Journal of Translational Medicine

RESEARCH Open Access

Multidimensional bioinformatics perspective on smoking-linked driver genes and immune regulatory mechanisms in non-small cell lung cancer

Can Ouyang[1,2], Xiaopeng Yu[2], Huazhong Wang[1] and Puhua Zeng[1,3*]

Abstract

Background Lung cancer, one of the leading causes of cancer-related morbidity and mortality worldwide, is strongly associated with smoking as its primary carcinogenic factor. However, despite the strong link between smoking and lung cancer, not all smokers develop the disease, suggesting that individual genetic susceptibility and molecular mechanisms may play a critical role in the onset of lung cancer. Understanding the gene-driving mechanisms and immune regulatory pathways involved in smoking-related lung cancer remains one of the key challenges in current lung cancer research.

Methods This study employs an integrative bioinformatics approach to explore gene expression differences and immune microenvironment characteristics between smokers with non-small cell lung cancer (NSCLC) and normal individuals. First, smoking-linked lung cancer driver genes (SLDCGs) were identified, followed by Mendelian Randomization (MR) and Summary-based Mendelian Randomization (SMR) analyses to further validate their causal relationships. Next, public databases, including TCGA, GEO, and GTEx, were used to systematically analyze the expression differences of SLDCGs across various clinical subgroups, and immune infiltration analysis was conducted to explore their potential roles in the immune microenvironment of NSCLC.

Results The study identified HLA-J and PRMT7 as core driver genes for smoking-associated NSCLC. MR analysis confirmed the potential causal relationship of HLA-J and PRMT7 in the development of NSCLC. Specifically, high expression of PRMT7 was closely associated with the occurrence of NSCLC, while low expression of HLA-J was implicated in immune evasion mechanisms in NSCLC. Additionally, immune microenvironment analysis revealed that HLA-J enhances the activity of immune cells, particularly T cells, to promote tumor immune recognition, whereas PRMT7 suppresses immune cell function, weakening immune surveillance and facilitating immune evasion.

Conclusion This study systematically reveals the molecular mechanisms of smoking-linked NSCLC through multidimensional bioinformatics analysis, highlighting the key roles of SLDCGs in immune evasion. The discovery of HLA-J and PRMT7 provides new theoretical foundations for targeted immunotherapy, with significant potential for

*Correspondence:
Puhua Zeng
zph120@126.com

Full list of author information is available at the end of the article

early diagnosis and personalized treatment of smoking-induced NSCLC. Future research should focus on validating these genes in clinical samples and exploring their potential in immunotherapy.

Keywords Non-Small cell lung Cancer, Smoking-Linked lung Cancer driver genes, Immune microenvironment, Mendelian randomization, Bioinformatics

Background

Lung cancer remains the leading cause of cancer-related incidence and mortality worldwide, posing a significant threat to human health [1]. Smoking, recognized as the primary risk factor for lung cancer, has been thoroughly confirmed as a major carcinogenic influence [2, 3]. However, clinical observations reveal that not all long-term smokers develop lung cancer. A subset of individuals exposed to tobacco carcinogens over extended periods does not manifest the disease [4]. This suggests that lung cancer development is not solely attributed to external environmental factors, like smoking but is also significantly influenced by individual genetic backgrounds and molecular regulatory mechanisms [5, 6]. Investigating the genetic differences between smokers who develop lung cancer and those who do not, and identifying potential molecular driving mechanisms, is a crucial pathway to better understanding the etiology of lung cancer.

Most existing research focuses on comparing the gene expression profiles of lung cancer patients to those of non-cancer individuals, aiming to uncover potential pathogenic mechanisms through differential expression analyses [7]. While these studies provide foundational knowledge of lung cancer-related molecules in the general population, they fall short in elucidating the role of smoking as a specific environmental factor, especially regarding the genetic characteristics distinguishing smokers with lung cancer from those without. Substantial evidence indicates that tobacco carcinogens (e.g., benzo[a]pyrene and nitrosamines) contribute to lung cancer development by inducing DNA damage and inflammatory responses [8, 9]. However, individual variability in sensitivity to tobacco carcinogens implies that genetic background may regulate the carcinogenic effects of smoking through complex molecular mechanisms.

With the rapid advancement of high-throughput sequencing technologies, databases such as TCGA, GEO, and GTEx have laid a robust foundation for the systematic analysis of gene expression patterns and clinical features in large-scale samples [10]. Utilizing these resources enables the identification of differentially expressed genes (DEGs) across various subgroups within the smoking population and the exploration of their potential pathogenic mechanisms. This study focuses on NSCLC in smokers to investigate SLDCGs. By comprehensively analyzing multidimensional data, this research delves into the differences in gene expression between smokers with NSCLC and healthy individuals, alongside their clinical characteristics. Additionally, the MR method and bioinformatics tools are employed to analyze SLDCGs from multiple perspectives, verifying their roles in NSCLC development. This provides new evidence for understanding the complex relationship between smoking and NSCLC and offers a scientific basis for the precise diagnosis and treatment of smoking-induced NSCLC.

Methods

The research methodology for this study is outlined in Fig. 1.

Identification of SLDCGs

To identify SLDCGs, we performed differential expression analysis of gene expression data between NSCLC patients and non-cancer smokers within the smoking population and selected differentially expressed genes (DEGs). Several methods were employed for validation, including MR analysis, SMR, and validation using the TCGA dataset. The following selection criteria were applied to the DEGs:

1) Initially, differentially expressed genes were identified.
2) In the MR analysis, the inverse variance weighted method (IVW) was used to assess the causal relationship between genes and smoking-related NSCLC. Genes showing significant causal effects and consistent differential expression direction were selected (i.e., upregulated DEGs with OR > 1, and downregulated DEGs with OR < 1).
3) In the SMR analysis, further selection of DEGs was conducted based on consistency with MR results and significant causal relationships.
4) Differential expression analysis of the selected DEGs was conducted using TCGA data, ensuring that the expression direction in GEO data was consistent with that in TCGA.

Acquisition of DEGs

We downloaded the GSE19027 dataset from the GEO database (https://www.ncbi.nlm.nih.gov/geo/), which includes gene expression data from smokers with and without NSCLC. The dataset consists of 21 smoking NSCLC samples, 30 smoking non-lung cancer samples, and 9 non-smoker samples. Differential expression analysis was performed using the DESeq2 package in R.

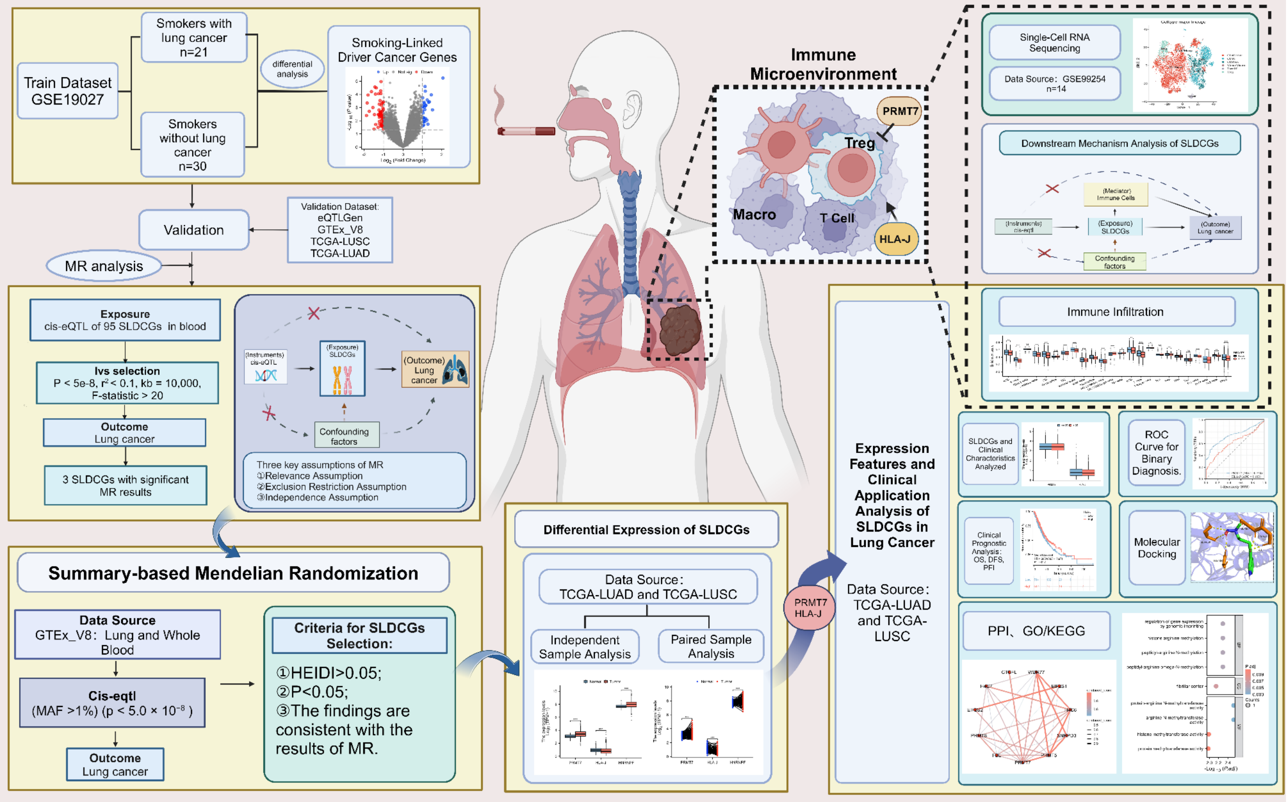

Train Dataset GSE19027
Smokers with lung cancer n=21
Smokers without lung cancer n=30
differential analysis
Smoking-Linked Driver Cancer Genes
Validation
Validation Dataset: eQTLGen GTEx_V8 TCGA-LUSC TCGA-LUAD
MR analysis
Exposure cis-eQTL of 95 SLDCGs in blood
Ivs selection P < 5e-8, r² < 0.1, kb = 10,000, F-statistic > 20
Outcome Lung cancer
3 SLDCGs with significant MR results
(Instruments) cis-eQTL
(Exposure) SLDCGs
(Outcome) Lung cancer
Confounding factors
Three key assumptions of MR
①Relevance Assumption
②Exclusion Restriction Assumption
③Independence Assumption
Summary-based Mendelian Randomization
Data Source GTEx_V8: Lung and Whole Blood
Cis-eqtl (MAF >1%) (p < 5.0 × 10⁻⁸)
Outcome Lung cancer
Criteria for SLDCGs Selection:
①HEIDI>0.05;
②P<0.05;
③The findings are consistent with the results of MR.
Immune Microenvironment
PRMT7
Treg
Macro
T Cell
HLA-J
Differential Expression of SLDCGs
Data Source: TCGA-LUAD and TCGA-LUSC
Independent Sample Analysis
Paired Sample Analysis
PRMT7 HLA-J
Expression Features and Clinical Application Analysis of SLDCGs in Lung Cancer
Data Source: TCGA-LUAD and TCGA-LUSC
Single-Cell RNA Sequencing
Data Source: GSE99254 n=14
Downstream Mechanism Analysis of SLDCGs
(Mediator) Immune Cells
(Instruments) cis-eqtl
(Exposure) SLDCGs
(Outcome) Lung cancer
Confounding factors
Immune Infiltration
SLDCGs and Clinical Characteristics Analyzed
ROC Curve for Binary Diagnosis.
Clinical Prognostic Analysis: OS, DFS, PFI
Molecular Docking
PPI、GO/KEGG

Fig. 1 Research methodology flowchart

First, we analyzed the gene expression differences between smoking NSCLC tissue (experimental group) and smoking normal tissue (control group). Next, we excluded the differentially expressed genes (DEGs) between smoking NSCLC samples (experimental group) and non-smoking normal tissue (control group). The criteria for selecting DEGs were |log2FC| > 1 and a corrected p-value < 0. 05. Finally, we identified the DEGs between NSCLC patients and non-cancer smokers in the smoking population.

MR analysis of DEGs

We conducted a two-sample MR analysis to investigate the relationship between DEGs (exposure) and NSCLC(outcome). To ensure the reliability and accuracy of the results, the MR analysis must satisfy three key assumptions:

1) The instrument variables must be strongly correlated with the exposure, ensuring that they can effectively reflect changes in exposure and thus infer causal relationships.
2) There should be no confounding factors between the instrument variables and the outcome, ensuring that causal inference is not affected by unobserved factors.
3) The instrument variables should only influence the outcome through the exposure, satisfying the "exclusion restriction assumption" to exclude potential pleiotropic effects.

Exposure data sources This study utilized data provided by the eQTLGen Consortium [11], a large-scale collaborative project designed to identify and analyze cis-expression quantitative trait loci (cis-eQTLs) by integrating whole-genome expression data from multiple European blood samples. These data, derived from joint analyses of several cohorts, have a large sample size, enhancing statistical power and the reliability of the results. We obtained the latest DEGs cis-eQTL data from the eQTLGen Consortium (https://eqtlgen.org/cis-eqtls. html), which prima rily covers gene expression and genetic variation associations in blood samples.

Outcome data sources For the outcome data of NSCLC, a genome-wide association study (GWAS) was conducted using a generalized linear mixed model to analyze NSCLC (including bronchial and NSCLCs). The study was based on 2, 120 NSCLC cases and 454, 228 control samples of European ancestry, with a total sample size of 456, 348 individuals, primarily from the UK. The study employed genome-wide genotyping arrays and genotype imputation techniques to identify genetic loci significantly associated with NSCLC.

Data selection and MR analysis In this study, we strictly followed the following criteria to select DEGs cis-eQTLs as instrument variables for the MR analysis:

1) Selection of SNPs: SNPs (cis-eQTLs) significantly associated with DEGs expression were chosen, with p-values below the genome-wide significance threshold ($p < 5.\ 0 \times 10^{-8}$), ensuring that the selected SNPs had sufficient statistical significance in their association with gene expression, thus reducing the risk of false positives.
2) Linkage Disequilibrium (LD) Control: To minimize the impact of LD on multiple correlations, the LD threshold was set at $r^2 < 0.\ 1$, retaining only those SNPs that did not have strong correlations with the target SNP, thereby avoiding confounding due to LD.
3) Clustering Window: The clustering window was set to 10, 000 kb to ensure that the selected SNPs were physically close to the target gene, thus increasing the credibility of their role as regulatory SNPs [12].
4) F-Statistic Selection: To prevent "weak instrument bias, " only SNPs with an F-statistic greater than 20 were retained as instrument variables. A higher F-statistic indicates a stronger association between the SNP and the exposure, improving the reliability of causal inference [13].

In this study, we employed the IVW method [14]as the primary analytical approach to assess the causal relationship between the exposure and outcome. Additionally, four supplementary methods were used: MR-Egger regression [15], weighted median method, weighted mode, and simple mode [16]. In the primary analysis, we calculated the Wald ratio estimate for each genetic variant and assessed potential heterogeneity using Cochran's Q test. If the p-value > 0. 05. we assumed no heterogeneity between SNPs and used the fixed-effect IVW model. If the p-value < 0. 05, we assumed heterogeneity and used the random-effect IVW model [17]. Moreover, a leave-one-out analysis was conducted to evaluate the impact of removing each SNP on the results, testing the sensitivity of the findings. To further assess pleiotropy. The intercept from MR-Egger regression and potential pleiotropy were assessed using the MR-Presso method, which evaluates residuals and outliers. When the P-value is greater than 0.05, it indicates the absence of pleiotropy.

SMR analysis of DEGs

Building on the significant results from the MR analysis, this study used the SMR method to further validate the reliability of DEGs as SLDCGs. The data for the SMR analysis were sourced from the 8th version of the GTEx project (GTEx_V8). The GTEx project aims to reveal how genomic variations influence gene expression by analyzing gene expression across various human tissues. The GTEx_V8 dataset includes gene expression data from 54 different tissues across nearly 1, 000 donors and is one of the most comprehensive cross-tissue gene expression resources available. This study specifically focused on the eQTL data from blood samples and lung tissues in GTEx_V8 to investigate the relationship between gene expression and genetic variation in these tissues.

In the analysis, we selected SNPs(p-value $< 5.0 \times 10^{-8}$, MAF > 1%) significantly associated with the expression of the target genes as instrument variables to infer the causal relationship between gene expression and disease. The HEIDI test was used to compare the contribution of different SNP instrument variables to the association between target gene expression and phenotypes (such as disease), evaluating whether these SNPs share the same causal effect. If the HEIDI test shows significant heterogeneity (p-value < 0. 01), it suggests that the signal may be influenced by pleiotropy.

DEGs expression differential analysis and identification of SLDCGs

In this study, RNA-seq data processed using the STAR pipeline were downloaded and organized from the TCGA database (https://portal.gdc.cancer. gov) for the TCGA-LUAD and TCGA-LUSC projects, and TPM format data were extracted for analysis. The study performed separate analyses for non-paired and paired samples.

1) Non-paired sample analysis: This analysis compared gene expression differences between tumor tissues from NSCLC patients and normal tissues from healthy control groups. Independent T-tests were used when the assumptions of normality and homogeneity of variance were met. If normality was satisfied but homogeneity of variance was not, Welch's T-test was applied. When normality was not met, the Wilcoxon rank-sum test was used.
2) Paired sample analysis: This analysis compared gene expression differences between tumor tissues and adjacent non-tumor tissues from the same patient. A paired T-test was used when both normality and homogeneity of variance assumptions were met. If normality was satisfied but homogeneity of variance was not, a paired Welch's T-test was applied. When normality was not met, the Wilcoxon signed-rank test was used. Finally, the DEGs that met the selection criteria were identified as SLDCGs.

Expression characteristics and clinical application analysis of SLDCGs in NSCLC

Correlation analysis of SLDCGs expression with different populations and clinical features

To further investigate the expression specificity of SLDCGs in different populations and their pathogenic relevance across various cancer types, we analyzed the correlation between SLDCGs expression and multiple clinical features using the TCGA-LUAD and TCGA-LUSC datasets. The clinical variables included smoking status (smokers: 95 cases; non-smokers: 928 cases), years of smoking (≥ 40 years: 476 cases; <40 years: 325 cases), tumor type (squamous cell carcinoma: 505 cases; adenocarcinoma: 544 cases), age (> 65 years: 572 cases; ≤65 years: 449 cases), sex (male: 626 cases; female: 423 cases), and race (Asian: 17 cases; Black or African American: 87 cases; White: 764 cases). When the variables met the assumptions of normality and homogeneity of variance, an independent T-test was used. For variables meeting normality but failing the homogeneity of variance assumption, Welch's T-test was applied. Non-normally distributed variables were analyzed using the Wilcoxon rank-sum test.

Clinical prognosis analysis

To assess the potential impact of SLDCGs on patient prognosis, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) data were extracted from the TCGA-LUAD and TCGA-LUSC datasets to standardize gene expression levels. Additionally, detailed clinical information and prognosis data for the patients, including overall survival (OS), disease-free survival (DFS), and progression-free interval (PFI), were collected. Statistical analysis was performed using the survival package in R, and the Cox proportional hazards model was used to evaluate the impact of SLDCGs on patient prognosis.

Binary classification diagnosis

The potential diagnostic value of SLDCGs in NSCLC was assessed by constructing receiver operating characteristic (ROC) curves based on the TCGA-LUAD and TCGA-LUSC datasets and calculating the area under the ROC curve (AUC). The AUC value ranges from 0. 5 to 1, with an AUC closer to 1 indicating a higher diagnostic efficiency in distinguishing the normal group from the tumor group. The analysis was performed using the pROC package in R.

Enrichment analysis and Protein-Protein interaction network construction

The potential biological functions of SLDCGs in NSCLC were explored through functional enrichment analysis using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). The GO analysis covered three aspects: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC), to systematically reveal the molecular-level functions of SLDCGs and the key biological processes they are involved in. The KEGG pathway analysis focused on the enrichment of SLDCGs in known metabolic and signaling pathways, providing support for their potential mechanisms. Additionally, a Protein-Protein Interaction (PPI) network was constructed to further elucidate the interaction patterns of proteins encoded by SLDCGs. The PPI network was generated using the STRING database (https://string-db.org), with a confidence score threshold set at 0. 5 as the minimum interaction threshold, and all other parameters set to default [18], This network serves as a candidate resource for subsequent functional validation.

Molecular docking

This study utilized molecular docking simulation methods to investigate the interactions between several tobacco carcinogens, including 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone, 4-Aminobiphenyl, Acrolein, and N'-Nitrosonornicotine, with PRMT7 and HLA-J proteins. The molecular structures of all carcinogen compounds were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov/), and the protein structures of SLDCGs were retrieved from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). Molecular docking simulations were performed using AutoDockTools-1.5.7 software to calculate the binding modes, binding free energies (ΔG), and potential functional impacts of these compounds with the target proteins.

SLDCGs in NSCLC immune microenvironment and immune downstream mechanisms: mediation MR analysis

SLDCGs immune infiltration analysis

Based on RNA sequencing data from the TCGA datasets for LUAD and LUSC projects, we extracted gene expression data in TPM format and conducted an in-depth analysis of the immune microenvironment using immune infiltration algorithms. Immune infiltration scores were calculated using the Single-sample Gene Set Enrichment Analysis (ssGSEA) algorithm, as proposed by Hänzelmann et al. [19]. implemented through the GSVA R package. The analysis was based on 24 immune cell marker gene sets defined by Bindea et al. [20]. The immune cell infiltration scores were calculated for each sample. Correlation analysis was performed using Spearman's rank correlation test, associating the immune infiltration score matrix with SLDCGs gene expression data. This analysis aimed to assess the potential regulatory role of differentially expressed genes in immune cell infiltration. To ensure the reliability of the results, multiple testing correction was applied to the significantly correlated results.

Expression and annotation of SLDCGs in single-cell subpopulations

This study analyzed single-cell RNA sequencing data from 14 untreated NSCLC patients in the GSE99254 dataset from the GEO database. The dataset contains a total of 12, 346 T cells from tumor, adjacent normal tissue, and peripheral blood samples. Data preprocessing was performed using the Seurat package, which involved filtering low-quality cells and normalizing the data. Principal component analysis (PCA) was then used to extract the primary features, followed by t-SNE dimensionality reduction for cell visualization and clustering. The T cells were categorized into different subpopulations. The expression differences of SLDCGs across different cell subpopulations were analyzed using the Wilcoxon rank-sum test, and the mean expression distribution of genes within each subpopulation was visualized using bar plots.

Mediation MR analysis of SLDCGs influencing immune downstream mechanisms in NSCLC

Multivariable MR analysis was employed to assess the potential mediating effect of immune cells in the relationship between SLDCGs and NSCLC Specifically, the IVW method was used as the primary analysis tool to estimate the impact of SLDCGs on immune cells (β_1). A multivariable MR model was then applied to assess the role of each immune cell in modulating NSCLC risk, while adjusting for the genetic effects of SLDCGs (β_2).

To calculate the indirect mediation effect of SLDCGs on NSCLC, the product of coefficients method was used. This approach evaluated the causal effect of SLDCGs on NSCLC through immune cells as intermediaries ($\beta_1 \times \beta_2$). The direct effect (β_3) represents the total effect of SLDCGs on NSCLC, while the indirect effect quantifies the influence of immune cells on this relationship. The total effect is the sum of the direct and indirect effects ($\beta_3 + \beta_1 \times \beta_2$). Additionally, the ratio of the indirect effect to the total effect ($[\beta_1 \times \beta_2 / (\beta_3 + \beta_1 \times \beta_2)]$) was calculated to quantify the mediating contribution of each immune cell in the relationship between SLDCGs and NSCLC. This ratio reflects the immune cells' mediation effect in the connection between SLDCGs and NSCLC.

The exposure data for 731 immune cells were sourced from GWAS databases. Tool variables were selected based on $P<5\times10^{-8}$, $r^2 < 0.01$, and a distance window set at 10, 000 kb, ensuring the reliability and validity of the selected instrumental variables. By combining multivariable Mendelian Randomization methods, this study deeply explored the regulatory role of SLDCGs in the immune environment and their involvement in the development and progression of NSCLC.

Results

Identification of SLDCGs

Acquisition of DEGs

Using the selection criteria of |log2FC| > 1 and corrected $P<0.05$, a total of 95 differentially expressed genes (DEGs) were initially identified. Among these, 30 genes were upregulated and 65 genes were downregulated in NSCLC Volcano plots (Fig. 2a), differential ranking plots (Fig. 2b), and heatmaps (Fig. 2d) were generated to visualize these results. A Venn diagram (Fig. 2c) displayed the overlap of DEGs between smoking NSCLC tissues and smoking normal tissues, excluding 13 DEGs from the comparison between NSCLC and non-smoking normal tissues. Detailed results of the differential analysis are provided in Supplementary Table 1.

MR analysis of DEGs

A MR analysis was performed on the 82 significantly differentially expressed genes identified in the initial screening. The results showed that 41 of the DEGs did not have valid instrument variables. Among the remaining 41 DEGs, 3 genes were selected for further analysis based on the screening criteria: HLA-J (OR = 0. 889, 95% CI = 0. 792–0. 997, $P=0.045$) was identified as a protective factor with decreased expression in NSCLC tissues; HNRNPF (OR = 0. 791, 95% CI = 0. 655–0. 955, $P=0.015$) was also identified as a protective factor with decreased expression in NSCLC tissues; PRMT7 (IVW OR = 1. 326, 95% CI = 1. 152–1. 527, $P<0.001$) was identified as a risk factor with increased expression in NSCLC tissues. Figure 2e presents these results, and detailed MR, heterogeneity, and horizontal pleiotropy results are provided in Supplementary Table 2.

SMR analysis of DEGs

The SMR results indicated that the expression of HLA-J and PRMT7 in lung and blood tissues was significantly associated with NSCLC susceptibility, and the direction of these results was consistent with those from the MR analysis. Additionally, the HEIDI test did not detect significant heterogeneity (detailed information in Supplementary Table 3). However, no relevant data for the HNRNPF gene were found in either lung or blood tissue.

DEGs differential expression analysis and identification of SLDCGs

The expression analysis of HLA-J, PRMT7, and HNRNPF showed that the expression of HLA-J was significantly lower in NSCLC tissues compared to normal tissues ($P<0.001$), while PRMT7 expression was significantly higher in NSCLC tissues ($P<0.001$). Additionally, HNRNPF expression was also significantly higher in NSCLC tissues ($P<0.001$). Figure 2f visually presents the median expression differences of these three genes

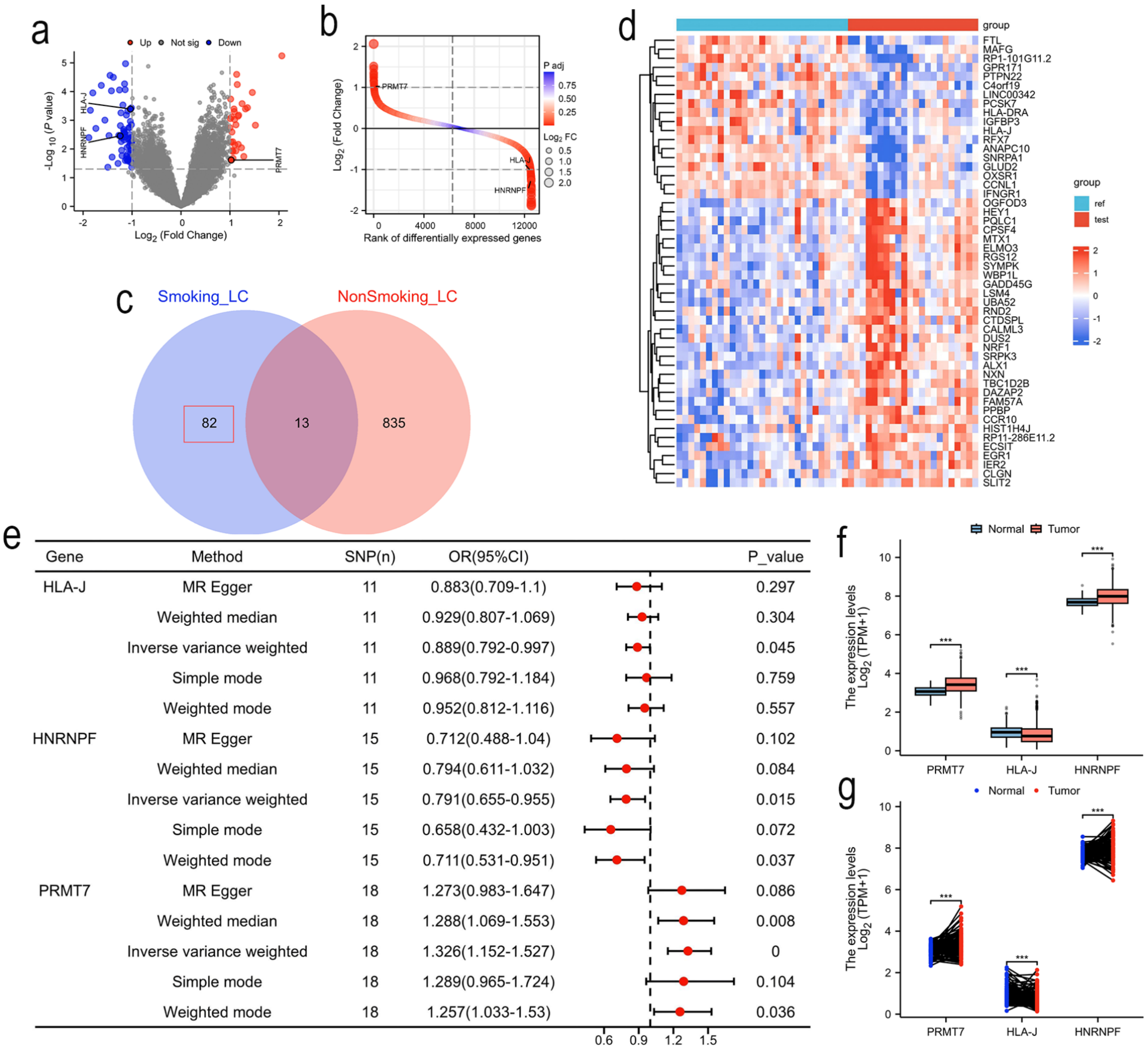


Fig. 2 Overview of SLDCGs Screening Results in Smokers with NSCLC and Smokers with Normal Lung Tissues. (**a**) Volcano Plot displaying the magnitude of gene expression changes (Log2 Fold Change) against significance levels (-log10 p-value), where red and blue represent upregulated and downregulated genes, respectively. (**b**) Ranking Plot sorted by gene significance, with bubble color representing the corrected p-value (from blue to red, decreasing significance), and bubble size reflecting the fold change in expression. (**c**) Venn Diagram illustrating the differentially expressed genes (DEGs) between smoking NSCLC tissues (Smoking_LC) and smoking normal tissues, as well as between NSCLC tissues and non-smoking normal tissues (Nonsmoking_LC). (**d**) Heatmap displaying the expression patterns of significantly differentially expressed genes across the two sample groups, with color coding for expression levels—red for high expression and blue for low expression. Clustering analysis clearly separates the NSCLC group from the control group. (**e**) MR Results for the genes HLA-J, PRMT7, and HNRNPF, with the x-axis representing Odds Ratios (OR) and the vertical line representing the null effect line (OR = 1). Red dots represent the OR values for each analysis method, with the horizontal line indicating the 95% confidence interval. (**f**) Box Plot of gene expression in non-paired samples, showing the median gene expression levels and inter-group differences. (**g**) Scatter Plot of gene expression in paired samples, where each tumor tissue sample (red dot) is connected to its corresponding normal tissue sample (blue dot), further clarifying the significant intra-individual gene expression differences

between the two groups, and Fig. 2g shows the expression trends in paired samples, both demonstrating statistically significant differences. While HNRNPF expression was significantly higher in NSCLC tissues compared to normal tissues, its directional findings were inconsistent with those from the MR analysis and differential expression analysis. As a result, HNRNPF was excluded from subsequent SLDCGs screening. In contrast, the expression patterns of PRMT7 and HLA-J further support their potential roles in NSCLC development, reinforcing the reliability of HLA-J and PRMT7 as potential NSCLC susceptibility genes. Detailed statistical descriptors are provided in Supplementary Table 4.

Expression characteristics and clinical application analysis of SLDCGs in NSCLC

Correlation analysis of SLDCGs expression with different populations and clinical features

The results of the correlation analysis between SLDCGs and clinical features showed that the expression of HLA-J was significantly higher in female patients compared to male patients (*P*<0. 001). Additionally, low expression of HLA-J was significantly associated with smokers and individuals with a longer smoking duration (≥40 years) (*P*<0. 001, *P*<0. 05). Regarding tumor types, high expression of PRMT7 was significantly associated with squamous cell carcinoma (*P*<0. 01), while high expression of HLA-J was significantly associated with adenocarcinoma (*P*<0. 001). However, no significant expression differences of SLDCGs were observed across racial or age groups (Fig. 3). Detailed results are provided in Supplementary Table 5.

Clinical prognosis analysis

Survival prognosis results indicated that high expression of HLA-J significantly prolonged patients' overall survival (OS) (HR=0. 80, 95% CI: 0. 65–0. 97, *P*=0. 027). Specifically, the median OS for patients with low expression of HLA-J was 48. 7 months (95% CI: 40. 3–56. 9), while the median OS for patients with high expression of HLA-J was 56. 6 months (95% CI: 48. 2–85. 8). These findings suggest that high expression of HLA-J may serve as a protective factor for better prognosis in NSCLC patients. Detailed results are shown in Fig. 4.

Enrichment analysis, Protein-Protein interaction network construction, and binary classification diagnosis

GO functional enrichment analysis (Fig. 5a) further revealed the significant biological functions of the differentially expressed genes. SLDCGs were significantly involved in key processes such as histone methylation, arginine methyltransferase activity, and genomic imprinting regulation. These functional enrichment results suggest that the differentially expressed genes may regulate NSCLC-related gene expression and cellular functions through epigenetic pathways. Detailed results are provided in Supplementary Table 6.

In the PPI network analysis (Fig. 5b), PRMT7 formed a dense interaction network with several epigenetically related proteins, such as PRMT5 and FBL, indicating that PRMT7 may play a role in NSCLC development and progression by regulating epigenetic pathways, including histone modification and RNA methylation. Detailed results are provided in Supplementary Table 7. HLA-J did not show significant regulatory protein associations in the PPI network.

In the diagnostic ROC curve analysis, the AUC values for PRMT7 and HLA-J were 0. 740 and 0. 789,

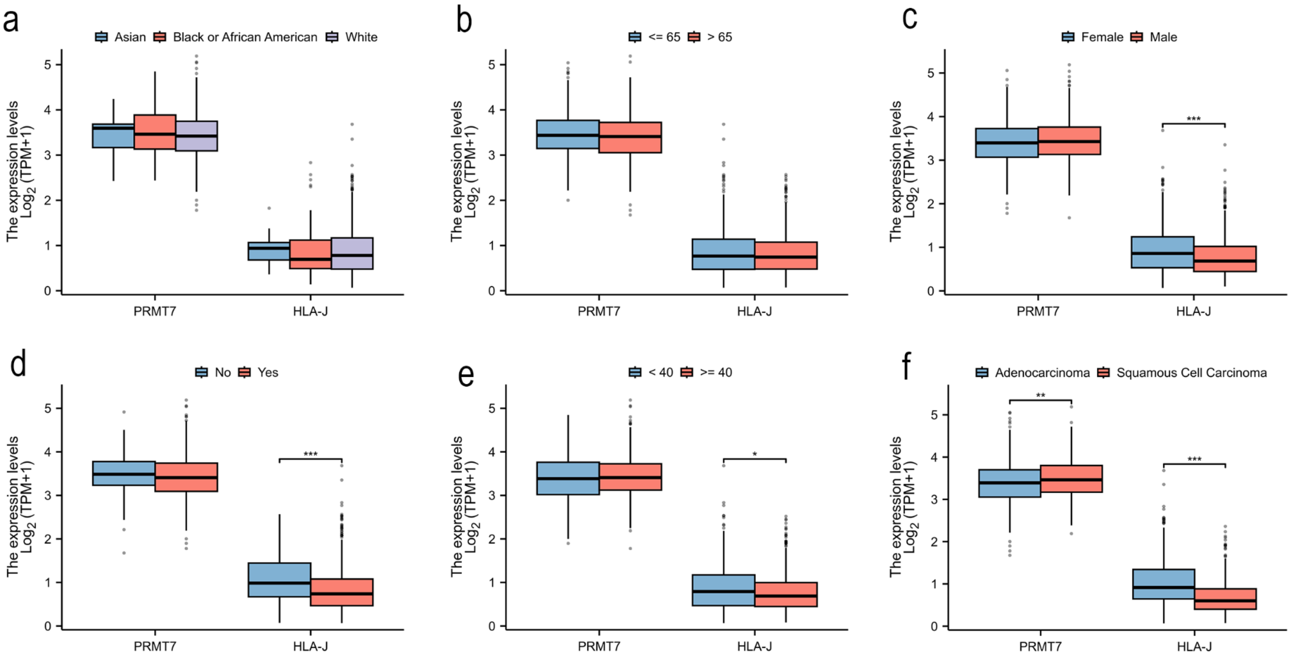


Fig. 3 Correlation Analysis of HLA-J and PRMT7 Gene Expression with Clinical Features in TCGA-LUAD and TCGA-LUSC Datasets. (**a**) Expression levels of HLA-J and PRMT7 genes across different racial groups. (**b**) Expression levels of HLA-J and PRMT7 genes across different age groups. (**c**) Expression levels of HLA-J and PRMT7 genes across different sexes. (**d**) Correlation between HLA-J and PRMT7 gene expression and smoking status. (**e**) Correlation between HLA-J and PRMT7 gene expression and years of smoking. (**f**) Expression levels of HLA-J and PRMT7 genes in different NSCLC types

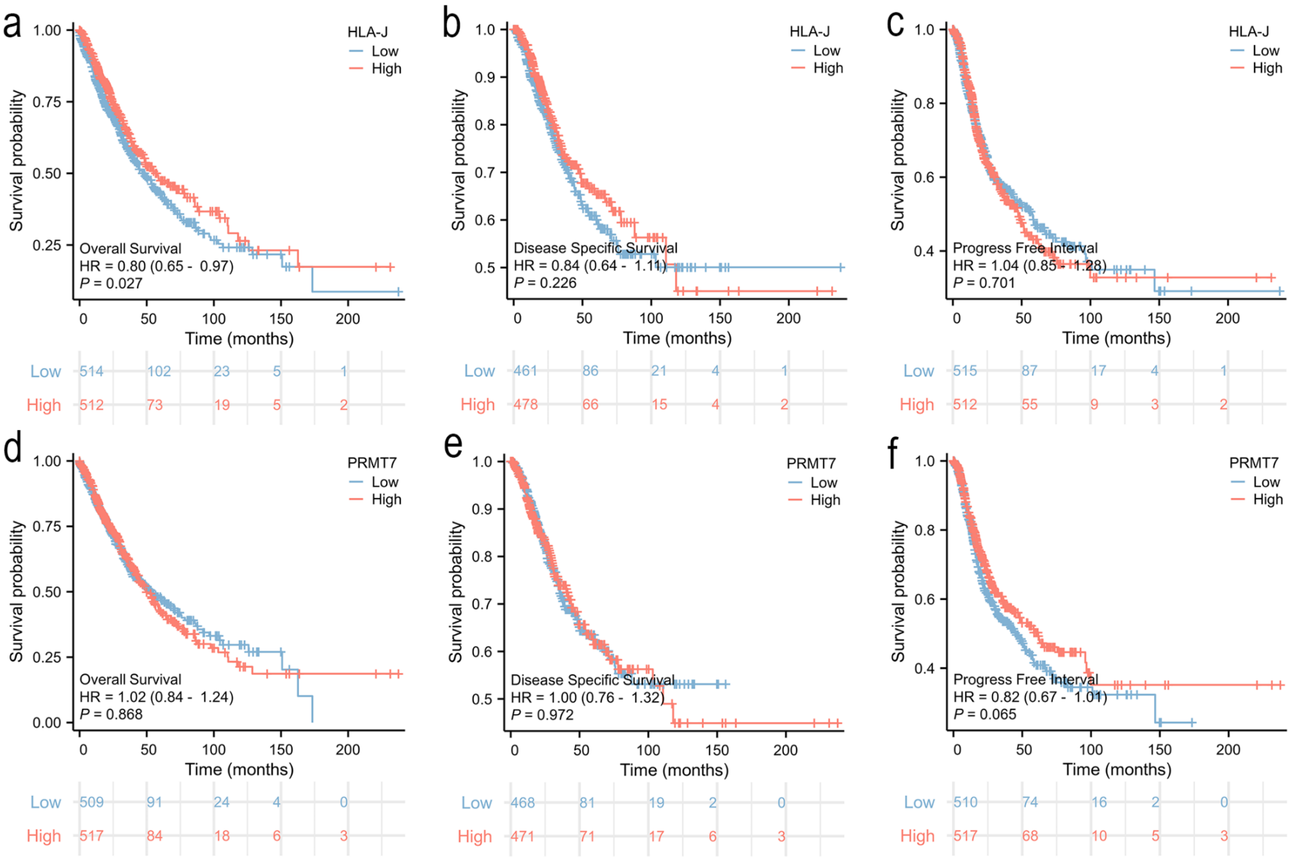


Fig. 4 The Role of SLDCGs Expression Levels in Survival Analysis of NSCLC Patients. (**a**) Relationship between HLA-J gene expression and overall survival (OS). (**b**) Relationship between HLA-J gene expression and disease-specific survival (DSS). (**c**) Relationship between HLA-J gene expression and progression-free interval (PFI). (**d**) Relationship between PRMT7 gene expression and OS. (**e**) Relationship between PRMT7 gene expression and DSS. (**f**) Relationship between PRMT7 gene expression and PFI

respectively, indicating that both genes have strong predictive power for NSCLC diagnosis (Fig. 5c).

Molecular Docking

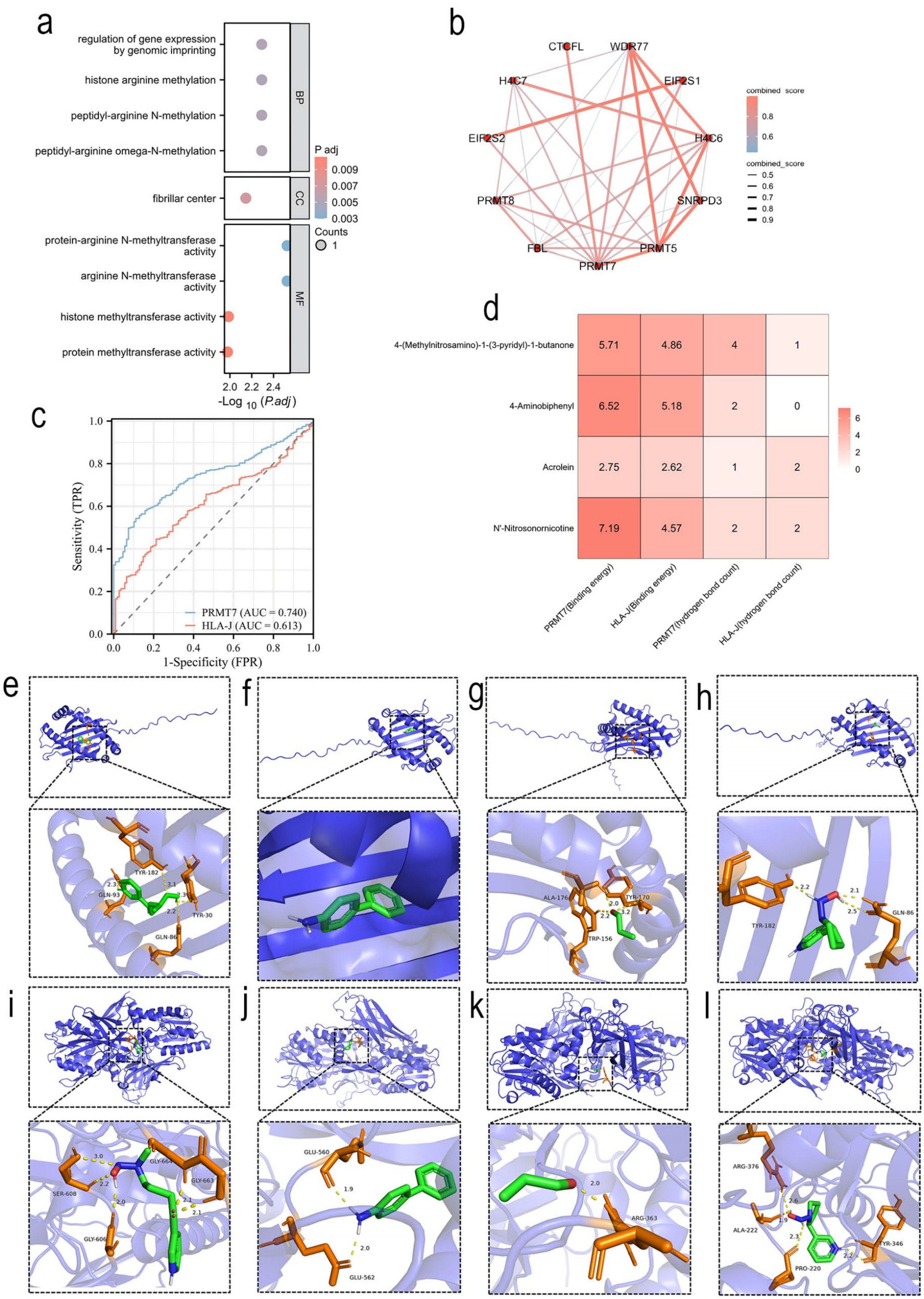
The molecular docking simulation results showed that the binding free energy of 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone with the PRMT7 protein was ΔG = -5. 71 kcal/mol, 4-Aminobiphenyl with PRMT7 was ΔG = -6. 52 kcal/mol, and N'-Nitrosonornicotine with PRMT7 was ΔG = -7. 19 kcal/mol. Similarly, the binding free energy of 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone with HLA-J was ΔG = -4. 86 kcal/mol, 4-Aminobiphenyl with HLA-J was ΔG = -5. 18 kcal/mol, and N'-Nitrosonornicotine with HLA-J was ΔG = -4. 57 kcal/mol. These results suggest that tobacco carcinogenic compounds interact with key residues of SLDCGs. Detailed values of binding energy and the number of hydrogen bonds for the molecular docking are presented in Fig. 5d, with docking visualization shown in Fig. 5e and l.

SLDCGs in NSCLC immune microenvironment and immune downstream mechanisms: mediation MR analysis

SLDCGs immune infiltration analysis

The immune infiltration analysis revealed that in samples with high expression of HLA-J, the infiltration levels of antigen-presenting cells (including aDC and iDC) and macrophages were significantly increased ($P<0.$ 001). Additionally, high expression of HLA-J was significantly associated with an increased level of regulatory T cell (Treg) enrichment ($P<0.$ 001). Simultaneously, HLA-J high expression was also accompanied by a significant increase in the infiltration levels of effector immune cells, such as CD8 + T cells and effector memory T cells (Tem) ($P<0.$ 01). Overall, HLA-J may play an important role in tumor immune recognition and cytotoxicity by promoting the function of antigen-presenting cells and enhancing the activity of effector T cells.

In contrast, in the PRMT7 high-expression group, the infiltration levels of effector immune cells, such as CD8 + T cells and effector memory T cells (Tem), were



	PRMT7(Binding energy)	HLA-J(Binding energy)	PRMT7(hydrogen bond count)	HLA-J(hydrogen bond count)
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone	5.71	4.86	4	1
4-Aminobiphenyl	6.52	5.18	2	0
Acrolein	2.75	2.62	1	2
N'-Nitrosonornicotine	7.19	4.57	2	2

Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 Analysis of SLDCGs and Molecular Docking Results. (**a**) GO Enrichment Analysis of SLDCGs: Includes Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). In the bubble chart, the size of each bubble represents the number of genes associated with each GO term. Larger bubbles indicate a greater number of genes enriched in the term, suggesting a closer relationship with the study objectives. The color of the bubbles reflects the significance level, with darker blue indicating a higher likelihood of statistical significance (smaller P value). (**b**) PPI Network of PRMT7: The color and thickness of the lines represent the combined_score values of the interactions (ranging from 0. 5 to 0. 9). (**c**) ROC Curve Analysis for HLA-J and PRMT7 Genes: The AUC value ranges from 0. 5 to 1, where a higher AUC value closer to 1 indicates greater diagnostic ability to distinguish between the normal and tumor groups. (**d**) Visualization heatmap of binding energy absolute values and hydrogen bond counts for molecular docking. (**e**) Docking results of 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone with HLA-J protein. (**f**) Docking results of 4-Aminobiphenyl with HLA-J protein. (**g**) Docking results of Acrolein with HLA-J protein. (**h**) Docking results of N′-Nitrosonornicotine with HLA-J protein. (**i**) Docking results of 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone with PRMT7 protein. (**j**) Docking results of 4-Aminobiphenyl with PRMT7 protein. (**k**) Docking results of Acrolein with PRMT7 protein. (**l**) Docking results of N′-Nitrosonornicotine with PRMT7 protein

significantly reduced ($P<0.\ 01$). The infiltration levels of immune-suppressive cells, such as macrophages and aDC, were also reduced, and the enrichment of Treg cells significantly decreased ($P<0.\ 05$). These results suggest that PRMT7 may play a crucial role in tumorigenesis and progression by promoting immune suppression. Detailed results are shown in Fig. 6a and e, with data presented in Supplementary Table 8.

Expression and annotation of SLDCGs in Single-Cell subpopulations

In the single-cell subpopulation expression analysis of the HLA-J gene, the results in Fig. 6f and g showed that HLA-J expression was not significantly concentrated in a specific type of T cell subpopulation but exhibited high expression in certain single cells in localized regions. Figure 6g shows that HLA-J expression was lowest in the monocyte/macrophage group (Mono/Macro), while the average expression level in Treg cells was significantly higher than in other cell populations. These findings suggest that HLA-J may participate in immune regulation within the tumor microenvironment through finely tuned high expression in specific cell populations.

The single-cell analysis of PRMT7 expression further revealed its distribution characteristics in different cell subpopulations. Results in Fig. 6f and h show that PRMT7 expression was significantly higher in the Tprolif (proliferating T cells) population compared to other groups. Taken together, PRMT7 may enhance the activity of proliferating T cells, regulate the immune function of exhausted T cells, and influence the behavior of immune-suppressive Treg cells.

Mediation MR analysis of SLDCGs influencing immune downstream mechanisms in NSCLC

In the MR analysis based on the IVW method, two immune cell subpopulations, “HLA DR on B cells” and “CD8 on CD39 + CD8br”, were identified as having a significant causal relationship with HLA-J and NSCLC. “HLA DR on B cells” belongs to the Treg subpopulation, and “CD8 on CD39 + CD8br” belongs to another subpopulation. In the mediation effect analysis, the study explored the regulatory mechanism by which the HLA-J gene influences the risk of NSCLC through “HLA DR on B cells” and “CD8 on CD39 + CD8br”. The analysis revealed that the total effect of HLA-J on NSCLC risk was negative (beta_all = -0. 1179). The indirect effect through the mediator “HLA DR on B cells” was − 0. 032 (95% CI: -0. 057, -0. 007, $P=0.\ 012$), which accounted for 21. 25% of the total effect, indicating that the indirect effect is significant, though relatively small. The direct effect was − 0. 086, accounting for the majority of the total effect. Similarly, in the analysis with “CD8 on CD39 + CD8br” as the mediator, the indirect effect was − 0. 0330 (95% CI: -0. 0628, -0. 0031, $P=0.\ 031$), accounting for 21. 85% of the total effect, which was statistically significant. The direct effect was − 0. 0849, remaining the dominant contribution pathway, as shown in Table 1. These results suggest that the protective effect of HLA-J on NSCLC risk is mainly mediated through the direct effect. Although the mediation effect is not the primary pathway through which HLA-J influences NSCLC, it still plays an important role in the immune regulatory mechanisms.

Discussion

This study, by integrating data from multiple public databases, provides a deep exploration of the potential molecular mechanisms underlying smoking-related NSCLC. By analyzing the genetic differences between smoking NSCLC patients and non-cancerous smokers, we identified potential SLDCGs (HLA-J and PRMT7), further revealing the critical role of the immune microenvironment in smoking-related NSCLC. The key contribution of this study lies in systematically integrating data and using causal inference to clarify the potential role of SLDCGs in the development of NSCLC, thereby offering potential new targets for immunotherapy and targeted therapies. Through this research, we aim to provide novel biomarkers for the early diagnosis and personalized treatment of NSCLC, as well as important theoretical support for the study of molecular mechanisms in smoking-related NSCLC.

HLA-J is a relatively under-researched gene and a member of the HLA family. HLA molecules play a key role in tumor immune surveillance and antigen presentation, acting as critical factors in T-cell recognition of

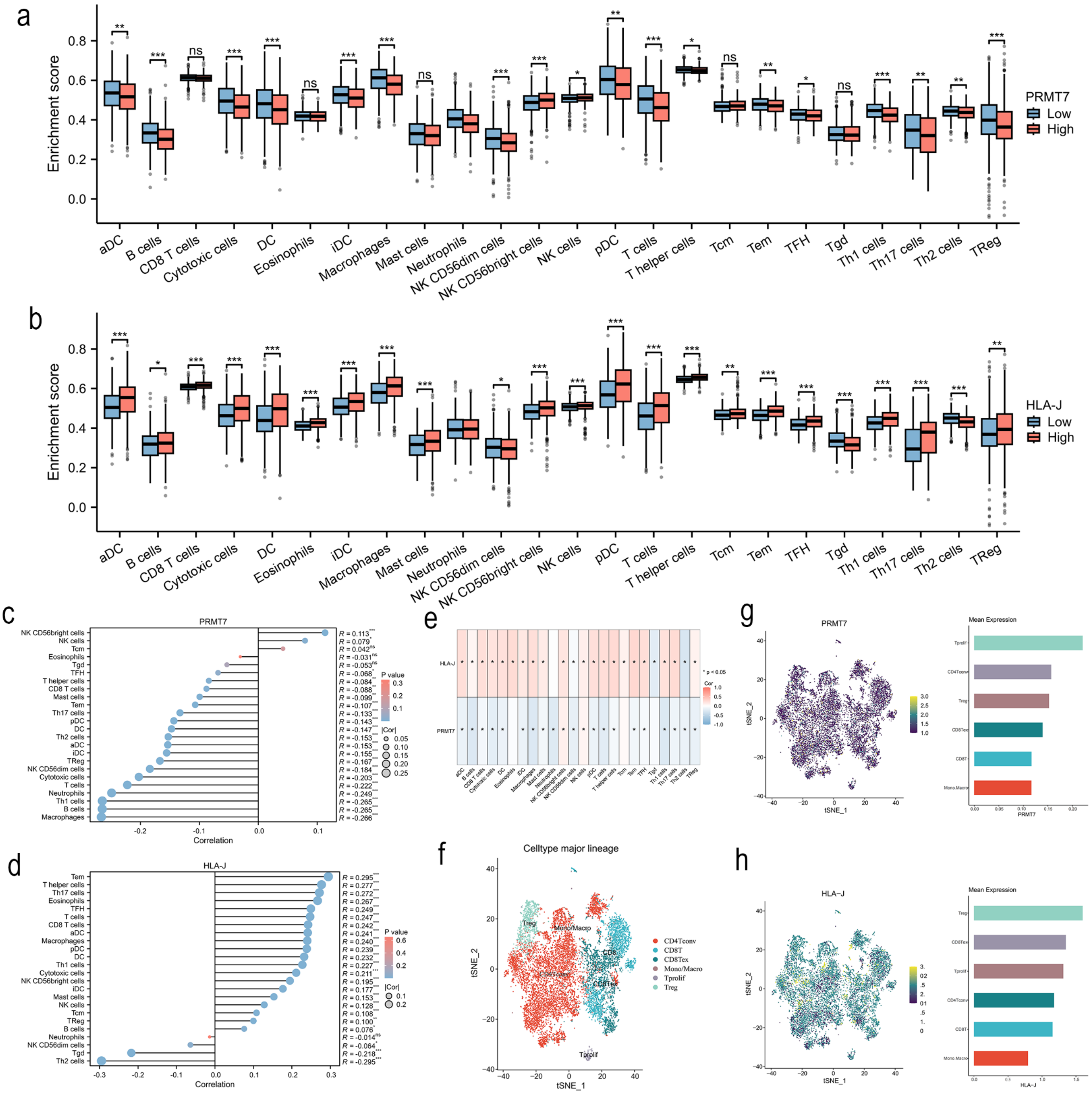


Fig. 6 Immune Cell Infiltration Analysis and Correlation Study of PRMT7 and HLA-J Genes. (**a-b**) Box plots comparing the immune cell enrichment levels between high and low expression groups of PRMT7 (**a**) and HLA-J (**b**). Each box plot represents the immune cell infiltration scores for the high expression group (red) and low expression group (blue). The middle line indicates the median, and the upper and lower box boundaries represent the upper and lower quartiles. (**c-d**) Scatter plots showing the correlation between PRMT7 (**c**) and HLA-J (**d**) gene expression and immune cell infiltration levels. Each point represents an immune cell type, with the point size indicating the significance of the correlation (-log10 of P value). The color of the points represents the direction and strength of the correlation: red indicates a positive correlation, and blue indicates a negative correlation. The intensity of the color reflects the strength of the correlation (the higher the absolute value, the darker the color). The x-axis represents the correlation coefficient. (**e**) Immune Correlation Heatmap: Each small square represents the correlation coefficient between molecules and cells. Darker colors indicate stronger correlations between the variables. Significance levels are indicated as * ($P < 0.05$), ** ($P < 0.01$), *** ($P < 0.001$). (**f**) t-SNE plot showing the distribution of major cell types in the single-cell RNA sequencing data. Different colors represent the six major cell subpopulations. (**g**) t-SNE plot showing the expression distribution of the HLA-J gene in single cells. The bar graph quantifies the average expression levels of HLA-J in different cell subpopulations. (**h**) t-SNE plot showing the expression distribution of the PRMT7 gene in single cells. The color gradient from purple (low expression) to yellow (high expression) represents the relative expression levels of the gene. The bar graph quantifies the average expression levels of PRMT7 in different cell subpopulations

Table 1 Mediation MR analysis of SLDCGs influencing immune downstream mechanisms in NSCLC

Effect Type	HLA DR on B cells (Treg)	CD8 on CD39 + CD8br (TBNK)
β_3[a]	−0. 1179	
β_1[b]	−0. 223,874,915	0. 182,144,243
β_2[c]	0. 142,144,391	−0. 180,979,521
Mediation Effect β (95% CI)[d]	−0. 032 (−0. 057,−0. 007)	−0. 0330 (−0. 063,−0. 003)
Mediated Proportion (%)[e]	21. 25%	21. 85%
P_value[f]	0. 012	0. 031
Total Effect[g]	−0. 1497	−0. 1509

a: β_3 = The causal effect of HLA-J on NSCLC

b: β_1= The causal effect of HLA-J on HLA-DR expression on B cells; The causal effect of HLA-J on CD8 + CD39 + CD8br cells

c: β_2 = The causal effect of HLA-DR expression on B cells on NSCLC; The causal effect of CD8 + CD39 + CD8br cells on NSCLC

d: Mediation effect = The indirect causal effect of SLDCGs on NSCLC through immune cells, represented by the product of β_1 and β_2

e: The mediated proportion = The proportion of the indirect effect relative to the total effect, calculated as the ratio of indirect effect / total effect, $[\beta_1 \times \beta_2 / (\beta_3 + \beta_1 \times \beta_2)]$

f: P_value = The significance of the indirect effect, with $P < 0.05$ supporting the alternative hypothesis (presence of mediation effect)

g: Total Effect= $(\beta_3 + \beta_1 \times \beta_2)$, the sum of the direct and indirect effects

tumor cells [21, 22]. These molecules are involved in the immune evasion mechanisms of tumor cells, regulating the activity of T cells and B cells, thereby influencing tumorigenesis and development [23]. As research into immunotherapy deepens, HLA molecules, as potential biomarkers and therapeutic targets, may open up new pathways for improving treatment outcomes and survival prognosis in lung cancer patients [24]. The results of this study show that HLA-J is a protective factor for lung cancer. Its expression is significantly lower in tumor tissues compared to normal tissues, and the expression of HLA-J is significantly correlated with immune cell infiltration in tumor tissues. Similar to other HLA molecules, the low expression of HLA-J may enable tumor cells to escape host immune surveillance, promoting tumor growth and metastasis [22]. As a relatively novel HLA molecule, the immune-regulatory and anti-tumor roles of HLA-J are particularly noteworthy and warrant further investigation.

PPRMT7 is a member of the protein arginine methyltransferases (PRMTs) family. PRMTs regulate various biological processes in cells through the arginine methylation of proteins, including gene expression, RNA splicing, and cellular metabolism [25]. Alterations in these processes are closely linked to the onset of various diseases, especially in cancer development, where the role of PRMTs cannot be overlooked [26]. Research shows that PRMTs play an important role in processes such as tumor cell proliferation, apoptosis, and metastasis. Therefore, targeting PRMTs for cancer therapy has become a key area of research [27]. PRMT7, as an important enzyme, has been receiving increasing attention. Low expression of PRMT7 is associated with tumor size, degree of differentiation, and lymph node metastasis [28]. Previous studies have found that PRMT7 interacts with the PTEN protein, promoting its arginine methylation, which in turn inhibits the activation of the PI3K/AKT signaling pathway, thereby suppressing cell proliferation and migration [28]. In this study, the high expression of PRMT7 in lung cancer is particularly significant. High expression of PRMT7 serves as a risk factor for lung cancer and is highly expressed in lung cancer tissues. PRMT7 may promote immune escape by methylating specific transcription factors and inhibiting anti-tumor immune responses, leading to the evasion of immune surveillance by tumor cells [29].

This study, by analyzing the expression of SLDCGs genes and their relationship with clinical features in the TCGA dataset, reveals the expression specificity of HLA-J and PRMT7 in different populations. For example, HLA-J expression was significantly higher in female patients compared to male patients. The influence of estrogen in women may contribute to the differences in HLA gene expression between males and females. This discrepancy may lead to a stronger immune response in women when facing NSCLC [30]. The low expression of HLA-J is also significantly associated with smoking behavior and longer smoking duration, suggesting that HLA-J gene polymorphisms may play an important role in the pathogenesis of smoking-related NSCLC [31]. Regarding tumor types, high expression of PRMT7 was significantly associated with squamous cell carcinoma, while high expression of HLA-J was significantly associated with adenocarcinoma. This finding suggests that the expression of HLA-J and PRMT7 is also subtype-specific in NSCLC, which may contribute to the ability of tumor cells to evade host immune surveillance [32]. The diagnostic ROC analysis showed that SLDCGs have high predictive value for early NSCLC diagnosis. Studies have found that in NSCLC, 61% of cases exhibit HLA transcriptional repression [32]. In the Cox proportional hazards regression model, high expression of HLA-J was significantly associated with prolonged overall survival (OS) in patients, suggesting that high expression of HLA-J may be a protective factor for better prognosis in NSCLC patients, similar to the behavior of other HLA molecules [33, 34]. Tumors with high HLA expression are often associated with better immune therapy responses [35]. Functional analysis revealed that SLDCGs significantly participate in key epigenetic processes, such as histone methylation, arginine methyltransferase activity, and genomic imprinting regulation in NSCLC. PRMT7 forms an intensive interaction network with various epigenetically related proteins (such as PRMT5 and FBL), suggesting that it may regulate the transcriptional activity of NSCLC cells through epigenetic mechanisms, thus promoting tumor progression [36]. Molecular docking indirectly confirmed the interaction between tobacco carcinogens and key residues of SLDCGs proteins, indicating that these interactions may alter the protein structure, disrupt its normal biological function, and promote tumorigenesis and immune escape [37].

This study demonstrated the role of SLDCGs in the NSCLC immune microenvironment through immune infiltration analysis. High expression of HLA-J was significantly associated with the infiltration of antigen-presenting cells and effector immune cells, further revealing that HLA-J, as a member of the HLA family, enhances immune cell functions, thereby promoting tumor immune recognition and immune effects [38]. In contrast, in the PRMT7 high-expression group, the infiltration of effector immune cells was significantly reduced, suggesting that PRMT7 may promote immune escape by inhibiting immune cell activity and weakening tumor immune surveillance [39]. The annotation of T cell subpopulations also suggested that HLA-J is primarily highly expressed in Treg cells, while PRMT7 is significantly expressed in proliferating T cells. Studies have found that Treg cells can regulate the expression of HLA molecules on antigen-presenting cells through cytokine secretion and direct cell-cell contact. For example, in transplantation models, Treg cells inhibited the expression of HLA-DR and HLA-DQ on microvascular endothelial cells, reducing T cell stimulation and thereby alleviating graft rejection responses [40]. Mediation MR analysis also found that HLA-J exerts a significant protective effect on NSCLC risk through both direct effects and immune cell-mediated effects. Although the direct effect dominates, the indirect effect mediated by immune cells also plays an important role in the persistence and efficacy of the anti-tumor immune response. These results provide theoretical support for targeted HLA-J tumor immunotherapy and further deepen the understanding of the NSCLC immune microenvironment [41].

This study systematically reveals the role of SLDCGs in the development of NSCLC and in the immune microenvironment, exploring and validating SLDCGs as carcinogenic biomarkers and functional mechanisms in smoking populations. This finding not only uncovers the gene specificity in populations affected by smoking-related NSCLC but also deepens our understanding of the immune microenvironment in NSCLC development. Furthermore, it provides new theoretical support for immune therapeutic strategies targeting HLA-J and PRMT7. However, there are certain limitations in this study. First, while this study utilized publicly available databases, the sample size analyzed was relatively small. The limited sample size could affect the generalizability and statistical power of the results. Future research should increase the sample size to enhance the robustness and reliability of the conclusions. Second, the patient data from the public databases may include geographic and ethnic differences, which limits the external validity of the study's results. Patients from different regions may exhibit distinct gene expression profiles, which could influence the outcomes. Lastly, although this study explored the mechanisms of SLDCGs gene function in NSCLC development and their specificity in different populations using a multi-dimensional approach, further experimental validation is needed, especially in clinical samples. Therefore, future studies should validate the role of SLDCGs in the pathogenesis of smoking-related NSCLC in larger clinical cohorts and explore their potential application in other types of cancer.

Conclusion

This study systematically explores the role of SLDCGs in the development of NSCLC, their expression in specific populations, and their involvement in the immune microenvironment. These findings provide new molecular insights into the understanding of NSCLC development and immune escape mechanisms in smoking populations and lay a theoretical foundation for the development of SLDCGs-based immunotherapy and targeted treatment strategies. However, further clinical validation and experimental studies are required to explore their potential applications in early diagnosis, prognosis assessment, and treatment of NSCLC.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06301-z.

Supplementary Material 1

Acknowledgements

We would like to express our sincere gratitude to the following institutions and individuals for their invaluable support in this study:

Author contributions

All authors have access to the manuscript's data and all drafts. The specific contributions are as follows: Can ouyang drafted the manuscript. Data collection, data management, and data analysis: Can ouyang and Xiaopeng Yu. Research Design: Can ouyang, Xiaopeng Yu, and Huazhong Wang. Image editors: Can ouyang and Xiaopeng Yu. Statistical analysis: Can ouyang, Xiaopeng Yu, Huazhong Wang, and Puhua Zeng. Manuscript revisions: Puhua Zeng. Manuscript review: All authors. All authors have reviewed and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China (82074425), Natural Science Foundation of Hunan Provincial(2023JJ30364 and 2023JJ30361), Key Research and Development Program of Hunan Provincial (2023SK2057), Key Project of Hunan Provincial Administration of Traditional Chinese Medicine (A2023042), Graduate Research Innovation Program of Hunan Provincial(CX20230835), Health and Wellness Research Program of Hunan Provincial(20232605), Major Scientific Research Project for High-Level Talent of Hunan Provincial(R2023081); Clinical Research Center for the Prevention and Control of Cancer Integrating Traditional Chinese and Western Medicine (Hunan Traditional Chinese Medicine Letter [2022] No. 93), Traditional Chinese Medicine Shennong Leading Talent Program of the Hunan Provincial Health Commission, and Hunan Provincial High-Tech Talent Program of the Hunan Provincial Science and Technology Department. We gratefully acknowledge the financial support which made this research possible.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request. The publicly available data used in this study are as follows:
1) Gene Expression Omnibus (GEO): The GEO datasets (e.g., GSE99254, GSE19027) used in this study are publicly available in the GEO database, accessible at: https://www.ncbi.nlm.nih.gov/geo/.
2) The Cancer Genome Atlas (TCGA): Data from TCGA, including gene expression and clinical information for lung cancer, are accessible via the TCGA portal at: https://portal.gdc.cancer.gov/.
3) Genotype-Tissue Expression (GTEx): The GTEx data for tissue-specific gene expression can be accessed at: https://gtexportal.org/home/.
4) eQTLGen Consortium: Data from the eQTLGen consortium, which contains information on gene expression and genetic variation, can be accessed at: https://eqtlgen.org/.
All data used in this study were publicly available and accessed according to their respective access policies. For additional data or materials, requests can be made to the corresponding author.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the ethical standards established by the Declaration of Helsinki. As all the data used in this study are publicly available and anonymized, there was no requirement for informed consent from individual participants.

Competing interests

The authors declare that they have no competing interests.

Peer reviewers

Special thanks to the anonymous peer reviewers for their constructive comments and suggestions, which significantly improved the quality of this manuscript.

Author details

[1]Hunan Provincial Hospital of Integrated Traditional Chinese and Western Medicine, Changsha, Hunan 410006, People's Republic of China
[2]School of Integrated Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha, Hunan 410208, People's Republic of China
[3]Cancer Research Institute of Hunan Academy of Traditional Chinese Medicine, Changsha, Hunan 410006, People's Republic of China

Received: 23 December 2024 / Accepted: 23 February 2025
Published online: 14 March 2025

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. Zhang P, Chen PL, Li ZH, et al. Association of smoking and polygenic risk with the incidence of lung cancer: a prospective cohort study. Br J Cancer. 2022;126(11):1637–46.
3. Shiels MS, Graubard BI, McNeel TS, Kahle L, Freedman ND. Trends in smoking-attributable and smoking-unrelated lung cancer death rates in the united States, 1991–2018. J Natl Cancer Inst. 2024;116(5):711–6.
4. LoPiccolo J, Gusev A, Christiani DC, Jänne PA. Lung cancer in patients who have never smoked - an emerging disease. Nat Rev Clin Oncol. 2024;21(2):121–46.
5. Benusiglio PR, Fallet V, Cadranel J. Invited editorial: Q and A on hereditary lung cancer. Respir Med Res. 2022;81:100881.
6. Benusiglio PR, Fallet V, Sanchis-Borja M, Coulet F, Cadranel J. Lung cancer is also a hereditary disease. Eur Respir Rev. 2021;30(162):210045.
7. Gesthalter YB, Vick J, Steiling K, Spira A. Translating the transcriptome into tools for the early detection and prevention of lung cancer. Thorax. 2015;70(5):476–81.
8. Zhang J, Chang L, Jin H, et al. Benzopyrene promotes lung cancer A549 cell migration and invasion through up-regulating cytokine IL8 and chemokines CCL2 and CCL3 expression. Exp Biol Med (Maywood). 2016;241(14):1516–23.
9. Qu Z, Tian J, Sun J, et al. Diallyl trisulfide inhibits 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone-induced lung cancer via modulating gut microbiota and the PPARγ/NF-κB pathway. Food Funct. 2024;15(1):158–71.
10. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97.
11. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53(9):1300–10.
12. Gkatzionis A, Burgess S, Newcombe PJ. Statistical methods for cis-Mendelian randomization with two-sample summary-level data. Genet Epidemiol. 2023;47(1):3–25.
13. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int J Epidemiol. 2011;40(3):740–52.
14. Lee CH, Cook S, Lee JS, Han B. Comparison of two Meta-Analysis methods: Inverse-Variance-Weighted average and weighted sum of Z-Scores. Genomics Inf. 2016;14(4):173–80.
15. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect Estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512–25.
16. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal Pleiotropy assumption. Int J Epidemiol. 2017;46(6):1985–98.
17. Yuan S, Kim JH, Xu P, Wang Z. Causal association between Celiac disease and inflammatory bowel disease: A two-sample bidirectional Mendelian randomization study. Front Immunol. 2022;13:1057253.
18. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46.
19. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
20. Bindea G, Mlecnik B, Tosolini M, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39(4):782–95.
21. Zhou H, Chan KC, Buratto D, Zhou R. The rigidity of a structural Bridge on HLA-I binding groove explains its differential outcome in cancer immune response. Int J Biol Macromol. 2023;253(Pt 7):127199.
22. Datar IJ, Hauc SC, Desai S, et al. Spatial analysis and clinical significance of HLA Class-I and Class-II subunit expression in Non-Small cell lung Cancer. Clin Cancer Res. 2021;27(10):2837–47.
23. Li Q, Jia C, Pan W, et al. Multi-omics study reveals different pathogenesis of the generation of skin lesions in SLE and IDLE patients. J Autoimmun. 2024;146:103203.
24. Zhang B, Ren Z, Zhao J, et al. Global analysis of HLA-A2 restricted MAGE-A3 tumor antigen epitopes and corresponding TCRs in non-small cell lung cancer. Theranostics. 2023;13(13):4449–68.

25. Gao Y, Feng C, Ma J, Yan Q. Protein arginine methyltransferases (PRMTs): orchestrators of cancer pathogenesis, immunotherapy dynamics, and drug resistance. Biochem Pharmacol. 2024;221:116048.
26. Jarrold J, Davies CC. PRMTs and arginine methylation: Cancer's Best-Kept secret. Trends Mol Med. 2019;25(11):993–1009.
27. Bryant JP, Heiss J, Banasavadi-Siddegowda YK. Arginine methylation in brain tumors: tumor biology and therapeutic strategies. Cells. 2021;10(1):124.
28. Wang X, Xu W, Zhu C, Cheng Y, Qi J. PRMT7 inhibits the proliferation and migration of gastric Cancer cells by suppressing the PI3K/AKT pathway via PTEN. J Cancer. 2023;14(15):2833–44.
29. Halabelian L, Barsyte-Lovejoy D. Structure and function of protein arginine methyltransferase PRMT7. Life (Basel). 2021;11(8):768.
30. Bettens F, Ongen H, Rey G, et al. Regulation of HLA class I expression by non-coding gene variations. PLoS Genet. 2022;18(6):e1010212.
31. Hedström AK, Rönnelid J, Klareskog L, Alfredsson L. Complex relationships of smoking, HLA-DRB1 genes, and serologic profiles in patients with early rheumatoid arthritis: update from a Swedish Population-Based Case-Control study. Arthritis Rheumatol. 2019;71(9):1504–11.
32. Puttick C, Jones TP, Leung MM, et al. MHC hammer reveals genetic and non-genetic HLA disruption in cancer evolution. Nat Genet. 2024;56(10):2121–31.
33. Huang YJ, He JK, Duan X, Hou R, Shi J. Prognostic gene HLA-DMA associated with cell cycle and immune infiltrates in LUAD. Clin Respir J. 2023;17(12):1286–300.
34. Zhang X, Tang H, Luo H, et al. Integrated investigation of the prognostic role of HLA LOH in advanced lung cancer patients with immunotherapy. Front Genet. 2022;13:1066636.
35. Lau D, Khare S, Stein MM, et al. Integration of tumor extrinsic and intrinsic features associates with immunotherapy response in non-small cell lung cancer. Nat Commun. 2022;13(1):4053.
36. Fulton MD, Cao M, Ho MC, Zhao X, Zheng YG. The macromolecular complexes of histones affect protein arginine methyltransferase activities. J Biol Chem. 2021;297(4):101123.
37. Chen IL, Todd I, Tighe PJ, Fairclough LC. Electronic cigarette vapour moderately stimulates pro-inflammatory signalling pathways and interleukin-6 production by human monocyte-derived dendritic cells. Arch Toxicol. 2020;94(6):2097–112.
38. Teng Y, Quah HS, Suteja L, et al. Analysis of T cell receptor clonotypes in tumor microenvironment identifies shared cancer-type-specific signatures. Cancer Immunol Immunother. 2022;71(4):989–98.
39. Srour N, Villarreal OD, Hardikar S, et al. PRMT7 ablation stimulates anti-tumor immunity and sensitizes melanoma to immune checkpoint Blockade. Cell Rep. 2022;38(13):110582.
40. Cross AR, Lion J, Poussin K, Glotz D, Mooney N. Inflammation determines the capacity of allogenic endothelial cells to regulate human Treg expansion. Front Immunol. 2021;12:666531.
41. Lin A, Yan WH. HLA-G/ILTs targeted solid Cancer immunotherapy: opportunities and challenges. Front Immunol. 2021;12:698677.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.