IVYSPRING INTERNATIONAL PUBLISHER

International Journal of Medical Sciences

2025; 22(15): 4036-4048. doi: 10.7150/ijms.107109

Research Paper

Plasma Proteome Profiling Identifies Biomarkers and Potential Drug Targets for Non-Small Cell Lung Cancer

Minghui Zhao[1], Xiaoke Di[1], Yucui Zhao[2]✉

1. Department of Radiation Oncology, The First Affiliated Hospital with Nanjing Medical University, Nanjing, China.
2. Department of Radiation Oncology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China.

✉ Corresponding author: Yucui Zhao, Department of Radiation Oncology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. Email: 2321130@zju.edu.cn.

Received: 2024.11.07; Accepted: 2025.08.25; Published: 2025.09.12

Abstract

Non-small cell lung cancer (NSCLC), as one of the most commonly diagnosed cancers globally, requires expedited identification of new drug targets. We conducted proteome-wide MR using genetic data for 4,853 plasma proteins. Summary-level data on lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) were extracted from GWAS meta-analyses (11,273 and 7,426 cases, respectively) and FinnGen cohort (1,590 and 1,510 cases, respectively). We genetically identified eight proteins with a causal role in the etiology of NSCLC. Lower levels of five proteins (CDH17, CXADR, FAM3D, POGLUT3, SFTPB) and higher levels of two proteins (CEACAM5, KLK1) were linked to increased LUAD risk, while higher CD14 levels were associated with elevated LUSC risk. Two proteins, POGLUT3 and SFTPB were validated through Bayesian colocalization. One protein SFTPB was identified using SMR and HEIDI tests. Bidirectional MR found no reverse causality. The primary findings were validated through scRNA-seq, GeneMANIA, GO analysis, druggability assessments and PheWAS analysis. These protein-coding genes are primarily expressed in epithelial cells, macrophages, monocytes, and endothelial cells. Furthermore, CEACAM5, KLK1, and CD14 correspond to existing drugs. These proteins may deepen our comprehension of the etiology and could serve as appealing novel biomarkers and drug targets for NSCLC management.

Keywords: proteome-wide Mendelian randomization, biomarker, drug target, non-small cell lung cancer.

Introduction

Lung cancer is the most frequently diagnosed malignancy globally and is the leading cause of cancer-related death. Annually, it is estimated that around 2.5 million new cases arise and over 1.8 million fatalities occur [1]. Approximately 80% to 85% of lung cancers are non-small cell lung cancer (NSCLC), with lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) being the two predominant subtypes. LUAD has become the most common subtype, with incidence rates exceeding those of LUSC in recent years [2]. Most patients are diagnosed at an advanced stage, where curative treatment options are limited, resulting in a 5-year survival rate that hovers just above 20% [3, 4]. Therefore, there is an urgent need to discover sensitive and specific markers for the early prediction of lung cancer to improve outcomes. Additionally, the development of novel therapeutic strategies is essential, either as standalone interventions or to enhance current prevention programs such as smoking cessation [5].

Proteins serve as the cornerstone of human metabolism, translating genomic information into the processes of growth, development, and homeostasis. The levels and characteristics of individual plasma proteins reflect a range of physiological or pathological states. Their accessibility via minimally invasive and inexpensive methods renders them ideal candidates for biomarker identification and as targets for drug development [6]. The advent of high-throughput technologies has propelled research into the potential of plasma proteins as predictors of lung cancer [7, 8]. These proteins are particularly promising as drug targets, as their interaction with therapeutic agents can be efficiently monitored in the bloodstream during randomized controlled trials (RCTs), thereby accelerating the drug development pipelines. Despite these advances, there remains a significant gap in meeting the clinical demand for novel biomarkers.

Mendelian randomization (MR) is a technique that leverages genetic variants as natural instruments to establish putative causality between phenotypes, thereby minimizing the potential for residual confounding and reverse causation bias in observational studies [9]. It relies on three key hypotheses: (i) the instrumental variable is associated with the exposure (relevance), (ii) the instrumental variable is not associated with any confounders (independence), (iii) the genetic instrument should only influence the outcome through the exposure, not through any other pathways, thus isolating the effect of the exposure on the outcome (exclusion restriction).

Human genetic variation has wide-ranging effects on protein structure and function, with important implications for health and disease. Proteomics data research through protein quantitative trait loci (pQTL) has opened new avenues for biomarker discovery and revealing disease mechanisms [10]. Integrating proteomics with genomics and MR has facilitated the identification of the causal roles of proteins in diseases, providing insights into etiology and potential therapeutic targets for colorectal cancer, ovarian cancer, stroke, and psychiatric disorders [11-14].

In the current study, we conducted proteome-wide MR analyses utilizing pQTL summary statistics from seven large-scaled proteomic studies. The study design is illustrated in Figure 1. This approach was designed to pinpoint circulating protein markers with a causal link to NSCLC risk. Bayesian colocalization, SMR (summary-data-based Mendelian randomization) and HEIDI (heterogeneity in dependent instruments) tests were also used to robustly identify cancer-risk proteins. Then single cell-type expression analysis was employed to detect the specific enrichment of these proteins in lung cancer tissue. Furthermore, we assessed the expression levels of these protein-coding genes between cancer and normal lung tissue using the GEPIA database. GeneMANIA and Gene Ontology (GO) analysis were conducted to explore the potential genetic interactions and functions of identified protein-coding genes. This was followed by a thorough druggability evaluation aimed at determining their potential as therapeutic targets for lung cancer. Lastly, risk proteins were mapped to potential therapeutic interventions and phenome-wide association study (PheWAS) was employed to elucidate the potential benefits and challenges of intervening on these proteins, including the possibility of unforeseen adverse outcomes. These multifaceted analyses not only provide compelling causal evidence for the role of proteins in lung cancer genesis, but also offer a new perspective on the predictive and preventive strategies for lung cancer.

Materials and Methods

We performed proteome-wide MR based on pQTLs from 7 large-scale proteomic studies and summary genetic association data of NSCLC from previously published GWAS, to examine the causal proteins associated with LUAD and LUSC. All contributing studies included in this analysis were approved by the relevant institutional review board from each country and all participants provided informed consent.

Study data sources and selection of instrumental variables (IVs)

The GWAS data for NSCLC in the primary outcome were obtained from a large-scale GWAS study of European descent, including 66,756 individuals of LUAD (11,273 cases and 55,483 controls) and 63,053 individuals of LUSC (7,426 cases and 55,627 controls), respectively [15]. In addition, for external validation, we used the latest release data on NSCLC from the FinnGen study R10 in this analysis, comprising 1,590 cases and 314,193 controls for LUAD, and 1,510 cases and 314,193 controls for LUSC. We obtained pQTL data from 7 large-scale proteomic studies [16-22]. These pQTLs were selected based on the following criteria: (i) single nucleotide polymorphisms (SNPs) associated with plasma proteins levels at a genome-wide significance level ($P < 5 \times 10^{-8}$); (ii) exclusion of SNPs within the Major Histocompatibility Complex (MHC) region (chr6:25.5–34.0Mb); (iii) SNPs that are independent of each other, with linkage disequilibrium (LD) clumped at $r^2 < 0.001$ within 10,000 kb window using the 1000 Genomes reference panel; (iv) utilization of F-statistics to evaluate the potential influence of weak instrument bias on the estimated effects of causal associations, wherein R^2 denotes the variance explained by the IVs [23, 24]. To eliminate duplicate proteins, we selected those with the highest cumulative of R^2 values.

In total, 13,236 SNPs representing 4,853 unique plasma proteins were included in the analysis. Details regarding the data sources utilized in this study are provided in Table S1-S2. The instrumental variables are listed in Table S3.

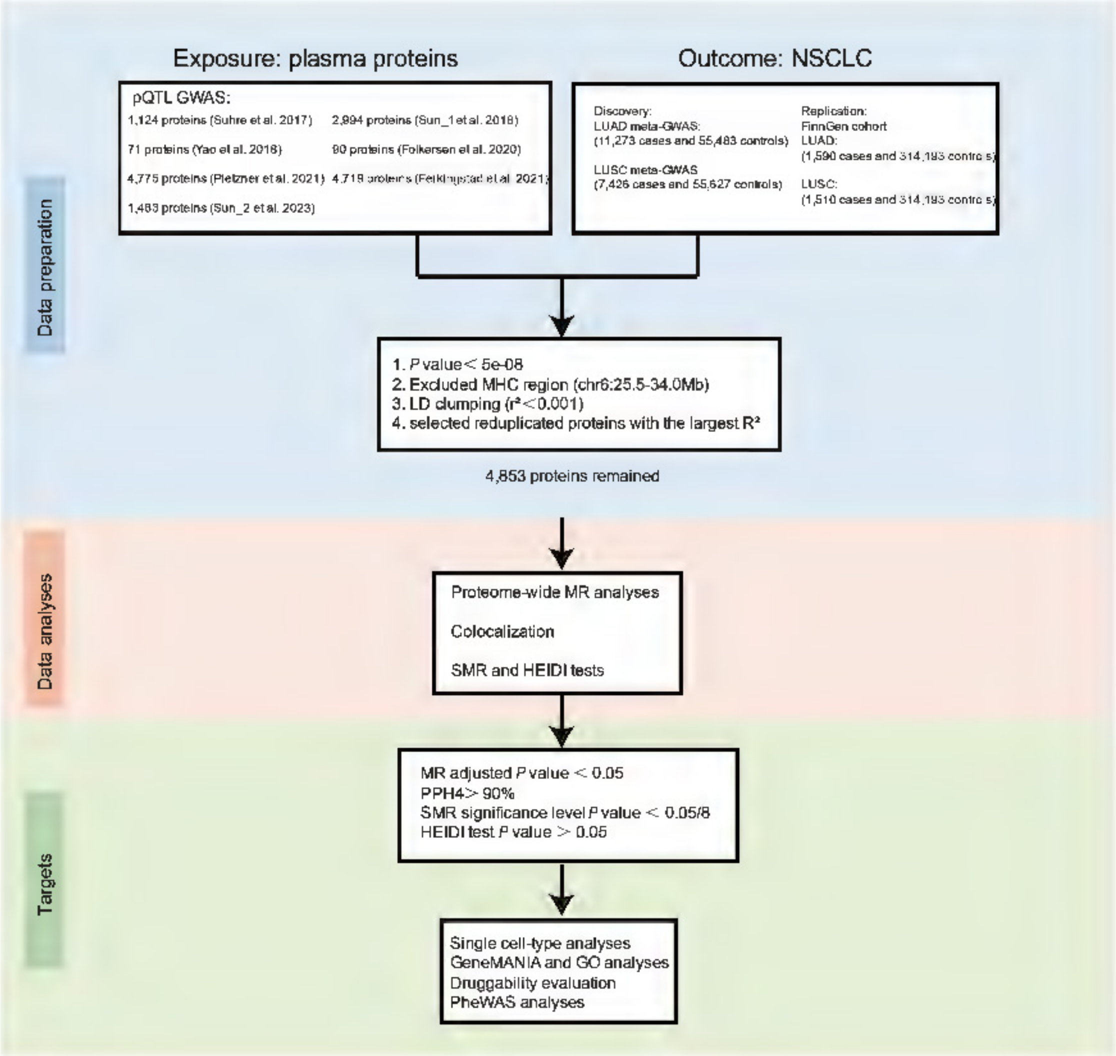

Exposure: plasma proteins
Outcome: NSCLC
pQTL GWAS:
1,124 proteins (Suhre et al. 2017)
2,994 proteins (Sun_1 et al. 2018)
71 proteins (Yao et al. 2018)
90 proteins (Folkersen et al. 2020)
4,775 proteins (Pietzner et al. 2021)
4,719 proteins (Ferkingstad et al. 2021)
1,463 proteins (Sun_2 et al. 2023)
Discovery:
LUAD meta-GWAS:
(11,273 cases and 55,483 controls)
LUSC meta-GWAS
(7,426 cases and 55,627 controls)
Replication:
FinnGen cohort
LUAD:
(1,590 cases and 314,193 controls)
LUSC:
(1,510 cases and 314,193 controls)
Data preparation
1. P value < 5e-08
2. Excluded MHC region (chr6:25.5-34.0Mb)
3. LD clumping (r² < 0.001)
4. selected reduplicated proteins with the largest R²
4,853 proteins remained
Data analyses
Proteome-wide MR analyses
Colocalization
SMR and HEIDI tests
Targets
MR adjusted P value < 0.05
PPH4> 90%
SMR significance level P value < 0.05/8
HEIDI test P value > 0.05
Single cell-type analyses
GeneMANIA and GO analyses
Druggability evaluation
PheWAS analyses

Figure 1. Overall flow chart of the study.

Proteome-wide MR analyses

We treated plasma proteins as exposures and the two subtypes of NSCLC as outcomes. The "TwoSampleMR" package [25] was used to conduct MR analysis in R. For proteins with a single SNP instrument, the Wald ratio method was employed to estimate the causal effect of circulating protein levels on the development of lung cancer. For the primary MR analyses, we applied the random-effects inverse variance weighted (IVW) method, which is appropriate for estimating MR effects for proteins with multiple SNP instruments. The MR-Egger, weighted median, simple mode, and weighted mode methods were also employed as sensitivity analyses to complement the IVW method for univariable models [26]. MR-Egger regression was utilized to assess horizontal pleiotropy in MR analysis. We also conducted Cochran's Q test to evaluate heterogeneity among SNPs and assess the consistency of MR assumptions and analyses. The *P* value was adjusted using the false discovery rate (FDR) procedure; an adjusted *P* value (adj. *P*) of < 0.05 was considered statistically significant. Additionally, a *P* value < 0.05 was defined as the nominal significance level. Replication MR analyses were also conducted for the identified plasma proteins using LUAD and LUSC GWAS data from FinnGen; a *P* value < 0.05 was considered statistically significant in the replication group.

Identified protein markers considered as causal inferences must satisfy the following criteria. (i) The MR results using the Wald ratio or IVW methods show a nominally significant association with LUAD or LUSC risk (*P* value < 0.05). (ii) The Wald ratio or IVW methods, or weighted median, exhibit a significant association with LUAD or LUSC risk after the Benjamini-Hochberg test (adj. *P* < 0.05). (iii) The MR-Egger regression test suggests the absence of horizontal pleiotropy. All these procedures were carried out using R software version 4.3.0.

Bayesian colocalization analyses

To investigate whether the observed association signals in two traits (trait 1: identified proteins, trait 2: NSCLC-related traits) are consistent with a shared genetic variant, we employed Bayesian colocalization analyses using the "coloc" package [27]. This analysis encompassed the posterior probabilities of five hypotheses: (i) no causal genetic variant exists for either trait in the genomic locus (H0); (ii) a single causal genetic variant for trait 1 (H1); (iii) a single causal genetic variant for trait 2 (H2); (iv) two different causal genetic variants for both traits (H3); (v) shared causal genetic variants for both traits (H4). A posterior probability exceeding 90% for H4 (PPH4) was considered strong evidence, suggesting that the proteins and NSCLC potentially shared identical genetic variants.

Summary-data-based MR (SMR) analyses and reverse MR analyses

Additionally, the SMR test was utilized to confirm the causal relationship between proteins and NSCLC risk [28]. The heterogeneity in dependent instruments (HEIDI) test was conducted to perform a sensitivity analysis using SNPs located within a defined genomic region. This analysis aimed to determine whether the association between specific proteins and NSCLC risk arises from shared genetic variants rather than genetic linkage. A P value < 6.25×10^{-3} (0.05/8) was set as the significance threshold for the SMR analysis. Furthermore, a P value > 0.05 in the HEIDI test indicated that the association between the protein and NSCLC was not influenced by linkage disequilibrium. The SMR and HEIDI tests were conducted using the SMR 1.3.1 software [28].

We investigated the reverse causality relationship between the potential influence of LUAD and LUSC on protein levels. Proteins identified in the primary analyses were examined utilizing five MR methods (IVW, MR-Egger, weighted median, simple mode, and weighted mode). Furthermore, we considered GWAS significant genetic variants ($P < 5 \times 10^{-08}$) and independent genetic variants (clumped at $r^2 < 0.001$), with an adjusted P value < 0.05 established as the threshold for statistical significance.

Single cell-type expression analyses and protein-coding genes expression pattern analyses

We further investigated the gene expression levels of plasma proteins across different cell types, specifically targeting those with a causal role in lung cancer development. This analysis leveraged single-cell RNA sequencing (scRNA-seq) data from the two major histopathological subtypes of NSCLC, and the adjacent non-cancerous tissue. The scRNA-seq data were sourced from the Gene Expression Omnibus (GEO) database. GSE149655 contains two LUAD tissue samples and two distal normal tissue samples. Additionally, we acquired two LUSC tissue samples, GSM3635278 and GSM3635285, from GSE127465. Using the "Seurat" package [29] in R, we implemented a standard workflow for data preprocessing and cell clustering based on the raw scRNA-seq data. The datasets for the two NSCLC subtypes and normal lung tissue underwent individualized analysis. During this process, genes with expression counts fewer than three in a single cell and cells with fewer than 200 unique features were excluded. Subsequently, we employed the NormalizeData and ScaleData functions to normalize and scale the transcripts per million (TPM) of RNA. The optimal number of principal components (PCs) for further analysis was determined using the RunPCA function, supplemented by constructing Elbow plots for each dataset. Cell clustering was conducted using the FindNeighbors and FindClusters functions, and non-linear dimensional reduction was performed using the RunUMAP function. To annotate various cell types, the "SingleR" package [30] was utilized with reference datasets HumanPrimaryCellAtlasData and BlueprintEncodeData. Utilizing GEPIA, a platform that combines extensive data from TCGA cancer and GTEx normal tissues, we created box plots to visualize the distribution of protein-coding genes in normal lung tissue and in LUAD or LUSC cancer tissues [31].

GeneMANIA and Metascape analyses and proteins druggability evaluation

The GeneMANIA platform [32] (https://genemania.org/) was employed to investigate the potential genetic interactions and functions of the identified protein-coding genes. We also employed the Metascape database (http://metascape.org/) to conduct gene ontology (GO) biological process enrichment analysis. Drug repositioning, which involves utilizing the molecular structure, therapeutic indications, and adverse effects of an existing drug, represents a promising strategy for developing novel therapeutic functions. This approach has the potential to significantly reduce the costs and time associated with the discovery and approval of new treatments for various diseases [33]. To determine whether the identified plasma proteins could serve as viable therapeutic targets for NSCLC, we searched three widely used drug-related databases: ChEMBL, DrugBank, and the Therapeutic Target Database (TTD). We documented the specific drug-gene interactions and outlined the drug development processes in this study.

Protein associations with other traits

To deepen our understanding of the specificity of the identified cancer risk proteins, we undertook additional analytical steps using PheWAS and consulted several public databases. Our goal was to gather comprehensive insights into the potential risks or benefits associated with modulating the expression of theses protein in human populations. Firstly, we evaluated the score of the probability of being loss-of-function intolerant (pLI) for each cancer risk protein on the Exome Aggregation Consortium platform using the gnomAD browser. Secondly, we explored the UK Biobank using Genebass, which included exome-sequencing studies, rare-variant association studies, and Mendelian genetics research. To conduct a thorough assessment of the horizontal pleiotropy of potential drug targets and their possible side effects, we conducted a PheWAS using the AstraZeneca PheWAS Portal (https://azphewas.com/) [34]. The study utilized a dataset from the UK Biobank, comprising approximately 15,500 binary and 1,500 continuous phenotypes derived from a subset of nearly 450,000 exome sequencing participants. To ensure the robustness of our findings, we applied rigorous statistical corrections and set a stringent threshold of 2E−9 (the default in the AstraZeneca PheWAS Portal) to mitigate the likelihood of false positives.

Results

Causal effects of 8 plasma proteins on NSCLC risk

Proteome-wide MR analysis genetically identified eight plasma proteins associated with NSCLC, with seven linked to LUAD and one to LUSC. These proteins comprise Cadherin-17 (CDH17), Carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5), Coxsackievirus and adenovirus receptor (CXADR), Family with sequence similarity 3, member D (FAM3D), Kallikrein-1 (KLK1), Protein O-glucosyltransferase 3 (POGLUT3), Pulmonary surfactant-associated protein B (SFTPB), and CD14. Higher levels of CDH17 (odds ratio (OR) = 0.89; 95% confidence interval (CI) 0.83–0.95), CXADR (OR = 0.74; 95% CI 0.64–0.84), FAM3D (OR = 0.90; 95% CI 0.85–0.95), POGLUT3 (OR = 0.84; 95% CI 0.78–0.91), and SFTPB (OR = 0.87; 95% CI 0.82–0.93) were associated with a reduced risk of LUAD. Conversely, CEACAM5 (OR = 1.18; 95% CI 1.02–1.35) and KLK1 (OR = 1.08; 95% CI 1.02–1.14) were linked to an elevated risk of LUAD. Furthermore, higher levels of CD14 (OR = 1.30; 95% CI 1.16–1.45) were associated with an increased risk of LUSC (Figure 2, Table S4-S6).

We validated our primary findings in the replication stage. Five proteins were successfully confirmed in the FinnGen cohort. The OR (95% CI) of NSCLC per SD increase in genetically predicted levels of protein was 0.73 (0.54–0.99) for CXADR, 0.85 (0.74–0.97) for FAM3D, 1.08 (1.00–1.17) for KLK1, 0.65 (0.51–0.83) for POGLUT3 in LUAD, and 1.53 (1.23–1.91) for CD14 in LUSC (Figure S1-S2, Table S7-S8).

Sensitivity analyses for non-small cell lung cancer causal proteins

To further scrutinize the robustness of our primary MR analyses, Cochran's Q test was performed to assess heterogeneity across the identified eight causal proteins using the IVW or MR-Egger method. A significance level of $P \leq 0.05$ would indicate the presence of heterogeneity, leading to the adoption of a random-effects IVW approach. Notably, four out of the eight proteins (CDH17, CXADR, FAM3D, CD14) exhibited no evidence of heterogeneity. Moreover, no signs of horizontal pleiotropy were identified in any of the eight proteins, as determined by the MR-Egger intercept method ($P_{\text{pleiotropy}} > 0.05$) (Table S4). During the replication stage, no heterogeneity or pleiotropy of the eight proteins was observed in either LUAD or LUSC ($P_{\text{heterogeneity}} > 0.05$, $P_{\text{pleiotropy}} > 0.05$) (Table S7, S8).

For investigating potential reverse causality between the identified plasma proteins and different pathological types of NSCLC, we identified LUAD and LUSC as exposures and eight proteins as outcomes. Bidirectional MR analyses provided no evidence of a reverse causal relationship between the identified plasma proteins and different pathological types of NSCLC (Table S10).

Causality Verification through Colocalization and SMR Analyses

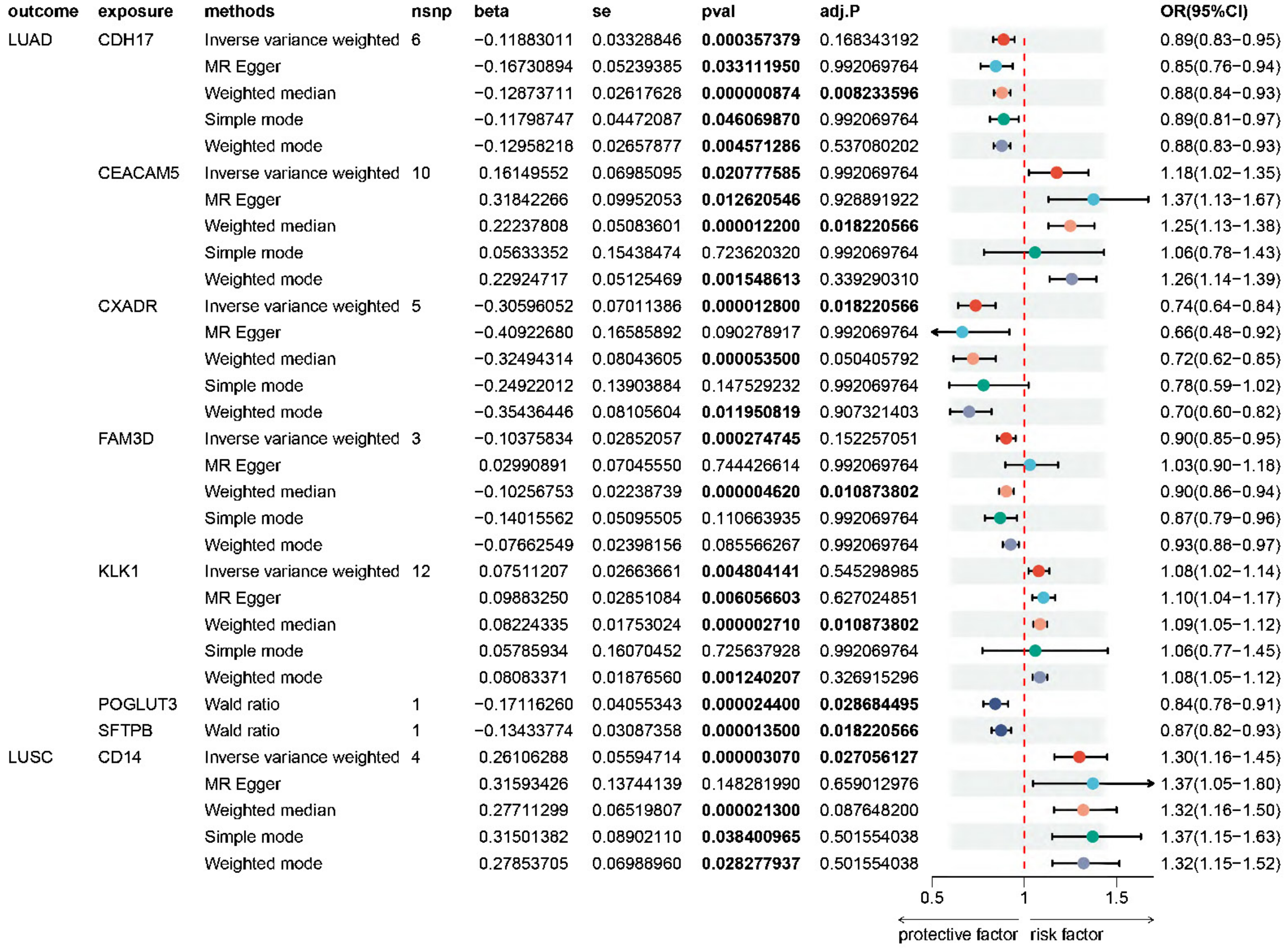
Among the eight causal proteins identified through proteome-wide MR, POGLUT3 and SFTPB stand out, supported by robust evidence of genetic colocalization, with a PPH4 > 90% (Table 1, Table S9). This suggests a high likelihood of a shared causal variant influencing both the protein level and the risk of LUAD.

To distinguish pleiotropy from linkage and validate the primary proteome-wide MR findings, we conducted the SMR and HEIDI tests utilizing proteins with complete summary-level data. Table 1 indicates that only the protein SFTPB successfully passed both the SMR test ($P < 6.25\times10^{-3}$) and the HEIDI test ($P > 0.05$). This result suggests the lack of pleiotropy in the SNPs associated with SFTPB.

Table 1. Summary results from Mendelian randomization (MR), colocalization, and SMR for 8 proteome-wide MR-identified proteins.

Protein	Protein full name	MR		Colocalization	SMR		
		Beta	P	PPH4>0.9	Beta	P	PHEIDI
CDH17	Cadherin-17	-0.12	3.57E-04	No	-	-	-
CEACAM5	Carcinoembryonic antigen-related cell adhesion molecule 5	0.16	2.08E-02	No	-	-	-
CXADR	Coxsackievirus and adenovirus receptor	-0.31	1.28E-05	No	-	-	-
FAM3D	Family with sequence similarity 3, member D	-0.10	2.75E-04	No	-0.129963	5.09E-02	0.4
KLK1	Kallikrein-1	0.08	4.80E-03	No	-	-	-
POGLUT3	Protein O-glucosyltransferase 3	-0.17	2.44E-05	Yes	-	-	-
SFTPB	Pulmonary surfactant-associated protein B	-0.13	1.35E-05	Yes	-0.135282	1.50E-05	0.3
CD14	Monocyte differentiation antigen CD14	0.26	3.07E-06	No	-	-	-

Abbreviations: HEIDI, heterogeneity in dependent instruments; MR, Mendelian randomization; PPH4, posterior probability exceeding 90% for H4; SMR, summary-data-based Mendelian randomization.



outcome	exposure	methods	nsnp	beta	se	pval	adj.P	OR(95%CI)
LUAD	CDH17	Inverse variance weighted	6	−0.11883011	0.03328846	**0.000357379**	0.168343192	0.89(0.83−0.95)
		MR Egger		−0.16730894	0.05239385	**0.033111950**	0.992069764	0.85(0.76−0.94)
		Weighted median		−0.12873711	0.02617628	**0.000000874**	**0.008233596**	0.88(0.84−0.93)
		Simple mode		−0.11798747	0.04472087	**0.046069870**	0.992069764	0.89(0.81−0.97)
		Weighted mode		−0.12958218	0.02657877	**0.004571286**	0.537080202	0.88(0.83−0.93)
	CEACAM5	Inverse variance weighted	10	0.16149552	0.06985095	**0.020777585**	0.992069764	1.18(1.02−1.35)
		MR Egger		0.31842266	0.09952053	**0.012620546**	0.928891922	1.37(1.13−1.67)
		Weighted median		0.22237808	0.05083601	**0.000012200**	**0.018220566**	1.25(1.13−1.38)
		Simple mode		0.05633352	0.15438474	0.723620320	0.992069764	1.06(0.78−1.43)
		Weighted mode		0.22924717	0.05125469	**0.001548613**	0.339290310	1.26(1.14−1.39)
	CXADR	Inverse variance weighted	5	−0.30596052	0.07011386	**0.000012800**	**0.018220566**	0.74(0.64−0.84)
		MR Egger		−0.40922680	0.16585892	0.090278917	0.992069764	0.66(0.48−0.92)
		Weighted median		−0.32494314	0.08043605	**0.000053500**	0.050405792	0.72(0.62−0.85)
		Simple mode		−0.24922012	0.13903884	0.147529232	0.992069764	0.78(0.59−1.02)
		Weighted mode		−0.35436446	0.08105604	**0.011950819**	0.907321403	0.70(0.60−0.82)
	FAM3D	Inverse variance weighted	3	−0.10375834	0.02852057	**0.000274745**	0.152257051	0.90(0.85−0.95)
		MR Egger		0.02990891	0.07045550	0.744426614	0.992069764	1.03(0.90−1.18)
		Weighted median		−0.10256753	0.02238739	**0.000004620**	**0.010873802**	0.90(0.86−0.94)
		Simple mode		−0.14015562	0.05095505	0.110663935	0.992069764	0.87(0.79−0.96)
		Weighted mode		−0.07662549	0.02398156	0.085566267	0.992069764	0.93(0.88−0.97)
	KLK1	Inverse variance weighted	12	0.07511207	0.02663661	**0.004804141**	0.545298985	1.08(1.02−1.14)
		MR Egger		0.09883250	0.02851084	**0.006056603**	0.627024851	1.10(1.04−1.17)
		Weighted median		0.08224335	0.01753024	**0.000002710**	**0.010873802**	1.09(1.05−1.12)
		Simple mode		0.05785934	0.16070452	0.725637928	0.992069764	1.06(0.77−1.45)
		Weighted mode		0.08083371	0.01876560	**0.001240207**	0.326915296	1.08(1.05−1.12)
	POGLUT3	Wald ratio	1	−0.17116260	0.04055343	**0.000024400**	**0.028684495**	0.84(0.78−0.91)
	SFTPB	Wald ratio	1	−0.13433774	0.03087358	**0.000013500**	**0.018220566**	0.87(0.82−0.93)
LUSC	CD14	Inverse variance weighted	4	0.26106288	0.05594714	**0.000003070**	**0.027056127**	1.30(1.16−1.45)
		MR Egger		0.31593426	0.13744139	0.148281990	0.659012976	1.37(1.05−1.80)
		Weighted median		0.27711299	0.06519807	**0.000021300**	0.087648200	1.32(1.16−1.50)
		Simple mode		0.31501382	0.08902110	**0.038400965**	0.501554038	1.37(1.15−1.63)
		Weighted mode		0.27853705	0.06988960	**0.028277937**	0.501554038	1.32(1.15−1.52)

Figure 2. Forest plot of Mendelian randomization (MR) analyses on the association of plasma proteins and the risk of non-small cell lung cancer. MR point estimates and 95% confidence intervals (CIs) are provided for the 8 specific causal proteins in association with the risk of LUAD and LUSC. An adjusted *P* value (adj. *P*) below 0.05 indicated statistical significance, while a *P* value < 0.05 was set as the threshold for nominal significance level. OR, odds ratio; CIs, confidence intervals; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma. Full name of proteins: CDH17, Cadherin-17; CEACAM5, Carcinoembryonic antigen-related cell adhesion molecule 5; CXADR, Coxsackievirus and adenovirus receptor; FAM3D, Family with sequence similarity 3, member D; KLK1, Kallikrein-1; POGLUT3, Protein O-glucosyltransferase 3; SFTPB, Pulmonary surfactant-associated protein B; CD14, Monocyte differentiation antigen CD14.

Cell-type specificity expression in the NSCLC tissue

To investigate the cell-type specificity of coding genes corresponding to the 8 plasma proteins in NSCLC tissue, we performed single-cell data analyses using the GEO database. In LUAD tumor tissue, using the FindCluster() function, we identified 13 clusters and annotated 5 cell types (T cells, epithelial cells, macrophages, fibroblasts, and endothelial cells) (Figure S3A, Figure 3A). Except for the *POGLUT3* gene, we detected 7 of the 8 protein-coding genes in

LUAD tissue. Figure 3 (B-C) shows the single-cell expression of these 7 coding genes in each cluster. *CEACAM5*, *CXADR*, *FAM3D*, and *SFTPB* were mainly enriched in epithelial cells, and *CD14* was associated with macrophages, whereas *CDH17* was deficient in all clusters. We annotated 7 cell types (T cells, macrophages, monocytes, epithelial cells, neutrophils, B cells, and tissue stem cells) in LUSC tissue (Figure 4A, Figure S3B). *CXADR* and *SFTPB* were mainly enriched in epithelial cells, while *CD14* was enriched in both macrophages and monocytes (Figure 4B-4C). Moreover, in normal lung tissue, 6 protein-coding genes had expression data, while *POGLUT3* and *CDH17* were undetected. *CXADR* and *SFTPB* were mainly enriched in epithelial cells, while *FAM3D* and *CD14* were mainly enriched in endothelial cells (Figure S4). The expression levels of these 6 protein-coding genes in both NSCLC and normal tissue were also depicted in Figure S5.

GeneMANIA analyses and druggability evaluation for candidate protein targets

We employed GeneMANIA to construct a comprehensive gene network by inputting the 8 protein-coding genes of interest. This analysis identified 20 correlated genes, predicated based on co-expression, co-localization, and shared protein domains. This expanded network, as depicted in Figure 5A, encompasses about a total of 510 interaction links. These interactions are categorized into co-expression (90.73%), co-localization (5.44%) and common protein structural domains (3.83%). A functional analysis of the network elucidates the roles of the drug targets and their associated genes, revealing significant involvement in cell-cell adhesion, tissue homeostasis, and protein glycosylation processes (Figure 5B).

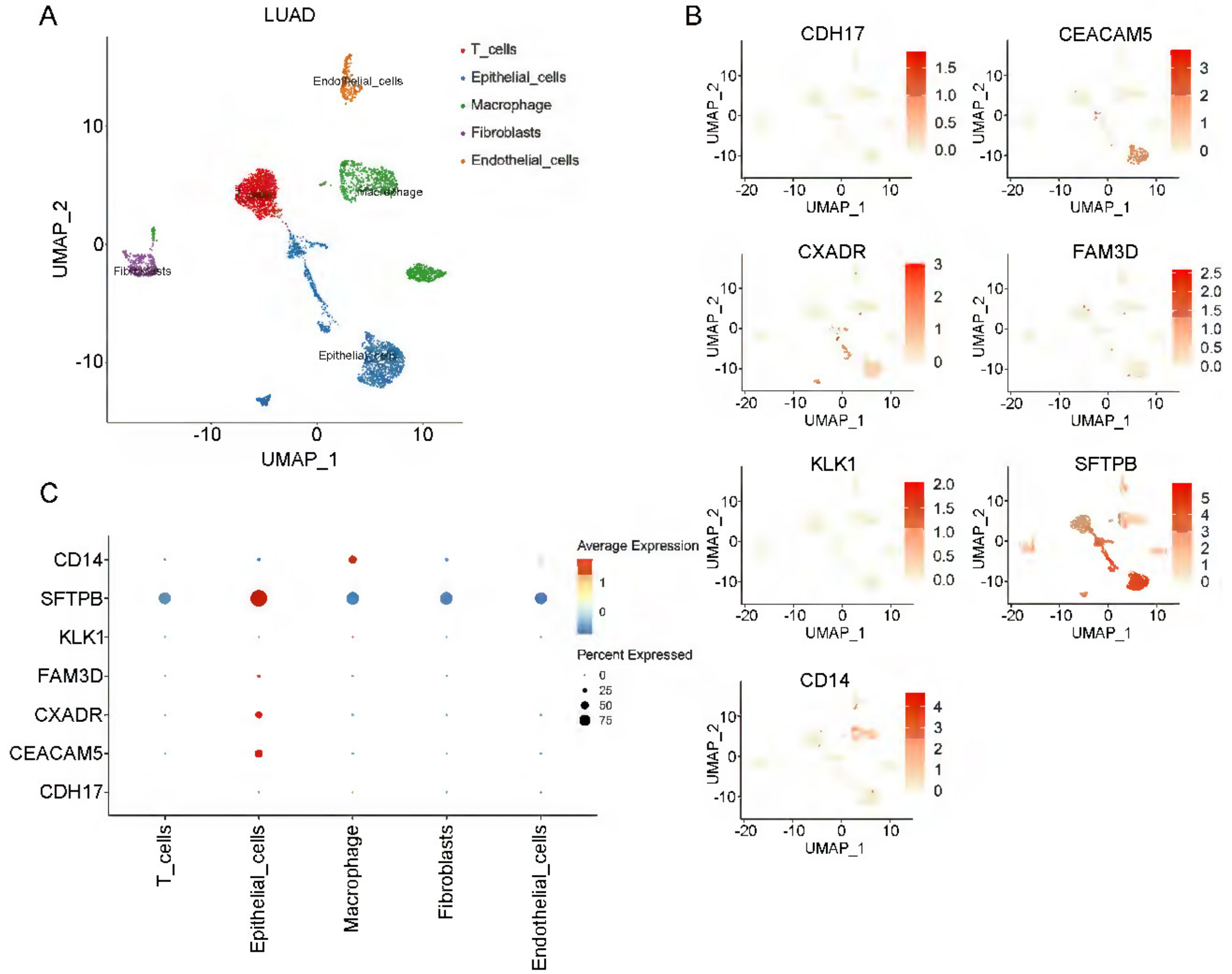


Figure 3. Single-cell data analyses of lung adenocarcinoma. A: The LUAD cell clusters were categorized into five distinct cell types; B and C display the expression patterns of the identified causal proteins within each cell type.

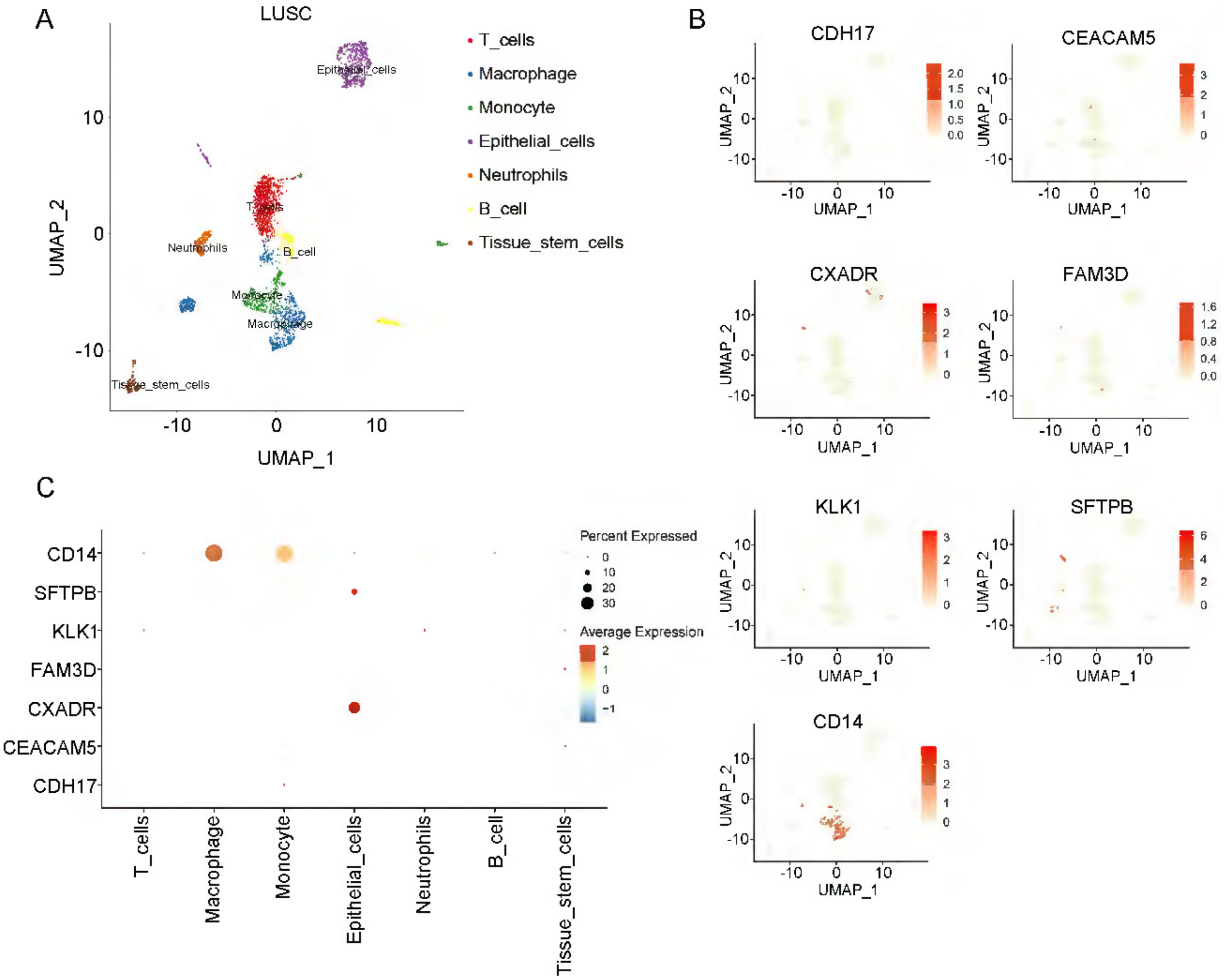


Figure 4. Single-cell data analyses of lung squamous cell carcinoma. A: The LUSC cell clusters were categorized into seven distinct cell types; B and C illustrate the expression of the identified causal proteins in each cell type.

In the assessment of the druggability and drug development potential of the eight candidate plasma proteins, we identified three proteins (CEACAM5, KLK1, CD14) that have been acknowledged as druggable targets (Table S11). CEACAM5, has been the focus of clinical trials for tusamitamab ravtansine, a drug designed for a spectrum of cancer types. Promising data from non-squamous NSCLC patients indicated a correlation between the antitumor activity and the expression level of CEACAM5 (Clinical trial information: NCT02187848; Registration Date: 2014-07-11) [35]. Drugs targeting KLK1 have been researched for various applications, including lanoteplase for myocardial infarction, aprotinin for reducing bleeding and transfusion needs during surgery, nafamostat for anticoagulant therapy in acute kidney injury and liver transplantation scenarios, and aniline for the treatment of multiple myeloma. Drugs targeting CD14 have been applied in the treatment of sepsis with atibuclimab and autoimmune diseases, such as psoriasis and ulcerative colitis with the example of VB-201.

Identifying loss-of-function variants and PheWAS analyses for causal proteins in non-small cell lung cancer

In our study, seven of the cancer risk proteins exhibited pLI scores below 0.1, implying a high tolerance for loss of function (LOF) variation. In contrast, CXADR showed a slightly lower intolerance of LOF variation, with a pLI score of 0.34. We conducted a gene-level PheWAS analysis using an extensive dataset from the AstraZeneca PheWAS Portal [34], which included 17,361 dichotomous and 1,419 quantitative phenotypes. As detailed in the Table S12, none of the eight drug targets showed significant associations with other traits or diseases at the gene level, with a rigorous significance threshold of P < 1E−05 for genomic association. We observed only limited evidence for the association of pLOF variants in cognate genes for cancer risk proteins with other traits, none of which were cancer endpoints. The pLOF burden in CXADR was associated with a lack of interest in doing things ($P = 3.39^{-5}$, Beta = 2.82^{-2}) and back pain ($P = 8.88^{-5}$, Beta = 1.7^{-1}), and FAM3D pLOF burden with chickenpox ($P = 6.05^{-5}$, Beta = 1.69^{-1}). Additionally, protein-damaging missense variation in CXADR was linked to cervical spondylosis ($P = 5.19^{-5}$, Beta = 3.88^{-2}), FAM3D was associated with B44 aspergillosis ($P = 3.35^{-5}$, Beta = 7.14^{-2}), and POGLUT3 with xenograft replacement of aortic valve ($P = 3.18^{-5}$, Beta = 4.31^{-2}), stroke ($P = 3.22^{-5}$, Beta = 1.45^{-2}) and vascular/heart problem ($P = 6.65^{-5}$, Beta = 1.31^{-2}). This finding suggests a low probability of substantial side effects associated with drugs targeting these genes or the presence of significant horizontal pleiotropy.

Discussion

In this study, we systematically evaluated the causal relationship between plasma proteins and NSCLC. Using pQTL data from up to 4,853 plasma proteins, we identified eight proteins that likely play a role in the etiology of NSCLC. Specifically, genetically determined lower levels of proteins such as CDH17, CXADR, FAM3D, POGLUT3, and SFTPB, along with higher levels of CEACAM5 and KLK1, were associated with increased risk of LUAD, while higher levels of CD14 were associated with increased risk of LUSC. Additionally, five proteins (CXADR, FAM3D, KLK1, POGLUT3, CD14) were corroborated in external cohorts; two proteins (POGLUT3, SFTPB) were verified by Bayesian colocalization; and one protein (SFTPB) was identified through SMR and HEIDI tests, underlining their potential therapeutic relevance. Subsequently, bidirectional MR was conducted and no proteins revealed reverse causality. We further verified the differential expressions of these protein-coding genes in various cell types,

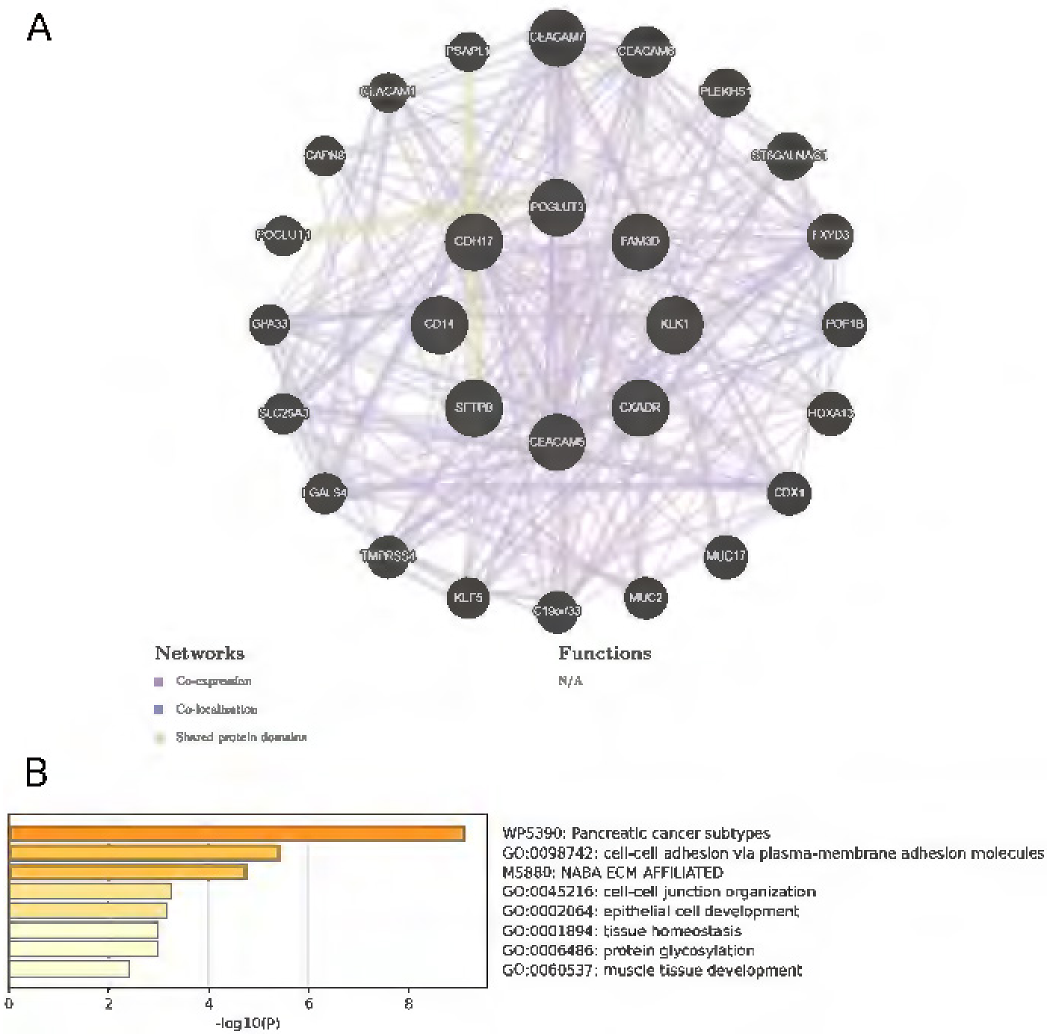


Figure 5. The GeneMANIA and Metascape analyses for the eight protein-coding genes. A: Eight identified protein-coding genes were placed in the inner circle, while twenty genes associated with them were situated in the outer circle. B: Metascape provides a bar graph for visualizing Gene Ontology (GO) enrichment analysis and interpreting metabolomic data related to the identified protein-coding genes.

including epithelial cells, macrophages, monocytes, and endothelial cells. We also identified causal proteins that mediate the therapeutic effects of specific drugs that may lead to drug repurposing. Moreover, we identified proteins as novel targets that were previously unexplored in the context of NSCLC therapy, thereby paving the way for new investigational pathways. Lastly, we employed phenome-wide association analysis to delve deeper into the potential pleiotropic effects of the target genes and to foresee possible drug side effects, ensuring a comprehensive evaluation of our findings.

The significance of our findings lies in their extension of current evidence associating these proteins with lung cancer, based on studies of gene polymorphisms, mRNA, or protein levels. It is notable that SFTPB was identified based on the strongest evidence from colocalization, SMR and HEIDI tests. SFTPB is an essential protein for maintaining normal lung function [36]. SFTPB is first synthesized as a hydrophilic 42-kD protein by type 2 alveolar pneumocytes and nonciliated bronchiolar cells in the form of pro-SFTPB. Following its synthesis, pro-SFTPB rapidly undergoes proteolytic cleavage by cysteine proteases within the endoplasmic reticulum, resulting in the production and secretion of a 9-kD non-collagenous hydrophobic SFTPB, which represents the functional mature form of the protein. Research has indicated that pro-SFTPB may serve as a potential biomarker for lung cancer, with elevated levels possibly signifying an increased risk of lung cancer [37, 38]. The mature SFTPB has also been recognized as a potential biomarker for identifying metastatic lymph node involvement in patients with NSCLC [39]. Existing assays are unable to differentiate between precursor and mature forms; however, numerous studies have established that total SFTPB is associated with lung cancer prognosis [37]. Our study provides population-level evidence confirming the causal relationship between reduced SFTPB protein levels and NSCLC risk, suggesting that SFTPB agonists could be a promising avenue for therapeutic prevention in high-risk individuals with lung cancer.

KLK1, a member of the kallikrein-related peptidase family, has been implicated in regulating vascular tone, inflammation, and tumor microenvironment [40, 41]. Our findings align with recent studies suggesting a causal role for KLK1 in cancer prognosis and carcinogenesis, indicating that drugs targeting KLK1 may offer an avenue for lung cancer prevention. CEACAM5, a cell-surface glycoprotein, has emerged as a promising therapeutic target in preclinical models [42]. Our findings support the idea that CEACAM5 may be a risk factor for LUAD, suggesting that high-risk individuals with LUAD may benefit from targeted therapeutic interventions. CD14, a co-receptor involved in innate immune responses, has been linked to disease aggressiveness in different cancers [43-45]. Elevated levels of CD14 have been observed in NSCLC patients [46]. Our research supports the role of CD14 in promoting LUSC carcinogenesis, however, the mechanisms are yet to be elucidated. POGLUT3, an O-glucosyltransferase, plays a pivotal role in targeting secreted proteins crucial for the assembly and function of the extracellular matrix [47, 48]. Our findings indicate that a genetic predisposition to elevated POGLUT3 levels is linked to reduced LUAD risk, underscoring the necessity for additional epidemiological studies and experimental investigations. While the effect sizes observed in our MR analysis—specifically, the association between a 1- SD increase in CEACAM5 levels and LUAD risk (OR = 1.18, 95% CI 1.02–1.35)—are indeed more modest than those of established epidemiological risk factors, such as smoking (e.g., a 15-fold increased risk for childhood-onset smokers in the UK Biobank [49]), this observation aligns with the biological role of plasma proteins as intermediate mediators in complex disease pathways. Furthermore, the synergistic potential of multiple proteins (e.g., CDH17 and CEACAM5) may amplify risk through either additive or interactive mechanisms, which is consistent with the rationale for multi-target combination therapies in cancer [50]. The lack of statistical significance for CEACAM5, CDH17, and KLK1 in the FinnGen cohort may stem from population differences or environmental interactions. The distinct genetic ancestry of the FinnGen cohort (Finnish-specific haplotype frequencies) [51] and its environmental context likely contribute to challenges in replication.

Proteins serve as essential targets for drug development. We identified three proteins that are the focus of drugs currently under investigation in phase I clinical trials or higher, indicating that their potential for drug-ability is actively being evaluated. Notably, the majority of drugs targeting cancer risk proteins are typically either monoclonal antibodies or small molecular inhibitors (SMIs). For example, labetuzumab, a monoclonal antibody, directly inhibits CEACAM5 and is utilized in the treatment of colorectal cancer [52]. KLK1 is inhibited by several SMIs in the treatment of cardiovascular diseases [53], while the anti-CD14 antibody atibuclimab is administered to patients with amyotrophic lateral sclerosis [54]. The identification of CEACAM5, KLK1, and CD14 as established drug targets for NSCLC presents significant therapeutic potential. These proteins are currently the focus of clinical trials and are involved in ongoing drug development efforts targeting both cancer and inflammatory diseases. Our study's findings further strengthen the argument for their role in NSCLC and suggest that these targets could be utilized in the development of novel treatment strategies, either individually or in combination with current therapeutic modalities, including immunotherapy, chemotherapy, or anti-angiogenic therapies. We propose that further exploration of these targets in clinical trials will facilitate the translation of these findings into effective therapeutic options for patients with NSCLC.

Our research demonstrates several strengths, such as a robust repository of plasma proteins, a substantial cohort of LUAD and LUSC cases, a minimal risk of reverse causation, and comprehensive colocalization analyses. These elements highlight the effectiveness of our methodology in elucidating the fundamental mechanisms governing NSCLC pathogenesis. The supplementary insights obtained from single-cell expression analysis, PPI studies, and drug target assessments have shed light on the potential pathogenic roles of candidate proteins in NSCLC. This has not only helped to delineate priority targets for drug development but also underscored the potential of our investigation to augment genetic screening methodologies for the early detection of NSCLC. Our work provides essential insights into the development of screening and preventative strategies, particularly for those with identified genetic predispositions. Furthermore, we underscore the potential for drug repurposing, acknowledging that high-risk proteins not currently the subject of any drug intervention may offer promising avenues for future drug discovery. Additionally, the identification of previously uninvestigated cancer-associated proteins unveils novel opportunities for transformative approaches in cancer treatment and prevention.

Our study presents several limitations that necessitate acknowledgment. Firstly, our study lacked a comprehensive assessment of the entire proteome, as it was restricted to measuring proteins exclusively via blood-based multiplex affinity platforms. Secondly, our analysis concentrated solely on individuals of European descent, thereby hindering the generalizability of our findings in the genetic landscape of diverse populations. To this end, we intend to utilize GWAS from East Asian populations, such as the China Kadoorie Biobank and Biobank Japan, to conduct cross-ethnic MR analysis in our future research. Thirdly, the statistical power might be insufficient in analyzing large cell lung cancer due to the absence of genome-wide association study (GWAS) data for this histological subtype of NSCLC. To evaluate the histological heterogeneity of NSCLC in greater detail, we intend to integrate single-cell multi-omics approaches in our subsequent studies. Fourthly, the GWAS summary data for LUAD and LUSC from public datasets exhibited imbalanced stratification across demographic covariates such as gender, age, and other crucial demographic factors. This imbalance constrained our ability to investigate potential causal links within distinct subgroups. Fifth, to address the absence of experimental validation, we plan to incorporate experimental validation through CRISPR-based protein knockdowns of SFTPB and CEACAM5 in NSCLC cell lines in our future work. Lastly, we conducted rigorous sensitivity analyses, including MR-Egger regression ($P_{\text{pleiotropy}} > 0.05$ for all proteins; see Table S4), weighted median methods, and Cochran's Q tests, to minimize bias from horizontal pleiotropy. However, due to the prevalent occurrence of horizontal pleiotropy in MR analysis, we were unable to sufficiently attenuate its impact in our study; nevertheless, this phenomenon may reveal alternative pathways associated with the traits under investigation.

Conclusions

Our study identified several plasma proteins associated with NSCLC risk, providing novel insights into its etiology and highlighting potential targets for diagnostic and therapeutic advancements. Further experimental and clinical investigations are imperative to assess the efficacy and significance of the eight identified proteins, with the goal of developing preventive strategies and clinical approaches to address the anticipated increase in cancer prevalence.

Abbreviations

CDH17: Cadherin-17; CEACAM5: Carcinoembryonic antigen-related cell adhesion molecule 5; CXADR: Coxsackievirus and adenovirus receptor; FAM3D: Family with sequence similarity 3, member D; FDR: False discovery rate; GO: Gene ontology; GWAS: Genome-wide association study; HEIDI: Heterogeneity in dependent instruments; IVs: Instrumental variables; IVW: Inverse variance weighted; KLK1: Kallikrein-1; LD: Linkage disequilibrium; LOF: Loss of function; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; MHC: Major histocompatibility complex; MR: Mendelian randomization; NSCLC: Non-small cell lung cancer; PheWAS: Phenome-wide association analyses; pLI: Probability of being loss-of-function intolerant; POGLUT3: Protein O-glucosyltransferase 3; pQTL: Protein quantitative trait loci; RCTs: Randomized controlled trials; SFTPB: Surfactant-associated protein B; SMR: Summary-data-based

Mendelian randomization; SNPs: Single nucleotide polymorphisms.

Supplementary Material

Supplementary figures and tables.
https://www.medsci.org/v22p4036s1.zip

Acknowledgements

We are grateful to all the participants, the institutions, and staff from the proteomic studies, the NSCLC GWASs, and the FinnGen cohort for providing the summary statistics utilized in this study for public access.

Funding

This work was supported by the National Natural Science Foundation of China (82102830, 82403772).

Author contributions

MHZ and YCZ conceived of and designed the research. MHZ and YCZ performed data analyses and drafted the manuscript. XKD helped to polish the manuscript. All authors read and approved the final version of this manuscript.

Availability of data and materials

The original contributions presented in the study are detailed in the article and Additional file; further inquiries can be directed to the corresponding author. The GWAS summary data of LUAD and LUSC are available at https://www.ebi.ac.uk/gwas/publications/28604730 [15]. The GWAS summary data of FinnGen are available at https://www.finngen.fi/en/access_results [51]. The GWAS summary data of proteins are available at https://omicscience.org/apps/pgwas/ [16], https://www.synapse.org/Synapse:syn51364943/ [19], https://www.decode.com/summarydata/ [17], http://www.phpc.cam.ac.uk/ceu/proteins/ [18], https://metabolomips.org/pgwas/ [20], https://preview.ncbi.nlm.nih.gov/gap/eqtl/studies/ (dbGaP Study Accession: phs000363) [22], and https://doi.org/10.5281/zenodo.2615265 [21].

Ethics approval and consent to participate

The ethics approval was obtained from the relevant authorities, specifically the coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa for FinnGen (HUS/990/2017). The study was conducted in strict adherence to the Declaration of Helsinki, and all participants provided informed consent.

Competing Interests

The authors have declared that no competing interest exists.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2024; 74: 229-63.
2. Zhang Y, Vaccarella S, Morgan E, Li M, Etxeberria J, Chokunonga E, et al. Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. The Lancet Oncology. 2023; 24: 1206-18.
3. Luo YH, Luo L, Wampfler JA, Wang Y, Liu D, Chen YM, et al. 5-year overall survival in patients with lung cancer eligible or ineligible for screening according to US Preventive Services Task Force criteria: a prospective, observational cohort study. The Lancet Oncology. 2019; 20: 1098-108.
4. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet. 2018; 391: 1023-75.
5. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nature reviews Clinical oncology. 2023; 20: 624-39.
6. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature. 2008; 452: 571-9.
7. Meaney CL, Mitchell KA, Zingone A, Brown D, Bowman E, Yu Y, et al. Circulating Inflammation Proteins Associated With Lung Cancer in African Americans. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2019; 14: 1192-203.
8. Li N, Tan F, Chen W, Dai M, Wang F, Shen S, et al. One-off low-dose CT for lung cancer screening in China: a multicentre, population-based, prospective cohort study. The Lancet Respiratory medicine. 2022; 10: 378-91.
9. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafo MR, et al. Mendelian randomization. Nature reviews Methods primers. 2022; 2.
10. Pietzner M, Wheeler E, Carrasco-Zanini J, Kerrison ND, Oerton E, Koprulu M, et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. Nature communications. 2021; 12: 6822.
11. Sun J, Zhao J, Jiang F, Wang L, Xiao Q, Han F, et al. Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. Genome medicine. 2023; 15: 75.
12. Considine DPC, Jia G, Shu X, Schildkraut JM, Pharoah PDP, Zheng W, et al. Genetically predicted circulating protein biomarkers and ovarian cancer risk. Gynecologic oncology. 2021; 160: 506-13.
13. Mishra A, Malik R, Hachiya T, Jurgenson T, Namba S, Posner DC, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. Nature. 2022; 611: 115-23.
14. Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nature genetics. 2020; 52: 1122-31.
15. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature genetics. 2017; 49: 1126-32.
16. Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, Worheide MA, et al. Mapping the proteo-genomic convergence of human diseases. Science. 2021; 374: eabj1541.
17. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. Nature genetics. 2021; 53: 1712-21.
18. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018; 558: 73-9.
19. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023; 622: 329-38.
20. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nature communications. 2017; 8: 14357.
21. Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman AK, Schork A, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nature metabolism. 2020; 2: 1135-48.
22. Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nature communications. 2018; 9: 3268.
23. Papadimitriou N, Dimou N, Tsilidis KK, Banbury B, Martin RM, Lewis SJ, et al. Physical activity and risks of breast and colorectal cancer: a Mendelian randomisation analysis. Nature communications. 2020; 11: 597.
24. Burgess S, Thompson SG, Collaboration CCG. Avoiding bias from weak instruments in Mendelian randomization studies. International journal of epidemiology. 2011; 40: 755-64.

25. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018; 7.
26. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic epidemiology. 2016; 40: 304-14.
27. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS genetics. 2014; 10: e1004383.
28. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nature communications. 2018; 9: 918.
29. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015; 33: 495-502.
30. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature immunology. 2019; 20: 163-72.
31. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research. 2017; 45: W98-W102.
32. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome biology. 2008; 9 Suppl 1: S4.
33. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nature reviews Drug discovery. 2004; 3: 673-83.
34. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021; 597: 527-32.
35. Gazzah A, Ricordel C, Cousin S, Cho BC, Calvo E, Kim TM, et al. Efficacy and safety of the antibody-drug conjugate (ADC) SAR408701 in patients (pts) with non-squamous non-small cell lung cancer (NSQ NSCLC) expressing carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5). J Clin Oncol. 2020; 38: 9505-.
36. Melton KR, Nesslein LL, Ikegami M, Tichelaar JW, Clark JC, Whitsett JA, et al. SP-B deficiency causes respiratory failure in adult mice. American journal of physiology Lung cellular and molecular physiology. 2003; 285: L543-9.
37. Taguchi A, Hanash S, Rundle A, McKeague IW, Tang D, Darakjy S, et al. Circulating pro-surfactant protein B as a risk biomarker for lung cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2013; 22: 1756-61.
38. Sin DD, Tammemagi CM, Lam S, Barnett MJ, Duan X, Tam A, et al. Pro-surfactant protein B as a biomarker for lung cancer prediction. J Clin Oncol. 2013; 31: 4536-43.
39. Xi L, Coello MC, Litle VR, Raja S, Gooding WE, Yousem SA, et al. A combination of molecular markers accurately detects lymph node metastasis in non-small cell lung cancer patients. Clinical cancer research : an official journal of the American Association for Cancer Research. 2006; 12: 2484-91.
40. Madeddu P, Emanueli C, El-Dahr S. Mechanisms of disease: the tissue kallikrein-kinin system in hypertension and vascular remodeling. Nature clinical practice Nephrology. 2007; 3: 208-21.
41. Srinivasan S, Kryza T, Batra J, Clements J. Remodelling of the tumour microenvironment by the kallikrein-related peptidases. Nature reviews Cancer. 2022; 22: 223-38.
42. Decary S, Berne PF, Nicolazzi C, Lefebvre AM, Dabdoubi T, Cameron B, et al. Preclinical Activity of SAR408701: A Novel Anti-CEACAM5-maytansinoid Antibody-drug Conjugate for the Treatment of CEACAM5-positive Epithelial Tumors. Clinical cancer research : an official journal of the American Association for Cancer Research. 2020; 26: 6589-99.
43. Sharygin D, Koniaris LG, Wells C, Zimmers TA, Hamidi T. Role of CD14 in human disease. Immunology. 2023; 169: 260-70.
44. Fang H, Huang Y, Luo Y, Tang J, Yu M, Zhang Y, et al. SIRT1 induces the accumulation of TAMs at colorectal cancer tumor sites via the CXCR4/CXCL12 axis. Cellular immunology. 2022; 371: 104458.
45. Gustafson MP, Lin Y, Bleeker JS, Warad D, Tollefson MK, Crispen PL, et al. Intratumoral CD14+ Cells and Circulating CD14+HLA-DRlo/neg Monocytes Correlate with Decreased Survival in Patients with Clear Cell Renal Cell Carcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015; 21: 4224-33.
46. Qian X, Zhang HY, Li QL, Ma GJ, Chen Z, Ji XM, et al. Integrated microbiome, metabolome, and proteome analysis identifies a novel interplay among commensal bacteria, metabolites and candidate targets in non-small cell lung cancer. Clinical and translational medicine. 2022; 12: e947.
47. Williamson DB, Sohn CJ, Ito A, Haltiwanger RS. POGLUT2 and POGLUT3 O-glucosylate multiple EGF repeats in fibrillin-1, -2, and LTBP1 and promote secretion of fibrillin-1. The Journal of biological chemistry. 2021; 297: 101055.
48. Williamson DB, Haltiwanger RS. Identification, function, and biological relevance of POGLUT2 and POGLUT3. Biochemical Society transactions. 2022; 50: 1003-12.
49. He H, He MM, Wang H, Qiu W, Liu L, Long L, et al. In Utero and Childhood/Adolescence Exposure to Tobacco Smoke, Genetic Risk, and Lung Cancer Incidence and Mortality in Adulthood. American journal of respiratory and critical care medicine. 2023; 207: 173-82.
50. Wong KK. Integrated transcriptomics and proteomics data analysis identifies CDH17 as a key cell surface target in colorectal cancer. Computational biology and chemistry. 2023; 105: 107897.
51. Kurki MI, Karjalainen J, Palta P, Sipila TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023; 613: 508-18.
52. Dotan E, Cohen SJ, Starodub AN, Lieu CH, Messersmith WA, Simpson PS, et al. Phase I/II Trial of Labetuzumab Govitecan (Anti-CEACAM5/SN-38 Antibody-Drug Conjugate) in Patients With Refractory or Relapsing Metastatic Colorectal Cancer. J Clin Oncol. 2017; 35: 3338-46.
53. Bhana N, Spencer CM. Lanoteplase. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy. 2000; 13: 217-24.
54. Gelevski D, Addy G, Rohrer M, Cohen C, Roderick A, Winter A, et al. Safety and activity of anti-CD14 antibody IC14 (atibuclimab) in ALS: Experience with expanded access protocol. Muscle & nerve. 2023; 67: 354-62.