), achieving 75% sensitivity compared to 50% for the hotspot assay and SPOT-MAS in stages I-II, and 91% compared to 27% for the hotspot assay and 64% for SPOT-MAS in stage IIIA. Interestingly, all LC samples positive for hotspot mutations were detected by the multi-feature model, while 24 out of 28 (86%) were detected by the SPOT-MAS assay (Figure 5(b)). This suggests that combining the hotspot mutation assay with the multi-feature model may not provide a significant improvement, whereas integrating the hotspot mutation assay with the SPOT-MAS could slightly increase overall sensitivity. These findings suggest that our LC multi-feature model, built from LC-specific signatures profiled through a simpler protocol, may offer superior detection of ctDNA shed by non-metastatic LC tumors compared to the hotspot mutation-based approach or the more complex MCED assay, SPOT-MAS.

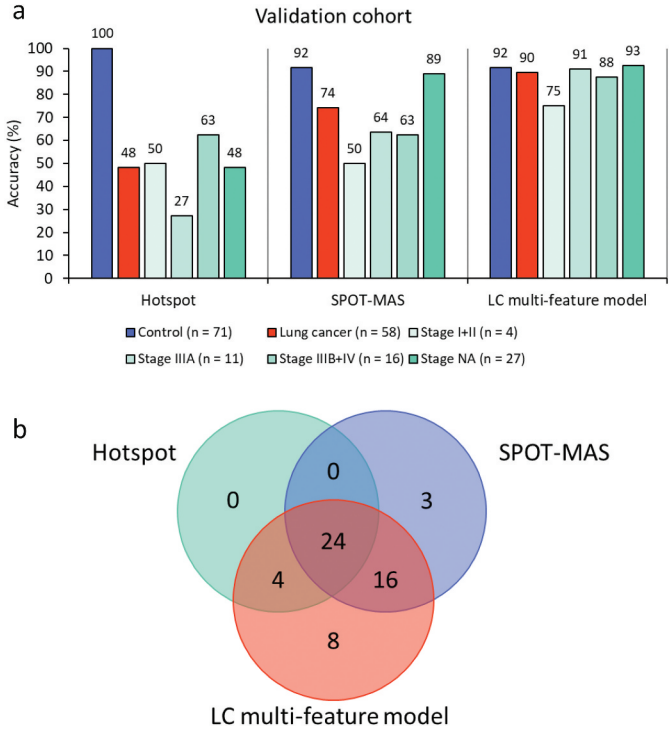

Figure 5. Performance comparison of hotspot, SPOT-MAS and Multi-feature lung model on the validation cohort.

(a) Accuracy of the three assays on control and lung cancer samples. (b) Venn diagram showing the overlap of detected lung cancer samples by three assays.

4. Discussion

Current screening methods for LC, particularly LDCT scans, have notable limitations, including high false-positive rates and low patient compliance [3]. Recently, ctDNA-based tests have emerged as a promising alternative. However, these tests face technological challenges, including low sensitivity for early-stage tumor detection and high costs. In this study, we present a novel approach utilizing a cost-effective, shallow genome-wide sequencing technique to profile multiple lung cancer-specific features in cfDNA. By integrating these diverse features, we developed a robust and accurate model for lung cancer detection, surpassing traditional methods that typically focus on single features or rely on pan-cancer feature selection.

Previous studies have demonstrated that cfDNA from LC patients is enriched with shorter fragments [27,34], and our fragmentomic profiling confirms these findings [14]. This enrichment arises primarily from altered nucleosome positioning and chromatin remodeling in cancer cells. In healthy cells, nucleosomes are regularly spaced, protecting the DNA from degradation and resulting in longer cfDNA fragments. In contrast, cancer cells exhibit increased transcriptional activity, disrupted nucleosome positioning, and chromatin disorganization. This results in nucleosome depletion near transcription start sites (TSS), exposing DNA to enzymatic cleavage and generating shorter cfDNA fragments [35]. Furthermore, rapid cell proliferation and apoptosis in cancer lead to an increased release of fragmented DNA into the bloodstream, amplifying the abundance of shorter cfDNA fragments [36]. These characteristic patterns of cfDNA fragmentation offer valuable insights into the molecular mechanisms driving this process and underscore its potential as a biomarker for early detection.

The nucleosome footprinting profiles observed in this study are consistent with previous research [21], revealing distinct differences between LC patients and healthy controls. Variations in nucleosome positioning likely reflect underlying disparities in epigenetic regulation and gene expression within lung tumors. In cancer, nucleosome positioning is frequently disrupted by epigenetic alterations and oncogenic mutations, leading to distinct cfDNA fragmentation patterns. In normal cells, well-positioned nucleosomes flank transcription start site (TSS) regions, creating a "footprint" of cfDNA depletion at highly expressed genes. In contrast, cancer-associated nucleosome shifts or depletion near oncogene TSS regions reflect transcriptional dysregulation. Studies have shown that nucleosome footprints around key oncogenes (e.g., *MYC, EGFR, KRAS*) shift in response to increased transcriptional activity, whereas the silencing of tumor suppressor genes (e.g., *TP53, CDKN2A*) leads to nucleosome repositioning, thereby reducing the accessibility of their cfDNA fragments [37,38]. Notably, shorter cfDNA fragments contain transcription factor footprints, providing further evidence of these regulatory differences [14]. These findings highlight the potential of cfDNA fragmentomics as a noninvasive tool to decode the molecular landscape of LC and enhance early detection strategies.

In line with the observed differences in fragment length profiles, we identified unique patterns in EM. Notably, cfDNA from LC patients exhibited significantly higher frequencies of motifs starting with thymine (T*) (Figure 2(e)). This finding aligns with a previous study by Jiang et al. [29], which reported significant enrichment of TAAA and TTTA motifs in hepatocellular carcinoma (HCC) patients. However, we observed lower frequencies of motifs starting with adenine (A*) in LC patients (Figure 2(f)), contrasting with Jiang et al.'s findings of reduced frequencies of cytosine-initiated (C*) motifs. These results suggest that EM profiles may represent cancer-specific signatures, providing insights into tumor-specific DNA fragmentation patterns.

Somatic CNAs in the cancer genome can influence the transcriptional regulation of tumor suppressor genes and oncogenes, driving cancer progression and metastasis [39]. In our LC study, prominent CNA gains were observed on chromosomes 1, 7, and 20, with chromosome 7 showing the most frequent gains (Figure 3(a)). Conversely, CNA losses were predominantly seen on chromosomes 4, 9, 13, and 22, with chromosome 13 exhibiting the most significant losses (Figure 3(b)). These results partially align with our pan-cancer analysis, which identified chromosome 22 as the site of the most frequent CNA losses [14]. Interestingly, the CNA gain/loss ratio in this study corroborates findings by Bjaanæs et al. [40], who analyzed the whole-genome copy number profiles of TP53-mutant lung adenocarcinomas. This congruence suggests the presence of universal CNA signatures in LC.

In this study, we hypothesized that integrating multiple lung cancer-specific molecular signatures in cfDNA could improve the performance of liquid biopsy assays, enhancing their accuracy for cancer detection. As anticipated, the integrated model combining all four features outperformed individual models, achieving consistent classification performance across both the discovery and validation cohorts, with an AUC of 0.97 in both cases (Figure 4(b,c)). The superior performance of the combined model can be attributed to its ability to capture a broader spectrum of cancer-related signals, effectively addressing the heterogeneity of ctDNA in cancer patients [41]. By integrating diverse molecular features – including CNA, FLEN, NFT, and EM profiles – the model utilizes a comprehensive set of biomarkers across LC patients and subtypes. Furthermore, our findings align with a recent study that underscores the potential of fragmentation features as biomarkers for ctDNA detection in lung cancer patients [42]. This integrative approach enhances robustness and accuracy, offering a promising tool for noninvasive cancer detection.

Compared to the DELFI test, our multi-feature model demonstrated a slightly lower sensitivity (90% vs. 94%) but achieved a higher specificity (92% vs. 80%) in the validation cohort [10]. Given the high tumor mutation burden in LC, mutations in plasma cfDNA are increasingly being utilized not only for selecting patients for targeted therapies but also for early detection [32]. To further evaluate our model, we compared its performance against a mutation-based assay targeting 700 hotspot mutations derived from recurrent mutations in LC patients from the TCGA database [26]. Our model demonstrated superior sensitivity (90% vs. 48%) but slightly lower specificity (92% vs. 100%) compared to the hotspot mutation-based assay (Figure 5(a)). Interestingly, all LC patients who tested positive using the hotspot mutation assay were also predicted positive by our multi-feature model (Figure 5(b)). This concordance suggests a potential relationship between mutations and cfDNA fragmentomic signatures captured by our assay. These findings align with prior studies [43–45], which reported that tumor-derived mutation fragments in cancer patients show an increased proportion of short cfDNA fragments.

Previously, our group developed the SPOT-MAS assay, which integrates deep targeted sequencing with genome-wide sequencing to profile fragmentomic and methylomic signatures for detecting five common cancers, including LC [14]. Compared to the SPOT-MAS assay, our multi-feature model demonstrated comparable specificity (92%) but higher sensitivity (90% vs.74%). These results highlight the advantages of focusing on lung cancer-specific biomarkers and features, allowing for more accurate detection of this cancer type [46,47]. In addition, our assay provides several practical benefits over MCED tests like SPOT-MAS or Galleri [9]. It requires only a single blood draw and utilizes shallow sequencing at 0.5× coverage, significantly reducing the complexity and cost typically associated with deep sequencing methods. Despite its simplicity, the assay delivers reliable and accurate results for detecting lung cancer. Notably, the detection rate of our assay was influenced by tumor stage rather than demographic factors such as gender or age, likely reflecting the increased ctDNA release into the bloodstream as tumors progress. This stage-dependent performance underscores the importance of ctDNA burden in enhancing detection sensitivity [48].

To further assess the performance of our assay in early-stage lung cancer, we established an external validation cohort comprising healthy individuals and patients with localized lung cancer (stages I, II, and IIIA). As of manuscript resubmission, we have recruited 200 healthy subjects and 19 lung cancer patients. Our model achieved an AUC of 0.93 (Figure S5A), with an overall sensitivity of 84% and specificity of 90%, demonstrating higher sensitivity for stage IIIA compared to stage I/II (91% vs. 75%, Figure S5B). These findings indicate that our model maintains robust performance in early-stage lung cancer when validated in an external cohort.

A key limitation of this study is its retrospective design, which may introduce biases related to sample collection, storage, and processing. As a case-control study, it does not reflect the prevalence of lung cancer in a typical screening population, and the results should be considered exploratory. Future prospective cohort studies are needed to rigorously assess clinical performance and mitigate potential biases. Another limitation is that our model's ability to distinguish lung cancer from benign pulmonary conditions, such as tuberculosis, has not been evaluated. This capability would offer significant clinical benefits by reducing unnecessary invasive procedures, such as biopsies or surgeries, for patients with benign nodules. In future studies, we aim to recruit patients with benign lesions and perform feature analysis to determine whether our assay can effectively differentiate between these two groups. Additionally, incomplete clinical staging information for some patients posed challenges, as those initially diagnosed at the study hospital but treated elsewhere were

lost to follow-up. Furthermore, late-stage lung cancer cases were overrepresented in this study, which may limit the generalizability of the findings to early-stage disease detection.

Therefore, the clinical implementation of our model presents several key challenges. First, false negatives may occur in early-stage cancers, where ctDNA levels are below the assay's detection limit, underscoring the need for additional molecular biomarkers or advanced ctDNA enrichment protocols to improve detection rates. Second, false positives may arise from benign lesions that share partial molecular similarities with tumors. To address this, incorporating a larger cohort of benign samples during model training and identifying distinguishing molecular features could enhance specificity. Most importantly, a longitudinal prospective study with randomized controls is crucial to assess the model's overall benefits, limitations, and effectiveness compared to conventional screening methods.

5. Conclusion

This study introduces a streamlined workflow that comprehensively profiles nucleosome footprints, fragmentomics, CNAs, and motif-end signatures from plasma cfDNA. By leveraging a unique multimodal approach, we demonstrate the potential for detecting LC with high accuracy using low-cost, shallow sequencing. This approach offers a promising complementary tool to traditional screening methods, addressing the high false-positive rates and low patient compliance associated with LDCT scans. Future work will focus on validating the model in prospective cohorts and evaluating its ability to differentiate malignant from benign conditions, ensuring broader clinical applicability.

Funding

This work was supported by Gene Solutions. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of interest

The authors including LST, HG, MDP, HNN and DSN hold equity in Gene Solutions. VTCN, DHV, TTT, TTT, THHN, TDHV, TTVV, TLV, MQL, GTHN, THT, NTP, QTT, THN, TVP, THD, HTPN, LHDN and HST received salaries from the project funded by the Medical Genetics Institute, Ho Chi Minh City, Vietnam. We confirm that this does not alter our adherence to journal policies on sharing data and materials. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Author contribution

VTCN, DHV, TTT, TTT, THHN, TDHV, TTVV, TLV, MQL, THT, NTP, QTT, THN perfomed formal analysis.
THD, HTPN, TVP, LHDN performed patient consultancy and screening.
VTCN, DHV, THT, NTP, QTT, THN performed data curation.
DSN, HST, MDP, HG, HNN, LST performed the methodology.
DSN, HST, MDP, HG, HNN, LST performed conceptualization.
VTCN, DHV, GTHN, LST performed writing-original draft.
VTCN, DHV, GTHN, LST performed writing-review and editing.

Data availability statement

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

Ethical disclosure

This study was approved by the Ethics Committee of University of Medicine and Pharmacy and the Medical Genetics Institute, Ho Chi Minh City, Vietnam. Written informed consent was obtained from each participant in accordance with the Declaration of Helsinki.

Acknowledgments

We thank all participants who agreed to take part in this study.

The authors utilized ChatGPT for language polishing and editing during the preparation of this work. After incorporating the tool's suggestions, they carefully reviewed and revised the content as needed, taking full responsibility for the final publication.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–263. doi: 10.3322/caac.21834
2. Pinsky PF. Lung cancer screening with low-dose CT: a world-wide view. Transl Lung Cancer Res. 2018 Jun;7(3):234–242. doi: 10.21037/tlcr.2018.05.12
3. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395–409.
4. Schabath MB, Cote ML. Cancer progress and priorities: lung cancer. Cancer Epidemiol Biomarkers Prev. 2019 Oct;28(10):1563–1579. doi: 10.1158/1055-9965.EPI-19-0221
5. Silvestri GA, Goldman L, Tanner NT, et al. Outcomes from more than 1 million people screened for lung cancer with low-dose CT imaging. Chest. 2023 Jul;164(1):241–251. doi: 10.1016/j.chest.2023.02.003
6. Cagle PT, Allen TC, Olsen RJ. Lung cancer biomarkers: present status and future developments. Archiv Pathol Lab Med. 2013;137(9):1191–1198. doi: 10.5858/arpa.2013-0319-CR
7. Lee H, Um SW. Probability of lung cancer based on the size threshold and volume-doubling time for lung nodules detected in low-dose CT screening. Ann Transl Med. 2015 Feb;3(2):21.
8. Li J, Han X, Yu X, et al. Clinical applications of liquid biopsy as prognostic and predictive biomarkers in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. J Exp Clin Cancer Res. 2018 Sep 3;37(1):213. doi: 10.1186/s13046-018-0893-1
9. Liu MC, Oxnard GR, Klein EA, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020 Jun;31(6):745–759.
10. Mathios D, Johansen JS, Cristiano S, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021 Aug 20;12(1):5060. doi: 10.1038/s41467-021-24994-w
11. Bruhm DC, Mathios D, Foda ZH, et al. Single-molecule genome-wide mutation profiles of cell-free DNA for non-invasive detection of cancer. Nat Genet. 2023 Aug;55(8):1301–1310. doi: 10.1038/s41588-023-01446-3
12. Chabon JJ, Hamilton EG, Kurtz DM, et al. Integrating genomic features for non-invasive early lung cancer detection. Nat. 2020 Apr;580(7802):245–251. doi: 10.1038/s41586-020-2140-0
13. Duffy MJ. Circulating tumor DNA (ctDNA) as a biomarker for lung cancer: early detection, monitoring and therapy prediction. Tumour Biol. 2024;46(s1):S283–s295. doi: 10.3233/TUB-220044
14. Nguyen VTC, Nguyen TH, Doan NNT, et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for

multi-cancer early detection and localization. Elife. 2023 Oct 11;12:RP89083. doi: 10.7554/eLife.89083.3
15. Van Thien Chi Nguyen DHV, Tran TT, Tran TT, et al. Cost-effective shallow genome-wide sequencing for profiling plasma cfDNA signatures to enhance lung cancer detection. 2025 Jan 11. doi: 10.2139/ssrn5094108
16. Master SR, Badrick TC, Bietenbeck A, et al. Machine learning in laboratory medicine: recommendations of the IFCC working group. Clin Chem. 2023;69(7):690–698. doi: 10.1093/clinchem/hvad055
17. Nguyen THH, Lu YT, Le VH, et al. Clinical validation of a ctDNA-based assay for multi-cancer detection: an interim report from a Vietnamese longitudinal prospective cohort study of 2795 participants. Cancer Invest. 2023 Mar 16;41(3):232–248.
18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120. doi: 10.1093/bioinformatics/btu170
19. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–2079. doi: 10.1093/bioinformatics/btp352
20. Snyder MW, Kircher M, Hill AJ, et al. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016 Jan 14;164(1–2):57–68. doi: 10.1016/j.cell.2015.11.050
21. Vanderstichele A, Busschaert P, Landolfo C, et al. Nucleosome footprinting in plasma cell-free DNA for the pre-surgical diagnosis of ovarian cancer. NPJ Genom Med. 2022 Apr 28;7(1):30. doi: 10.1038/s41525-022-00300-5
22. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene. 2002 Aug 12;21(35):5400–5413. doi: 10.1038/sj.onc.1205651
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2012 Jan 2;12:2825–2830.
24. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020 Mar 1;17(3):261–272.
25. Skipper SJ. Statsmodels: econometric and statistical modeling with Python. SciPy Proc. 2010 Jun 28. doi: 10.25080/Majora-92bf1922-011
26. Nguyen THH, Vu GH, Nguyen TT, et al. Combination of hotspot mutations with methylation and fragmentomic profiles to enhance multi-cancer early detection. Cancer Med. 2025;14(1):e70575. doi: 10.1002/cam4.70575
27. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nat. 2019 Jun 1;570(7761):385–389. doi: 10.1038/s41586-019-1272-6
28. Budhraja KK, McDonald BR, Stephens MD, et al. Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. Sci Transl Med. 2023 Jan 11;15(678):eabm6863. doi: 10.1126/scitranslmed.abm6863
29. 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 May;10(5):664–673. doi: 10.1158/2159-8290.CD-19-0622
30. Dereli-Öz A, Versini G, Halazonetis TD. Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. Mol Oncol. 2011 Aug;5(4):308–314. doi: 10.1016/j.molonc.2011.05.002
31. Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan;3(1):32–35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
32. Sha D, Jin Z, Budczies J, et al. Tumor mutational burden as a predictive biomarker in solid tumors. Cancer Discov. 2020 Dec;10(12):1808–1825. doi: 10.1158/2159-8290.CD-20-0522
33. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Sci. 2018 Feb 23;359(6378):926–930. doi: 10.1126/science.aar3247
34. Kim M, Park J, Seonghee O, et al. Deep learning model integrating cfDNA methylation and fragment size profiles for lung cancer diagnosis. Sci Rep. 2024 Jun 26;14(1):14797. doi: 10.1038/s41598-024-63411-2
35. Shi J, Zhang R, Li J, et al. Size profile of cell-free DNA: a beacon guiding the practice and innovation of clinical testing. Theranostics. 2020;10(11):4737–4748. doi: 10.7150/thno.42565
36. Stejskal P, Goodarzi H, Srovnal J, et al. Circulating tumor nucleic acids: biology, release mechanisms, and clinical relevance. Mol Cancer. 2023 Jan 21;22(1):15. doi: 10.1186/s12943-022-01710-w
37. Seres M, Spacayova K, Sulova Z, et al. Dynamic multilevel regulation of EGFR, KRAS, and MYC oncogenes: driving cancer cell proliferation through (Epi)genetic and post-transcriptional/translational pathways. Cancers (Basel). 2025 Jan 14;17(2):248. doi: 10.3390/cancers17020248
38. Adalsteinsson VA, Ha G, Freeman SS, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017 Nov 6;8(1):1324. doi: 10.1038/s41467-017-00965-y
39. Shao X, Lv N, Liao J, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019 Nov 9;20(1):175. doi: 10.1186/s12881-019-0909-5
40. Bjaanæs MM, Nilsen G, Halvorsen AR, et al. Whole genome copy number analyses reveal a highly aberrant genome in TP53 mutant lung adenocarcinoma tumors. BMC Cancer. 2021 Oct 9;21(1):1089. doi: 10.1186/s12885-021-08811-7
41. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014 Feb 19;6(224):224ra24.
42. Meshkov IO, Koturgin AP, Ershov PV, et al. Diagnostics of lung cancer by fragmentated blood circulating cell-free DNA based on machine learning methods. Front Med. 2025;12:1435428. doi: 10.3389/fmed.2025.1435428
43. Nguyen VC, Nguyen TH, Phan TH, et al. Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. BMC Cancer. 2023 Mar 13;23(1):233. doi: 10.1186/s12885-023-10681-0
44. Liu X, Liu L, Ji Y, et al. Enrichment of short mutant cell-free DNA fragments enhanced detection of pancreatic cancer. EBioMedicine. 2019 Mar 1;41:345–356. doi: 10.1016/j.ebiom.2019.02.010
45. Maansson CT, Thomsen LS, Stokkebro L, et al. In vitro size-selection of short circulating tumor DNA fragments from late-stage lung cancer patients enhance the detection of mutations and aneuploidies. J Liquid Biopsy. 2024 Jun 1;4:100141. doi: 10.1016/j.jlb.2024.100141
46. Shao SH, Allen B, Clement J, et al. Multi-cancer early detection test sensitivity for cancers with and without current population-level screening options. Tumori. 2023 Jun;109(3):335–341. doi: 10.1177/03008916221133136
47. Barta JA, Mazzone PJ, Nair VS. Multi-cancer and single-cancer early detection testing: opportunities and challenges. Chest. 2024 Sep;166(3):425–428. doi: 10.1016/j.chest.2024.03.044
48. Alese OB, Cook N, Ortega-Franco A, et al. Circulating tumor DNA: an emerging tool in gastrointestinal cancers. Am Soc Clin Oncol Educ Book. 2022;(42):279–298. doi: 10.1200/EDBK_349143