

ORIGINAL ARTICLE

Development of a genome atlas for discriminating benign, preinvasive, and invasive lung nodules

Peng Liang^{1,#} | Minhua Peng^{2,#} | Jinsheng Tao^{2,#} | Bo Wang^{2,#} |
 Jinwang Wei^{3,4,#} | Lixuan Lin¹ | Bo Cheng¹ | Shan Xiong¹ | Jianfu Li¹ |
 Caichen Li¹ | Ziwen Yu¹ | Chunyan Li¹ | Jun Wang² | Hui Li² |
 Zhiwei Chen^{2,5} | Jian-Bing Fan^{2,6,*} | Wenhua Liang^{1,*}  | Jianxing He^{1,*} 

¹Department of Thoracic Surgery and Oncology, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou Institute of Respiratory Health, State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou, Guangdong, China

²AnchorDx Medical Co., Ltd, Guangzhou, Guangdong, China

³Department of Data Science, Genomicare Biotechnology (Shanghai) Co., Ltd., Shanghai, China

⁴Department of Data Science, Shanghai CreateCured Biotechnology Co., Ltd., Shanghai, China

⁵AnchorDx Inc., Fremont, California, USA

⁶Department of Pathology, Southern Medical University, Guangzhou, Guangdong, China

*Correspondence

Jianxing He, MD, and Wenhua Liang, MD, Department of Thoracic Surgery and Oncology, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou Institute of Respiratory Health, State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou, Guangdong, China.
 Email: drjianxing.he@gmail.com; liangwh1987@163.com

Jian-Bing Fan, AnchorDx Medical Co., Ltd, Guangzhou, Guangdong, China.
 Email: jianbingfan1115@smu.edu.cn

Abstract

To tackle misdiagnosis in lung cancer screening with low-dose computed tomography (LDCT), we aimed to compile a genome atlas for differentiating benign, preinvasive, and invasive lung nodules and characterize their molecular pathogenesis. We collected 432 lung nodule tissue samples from Chinese patients, spanning benign, atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IA). We performed comprehensive sequencing, examining somatic variants, gene expressions, and methylation levels. **Our findings uncovered EGFR and TP53 mutations as key drivers in - early lung cancer development, with EGFR mutation frequency increasing with disease progression.** Both

[#]Peng Liang, Minhua Peng, Jinsheng Tao, Bo Wang, Jinwang Wei, Jian-Bing Fan, Wenhua Liang and Jianxing He are the co-first authors and contributed equally to this study.

Funding information: China National Science Foundation, Grant/Award Numbers: 82022048, 81871893; Research Project of Guangzhou Lab, Grant/Award Numbers: SRPG22-017, GZNL2023A02007; Scheme of Guangzhou Economic and Technological Development District for Leading Talents in Innovation and Entrepreneurship, Grant/Award Number: 2017-L152; Scheme of Guangzhou for Leading Talents in Innovation and Entrepreneurship, Grant/Award Number: 2016007; Scheme of Guangzhou for Leading Team in Innovation, Grant/Award Number: 201909010010; 2020 Guangzhou Development Zone International Science and Technology Cooperation Project, Grant/Award Number: 2020GH15; Science and Technology Planning Project of Guangzhou, Grant/Award Number: 202206080013; Guangzhou Development Zone International Science and Technology Cooperation Project, Grant/Award Number: 2021GH17; Key Project of Guangzhou Scientific Research Project, Grant/Award Number: 201804020030; National Key Research & Development Programme, Grant/Award Numbers: 2022YFC2505100, 2017YFC0112704; High-Level University Construction Project of Guangzhou Medical University, Grant/Award Numbers: 20182737, 201721007, 201715907, 2017160107

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). MedComm published by Sichuan International Medical Exchange & Promotion Association (SCIMEA) and John Wiley & Sons Australia, Ltd.

EGFR mutations and EGF/EGFR hypo-methylation activated the EGFR pathway, fueling cancer growth. Transcriptome analysis identified four lung nodule subtypes (G1-4) with distinct molecular features and immune cell infiltrations: EGFR-driven G1, EGFR/TP53 co-mutation G2, inflamed G3, stem-like G4. Estrogen/androgen response was associated with the EGFR pathway, proposing a new therapy combining tyrosine kinase inhibitors with antiestrogens. Preinvasive nodules exhibited stem cell pathway enrichment, potentially hindering invasion. Epigenetic regulation of various genes was essential for lung cancer initiation and development. This study provides insights into the molecular mechanism of neoplastic progression and identifies potential diagnostic biomarkers and therapeutic targets for lung cancer.

KEYWORDS

EGFR/TP53 mutations, epigenetic regulation, genome atlas, lung adenocarcinoma, molecular pathogenesis, therapeutic targets

1 | INTRODUCTION

Lung cancer is the most common malignancy and the primary cause of cancer-related death worldwide.¹ The risk factors for lung cancer include tobacco smoking, exposure to environmental chemicals, genetic predisposition, etc.^{2,3} Early diagnosis is associated with a favorable prognosis in lung cancer, as evidenced by the significantly increased 5-year overall survival rate for stage I (68% to 92%) compared to stage IV (<10%).⁴

Currently, LDCT is extensively utilized for detecting small lung nodules that might be overlooked by conventional X-ray.⁵ Lung nodules less than 30 mm in size pose a significant challenge in assessing their malignant potential, often leading to unnecessary diagnostic work-ups.⁶ Adenocarcinoma is the predominant type of lung cancer. The most prevalent subset, lung adenocarcinoma (LUAD), is believed to progress stepwise from preinvasive lesions of atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), to invasive adenocarcinoma (IA) with a lepidic growth pattern.^{7,8} Although early screening of lung cancers through LDCT reduces mortality, the resulting high false-positive rate and misdiagnosis emphasize the need for a deeper comprehension of the molecular mechanisms underlying benign, preinvasive, and invasive lung nodules. This understanding is crucial for discovering diagnostic biomarkers that can accurately identify early-stage lung cancer.

Accumulated evidence from next-generation sequencing studies has shed light on the distinct genetic characteristics of various subtypes of lung adenocarcinoma nodules.^{9–12} Hu et al. have shown the multiregion exome

sequencing of 116 AAH, AIS, MIA, and ADC (lung adenocarcinoma) samples, observing genomics evolution at both the single nucleotide level and chromosomal levels.^{7,9} They also noted a progressive increase in methylation aberrations from AAH to AIS, MIA, and ADC, suggesting that global hypomethylation might influence chromosomal instability, mutagenesis and tumor immune microenvironment during precancer progression. Chen et al. have reported on the exome and transcriptome sequencing of preinvasive and invasive lung adenocarcinoma, identifying significantly mutated genes such as *EGFR* (Epidermal Growth Factor Receptor), *RBM10* (RNA binding motif protein 10), *BRAF* (B-Raf proto-oncogene), *ERBB2* (erb-b2 receptor tyrosine kinase 2), *TP53* (tumor protein p53), *KRAS* (KRAS proto-oncogene, GTPase), *MAP2K1* (mitogen-activated protein kinase kinase 1), and *MET* (mesenchymal to epithelial transition factor) in the pre/minimally invasive group.¹¹ Dejima et al. have characterized the immune contexture of invasive ADC and its precursors using transcriptomic immune profiling, T cell receptor sequencing, and multiplex immunofluorescence. Their findings demonstrated that anti-tumor immunity evolved as a continuum with a gradually less effective and more intensively regulated immune response from AAH to AIS, MIA, and ADC.¹² However, to fully understand the underlying biology and altered pathways involved in the malignancy development of lung nodules, comprehensive analyses are required. The use of systematic multiomics integrative approaches could significantly enhance our understanding of lung cancer stratification and facilitate the discovery of biomarkers for anti-cancer drugs.^{13,14}

In this study, we aimed to distinguish between benign, preinvasive (AAH, AIS), and invasive (MIA, IA) nodules.

To achieve this, we conduct a comprehensive analysis of data from whole exome sequencing (WES), transcriptomics, and DNA methylation sequencing of lung nodules samples from Chinese patients who had undergone surgical resection. Our findings offer a detailed genetic and epigenetic profile of benign nodules and early-stage lung cancer, along with novel insights into the molecular mechanisms involved in the development of lung cancer.

2 | RESULTS

2.1 | Patient samples and characteristics

We collected 432 archived Formalin-fixed paraffin-embedded (FFPE) tissue samples from lung nodules. Samples with insufficient Genomic DNA (gDNA) amounts, low gDNA quality, low library yields, or failed sequencing quality control metrics were excluded. After thorough analysis, we obtained a final set of 422 samples, which were classified into three groups: benign ($n = 124$), preinvasive (AAH+AIS, $n = 32$), and invasive (MIA+IA, $n = 266$). The average age of the benign group was 50 years, younger than the preinvasive group (55 years) and the invasive group (57 years) ($p < 0.05$). In terms of gender, 59.4% of the preinvasive and invasive groups were female, while 62.1% of the benign group were male. Specifically, 78.2%, 90.6%, and 81.2% of the benign, preinvasive, and invasive groups were nonsmokers, respectively. Regarding nodule density, 51.9% of all nodules were solitary nodules (SNs), 37.9% were pure ground glass nodules (pGGNs), and 9.2% were part-solid nodules (PSNs). Remarkably, 62.5% of preinvasive and 49.2% of invasive nodules were pGGNs, while only 7.3% of benign nodules were pGGNs. Additionally, 87.9% of benign patients had SNs, but this percentage decreased to 41.0% among the invasive groups. Among the benign nodules, 29.8% were inflammation, while 17.7% were tuberculosis. It is worth noting that 97.7% of invasive nodules were very early lung cancers, classified as American Joint Committee on Cancer (AJCC) stage I (Table 1).

2.2 | Somatic mutation landscape of lung nodules

In the WES study, we analyzed a total of 422 lung nodule tissues (124 benign, 32 AAH and AIS, 266 MIA and IA) and their matched white blood cells (WBCs). Twenty-two significantly mutated genes were identified in the invasive samples (Figure 1A), among which the two most frequently mutated genes were *EGFR* (61%) and *TP53* (24%), consistent with previous reports.¹¹

The frequency of *EGFR* mutations increased from benign (3%), to preinvasive (AAH+AIS, 16%), MIA (35%), and IA (73%) (Figure 1A). The *EGFR* mutation frequencies were statistically higher in females (vs. male, $p = 0.0016$) and nonsmokers (vs. smokers, $p = 0.0004$) (Figures 1A and S1). The most prevalent types of *EGFR* mutations identified were exon 21 p.L858R (39%) and exon19 deletions (36%) (Figure S2A), and these mutations were highly prevalent in MIA (81.25%, 50% exon 21 p. L858R, 31.25% exon 19 deletions) and IA (78.91%, 39.46% exon 21 p. L858R, 39.46% exon 19 deletions) (Figure S2B). These hotspots *EGFR* mutations observed in our early-stage lung cancer cohort were consistent with those found in late-stage lung cancers in another Chinese lung cancer cohort.^{15,16} Other *EGFR* mutations, such as exon 20 insertions, exon 18 p.G719S/A, exon 18 p.E709A, exon 21 p.L861Q/R, and exon 20 p.T790M, were also detected (Figure S2A and B). Notably, *EGFR* mutations in benign nodules were infrequent and inactive, as demonstrated by functional annotation of mutated genes (Figure S6).

TP53 mutations were significantly more prevalent in IA (35.43%) compared to MIA (2.2%) (Figure 1A) and were scarcely detected in preinvasive and benign samples. This suggests that *TP53* mutations might not influence nodule growth but could play an important role in promoting invasiveness development, which is in line with a recent finding.¹⁷ Patients harboring *TP53* mutations were more likely to be smokers, strongly associated with older age (≥ 55 years), male gender, solid nodules, and higher TMB (≥ 1.39 /Mb) (Figure S1).

The mean TMB of this cohort was 3.34/Mb, which was lower than the studies of late-stage lung cancers.¹⁸ The average TMB in invasive samples (MIA, IA) was 2.94/Mb. When comparing within nodule types, the TMB of solid nodules (S) was significantly higher than that of pure GGN (G) ($p < 0.001$) (Figure 1B). Furthermore, when TMB was compared within different pathological subtypes, we observed that it was higher in invasive stage compared to benign and preinvasive stage (Figure 1C). Interestingly, within the IA groups, the TMBs of pure GGN, part-solid, and solid nodules gradually increased as the nodule density increased (Figure 1C). When intersecting pathological subtypes with *EGFR* mutation status, the median TMB in the *EGFR* mutation (MUT) group was significantly lower compared to the *EGFR* wild type (WT) group in IA ($p < 0.01$), consistent with a previous study.¹⁹ However, this trend was reversed among benign nodules, possibly due to the very low proportion of the *EGFR* MUT group in benign group (Figure 1D). We further compared *EGFR* mutation status among nodule types based on TMB and found that the median TMB in the *EGFR* mutation group was statistically higher than that in the *EGFR* WT group with pure GGNs (G) and solid nodules (S) (Figure 1E)

TABLE 1 Clinical and demographic characteristics of patients.

Characteristics	Benign (<i>n</i> = 124)	Preinvasive (AAH+AIS) (<i>n</i> = 32)	Invasive (MIA+IA) (<i>n</i> = 266)	Total (<i>n</i> = 422)
Age (mean ± SD)	50.77 ± 13.41	55.31 ± 9.46	56.8 ± 11.44	54.92 ± 12.2
Gender				
Female	47 (37.9%)	19 (59.4%)	158 (59.4%)	
Male	77 (62.1%)	13 (40.6%)	108 (40.6%)	
Smoking history				
Nonsmokers	97 (78.2%)	29 (90.6%)	216 (81.2%)	342 (81.0%)
Smokers	27 (21.8%)	3 (9.4%)	50 (18.8%)	80 (19.0%)
Nodule type				
pGGN (G)	9 (7.3%)	20 (62.5%)	131 (49.2%)	160 (37.9%)
PSN (P)	6 (4.8%)	7 (21.9%)	26 (9.8%)	39 (9.2%)
SN (S)	109 (87.9%)	1 (3.1%)	109 (41.0%)	219 (51.9%)
NA		4 (12.5%)		4 (0.9%)
Pathological Subtype				
INF	37 (29.8%)			
TB	22 (17.7%)			
FUN	18 (14.5%)			
HAM	13 (10.5%)			
GRAN	8 (6.5%)			
FIBROSIS	7 (5.6%)			
PSP	7 (5.6%)			
other	12 (9.6%)			
IA			175 (65.8%)	
AIS		26 (81.3%)		
MIA			91 (34.2%)	
AAH		6 (18.8%)		
AJCC stage				
0		26 (81.3%)		
I			260 (97.7%)	
II			1 (0.4%)	
III			1 (0.4%)	
IV			3 (1.1%)	
NA		6 (18.8%)	1 (0.4%)	

Abbreviations: AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; FUN, fungus infection; GRAN, granulomas; HAM, hamartoma; IA, invasive adenocarcinomas; INF, inflammation; MIA, minimally invasive adenocarcinoma; NA, not applicable; pGGN, pure ground-glass nodule; PSN, part solid nodule; PSP, pulmonary sclerosing pneumocytoma; SN, solid nodule; TB, tuberculosis.

($p < 0.001$). Considering the impact of *TP53* mutation on TMB in IA nodules, a group of *TP53* MUT & *EGFR* WT samples appeared to have the highest TMB compared with other combinations of *TP53* WT/MUT and *EGFR* WT/MUT (Figure S3), which is in agreement with other studies^{16,20} showing that TMB was higher in tumors with *TP53* MUT or *EGFR* WT.

To assess the functional significance of mutated splicing genes, nodule samples were divided into splicing gene MUT and WT groups, and alternative splicing events from RNA-Seq data were analyzed (Figure 2A). In inva-

sive samples, a significantly higher number of splicing events were detected in samples carrying splicing gene mutation (Figure 2B). A total of 417 genes were identified with splicing events (splicing events ≥ 10). Pathway enrichment analysis revealed that these genes with splicing events were primarily involved in the cell cycle and other cancer-related pathways, such as *KRAS* signaling and heme metabolism (Figure S4A). *RBM10* (12%) and *RBMX* (RNA Binding Motif Protein X-Linked) (7%) were the most common splicing mutation genes. However, no significant difference in alternative splicing events was observed

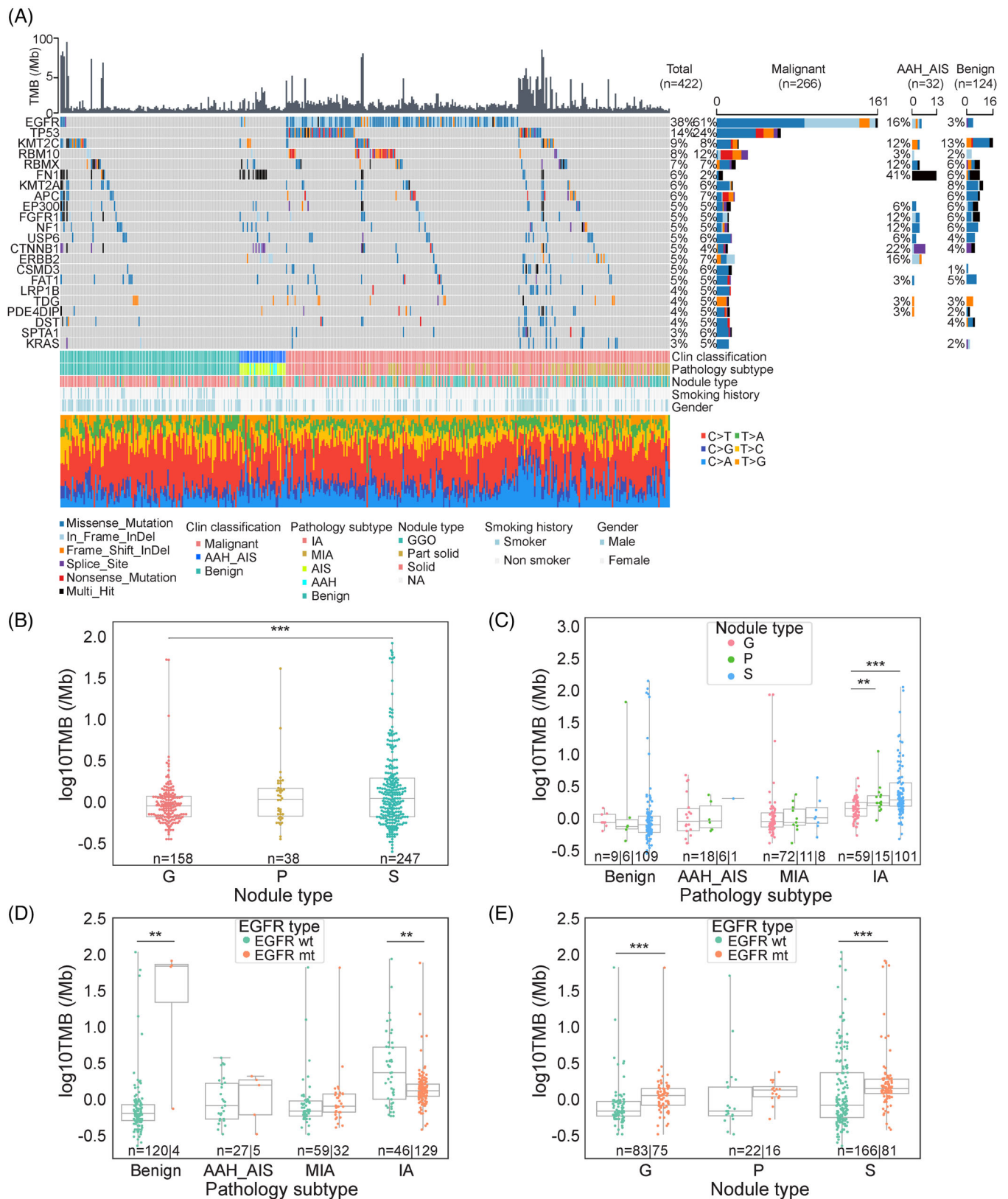


FIGURE 1 Landscape of somatic mutation in lung nodules. (A) Oncoprint of sequenced nodule samples including benign, AIS, AAH, MIA, and IA. Mutation type breakdowns are shown on the right. TMB is shown on the top. (B) Comparison of nodule types according to TMB. G: pure GGN; P: part-solid nodule; S: solid nodule. (C) Comparison of pathological subtypes among nodule types, stratified by TMB. (D) Comparison of *EGFR* mutation status in different pathological subtypes, stratified by TMB. (E) Comparison of *EGFR* mutation status in different nodule types, stratified by TMB. In the boxplots, the upper and lower lines represent the first and third quartiles, and the center lines represent the median. ns, not significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$, Two-sided Mann-Whitney U test.

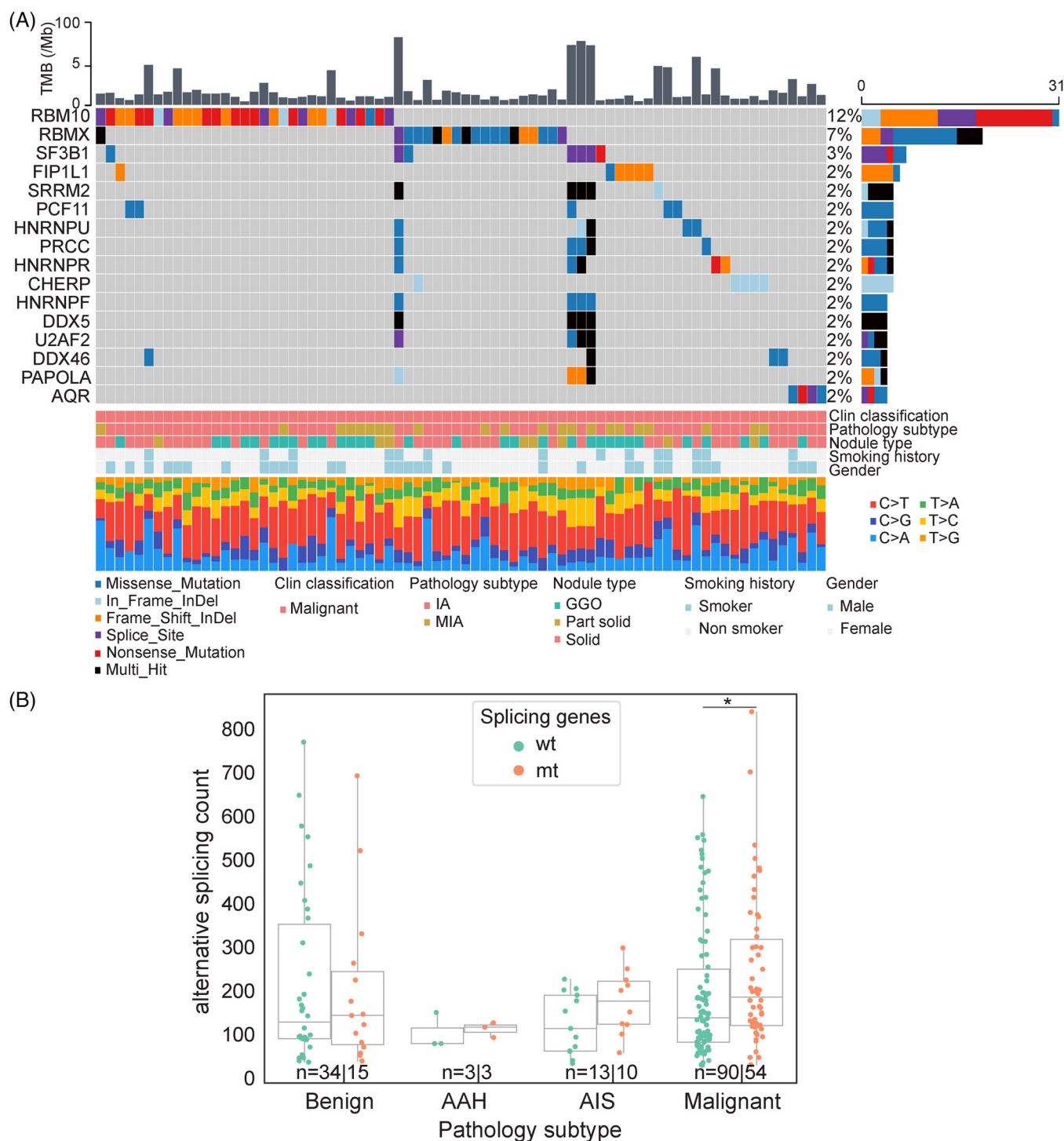


FIGURE 2 Landscape of Splicing Mutations in lung nodules. (A) Oncoplot of frequently splicing mutated genes in malignant nodules. The barplot above represents the TMB. The barplot on the right represents the type of mutations detected for each gene. (B) Comparison of alternative splicing events in different pathological subtypes of nodules with wild type or mutation in splicing genes. In the boxplots, the upper and lower lines represent the first and third quartiles, and the center lines represent the median. ns, not significant, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$, Two-sided Mann-Whitney U test.

across pathological subtypes or between MUT and WT groups (excluding any splicing mutation) (Figure S4B and C).

Furthermore, we analyzed the detected CNVs by gene annotation based on the significance of differential genes between malignant and benign nodules. However, no lung cancer-related gene was identified (data not shown).

2.3 | The transcriptomic features and immune microenvironment of lung nodules

To elucidate the transcriptomic landscape of lung nodules, RNA-Seq was conducted on 226 lung nodules samples, comprising 52 benign, 29 AAH+AIS, and 145 MIA+IA cases. Through the integration of WES and RNA-Seq data, four distinct molecular subgroups were identified based on unsupervised hierarchical clustering analysis of the gene expression profiles (Figure 3A).

Group 1, classified as the *EGFR*-driven group, was predominantly composed of invasive nodules harboring *EGFR* mutations. Markedly increased expression of early and late estrogen response (ER) and androgen response (AR) genes was observed. Estrogen and its receptors have been postulated to play a role in the development and progression of lung cancer.²¹ Group 2, referred to as the *EGFR/TP53* co-mutation group, was enriched for IA nodules carrying *EGFR/TP53* co-mutations. These nodules exhibited upregulation of genes associated with mitotic spindles, the G2/M checkpoint, and E2F targets. Most of these genes are linked to cell cycle regulation and tumor proliferation. Notably, the glycolysis pathway was also upregulated in Group 2, aligning with the Warburg effect in tumor metabolism and supporting the idea that tumors generate energy through glycolysis to fuel proliferation and metastasis via metabolic reprogramming.²² Group 3, designated as the “inflamed group,” was enriched with benign nodules and included some MIA and IA samples. Increased expression of numerous inflammation-related genes, such as *IL-2* (Interleukin-2), *IL-6* (Interleukin-6), and *IFN* (Interferons), was observed. Group 4 represented the stem-like group, which comprised the majority of preinvasive (AAH, AIS) nodules and some MIA nodules. This group exhibited upregulation of stem cell pathways, including NOTCH, Wnt/ β -catenin, and hedgehog signaling, implicating the important role of stem cell pathways in tumor initiation (Figure 3A).

To explore the immune infiltration in lung nodules, we computed the immunophenotype score and immune cell infiltration score using single-sample Gene Set Enrichment Analysis based on the RNA-seq data, similar to a previous study.²³ Notably, immune cell infiltration seemed

more pronounced in the inflamed (Group 3) and stem-like (Group 4) groups, where benign and preinvasive nodules tend to cluster. In contrast, *EGFR*-driven (Group 1) and *EGFR/TP53* co-mutation (Group 2) groups appeared relatively “cold” with fewer immune infiltrates (Figure 3A and B). The *EGFR*-driven group (Group 1) exhibits a lack of T cell infiltration, especially CD8 T cell infiltration, which is consistent with the clinical observation that lung cancer patients with *EGFR* mutation tend to have an unfavorable response to PD-1/PD-L1 checkpoint inhibitors.²⁴ In the *EGFR/TP53* co-mutation group (Group 2), CD8 T cell infiltration was also low. Conversely, Group 2 showed higher levels of Th2 and Treg cell infiltration. A recent study revealed that the loss of *TP53* function in tumor cells results in the accumulation of Treg cells.²⁵ Due to the inflammatory characteristics of benign nodules, immune cell infiltration in the inflamed group was significant, which aligns with the observation that this group demonstrated activation of numerous inflammatory pathways, such as *IL-2/STAT5* and *IL-6/JAK/STAT3*. Interestingly, the stem-like group (Group 4) exhibited very high natural killer (NK) cell infiltration (Figure 3B). We also identified a strong correlation between the Wnt/ β -catenin pathway (Pearson’s correlation coefficient (PCC) = 0.8) and a moderate correlation of the NOTCH pathway (PCC = 0.48) in relation to NK cells (Figure 4A and C). NK cells are a crucial component of the innate immune system, contributing to immunosurveillance.²⁶ The Wnt/ β -catenin signaling pathway is involved in NK cell development and activates the development and function of natural killer T (NKT) cells.²⁷ We postulated that NK cells infiltrating preinvasive nodules were part of the host’s innate immunosurveillance mechanism to prevent invasion. To evaluate whether this association is specific to preinvasive nodules, we examined it using TCGA data. The results indicated that the correlation was moderate for both the Wnt/ β -catenin (PCC = 0.37) and NOTCH (PCC = 0.56) signaling pathways (Figure 4B and D).

2.4 | The DNA methylation characterization of lung nodules

To investigate the DNA methylation patterns in lung cancer, 83 DNA samples (comprising 18 benign, 19 MIA, 46 IA) were subjected to targeted DNA methylation sequencing. The data from a custom panel, containing 12,899 preselected probes, were clustered into four groups (M1, M2, M3, and M4) using unsupervised hierarchical clustering. Consequently, the M2 and M4 groups exhibited distinct methylation patterns across pathology subtypes. In the M2 group, markers were hypomethylated in benign and MIA nodules, while they were hypermethylated in IA nodules.

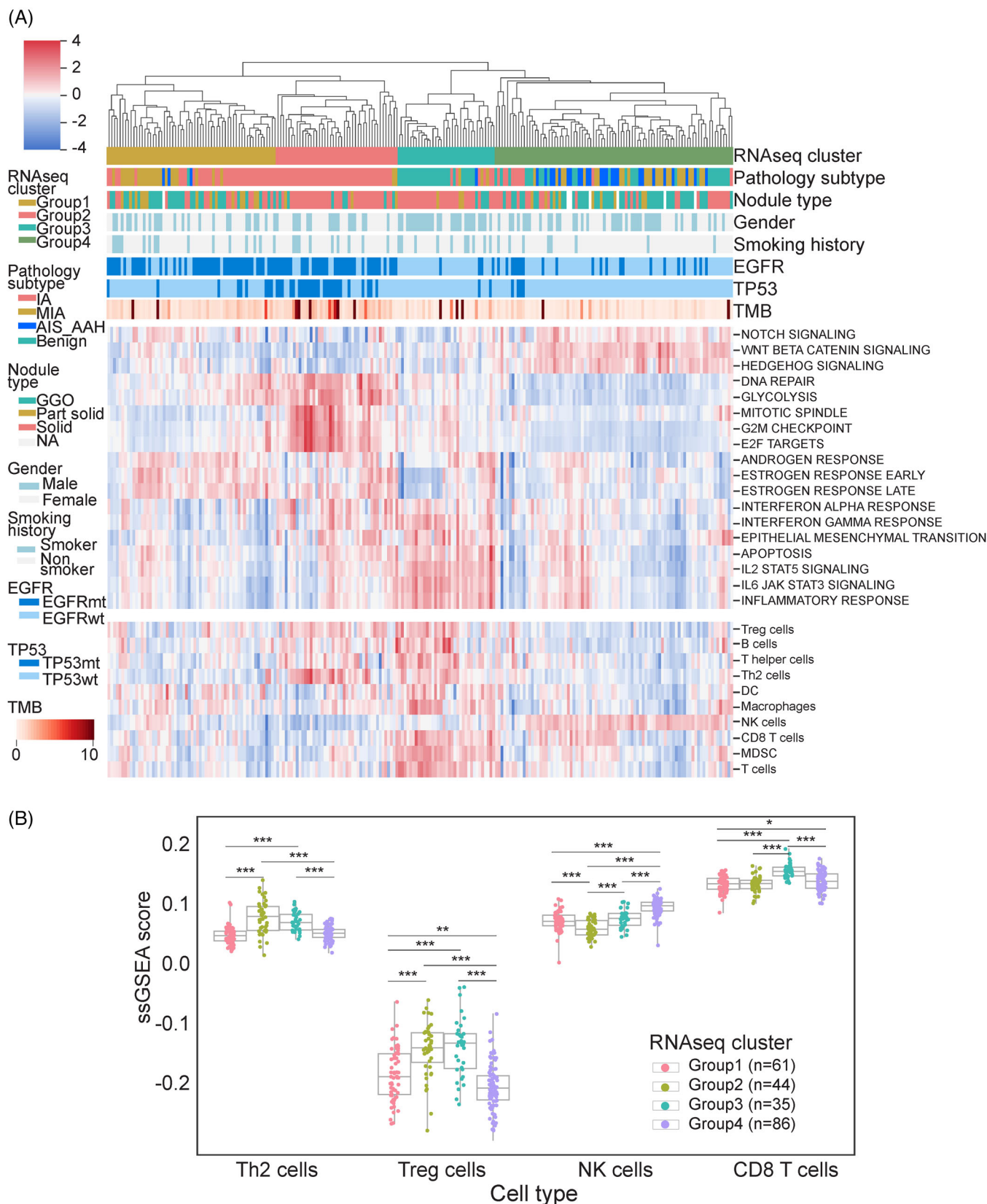


FIGURE 3 The transcriptomic features and immune microenvironment of lung nodules. (A) Phenotypes of RNA cluster subgroups. The cluster assignments of patients are shown on the top. The immune signature is on the bottom. (B) Immune cell comparison between RNA cluster subgroups.

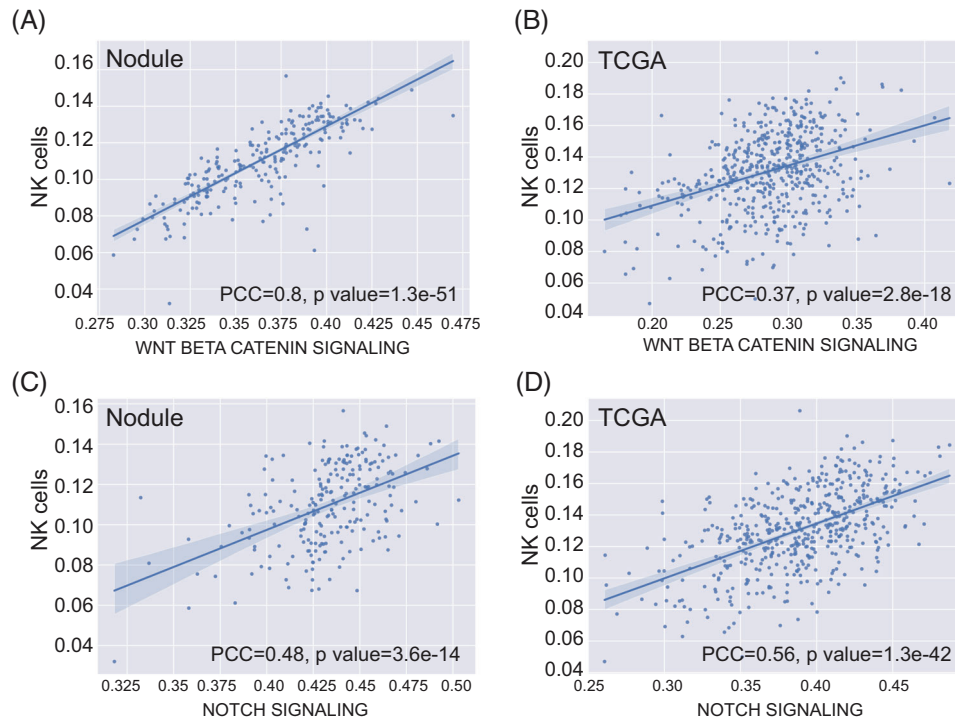


FIGURE 4 The correlation between NK cells and stem cell signaling pathway. (A–B) The correlation between NK cells and WNT Beta Catenin signaling in lung nodules (A) or based on TCGA data (B). (C–D) The correlation between NK cells and Notch signaling in lung nodules (C) or based on TCGA data (D). ns, not significant, * $p < 0.05$, ** $p < 0.01$; *** $p < 0.001$, Two-sided Mann–Whitney U test.

In contrast, markers in the M4 group were hypermethylated in benign and MIA nodules but hypomethylated in IA nodules (Figure 5A).

To investigate gene functions, a hallmark gene set was utilized to conduct pathway enrichment analysis (Figure 5B). Genes within the M2 group were enriched in myogenesis, epithelial mesenchymal transition (EMT), IL2/STAT5 signaling, and the upregulation of KRAS signaling. EMT plays a role in tumor invasion and metastasis.²⁸ IL2/STAT5 signaling regulates Treg cell differentiation and maturation.²⁹ KRAS signaling was related to cell proliferation. Genes in the M4 group were enriched in ER and AR pathways, TGF β signaling, and the downregulation of KRAS signaling. TGF β signaling modulates tumor invasion and metastasis through EMT, angiogenesis, and immunosuppression.³⁰ The enrichment of ER and AR pathways in M4 was consistent with the transcriptomics analysis of *EGFR*-driven group (Group 1) in this study, further supporting the involvement of hormonal pathways in lung cancer development (Figure 3A).

2.5 | The integration of gene expression and DNA methylation

There is evidence of both strong positive and strong negative correlations between gene methylation and gene

expression.³¹ To pinpoint potentially relevant DNA methylation and gene expression alterations during the progression of lung nodules, an integrated analysis of DNA methylation and gene expression was conducted. Pearson's correlations were calculated to identify statistically significant inverse correlations between DNA methylation and gene expression of enriched genes from pathway analysis, yielding fourteen genes with DNA methylation alterations situated at the promoter regions. **In the M2 group, five genes, including *BDNF* (brain-derived neurotrophic factor), *CD48* (CD48 molecule), *COX7A1* (cytochrome c oxidase subunit 7A1), *IKZF1* (IKAROS family zinc finger 1), and *TNFRSF1B* (TNF receptor superfamily member 1B), displayed progressively increasing methylation levels, accompanied by corresponding decreases in gene expression levels from benign, MIA, to IA (Figure 6).** The highest inverse correlation values in the M2 group were observed for *CD48* (−0.7), *COX7A1* (−0.65), and *TNFRSF1B* (−0.61). Our findings regarding *COX7A1* were consistent with recent studies showing that *COX7A1* had higher methylation and lower expression in the NSCLC cell line,³² which suppressed cell viability by downregulating tumor autophagy in vitro.³³ Previous studies have demonstrated that *TNFRSF1B*, a gene associated with apoptosis, had mutations and lower expression in NSCLC patients,^{34,35} which is supported by our findings. However, it has been reported that *BDNF* had a high expression level in lung

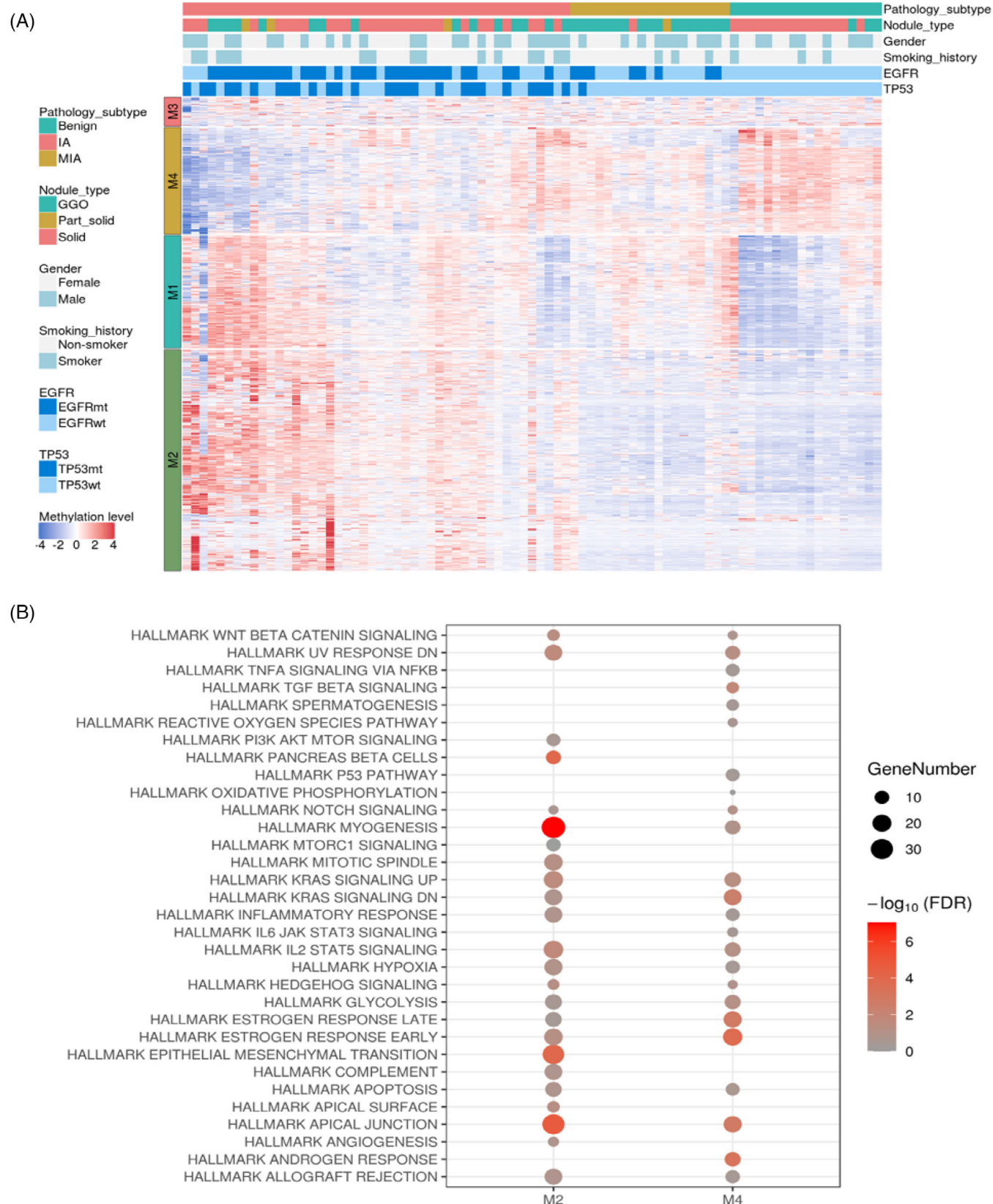


FIGURE 5 The DNA methylation characterization of lung nodules. (A) Phenotypes of DNA methylation cluster subgroups. The cluster assignments of patients are shown on the left. Red color indicates hypermethylation, and blue color indicates hypomethylation. (B) Pathway enrichment analysis for DNA methylation cluster subgroups M2 and M4.

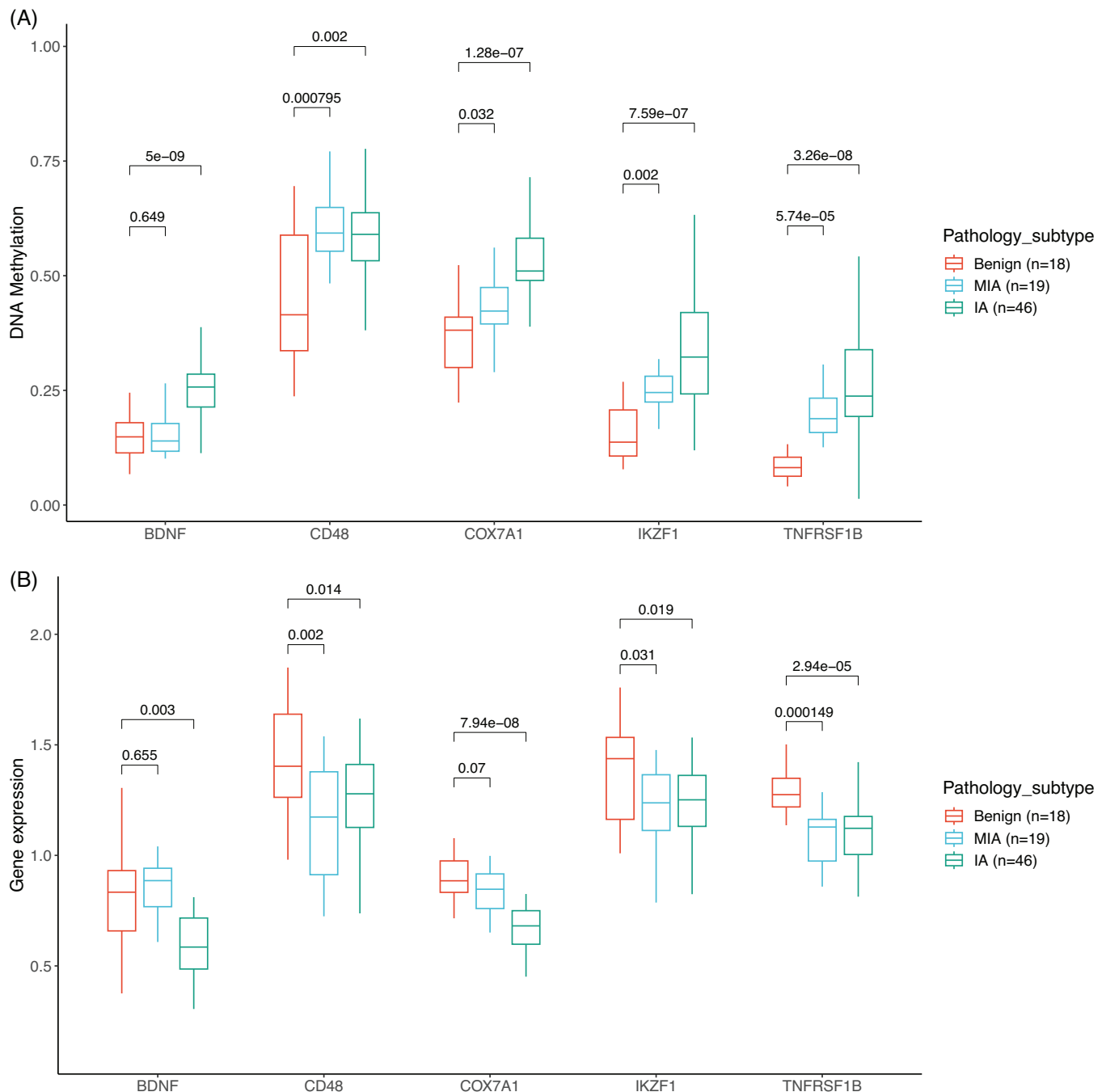


FIGURE 6 The integration of gene expression and DNA methylation in M2. (A) Comparison of DNA methylation levels between benign, MIA, and IA in the M2 group. (B) Comparison of gene expression levels between benign, MIA, and IA in the M2 group. ns, not significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$, Two-sided Mann-Whitney U test.

squamous cell carcinoma and NSCLC, linked to tumor proliferation and invasion.³⁶ *BDNF* can bind to TrkB receptors to activate ERK/CREB, PI3K/AKT, and NF- κ B pathways, ultimately enhancing tumor survival, proliferation, and migration.³⁷ In contrast, our study revealed that *BDNF* had higher methylation and lower expression in IA, suggesting that *BDNF* may have distinct functions in early-stage versus late-stage lung cancers.

Conversely, in the M4 group, nine genes, including *CDC6* (cell division cycle 6), *CLDN4* (claudin 4), *EGF* (epidermal growth factor), *EGFR*, *ESRP2* (epithelial splicing regulatory protein 2), *GRB7* (growth factor receptor bound protein 7), *KRT8* (keratin 8), *SFN* (stratifin), and *SLC22A18* (solute carrier family 22 member 18), exhibited progressively decreasing methylation levels and increasing levels of gene expression from benign, MIA, to IA (Figure 7).

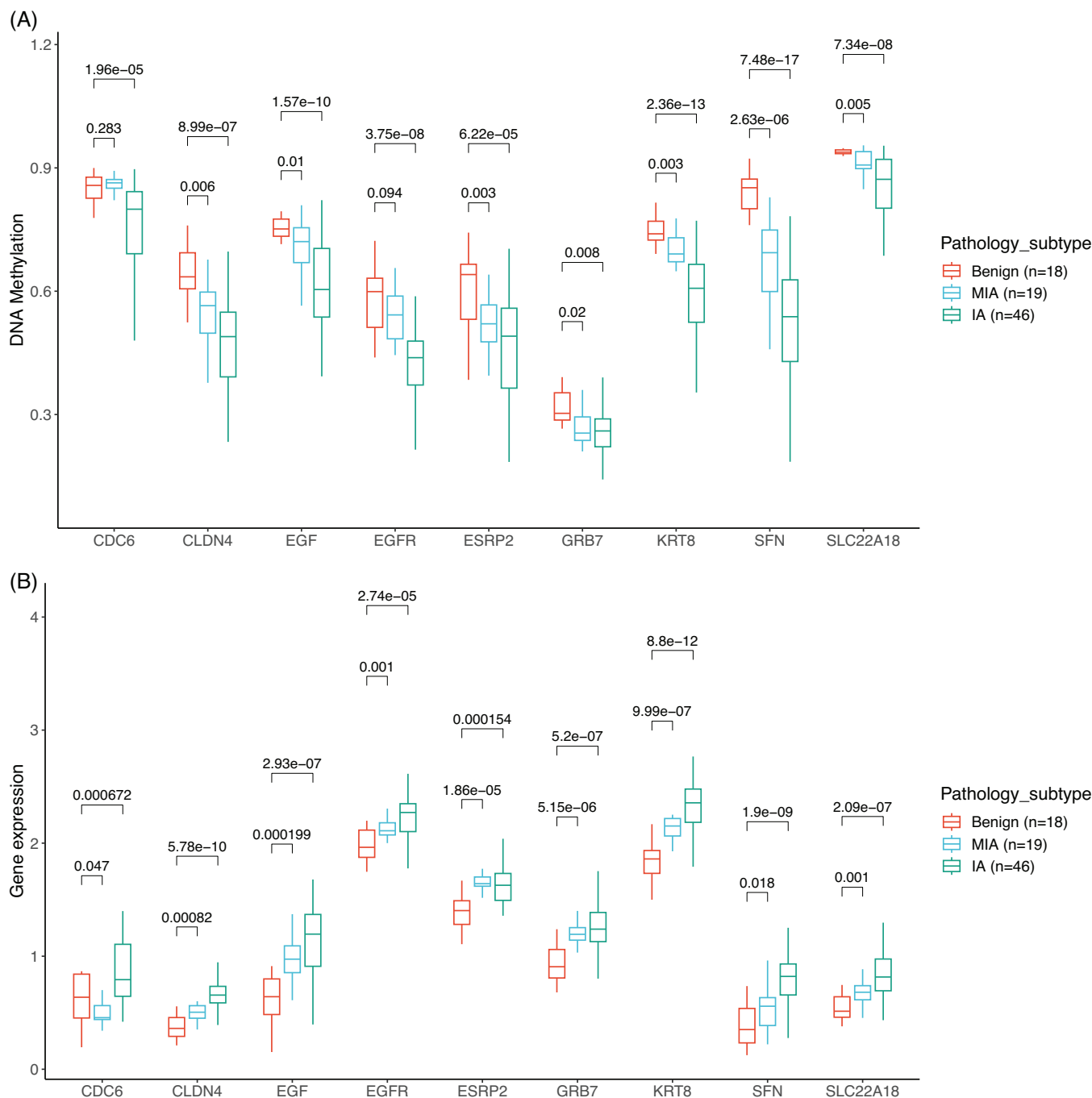


FIGURE 7 The integration of gene expression and DNA methylation in M4. (A) Comparison of DNA methylation levels between benign, MIA, and IA in the M4 group. (B) Comparison of gene expression levels between benign, MIA, and IA in the M4 group. ns, not significant, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$, Two-sided Mann-Whitney U test.

The highest inverse correlation values were observed for *EGF* (-0.78), *ESRP2* (-0.69), and *CDC6* (-0.63). *EGF* and *EGFR* had lower DNA methylation and higher expression in MIA and IA. Studies have suggested that the stimulation of *EGFR* by its ligand, *EGF*, may have pro-tumorigenic effects, such as enhancing cell migration, with clinical implications.³⁸ Upregulation of *EGFR* leads to activation of *EGFR* signaling and promotes tumor progression through

EGFR downstream pathways, including PI3K/Akt/mTOR and RAS/MEK/ERK pathways.^{39,40} Our findings emphasize the importance of the *EGFR* pathway and uncover a novel mechanism regulating this pathway during the neoplastic transformation from benign to IA. A previous study reported that NSCLC patients with strong overexpressing of *CDC6* had a significantly poorer prognosis compared to those with normal expression levels,⁴¹ consistent with our

research. *KRT8*, a type II intermediate filament protein, is over-expressed in various cancers, including lung cancer.⁴² While overexpression of *KRT8* has been attributed to DNA copy number gains,⁴³ our results suggested that DNA hypomethylation may represent another mechanism for the upregulation of *KRT8*. Collectively, these genes hold promise as potential biomarkers for diagnostics and drug targets for the early interception of lung cancer.

3 | DISCUSSION

Despite significant advancements in the genomic profiling of lung cancer in recent years, the genetic characteristics of benign, preinvasive, and invasive lung nodules remain largely unexplored. To the best of our knowledge, this study represents the most comprehensive multiomics investigation to date, aiming to elucidate the molecular pathogenesis of benign lung nodules and early-stage lung cancers within the Chinese population.

In this study, we identified a high mutation frequency in genes such as *EGFR*, *TP53*, and *RBM10* in invasive nodules. Similarly, a previous investigation on pre/minimally invasive lung adenocarcinoma in the Chinese population revealed mutated genes like *EGFR*, *RBM10*, *BRAF*, *ERBB2*, *TP53*, *KRAS*, *MAP2K1*, and *MET*.¹¹ Concordantly, the integrative proteomic characterization of LUAD revealed that clinical outcomes in early stage patients were linked to *EGFR* and *TP53* mutations.⁴⁴ Our study corroborates these previous observations and demonstrates that exon 21 p.L858R and exon19 deletions were the predominant *EGFR* mutations. This study also supported previous reports on the presence of somatic mutations in the tyrosine kinase (TK) domain of *EGFR*, short deletions in exon 19, and point mutations (p.G719S, p.L858R, and p.L861Q) in exons 18 and 21.^{45,46} Zhang's study reported the mutation rate of *EGFR* in AIS (31.3%, 5/16) was comparable to IAC (50%, 7/14) ($p = 0.46$),⁴⁷ while Zhu's study found that the *EGFR* mutation frequency increased from AIS (33.3%) to MIA (50.8%) but remained stable from MIA to IAC (50.2%) stage.⁴⁸ However, in our study, *EGFR* mutation frequency increased from AAH+AIS (16%), MIA (35%), to IA (73%), confirming the mutations of *EGFR* drive the transformation from preinvasive to invasive stages.

Interestingly, the TMB of nonsmokers was lower than that of smokers in both MIA and IA groups (Figure S5A), consistent with the previous report.⁴⁹ Given that our cohort consisted of early-stage lung cancer, we aimed to investigate the comparability of *EGFR* mutation frequency as a reference point by comparing the WES data of nonsmokers in our current cohort (early stage, MIA/IA) with the CHOICE cohort (late-stage). The results showed that the mutation frequencies of the most commonly

mutated genes, such as *EGFR* and *TP53*, were highly similar, with the exception of a few genes that exhibited notable differences (Figure S5B).

Abnormal regulations of alternative splicing can lead to the production of multiple transcriptional isoforms and diversified proteins, which often contribute to cancer development and progression. We identified frequent mutations in *RBM10* and *RBMX* among malignant nodules. *RBM10* mutations detected in LUADs disrupt the splicing of *NUMB* (NUMB endocytic adaptor protein), a NOTCH signaling pathway regulator, thereby promoting cell growth.^{50,51} Truncation mutations of *RBMX* identified in lung cancer suggest it as a potential tumor suppressor gene.^{52,53} *RBMX* mutations induced by cigarettes smoking may drive lung cancer development in smokers.⁵⁴ Although no significant difference was observed between MUT and WT groups among pathological subtypes in terms of alternative splicing events for *RBM10* and *RBMX*, their significant roles in lung cancer progression warrant further investigation in larger cohorts.

Besides *EGFR* mutations, our study uncovered a novel mechanism for activating the *EGFR* pathway in lung cancer. Hypomethylation of *EGF* and *EGFR* leads to increased expression levels of these genes, presumably leading to higher activity of this vital pathway and driving lung cancer development. The *ER/AR* pathway was also upregulated in the *EGFR*-driven group from the transcriptomic analysis. Estrogen and its receptors have been postulated to involve in the development and progression of lung cancer.⁵⁵ A study reported that estrogen induces NSCLC cell proliferation by promoting *c-Myc* and *Cyclin D1* expressions.⁵⁶ Furthermore, estrogen signaling through *ER β* regulates cell proliferation via the activation of *PI3K/IKK/NF- κ B*, *PI3K/AKT/Bcl-XL*, and *RAS/RAF/MEK/ERK* signaling pathways.⁵⁷ Estrogen has been reported to activate *ER* transactivation of the *EGFR* pathway and the nuclear localization of *Er β* , which is significantly associated with *EGFR* mutations.⁵⁷ The potential crosstalk between *EGFR* and *ER* signaling pathways has been documented in previous studies, leading to emerging clinical trials combining *EGFR* tyrosine kinase inhibitors (TKIs) with antiestrogen treatment.^{56,58} Our study implies that the interaction of *EGFR* with *ER* signaling pathway may be pivotal in the tumorigenesis process and could aid in identifying lung cancer patients who would benefit most from the combination treatment.

Understanding the immune landscape within the tumor microenvironment during lung nodule progression is essential, as it may reveal the underlying molecular mechanism of susceptibility and resistance to immunotherapy and inform more effective strategies to enhance treatment outcomes. The RNA-Seq data from the stem-like group, enriched with preinvasive nodules, suggested a

relationship between tumor stemness and the development of invasiveness. A recent study discovered that cancer stem-like cells (CSCs) enhanced *PD-L1* expression via *ALDH3A1* isozyme to promote immune evasion in lung tumors.⁵⁹ Our study implies a potential link between NK cell infiltration and tumor stemness. NK cells can target and shape CSC-nondifferentiated tumors, leading to the selection of differentiated tumor subpopulations with reduced proliferation rates, incapable of invasion or metastasis, and increased susceptibility to chemotherapy and radiotherapy.^{60,61} We also found that Treg cell infiltration was associated with *TP53* mutations. The loss of *TP53* function in tumors results in the accumulation of Treg cells.²⁵ Moreover, *TP53*-null cancer cells promote *CXCL17* (C-X-C motif chemokine ligand 17) secretion, which is also associated with the expansion of immunosuppressive Treg cells.⁶² Based on these findings, remodeling the tumor immune microenvironment in patients at preinvasive stages prior to immunosuppression, may represent a beneficial therapeutic strategy.

Besides the aberrant DNA methylation of *EGFR*, our study revealed that other methylation changes also contribute to lung cancer initiation and development. **Hypermethylation of genes such as *COX7A1*, *TNFRSF1B*, and *BDNF*, as well as hypomethylation of genes such as *CDC6* and *KRT8*, may be associated with lung cancer compared to benign nodules.** *COX7A1* can suppress the viability of human NSCLC cells by regulating autophagy through the downregulation of *PGC-1 α* and upregulation of *NOX2*.³³ As a critical mediator of *TNF/TNFR* signaling pathway, *TNFRSF1B* can promote apoptosis⁶³ and necroptosis,⁶⁴ which may explain the association between *TNFRSF1B* polymorphisms and the survival of lung cancer patients treated with chemoradiotherapy.^{35,64} Apart from its role as a neurotrophins, *BDNF* can also enhance tumor survival, proliferation, and migration by activating the ERK/CREB, PI3K/Akt, and NF- κ B pathways via binding to the TrkB receptor.⁶⁵ As a regulator of DNA replication, downregulation of *CDC6* can inhibit the DNA damage response activity of cancer stem-like cells.⁶⁶ *KRT8* has been identified as a target of cisplatin, and the *KRT8/AKT* signaling pathway plays a crucial role in lung cancer metastasis by regulating the function of cancer-associated fibroblast in the tumor microenvironment.⁶⁷ These genes may play a critical role in the initiation and development of cancer and could potentially be used as diagnostic biomarkers and drug targets for early-stage lung cancers.

This study has several limitations that should be acknowledged. First, we included benign nodules instead of normal lung tissues, as benign nodules in CT images can be easily confused with malignant nodules, making it

challenging to differentiate between them. Due to their frequent overdiagnosis and overtreatment in clinical practice, our goal was to distinguish the molecular characteristics of benign and malignant nodules. Second, since AAH and AIS are defined as precursor glandular lesions and are rarely surgically resected in clinical practice, it is regrettable that the AAH+AIS group was relatively small and lacked epigenetic data in the entire cohort. Third, we have not conducted in vitro/vivo experiments to investigate the molecular functions, such as the interaction between the *EGFR* pathway and the estrogen/androgen pathways. Future studies will further validate these findings in a more careful manner. Lastly, comparing different patient samples is influenced by various factors; however, investigating the molecular pathogenesis of nodule progression within an individual patient or within a single tumor is the best way to proceed. It is challenging to track a patient from the initial appearance of a lung nodule to the development of lung cancer and obtain tissue samples at various stages of the disease. Moreover, cases with multiple lesions containing benign, preinvasive, and invasive subtypes are uncommon. Therefore, larger cohorts with long-term follow-up are essential to validate our findings.

In summary, our study has presented the genomic landscape of lung nodules through an integrated analysis of whole exome sequencing, RNA-Seq, and DNA methylation sequencing. This comprehensive approach enhances our understanding of the distinct molecular mechanisms underlying benign, preinvasive, and invasive lung nodules. Moreover, we have identified potential biomarkers for the early detection of lung cancer and drug targets for lung cancer treatments. Our study offers an unprecedented resource for unraveling cancer evolution and developing interception solutions.

4 | METHODS AND MATERIALS

4.1 | Participating patients and sample collection

This study enrolled patients with lung nodules (5–30 mm in diameter) detected by CT/LDCT scan at The First Affiliated Hospital of Guangzhou Medical University in China between 2018 and 2019. None of the patients had received any preoperative cancer treatments. FFPE tissue samples were acquired from subsequent surgical resections, and matched WBCs were also collected to subtract germline mutations. Both genders were included in the study, and smoking history information was gathered. The pathological results of all samples were determined based on surgically resected tissue sections according to the 2015

WHO Histological Classification of Lung Cancer.⁶⁸ Written informed consents was obtained from all participants.

Sections of 5–10 μ m in thickness were cut from the FFPE tissue samples of 432 patients with lung nodules. Approximately 8–10 mL of whole blood sample was drawn from each patient using a Cell-Free DNA BCT blood collection tube (Streck, Inc. Cat# 218962). WBCs were separated immediately upon receipt of the whole blood samples according to the standard protocol and stored at -80°C until DNA isolation. Purity of tumor cells were estimated by two independent pathologists by reviewing Hematoxylin and Eosin (H&E) staining slides, with a minimum tumor cell content of 10% required for subsequent sequencing.

4.2 | Nucleic acid isolations

gDNA and total RNA were simultaneously extracted from FFPE tissue samples using AllPrep FFPE DNA/RNA Kit (Qiagen, Cat# 80234) according to the manufacturer's instructions. gDNA from WBC samples was isolated using the DNeasy Blood & Tissue Kit (Qiagen, Cat# 69504). The concentrations of DNA and RNA were measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Cat# Q32854) and Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Cat# Q32855), respectively. The quality of DNA and RNA were assessed using the Agilent High Sensitivity DNA Kit (Cat# 5067-4626) and the Agilent RNA 6000 Pico Kit (Cat# 5067-1513) on a 2100 Bioanalyzer Instrument (Agilent Technologies, Inc.), respectively. Only samples with DV200 values (the percentage of fragments > 200 nucleotides) $\geq 30\%$ were selected for RNA-Seq.

4.3 | Whole exome sequencing and RNA-Seq

WES and RNA-Seq were carried out by Mingma Technologies Co., Ltd. (Shanghai, China). In brief, 300 ng of gDNA from FFPE tissue or WBC samples were fragmented into 150–200 bp pieces (mean size) using a LE220-plus Focused-ultrasonicator (Covaris, Inc.). DNA libraries were prepared using SureSelectXT Reagent Kit (Agilent, Cat# G9611A) following the manufacturer's instructions. Whole exome enrichment was conducted using SureSelectXT Human All Exon V6 Kit (Agilent, Cat# 5190–8864). For RNA from FFPE tissue, 200 ng was converted into sequencing libraries using the TruSeq RNA Exome Kits (Illumina, Cat# 20020189, 20020490, 20020492/20020493, 20024145). The Resulting libraries were quantified using the Qubit dsDNA HS Assay Kit, and their size distributions were analyzed by Agilent DNA 1000 Kit (Cat#

5067-1504) on a 4200 TapeStation System (Agilent Technologies, Inc.). Sequencing was performed in the 2×150 -bp paired-end mode on the HiSeq X Ten or NovaSeq 6000 platform (Illumina, Inc.).

4.4 | Targeted DNA methylation assay

Full details of DNA fragmentation, bisulfite conversion, and targeted methylation sequencing were described previously.⁶⁹ Briefly, gDNA from 83 selected samples were fragmented into approximately 200 bp (peak size) using an M220 Focused-ultrasonicator (Covaris, Inc.) following the manufacturer's instructions. 50 ng of purified, fragmented gDNA was used for the subsequent bisulfite conversion step. Bisulfite conversion was carried out using the EZ DNA Methylation-Lightning Kit (Zymo Research, Cat# D5031) according to the manufacturer's protocol. The AnchorIRIS library preparation technology was employed for targeted methylation analysis. AnchorIRIS prehybridization library construction was performed using the AnchorDx EpiVisio Methylation Library Prep Kit (AnchorDx, Cat# A0UX00019) and AnchorDx EpiVisio Indexing PCR Kit (AnchorDx, Cat# A2DX00025). Target enrichment was performed using AnchorDx EpiVisio Target Enrichment Kit (AnchorDx, Cat# A0UX00031). A custom-made lung cancer methylation panel (LC Panel), which includes 12,899 preselected probes enriched for lung cancer-specific methylation regions, was utilized in the present study.⁷⁰ After probe hybridization, the resulting libraries were sequenced in the 2×150 -bp paired-end mode on the NovaSeq 6000 platform.

4.5 | Data quality control

We used the default parameters of the FastQC (version 0.11.2) (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) software to perform quality control (QC) on all samples. Skewer (v0.2.2)⁷¹ software was applied to remove splice sequences from the sequencing platform raw data (Fastq). Sequences with a length of at least 75 bp were filtered by the splice removal process, and the resulting filtered data (clean reads) were used for downstream bioinformatic analyses. The exome capture efficiency and targeted region coverage were evaluated across the entire sample set. Lung nodules were sequenced at a mean targeted region coverage of 417 $\times$, with 97.7% coverage of $\geq 20\times$; WBCs were sequenced at a mean targeted region coverage of 215 $\times$, with 96.7% coverage of $\geq 20\times$, qualifying for subsequent variant analysis. The quality scores of 30 (%) for RNAseq was greater than 93.5. The average sequencing data amounts were 62.3 M, the average clean

reads obtained after removing splices and low quality were 62.2 M, the average clean rate was 99.2%, and the average mapping rate was greater than 95%. For DNA methylation sequencing, the bisulfite conversion efficiency was greater than 99.2%, and the mean targeted region coverage was 785×, with 99% coverage of > 30×.

Three samples were excluded from further analysis due to low coverage in WES, two samples due to low coverage in RNA-seq, and five samples due to a mismatch between tumor and normal samples. Therefore, 422 patients were included in the final analysis. Among them, 422 had qualified WES data, 228 had qualified RNA-seq data, and 226 had both WES and RNA-seq data (paired). For DNA methylation sequencing, 20 benign and 80 malignant samples (including 20 with *TP53* mutations, 20 with *EGFR* mutations, 20 with *TP53* & *EGFR* co-mutations, and 20 without mutations) were included. Seventeen samples were excluded due to failing sequencing QC. Finally, 83 samples had qualified DNA methylation sequencing data.

4.6 | Somatic variant identification

After removing adapters and low-quality reads, the Sentieon (version 201911)⁶⁹ pipeline with default parameters was implemented to process the following steps sequentially: reads alignment to the human genome reference assembly hg19 using the Burrows–Wheeler Aligner (BWA) algorithm and sorting, removal of PCR duplicates, InDel realign, base quality score recalibration, somatic mutation calling including single nucleotide variations (SNVs) and short insertion/deletions (indels), and variant quality score recalibration. During the mutation calling stage, the reads from the tumor sample were compared with the paired blood sample from the same patient to generate the somatic mutation list. Somatic mutations were then filtered and retained only with the variant allele frequency (VAF) ≥ 0.05 and supported by at least three reads. They were further annotated using the Variant Effect Predictor (VEP) package.⁷⁰

4.7 | Tumor mutation burden (TMB)

TMB was defined as the total number of somatic nonsynonymous mutations (SNVs or indels) in the tumor exome for each patient. The number was divided by the total size of the targeted regions to calculate the TMB score in counts/Mb. The Agilent SureSelect Human All Exon V6 kit was used, and its estimated total targeting size (exome) is 38 Mb. TMB was regressed on each gene to test for a statistical association between genes and TMB. Genes were

identified according to the following criteria: correlation $|r| > 0.4$ and false discovery rate (FDR) < 0.05 .

4.8 | Copy number variation (CNV)

CNVkit v0.8.5 (<https://cnvkit.readthedocs.io/en/stable/index.html>) was used to analyze the copy number variation. In particular, for tumor/normal pairwise analysis, all normal samples were pooled together and then each tumor sample was compared with the normal pool to obtain tumor-specific CNV, with filters of $|\log_2 \text{ratio}| \geq 0.4$ and bin number ≥ 5 .

4.9 | Splicing event counting

After assigning transcripts, alternative splicing events were identified using vast-tools. Percent Splice In (PSIs) calculated by vast-tools was then filtered using the cutoff $\text{psi} > 80$ (for IR, Alt5, Alt3) or $\text{psi} < 20$ (for EX) to count event number for each splicing type. Only high-frequency splicing events (occurring in ≥ 10 samples) were preserved. The total splicing event number of a sample is the sum of the event number for each splicing type. The splicing gene list was curated based on the mRNA splicing gene set from Reactome.⁷²

4.10 | Differential gene expression analysis

RNA-seq reads were mapped to human genome reference assembly hg19 using STAR (version 020201).⁷³ The raw read counts were then normalized by log2-counts per million normalization. The list of differentially expressed genes between groups was generated using DESeq2.⁷⁴ Biological pathway enrichment was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and MSigDB hallmark gene sets. The FDR was calculated using the Benjamini and Hochberg (BH) false discovery rate (FDR) algorithm. We used an FDR ≤ 0.05 as the criterion to select significantly enriched KEGG pathways.

4.11 | KEGG pathway enrichment analysis

The KEGG pathway database was used to perform pathway enrichment analysis. ClusterProfiler was used to obtain the latest online database and conduct an over-representation test with the “enrichKEGG” function.⁷⁵

4.12 | Clustering analysis of ssGSEA scores based on hallmark gene set

The single sample gene set enrichment analysis (ssGSEA) scores of each sample based on Hallmark 50 gene sets were calculated, and then unsupervised hierarchical clustering of samples was performed based on the Hallmark ssGSEA scores to generate subgroups using the *seaborn* (version 0.11.2) package. The clustering result was then visualized by *matplotlib* in Python, and the immune cell ssGSEA scores corresponding to each sample were also plotted in the clustering heatmap. The Hallmark gene set used here was downloaded from the Gene set enrichment analysis (GSEA) molecular signatures database (gsea-msigdb.org).

4.13 | Immune cell signature score

RNA-seq reads were assembled using *StringTie2* (version 1.3.5), and FPKM and TPM data matrices were generated. TPM data were used to calculate the immunopheno score using the immunophenogram R package.⁷⁶ FPKM data were used to calculate the immune infiltration score using the ssGSEA method.²³ Infiltration cell types were clustered by heatmap analysis with the *seaborn* package and visualized by *matplotlib* in Python.

4.14 | DNA methylation analysis

Sequencing adapters and three prime low-quality bases were trimmed from raw sequencing reads using *fastp* (v0.19.6) software.⁷⁷ *In silico* converted hg19 was used as reference genome and cleaned reads were mapped to reference genome using *Bismark* (v0.18.2) software.⁷⁸ Aligned reads were evaluated by *Picard* (v2.5.0) for metrics that measured the performance of target-capture based bisulfite sequencing assays (<http://broadinstitute.github.io/picard>). The “*bismark_methylation_extractor*” tool was used to calculate methylation level for each targeted CpG site.

4.15 | Differential methylation analysis

Differential methylation analysis was conducted using R package *DSS* (version 2.14.0).⁷⁹ Differentially methylated CpGs (DMs) were first identified (criteria: FDR < 0.05, δ > 0.05), and adjacent DMs were further merged into differentially methylated regions (DMRs). DMRs were intersected with protein-coding genes (hg19 Ensembl (v75), $n = 20,232$) by using *annovar*.⁸⁰

4.16 | Unsupervised hierarchical clustering of DNA methylation datasets

The preselected 12,899 probes were used for unsupervised hierarchical clustering. The methylation level of each targeted regions was calculated as the ratio of the methylated CpGs and the total sequenced CpGs (sum of methylated and unmethylated CpGs). Before clustering, the methylation levels of each targeted region were Z-score normalized. The R function “*hclust*” was used to perform hierarchical clustering with “*ward.D2*” as the clustering algorithm. GSEA and MSigDB with BioCarta gene set, oncogenic signature gene sets and Hallmark gene sets were also used for biological pathway enrichment analysis.

4.17 | Statistics analysis

Unless specified otherwise, Pearson's chi-square test was used for *p*-value calculations between two categorical variables. The two-sided Mann-Whitney *U* test was used for comparisons between two continuous variables. The most frequently mutated genes and their distribution across different clinical features, the significantly different mutation genes between groups, and the mutually exclusive or co-occurring mutated genes were calculated and visualized using the *MAFtools* package in R.⁸¹ GSEA was performed using the GSEA software.⁸² Pearson's correlation analysis was employed to calculate the coefficient between gene expression and methylation data. The *p*-value < 0.05 was determined to be statistically significant for all analyses.

AUTHOR CONTRIBUTIONS

Wenhua Liang, Jianbing Fan, Bo Wang, and Jianxing He conceived the study. Wenhua Liang, Jianbing Fan, Bo Wang, and Peng Liang designed the experiments. Bo Wang, Jun Wang, Hui Li, and Zhiwei Chen performed the experiments. Wenhua Liang, Lixuan Lin, Bo Cheng, Shan Xiong, Jianfu Li, Caichen Li, Ziwen Yu, and Chunyan Li contributed to data acquisition. Peng Liang, Minhua Peng, Jinsheng Tao, and Jinwang Wei contributed to data visualization. Wenhua Liang, Minhua Peng, Bo Wang, Jinwang Wei, and Jinsheng Tao wrote the manuscript. Jianxing He, Jianbing Fan, and Zhiwei Chen contributed substantially to the development of this manuscript. All authors reviewed and approved the manuscript.

ACKNOWLEDGMENTS

This study is supported by the China National Science Foundation (NO: 82022048 & 81871893); Research Project of Guangzhou Lab (NO: SRPG22-017 & GZNL2023A02007); Scheme of Guangzhou Economic

and Technological Development District for Leading Talents in Innovation and Entrepreneurship (NO: 2017-L152); Scheme of Guangzhou for Leading Talents in Innovation and Entrepreneurship (NO: 2016007); Scheme of Guangzhou for Leading Team in Innovation (NO: 201909010010); 2020 Guangzhou Development Zone International Science and Technology Cooperation Project (NO: 2020GH15); Science and Technology Planning Project of Guangzhou (NO: 202206080013); Guangzhou Development Zone International Science and Technology Cooperation Project (NO: 2021GH17), Key Project of Guangzhou Scientific Research Project (NO: 201804020030), National Key Research & Development Programme (NOs: 2022YFC2505100, 2017YFC0112704), High-Level University Construction Project of Guangzhou Medical University (NOs: 20182737, 201721007, 201715907, 2017160107).

CONFLICT OF INTEREST STATEMENT

Authors Minhua Peng, Jinsheng Tao, Jun Wang, Hui Li, Zhiwei Chen, Jianbing Fan, and Bo Wang are current employees of AnchorDx Medical Co., Ltd, or AnchorDx, Inc. Jinwang Wei is a current employee of Genomicare Biotechnology (Shanghai) Co., Ltd and Shanghai CreateCured Biotechnology Co., Ltd. All other authors declare no competing financial interest.

DATA AVAILABILITY STATEMENT

AnchorDx Medical Co., Ltd provides access to the study protocol, the statistical analysis plan, the clinical study report, and all individual participant data except genetic data with academic researchers. Access is provided after a proposal has been approved by an Independent Review Committee identified for this purpose and after receipt of a signed Data Use Agreement. Proposals should be directed to contact-us@anchordx.com.

ETHICS STATEMENT

This study was approved by the Ethical Committees of The First Affiliated Hospital of Guangzhou Medical University (No. 2021–145). Written informed consents was obtained from all participants.

ORCID

Wenhua Liang  <https://orcid.org/0000-0002-1391-8238>
Jianxing He  <https://orcid.org/0000-0003-1737-8192>

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424. doi:10.3322/caac.21492
- Novello S, Barlesi F, Califano R, et al. Metastatic non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2016;27(5):v1–v27. suppl. doi:10.1093/annonc/mdw326
- Kharazmi E, Fallah M, Sundquist K, Hemminki K. Familial risk of early and late onset cancer: nationwide prospective cohort study. *BMJ*. 2012;345:e8076. doi:10.1136/bmj.e8076
- Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol*. 2017;7(9):170070. doi:10.1098/rsob.170070
- Tanoue LT, Tanner NT, Gould MK, Silvestri GA. Lung cancer screening. *Am J Respir Crit Care Med*. 2015;191(1):19–33. doi:10.1164/rccm.201410-1777CI
- Mazzone PJ, Silvestri GA, Patel S, et al. Screening for lung cancer: CHEST guideline and expert panel report. *Chest*. 2018;153(4):954–985. doi:10.1016/j.chest.2018.01.016
- Inamura K. Clinicopathological characteristics and mutations driving development of early lung adenocarcinoma: tumor initiation and progression. *Int J Mol Sci*. 2018;19(4):1259. doi:10.3390/ijms19041259
- Suzuki K, Watanabe SI, Wakabayashi M, et al. A single-arm study of sublobar resection for ground-glass opacity dominant peripheral lung cancer. *J Thorac Cardiovasc Surg*. 2022;163(1):289–301. e2. doi:10.1016/j.jtcvs.2020.09.146
- Hu X, Fujimoto J, Ying L, et al. Multi-region exome sequencing reveals genomic evolution from preneoplasia to lung adenocarcinoma. *Nat Commun*. 2019;10(1):2978. doi:10.1038/s41467-019-10877-8
- Hu X, Estecio MR, Chen R, et al. Evolution of DNA methylome from precancerous lesions to invasive lung adenocarcinomas. *Nat Commun*. 2021;12(1):687. doi:10.1038/s41467-021-20907-z
- Chen H, Carrot-Zhang J, Zhao Y, et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. *Nat Commun*. 2019;10(1):5472. doi:10.1038/s41467-019-13460-3
- Dejima H, Hu X, Chen R, et al. Immune evolution from preneoplasia to invasive lung adenocarcinomas and underlying molecular features. *Nat Commun*. 2021;12(1):2722. doi:10.1038/s41467-021-22890-x
- The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519–525. doi:10.1038/nature11404
- The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543–550. doi:10.1038/nature13385
- Yoon HY, Ryu JS, Sim YS, et al. Clinical significance of EGFR mutation types in lung adenocarcinoma: a multi-centre Korean study. *PLoS One*. 2020;15(2):e0228925. doi:10.1371/journal.pone.0228925
- Offin M, Rizvi H, Tenet M, et al. Tumor mutation burden and efficacy of EGFR-tyrosine kinase inhibitors in patients with EGFR-mutant lung cancers. *Clin Cancer Res*. 2019;25(3):1063–1069. doi:10.1158/1078-0432.CCR-18-1102
- Li Y, Li X, Li H, et al. Genomic characterisation of pulmonary subsolid nodules: mutational landscape and radiological features. *Eur Respir J*. 2020;55(2):1901409. doi:10.1183/13993003.01409-2019
- Chen J, Yang H, Teo ASM, et al. Genomic landscape of lung adenocarcinoma in East Asians. *Nat Genet*. 2020;52(2):177–186. doi:10.1038/s41588-019-0569-6

19. Jiao XD, He X, Qin BD, et al. The prognostic value of tumor mutation burden in EGFR-mutant advanced lung adenocarcinoma, an analysis based on cBioPortal data base. *J Thorac Dis.* 2019;11(11):4507-4515. doi:[10.21037/jtd.2019.11.04](https://doi.org/10.21037/jtd.2019.11.04)
20. Fu J, Li Y, Li C, Tong Y, Li M, Cang S. A special prognostic indicator: tumor mutation burden combined with immune infiltrates in lung adenocarcinoma with TP53 mutation. *Transl Cancer Res.* 2021;10(9):3963-3978. doi:[10.21037/tcr-21-565](https://doi.org/10.21037/tcr-21-565)
21. Fan S, Liao Y, Liu C, et al. Estrogen promotes tumor metastasis via estrogen receptor beta-mediated regulation of matrix-metalloproteinase-2 in non-small cell lung cancer. *Oncotarget.* 2017;8(34):56443-56459. doi:[10.18632/oncotarget.16992](https://doi.org/10.18632/oncotarget.16992)
22. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science.* 2009;324(5930):1029-1033. doi:[10.1126/science.1160809](https://doi.org/10.1126/science.1160809)
23. Zhang XC, Wang J, Shao GG, et al. Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients. *Nat Commun.* 2019;10(1):1772. doi:[10.1038/s41467-019-09762-1](https://doi.org/10.1038/s41467-019-09762-1)
24. Dong ZY, Zhang JT, Liu SY, et al. EGFR mutation correlates with uninflamed phenotype and weak immunogenicity, causing impaired response to PD-1 blockade in non-small cell lung cancer. *Oncoimmunology.* 2017;6(11):e1356145. doi:[10.1080/2162402X.2017.1356145](https://doi.org/10.1080/2162402X.2017.1356145)
25. Blagih J, Zani F, Chakravarty P, et al. Cancer-specific loss of p53 leads to a modulation of myeloid and T cell responses. *Cell Rep.* 2020;30(2):481-496. e6. doi:[10.1016/j.celrep.2019.12.028](https://doi.org/10.1016/j.celrep.2019.12.028)
26. Chockley PJ, Chen J, Chen G, Beer DG, Standiford TJ, Keshamouni VG. Epithelial-mesenchymal transition leads to NK cell-mediated metastasis-specific immunosurveillance in lung cancer. *J Clin Invest.* 2018;128(4):1384-1396. doi:[10.1172/JCI97611](https://doi.org/10.1172/JCI97611)
27. Grzywacz B, Kataria N, Kataria N, Blazar BR, Miller JS, Verneris MR. Natural killer-cell differentiation by myeloid progenitors. *Blood.* 2011;117(13):3548-3558. doi:[10.1182/blood-2010-04-281394](https://doi.org/10.1182/blood-2010-04-281394)
28. Singh M, Yelle N, Venugopal C, Singh SK. EMT: mechanisms and therapeutic implications. *Pharmacol Ther.* 2018;182:80-94. doi:[10.1016/j.pharmthera.2017.08.009](https://doi.org/10.1016/j.pharmthera.2017.08.009)
29. Shi H, Liu C, Tan H, et al. Hippo kinases Mst1 and Mst2 sense and amplify IL-2R-STAT5 signaling in regulatory T cells to establish stable regulatory activity. *Immunity.* 2018;49(5):899-914. e6. doi:[10.1016/j.immuni.2018.10.010](https://doi.org/10.1016/j.immuni.2018.10.010)
30. Neuzillet C, Tijeras-Raballand A, Cohen R, et al. Targeting the TGFbeta pathway for cancer therapy. *Pharmacol Ther.* 2015;147:22-31. doi:[10.1016/j.pharmthera.2014.11.001](https://doi.org/10.1016/j.pharmthera.2014.11.001)
31. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. *Oncotarget.* 2012;3(4):462-474. doi:[10.18632/oncotarget.497](https://doi.org/10.18632/oncotarget.497)
32. Wang J, Yu X, Ouyang N, et al. Multi-platform analysis of methylation-regulated genes in human lung adenocarcinoma. *J Toxicol Environ Health A.* 2019;82(1):37-45. doi:[10.1080/15287394.2018.1551645](https://doi.org/10.1080/15287394.2018.1551645)
33. Zhao L, Chen X, Feng Y, et al. COX7A1 suppresses the viability of human non-small cell lung cancer cells via regulating autophagy. *Cancer Med.* 2019;8(18):7762-7773. doi:[10.1002/cam4.2659](https://doi.org/10.1002/cam4.2659)
34. Guo Y, Feng Y, Liu H, et al. Potentially functional genetic variants in the TNF/TNFR signaling pathway genes predict survival of patients with non-small cell lung cancer in the PLCO cancer screening trial. *Mol Carcinog.* 2019;58(7):1094-1104. doi:[10.1002/mc.23017](https://doi.org/10.1002/mc.23017)
35. Guan X, Liao Z, Ma H, et al. TNFRSF1B +676 T>G polymorphism predicts survival of non-small cell lung cancer patients treated with chemoradiotherapy. *BMC Cancer.* 2011;11:447. doi:[10.1186/1471-2407-11-447](https://doi.org/10.1186/1471-2407-11-447)
36. Zhang SY, Hui LP, Li CY, Gao J, Cui ZS, Qiu XS. More expression of BDNF associates with lung squamous cell carcinoma and is critical to the proliferation and invasion of lung cancer cells. *BMC Cancer.* 2016;16:171. doi:[10.1186/s12885-016-2218-0](https://doi.org/10.1186/s12885-016-2218-0)
37. Chen B, Liang Y, He Z, An Y, Zhao W, Wu J. Autocrine activity of BDNF induced by the STAT3 signaling pathway causes prolonged TrkB activation and promotes human non-small-cell lung cancer proliferation. *Sci Rep.* 2016;6:30404. doi:[10.1038/srep30404](https://doi.org/10.1038/srep30404)
38. Oda K, Matsuoka Y, Funahashi A, Kitano H. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol.* 2005;1:2005. 0010. doi:[10.1038/msb4100014](https://doi.org/10.1038/msb4100014)
39. Yuan M, Huang LL, Chen JH, Wu J, Xu Q. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. *Signal Transduct Target Ther.* 2019;4:61. doi:[10.1038/s41392-019-0099-9](https://doi.org/10.1038/s41392-019-0099-9)
40. Malapelle U, Passiglia F, Cremolini C, et al. RAS as a positive predictive biomarker: focus on lung and colorectal cancer patients. *Eur J Cancer.* 2021;146:74-83. doi:[10.1016/j.ejca.2021.01.015](https://doi.org/10.1016/j.ejca.2021.01.015)
41. Allera-Moreau C, Rouquette I, Lepage B, et al. DNA replication stress response involving PLK1, CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients. *Oncogenesis.* 2012;1(10):e30. doi:[10.1038/oncsis.2012.29](https://doi.org/10.1038/oncsis.2012.29)
42. Selamat SA, Chung BS, Girard L, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res.* 2012;22(7):1197-1211. doi:[10.1101/gr.132662.111](https://doi.org/10.1101/gr.132662.111)
43. Xie L, Dang Y, Guo J, et al. High KRT8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes (Basel).* 2019;10(1):36. doi:[10.3390/genes10010036](https://doi.org/10.3390/genes10010036)
44. Xu JY, Zhang C, Wang X, et al. integrative proteomic characterization of human lung adenocarcinoma. *Cell.* 2020;182(1):245-261. e17. doi:[10.1016/j.cell.2020.05.043](https://doi.org/10.1016/j.cell.2020.05.043)
45. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med.* 2004;350(21):2129-2139. doi:[10.1056/NEJMoa040938](https://doi.org/10.1056/NEJMoa040938)
46. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science.* 2004;304(5676):1497-1500. doi:[10.1126/science.1099314](https://doi.org/10.1126/science.1099314)
47. Zhang C, Zhang J, Xu FP, et al. Genomic landscape and immune microenvironment features of preinvasive and early invasive lung adenocarcinoma. *J Thorac Oncol.* 2019;14(11):1912-1923. doi:[10.1016/j.jtho.2019.07.031](https://doi.org/10.1016/j.jtho.2019.07.031)
48. Zhu J, Wang W, Xiong Y, et al. Evolution of lung adenocarcinoma from preneoplasia to invasive adenocarcinoma. *Cancer Med.* 2023;12(5):5545-5557. doi:[10.1002/cam4.5393](https://doi.org/10.1002/cam4.5393)
49. Sun LY, Cen WJ, Tang WT, et al. Smoking status combined with tumor mutational burden as a prognosis predictor for combination immune checkpoint inhibitor therapy in non-small

- cell lung cancer. *Cancer Med*. 2021;10(19):6610-6617. doi:[10.1002/cam4.4197](https://doi.org/10.1002/cam4.4197)
50. Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012;150(6):1107-1120. doi:[10.1016/j.cell.2012.08.029](https://doi.org/10.1016/j.cell.2012.08.029)
 51. Seiler M, Peng S, Agrawal AA, et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Rep*. 2018;23(1):282-296. e4. doi:[10.1016/j.celrep.2018.01.088](https://doi.org/10.1016/j.celrep.2018.01.088)
 52. Yan Q, Zeng P, Zhou X, et al. RBMX suppresses tumorigenicity and progression of bladder cancer by interacting with the hnRNP A1 protein to regulate PKM alternative splicing. *Oncogene*. 2021;40(15):2635-2650. doi:[10.1038/s41388-021-01666-z](https://doi.org/10.1038/s41388-021-01666-z)
 53. Renieri A, Mencarelli MA, Cetta F, et al. Oligogenic germline mutations identified in early non-smokers lung adenocarcinoma patients. *Lung Cancer*. 2014;85(2):168-174. doi:[10.1016/j.lungcan.2014.05.020](https://doi.org/10.1016/j.lungcan.2014.05.020)
 54. Zhang D, Qu L, Zhou B, Wang G, Zhou G. Genomic variations in the counterpart normal controls of lung squamous cell carcinomas. *Front Med*. 2018;12(3):280-288. doi:[10.1007/s11684-017-0580-1](https://doi.org/10.1007/s11684-017-0580-1)
 55. Chakraborty S, Ganti AK, Marr A, Batra SK. Lung cancer in women: role of estrogens. *Expert Rev Respir Med*. 2010;4(4):509-518. doi:[10.1586/ers.10.50](https://doi.org/10.1586/ers.10.50)
 56. Rodriguez-Lara V, Hernandez-Martinez JM, Arrieta O. Influence of estrogen in non-small cell lung cancer and its clinical implications. *J Thorac Dis*. 2018;10(1):482-497. doi:[10.21037/jtd.2017.12.61](https://doi.org/10.21037/jtd.2017.12.61)
 57. Hsu LH, Chu NM, Kao SH. Estrogen, estrogen receptor and lung cancer. *Int J Mol Sci*. 2017;18(8):1713. doi:[10.3390/ijms18081713](https://doi.org/10.3390/ijms18081713)
 58. Mazieres J, Barlesi F, Rouquette I, et al. Randomized Phase II trial evaluating treatment with EGFR-TKI associated with anti-estrogen in women with nonsquamous advanced-stage NSCLC: IFCT-1003 LADIE trial. *Clin Cancer Res*. 2020;26(13):3172-3181. doi:[10.1158/1078-0432.CCR-19-3056](https://doi.org/10.1158/1078-0432.CCR-19-3056)
 59. Terzuoli E, Bellan C, Aversa S, et al. ALDH3A1 overexpression in melanoma and lung tumors drives cancer stem cell expansion, impairing immune surveillance through enhanced PD-L1 Output. *Cancers (Basel)*. 2019;11(12):1963. doi:[10.3390/cancers11121963](https://doi.org/10.3390/cancers11121963)
 60. Kozłowska AK, Topchyan P, Kaur K, et al. Differentiation by NK cells is a prerequisite for effective targeting of cancer stem cells/poorly differentiated tumors by chemopreventive and chemotherapeutic drugs. *J Cancer*. 2017;8(4):537-554. doi:[10.7150/jca.15989](https://doi.org/10.7150/jca.15989)
 61. Kozłowska AK, Kaur K, Topchyan P, Jewett A. Adoptive transfer of osteoclast-expanded natural killer cells for immunotherapy targeting cancer stem-like cells in humanized mice. *Cancer Immunol Immunother*. 2016;65(7):835-845. doi:[10.1007/s00262-016-1822-9](https://doi.org/10.1007/s00262-016-1822-9)
 62. Bezzi M, Seitzer N, Ishikawa T, et al. Diverse genetic-driven immune landscapes dictate tumor progression through distinct mechanisms. *Nat Med*. 2018;24(2):165-175. doi:[10.1038/nm.4463](https://doi.org/10.1038/nm.4463)
 63. Lee SY, Kang HG, Yoo SS, et al. Polymorphisms in DNA repair and apoptosis-related genes and clinical outcomes of patients with non-small cell lung cancer treated with first-line paclitaxel-cisplatin chemotherapy. *Lung Cancer*. 2013;82(2):330-339. doi:[10.1016/j.lungcan.2013.07.024](https://doi.org/10.1016/j.lungcan.2013.07.024)
 64. Simader E, Beer L, Laggner M, et al. Tissue-regenerative potential of the secretome of gamma-irradiated peripheral blood mononuclear cells is mediated via TNFRSF1B-induced necroptosis. *Cell Death Dis*. 2019;10(10):729. doi:[10.1038/s41419-019-1974-6](https://doi.org/10.1038/s41419-019-1974-6)
 65. Prakash YS, Martin RJ. Brain-derived neurotrophic factor in the airways. *Pharmacol Ther*. 2014;143(1):74-86. doi:[10.1016/j.pharmthera.2014.02.006](https://doi.org/10.1016/j.pharmthera.2014.02.006)
 66. Shi X, Chen S, Zhang Y, et al. Norcantharidin inhibits the DDR of bladder cancer stem-like cells through cdc6 degradation. *Oncotargets Ther*. 2019;12:4403-4413. doi:[10.2147/OTT.S209907](https://doi.org/10.2147/OTT.S209907)
 67. Li X, Song Q, Guo X, et al. the metastasis potential promoting capacity of cancer-associated fibroblasts was attenuated by cisplatin via modulating KRT8. *Oncotargets Ther*. 2020;13:2711-2723. doi:[10.2147/OTT.S246235](https://doi.org/10.2147/OTT.S246235)
 68. Travis WD, Brambilla E, Burke AP, Marx A, Nicholson AG. Introduction to The 2015 World Health Organization classification of tumors of the lung, pleura, thymus, and heart. *J Thorac Oncol*. 2015;10(9):1240-1242. doi:[10.1097/JTO.0000000000000663](https://doi.org/10.1097/JTO.0000000000000663)
 69. Liang W, Zhao Y, Huang W, et al. Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). *Theranostics*. 2019;9(7):2056-2070. doi:[10.7150/thno.28119](https://doi.org/10.7150/thno.28119)
 70. Liang W, Chen Z, Li C, et al. Accurate diagnosis of pulmonary nodules using a noninvasive DNA methylation test. *J Clin Invest*. 2021;131(10):e145973. doi:[10.1172/JCI145973](https://doi.org/10.1172/JCI145973)
 71. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-595. doi:[10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698)
 72. Tapial J, Ha KCH, Sterne-Weiler T, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res*. 2017;27(10):1759-1768. doi:[10.1101/gr.220962.117](https://doi.org/10.1101/gr.220962.117)
 73. Vineis P, Airolidi L, Veglia F, et al. Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ*. 2005;330(7486):277. doi:[10.1136/bmj.38327.648472.82](https://doi.org/10.1136/bmj.38327.648472.82)
 74. Mossman BT, Lippmann M, Hesterberg TW, Kelsey KT, Barchowsky A, Bonner JC. Pulmonary endpoints (lung carcinomas and asbestosis) following inhalation exposure to asbestos. *J Toxicol Environ Health B Crit Rev*. 2011;14(1-4):76-121. doi:[10.1080/10937404.2011.556047](https://doi.org/10.1080/10937404.2011.556047)
 75. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi:[10.1089/omi.2011.0118](https://doi.org/10.1089/omi.2011.0118)
 76. Oken MM, Hocking WG, Kvale PA, et al. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA*. 2011;306(17):1865-1873. doi:[10.1001/jama.2011.1591](https://doi.org/10.1001/jama.2011.1591)
 77. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:[10.1093/bioinformatics/bty560](https://doi.org/10.1093/bioinformatics/bty560)

78. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27(11):1571-1572. doi:[10.1093/bioinformatics/btr167](https://doi.org/10.1093/bioinformatics/btr167)
79. Feng H, Wu H. Differential methylation analysis for bisulfite sequencing using DSS. *Quant Biol*. 2019;7(4):327-334. doi:[10.1007/s40484-019-0183-8](https://doi.org/10.1007/s40484-019-0183-8)
80. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:[10.1093/nar/gkq603](https://doi.org/10.1093/nar/gkq603)
81. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 2018;28(11):1747-1756. doi:[10.1101/gr.239244.118](https://doi.org/10.1101/gr.239244.118)
82. Bai Y, Yang C, Wu R, et al. YTHDF1 regulates tumorigenicity and cancer stem cell-like activity in human colorectal carcinoma. *Front Oncol*. 2019;9:332. doi:[10.3389/fonc.2019.00332](https://doi.org/10.3389/fonc.2019.00332)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liang P, Peng M, Tao J, et al. Development of a genome atlas for discriminating benign, preinvasive, and invasive lung nodules. *MedComm*. 2024;5:e644. <https://doi.org/10.1002/mco2.644>