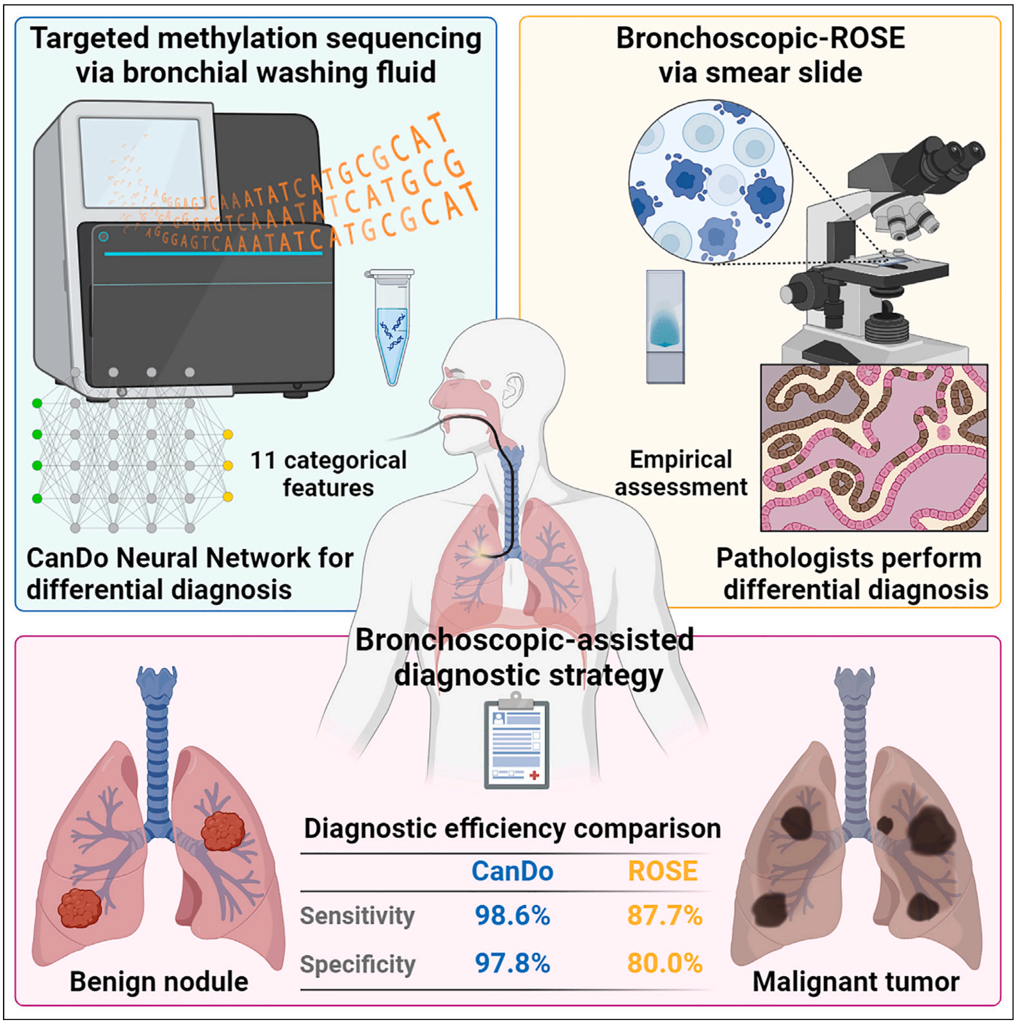
iScience

CellPress
OPEN ACCESS

Article

Lung tumor discrimination by deep neural network model CanDo via DNA methylation in bronchial lavage


Targeted methylation sequencing via bronchial washing fluid
Bronchoscopic-ROSE via smear slide
11 categorical features
Empirical assessment
CanDo Neural Network for differential diagnosis
Pathologists perform differential diagnosis
Bronchoscopic-assisted diagnostic strategy
Diagnostic efficiency comparison
CanDo
ROSE
Sensitivity
98.6%
87.7%
Specificity
97.8%
80.0%
Benign nodule
Malignant tumor

Zezhong Yu, Jieyi Li, Yi Deng, ..., Ziying Gong, Daoyun Zhang, Xin Zhang

gzy@yunyingmedicine.com (Z.G.)
zdy@yunyingmedicine.com (D.Z.)
xinhaier@sina.com (X.Z.)

Highlights

Meta-analysis found 13 genes significant in methylation for distinguishing lung cancer

Multiplex PCR with targeted methylation sequencing enhanced detection sensitivity

BWF methylation test had 98.6% sensitivity, 97.8% specificity in lung tumor diagnosis

TBR1 gene methylation may be important in the occurrence of lung cancer

Yu et al., iScience *27*, 110079
June 21, 2024 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110079

iScience

CellPress
OPEN ACCESS

Article

Lung tumor discrimination by deep neural network model CanDo via DNA methylation in bronchial lavage

Zezhong Yu,[1,7] Jieyi Li,[2,3,7] Yi Deng,[1] Chun Li,[1] Maosong Ye,[1] Yong Zhang,[1] Yuqing Huang,[3] Xintao Wang,[2,3] Xiaokai Zhao,[2,3] Jie Liu,[1] Zilong Liu,[1] Xia Yin,[1] Lijiang Mei,[1] Yingyong Hou,[4] Qin Hu,[1] Yao Huang,[5] Rongping Wang,[5] Huiyu Fu,[3] Rumeng Qiu,[3] Jiahuan Xu,[3] Ziying Gong,[2,3,6,*] Daoyun Zhang,[2,3,6,*] and Xin Zhang[1,5,8,*]

SUMMARY

Bronchoscopic-assisted discrimination of lung tumors presents challenges, especially in cases with contraindications or inaccessible lesions. Through meta-analysis and validation using the HumanMethylation450 database, this study identified methylation markers for molecular discrimination in lung tumors and designed a sequencing panel. DNA samples from 118 bronchial washing fluid (BWF) specimens underwent enrichment via multiplex PCR before targeted methylation sequencing. The Recursive Feature Elimination Cross-Validation and deep neural network algorithm established the CanDo classification model, which incorporated 11 methylation features (including 8 specific to the *TBR1* gene), demonstrating a sensitivity of 98.6% and specificity of 97.8%. In contrast, bronchoscopic rapid on-site evaluation (bronchoscopic-ROSE) had lower sensitivity (87.7%) and specificity (80%). Further validation in 33 individuals confirmed CanDo's discriminatory potential, particularly in challenging cases for bronchoscopic-ROSE due to pathological complexity. CanDo serves as a valuable complement to bronchoscopy for the discriminatory diagnosis and stratified management of lung tumors utilizing BWF specimens.

INTRODUCTION

The global cancer statistics highlight that lung cancer accounts for 12.2–27.2% of the total annual incidence of malignant tumors worldwide,[1,2] representing a significant 18% of cancer-related mortalities and leading to 1.8 million deaths annually.[3] Despite the 5-year survival rate of less than 13% for patients diagnosed with advanced-stage lung cancer,[4] the active promotion of low-dose chest computed tomography (LDCT) screening has contributed to a significant increase in the 5-year survival rate, rising from 6% for distant-stage disease to 33% for regional stage and 60% for localized-stage disease.[5,6] However, 20–50% of individuals undergoing LDCT screening may detect lung tumors with a diameter of less than 3 cm, and as much as 96.4% of these positive tumors are benign non-neoplastic lesions.[7,8] This necessitates careful, followed steps in clinical stratification management decisions to mitigate the risk of overdiagnosis, preventing patients from bearing unnecessary surgical and medication risks.[9,10]

A viable solution is to employ additional auxiliary diagnostic techniques, particularly molecular diagnostics based on genomic methylation biomarkers, which offer significant advantages in this context. In comparison to genomic mutations, alterations in the DNA methylation patterns of tumor cells not only occur earlier but also demonstrate a high degree of cell-type specificity.[11,12] This theoretically endows methylation markers with both tumor discrimination and cell lineage tracing capabilities, earning them the designation of "molecular fingerprints". Numerous studies have successively reported a series of potential methylation markers applicable to lung cancer diagnosis.[13] For instance, Rosa et al. conducted a small sample-size study involving the simultaneous detection of 10 methylation markers, achieving a sensitivity of up to 73% for the diagnosis of early-stage lung cancer in blood samples.[14] Additionally, Guo et al. confirmed the ability to accurately differentiate between lung cancer and colorectal cancer in a cohort of 59 cancer patients through the methylation patterns of plasma cfDNA.[15]

[1]Department of Respiratory Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[2]Jiaxing Key Laboratory of Precision Medicine and Companion Diagnostics, Jiaxing Yunying Medical Inspection Co., Ltd., Jiaxing 314033, China
[3]Department of R&D, Zhejiang Yunying Medical Technology Co., Ltd., Jiaxing 314006, China
[4]Department of Pathology Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[5]Department of Pulmonary, Taikang Xianlin Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210046, China
[6]Department of R&D, Shanghai Yunying Biopharmaceutical Technology Co., Ltd., Shanghai 201612, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: gzy@yunyingmedicine.com (Z.G.), zdy@yunyingmedicine.com (D.Z.), xinhaier@sina.com (X.Z.)

https://doi.org/10.1016/j.isci.2024.110079

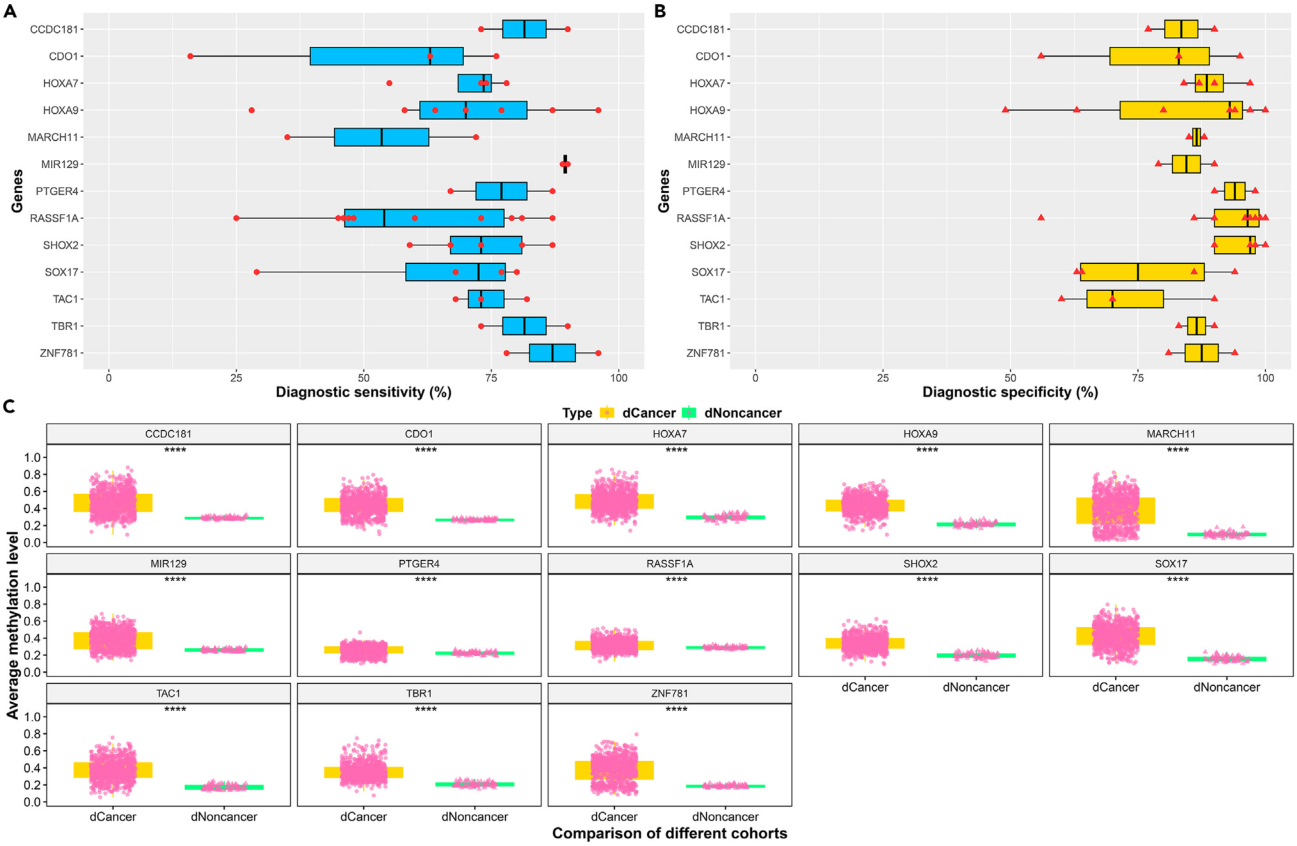


Figure 1. Panel design and database validation

(A) Diagnostic sensitivity reported in the literature for the 13 selected genes, meta-analysis; (B) Diagnostic specificity reported in the literature for the 13 selected genes, meta-analysis; (C) Average methylation levels of the 13 selected genes in public databases among lung tumor tissue and normal tissue adjacents (i.e., the average methylation level of all 450K BeadChip cg probes covered by each gene). Data are represented as mean ± SEM, and significance testing was performed using the Student's *t* test. ****, $p < 0.00005$.

Notably, the low abundance of circulating free DNA (cfDNA) derived from tumor cells in blood poses a major challenge to sensitivity in the detection process.[16] Another issue in plasma specimens is the heterogeneous interference of clonal hematopoietic mutations.[17] In comparison, bronchial washing fluid (BWF) obtained through saline flushing during bronchoscopy holds obvious value for companion diagnostic applications. Direct contact with the lesion and bypassing the circulatory process allows for the maximization of recovery of various cellular detritus and exudates, thereby obviously contributing to improved detection rates and accuracy. Furthermore, the collection approach for BWF is entirely practical for patients with contraindications to biopsy procedures, such as bleeding or pneumothorax, effectively avoiding additional pain and potential risks associated with invasive diagnostic interventions.[18,19]

In this study, a customized targeted methylation sequencing panel was developed to analyze the methylation characteristics of DNA molecules in BWF specimens. Utilizing the machine learning neural network algorithm for methylation feature selection, a classification model named CanDo (cancer diagnosis) was established, and a comprehensive comparison to evaluate the consistency among bronchoscopic rapid on-site evaluation (bronchoscopic-ROSE), CanDo, and histopathological diagnosis from the biopsy. We proposed a clinical strategy for auxiliary diagnosis of lung malignant tumors, aiming to guide more precise patient stratification management.

RESULTS

Candidate genes for the diagnosis of lung malignant tumors

The methylation patterns of 79 genes within 23 selected publications were considered as potential diagnostic markers for lung malignant tumors (Table S1), and 13 of them were included in the candidate panel due to being mentioned in at least two literatures. Among these, 12 genes exhibited excellent diagnostic specificity exceeding 90% when applied independently or in combination. The exception was *MARCH11*, which achieved 88%. However, it's important to note that five genes (*HOXA9*, *ZNF781*, *CCDC181*, *MIR129*, and *TBR1*) reached a sensitivity level of 90% or higher. Moreover, only *HOXA9*, *ZNF781*, and *MIR129* demonstrated diagnostic sensitivity and specificity exceeding 90% in the same study (Figures 1A and 1B).

To comprehensively evaluate the diagnostic relevance of these 13 genes in the context of malignant lung tumors, we analyzed their methylation patterns in both lung tumor tissues and adjacent non-cancerous specimens. Publicly available sequencing data from the HumanMethylation450 (450K) BeadChip platform, encompassing over 480,000 methylation sites across the human genome, were utilized. Of these, 356 CpG sites were mapped onto the 13 selected genes, with only 18 (5.1%) showing no significant difference in methylation levels between lung tumors and noncancer tissues (Table S2).

Additionally, we determined the average methylation levels for these 13 genes by calculating the mean values of methylation rates from all CpG probes targeting them. Our analysis revealed statistically significant differences in the methylation levels of these genes between lung malignant tumor tissues and their noncancer counterparts ($p < 0.00005$, Figure 1C, and Table S2), highlighting their robust potential as diagnostic biomarkers.

NGS-panel design and methodological assessment

Conducting a large-scale examination of CpG sites in a single sweep undoubtedly enhances the specificity and sensitivity of lung malignant tumor diagnosis. However, the increased cost may hinder the clinical feasibility of such testing. To strike an optimal balance between precision and cost-effectiveness, an efficient targeted amplification strategy was developed. In brief, segments of CpG-rich promoter regions from 13 genes that obtained from previous literature meta-analysis were carefully selected, with each fragment approximately 150 base pairs in length. Following two rounds of targeted enrichment via multiplex PCR, next-generation sequencing (NGS) was employed to analyze the DNA methylation levels of these genes. The advantage of this approach lies in its significantly improved cost-effectiveness due to the shorter and limited fragment lengths. Additionally, libraries enriched through this method effectively enhance sequencing depth and signal-to-noise ratio. Ultimately, a panel for further studies was customized, encompassing 13 genes and 151 CpG sites (Table S3).

To evaluate the methodological stability of the detection strategy, we divided BWF samples obtained from bronchoscopy examinations of 6 volunteer donors into two equal portions. One portion underwent processing within 24 h of collection and the other portion was stored in storage tubes at room temperature for 7 days before further processing. The differences in DNA methylation rates among samples subjected to each procedures were then assessed. Results demonstrated that, under standardized protocols and controlled environmental conditions, samples stored in lavage fluid storage tubes remained highly consistent with fresh samples even after 7 days at room temperature (Fig. S1). This indicates that the detection strategy is more robust to variations in sample storage time, effectively extending the sample preservation window.

Patient enrollment

From 2022 to 2023 in Zhongshan Hospital, 118 patients with suspicious lung tumors detected through CT scans and required bronchoscopy-guided biopsies were enrolled in this study. All participants strictly adhered to inclusion and exclusion criteria (Supplementary Methods), and detailed records of their baseline information were kept (Tables 1, and S4). Notably, following the diagnostic gold standard pathological examination, 73 cases were ultimately diagnosed with malignancies, while 45 were classified as benign cases (Table S4).

Methylation characteristics of lung malignant tumors

Out of the 118 clinical bronchoscopy lavage fluid specimens collected, the methylation patterns of three genes (*CDO1*, *SOX17*, and *MARCH11*) were not fully detected in over 10% of the samples. To ensure the robustness of subsequent diagnostic discrimination modeling, we excluded these three genes and proceeded with the remaining 115 CpGs (located on 10 genes) as the final testing panel. Our findings revealed a significant trend of increased methylation levels at most CpG sites in the 73 cancer patients compared to the 45 patients diagnosed with non-cancerous conditions through pathology (Figure S2, EXdata 1. Mendeley Data: https://data.mendeley.com/datasets/wcnzyth6vd).

To further explore the methylation characteristics distinguishing the cancer and noncancer cohorts, we computed the average methylation level of all CpG sites within each gene in the panel, defining it as the gene average methylation rate. Consistent with the findings from our previous analysis in the public database, hypomethylation was observed in all 10 genes in the cancer group. In contrast, the noncancer group exhibited lower methylation rates, showing a highly significant statistical difference ($p < 0.00005$, Figure 2A, EXdata 2. Mendeley Data: https://data.mendeley.com/datasets/wcnzyth6vd).

For a more in-depth examination of whether the observed statistical significance difference could be attributed to small sample bias resulting from our limited sample size, we conducted further analysis using publicly available DNA methylation data based on the HumanMethylation450 BeadChip. Through probe chromosome coordinate alignment, we identified a total of 13 CpG sites spanning seven genes included in our sequencing panel (Table S2). The methylation status of these 13 CpG sites between our dataset and a public database containing 837 lung cancer tissue samples and 74 adjacent non-cancer tissue samples were compared. The results revealed that, in both our dataset and the public database, these 13 CpG sites exhibited significantly elevated methylation levels in the cancer group compared to the noncancer group ($p < 0.00005$, Figure 2B). This suggests that the methylation pattern changes at these sites were not solely due to small sample effects. These findings underscore the potential of these CpG sites as promising molecular biomarkers for distinguishing lung malignant tumors in diagnostic applications.

Establishment and application evaluation of CanDo model

Individual receiver operating characteristic (ROC) curves were initially constructed for each gene, and optimal cut-off values were determined to evaluate the diagnostic performance of each gene separately. Notably, the *TBR1* gene exhibited outstanding performance, demonstrating significant discriminative ability at a methylation level cut-off of 0.2, with 86% sensitivity and 98% specificity (Table S5, and Figure 3A).

Table 1. Patient characteristics

Characteristics	Patient number (%)
All participants	118 (100)
Age at diagnosis (years)	
<65	49 (41.5)
≥65	69 (58.5)
Gender	
Male	75 (63.6)
Female	43 (36.4)
Tobacco Habits	
Smoker	45 (38.1)
Nonsmoker	73 (61.9)
Imaging nodule length (mm)	
<5	71 (60.2)
≥5	18 (15.3)
Pathological	
Malignant	73 (61.9)
Benign	45 (38.1)
Tumor Stage	
I-III	30 (25.4)
IV	34 (28.9)

To further enhance the discriminative diagnostic capability, two dimensions were selected for training the machine learning classification models: the methylation level at 115 individual CpG sites and the average methylation level of all methylated sites covered by each gene fragment (defined as the average methylation level of the gene). The recursive feature elimination (RFE) algorithm was employed to rank the importance of all 125 features based on their contribution to the classification ability. Finally, the top 11 feature combinations, displaying the highest classification accuracy, were determined as the optimal feature set by the cross-validation algorithm (Figure S3).

Subsequently, four types of machine learning classifiers were trained and compared: the least absolute shrinkage and selection operator (LASSO), the support vector machine (SVM), the extreme gradient boosting (XGBOOST), and the deep neural network (DNN). Due to the DNN algorithm exhibiting the highest area under the curve (AUC) and achieving the highest F1 score, indicating the most robust performance (Figure S4), it was selected to construct the diagnostic model named CanDo. Briefly, CanDo utilized these top 11 features to score and evaluate the probability of malignant tumors. When compared with pathological diagnostic results, the CanDo score achieved diagnostic sensitivity and specificity of 98.6% (72/73) and 97.8% (44/45), respectively. This represents an improvement of approximately 10–20% compared to bronchoscopic-ROSE diagnosis, which presented a sensitivity of 87.7% (64/73) and specificity of 80% (36/45).

Interestingly, within these 11 features, two correspond to CpG sites situated on *SHOX2*, one is linked to a CpG site on *PTGER4*, and the remaining eight features are all intricately connected to the *TBR1* gene (comprising 7 CpG sites on *TBR1* and the average *TBR1* methylation level). This alignment with the previous outstanding ROC curve in lung malignant tumor diagnosis by the *TBR1* single gene, strongly suggests the potential significance of *TBR1* in the discriminative diagnosis of malignant lung tumors.

Potential effects on CanDo diagnostic performance across various baselines

Following the grouping based on pathological diagnostic results, a non-uniform distribution of clinical baseline characteristics was observed between the two comparative cohorts. Due to these pronounced deviations, it is imperative to assess whether other potential confounding factors within the sample cohorts might influence the discriminative capacity of the CanDo model. Consequently, we investigated the correlation between the clinical baseline information of all participants and CanDo scores. The analysis revealed that either methylation level or CanDo scores exhibited the highest correlation with pathological diagnosis, and no statistically significant inter-group differences were observed in terms of patient gender and tobacco habits. However, both methylation level and CanDo scores displayed a highly significant inter-group difference between the elderly (≥65) and younger (<65) age groups, s suggesting that age could potentially significantly affect the discriminative capability of the CanDo model (Figure S5 and Figure 3C).

In order to precisely evaluate the impact of age on CanDo diagnostic ability, we further analyzed correlations between patient baseline characteristics, pathological diagnosis, model scores, and the methylation levels of each feature within the model. Consistent with previous

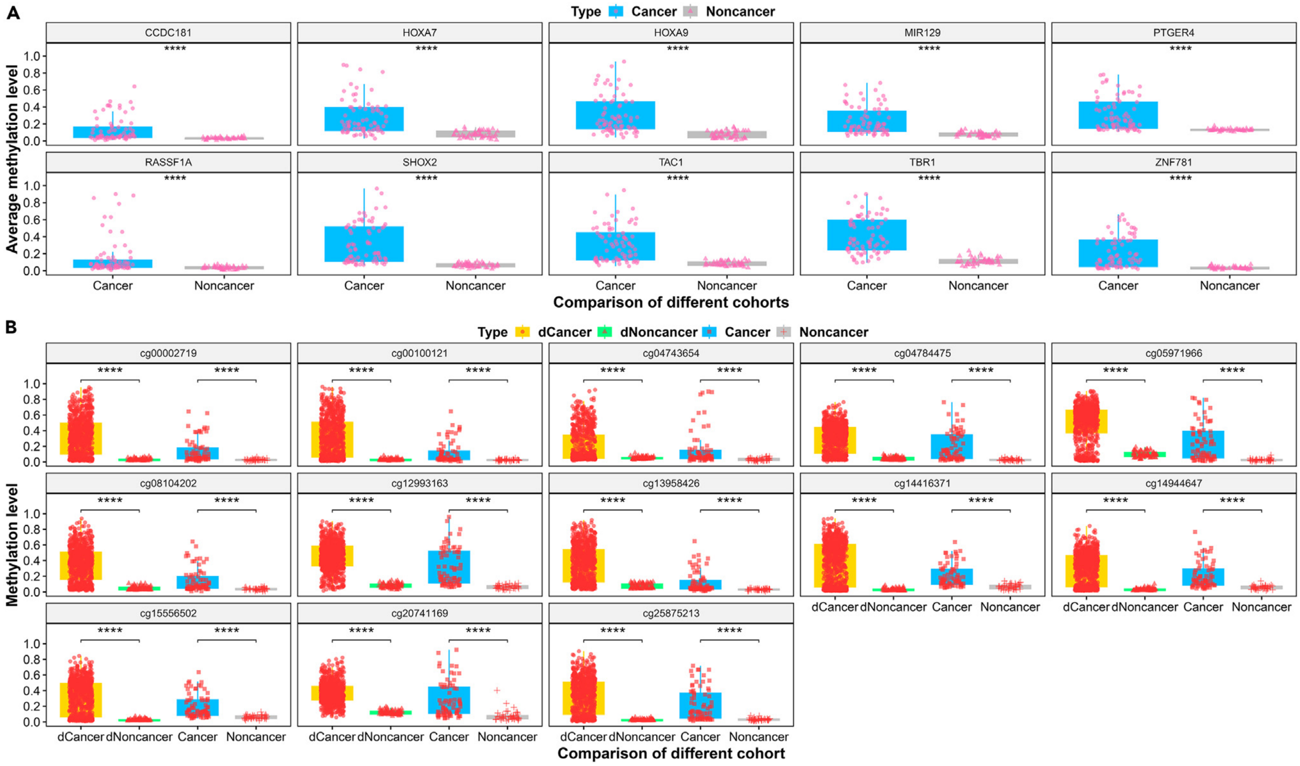


Figure 2. Methylation detection results of the 10-gene panel and database validation

(A) Comparison of the average methylation levels of the 10 genes in the panel between the cancer and non-cancer groups (i.e., the average methylation level of all methylated sites covered by each gene fragment); (B) Comparison of the methylation levels quantified in the panel detection results of 13 common CpG sites covered by both the panel and the public database's 450K BeadChip cg probes. Cancer, cancer cohort in this study; Noncancer, non-cancer cohort in this study; dCancer, lung cancer samples in the public database; dNoncancer, lung normal tissue adjacent samples in the public database. Data are represented as mean ± SEM, and significance testing was performed using the Student's *t* test. ****, $p < 0.00005$.

analyses, gender and smoking history showed low correlation with CanDo diagnostic, while age exhibited a higher correlation with both pathological diagnosis and CanDo scores. It is noteworthy that a highly significant strong correlation among the methylation patterns of the 11 features was observed, whereas the correlation between age and all 11 features did not reach statistical significance (Figure 3D, and Table S6). This suggests that despite prominent age disparities between the two cohorts, age does not seem to significantly interfere with the functionality of the model.

Considering the noted molecular characteristic differences between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) patients, such as gene expression profiles, the potential diagnostic performance of the CanDo model among patients with different subtypes was investigated. It was found that the mean or median CanDo scores did not exhibit significant differences between LUAD or LUSC patients. Furthermore, no significant differences in the methylation levels of these 11 features were observed among different pathological subtypes (Figure S6). The methylation levels of these 11 features appeared to increase with tumor stage, although most of these differences did not reach statistical significance. However, stage IV tumors had higher CanDo scores than stage III tumors and showed statistical significance ($p < 0.005$), although no significant differences were observed in earlier tumor stages (stage I or II) due to the limited sample size in the cohort (Figure S7). These findings suggest that CanDo scores are mainly associated with the benign or malignant nature of suspicious nodules, with higher scores potentially correlated with advanced tumor stages. However, CanDo scores are less influenced by pathological subtypes or other factors, making it a relatively robust model with clinical utility for assisting in diagnosis.

Independent and single-blind validation of CanDo diagnostic accuracy

To rigorously control confounding factors and demonstrate their limited impact on the CanDo model, a prospectively designed, single-blind, unbiased independent validation cohort with more stringent inclusion criteria was recruited (Figure S8). In total, 33 participants with no history of prior smoking were enrolled (Table S7). Among them, 22 patients were pathologically diagnosed with cancer, and the remaining 11 served as age- and sex-matched controls. Briefly, a balanced gender ratio (cancer, 11:11; noncancer, 5:6) and consistent age distribution (cancer: max 77, min 36, average 60.1, median 60.5; noncancer: max 78, min 33, average 60.2, median 60) were achieved in two cohorts, which is sufficient to evaluate diagnostic accuracy with desired statistical errors.

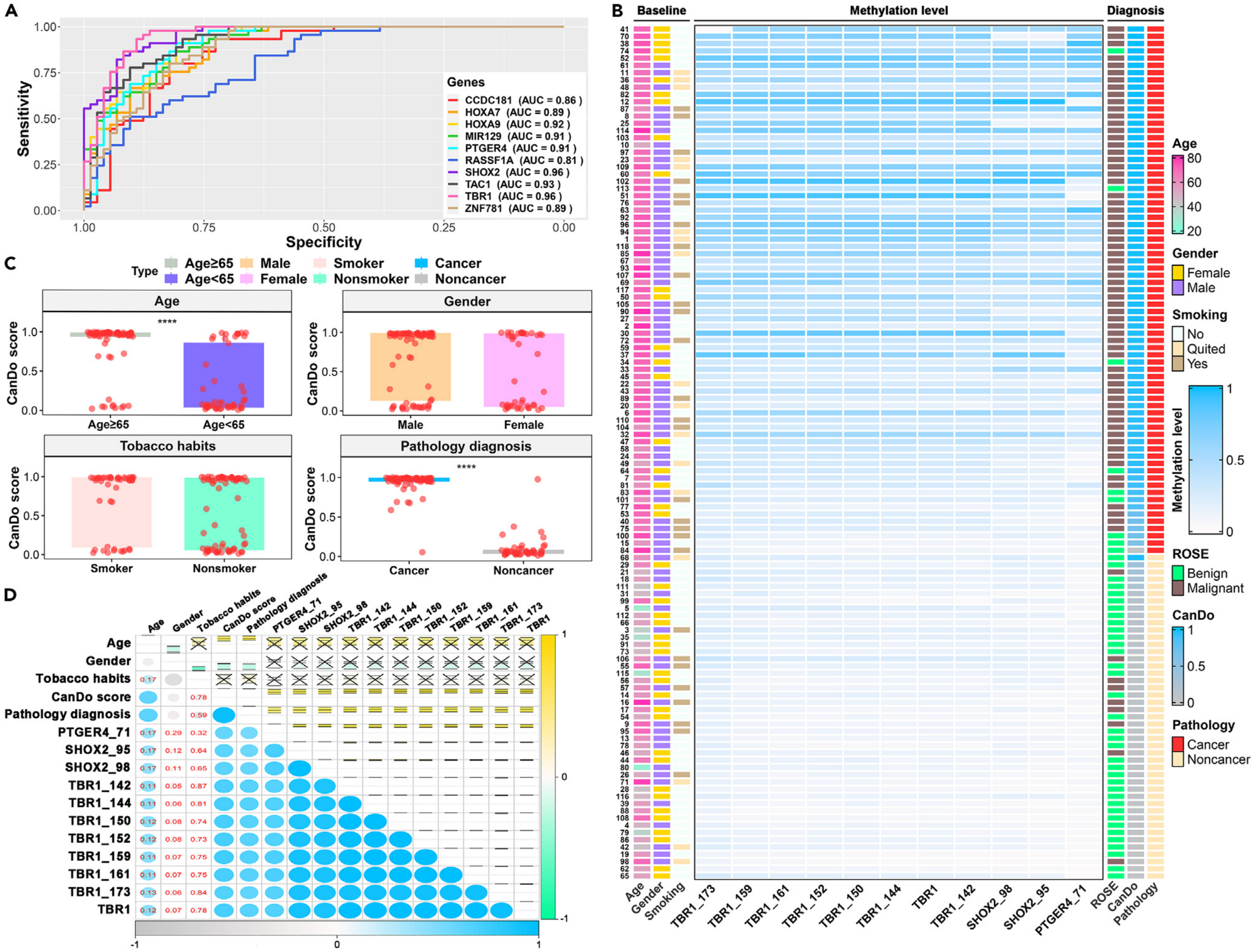


Figure 3. Characteristic analysis of the CanDo model

(A) ROC curves using the average methylation levels of the 10 genes individually for distinguishing between benign and malignant lung tumors (i.e., the average methylation level of all methylated sites covered by each gene fragment); (B) Heatmap comparing the methylation levels of the 11 CanDo model features with clinical baselines, bronchoscopic-ROSE, CanDo scores, and pathological diagnosis results; (C) Correlation between CanDo scores and baselines in different groups, data are represented as mean ± SEM, significance testing was performed using the Wilcoxon rank-sum test; (D) Heatmap showing the correlation between the 11 CanDo model features and baselines. The lower left corner represents the magnitude of the correlation coefficients, and *p*-values are highlighted in red font when the correlation is not significant. The upper right corner represents the 95% confidence interval range, marked with an "×" when not significant. ****, $p < 0.00005$.

In this rigorously controlled cohort, CanDo displayed a diagnostic sensitivity of 95.5% (21/22) and specificity of 100% (11/11). In comparison, the sensitivity and specificity of bronchoscopic-ROSE maintained 86.4% (19/22) and 81.8% (9/11), respectively (Figure 4A, EXdata 3, EXdata 4. Mendeley Data: https://data.mendeley.com/datasets/wcnzyth6vd). CanDo demonstrated an improvement of approximately 10–20% in both sensitivity and specificity, affirming its consistent and enhanced companion diagnostic performance in lung malignant tumors.

Valuable case reports on clinical applications

In five cases of misjudgment during bronchoscopic-ROSE, one case was also misjudged by CanDo. We carefully assessed the remaining four cases, in which CanDo made accurate diagnoses while ROSE misdiagnosed. Among them, two true benign cases were misjudged as malignant due to various interfering factors. For example, in Case 1 (ID129), a 51-year-old male exhibited a suspicious lesion in the right middle lobe on CT (Figures 4B–1). ROSE initially reported squamous epithelial proliferation as malignant (Figures 4B–2). However, subsequent histopathology diagnosed benign squamous epithelial proliferation with chronic inflammation (Figure 4B-3). Similarly, in Case 2 (ID143), a 70-year-old female with a CT-suggested irregular patchy lesion in the right middle lobe was initially considered suspicious for squamous cell carcinoma (Figures 4B and 5), despite histopathology ultimately confirming atypical squamous epithelial proliferation (Figures 4B–6).

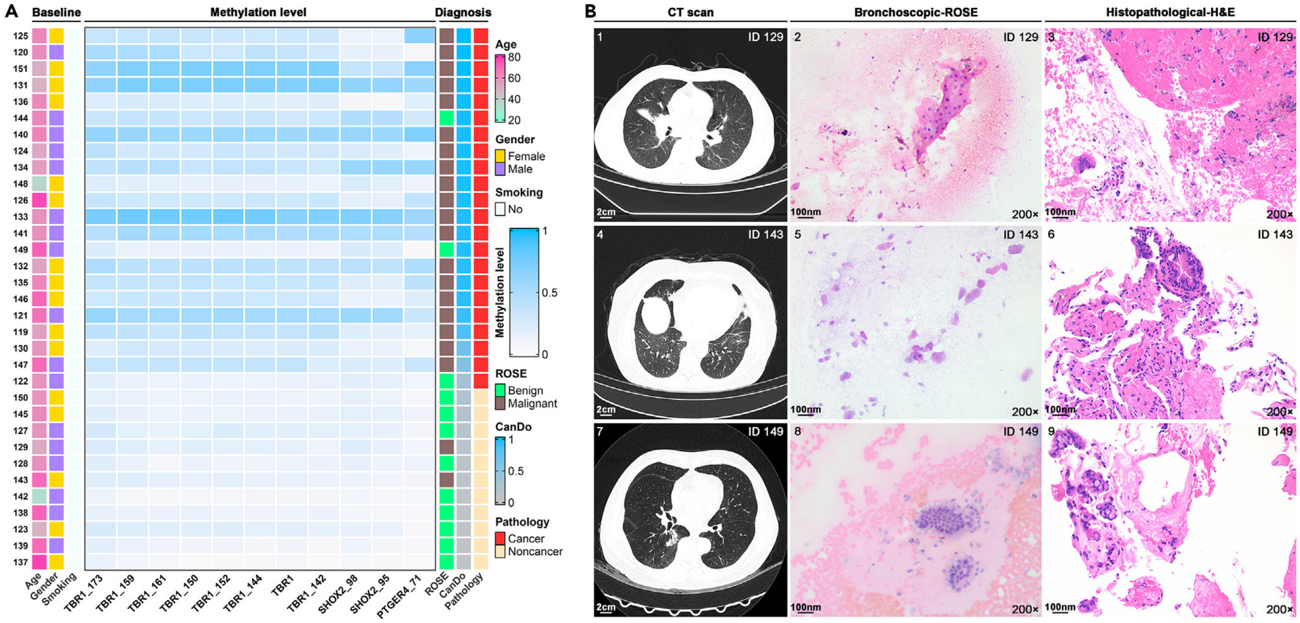


Figure 4. Validation and Clinical Application Case Reports of the CanDo Model

(A) Heatmap comparing the methylation levels of the 11 CanDo model features with clinical baselines, bronchoscopic-ROSE, CanDo scores, and pathological diagnosis results in the validation cohort; (B) CT scan image, bronchoscopic-ROSE image, and histopathological image of three representative cases. The length represented by the scale bar is indicated on each figure: for CT images (numbered 1, 4, and 7), the scale bar represents 2 cm; for the remaining microscopic images enlarged 200 times, the scale bar represents 100 nm.

Notably, both true malignant cases were misjudged by ROSE as benign due to the absence of clear malignant tumor cells. A representative case is Case 3 (ID 149), a 72-year-old male with a CT scan indicating suspicious shadows in the right lower lobe (Figures 4B–7). Bronchoscopy with peripheral ultrasound in the back segments b6a and b6b of the right lower lobe revealed heterogeneous echoes. ROSE microscopy showed a small amount of proliferative mucous gland epithelial cells, initially assessed as benign proliferation (Figures 4B–8). Histopathological biopsy revealed infiltration of a small number of inflammatory cells and concurrent malignant lung adenocarcinoma cells, confirming the diagnosis of lung adenocarcinoma (Figures 4B–9). These findings underscore the clinical value and potential application of CanDo as an alternative to ROSE in aiding the differential diagnosis during bronchoscopic examinations, particularly in cases where proliferation or the presence of malignant cells is not readily apparent.

DISCUSSION

Methylation patterns serve as both molecular fingerprints for cells and relatively stable signals within the genome.[20] Methylated cytosine stabilizes the DNA helix, specifically increasing the DNA melting temperature, thereby effectively extending the half-life of DNA in clinical samples.[21] Due to being perfectly devoid of additional adverse risks such as hemolysis, the role of this molecular modification in enhancing structural stability is particularly evident in BWF samples. Our study observed that BWF specimens maintained highly consistent DNA methylation levels even after prolonged storage in room temperature, despite the complexity of collection process compared to the convenience of blood. Therefore, BWF not only offered a higher initial concentration of target molecules[19] but also extended preservation duration, which is advantageous in supporting precision in detection capabilities.

Given the promising potentials exhibited by genomic methylation, numerous studies have attempted clinical tumor screening and differential diagnosis using whole-genome methylation maps or single-gene methylation-specific PCR (MS-PCR).[22,23] Comparison by some researchers revealed that while whole-methylome sequencing (WMS) enables excellent resolution accuracy in detection, its limitation primarily resides in the sequencing depth, often constrained to 30×,[24] potentially leading to the oversight of rare methylation signals originating from early-stage tumors. Additionally, its high cost impedes its broader clinical translational application and widespread implementation. Single-gene MS-PCR detection can effectively amplify target signals, and its low-cost favors rapid clinical deployment; however, the discovery of genetic markers with diagnostic value depends on whole-genome methylation data.[25] For example, the methylation levels of *SDC2* and *SEPTIN9* have provided auxiliary references for minimally invasive colorectal cancer screening, despite the detection sensitivity and specificity remaining relatively limited.[26,27] Here, to the best of our knowledge, we present a targeted methylation detection method for the first time, involving two rounds of multiplex PCR enrichment of target molecules followed by NGS sequencing. This approach effectively combines the strengths of existing detection methods. Trace molecular signals within samples are efficiently captured for sequencing after undergoing two rounds of PCR amplification, resulting in a sequencing depth exceeding 1000× for individual CpG sites, significantly enhancing detection

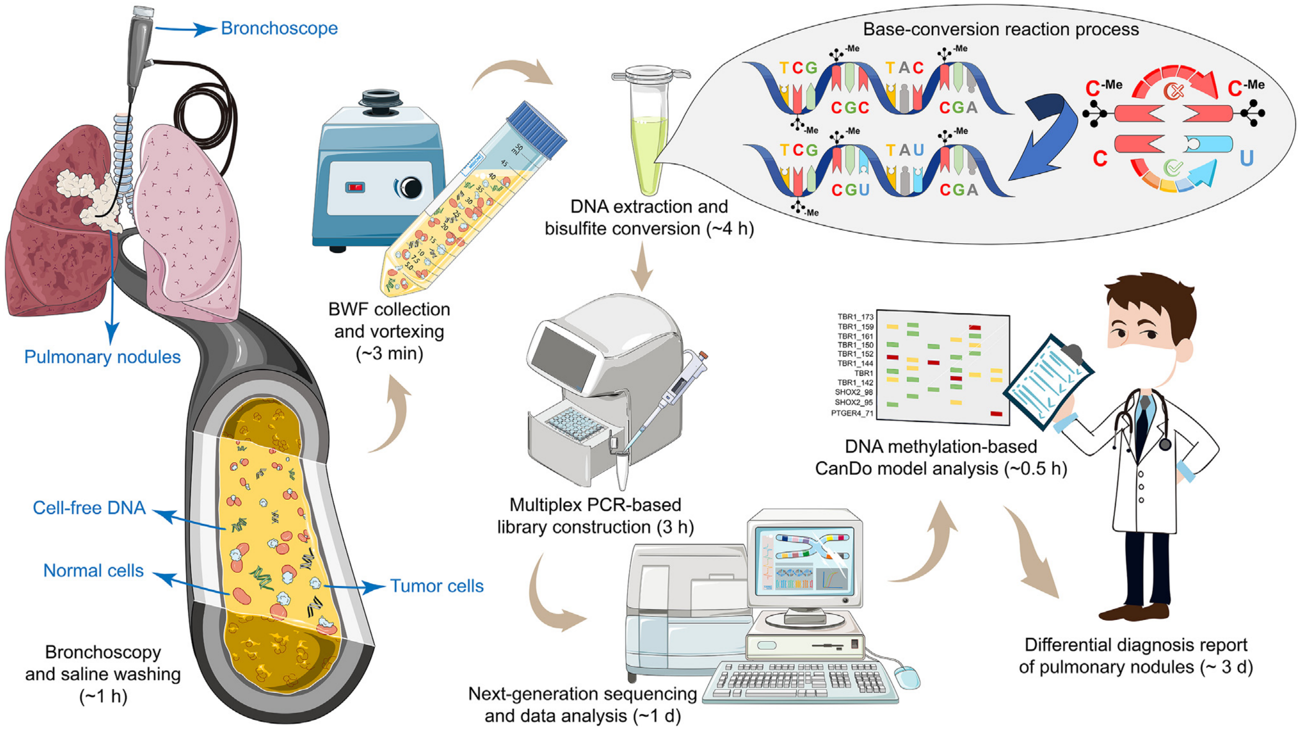

Bronchoscope
Pulmonary nodules
Cell-free DNA
Normal cells
Tumor cells
Bronchoscopy and saline washing (~1 h)
BWF collection and vortexing (~3 min)
DNA extraction and bisulfite conversion (~4 h)
Base-conversion reaction process
TCG
TAC
CGC
CGA
TCG
TAU
CGU
CGA
-Me
C-Me
C
U
Multiplex PCR-based library construction (3 h)
Next-generation sequencing and data analysis (~1 d)
TBR1_173
TBR1_159
TBR1_161
TBR1_150
TBR1_152
TBR1_144
TBR1
TBR1_142
SHOX2_98
SHOX2_95
PTGER4_71
DNA methylation-based CanDo model analysis (~0.5 h)
Differential diagnosis report of pulmonary nodules (~ 3 d)

Figure 5. Study summary

sensitivity. The three genes included in our 11 features (*TBR1*, *PTGER4*, and *SHOX2*) have been reported in previous literature with average sensitivities for lung cancer diagnosis of 81.5%, 77%, and 73.4%, and average specificities of 86.5%, 94%, and 95%, respectively (Figures 1A, 1B, and Table S1). In contrast, our method has a sensitivity of 98.6% and a specificity of 97.8%, demonstrating a significant improvement in diagnostic performance.

Another outstanding advantage of our targeted methylation sequencing technology lies in its cost-effectiveness during development. Building upon the meta-analysis performed before public database verification, this study promptly focused on a concise yet potent gene set, avoiding the use of costly WMS and escaping the tedious process of validating tens or even hundreds of thousands of potentially valuable CpG sites directly sourced from databases. It is essential to recognize that the traditional approach enables a more comprehensive identification of methylation markers and achieves a detection sensitivity of 52–81% in early lung cancer screening.[28] However, classic strategies like WMS can lead to unnecessary resource wastage and economic burden for BWF-based auxiliary diagnostic techniques during bronchoscopy intervention. Hence, our approach provides a valuable model reference for the development of simplified, efficient, and cost-effective rapid response technologies tailored to address specific clinical challenges.

After excluding three genes with insufficient detection quality, this study included a total of 115 CpG sites covered 10 genes. Among them, the CanDo diagnostic established based on the neural network classifier suggested that 11 features associated with *SHOX2*, *PTGER4*, and *TBR1* genes were sufficient to support BWF-dependent precise discrimination of malignant lung tumors. This contributes to the subsequent development of a more affordable and convenient MS-PCR detection method, aiming to further reduce the cost of single tests while enhancing the clinical application and widespread implementation value of CanDo. Notably, among these 11 features, 8 (comprising 7 CpG sites and the average methylation level) were significantly correlated with the *TBR1* (T-box brain transcription factor 1) gene. Although *TBR1* expression is well-documented for its crucial role in brain development,[29,30] its hypermethylation and association with lung cancer occurrence remain unclear, despite being previously identified as one of the joint diagnostic molecular markers for lung cancer.[31] This study, for the first time, underscores the importance of *TBR1* in the occurrence of lung cancer, warranting further research to elucidate specific molecular mechanisms.

Age-related biases between cohorts may influence outcomes, particularly given the strong correlation between age and genomic methylation levels.[32,33] Unfortunately, age stands as a known risk factor for lung cancer,[34] leading to heightened detection rates among older individuals. Consequently, our study encountered inevitable age discrepancies between cancer and non-cancer groups among bronchoscopy participants. Believing that cohorts with baseline biases better mirrored real-world scenarios, we opted against manually matching and balancing various baseline characteristics across cohorts, despite the potential statistical advantages in controlling confounding factors.[35] The established CanDo model demonstrated 98.6% sensitivity and 97.8% specificity, marking a 10–20% improvement over bronchoscopic-ROSE (sensitivity 87.7%, specificity 80%, similar with previous reports[36]) in this context. While subsequent analyses suggested a significant correlation between age groups and CanDo scores, further detailed examinations of the correlation between age and specific

CpG sites raised doubts about the impact of age bias on CanDo scores. Validation using an impartial cohort reaffirmed that CanDo scores primarily reflect patient pathology rather than other baseline confounding factors. In this validation, neither the diagnostic accuracy of bronchoscopic-ROSE nor CanDo showed significant changes, with CanDo maintaining a 10–20% improvement (sensitivity 95.5% vs. 86.4%, specificity 100% vs. 81.8%).

In conclusion, this study has developed a cost-effective research strategy, investigating 115 methylation sites on 10 genes. We established the CanDo diagnostic method with 11 features. Additionally, CanDo demonstrates a sensitivity of 95.5–98.6% and specificity of 97.8–100% in discriminating between benign and malignant lung tumors based on BWF specimens. Compared to the sensitivity and specificity of bronchoscopic-ROSE diagnosis, CanDo shows an improvement of 10–20% and excels in making accurate judgments that are challenging to diagnose with bronchoscopic-ROSE. Importantly, our results provide a method to mitigate potential risks and discomfort associated with biopsy, making it clinically valuable and offering guidance for precise patient stratification (Figure 5).

Limitations of the study

First, under conditions of oxidative stress, cytosine has the potential to convert into uracil, leading to the possibility of false positive errors in bisulfite sequencing. Second, due to the single-center study design, participants who rigorously controlled to minimize baseline deviations in the CanDo validation cohort were limited in number, which may introduce potential statistical errors. Lastly, this study did not investigate the influence of heterogeneity in cell composition or cell content in BWF samples from different sources, it is necessary to consider eliminating this potential impact in subsequent analyses. Therefore, further prospective multicenter studies with a large sample-size should be undertaken to expand the understanding of the ability of CanDo to distinguish between benign and malignant lung tumor patients. Additionally, technical interventions incorporated into the methodology are required in further investigation to establish a more robust method for homogeneous BWF sample collection, as well as mitigate the risk of cytosine unintended conversion into uracil.[37]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Panel design
  - ○ The biomedical database resource acquisition and management
  - ○ Inclusion criteria of participants
  - ○ Exclusion criteria of participants
  - ○ Patient population and diagnosis
  - ○ Bronchoscopy, bronchial washing and biopsy procedure
  - ○ Sample preprocessing and DNA extraction
  - ○ Library preparation and targeted methylation DNA sequencing
  - ○ Sequencing data processing
  - ○ Methylation feature selection using RFE and RFECV
  - ○ Machine learning classifier evaluation
  - ○ CanDo model development using DNN
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110079.

ACKNOWLEDGMENTS

The author extends gratitude to Xiaozhi Wang (Department of Bioinformatics, Jiaxing Yunying Medical Inspection Co., Ltd., Zhejiang, China) for insightful suggestions during the data analysis process. Appreciation is also expressed to Xiaolei Su (Department of Pathology, Zhongshan Hospital, Fudan University, Shanghai, China) for expert assistance in pathological slide staining, and special thanks to Qing Su (Molecular Pathology Laboratory, Jiaxing Yunying Medical Inspection Co., Ltd., Zhejiang, China) for skillfully capturing microscopic images of the pathological slides. Additionally, during the preparation of this work the authors used ChatGPT (https://chat.openai.com) in order to language editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

AUTHOR CONTRIBUTIONS

Conceptualization: Z.G., D.Z, and X.Z. (2); Data curation: Z.Y., Y.D., C.L., M. Y., Y.Z., J.L. (2), and Z. L.; Formal analysis: Z.Y., J.L.(1), X.W., and X.Z. (1); Funding acquisition: D. Z. and X.Z. (2); Investigation: Z.Y., J.L. (1), M.Y., Y. H. (1), J.L. (2), Z. L., X.Y., L. M., Q.H., Y.H. (3), H. F., R. Q., and J. X.; Methodology: Z. Y., J. L.(1), Y. H. (1), X. Z. (1), Z.G., D. Z., and X.Z. (2); Project administration: Y. H. (1), Z. G., D. Z., and X.Z. (2); Resources: C.L., X.Y., L. M., Y. H., and R. W.; Software: Z. Y., J. L.(1), and X. W.; Supervision: Y. H.(2), Z. G., D. Z., and X.Z.(2); Validation: Z. Y., J. L. (1), Y.D., Y.Z., Q.H., Y.H. (3), R. W., H. F., R. Q., and J. X.; Visualization: J. L. (1); Writing - original draft: Z. Y. and J. L. (1); Writing - review and editing: J. L. (1), D. Z., and X.Z.(2). All authors read and approved the final manuscript.

This work was supported by the Zhongshan Hospital Clinical Research Fund [2018ZSLC12], the Key Research and Development Program of Zhejiang province [2023C03057], the Zhejiang Leading Talent Entrepreneurship Project [2021R02019], the Jiaxing Leading Talent Entrepreneurship Project, and the Key Technology Innovation Projects of Jiaxing [2024BZ20002].

DECLARATION OF INTERESTS

J.L. (1), X.W., and X.Z. (1) are employed by the Jiaxing Yunying Medical Inspection Co., Ltd. and Zhejiang Yunying Medical Technology Co., Ltd. Y.H. (1), H.F., R.Q., and J.X. are employed by the Zhejiang Yunying Medical Technology Co., Ltd. Z.G. and D.Z. are employed by the Jiaxing Yunying Medical Inspection Co., Ltd., Zhejiang Yunying Medical Technology Co., Ltd., and Shanghai Yunying Biopharmaceutical Technology Co., Ltd. Besides, Z.G. and D.Z. hold shares of these companies. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: January 8, 2024
Revised: May 2, 2024
Accepted: May 20, 2024
Published: May 22, 2024

REFERENCES

1. Siegel, R.L., Miller, K.D., Wagle, N.S., and Jemal, A. (2023). Cancer statistics, 2023. CA. Cancer J. Clin. *73*, 17–48. https://doi.org/10.3322/caac.21763.
2. Zheng, R., Zhang, S., Zeng, H., Wang, S., Sun, K., Chen, R., Li, L., Wei, W., and He, J. (2022). Cancer incidence and mortality in China, 2016. Journal of the National Cancer Center *2*, 1–9. https://doi.org/10.1016/j.jncc.2022.02.002.
3. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
4. Woodard, G.A., Jones, K.D., and Jablons, D.M. (2016). Lung Cancer Staging and Prognosis. Cancer Treat Res. *170*, 47–75. https://doi.org/10.1007/978-3-319-40389-2_3.
5. Oudkerk, M., Liu, S., Heuvelmans, M.A., Walter, J.E., and Field, J.K. (2021). Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. Nat. Rev. Clin. Oncol. *18*, 135–151. https://doi.org/10.1038/s41571-020-00432-6.
6. Jonas, D.E., Reuland, D.S., Reddy, S.M., Nagle, M., Clark, S.D., Weber, R.P., Enyioha, C., Malo, T.L., Brenner, A.T., Armstrong, C., et al. (2021). Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA *325*, 971–987. https://doi.org/10.1001/jama.2021.0377.
7. National Lung Screening Trial Research Team, Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., and Sicks, J.D. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. *365*, 395–409. https://doi.org/10.1056/NEJMoa1102873.
8. Au-Yong, I.T.H., Hamilton, W., Rawlinson, J., and Baldwin, D.R. (2020). Pulmonary nodules. BMJ *371*, m3673. https://doi.org/10.1136/bmj.m3673.
9. Callister, M.E.J., Sasieni, P., and Robbins, H.A. (2021). Overdiagnosis in lung cancer screening. Lancet Respir. Med. *9*, 7–9. https://doi.org/10.1016/S2213-2600(20)30553-1.
10. Mazzone, P.J., and Lam, L. (2022). Evaluating the Patient With a Pulmonary Nodule: A Review. JAMA *327*, 264–273. https://doi.org/10.1001/jama.2021.24287.
11. Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., Yuan, Z., Yang, X., Jiang, Y., Zhang, T., et al. (2020). Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat. Commun. *11*, 3475. https://doi.org/10.1038/s41467-020-17316-z.
12. Sun, K., Jiang, P., Chan, K.C.A., Wong, J., Cheng, Y.K.Y., Liang, R.H.S., Chan, W.K., Ma, E.S.K., Chan, S.L., Cheng, S.H., et al. (2015). Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc. Natl. Acad. Sci. USA *112*, E5503–E5512. https://doi.org/10.1073/pnas.1508736112.
13. Constâncio, V., Nunes, S.P., Henrique, R., and Jerónimo, C. (2020). DNA Methylation-Based Testing in Liquid Biopsies as Detection and Prognostic Biomarkers for the Four Major Cancer Types. Cells *9*, 624. https://doi.org/10.3390/cells9030624.
14. Rosa, K. (2021). cfDNA Methylation Assay Allows for Early Lung Cancer Detection. https://www.onclive.com/view/cfdna-methylation-assay-allows-for-early-lung-cancer-detection.
15. Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., and Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat. Genet. *49*, 635–642. https://doi.org/10.1038/ng.3805.
16. Pantel, K., and Alix-Panabières, C. (2019). Liquid biopsy and minimal residual disease - latest advances and implications for cure. Nat. Rev. Clin. Oncol. *16*, 409–424. https://doi.org/10.1038/s41571-019-0187-3.
17. Razavi, P., Li, B.T., Brown, D.N., Jung, B., Hubbell, E., Shen, R., Abida, W., Juluru, K., De Bruijn, I., Hou, C., et al. (2019). High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nat. Med. *25*, 1928–1937. https://doi.org/10.1038/s41591-019-0652-7.
18. Zhang, X., Yu, Z., Xu, Y., Chao, Y., Hu, Q., Li, C., Ye, M., Zhu, X., Cui, L., Bai, J., et al. (2022). Utility of cell-free DNA from bronchial washing fluid in diagnosis and genomic determination for radiology-suspected pulmonary nodules. Br. J. Cancer *127*, 2154–2165. https://doi.org/10.1038/s41416-022-01969-2.
19. Zhang, X., Li, C., Ye, M., Hu, Q., Hu, J., Gong, Z., Li, J., Zhao, X., Xu, Y., Zhang, D., et al. (2021). Bronchial Washing Fluid Versus Plasma and Bronchoscopy Biopsy Samples for Detecting Epidermal Growth Factor Receptor Mutation Status in Lung Cancer. Front. Oncol. *11*, 602402. https://doi.org/10.3389/fonc.2021.602402.
20. Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. (2023). A DNA methylation atlas of normal

human cell types. Nature *613*, 355–364. https://doi.org/10.1038/s41586-022-05580-6.

21. Rausch, C., Zhang, P., Casas-Delucchi, C.S., Daiß, J.L., Engel, C., Coster, G., Hastert, F.D., Weber, P., and Cardoso, M.C. (2021). Cytosine base modifications regulate DNA duplex stability and metabolism. Nucleic Acids Res. *49*, 12870–12894. https://doi.org/10.1093/nar/gkab509.
22. Liang, C., Liu, N., Zhang, Q., Deng, M., Ma, J., Lu, J., Yin, Y., Wang, J., Miao, Y., She, B., et al. (2022). A detection panel of novel methylated DNA markers for malignant pleural effusion. Front. Oncol. *12*, 967079. https://doi.org/10.3389/fonc.2022.967079.
23. Adusumalli, S., Mohd Omar, M.F., Soong, R., and Benoukraf, T. (2015). Methodological aspects of whole-genome bisulfite sequencing analysis. Brief. Bioinform. *16*, 369–379. https://doi.org/10.1093/bib/bbu016.
24. Ziller, M.J., Hansen, K.D., Meissner, A., and Aryee, M.J. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nat. Methods *12*, 230–232. 231 p following 232. https://doi.org/10.1038/nmeth.3152.
25. Jamshidi, A., Liu, M.C., Klein, E.A., Venn, O., Hubbell, E., Beausang, J.F., Gross, S., Melton, C., Fields, A.P., Liu, Q., et al. (2022). Evaluation of cell-free DNA approaches for multi-cancer early detection. Cancer Cell *40*, 1537–1549.e12. https://doi.org/10.1016/j.ccell.2022.10.022.
26. Müller, D., and Győrffy, B. (2022). DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer. Biochim. Biophys. Acta. Rev. Cancer *1877*, 188722. https://doi.org/10.1016/j.bbcan.2022.188722.
27. Han, Y.D., Oh, T.J., Chung, T.H., Jang, H.W., Kim, Y.N., An, S., and Kim, N.K. (2019). Early detection of colorectal cancer based on presence of methylated syndecan-2 (SDC2) in stool DNA. Clin. Epigenetics *11*, 51. https://doi.org/10.1186/s13148-019-0642-0.
28. Liang, N., Li, B., Jia, Z., Wang, C., Wu, P., Zheng, T., Wang, Y., Qiu, F., Wu, Y., Su, J., et al. (2021). Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. Nat. Biomed. Eng. *5*, 586–599. https://doi.org/10.1038/s41551-021-00746-5.
29. Crespo, I., Pignatelli, J., Kinare, V., Méndez-Gómez, H.R., Esgleas, M., Román, M.J., Canals, J.M., Tole, S., and Vicario, C. (2022). Tbr1 Misexpression Alters Neuronal Development in the Cerebral Cortex. Mol. Neurobiol. *59*, 5750–5765. https://doi.org/10.1007/s12035-022-02936-x.
30. Fazel Darbandi, S., Robinson Schwartz, S.E., Pai, E.L.L., Everitt, A., Turner, M.L., Cheyette, B.N.R., Willsey, A.J., State, M.W., Sohal, V.S., and Rubenstein, J.L.R. (2020). Enhancing WNT Signaling Restores Cortical Neuronal Spine Maturation and Synaptogenesis in Tbr1 Mutants. Cell Rep. *31*, 107495. https://doi.org/10.1016/j.celrep.2020.03.059.
31. Vrba, L., Oshiro, M.M., Kim, S.S., Garland, L.L., Placencia, C., Mahadevan, D., Nelson, M.A., and Futscher, B.W. (2020). DNA methylation biomarkers discovered in silico detect cancer in liquid biopsies from non-small cell lung cancer patients. Epigenetics *15*, 419–430. https://doi.org/10.1080/15592294.2019.1695333.
32. Seale, K., Horvath, S., Teschendorff, A., Eynon, N., and Voisin, S. (2022). Making sense of the ageing methylome. Nat. Rev. Genet. *23*, 585–605. https://doi.org/10.1038/s41576-022-00477-6.
33. Noroozi, R., Ghafouri-Fard, S., Pisarek, A., Rudnicka, J., Spólnicka, M., Branicki, W., Taheri, M., and Pośpiech, E. (2021). DNA methylation-based age clocks: From age prediction to age reversion. Ageing Res. Rev. *68*, 101314. https://doi.org/10.1016/j.arr.2021.101314.
34. Schneider, J.L., Rowe, J.H., Garcia-de-Alba, C., Kim, C.F., Sharpe, A.H., and Haigis, M.C. (2021). The aging lung: Physiology, disease, and immunity. Cell *184*, 1990–2019. https://doi.org/10.1016/j.cell.2021.03.005.
35. Monti, S., Grosso, V., Todoerti, M., and Caporali, R. (2018). Randomized controlled trials and real-world data: differences and similarities to untangle literature data. Rheumatology *57*, vii54–vii58. https://doi.org/10.1093/rheumatology/key109.
36. Jain, D., Allen, T.C., Aisner, D.L., Beasley, M.B., Cagle, P.T., Capelozzi, V.L., Hariri, L.P., Lantuejoul, S., Miller, R., Mino-Kenudson, M., et al. (2018). Rapid On-Site Evaluation of Endobronchial Ultrasound-Guided Transbronchial Needle Aspirations for the Diagnosis of Lung Cancer: A Perspective From Members of the Pulmonary Pathology Society. Arch. Pathol. Lab Med. *142*, 253–262. https://doi.org/10.5858/arpa.2017-0114-SA.
37. Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67. https://doi.org/10.1093/nar/gks1443.
38. Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta *2*, e107. https://doi.org/10.1002/imt2.107.
39. Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.
40. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
41. Robinson, J.T., Thorvaldsdottir, H., Turner, D., and Mesirov, J.P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics *39*, btac830. https://doi.org/10.1093/bioinformatics/btac830.
42. Steinfort, D.P., Leong, T.L., Laska, I.F., Beaty, A., Tsui, A., and Irving, L.B. (2015). Diagnostic utility and accuracy of rapid on-site evaluation of bronchoscopic brushings. Eur. Respir. J. *45*, 1653–1660. https://doi.org/10.1183/09031936.00111314.
43. Nadig, T.R., Thomas, N., Nietert, P.J., Lozier, J., Tanner, N.T., Wang Memoli, J.S., Pastis, N.J., and Silvestri, G.A. (2023). Guided Bronchoscopy for the Evaluation of Pulmonary Lesions: An Updated Meta-analysis. Chest *163*, 1589–1598. https://doi.org/10.1016/j.chest.2022.12.044.
44. Thiboutot, J., Pastis, N.J., Akulian, J., Silvestri, G.A., Chen, A., Wahidi, M.M., Gilbert, C.R., Lin, C.T., Los, J., Flenaugh, E., et al. (2023). A Multicenter, Single-Arm, Prospective Trial Assessing the Diagnostic Yield of Electromagnetic Bronchoscopic and Transthoracic Navigation for Peripheral Pulmonary Nodules. Am. J. Respir. Crit. Care Med. *208*, 837–845. https://doi.org/10.1164/rccm.202301-0099OC.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Bronchial washing fluid of 118 participants	Zhongshan hospital, Fudan University	NA
Bronchial washing fluid of 33 participants in the validation cohort	Zhongshan hospital, Fudan University	NA
Chemicals, peptides, and recombinant proteins		
Saline	Kelun Pharmaceutical	H51021158
Critical commercial assays		
Bronchial washing fluid storage tubes	Yunying Medicine	C102-20220108
Human DNAplus kit	Yunying Medicine	C105-20230020
MetPro DNA bisulfite conversion kit	Yunying Medicine	C21-20210169
DNA magnetic bead purification kit	Yunying Medicine	C17-20210171
Equalbit dsDNA HS Assay Kit	Vazyme	EQ111-02
AceTaq DNA Polymerase	Vazyme	P401-d2
MiniSeq Mid Output Kit (300-cycles)	Illumina	FC-420-1004
Deposited data		
Raw and analyzed data	This paper; Mendeley Data	https://doi.org/10.17632/wcnzyth6vd
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
GDC TCGA Lung Adenocarcinoma (LUAD), Illumina Human Methylation 450	NCI Genomic Data Commons	https://xenabrowser.net/datapages/
GDC TCGA Lung Squamous Cell Carcinoma (LUSC), Illumina Human Methylation 450	NCI Genomic Data Commons	https://xenabrowser.net/datapages/
Oligonucleotides		
Primers for 13 selected genes, see Table S3	This paper	N/A
Adapter primer: AATGATACGGCGACCACCGAGATCTACACT CTTTCCCTACACGACGCTCTTCCGATCT	This paper	N/A
Index primer: CAAGCAGAAGACGGCATACGAGATNNNNNN NNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	This paper	N/A
Software and algorithms		
Sequencing Analysis Viewer (V1.8)	Illumina	https://support.illumina.com/downloads/sequencing-analysis-viewer-software.html
bcl2fastq (V2.20.0.422)	Illumina	https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html
fastp (V0.20.1)	Chen et al.[38]	https://github.com/OpenGene/fastp
Burrows-Wheeler aligner (V0.7.17)	Li et al.[39]	https://github.com/lh3/bwa
samtools (V1.2)	Danecek et al.[40]	https://www.htslib.org/
igvtools (V2.3.98)	Robinson et al.[41]	https://igv.org/doc/desktop/
MetSeq module	Yunying Medicine	http://10.168.10.102:8080/complex/

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Scikit-learn (V1.3.2), sklearn.feature_selection.RFE	Python (V3.9)	https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
Scikit-learn (V1.3.2), sklearn.feature_selection.RFECV	Python (V3.9)	https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
Scikit-learn, V1.3.2, sklearn.linear_model.LassoCV	Python (V3.9)	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.htmlScikit-learn
Scikit-learn (V1.3.2), sklearn.svm.SVR	Python (V3.9)	https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
Scikit-learn (V1.3.2), sklearn.model_selection.GridSearchCV	Python (V3.9)	https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn-model-selection-gridsearchcv
XGBOOST (V 2.0.3)	Python (V3.9)	https://pypi.org/project/xgboost/
KERAS (V2.15)	Python (V3.9)	https://pypi.org/project/keras/
pROC	R (v4.2.3)	https://cran.r-project.org/web/packages/pROC/index.html
CBCgrps	R (v4.2.3)	https://CRAN.R-project.org/package=CBCgrps
Corrplot	R (v4.2.3)	https://cran.r-project.org/web/packages/corrplot/index.html
ggplot2	R (v4.2.3)	https://cran.r-project.org/web/packages/ggplot2/index.html
Other		
Count data related to the target methylation sequencing of the 118 patients and the validation cohort	This paper (EXdata 1, EXdata 3); Mendeley Data	https://doi.org/10.17632/wcnzyth6vd
Count data related to the calculated gene average methylation levels of the 118 patients and the validation cohort	This paper (EXdata 2, EXdata 4); Mendeley Data	https://doi.org/10.17632/wcnzyth6vd
Code utilized in constructing and validating the CanDo model	This paper; Mendeley Data	https://doi.org/10.17632/wcnzyth6vd

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Xin Zhang (xinhaier@sina.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Target methylation sequencing raw/count data has been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- All original code utilized in constructing and validating the CanDo model has been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- Human reference genome and Illumina Human Methylation 450 data were sourced from public database, accessible openly. The URL is provided in the key resources table.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study was conducted according to the principles of the World Medical Association Declaration of Helsinki. The study was approved specifically by the Internal Review Board of the Affiliated Zhongshan Hospital, Fudan University. All participants provided written informed consent for this study (IRB approval No. B2018-027R).

All participants included in this study were of Han Chinese ethnicity. Detailed information regarding cohort demographics such as age and gender can be found in Tables S4 and S7 of the manuscript. Potential impacts of age and gender on the outcomes are delineated in Figure 3D and Table S6.

METHOD DETAILS

Panel design

A comprehensive literature search was conducted employing the search terms "gene methylation," "lung cancer," "diagnosis," and "sensitivity and specificity" across various databases, including PubMed, Google Scholar, and CNKI. The search results were filtered based on publication dates, with an emphasis on articles published within the last decade (2013-2022). Out of a total of 184 records, we initially excluded review articles, meta-analyses, and duplicate publications. Subsequent exclusions were applied to studies concerning clinical prognosis, recurrence prediction, and those lacking explicit information regarding the specificity and sensitivity of methylation biomarkers in the final diagnostic model. Additionally, publications that combined methylation analysis with other biomarkers in multi-omics studies were omitted due to the challenge of independently assessing the diagnostic efficacy of methylation markers. Furthermore, studies relying solely on publicly available databases for analysis without validation were not considered.

Among the remaining eligible publications, a detailed analysis was conducted to assess the reported diagnostic sensitivity and specificity achieved by each methylation panel in the context of lung cancer. All mentioned genes were compiled, and those mentioned at least twice were selected to form the initial testing panel. Specific amplification primers targeting the CpG-rich promoter regions of these selected genes were carefully designed using Primer Premier 5 to perform downstream experimental strategies.

The biomedical database resource acquisition and management

The GDC TCGA public lung cancer database was accessed via the UCSC Xena Browser (https://xenabrowser.net/datapages/). Briefly, Illumina HumanMethylation450 (450K) raw data, phenotype information, and probeMap mapping data for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were downloaded. Subsequently, the data were organized and aligned using an in-house Python script, and TCGAutils was employed to convert UUIDs into Barcode files containing annotated phenotype information. LUAD and LUSC samples were then combined into “cancer” and “noncancer” groups based on phenotype. Probe sets aligning to the specified 13 genes were filtered using genomic information obtained from the probe mapping. The inter-group differences in methylation levels for each probe covering these 13 genes were computed. Concurrently, the average methylation level of all probes within each gene was calculated to denote the gene’s mean methylation level. The t-test was applied to assess the significance of inter-group differences. Additionally, the genomic coordinates of these gene loci were converted from the GRCh38 version to GRCh37 using the UCSC LiftOver tool. Following this conversion, these coordinates were cross-referenced with our designed primer target segments, and the shared CpG sites were annotated for subsequent analysis.

Inclusion criteria of participants

(1) Patients with peripheral pulmonary lesions necessitating bronchoscopy-guided biopsy for clinical assessment.
(2) Patients without contraindications to bronchoscopy examination or bronchoalveolar lavage.
(3) Patients with comprehensive clinical imaging data (such as CT scan images), baseline information (including age at diagnosis, gender, and smoking habits), and other pertinent details.
(4) Patients expressing a willingness to collaborate with the study, engage in potential follow-ups, and sign an informed consent form either personally or through a designated representative. For minors, the informed consent form should be signed by a legal guardian.

Exclusion criteria of participants

(1) Patients with an ambiguous final diagnosis for various reasons, including cases where the patient did not continue treatment at Zhongshan Hospital or demonstrated poor clinical adherence.
(2) Patients clinically diagnosed with contagious diseases or necessitating preventive control measures for infectious diseases (e.g., COVID-19, Pulmonary anthrax, Avian Influenza) prior to bronchoscopy examination.
(3) Patients lacking bronchoscopic rapid on-site evaluation diagnosis or histopathological staining of biopsy tissue will be excluded from the study.
(4) Inadequate collection of bronchial washing fluid specimens or samples of substandard quality.

Patient population and diagnosis

Between September 2022 to May 2023, volunteers with suspected lung nodules detected through CT scans and required bronchoscopy-guided biopsies were enrolled at Zhongshan Hospital, Fudan University (Shanghai, China). All enrolled participants strictly adhered to the predetermined inclusion and exclusion criteria, underwent bronchoscopy, and provided specimens from both bronchoscopic biopsy and lavage fluid in accordance with the established protocol. The clinical diagnostic strategy relied on a combination of clinical manifestations, digital chest X-ray findings, bronchoscopic-ROSE, and confirmation through histopathological biopsy. The histopathology-biopsy procedure followed the criteria outlined in the NCCN (National Comprehensive Cancer Network) or CSCO (Chinese Society of Clinical Oncology) diagnostic guidance for lung malignant tumors. Data regarding clinical characteristics of enrolled patients were carefully collected.

Moreover, an additional validation queue with rigorous control over baseline information was introduced in this study. All patients enrolled in this cohort reported no history of smoking, and meticulous measures were implemented to ensure uniformity in age and gender distribution across the groups. The BWF and paired biopsy samples were prospectively collected from each patient, and all examinations were consistent with the previously mentioned cohort.

Bronchoscopy, bronchial washing and biopsy procedure

The Fibrobronchoscopy (BF-1TQ260; Olympus, Tokyo, Japan) was used for a comprehensive airway examination and the collection of bronchial washing fluid (BWF). All bronchoscopies were performed under topical anesthesia and conscious sedation. During bronchoscopy examinations, two experienced physicians independently assessed the malignancy of the nodular phenotypes observed in the bronchoscopic field of view. Additionally, all patients underwent a standardized bronchoscopic rapid on-site evaluation (ROSE) process before biopsy for auxiliary diagnosis. For this, a cytology brush (Olympus, Japan) was used to collect cytological specimens from the target lesion, with each brushing consisting of 10-20 back-and-forth strokes, repeated 2-5 times. The collected material was then smeared onto three slides for ROSE examination.[36,42] The primary decision was based on the results of bronchoscopic-ROSE.

In cases of endobronchial visible lesions, BWF specimens were obtained from the subsegmental bronchus where the lesion was located, following saline irrigation of the lesion surface. Subsequently, endobronchial biopsy (EBB) was performed through the working channel of the wedged bronchoscope in the specified segmental bronchus. For endobronchial invisible peripheral lesions, transbronchial lung biopsy (TBLB) was performed with the assistance of endobronchial ultrasound. The Endoscopic Ultrasonography (EBUS) model included an EBUS probe (20MHz mechanical-radial type, UM-S20-20R or UM-S20-17S; Olympus, Japan) and a guide sheath (GS) kit (K-203 of Olympus, Japan). During EBUS, the ultrasound probe along with the GS was inserted through the working channel of the bronchoscope into the target bronchus. After sonographic confirmation of the biopsy site, saline was instilled through the GS channel using a connected syringe. In each examination, 20-40 mL of saline was used to irrigate the lesion before biopsy, flushed for 3-5 seconds, and then aspirated back, resulting in a minimum fluid recovery of 6 mL. Bronchoscopic forceps were then advanced through the working channel on the guide sheath, and multiple forceps biopsies were conducted with additional X-ray fluoroscopy assistance. All procedures were carried out by experienced practitioners following established protocols.[43,44]

Sample preprocessing and DNA extraction

Biopsy samples were sent to the pathology department for formalin-fixed paraffin-embedded processing, routine staining, and assessment by two experienced pathologists independently. To maintain the stability of DNA molecules during the stages of clinical collection and preprocessing, BWF specimens were promptly transferred to dedicated tubes (Yunying, Zhejiang, China) following sample collection. The tubes were then inverted 3-5 times to ensure comprehensive mixing of the preservative with the samples and were stored at 4°C until use.

The storage tubes were vortexed for 5 seconds before DNA extraction, and then 2 mL of BWF was aspirated after pipetting three times. Both the Human DNAplus kit (Yunying, Zhejiang, China) and the Auto-Pure20 system (Allsheng, Zhejiang, China) were employed for automated DNA extraction and purification. Subsequent to purification, sulfite conversion was conducted using the MetPro DNA bisulfite conversion kit (Yunying, Zhejiang, China). Each process strictly adhered to the manufacturer's protocol, and the products were harvested and stored at -20°C until use.

Library preparation and targeted methylation DNA sequencing

The bisulfite-treated DNA mentioned above was utilized in a multiplex PCR assay to generate barcoded sequencing libraries. The total reaction mixture was 30 μL, consisting of 19.5 μL mixed reaction buffer (Vazyme, Jiangsu, China), 5 μL template DNA, 5 μL of the pre-designed primer mix (synthesized by Sangon, Shanghai, China. 2 μM for each primer, and 0.5 μL AceTaq DNA polymerase (Vazyme, Jiangsu, China), underwent the first round of multiplex PCR in a CFX96 PCR machine (Eastwin, Beijing, China). The reaction conditions were set as follows: 95°C for 10 minutes, 35 cycles with an increment of 0.2°C per cycle (95°C for 30 seconds, 46-53°C for 30 seconds, 72°C for 30 s), followed by 72°C for 5 minutes, and a final hold at 4°C.

For the second round of multiplex PCR, the reaction mixture comprised 20.5 μL mixed reaction buffer (Vazyme, Jiangsu, China), 2 μL adapter primer (synthesized by Sangon, Shanghai, China), 2 μL index primer (synthesized by Sangon, Shanghai, China), 0.5 μL AceTaq DNA polymerase (Vazyme, Jiangsu, China), and 5 μL products obtained from the first round. The reaction conditions were set as follows: 95°C for 5 minutes, 20 cycles (95°C for 30 seconds, 55°C for 30 seconds, 72°C for 30 seconds), followed by 72°C for 5 minutes, and a final hold at 4°C.

The products from the second round of multiplex PCR were purified using the DNA magnetic bead purification kit (Yunying, Zhejiang, China). The concentration of barcoded libraries was determined using the Equalbit 1× dsDNA HS assay kit (Vazyme, Jiangsu, China). The 150-bp paired-end next-generation sequencing strategy was performed on the MiniSeq system (Illumina, California, United States) using the Miniseq Mid Output Reagent Cartridge (Illumina, California, United States), strictly following the manufacturer's protocols.

Sequencing data processing

The raw data BCL files were subjected to quality control using Sequencing Analysis Viewer V1.8 (Illumina, California, United States). Sequences with a Q30 base percentage exceeding 70% were considered as high-quality sequences. The bcl2fastq V2.20.0.422 (Illumina, California, United States) was employed to converted the qualified BCL files to FASTQ files. Followed by the fastp (V0.20.1)[38] removing the adapter sequences, low-quality base fragments, the Burrows-Wheeler aligner (BWA, V0.7.17)[39] was utilized to align the sequences from the quality-controlled FASTQ files to a custom methylated reference sequence (GRCh37). The aligned sequences were then sorted based on genomic coordinates using samtools (V1.2)[40] to generate BAM files and construct file indices. The igvtools (V2.3.98)[41] was applied for base depth analysis at each position in the resulting BAM files, and sites with sequencing depth below 1000× were filtered out. Finally, the internally developed MetSeq module (Yunying, Zhejiang, China) was employed to calculate the methylation levels at each CpG and non-CpG site. A comprehensive set of quality control metrics, including sample Q30 base percentage, sequencing depth at each position, and methylation rates, was aggregated and output as a tsv table file.

Methylation feature selection using RFE and RFECV

The methylation status of 151 CpG sites across 13 genes was assessed in 118 samples. Genes with over 10% of samples unable to detect methylation across all CpG sites in amplification fragments were excluded. The final dataset included 125 methylation features for each sample, comprising the methylation levels of 115 CpG sites and the average methylation levels of 10 genes (calculated as the mean methylation across all CpG sites covered by each gene fragment). The Recursive Feature Elimination (RFE) algorithm (Scikit-learn, V1.3.2, sklearn.feature_selection.RFE) was used for feature selection, removing the weakest features. This decision was validated through 10-fold cross-validation, with favorable outcomes observed when the feature number ranged from 5 to 17. The final number of selected features, denoted as "n_feature_to_select," was determined by the Recursive Feature Elimination Cross-Validation (RFECV) algorithm (Scikit-learn, V1.3.2, sklearn.feature_selection.RFECV), with the "step" option set to 1, the "cv" option set to "StratifiedKFold(2)," and the parameter "scoring" set to "accuracy." Ultimately, 11 features were identified as optimal and deemed most relevant for constructing the diagnostic model.

Machine learning classifier evaluation

The LASSO model extends the classic linear regression model by incorporating an additional regularization term in the loss function to constrain the weights and mitigate overfitting. The alpha value of the LASSO model is determined using the LassoCV algorithm with 5-fold cross-validation (Scikit-learn, V1.3.2, sklearn.linear_model.LassoCV). The SVM model is employed for group classification using a hyperplane and is often utilized as a regression model due to its high flexibility. The parameters (C, gamma, epsilon) of the SVM model are automatically selected using the GridSearchCV algorithm (Scikit-learn, V1.3.2, sklearn.model_selection.GridSearchCV) with 5-fold cross-validation. XGBOOST algorithm (XGBOOST, V 2.0.3) is based on the decision tree model but offers a parallel tree boosting method, enabling rapid and stable resolution of regression problems. The optimal settings for XGBOOST, including maximum depth, minimum sub-depth, gamma, subsample, colsample bytree, and reg alpha, are determined using the GridSearchCV (Scikit-learn, V1.3.2, sklearn.model_selection.GridSearchCV) algorithm with 5-fold cross-validation. Other parameters (eta, nthread, etc.) are set to their default values as specified in the original documentation. Receiver Operating Characteristic (ROC) curves were employed to evaluate the classification performance of these machine learning algorithms, and the F1 score was calculated using the "f1_score" function imported from the "sklearn.metrics" module. In brief, the F1 score is calculated as 2*(precision*recall)/(precision+recall), where precision = TP/(TP+FP) and recall = TP/(TP+FN). In the formula, TP represents true positive, FP represents false positive, and FN represents false negative.

CanDo model development using DNN

The CanDo model was developed using a DNN algorithm (KERAS, V2.15) with three fully connected hidden layers. The model takes 11 methylation percentage values of each sample as inputs, with the output ranging from 0 to 1, where a higher value indicates a higher likelihood of the patient having cancer. The first hidden layer consists of 50 nodes, the second has 10 nodes, and the third has 5 nodes. Dropout layers with a dropout rate of 0.2 were added to all hidden layers, and batch normalization methods were applied. Rectified Linear Units (ReLU) were used as the activation function for the hidden layers, while the output layer utilized the sigmoid function. Additionally, L2 regularizers (lambda = 0.001) were applied to adjust the weights smoothly.

During the training validation phase, the network underwent 500 epochs of training, utilizing Adam as the optimizer with a learning rate of 0.0002, a batch size of 8, a validation split of 0.2, and Mean Squared Error (MSE) as the loss function. Model performance was evaluated based on both accuracy and loss for each epoch.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis and data visualization were conducted using R (v4.2.3). One-way analysis of variance (ANOVA) was employed to compare differences in methylation detection among different groups. Significance analysis of intergroup differences was conducted based on the normality of data distribution, using either Student's t-test or Wilcoxon rank-sum test. Statistical significance was assessed using a two-tailed *p* value < 0.05. The 'pROC' package was used to calculate ROC curves for the methylation levels of different genes, and the 'CBCgrps' package was employed for correlation analysis between various factors. The "corrplot" package was used to generate a correlation heatmap. Various plots, including scatter plots, boxplots, and sequencing heatmaps, were created using the "ggplot2" package.