Am J Cancer Res 2025;15(3):855-875
www.ajcr.us /ISSN:2156-6976/ajcr0161352

Original Article

Causal effect of interstitial lung disease on lung cancer risk in never-smokers and prognostic insights from Mendelian randomization and transcriptome analysis

Limin Chi[1*], Mengyan Li[1*], Hanxing Zhou[1*], Mien-Chie Hung[2,3,4], Wei-Jan Wang[3,5], Bo Wang[1], Xian Sun[1,6]

[1]Department of Oncology, The Seventh Affiliated Hospital, Sun Yat-Sen University, Shenzhen, Guangdong, P. R. China; [2]Graduate Institute of Biomedical Sciences, Institute of Biochemistry and Molecular Biology, China Medical University, Taichung 406, Taiwan; [3]Research Center for Cancer Biology, Cancer Biology and Precision Therapeutics Center, China Medical University, Taichung 406, Taiwan; [4]Center for Molecular Medicine, China Medical University Hospital, Taichung 406, Taiwan; [5]Department of Biological Science and Technology, College of Life Sciences, China Medical University, Taichung, Taiwan; [6]Shenzhen Key Laboratory of Chinese Medicine Active Substance Screening and Translational Research, Shenzhen, Guangdong, P. R. China. [*]Equal contributors.

Received October 21, 2024; Accepted February 8, 2025; Epub March 15, 2025; Published March 30, 2025

Abstract: Introduction: The relationship between interstitial lung disease (ILD) and lung cancer in nonsmokers (LCINS) has garnered increasing interest. However, the causal associations and underlying pathogenesis between ILD and LCINS remain poorly understood. Methods: This research utilized a bidirectional two-sample Mendelian randomization (MR) method, utilizing forward MR analysis to assess the causal impact of ILD on LCINS and reverse MR analysis to evaluate the causal effect of LCINS on ILD. Additionally, transcriptome data and bioinformatics analyses were used to explore the associations between ILD and LCINS. An ILD-related gene signature (ILD risk score) was identified to examine its influence on the hallmark signaling pathways and the immune microenvironment in LCINS. Results: The study revealed a significant causal relationship between ILD and LCINS, with ILD increasing the risk of lung cancer in nonsmoking European populations. We developed a 5-gene risk model, which includes CD1A, CDH3, KRT6B, MMP1, and MMP10, via least absolute shrinkage and selection operator (LASSO) regression. The ILD risk score independently influences the prognosis of nonsmoking patients with lung cancer, and these five genes are also significantly associated with overall survival (OS) rates. Patients in the high-ILD risk subgroup exhibited significantly poorer survival rates. A highly accurate nomogram was developed to increase the clinical applicability of the ILD risk score. Additionally, the ILD risk scores were significantly correlated with hallmark signaling pathways and immune cell infiltration. Conclusions: This study suggested that ILD may have a positive causal effect on LCINS, with the ILD risk score serving as an effective predictor of the prognoses in LCINS patients. It is associated with tumor proliferation and the activation of metabolism-related signaling pathways. These findings also indicate that ILD may contribute to the occurrence and progression of LCINS through its influence on immune cell infiltration.

Keywords: Interstitial lung disease (ILD), lung cancer in never smokers (LCINS), Mendelian randomization (MR), transcriptomic data

Introduction

Lung cancer is the most common type of cancer worldwide and accounts for the highest number of cancer-related deaths [1, 2]. Although cigarette smoking is the predominant risk factor associated with lung cancer, 10% to 25% of lung cancer cases are observed in individuals who are nonsmokers [1, 3-5]. Lung cancer may be related to gene mutations, the tumor microenvironment, and inflammatory factors [6-9]. The incidence of lung cancer among individuals who have never smoked (LCINS) is notably elevated in East Asia, accounting for approximately one-third of all lung cancer cases [10]. In the USA, approximately 20,500 deaths from lung cancer not attributed to smoking rank as the eighth most prevalent cause of cancer-

https://doi.org/10.62347/NJCQ2464

related death [4]. LCINS is histologically and epidemiologically distinguishable from smoking-related lung cancer. It predominantly presents as adenocarcinoma and is most commonly found in women and individuals of Asian descent. This subtype also exhibits a relatively high prevalence of genetic mutations. Risk factors for LCINS include exposure to secondhand smoke, occupational hazards, family history of lung cancer, hormonal infuences, and preexisting diseases [11].

ILD is a group of various diffuse parenchymal lung diseases with increasing incidence. Idiopathic pulmonary fibrosis (IPF) is a typical form of ILD. In a retrospective study involving 938 patients with IPF, 135 (14.5%) developed lung cancer during follow-up, and the cumulative incidence rates of lung cancer at 1, 3, 5, and 10 years were 1.1%, 8.7%, 15.9%, and 31.1%, respectively [12]. HUBBARD [13] demonstrated that the occurrence of lung cancer in individuals diagnosed with cryptogenic fibrosing alveolitis is increasing, with an incidence ratio of 7.31, and this effect was found to be independent of smoking. Individuals diagnosed with IPF are found to develop lung cancer approximately 3.34 times more frequently than those in the general population. Furthermore, their prognosis is poorer than that of IPF patients without lung cancer. However, research has revealed no evidence of an increased risk of lung cancer among individuals who never smoke [14].

The association between ILD and lung cancer has been explored in observational studies and meta-analyses. However, conflicting conclusions have been derived from observational studies due to the potential influence of numerous confounding factors. Therefore, a more carefully designed approach is needed to assess the causal relationship between ILD and lung cancer risk. This research employed Mendelian randomization to discover specific single-nucleotide polymorphisms (SNPs) linked to ILD by analyzing summary-level data from previous genome-wide association studies (GWASs). The SNPs were then utilized to assess the impact of ILD on susceptibility to LCINS. Furthermore, transcriptome data and bioinformatics analyses were employed to explore the relationship between ILD and LCINS, as well as to investigate the hallmark signaling pathways and immune infiltration in LCINS. These findings contribute to an enhanced understanding of the correlation between ILD and LCINS, providing potential targets and predictive models for nonsmoking-related lung cancer patients with coexisting ILD.

Method

The overview design of our work is depicted in **Figure 1**. The data sources for MR and transcriptome data are accessible to the public online.

Mendelian randomization analysis

We performed a bidirectional two-sample MR analysis following the latest STROBE-MR (Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization) guidelines [15]. We employed a rigorous quality control procedure to identify relevant SNPs from the results of GWASs. We used MR analysis to assess the causal relationship between ILD and LCINS. With SNPs as Instrumental variables (IVs), MR studies need to satisfy three hypotheses: (1) IVs directly impact exposure; (2) IVs do not correlate with confounding factors; and (3) IVs influence outcome solely by influencing exposure [16].

Data sources: The GWAS summary dataset for ILD (finn-b-ILD), derived from the FINNGEN research, consists of 1,969 cases and 196,986 controls of European ancestry. The GWAS summary data for LCINS (ebi-a-GCST004747) originate from a comprehensive meta-analysis of GWAS, encompassing 2,355 cases of LCINS and 75,044 controls of European descent. The MRC IEU Open GWAS Project database (https://gwas.mrcieu.ac.uk/) provides access to the summary statistics for these GWASs (**Table 1**).

Selection of genetic instruments: For forward MR analysis, two-sample MR analysis was performed, with ILD as the exposure and LCINS as the outcome. We obtained genetic variants related to ILD, totaling 16380417 SNPs, from the GWAS summary data (GWAS ID: finn-b-ILD). Owing to the limited number of available SNPs, to guarantee an adequate quantity of IVs and to maintain the robustness of our research results, we used a more lenient threshold of a p value of 5×10e5. The impact of linkage disequi-

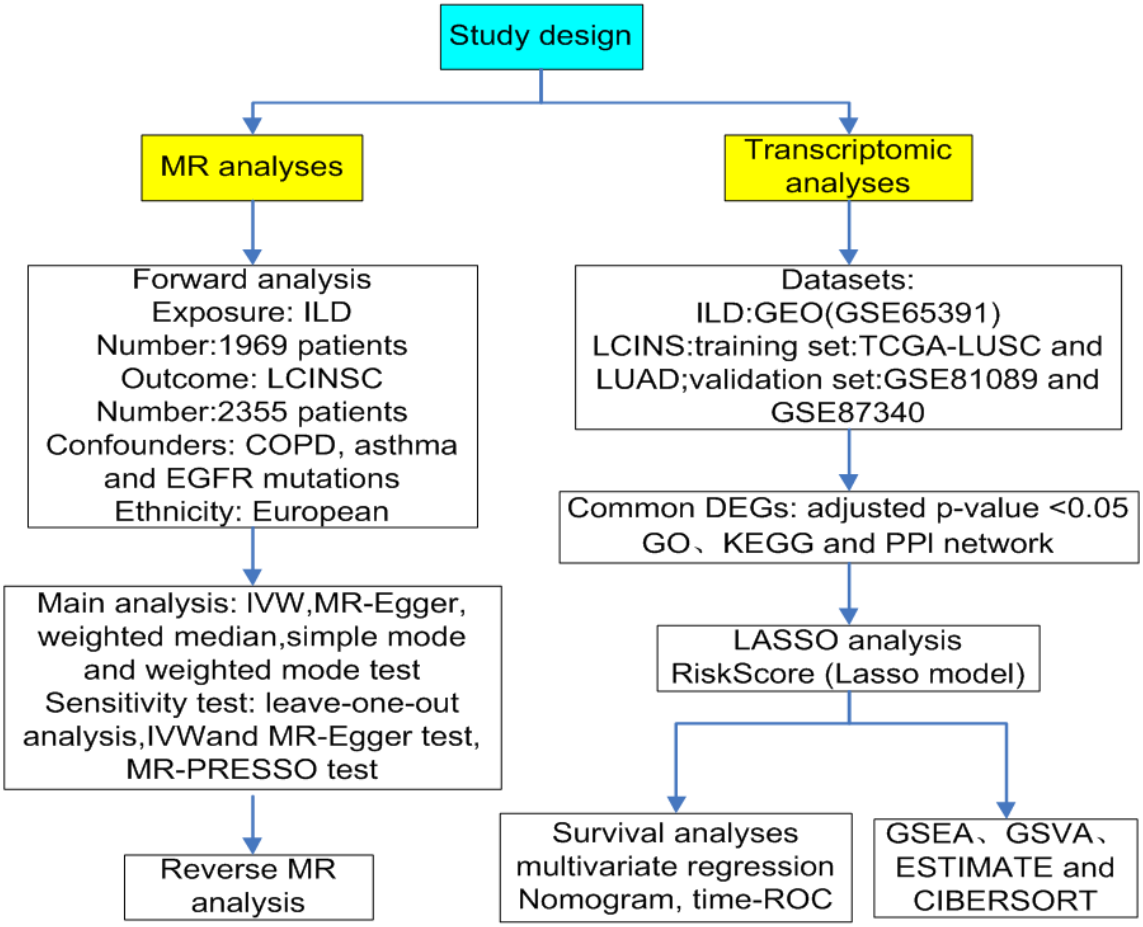

Figure 1. Overview of research design.

librium was addressed by choosing a clumping distance of 10,000 kb and setting a threshold of <0.001 for R^2. Furthermore, the potential instrumental bias was evaluated by assessing the F statistic via the formula, $F = \frac{R^2(N - k - 1)}{k(1 - R^2)}$ [17]. Here, the sample size, number of SNPs and fraction of variance explained by IVs are represented by N, k and R^2, respectively. When the F statistic was less than 10, the influence of IVs in this study was considered weak. COPD, asthma and EGFR mutations have been identified as potential confounding factors in the association with LCINS [11]. We excluded four SNPs associated with outcome and exposure after a manual review of potential confounding factors. We subsequently identified 90 SNPs that are associated with ILD as IVs.

Data harmonization: We harmonized the allelic orientation of the filtered SNPs in both the exposure and outcome datasets to guarantee compatibility and uniformity. We removed incompatible SNP (rs12934985) and palindromic SNPs (rs11126629 and rs9613668) to ensure that the effector alleles (and corresponding β and effector allele frequencies) in the outcome dataset were the same as the effector alleles reflected in the exposure data. After the harmonization analysis, the MR analysis included 75 SNPs in the study.

Two-sample MR analysis: To investigate the causal impact of ILD on the risk of LCINS, we employed five different analysis methods, including inverse variance weighting (IVW) [18], MR-Egger [19], and weighted median (WM) [20] approaches, as well as simple mode and weighted mode [21] methods. The IVW method served as our primary analysis, given its efficiency when all instrumental variables (IVs) are valid and the absence of horizontal pleiotropy is assumed. However, as IVW requires that the

Table 1. Information of GWASs analyzed in the current MR analyses

Phenotype	First author	Ncase	Ncontrol	Number of SNPs	Ethnicity	Trait ID in GWAS	Year
ILD	NA	1,969	196,986	16,380,417	European	finn-b-ILD	2021
LCINS	McKay JD	2,355	7,504	7,993,812	European	ebi-a-GCST004747	2017

intercept is constrained to zero, estimates can be biased if even minor pleiotropy exists [18]. To complement IVW, we implemented sensitivity analyses using MR-Egger regression and weighted median approaches [19]. MR-Egger accounts for horizontal pleiotropy under the Instrument Strength Independent of Direct Effect (InSIDE) assumption, offering flexibility in scenarios with pleiotropic effects. The weighted median method provides robust estimates even when up to half of the IVs violate the core assumptions of validity. By integrating these complementary methods, we ensured that our causal estimates remained reliable under varying conditions of pleiotropy or heterogeneity.

Sensitivity analyses: By utilizing IVW and MR Egger regression methods to identify heterogeneity, Cochran's Q statistic was computed to measure the level of heterogeneity [22]. A *P* value lower than 0.05 suggested the existence of heterogeneity, and a random effects model was applied to the MR analysis. The detection of horizontal pleiotropy can be achieved by assessing the intercept of MR-Egger regression [19]. A *P* value less than 0.05 indicated horizontal pleiotropy, which may compromise the reliability of the MR results. The MR-PRESSO global test can be utilized to identify potential outliers [23]. A *P* value lower than 0.05 indicated the existence of an outlier that should be removed before further MR analysis. Furthermore, we performed leave-one-out analyses to assess the impact of individual SNPs on our Mendelian randomization findings [24].

Reverse MR analysis: A reverse MR analysis was conducted to elucidate the causal relationship between ILD and LCINS. SNPs related to LCINS at a threshold of P<5×10e5 were selected as IVs, and we conducted a reverse MR study using the same MR methodology utilized in the forward analysis.

Transcriptomic analyses

Data acquisition: High-throughput sequencing count data from the GSE231693 dataset (20 ILD patients and 20 normal controls) were obtained from the Gene Expression Omnibus (GEO) database. We also acquired RNA sequencing data and clinical details of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) from The Cancer Genome Atlas (TCGA). Following the exclusion of duplicate records and patients lacking follow-up data, we incorporated 85 samples of LCINS (comprising 13 squamous cell carcinoma and 72 adenocarcinoma cases) along with 7 normal tissue RNA sequencing datasets as the training cohort for this study. Additionally, two GEO RNA-sequence datasets (GSE81089 and GSE87340), containing 46 LCINS cases with survival data, were used as an external validation cohort after removing batch effects among GEO datasets via the "Combat" function in the "sva" R package.

Detection of common DEGs: We utilized the "Deseq2" package within R software (version 4.3.6) to identify differentially expressed genes (DEGs) between ILD tissue and normal controls, as well as between LCINS tissue and normal controls. The predetermined criteria for determining DEGs were an adj. *P* value <0.05 and |fold change| ≥2. In addition, a volcano plot was subsequently generated via GraphPad Prism. Venn diagrams were employed to delineate the intersections of genes that are upregulated in both ILD and LCINS, as well as genes that are downregulated in these conditions, with the aim of identifying the genes that are commonly expressed.

Functional enrichment and protein-protein interaction: We further explored the biological functions of the DEGs via Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation via the R package "ClusterProfiler". We utilized the STRING database (https://string-db.org) for predicting protein-protein interaction (PPI) networks. Cytoscape was employed to visualize and further experimentally investigate the PPI network. Additionally, we used the cytoHubba plugin to identify critical hub genes.

Development and verification of the ILD-related prognostic model: The LASSO algorithm, implemented via the 'glmnet' R package, was employed to resolve redundancy issues among the 50 hub genes identified via PPI network analysis in the ILD-related DEGs from the TCGA dataset. The optimal set of DEGs that was associated with the prognoses of LCINS patients was determined by compressing the regression coefficients. We constructed an ILD-related prognostic model (ILD risk score) composed of five genes (CD1A, CDH3, KRT6B, MMP1 and MMP10). The ILD risk score was defined as follows: Risk score = $\sum_i$ Coefficient(i) × Expression of gene(i). The TCGA-LCINS cohort was separated into low- and high-ILD risk groups in light of the median ILD risk score, and Kaplan-Meier analysis was employed to compare the OS between these two subgroups. Multivariate Cox regression analysis was employed to assess the discriminative capacity of the clinical variables and the ILD risk score. Receiver operating characteristic (ROC) analysis combined with the area under the curve (AUC) were used to evaluate the prognostic performance of the ILD risk score. Additionally, a validation set containing 46 cases from the GEO dataset was used to validate the predictive value of the ILD risk score. The same formula derived from the TCGA cohort was used to estimate the ILD risk scores for the LCINS patients in the GEO cohort, which were then divided into two subsets, low ILD risk and high ILD risk, derived from the same cutoff value from the training set.

Nomogram based on the ILD score: We utilized the "rms" R package to construct a nomogram that integrates survival time, survival status, and 7 features obtained from the multivariate Cox analysis. This was done to evaluate the prognostic significance of these features in 85 samples from the TCGA. This allows for the estimation of the likelihood of OS at 1, 3, and 5 years. The calibration curve serves as an indicator of agreement between the predicted probabilities from the graph and the observed probabilities.

Pathways and molecular mechanisms: A collection of 50 hallmark gene sets (h.all.v7.4.symbols.gmt) was obtained from the Molecular Signatures Database (MSigDB) (http://www.gsea-msigdb.org/gsea/downloads.jsp). The "clusterProfiler" package was used to conduct gene set enrichment analysis (GSEA) on all genes in the high- and low-ILD risk groups, aiming to explore the functional and pathway variances between the two groups. The gene set sizes ranged from 5 to 5000, and one thousand resampling iterations were performed. If the *P* value was less than 0.05 and the NES was greater than 1, the result was considered statistically significant. Additionally, we conducted GSVA enrichment analysis on the genes from samples in the high- and low-ILD risk groups and employed the "limma" package to compare the differences in GSVA scores between these two groups. Furthermore, Spearman correlation analysis was performed to investigate the associations between the ILD risk scores and pathway enrichment scores.

Estimation of immune infiltration cells: The ESTIMATE algorithm was employed to compute the stromal score and immune score, allowing for the estimation of tumor purity by assessing the levels of stromal cell and immune cell infiltration [25]. The CIBERSORT algorithm utilizes support vector regression principles and is widely employed for assessing the presence of various immune cell types within the tumor microenvironment by deconvoluting the expression matrix of immune cell subtypes. This matrix consists of 547 biomarkers and can distinguish 22 human immune cell phenotypes [26]. In this study, the CIBERSORT algorithm was used to process the LCINS data from TCGA to estimate the relative proportions of 22 infiltrating immune cells in tumor tissue. The Wilcoxon test was used to assess the difference in immune cell abundance between the high- and low-ILD risk groups. Furthermore, Spearman correlation analysis was used to evaluate the correlations between the risk scores and the expressions of the 5 prognostic genes with immune cell abundance.

Statistics

In this study, the data analyses were conducted via R software (version 4.3.6). The MR study was performed following recommended guidelines, and all MR analyses utilized the "Two-Sample MR" and "MR-PRESSO" packages. We calculated odds ratios (ORs) and hazard ratios (HRs) with corresponding 95% confidence intervals (CIs). A result was considered statistically

exposure	outcome	method	nsnp	b	se	pval	OR(95%CI)
ILD	LCINS	MR Egger	75	0.092	0.033	0.007	1.096(1.028-1.169)
ILD	LCINS	Weighted median	75	0.093	0.035	0.007	1.098(1.025-1.176)
ILD	LCINS	IVW	75	0.081	0.022	2E-04	1.084(1.038-1.132)
ILD	LCINS	Simple mode	75	0.004	0.072	0.951	1.004(0.872-1.156)
ILD	LCINS	Weighted mode	75	0.099	0.039	0.012	1.104(1.024-1.191)
LCINS	ILD	MR Egger	95	0.02	0.047	0.666	1.02(0.931-1.118)
LCINS	ILD	Weighted median	95	0.031	0.031	0.315	1.031(0.971-1.095)
LCINS	ILD	IVW	95	0.013	0.022	0.574	1.013(0.969-1.058)
LCINS	ILD	Simple mode	95	0.052	0.076	0.497	1.053(0.907-1.222)
LCINS	ILD	Weighted mode	95	0.071	0.069	0.308	1.074(0.937-1.23)

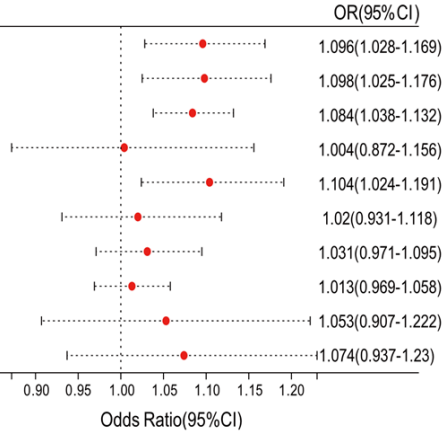

Figure 2. The forest plot displays the causal associations between ILD and LCINS.

significant when the two-sided *p* value was less than 0.05.

Results

Genetic susceptibility to ILD and LCINS

In the forward MR analysis, IVW, MR-Egger regression, the weighted model, and the weighted median indicated a remarkable causal association between ILD and increased risk of LCINS (IVW: OR: 1.084, 95% CI: 1.038-1.132, P=0.0002; MR-Egger: OR: 1.096, 95% CI: 1.028-1.169, P=0.007; weighted mode: OR: 1.104, 95% CI: 1.024-1.191, P=0.012; and weighted median: OR: 1.098, 95% CI: 1.025-1.176, P=0.007). The results of the simple mode also indicated the same direction of causal estimation but did not find a significant correlation between ILD and the occurrence of LCINS (OR: 1.004, 95% CI: 0.872-1.156; P>0.05) (**Figures 2**, **3A**).

Sensitivity analyses for MR estimates

The heterogeneity in IVW and MR Egger regression was assessed via Cochran's Q test, whereas horizontal pleiotropy was identified via the MR Egger intercept. The *p* values for the Cochrane Q test all exceeded 0.05 (Q value for the IVW test: 59.832, P=0.884 and Q value for the MR-Egger test: 59.626, P=0.87), indicating the absence of heterogeneity in the analyses (P>0.05) (**Table 2**). A funnel plot was used to assess the study heterogeneity (**Figure 3B**). Horizontal pleiotropy was not identified by the MR Egger intercept test (P>0.05) (**Table 2**). The outcome of the MR-PRESSO test agreed with the results of the IVW method without outliers, suggesting the reliability of the original results (**Table 2**). The results of the leave-one-out analysis indicated that no single SNP drove the outcomes of the MR analysis. A variety of sensitivity analyses confirmed the reliability of the MR results in this study. Additionally, forest plots illustrating the causal effect on outcome risk for each SNP as an instrumental variable (IV) were included in this study (**Figure 3C**, **3D**).

Results of the reverse MR analysis

A reverse MR analysis revealed that none of the five methods (IVW, MR-Egger regression, weighted median, weighted mode, and simple mode) were significantly correlated with LCINS and ILD risk (**Figures 2**, **4A**). The Cochran's Q test and MR-Egger intercept provided no evidence of heterogeneity or horizontal pleiotropy (P>0.05) (**Table 2**; **Figure 4B**). The leave-one-out analysis did not reveal any SNPs with a significant impact on the overall estimation (**Figure 4C**, **4D**). Overall, these results suggest that there is no significant reverse causality between ILD and LCINS.

Identification of ILD-related DEGs in patients with LCINS

The analysis of the differences revealed that there were 1799 differentially expressed genes (DEGs) in the LCINS cancer tissues compared with the normal adjacent tissues, with 1207 genes being upregulated and 592 genes being downregulated (**Figure 5A**). Compared with those in normal tissues, 1210 DEGs were identified in ILD tissues, with 988 genes upregulat-

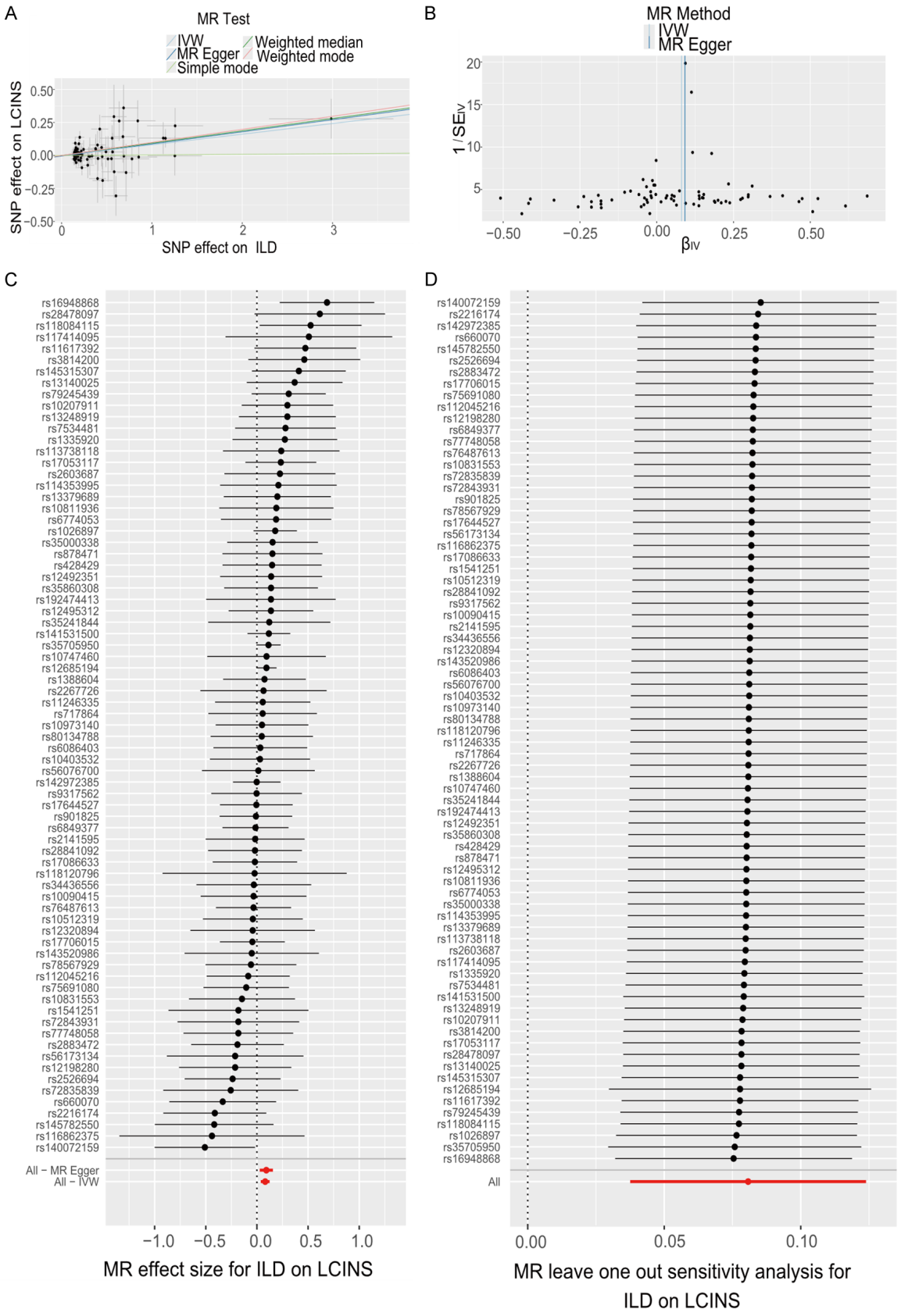

Figure 3. Causality of ILD on LCINS via Mendelian randomization. A. Scatter plot: The slope of each line represents the estimated causal effect derived from the five MR methods. B. Funnel plot: Vertical lines indicate the estimates

for all SNPs. Symmetry in the plot suggests no evidence of horizontal pleiotropy. C. Forest plot: Red points represent the overall causal estimate using all SNPs with the IVW method, while horizontal lines depict the 95% confidence intervals (CIs). D. Leave-one-out analysis: Black points represent the IVW estimates obtained by excluding one SNP at a time, while the red point indicates the IVW estimate using all SNPs.

Table 2. Reults of sensitivity analyses for MR estimates

Exposure	Outcome	Heterogeneity						Pleiotropy		
		MR Egger			IVW			MR Egger		
		Q	Q_df	Q_pval	Q	Q_df	Q_pval	egger_intercept	se	pval
ILD	LCINS	59.626	73	0.87	59.832	74	0.884	-0.004	0.01	0.651
LCINS	ILD	113.1	93	0.077	113.14	94	0.087	-0.002	0.013	0.853

ed and 222 genes downregulated (**Figure 5B**). We identified ILD-related DEGs in LCINS patients through intersection analysis. A total of 262 genes were identified as upregulated, and 61 were recognized as downregulated in both diseases simultaneously, resulting in a total of 323 intersecting genes (**Figure 5C**, **5D**).

Analysis of functional characteristics

We performed GO and KEGG enrichment analyses to gain a deeper understanding of the underlying function of the 323 ILD-DEGs in LCINS. The results of the GO analysis indicated that the DEGs were enriched primarily in cellular components (such as the collagen trimer, cornified envelope, and lamellar body); biological processes (including epidermis development, extracellular matrix organization, extracellular structure organization, and cornification); and molecular function (such as neurotransmitter receptor activity, metallopeptidase activity, and retinol binding) (**Figure 6A**). The results of the KEGG pathway analysis revealed that these genes were enriched primarily in neuroactive ligand - receptor interactions, the PI3K-Akt signaling pathway, and the cAMP signaling pathway (**Figure 6B**).

PPI network and analysis of the hub genes

The PPI network was first created from 323 common DEGs. Next, the Cytuhubba plug-in in Cytoscape was used to compute the top 50 hub genes (namely, KRT6B, KRT6A, DSG3, KRT17, KRT14, KRT16, SPRR1B, IVL, KRT5, DSC3, KRT13, SPRR1A, COL17A1, KRT15, TP63, SERPINB5, SPRR2D, CD19, CD79A, MS4A1, CXCR5, CR2, TNFRSF13C, S100A2, PAX5, CXCL13, FCRLA, MME, FCRL5, VPREB3, IGLL5, FCGR3B, CDH3, IGLL1, SERPINB13, MMP13, MMP3, MMP1, SPP1, MMP7, COL10A1, COMP, COL7A1, CD1A, MMP12, MMP10, MUC5AC, IBSP, KLK6, and COL9A1) (**Figure 7A**).

Construction and verification of the prognostic model

The LASSO regression method was employed to refine 50 hub genes, ultimately selecting the top five predictive genes (CD1A, CDH3, KRT6B, MMP1, and MMP10) for constructing the ILD risk score (**Figure 7B**, **7C**). The ILD risk score was calculated as follows: [-0.104 × Expression of CD1A] + [0.012 × Expression of CDH3] + [0.050 × Expression of KRT6B] + [0.023 × Expression of MMP1] + [0.124 × Expression of MMP10]. Patients in the training set with LCINS were stratified into two subgroups on the basis of their ILD risk scores. Elevated expression levels of four prognostic molecules, with the exception of CD1A, were associated with high ILD risk (**Figure 7D**). The expression levels of all 5 hub genes associated with the ILD risk score were markedly elevated in tumor tissues (**Figure 8A**). These 5 genes can be categorized into two groups: harmful factors (CDH3, KRT6B, MMP1 and MMP10) and a protective factor (CD1A). Elevated expression levels of CDH3, KRT6B, MMP1, and MMP10 were significantly associated with poor survival outcomes (CDH3: P=2.0e-3, KRT6B: P=3.9e-5, MMP1: P=2.3e-3, and MMP10: P=9.2e-4). Patients with elevated expression levels of CD1A, in turn, demonstrated significant improvements in survival (P=5.8e-4) (**Figure 8B-F**). The multivariate Cox regression analysis revealed that the ILD risk score was an independent prognostic factor for patients with LCINS (HR 1.722, 95% CI: 1.284-2.309; P= 0.0003) (**Figure 9A**). Compared with the low-ILD risk group, the high-ILD risk group exhibited poorer overall survival rates (P=2.5e-4) (**Figure 9B**). Furthermore, the ROC curve suggested

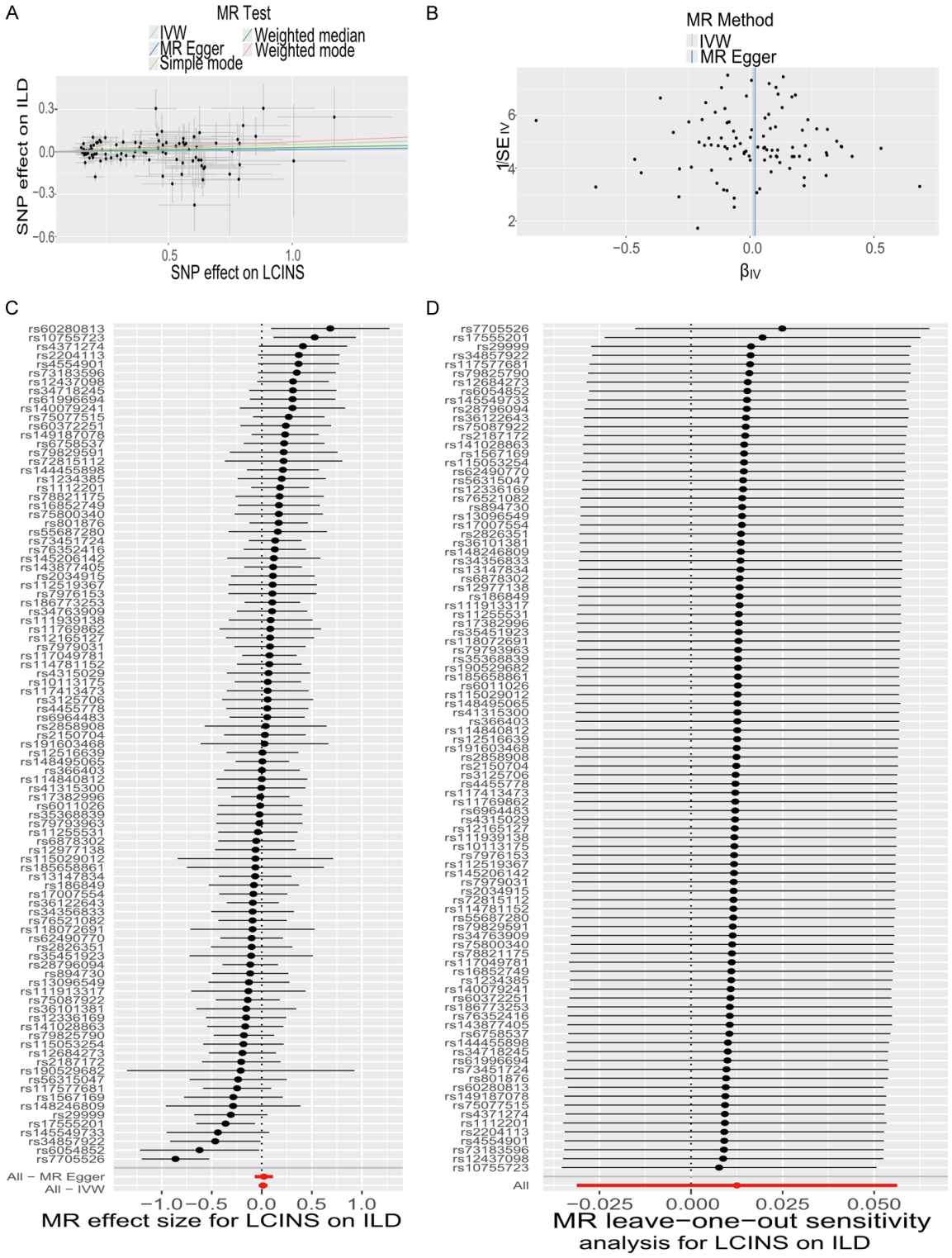

Figure 4. Causality of LCINS on ILD via Mendelian randomization. A. Scatter plot: The slope of each line represents the estimated causal effect derived from five MR methods. B. Funnel plot: Vertical lines indicate the estimates for all SNPs. Symmetry in the plot suggests no evidence of horizontal pleiotropy. C. Forest plot: Red points represent the overall causal estimate using all SNPs with the IVW method, while horizontal lines depict the 95% confidence intervals (CIs). D. Leave-one-out analysis: Black points represent the IVW estimates obtained by excluding one SNP at a time, while the red point indicates the IVW estimate using all SNPs.

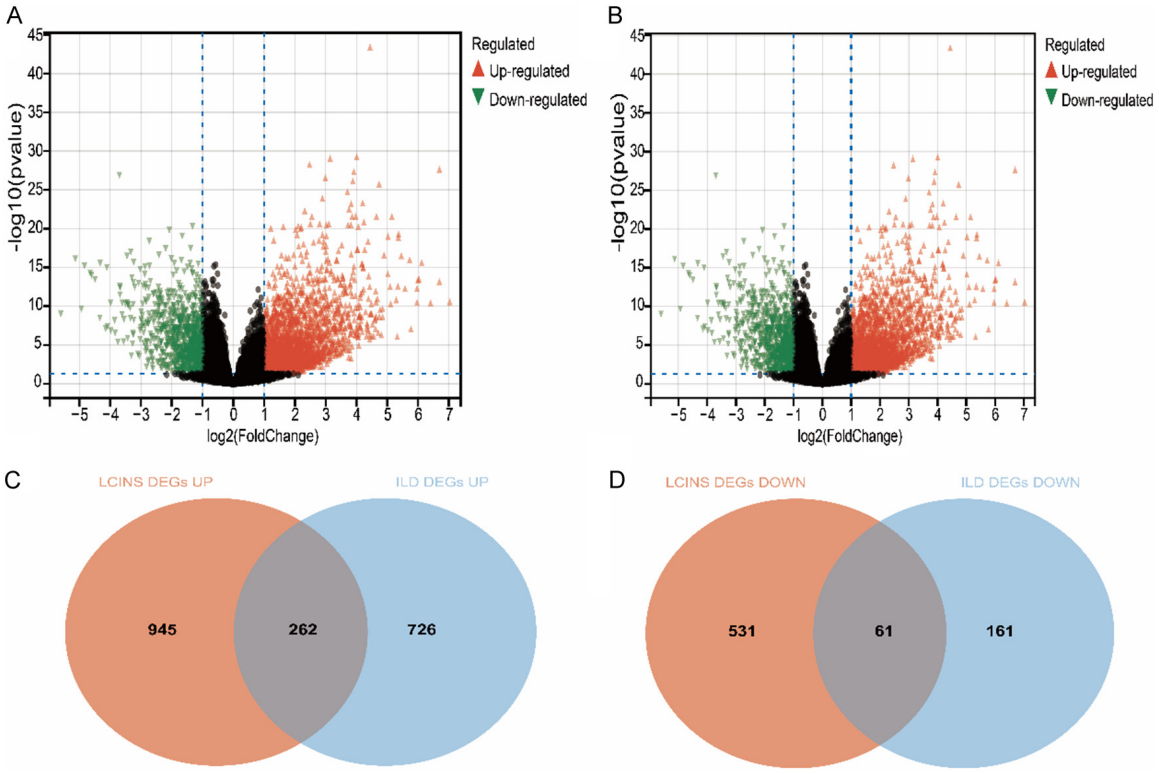

Figure 5. Identification of differentially expressed genes (DEGs). A. Volcano plot illustrating DEGs in LCINS from the TCGA dataset. Green dots represent downregulated genes, while red dots indicate upregulated genes. B. Volcano plot showing DEGs in ILD from the GEO dataset. Green dots represent downregulated genes, and red dots indicate upregulated genes. C. Venn diagram displaying intersecting upregulated genes between LCINS and ILD. D. Venn diagram displaying intersecting downregulated genes between LCINS and ILD.

that ILD risk scores have the potential to serve as a sensitive indicator for predicting OS in individuals with LCINS (AUC at 1 year: 0.82, AUC at 3 years: 0.77, and AUC at 5 years: 0.88) (**Figure 9C**). Further validation in the GEO cohort confirmed that the ILD risk scores obtained from the TCGA database were related to poor prognoses in LCINS patients (P=0.01) (**Figure 9D**). The nomogram was subsequently used to evaluate the prognostic significance of 7 features in 85 TCGA samples. The model showed strong predictive ability for the 1-, 3-, and 5-year overall survival rates in LCINS patients, with a C-index of 0.746 (95% CI: 0.628-0.863, *p* value =4.281e-05) (**Figure 9E**). The calibration curve demonstrated a strong correlation between the predicted values of the model and the observed values (**Figure 9F**).

Molecular pathway for the risk model

The GSEA results revealed significant enrichment of the MTORC1_SIGNALING, GLYCOLYSIS, G2M_CHECKPOINT, MYC_TARGETS_V1, E2F_TARGETS, MYC_TARGETS_V2, and ESTROGEN_RESPONSE_LATE pathways in the high-risk group (**Figure 10A-G**). The GSVA analysis indicated that the primary differences between the two cohorts were concentrated in the G2M_CHECKPOINT, E2F_TARGETS, MTORC1_SIGNALING, CHOLESTEROL_HOMEOSTASIS, GLYCOLYSIS, ESTROGEN_RESPONSE_LATE and MYC_TARGETS_V2 signaling pathways (**Figure 10H**). Spearman correlation analysis revealed that the ILD risk score was positively correlated with 22 signaling pathways, such as G2M_CHECKPOINT, MTORC1_SIGNALING, E2F_TARGETS, and CHOLESTEROL_HOMEOSTASIS but negatively correlated with the BILE_ACID_METABOLISM pathway. The expressions of MP10, MMP1, KRT6B and CDH3 were positively correlated with most pathways, negatively correlated with the BILE_ACID_METABOLISM pathway, and consistent with the impact of risk scores. In contrast, CD1A had the opposite effect (**Figure 10I**).

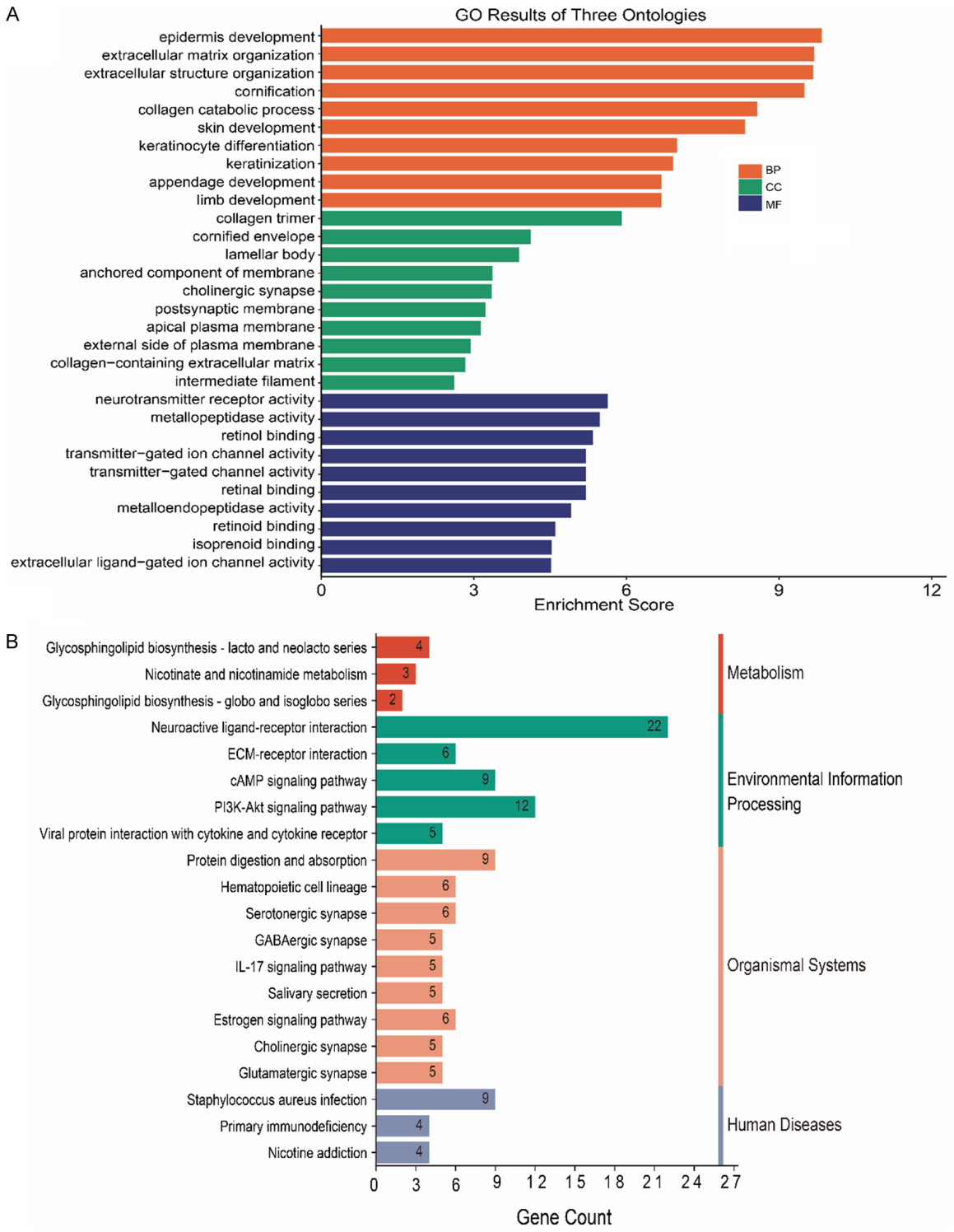

Figure 6. Functional enrichment analysis of the intersected DEGs. A. Gene Ontology (GO) analysis of intersected DEGs, highlighting enrichment in biological processes (e.g., epidermis development, extracellular matrix organization), cellular components (e.g., collagen trimer, lamellar body), and molecular functions (e.g., neurotransmitter receptor activity, metallopeptidase activity). B. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, showing significant enrichment in pathways such as Neuroactive ligand-receptor interaction, PI3K-Akt signaling pathway, and ECM-receptor interaction.

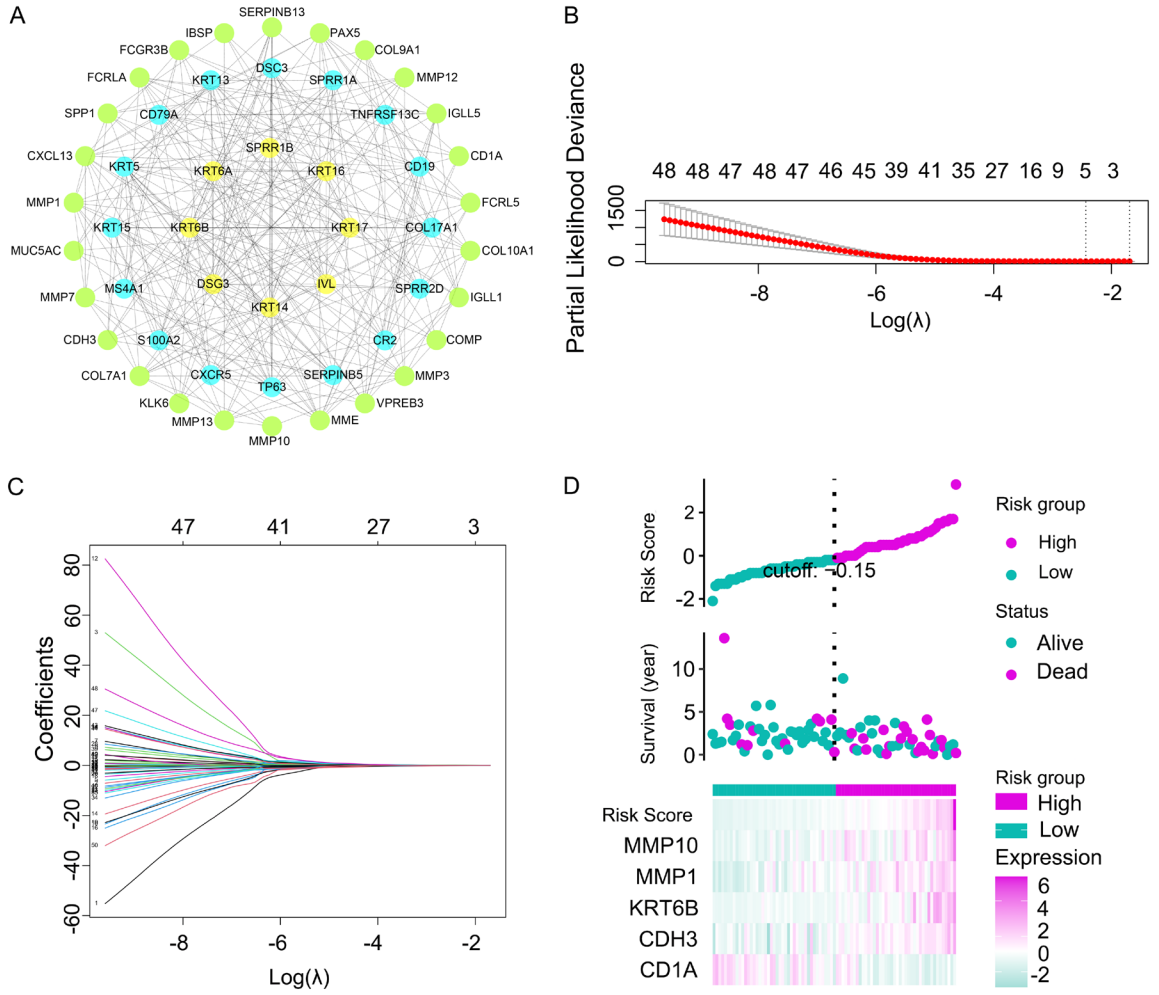

Figure 7. Identification and validation of the ILD risk score model using hub genes. A. Protein-protein interaction (PPI) network depicting interactions among the top 50 hub genes. B. C. LASSO regression analysis to minimize overfitting and refine key ILD-related DEGs for the prognostic model. D. Risk score distribution, survival status scatter plot, and heatmap of gene expression in the prognostic model, with stratification into high- and low-risk groups based on the ILD risk score.

Relationship between ILD risk and the immune microenvironment

The findings from the ESTIMATE analysis suggest that there are no notable disparities in the stromal score, immune score, or ESTIMATE score between the high- and low-ILD-risk groups (**Figure 11A**). The CIBERSORT results revealed differential levels of tumor-infiltrating immune cells between the low- and high-ILD risk groups. Specifically, in the high-ILD risk group, there were lower levels of memory B cells, activated NK cells, monocytes, resting dendritic cells, and resting mast cells, while higher levels of resting NK cells and M0 macrophages were observed (**Figure 11B**). Spearman correlation analysis revealed that the ILD risk score was negatively correlated with the proportions of memory B cells, activated NK cells, monocytes, resting dendritic cells and resting mast cells; moreover, it was positively correlated with the proportions of resting NK cells, M0 macrophages, M1 macrophages and plasma cells. The expressions of MP10, MMP1, KRT6B, and CDH3 were positively correlated with the proportions of resting NK cells and M0 macrophages. Conversely, the expression of CD1A was negatively correlated with these two cell types but positively correlated with the proportions of memory B cells, activated NK cells, monocytes, resting dendritic cells, and resting mast cells (**Figure 11C**).

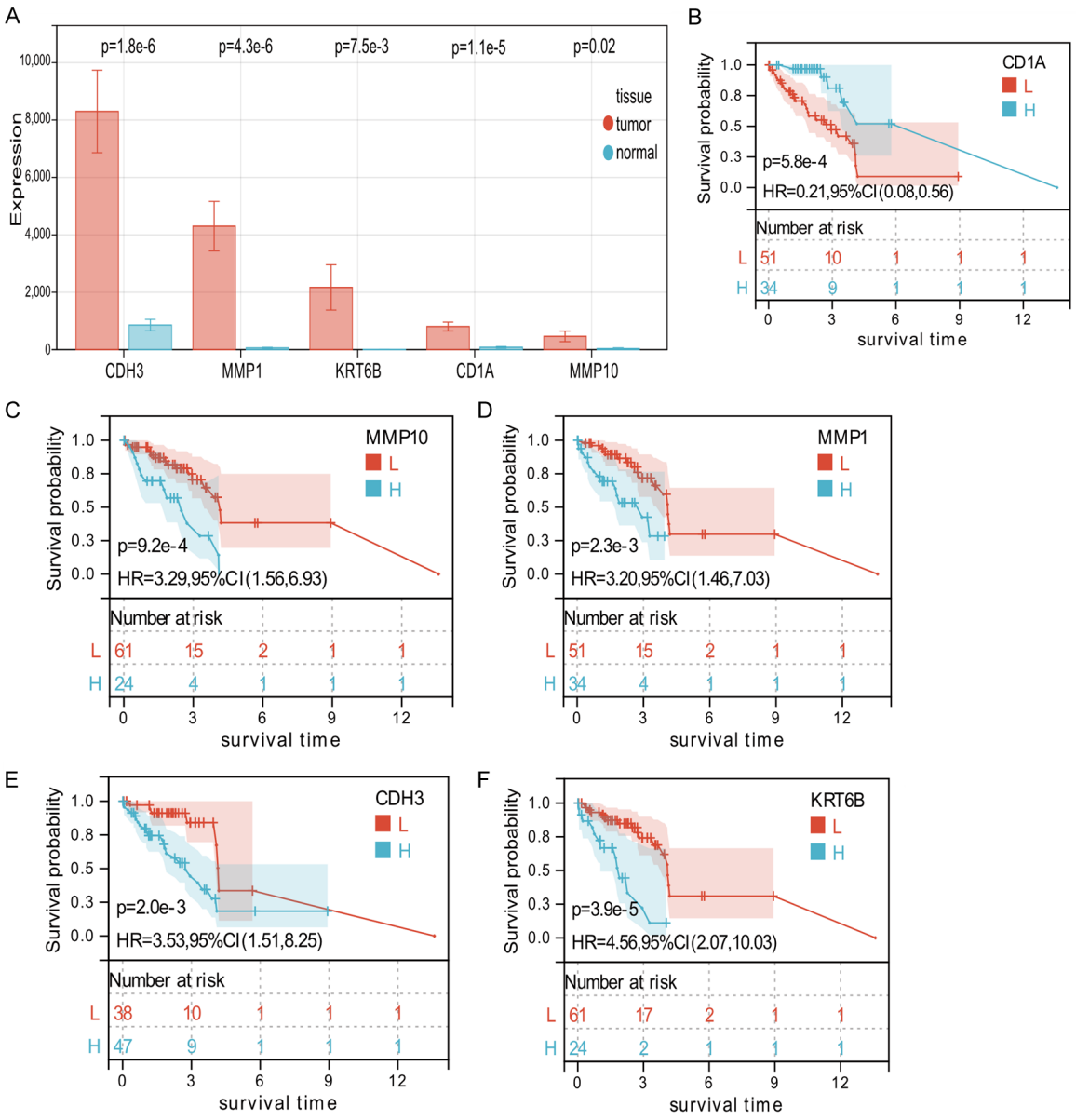

Figure 8. Expression of ILD-related genes and their impact on overall survival. (A) Comparison of expression levels of the five prognostic genes (CDH3, KRT6B, MMP1, MMP10, and CD1A) in normal tissues and cancer samples. (B-F) Kaplan-Meier survival curves showing the association between high (red) and low (blue) expression levels of the prognostic genes and overall survival. Each panel corresponds to a specific gene: (B) CD1A, (C) MMP10, (D) MMP1, (E) CDH3, and (F) KRT6B.

Discussion

The factors affecting LCINS are complex. In addition to genetic mutations, preexisting lung disease is a common risk factor [11]. Currently, there is no substantial evidence to suggest a direct link between ILD and lung cancer occurrence. To minimize potential confounding variables and reverse causality, this MR analysis aims to assess the causal relationship between ILD and LCINS. The findings of our MR analysis suggest that genetic susceptibility to ILD is causally associated with lung cancer in non-smoking European populations. Additionally, ILD-related DEG risk markers (ILD risk scores) can be used to categorize LCINS patients into two groups and predict clinical outcomes. Additional validation confirmed the correlation

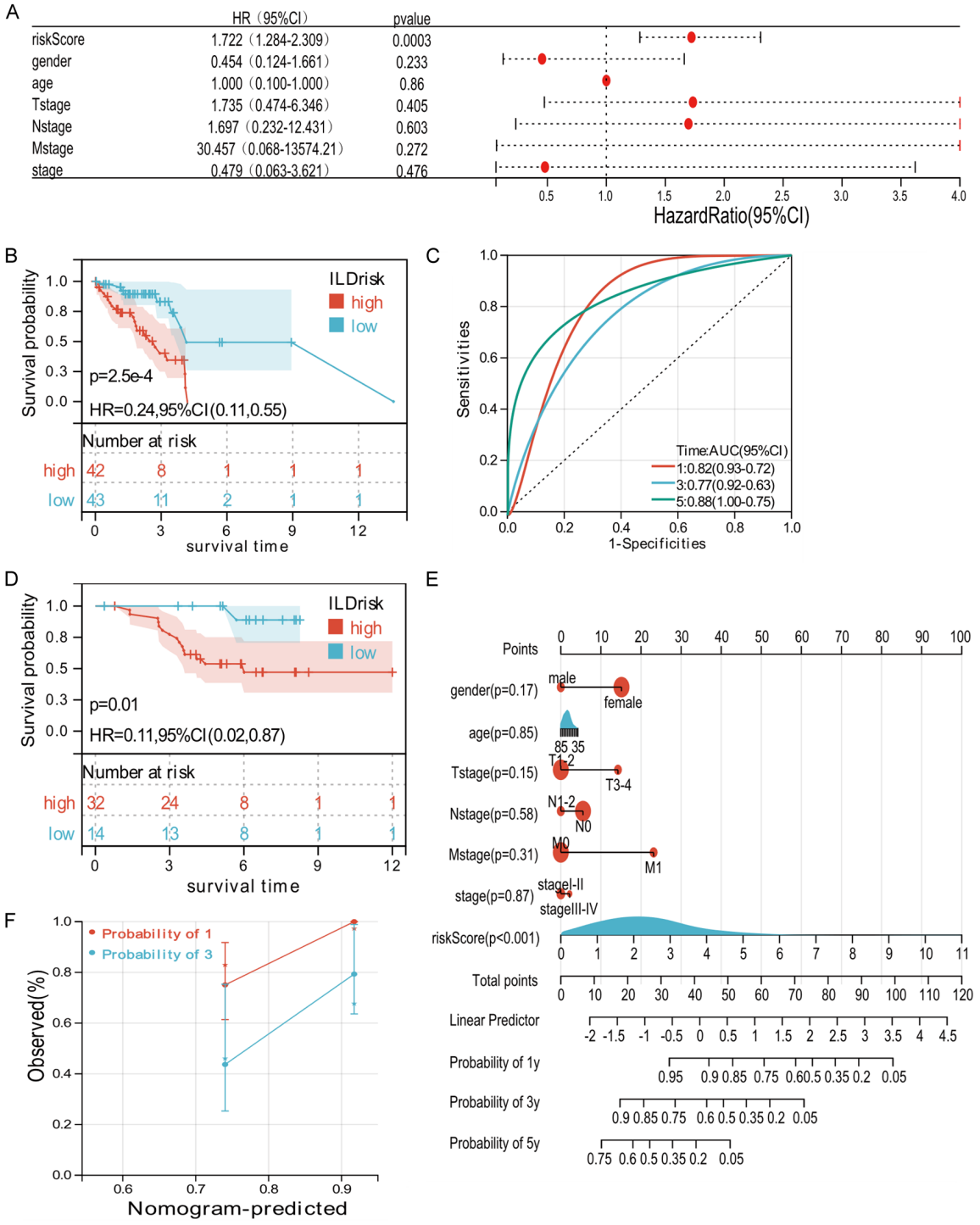

Figure 9. Prognostic analysis and validation of the ILD risk score model. A. Multivariate Cox regression analysis showing that the ILD risk score is an independent prognostic factor for overall survival. B. Kaplan-Meier survival analysis for high- and low-ILD risk score groups in the TCGA cohort. Red indicates the high-risk group, while blue represents the low-risk group. C. ROC curve demonstrating the predictive accuracy of the ILD risk score for 1-year (AUC=0.82), 3-year (AUC=0.77), and 5-year (AUC=0.88) overall survival. D. Kaplan-Meier survival analysis for high- and low-risk groups in the GEO cohort. Blue represents the low-risk group, and red indicates the high-risk group. E. Nomogram constructed to predict 1-, 3-, and 5-year overall survival in patients with LCINS. F. Calibration plot showing the agreement between predicted and observed survival probabilities at 1 and 3 years, indicating moderate predictive performance of the nomogram.

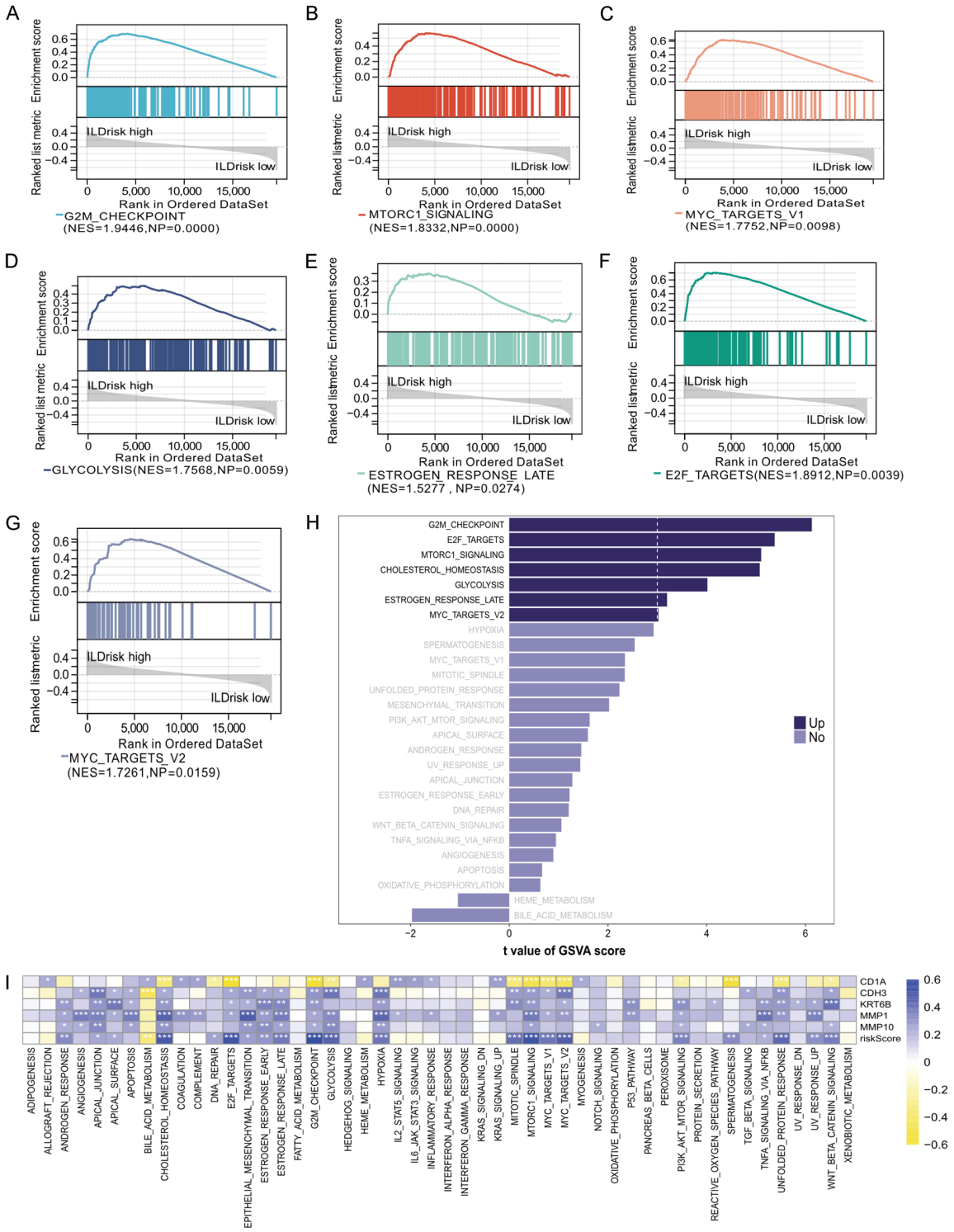

Figure 10. Correlations between signal transduction pathways and ILD risk. A-G. GSEA analysis of high and low ILD risk score groups, highlighting pathways such as G2M_CHECKPOINT, MTORC1_SIGNALING, MYC_TARGETS_V1, GLYCOLYSIS, ESTROGEN_RESPONSE_LATE, E2F_TARGETS, and MYC_TARGETS_V2. H. GSVA analysis comparing signaling pathway enrichment scores between high and low ILD risk groups. I. Spearman correlation analysis illustrating the association between 50 hallmark gene sets and prognostic genes. Purple represents a positive correlation, while yellow indicates a negative correlation. Significance levels are denoted as ***, P<0.001; **, P<0.01; *, P<0.05.

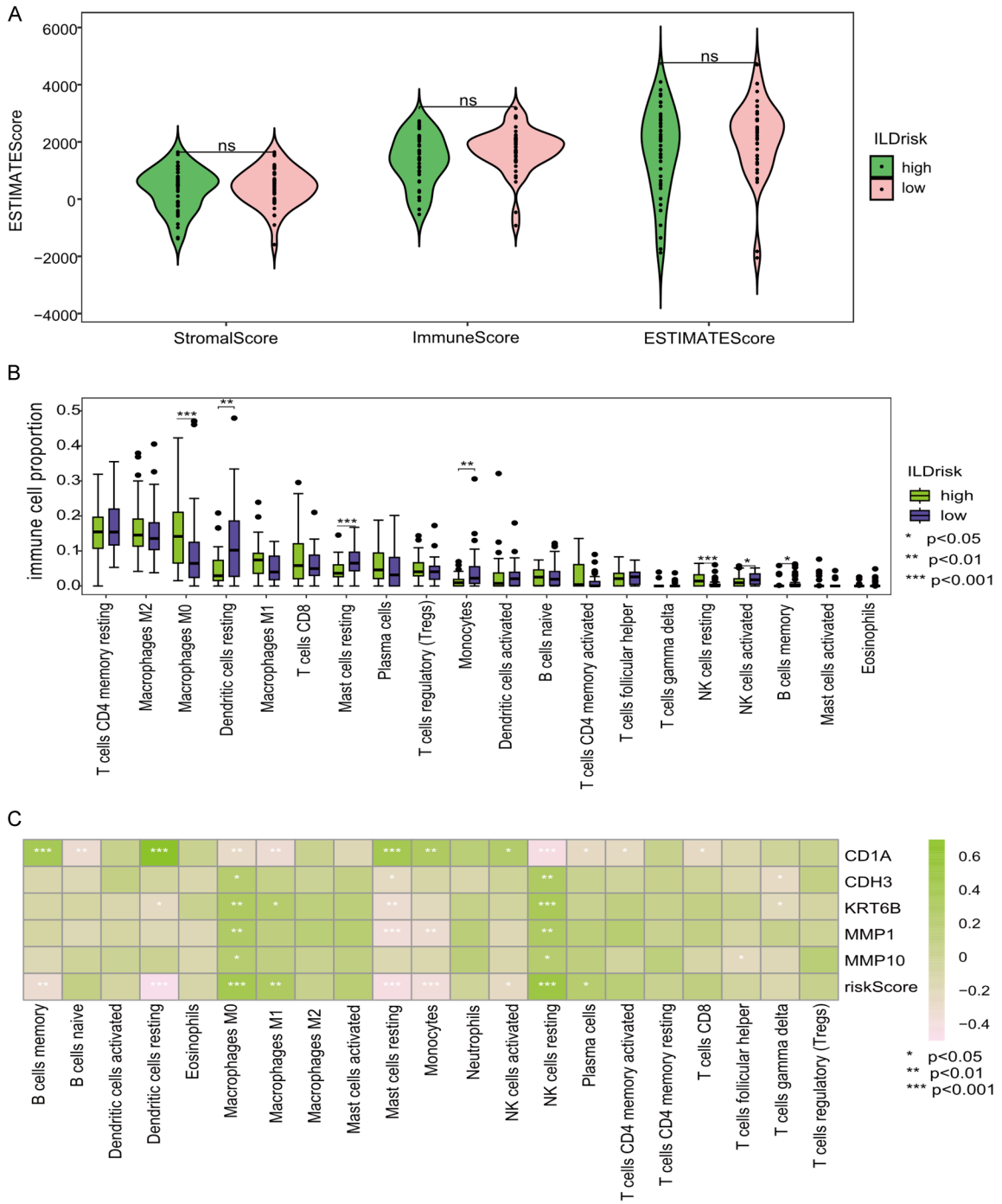

Figure 11. Associations between the tumor microenvironment (TME) and ILD risk scores. A. Comparison of ESTIMATE scores (stromal score, immune score, and total ESTIMATE score) between high and low ILD risk groups. B. Proportion of immune cell types compared between high and low ILD risk groups as determined by the CIBERSORT algorithm. Green bars represent the high ILD risk group, while purple bars represent the low ILD risk group. Significance levels: ***, P<0.001; **, P<0.01; *, P<0.05. C. Spearman correlation analysis between 21 immune cell types and the five key genes in the ILD risk model. Green indicates a positive correlation, while pink indicates a negative correlation. Significance levels: ***, P<0.001; **, P<0.01; *, P<0.05.

between ILD risk and hallmark signaling pathways and the level of tumor-infiltrating immune cells.

Our study revealed a robust genetic association between these two diseases, suggesting that poor prognoses in LCINS patients are associat-

ed with a high ILD risk, which aligns with findings from previous retrospective analyses. Sara Tomassetti and her colleagues [27] conducted a retrospective study of 260 patients with IPF, in which the incidence of lung cancer was 13%. The OS of LC-IPF patients was significantly shorter than that of IPF patients without lung cancer (mOS 38.7 m vs. 63.9 m; HR 5.0; 95% CI: 2.91-8.57; P<0.001). In the study group, the causes of death included respiratory failure (43%), progression of lung cancer (13%), and complications related to lung cancer treatment (17%). A study conducted in China [28] examined the characteristics, clinical and pathological features, and prognoses of non-small cell lung cancer (NSCLC) patients with interstitial lung abnormalities (ILA). During the study, ILA was detected in 101 out of 765 patients (13.2%) when they were diagnosed with NSCLC. The analysis revealed a significant association between the presence of ILA in NSCLC patients and a shorter OS period (751 days vs. 445 days, HR 0.6, P=0.001). Additionally, individuals with usual interstitial pneumonia (UIP) had a shorter OS than did those without UIP (HR 1.82, P=0.037). Patients with lung cancer and ILD who received platinum-based doublet chemotherapy had significantly shorter median progression-free survival (PFS) and overall survival (OS) times than did those without ILD (mPFS: 3.0 m vs. 7.0 m, P<0.001; mOS: 7.0 m vs. 15.0 m, P<0.001, respectively) [29]. The median survival times were not significantly different between the IPF-LC and non-IPF ILD-LC groups (26 m vs. 20 m, P=0.530) [30]. Currently, there is a lack of statistical data on the relationship between ILD and LCINS. In our study, we constructed an ILD risk score on the basis of ILD-related DEGs and constructed a nomogram to predict the prognoses of patients with LCINS. The results indicated that the high-ILD risk group had poor OS, and the ILD risk score was determined to be an independent prognostic factor for LCINS (HR 1.722, 95% CI: 1.284-2.309; P=0.0003). The nomogram and ROC curves exhibited robust predictive power for 1-, 3-, and 5-year OS in LCINS patients.

Currently, there is a lack of laboratory research that has investigated the molecular mechanisms by which ILD impacts LCINS. We identified 323 ILD-related differentially expressed genes (DEGs) in the LCINS samples, which are enriched in KEGG pathways, including neuroactive ligand - receptor interactions, the PI3K-Akt signaling pathway, and the cAMP signaling pathway. The top 50 genes with close interactions were selected from the PPI network in LCINS. By employing the LASSO algorithm, a 5-gene risk model related to ILD was developed in this study for predicting the prognoses of patients with LCINS. These genes can be categorized into two groups: harmful factors (CDH3, KRT6B, MMP1, and MMP10) and a protective factor (CD1A). Several experiments have demonstrated the roles of these five genes in cancer. MMP1 and MMP10 are members of the matrix metalloproteinase (MMP) family. In addition to their involvement in extracellular matrix remodeling and cancer cell migration, MMPs also play a role in regulating the signaling pathways that control cell growth, inflammation or angiogenesis and may even function through nonproteolytic mechanisms [31]. CDH3 is upregulated in LUAD tissue and is associated with poorer OS. Additionally, CDH3 expressions are positively correlated with the infiltration of CD4+ T cells, Treg cells and exhausted T cells but negatively correlated with B-cell infiltration. Furthermore, downregulation of CDH3 has been shown to inhibit cell proliferation and migration [32]. KRT6B expressions are elevated in bladder cancer and are inversely correlated with the infiltration of B cells and macrophages [33]. PDE3B can downregulate KRT6B expression, thereby suppressing the invasion and migration of bladder cancer cells [34]. CD1a molecules are capable of presenting glycolipid antigens, and the expression of CD1a on dendritic cells may play a crucial role in presenting tumor-derived glycolipid antigens to T cells. This process can lead to the generation of effective antitumor responses, potentially improving the prognoses of cancer patients [35].

There are many pathogenic similarities between ILD and lung cancer, with abnormal activation of multiple signaling pathways, such as the Wnt/β-catenin, transforming growth factor-β, phosphoinositide 3-kinase (PI3K)/protein kinase B and tyrosine kinase pathways, which are collectively involved in the pathogenesis of IPF and lung cancer [36-38]. Abnormally activated pathways overexpress their target genes and have been implicated in cancer invasion, lung remodeling, and epithelial-mesenchymal transition. In our study, we demonstrat-

ed that numerous pathways, including MTORC1_SIGNALING, GLYCOLYSIS, G2M_CHECKPOINT, MYC_TARGETS_V1, E2F_TARGETS, MYC_TARGETS_V2 and ESTROGEN_RESPONSE_LATE, were significantly enriched in the high-ILD risk group. In addition, we found a positive correlation between the ILD risk scores and the G2M_CHECKPOINT, MTORC1_SIGNALING, E2F_TARGETS, CHOLESTEROL_HOMEOSTASIS, MYC_TARGETS_V2, GLYCOLYSIS, UNFOLDED_PROTEIN_RESPONSE, MITOTIC_SPINDLE, HYPOX IA, MYC_TARGETS_V1, ESTROGEN_RESPONSE_LATE, WNT_BETA_CATENIN_SIGNALING, and PI3K_A KT_MTOR_SIGNALING pathways. On the basis of these findings, ILD may contribute to the development and progression of LCINS by modulating the cell cycle, proliferation, and glucose metabolism pathways.

The inflammatory microenvironment may promote the formation and progression of lung cancer [39], and fibrotic alterations in pulmonary fibrosis result in an overabundance of collagen and other components of the extracellular matrix, causing tissue restructuring and the formation of scars. This environment may facilitate the proliferation of cancer cells [37, 40]. In our study, the ESTIMATE results revealed no differences in the tumor stromal score or immune score between the high- and low-ILD risk groups. Moreover, we observed a significant inverse relationship between the ILD risk score and several TME infiltrates, such as B-cell memory, activated NK cells, monocytes, resting dendritic cells, and resting mast cells, while a positive correlation with resting NK cells, M0 macrophages, M1 macrophages, and plasma cells was detected. The results of our analysis suggest that ILD contributes to the occurrence and development of LCINS through its impact on immune cell infiltration rather than the level of stromal cells.

The use of MR in this research provides a significant advantage over traditional observational studies by reducing bias and reversing causality. Our findings have important implications for the prevention and treatment of ILD and LCINS. As the number of ILD and LCINS cases is increasing, understanding their causal relationship could lead to targeted interventions aimed at reducing the risk of lung cancer in nonsmoking ILD patients.

Our research has several limitations. First, our results are specific to the European population and may not be applicable to other ethnic populations. In addition, while MR can significantly reduce confounding factors, it cannot eliminate pleiotropy, which refers to genetic variations that influence outcomes through pathways unrelated to the exposure. Fortunately, the assessments of horizontal pleiotropy and sensitivity in this research produced dependable and consistent outcomes, with no indication of heterogeneity detected, validating the conclusions drawn from the MR analysis. The directions and magnitudes of the MR estimates in the IVW, weighted median, and MR-Egger methods were consistent with each other. Third, the results of the bidirectional MR method may be impacted by the availability and quality of GWAS data, which could influence their reliability. Finally, while a genetic association between ILD and an increased risk of lung cancer has been demonstrated in nonsmoking European populations, our results are currently limited to data analyses, and there are insufficient experimental data to confirm our findings. Therefore, it is necessary to carry out reasonable assays to verify our conjecture step by step.

Conclusion

The results of the bidirectional MR study indicate that ILD may have a positive causal effect on LCINS. A risk score model was developed in this study to accurately predict the prognoses of patients with LCINS and gain insight into the underlying molecular mechanisms. These findings enhance our understanding of the interactions between ILD and LCINS, potentially identifying targets for personalized treatments. Nevertheless, additional experimental validations and clinical studies are necessary to confirm these findings and address any potential limitations of the study.

Acknowledgements

This work was funded in part by the following: National Natural Science Foundation of China (82002414 to Xian Sun); Basic and Applied Basic Research Foundation of Guangdong Province (2023A1515010444 to Xian Sun); Education Department of Guangdong Province Clinical teaching base reform research project

(2021JD027 to Xian Sun); Shenzhen Natural Science Foundation Basic Research Surface Project (JCYJ20220530144811027 to Xian Sun); Shenzhen Science and Technology Program (JCYJ20240813150437049 to Xian Sun); Shenzhen Key Laboratory of Chinese Medicine Active substance screening and Translational Research (ZDSYS2022060610-0801003); National Science and Technology Council, Taiwan (NSTC 112-2320-B-039-012-MY3 to Wei-Jan Wang); China Medical University in Taiwan (CMU113-MF-69 to Wei-Jan Wang); National Science and Technology Council Taiwan (NSTC 113-2639-B-039-001 -ASP and T-Star Center NSTC 113-2634-F-039-001 to Mien-Chie Hung); Ministry of Health and Welfare Taiwan (MOHW114-TDU-B-222-144016 to Mien-Chie Hung); The Featured Areas Research Center Program by the Ministry of Education (MOE to Mien-Chie Hung) in Taiwan.

Disclosure of conflict of interest

None.

Address correspondence to: Xian Sun and Bo Wang, Department of Oncology, The Seventh Affiliated Hospital, Sun Yat-Sen University, Shenzhen, Guangdong, P. R. China. E-mail: sunxian@sysush.com (XS); wangbo@sysush.com (BW); Wei-Jan Wang, Research Center for Cancer Biology, Cancer Biology and Precision Therapeutics Center, China Medical University, Taichung 406, Taiwan. E-mail: cvcsky@cmu.edu.tw

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I and Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024; 74: 229-263.

[2] Cainap C, Pop LA, Balacescu O and Cainap SS. Early diagnosis and screening in lung cancer. Am J Cancer Res 2020; 10: 1993-2009.

[3] Scagliotti GV, Longo M and Novello S. Nonsmall cell lung cancer in never smokers. Curr Opin Oncol 2009; 21: 99-104.

[4] Siegel RL, Miller KD, Wagle NS and Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023; 73: 17-48.

[5] Midha A, Dearden S and McCormack R. EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII). Am J Cancer Res 2015; 5: 2892-2911.

[6] Du H, Liu L, Liu H, Luo S, Patz EF Jr, Glass C, Su L, Du M, Christiani DC and Wei Q. Genetic variants of DOCK2, EPHB1 and VAV2 in the natural killer cell-related pathway are associated with non-small cell lung cancer survival. Am J Cancer Res 2021; 11: 2264-2277.

[7] Li F, Zhao S, Cui Y, Guo T, Qiang J, Xie Q, Yu W, Guo W, Deng W, Gu C, Wu T and Wu T. α1,6-Fucosyltransferase (FUT8) regulates the cancer-promoting capacity of cancer-associated fibroblasts (CAFs) by modifying EGFR core fucosylation (CF) in non-small cell lung cancer (NSCLC). Am J Cancer Res 2020; 10: 816-837.

[8] Tang D, Liu H, Zhao Y, Qian D, Luo S, Patz EF Jr, Su L, Shen S, ChristianI DC, Gao W and Wei Q. Genetic variants of BIRC3 and NRG1 in the NLRP3 inflammasome pathway are associated with non-small cell lung cancer survival. Am J Cancer Res 2020; 10: 2582-2595.

[9] Mori S, Maiguma T, Yoshii K, Moriya Y, Takada R, Shinkai F, Haruki Y, Hashimoto H, Komoto A, Takayanagi K, Tamura K, Okura Y, Sugiyama T and Shimada K. Effect of the thyroid transcription factor 1 expression and treatment discontinuation due to adverse events on progression-free survival in patients with advanced non-squamous non-small cell lung cancer treated with pembrolizumab plus pemetrexed and platinum chemotherapy: a Japanese four-hospital, retrospective study. Am J Cancer Res 2024; 14: 3852-3858.

[10] Lan Q, Hsiung CA, Matsuo K, Hong YC, Seow A, Wang Z, Hosgood HD 3rd, Chen K, Wang JC, Chatterjee N, Hu W, Wong MP, Zheng W, Caporaso N, Park JY, Chen CJ, Kim YH, Kim YT, Landi MT, Shen H, Lawrence C, Burdett L, Yeager M, Yuenger J, Jacobs KB, Chang IS, Mitsudomi T, Kim HN, Chang GC, Bassig BA, Tucker M, Wei F, Yin Z, Wu C, An SJ, Qian B, Lee VH, Lu D, Liu J, Jeon HS, Hsiao CF, Sung JS, Kim JH, Gao YT, Tsai YH, Jung YJ, Guo H, Hu Z, Hutchinson A, Wang WC, Klein R, Chung CC, Oh IJ, Chen KY, Berndt SI, He X, Wu W, Chang J, Zhang XC, Huang MS, Zheng H, Wang J, Zhao X, Li Y, Choi JE, Su WC, Park KH, Sung SW, Shu XO, Chen YM, Liu L, Kang CH, Hu L, Chen CH, Pao W, Kim YC, Yang TY, Xu J, Guan P, Tan W, Su J, Wang CL, Li H, Sihoe AD, Zhao Z, Chen Y, Choi YY, Hung JY, Kim JS, Yoon HI, Cai Q, Lin CC, Park IK, Xu P, Dong J, Kim C, He Q, Perng RP, Kohno T, Kweon SS, Chen CY, Vermeulen R, Wu J, Lim WY, Chen KC, Chow WH, Ji BT, Chan JK, Chu M, Li YJ, Yokota J, Li J, Chen H, Xiang YB, Yu CJ, Kunitoh H, Wu G, Jin L, Lo YL, Shiraishi K, Chen YH, Lin HC, Wu T, Wu YL, Yang PC, Zhou B, Shin MH, Fraumeni JF Jr, Lin D, Chanock SJ and Rothman N. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet 2012; 44: 1330-1335.

[11] Daylan AEC, Miao E, Tang K, Chiu G and Cheng H. Lung cancer in never smokers: delving into epidemiology, genomic and immune landscape, prognosis, treatment, and screening. Lung 2023; 201: 521-529.
[12] Yoo H, Jeong BH, Chung MJ, Lee KS, Kwon OJ and Chung MP. Risk factors and clinical characteristics of lung cancer in idiopathic pulmonary fibrosis: a retrospective cohort study. BMC Pulm Med 2019; 19: 149.
[13] Hubbard R, Venn A, Lewis S and Britton J. Lung cancer and cryptogenic fibrosing alveolitis. A population-based cohort study. Am J Respir Crit Care Med 2000; 161: 5-8.
[14] Yoon JH, Nouraie M, Chen X, Zou RH, Sellares J, Veraldi KL, Chiarchiaro J, Lindell K, Wilson DO, Kaminski N, Burns T, Trejo Bittar H, Yousem S, Gibson K and Kass DJ. Characteristics of lung cancer among patients with idiopathic pulmonary fibrosis and interstitial lung disease - analysis of institutional and population data. Respir Res 2018; 19: 195.
[15] Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, VanderWeele TJ, Higgins JPT, Timpson NJ, Dimou N, Langenberg C, Golub RM, Loder EW, Gallo V, Tybjaerg-Hansen A, Davey Smith G, Egger M and Richards JB. Strengthening the reporting of observational studies in epidemiology using Mendelian randomization: the STROBE-MR statement. JAMA 2021; 326: 1614-1621.
[16] Emdin CA, Khera AV and Kathiresan S. Mendelian randomization. JAMA 2017; 318: 1925-1926.
[17] Burgess S and Thompson SG; CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. Int J Epidemiol 2011; 40: 755-764.
[18] Burgess S, Scott RA, Timpson NJ, Davey Smith G and Thompson SG; EPIC- InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol 2015; 30: 543-552.
[19] Bowden J, Davey Smith G and Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 2015; 44: 512-525.
[20] Bowden J, Davey Smith G, Haycock PC and Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 2016; 40: 304-314.
[21] Hartwig FP, Davey Smith G and Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol 2017; 46: 1985-1998.
[22] Burgess S, Butterworth A and Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013; 37: 658-665.
[23] Verbanck M, Chen CY, Neale B and Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet 2018; 50: 693-698.
[24] Burgess S, Bowden J, Fall T, Ingelsson E and Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. Epidemiology 2017; 18: 30-42.
[25] Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB and Verhaak RG. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013; 4: 2612.
[26] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M and Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015; 12: 453-457.
[27] Tomassetti S, Gurioli C, Ryu JH, Decker PA, Ravaglia C, Tantalocco P, Buccioli M, Piciucchi S, Sverzellati N, Dubini A, Gavelli G, Chilosi M and Poletti V. The impact of lung cancer on survival of idiopathic pulmonary fibrosis. Chest 2015; 147: 157-164.
[28] Zhu M, Yi J, Su Y, Zhang Y, Gao Y, Xu X, Zhang S, Zhang Y and Huang K. Newly diagnosed non-small cell lung cancer with interstitial lung abnormality: prevalence, characteristics, and prognosis. Thorac Cancer 2023; 14: 1874-1882.
[29] Xiu W, Zheng J, Zhou Y, Du H, Li J, Li W, Zhou F, Zhou C and Wu F. A nomogram for the prediction of the survival of patients with advanced non-small cell lung cancer and interstitial lung disease. Cancer Med 2023; 12: 11375-11384.
[30] Han SJ, Kim HH, Hyun DG, Ji W, Choi CM, Lee JC and Kim HC. Clinical characteristics and outcome of lung cancer in patients with fibrosing interstitial lung disease. BMC Pulm Med 2024; 24: 136.
[31] Kessenbrock K, Plaks V and Werb Z. Matrix metalloproteinases: regulators of the tumor microenvironment. Cell 2010; 141: 52-67.
[32] Ma W and Hu J. Downregulated CDH3 is correlated with a better prognosis for LUAD and decreases proliferation and migration of lung cancer cells. Genes Genomics 2024; 46: 713-731.

[33] Song Q, Yu H, Cheng Y, Han J, Li K, Zhuang J, Lv Q, Yang X and Yang H. Bladder cancer-derived exosomal KRT6B promotes invasion and metastasis by inducing EMT and regulating the immune microenvironment. J Transl Med 2022; 20: 308.
[34] Feng Y, Huang Z, Song L, Li N, Li X, Shi H, Liu R, Lu F, Han X, Ding Y, Ding Y, Wang J, Yang J and Jia Z. PDE3B regulates KRT6B and increases the sensitivity of bladder cancer cells to copper ionophores. Naunyn Schmiedebergs Arch Pharmacol 2024; 397: 4911-4925.
[35] Coventry B and Heinzel S. CD1a in human cancers: a new role for an old molecule. Trends Immunol 2004; 25: 242-248.
[36] Tzouvelekis A, Gomatou G, Bouros E, Trigidou R, Tzilas V and Bouros D. Common pathogenic mechanisms between idiopathic pulmonary fibrosis and lung cancer. Chest 2019; 156: 383-391.
[37] Drakopanagiotakis F, Krauss E, Michailidou I, Drosos V, Anevlavis S, Gunther A and Steiropoulos P. Lung cancer and interstitial lung diseases. Cancers (Basel) 2024; 16: 2837.
[38] Kinoshita T and Goto T. Molecular mechanisms of pulmonary fibrogenesis and its progression to lung cancer: a review. Int J Mol Sci 2019; 20: 1-16.
[39] O'Callaghan DS, ODonnell D, O'Connell F and O'Byrne KJ. The role of inflammation in the pathogenesis of non-small cell lung cancer. J Thorac Oncol 2010; 5: 2024-2036.
[40] Selman M and Pardo A. Role of epithelial cells in idiopathic pulmonary fibrosis: from innocent targets to serial killers. Proc Am Thorac Soc 2006; 3: 364-372.