Original Article

Identifying genetically-supported drug repurposing targets for non-small cell lung cancer through mendelian randomization of the druggable genome

Yi Feng[1,2#^], Caichen Li[1,2#], Bo Cheng[1,2#], Ying Chen[1,2#], Peiling Chen[1,2], Zixun Wang[1,2,3], Xiangyuan Zheng[1,2], Juan He[1,2,3], Feng Zhu[4], Wei Wang[1,2], Wenhua Liang[1,2,5]

[1]Dpartment of Thoracic Surgery and Oncology, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China; [2]Guangzhou Institute of Respiratory Health, Guangzhou, China; [3]Nanshan School, Guangzhou Medical University, Guangzhou, China; [4]Internal Medicine Department, Detroit Medical Center Sinai-Grace Hospital, Detroit, MI, USA; [5]Department of Oncology Medical Center, The First People's Hospital of Zhaoqing, Zhaoqing, China

Contributions: (I) Conceptualization and design: Y Feng, C Li, B Cheng, W Liang; (II) Administrative support: W Liang, W Wang; (III) Provision of study materials or patients: Y Feng, Y Chen, P Chen, Z Wang, X Zheng; (IV) Collection and assembly of data: Y Feng, B Cheng, J He, F Zhu; (V) Data analysis and interpretation: Y Feng, C Li, B Cheng, Y Chen, P Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Wenhua Liang, MD. Department of Thoracic Surgery and Oncology, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, No. 28 Qiaozhong Middle Road, Liwan District, Guangzhou 510120, China; Guangzhou Institute of Respiratory Health, No. 28 Qiaozhong Middle Road, Liwan District, Guangzhou 510120, China; Department of Oncology Medical Center, The First People's Hospital of Zhaoqing, No. 9 Donggang East Road, Duanzhou District, Zhaoqing 526000, China. Email: liangwh1987@163.com.

Background: Lung cancer is responsible for most cancer-related deaths, and non-small cell lung cancer (NSCLC) accounts for the majority of cases. Targeted therapy has made promising advancements in systemic treatment for NSCLC over the last two decades, but inadequate drug targets with clinically proven survival benefits limit its universal application in clinical practice compared to chemotherapy and immunotherapy. There is an urgent need to explore new drug targets to expand the beneficiary group. This study aims to identify druggable genes and to predict the efficacy and prognostic value of the corresponding targeted drugs in NSCLC.

Methods: Two-sample mendelian randomization (MR) of druggable genes was performed to predict the efficacy of their corresponding targeted therapy for NSCLC. Subsequent sensitivity analyses were performed to assess potential confounders. Accessible RNA sequencing data were incorporated for subsequent verifications, and Kaplan-Meier survival curves of different gene expressions were used to explore the prognostic value of candidate druggable genes.

Results: MR screening encompassing 4,863 expression quantitative trait loci (eQTL) and 1,072 protein quantitative trait loci (pQTL, with 453 proteins overlapping) were performed. Seven candidate druggable genes were identified, including *CD33*, *ENG*, *ICOSLG* and *IL18R1* for lung adenocarcinoma, and *VSIR*, *FSTL1* and *TIMP2* for lung squamous cell carcinoma. The results were validated by further transcriptomic investigations.

Conclusions: Drugs targeting genetically supported genomes are considerably more likely to yield promising efficacy and succeed in clinical trials. We provide compelling genetic evidence to prioritize drug development for NSCLC.

^ ORCID: 0000-0001-7998-656X.

Keywords: Mendelian randomization (MR); drug targets; non-small cell lung cancer (NSCLC)

Submitted Jan 18, 2024. Accepted for publication Jul 04, 2024. Published online Aug 28, 2024.
doi: 10.21037/tlcr-24-65
View this article at: https://dx.doi.org/10.21037/tlcr-24-65

Introduction

Lung cancer is the second most common cancer and the leading cause of cancer-related death worldwide (1). Despite the declining morbidity and mortality rates attributable to tobacco control measures, the incidence and prevalence of lung cancer remain high (2).

In recent years, the clinical application of targeted therapy has significantly extended the survival time of lung cancer patients, particularly those with non-small cell lung cancer (NSCLC) (3). NSCLC is characterized by possessing various genetic drivers that tyrosine kinase inhibitors can specifically target. Established actionable drivers include the epidermal growth factor receptor (*EGFR*), Kirsten rat sarcoma viral oncogene (*KRAS*), anaplastic lymphoma kinase (*ALK*), c-Ros oncogene 1 (*ROS1*), v-Raf murine sarcoma viral oncogene homolog B (*BRAF*), mesenchymal-to-epithelial transition (*MET*), and rearranged during transfection (*RET*) (4-9). Each of these gene alterations plays a crucial role in driving tumor growth and progression at the genetic level, and tyrosine kinase inhibitors tailored toward these genes have significantly improved patient outcomes by cutting off the driving power behind tumorigenesis. However, most targeted drugs under development have failed in preclinical trials, leading to a significant waste of resources. It is estimated that the average cost of successfully developing a new drug is up to $1.3 billion (10). Therefore, there is an urgent need to develop cost-effective methods to identify gene targets with greater potential for clinical success before devoting resources to clinical trials of their targeted drugs.

With the advancement of genomics, the concept of the "druggable genome" has been proposed. This concept encompasses human genes that encode drug targets, including proteins targeted by approved or drugs in clinical-trial phase, proteins similar to approved drug targets, and proteins accessible to monoclonal antibodies or drug-like small molecules *in vivo* (11). Drugs supported by genetic evidence are more likely to succeed in clinical treatment (12). Genome-wide association studies (GWAS) have been applied in drug development, increasing the proportion of preclinical stage drugs from 2.0% to 8.2% (13). However, GWAS could not directly illustrate the causal effects of genes or reliably pinpoint novel drug targets.

Two-sample mendelian randomization (MR) study is a novel approach that can evaluate causality between two traits by using genetic variants [single nucleotide polymorphisms (SNPs)] as instrumental variables (IVs). These SNPs are randomly allocated at conception and are largely independent of confounders, including environmental and other genetic variants. Therefore, the MR study may have a similar impact on evaluating causality as randomized controlled trials (14,15). A trait determined by multiple genes, each having a minor effect on the trait, is known as a quantitative trait loci (QTL). SNPs related to the expression level of druggable genes or the circulating level of encoded proteins are named expression quantitative

Highlight box

Key findings

- We identified druggable genes for non-small cell lung cancer (NSCLC), including *CD33*, *ENG*, *ICOSLG* and *IL18R1* for lung adenocarcinoma and *VSIR*, *FSTL1* and *TIMP2* for lung squamous cell carcinoma.

What is known and what is new?

- Targeted therapy has made promising advancements in systemic treatment for NSCLC over the last two decades. Current established targets include epidermal growth factor receptor (EGFR), Kirsten rat sarcoma viral oncogene (KRAS), anaplastic lymphoma kinase (ALK), c-Ros oncogene 1 (ROS1), v-Raf murine sarcoma viral oncogene homolog B (BRAF), mesenchymal-to-epithelial transition (MET), and rearranged during transfection (RET), which has benefited numerous patients with these gene alterations.
- There is an urgent need to explore other potential drug targets to expand the beneficiary group. This study identified several potential targets using the mendelian randomization of the druggable genomes.

What is the implication, and what should change now?

- Drugs with genetic support are more likely to be effective and successful in clinical trials. We provide compelling genetic evidence to prioritize drug development in NSCLC.

trait locus (eQTL) and protein quantitative trait locus (pQTL), respectively. The protein encoded by eQTL or pQTL might be analogous to a lifelong exposure that can be targeted by medication (16).

This study aimed to identify the expression levels of druggable genes and the circulating level of encoded proteins to predict the efficacy and prognostic value of the corresponding targeted drugs in NSCLC. Our research seeks to enhance the understanding of genetically repurposed drug targets for NSCLC and explore candidate druggable genes with causal association to NSCLC. We present this article in accordance with the STROBE-MR reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-65/rc) (17).

Methods

Ethical approval

As this study used publicly available data, requirement for separate ethical approval was waived, and the study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Study design

Figure 1 illustrates the entire research design scheme. We employed a two-sample MR design to identify eQTL and pQTL and predict the efficacy of the corresponding targeted drugs on NSCLC. To investigate the potential effects of druggable genes on different histological types of NSCLC, analyses were performed separately for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

Data sources

SNPs located within 1 Mb upstream and downstream of druggable genes (cis-SNPs) provide more credible genetic evidence for these genes (18). GWAS summary data of cis-eQTL were obtained from the eQTLGen consortium (https://eqtlgen.org/) (18). We obtained data on 4,863 druggable genes from the original publication (19) and selected independent cis-eQTL ($r^2 \leq 0.001$) related SNPs that are significantly associated with each druggable gene ($P<5\times10^{-8}$).

SNPs associated with circulating protein levels could model drug target effects (20). We identified three studies that met the following criteria: (I) reported significant pQTLs in individuals of European descent; (II) provided all the SNP information required for MR; and (III) included SNPs relevant to NSCLC outcomes. We obtained three independent pQTL GWAS summary datasets from INTERVAL, AGES Reykjavik, and KORA F4 studies, respectively (21-23). These large-scale pQTL studies (21-23) found that the genetic determinants of circulating proteins are located in cis to the encoding genes. This is because cis-acting SNPs strongly associated with the encoding gene tend to influence the gene's transcription and circulating protein level. The location proximity from each gene boundary to determine the cis-pQTL-related SNPs among the three studies were 10, 1 and 0.3 Mb in Suhre *et al.* (23), Sun *et al.* (21) and Emilsson *et al.* (22) respectively. The GWAS summary data for NSCLC were obtained from the Transdisciplinary Research In Cancer of the Lung and the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3) (24). These data represent the meta-analysis results aimed at identifying new lung cancer susceptibility loci.

Statistical analysis

MR study is based on three main assumptions: (I) the genetic variant(s) must be reliably associated with the exposure; (II) the genetic variant(s) must influence the outcome only through the exposure of interest; and (III) the genetic variant(s) must be independent of any measured or unmeasured confounders.

The same SNPs associated with the exposure can be extracted from NSCLC GWAS data to evaluate the causal effect. MR analyses were performed using the R package (Version 4.2.0) "TwoSampleMR" (Version 0.5.6) (25). Clumping was performed under linkage disequilibrium (LD) conditions with $r^2<0.001$ and 10,000 kb using European samples from the 1000 Genomes Project panel 3 of European-ancestry (26). The exposure and outcome data were harmonized using in-built functions. The Wald estimator ($\beta_{outcome}/\beta_{exposure}$) was used for single cis-SNPs, while multiple random-effect inverse variance weighted (IVW) analysis was used for multiple cis-SNPs (27). Bonferroni correction was applied to adjust for multiple eQTL SNPs (adjusted $P\leq0.05/4,863$).

Candidate NSCLC-influencing druggable genes supported by MR were evaluated via colocalization analyses using the coloc R package (28). Colocalization analysis assesses the probability that two traits share the same causal

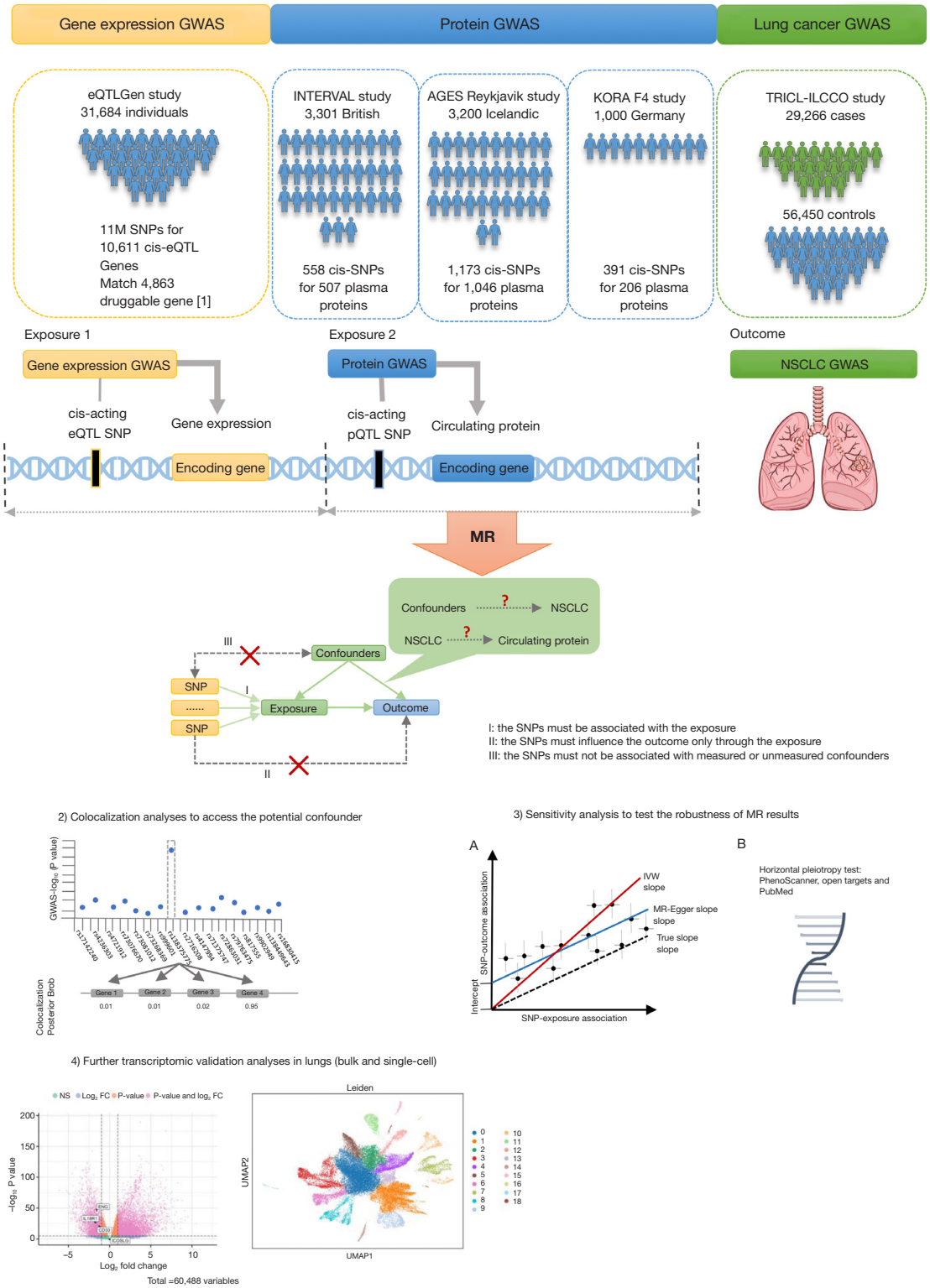

Figure 1 Overall study design. Full details of the analyses are provided in the main text and supplementary material. IVW, inverse variance weighted; MR, Mendelian randomization; GWAS, genome-wide association study; eQTL, expression quantitative trait locus; pQTL, protein quantitative trait locus; SNP, single nucleotide polymorphism; NSCLC, non-small cell lung cancer; UMAP, Uniform Manifold Approximation and Projection.

variant, rather than the variant being coincidentally shared due to correlation through LD (28). Default prior was used, including p1=10^{-4}, p2=10^{-4}, and p12=10^{-5}, where p1, p2 and p12 represent the prior probabilities that a SNP in the tested region is significantly associated with the exposure, the outcome, or both, respectively. The colocalization analysis yields posterior probabilities corresponding to one of the five hypotheses: PPH0, no association with either trait; PPH1, association with the exposure but not the outcome; PPH2, association with the outcome, but not the exposure; PPH3, association with both traits with distinct causal variants; and PPH4, association with both traits with a shared causal variant (29). The formula PPH4/(PPH3 + PPH4) is often used to represent the probability of PPH4 when only a few SNPs are included in the colocalization analysis (30). For colocalization, regions within 1 Mb of the SNPs were selected.

Sensitivity analysis

Sensitivity analyses were performed for multiple cis-SNPs following the MR analyses. Cochran's Q and I^2 methods were used to assess the overall heterogeneity among Wald ratios (31). The MR-Egger regression method relaxes the assumption by not restricting the y-intercept. Directional pleiotropy is indicated if the MR-Egger y-intercept significantly deviates from zero. If more than two SNPs were available, pleiotropic effects were evaluated using the MR-Egger regression method (32). To further detect the presence of horizontal pleiotropy, each cis-SNP of the druggable genes was examined using PhenoScanner (33) and Open Targets (34). The parameters set in the PhenoScanner database were "P value <5E–8, Proxies = EUR, $r^2 \geq 0.8$ and Build 37" for gene expression, proteins, traits, and diseases. MR analyses were then performed between these traits and NSCLC to explore potential confounders. Further assessment for potential horizontal pleiotropy and external validation of druggable genes was conducted by searching Open Targets (34). In the search result for Open Targets datasets, none of the traits related to these SNPs were associated with NSCLC. To verify whether the causal direction from the exposure to the outcome in MR results, the "mr_steiger" function in the "TwoSampleMR" package was used. The SNPs were removed if the direction was "FALSE". Bidirectional MR was performed to assess the orientation of causality.

Transcriptomic RNA sequencing (RNA-seq) data from lung tissue

The bulk RNA-seq analysis of The Cancer Genome Atlas (TCGA) pre-processed counts and fragments per kilobase million (FPKM) RNA-seq datasets for LUAD and LUSC, as well as the clinical information of each sample, were obtained from UCSC Xena platform (https://xenabrowser.net/datapages/) (35). The FPKM values were converted to the transcripts per kilobase million (TPM) value for further visualization and survival analysis. Differentially expressed genes (DEGs) were determined using R package "DESeq2" (36), all DEG analyses were conducted by comparing tumor tissue to normal tissue. Genes with $|\log_2\text{FoldChange}| > 1$ and adjusted P value <0.05 were considered as DEGs. The high- and low-expression groups were divided by the median of the TPM value of specific genes for subsequent survival analysis. Overall survival curves were constructed using the Kaplan-Meier methodology, and log-rank tests were performed to identify survival-related genes, with P value <0.05 was considered as statistically significant.

We further analyzed single-cell RNA sequencing (scRNA-seq) data. The Lambrechts *et al.* dataset (37) was downloaded from BioStudies website (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-6149). Raw scRNA-seq data were processed using CellRanger (v7.0.1, 10xGenomics). Reads were aligned to the human genome to produce gene counts across barcodes, with all CellRanger parameters set to default. Quality control (QC) and clustering were performed using Scanpy (v1.9) (38) in Python 3.8. The QC procedure was based on gene numbers, the percentage of mitochondrial genes, and ribosomal genes to remove cells with low-quality RNA-seq data. Basic filtering included genes expressed in ≥20 cells and cells with at least 200 detected genes. Doublets were removed using Scrublet (39) software. The origin counts were converted to counts per million (CPM) values and normalized for further analysis. The "pp.highly_variable_genes" function in Scanpy was used to screen the highly variable genes for Uniform Manifold Approximation and Projection (UMAP) dimension reduction. The Leiden method was adopted for clustering. Malignant cancer cells were identified using infercnvpy (https://github.com/broadinstitute/inferCNV). The marker genes of each cell population were determined using the "tl.rank_genes_groups"

function in Scanpy. Differential expression of the druggable genes was evaluated by Kruskal-Wallis' test with Dunn post hoc tests. P value was adjusted using Bonferroni method.

Results

Cohort data source

The GWAS summary data for eQTL were obtained from the eQTLGen study (18), which incorporated 37 datasets comprising a total of 31,684 individuals and 11 million SNPs. SNPs related to 4,863 druggable genes were used to model the exposure to medicines targeting the encoded proteins (19). The GWAS summary data for circulating proteins from the INTERVAL study by Sun *et al.* (21), included 3,301 participants of European descent from England. The GWAS summary data for circulating protein from the AGES Reykjavik study by Emilsson *et al.* (22), consisted of 3,200 Icelanders. The GWAS summary data for circulating protein from the KORA F4 study (23) included 1,000 Germans (Table S1).

The GWAS summary data for NSCLC were obtained from the TRICL-CCO and LC3 consortia (24). In total, 29,266 cases and 56,450 controls were included in the analyses. These data were genotyped using the OncoArray, followed by imputation and meta-analysis, with detailed procedures described previously (24). In this study, LUAD and LUSC were the two major histological subgroups in NSCLC, comprising 11,273 cases versus 55,483 controls, and 7,426 cases versus 55,627 controls, respectively. Demographic characteristics can be found in Table S2.

Statistical analysis

After anchoring 4,863 druggable genes using cis-SNPs from eQTLGen (18), the cis-SNPs of eQTL were used for MR scanning analyses in LUAD and LUSC respectively. In total, 2,491 druggable genes with available SNPs were included in the MR analyses. Of these, 106 and 136 druggable genes were causally related to LUAD and LUSC respectively (Bonferroni adjusted P value $\leq 1.03\times10^{-5}$). All analysis results are presented in the online tables (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-1.xlsx, and https://cdn.amegroups.cn/static/public/tlcr-24-65-2.xlsx). All the cis-eQTL SNPs included in this study can be found in the online tables (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-3.xlsx, and https://cdn.amegroups.cn/static/public/tlcr-24-65-4.xlsx). A total of 1,072 proteins, 453 of which were overlapping, were used for MR scanning analyses. According to MR estimates (Wald ratio or IVW) with a nominal P value <0.05, a total of 139 and 146 proteins were causally associated with LUAD and LUSC, respectively. The results can be found in the online tables (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-5.xlsx, and https://cdn.amegroups.cn/static/public/tlcr-24-65-6.xlsx), while the cis-pQTL SNPs included in this study were available at https://cdn.amegroups.cn/static/public/tlcr-24-65-7.xlsx. By intersecting the data to establish a strict causal association at both the eQTL and pQTL levels, seven candidate druggable genes were confirmed, including *CD33*, *ENG*, *ICOSLG*, *IL18R1* in LUAD and *VSIR* (also named *C10orf54* in the original study), *FSTL1*, *TIMP2* in LUSC. All these confirmed genes exhibited the same directional effect at both eQTL and pQTL level, except for *ICOSLG* and *TIMP2* (*Figure 2* and Table S3). For instance, a 1 – standard-deviation (1 – SD) genetically determined increase in CD33 expression and circulating protein levels was associated with an average 12% and 2% lower risk of developing LUAD, respectively (OR =0.88; 95% CI: 0.84–0.92; $P=2\times10^{-8}$; and OR =0.98; 95% CI: 0.97–0.99; P=0.02) (*Figure 2*). MR evidence for *ICOSLG* in LUAD was inconsistent between gene expression level (OR =1.05; 95% CI: 1.04–1.06; $P=1.39\times10^{-30}$) and circulating protein level (OR =0.992; 95% CI: 0.987–0.999; P=0.02). The causal effect of *TIMP2* in LUSC was also divided (OR =1.27; 95% CI: 1.14–1.41; $P=7.27\times10^{-6}$ at the gene expression level; OR =0.83; 95% CI: 0.72–0.95; P=0.007 at the protein level). Detailed results can be found in the online tables (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-1.xlsx, and https://cdn.amegroups.cn/static/public/tlcr-24-65-6.xlsx). It is worth mentioning that *CD33* and *VSIR* were causally related to both LUAD and LUSC at the gene expression level while *ENG* was causally associated with both LUAD and LUSC at the protein level (*Figure 2* and Table S3).

We performed colocalization analyses based on the GWAS summary data of the eQTL and pQTL for *CD33*, *ENG*, *ICOSLG*, *IL18R1 VSIR*, *FSTL1* and *TIMP2* to assess potential confounding due to LD. All these candidate genes were well colocalized with LUAD and LUSC using the coloc package (28) and the formula PPH4/(PPH3 + PPH4) with the lowest posterior probability being 98.98% (*Tables 1*,*2*).

Sensitivity analyses

After validating the druggable genes, we conducted

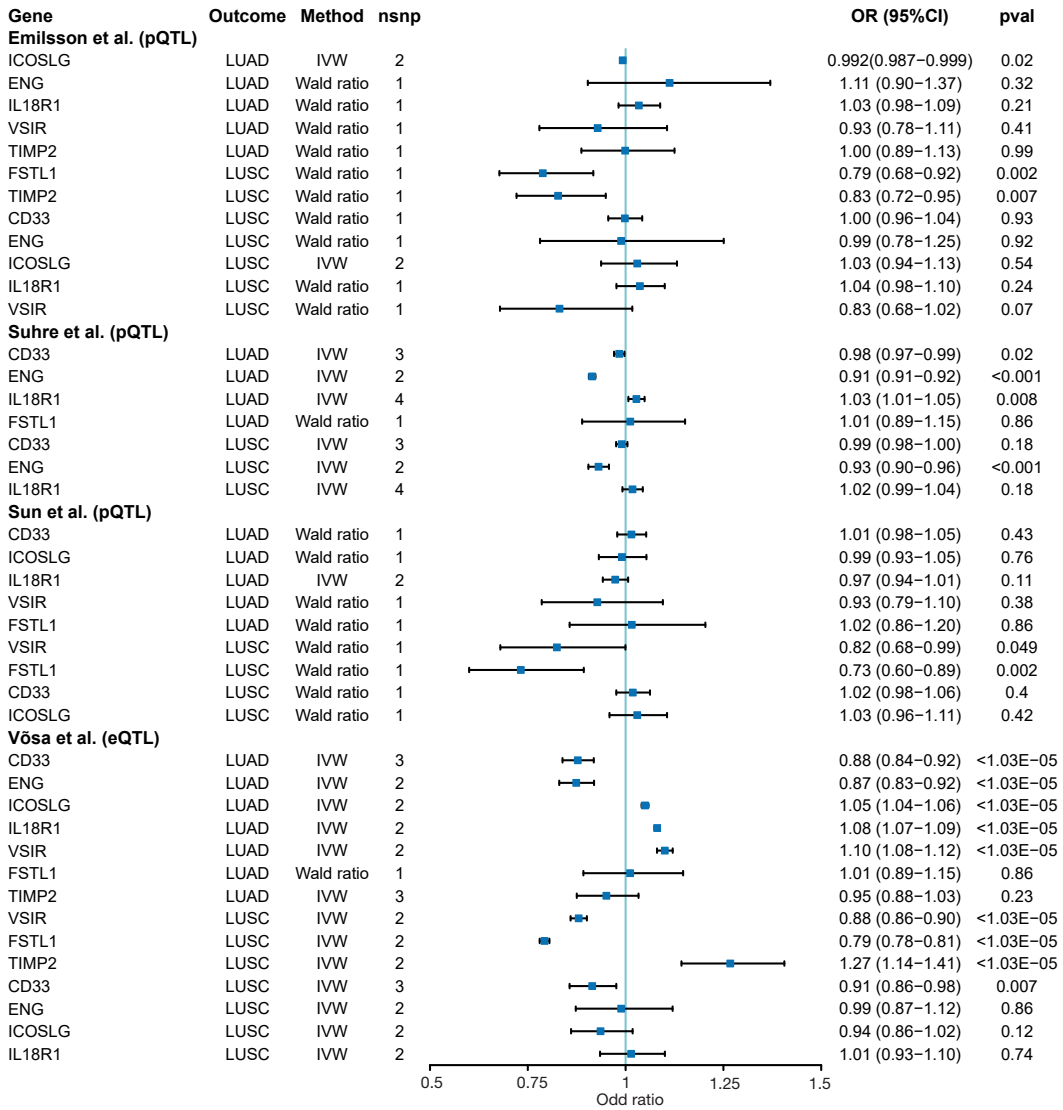

Figure 2 Genetically-supported druggable genes causally associated with LUAD and LUSC. pQTL, protein quantitative trait loci; eQTL, expression quantitative trait loci; IVW, inverse variance weighted; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OR, odds ratio; CI, confidence interval.

sensitivity analyses to assess potential heterogeneity and pleiotropy caused by confounders. All sensitivity analyses of druggable genes are shown in *Table 2*, with detailed results can be found in Table S4. All the genes passed Cochran's Q test (P>0.05). The MR-Egger intercept estimates for *CD33* and *IL18R1* were close to the null, suggesting no evidence of directional pleiotropy (P>0.05). However, *CD33*, *IL18R1* and *FSTL1* showed bias with I^2>0.50. Bidirectional MR, used as a sensitivity analysis to assess the correct orientation of MR estimates, provided no evidence that NSCLC influences these gene expression levels or protein levels in this study. The PhenoScanner dataset (33) identified some traits associated with druggable genes such as *CD33*, *ENG*, *IL18R1* and *VSIR*. For example, cis-SNP rs10421385 of *CD33* was closely related to the expression level of *SIGLEC20P*, along with rs12459419 associated with *SIGLEC22P* and *SIGLECL1*. The cis-SNP rs651007 of *ENG* was associated with multiple traits, including angiopoietin-1 receptor, beta-1,4-galactosyltransferase 1, and lithostathine-1-alpha. Detailed SNP-related traits can be found in the online table (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-8.xlsx). Since we

Table 1 MR evidence supporting druggable genes for existing drugs

Gene	Outcome (type)	eQTL GWAS					pQTL GWAS				
		Sign with Wald or IVW	Cochran's Q test	I^2 test	pQTL evidence	MR-Egger_ intercept	Sign with Wald or IVW	Cochran's Q test	I^2 test	MR-Egger_ intercept	Coloc
CD33	LUAD	√	√	×	√	√	√	√	√	√	√
ENG	LUAD	√	√	√	√	NA	√	√	×	NA	√
ICOSLG	LUAD	√	√	√	√	NA	√	√	√	NA	√
IL18R1	LUAD	√	√	×	√	NA	√	√	√	√	√
VSIR	LUSC	√	√	√	√	NA	√	√	NA	NA	√
FSTL1	LUSC	√	√	×	√	NA	√	√	NA	NA	√
TIMP2	LUSC	√	√	√	√	NA	√	√	NA	NA	√

The Wald ratio method was used when SNP =1, IVW method for SNP ≥2. These drugs are either approved or in clinical trial phase, the drug-gene interactions based on https://dgidb.org, Indications and Clinical phase based on https://clinicaltrials.gov/, the mechanisms and direction of effect were confirmed in https://go.drugbank.com/. √, pass; ×, fail to test. Coloc, colocalization; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; IVW, inverse variance weighted; MR, Mendelian randomization; eQTL, expression quantitative trait locus; GWAS, genome-wide association study; pQTL, protein quantitative trait locus; Sign, significant; SNP, single nucleotide polymorphism; NA, not applicable.

used cis-SNPs for these candidate genes, these pleiotropic effects on other molecules were more likely to represent vertical pleiotropy, where SNPs affect levels of other molecules. However, vertical pleiotropy does not violate the assumptions of MR. Hence, we conducted MR analyses between these traits (as exposure) and NSCLC (as outcome) to reduce the possibility of MR estimates bias caused by horizontal pleiotropy. When the MR estimated the expression levels of *SIGLEC20P*, *SIGLEC22P* and *SIGLECL1* on LUAD, none of these traits showed evidence of a causal effect on LUAD (P value for Wald ratio were: 0.14, 0.58 and 0.43, respectively). We did find some traits causally associated with LUAD, including: *ABO*, angiopoietin-1 receptor, and beta-1,4-galactosyltransferase 1, rs651007 and rs8176749 were identified for these traits. Detailed results can be found in the online table (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-9.xlsx).

Transcriptomic RNA sequencing analyses in lung tissue

In the bulk-RNA-seq based transcriptomic data from the TCGA dataset, the majority of candidate druggable genes were found to be differentially expressed between tumor and normal tissue, except for *ICOSLG* in LUAD and *FSTL1* in LUSC. In the LUAD dataset, *CD33*, *ENG* and *IL18R1* were significantly down-regulated in tumor tissue. In the LUSC dataset, *VSIR* and *TIMP2* were significantly down-regulated in tumor tissue (Figure S1A,S1B and Table S5). Kaplan-Meier survival analysis revealed that in LUAD, lower expression levels of *CD33* and *ENG* were associated with poor prognosis (log-rank test P=0.006 for *CD33* and 0.021 for *ENG*) (*Figure 3*), other genes-related survival analysis results can be found in Figure S2. In LUSC, higher expression levels of *TIMP2* indicated poor prognosis (log-rank test P=0.02) (*Figure 3*). Therefore, *CD33*, *ENG*, and *TIMP2* could serve as prognostic factors. ScRNA-seq analysis revealed that in certain cell clusters of the tumor microenvironment, candidate druggable genes showed significantly differential expression between tumor and normal tissue. This evidence further confirmed the drug-targeting potential of genes screened by MR methods.

Single-cell RNA sequencing data of lung tissue

Because of the lack of cellular-resolution in bulk RNA-seq analysis, we further evaluate the expression of the DEGs in scRNA-seq dataset (Figures S3-S5 and table available at https://cdn.amegroups.cn/static/public/tlcr-24-65-10.xlsx). The dataset from Lambrechts *et al.* (37) included 6 LUAD samples, 6 LUSC samples, and 4 normal tissue samples. ScRNA-seq analysis revealed that in LUAD, candidate druggable genes were mainly expressed in alveolar macrophages [with high expression of *MACRO*, *FABP4* and *MCEMP1* (40)], endothelial cells, fibroblasts, mast cells,

Table 2 Validation of druggable genes and repurposing opportunities for existing drugs using multi-dataset

Gene	PhenoScanner	Drug name	Query score	Interaction score	Indications/ usages	Clinical phase	ClinicalTrials.gov identifier/PMID	Supplement
CD33	√	Lintuzumab (SGN-33)	8.61	20.61	Acute myeloid leukemia	Phase 2	NCT00528333	
					Myelodysplastic syndrome	Phase 1	NCT00502112	
		M195	4.3	10.3	Relapsed and refractory myeloid leukemia	Phase 1B	8142644	
		AMV-564 (CNTO-3953)	4.3	10.3	Acute myeloid leukemia	Phase 1	NCT03144245	
					Locally advanced or metastatic solid tumors	Phase 1	NCT04128423	
					Myelodysplastic syndrome	Phase 1	NCT03516591	
		Oncolysin M	4.3	10.3	NA	NA	NA	
		Vadastuximab talirine (SGN-CD33A; 33A)	4.3	10.3	Acute myeloid leukemia	Phase 3	NCT02785900	Terminated (due to safety; specifically, a higher rate of deaths, including fatal infections, in the SGN33A arm versus the control arm)
					Myelodysplastic syndrome	Phase 2	NCT02706899	Terminated (sponsor decision based on portfolio prioritization)
		Gemtuzumab ozogamicin	2.15	5.15	Acute myeloblastic leukemia	Phase 4	NCT01041040	
ENG	×	Carotuximab (TRC105)	4.3	61.83	Non-small cell lung cancer	Phase 1	NCT05401110	
					Castration-resistant prostate cancer	Phase 2	NCT05534646	
					Advanced soft tissue sarcoma	Phase 2	NCT01975519	
ICOSLG	√	NA	NA	NA	NA	NA	NA	
IL18R1	√	NA	NA	NA	NA	NA	NA	
VSIR	√	Onvatilimab (JNJ-61610588)	4.3	61.83	Advanced cancer	Phase 1	NCT02671955	Terminated (Janssen business decision)
FSTL1	√	NA	NA	NA	NA	NA	NA	
TIMP2	√	NA	NA	NA	NA	NA	NA	

√, pass; ×, fail to test. NA, not applicable.

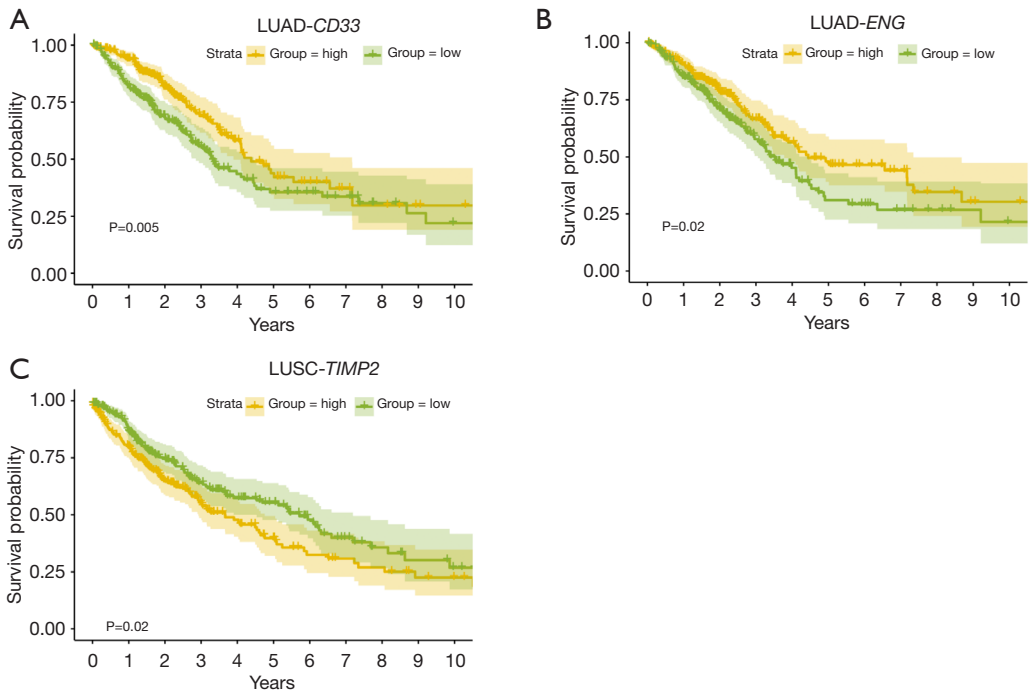

Figure 3 Overall survival curves comparing groups with high or low expression of druggable genes. (A) *CD33*; (B) *ENG*; (C) *TIMP2*. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

and monocyte-derived macrophages (40) (Figures S5A,S6). In some of these cells, druggable genes showed significantly differential expression. For instance, the expression level of *IL18R1* was lower in tumor tissue compared to normal tissue in both mast cells and endothelial cells (adjusted P value =6.79×10^{-14} for mast cell and 1.22×10^{-8} for endothelial cells). *CD33* was lower expressed in monocyte-derived macrophages in tumor tissue (adjusted P value =6.16×10^{-12}). However, *ENG* was highly expressed in endothelial cells in tumor tissue (adjusted P value =1.51×10^{-24}) (available at https://cdn.amegroups.cn/static/public/tlcr-24-65-10.xlsx). This challenged the previous conclusion obtained from bulk RNA-seq analysis; one possible explanation is the much lower abundance of alveolar macrophage in tumor tissue, leading to the overall low expression levels of *ENG* and *CD33*. In LUSC, candidate druggable genes were mainly expressed in non-tumor epithelial cells, monocyte-derived macrophage clusters [where *APOE* and *C1QB* were highly expressed, C1Q+ macrophages (41)], monocyte-derived macrophage cluster (where *CD163* was lowly expressed), alveolar macrophages, dendritic cells (DCs), and fibroblasts (Figures S5B,S7). Intriguingly, *TIMP2* was significantly lower expressed in cancer cells compared to non-tumor epithelial cells (adjusted P value =8.91×10^{-6}), suggesting that *TIMP2* may play a role in tumorigenesis and tumor cells growth. *TIMP2* was also expressed at lower levels in the macrophage cluster (adjusted P value =1.54×10^{-30}), alveolar macrophages (adjusted P value =2.01×10^{-13}) and DCs (adjusted P value =0.13). *VSIR* was down-regulated in the macrophage cluster as well (adjusted P value =8.10×10^{-7}). Overall, targeting *TIMP2* may directly affect the tumor cell in LUSC. Other druggable genes may play their roles by intervening with other tumor microenvironment components in both LUAD and LUSC.

Discussion

We conducted a two-sample MR analysis to simulate the causal effect of druggable genes on LUAD and LUSC using accessible GWAS summary data from multiple consortia. Our findings provide strong MR evidence for the repurposing value of candidate druggable genes in NSCLC, including *CD33*, *ENG*, *IL18R1* in LUAD and *VSIR*, *FSTL1* in LUSC (*Table 1* and *Figure 2*).

We identified two genes encoding the targets with existing drugs that warrant further discussion by searching DGIdb (42), Clinicaltrials, DrugBank (43) and PubMed. All results are summarized in *Table 1*. Gemtuzumab ozogamicin

is a recombinant humanized IgG4 kappa antibody against *CD33* conjugated with a calicheamicin derivative, a cytotoxic antitumor antibiotic isolated from the fermentation of *Micromonospora echinospora* ssp. (44). The Food and Drug Administration (FDA) approved this drug on May 17, 2000, as a first-line treatment for patients aged 60 years or older with *CD33*-positive acute myeloid leukemia (AML) (45). Subsequent studies have made significant progress in treating AML by targeting *CD33* in combination with other targets, such as *CD123* (46) and CAR-T cells (47). Our MR results showed that increased expression of *CD33* at the gene level can reduce the risk of NSCLC, a finding confirmed by transcriptome analysis. ScRNA-seq revealed that *CD33* is lowerly expressed in monocyte-derived macrophages of LUAD tumor tissue. However, the efficacy of the drugs targeting *CD33* in NSCLC and the specific mechanisms involved remain inexplicit. Carotuximab (TRC105) is a drug targeting the protein encoded by *ENG*, which is currently in phase I clinical trials and is also being investigated in NSCLC (NCT05401110). Other drugs or targets are either in the clinical trial stage or awaiting investigation (*Table 1*). Despite advancements in targeted therapy, discovering new genetic targets remains challenging. Many barriers to the clinical success of targeted drugs include tumor heterogeneity, genetic complexity, and technological limitations.

Since most clinically successful drugs target proteins rather than gene expression, genetic variants associated with protein levels may more accurately resemble drug target effects than eQTLs. Therefore, MR results for protein levels were prioritized for evaluation (20). Consequently, the MR evidence for *ICOSLG* in LUAD at the protein level is considered more significant.

When conducting a PhenoScanner search to assess potential confounders, we found that *CD33*, *IL18R1* and *VSIR* showed no evidence of horizontal pleiotropy. However, *ENG* influenced LUAD through diverse pathways (also known as confounders), such as ABO (48), angiopoietin-1 receptor (49), and beta-1,4-galactosyltransferase 1 (21). These analyses allowed us to explore pleiotropy due to confounders (28,32,50).

The two-sample MR design allows us to explore the expression levels and circulating proteins levels of candidate genes in LUAD (18,21-23) and LUSC (24). Each type of NSCLC has its own biological characteristics, suggesting that different druggable genes may have different causal effects. Notably, some candidate druggable genes were also supported by transcriptomic data. Moreover, *CD33*, *ENG* and *TIMP2* showed differences in the survival curves, and the direction of the effect was consistent with MR results. Therefore, the prognostic value of these genes in targeted therapy for NSCLC is worth further investigation.

Our study has several valuable advantages compared to previous MR studies. Firstly, the SNPs of eQTLs were sufficiently large to conduct robust MR analyses (18). Multiple sources of pQTL (21-23) could be mutually verified, and sufficient GWAS data were available when conducting external verification using PhenoScanner (33). Additionally, we further explored the role of druggable genes in different subtypes of the NSCLC, including LUAD and LUSC, since the current practices in LUAD and LUSC are significantly different. Secondly, this study employed a novel MR approach, using QTL-related SNPs as IVs and anchoring of the druggable gene sites to translate drug effects into gene-level effects (51). By utilizing genetic variants as IVs, MR can address confounding and reverse causation and provide more reliable insights into genetic associations. Lastly, the rapid development of genomics and the growth in publicly available druggable gene resources provide valuable genetic data for discovering novel drug targets (19,52,53). If genetic evidence from MR can shed more light on the clinical successful rate of potential targeted drugs and simplify the screening steps for the clinical trials, it could significantly lower the cost of drug development (16,54).

However, there are some limitations in this study. A key limitation is that MR cannot fully replicate the conditions of a randomized controlled trial (RCT). MR mimics lifelong low-dose exposure to a drug and assumes a linear relationship between exposure and outcome, whereas, RCTs typically investigate higher doses of a drug over a much shorter timeframe (55). The MR result may not directly correspond to the effect size in practice and does not perfectly predict the effect of a drug. Independent cohorts are required to validate scientific findings from the MR approach and eliminate false positives. Secondly, a preventative agent would need to have high tolerability and a reasonable safety profile. However, our approach is not well-suited for systematically evaluating the safety aspects of the proposed candidate genes in this study. Thirdly, the eQTL cohorts included some non-European individuals (18), and the sample size of the pQTL study is far smaller than that of the eQTL study (21-23). Larger sample sizes are needed to generate a better ability to detect QTLs. Lastly, the accuracy of the results is limited by the original data and the methodology since the causal link

in this study is based on the secondary analysis of mixed models. Despite of the limitations, two-sample MR analysis is a time- and cost-effective adjuvant to RCTs, considering the current success rate for drugs proceeding from phase I trials to approval is 13.8% approximately (56).

Conclusions

Our research provides genetic evidence of druggable genes in NSCLC, and we hope that these data will serve as a valuable resource for prioritizing drug development efforts.

Acknowledgments

Funding: This study was supported by the China National Science Foundation (82022048 & 81871893), the Key Project of Guangzhou Scientific Research Project (201804020030), National Key Research and Development Program (2022YFC2505100, 2022YFC2505105), and Guangdong Basic and Applied Basic Research Foundation (2023B1515120076).

Footnote

Reporting Checklist: The authors have completed the STROBE-MR reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-65/rc

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-65/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-65/coif). W.L. serves as the unpaid Associate Editor-in-Chief of *Translational Lung Cancer Research* from May 2024 to April 2025. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74:12-49.
2. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7-33.
3. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol 2016;11:39-51.
4. Planchard D, Jänne PA, Cheng Y, et al. Osimertinib with or without Chemotherapy in EGFR-Mutated Advanced NSCLC. N Engl J Med 2023;389:1935-48.
5. Wu YL, Dziadziuszko R, Ahn JS, et al. Alectinib in Resected ALK-Positive Non-Small-Cell Lung Cancer. N Engl J Med 2024;390:1265-76.
6. Drilon A, Camidge DR, Lin JJ, et al. Repotrectinib in ROS1 Fusion-Positive Non-Small-Cell Lung Cancer. N Engl J Med 2024;390:118-31.
7. Riely GJ, Smit EF, Ahn MJ, et al. Phase II, Open-Label Study of Encorafenib Plus Binimetinib in Patients With BRAF(V600)-Mutant Metastatic Non-Small-Cell Lung Cancer. J Clin Oncol 2023;41:3700-11.
8. Riedel R, Fassunke J, Scheel AH, et al. MET Fusions in NSCLC: Clinicopathologic Features and Response to MET Inhibition. J Thorac Oncol 2024;19:160-5.
9. Zhou C, Solomon B, Loong HH, et al. First-Line Selpercatinib or Chemotherapy and Pembrolizumab in RET Fusion-Positive NSCLC. N Engl J Med 2023;389:1839-50.
10. Galkina Cleary E, Jackson MJ, Zhou EW, et al. Comparison of Research Spending on New Drug Approvals by the National Institutes of Health vs the Pharmaceutical Industry, 2010-2019. JAMA Health Forum 2023;4:e230511.
11. Minikel EV, Painter JL, Dong CC, et al. Refining the impact of genetic evidence on clinical success. Nature 2024;629:624-9.
12. Su WM, Gu XJ, Dou M, et al. Systematic druggable genome-wide Mendelian randomisation identifies

therapeutic targets for Alzheimer's disease. J Neurol Neurosurg Psychiatry 2023;94:954-61.
13. Rasooly D, Peloso GM, Pereira AC, et al. Genome-wide association analysis and Mendelian randomization proteomics identify drug targets for heart failure. Nat Commun 2023;14:3826.
14. Zheng G, Chattopadhyay S, Sundquist J, et al. Antihypertensive drug targets and breast cancer risk: a two-sample Mendelian randomization study. Eur J Epidemiol 2024;39:535-48.
15. Bourgault J, Abner E, Manikpurage HD, et al. Proteome-Wide Mendelian Randomization Identifies Causal Links Between Blood Proteins and Acute Pancreatitis. Gastroenterology 2023;164:953-965.e3.
16. Ochoa D, Karim M, Ghoussaini M, et al. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov 2022;21:551.
17. Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA 2021;326:1614-21.
18. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet 2021;53:1300-10.
19. Finan C, Gaulton A, Kruger FA, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med 2017;9:eaag1166.
20. Folkersen L, Gustafsson S, Wang Q, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat Metab 2020;2:1135-48.
21. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. Nature 2018;558:73-9.
22. Emilsson V, Ilkov M, Lamb JR, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science 2018;361:769-73.
23. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun 2017;8:14357.
24. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 2017;49:1126-32.
25. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife 2018;7:e34408.
26. 1000 Genomes Project Consortium; Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.
27. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing Mendelian randomization investigations: update for summer 2023. Wellcome Open Res 2023;4:186.
28. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 2014;10:e1004383.
29. Wang G, Sarkar A, Carbonetto P, et al. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Series B Stat Methodol 2020;82:1273-300.
30. Zuber V, Grinberg NF, Gill D, et al. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. Am J Hum Genet 2022;109:767-82.
31. Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. Nat Rev Methods Primers 2022;2:6.
32. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32:377-89.
33. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019;35:4851-3.
34. Mountjoy E, Schmidt EM, Carmona M, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet 2021;53:1527-33.
35. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 2020;38:675-8.
36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
37. Lambrechts D, Wauters E, Boeckx B, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 2018;24:1277-89.
38. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
39. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 2019;8:281-291.e9.
40. Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun 2020;11:2285.
41. Revel M, Sautès-Fridman C, Fridman WH, et al. C1q+

macrophages: passengers or drivers of cancer progression. Trends Cancer 2022;8:517-26.
42. Freshour SL, Kiwala S, Cotto KC, et al. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res 2021;49:D1144-51.
43. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074-82.
44. Ali S, Dunmore HM, Karres D, et al. The EMA Review of Mylotarg (Gemtuzumab Ozogamicin) for the Treatment of Acute Myeloid Leukemia. Oncologist 2019;24:e171-9.
45. Bross PF, Beitz J, Chen G, et al. Approval summary: gemtuzumab ozogamicin in relapsed acute myeloid leukemia. Clin Cancer Res 2001;7:1490-6.
46. Pizzitola I, Anjos-Afonso F, Rouault-Pierre K, et al. Chimeric antigen receptors against CD33/CD123 antigens efficiently target primary acute myeloid leukemia cells in vivo. Leukemia 2014;28:1596-605.
47. Kim MY, Yu KR, Kenderian SS, et al. Genetic Inactivation of CD33 in Hematopoietic Stem Cells to Enable CAR T Cell Immunotherapy for Acute Myeloid Leukemia. Cell 2018;173:1439-1453.e19.
48. Joehanes R, Zhang X, Huan T, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol 2017;18:16.
49. Folkersen L, Fauman E, Sabater-Lleal M, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet 2017;13:e1006706.
50. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. Hum Mol Genet 2018;27:R195-208.
51. Schmidt AF, Finan C, Gordillo-Marañón M, et al. Genetic drug target validation using Mendelian randomisation. Nat Commun 2020;11:3255.
52. Gaziano L, Giambartolomei C, Pereira AC, et al. Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. Nat Med 2021;27:668-76.
53. Jacobs BM, Taylor T, Awad A, et al. Summary-data-based Mendelian randomization prioritizes potential druggable targets for multiple sclerosis. Brain Commun 2020;2:fcaa119.
54. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet 2019;15:e1008489.
55. Gill D, Cameron AC, Burgess S, et al. Urate, Blood Pressure, and Cardiovascular Disease: Evidence From Mendelian Randomization and Meta-Analysis of Clinical Trials. Hypertension 2021;77:383-92.
56. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA 2020;323:844-53.

Cite this article as: Feng Y, Li C, Cheng B, Chen Y, Chen P, Wang Z, Zheng X, He J, Zhu F, Wang W, Liang W. Identifying genetically-supported drug repurposing targets for non-small cell lung cancer through mendelian randomization of the druggable genome. Transl Lung Cancer Res 2024;13(8):1780-1793. doi: 10.21037/tlcr-24-65

Supplementary

Table S1 Demographic characteristics of the cohorts for exposure

GWAS	Sample size (n)	Ethnicity	Somkers (%)	Male (%)	Assay	Sample	PubMed ID
Gene expression GWAS							
eQTLGen study	31,684	Mixed	NA	NA	Illumina (55%), Illumina TruSeq (20.3%), Affymetrix U219 (8.7%), Affymetrix Hu-Ex v1.0ST (16%)	Whole blood, peripheral blood mononuclear cell	34475573
Proteome GWAS							
INTERVAL study	3,301	British	8.6[+]	51.1	SOMAscan	Plasma	29875488
AGES Reykjavik study	3,200	Icelandic	12[#]	42.7	SOMAscan	Serum	30072576
KORA F4 study	1,000	Germany	36.3[+]	49.8[$]	SOMAscan	Blood	28240269

[+], percentage of current smoker. [#], percentage were calculated using total participants in the AGES Reykjavik study (n=5,457). [$], percentage of total participants. GWAS, genome-wide association study.

Table S2 Characteristics of the study populations in the TRICL-ILLCO and OncoArray studies for the outcome

study	Overall		Histological types				Smoking status			
			LUAD		LUSC		Never smoking		Ever smoking	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
OnciArray	14,803	12,262	6,411	12,262	3,529	12,262	1,624	4,274	12,803	7,647
TRICL-ILCCO	14,463	44,188	4,862	43,221	3,897	43,365	731	3,230	10,420	9,317
deCODE	1,319	26,380	547	26,380	259	26,380				
GLC	481	478	186	478	97	478	35	220	433	258
Harvard	984	970	597	970	216	970	92	161	892	809
IARC	2,533	3,791	517	2,824	911	2,968	159	1,253	2,367	2,508
ICR	1,952	5,200	465	5,200	611	5,200				
MDACC	1,150	1,134	619	1,134	306	1,134			1,150	1,134
NCI	5,713	5,736	1,841	5,736	1,447	5,736	350	1,379	5,342	4,336
Toronto	331	499	90	499	50	499	95	217	236	272
Total	29,266	56,450	11,273	55,483	7,426	55,627	2,355	7,504	23,223	16,964

OnciArray, the detailed characteristics of included in OncoArray stuides can be found in original atricle's *Table S1*; deCODE, Icelandic Lung Cancer Study, Iceland; GLC, German Lung Cancer Study, US; Harvard, Harvard Lung Cancer Study, US; IRAC, the International Agency for Research on Cancer Genome-wide Association Study, France; ICR, the institute of Cancer Research Genome-wide Association Study, UK; MDACC, the MD Anderson Cancer Center Genome-wide Association Study, US; NCI, the National Cancer Institute Genome-wide Association Study, US; Toronto, the Lundenfeld-Tanenbaum Research Institute Genome-wide Association Study, Toronto, Canada. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

https://dx.doi.org/10.21037/tlcr-24-65

Table S3 Significant mendelian randomization analyses result of the druggable genes overlapping between eQTL (P≤1.03E−05) and pQTL (nominal P<0.05)

Exposure	Outcome	Gene	Method	nsnp	b	se	b_lci	b_uci	pval	OR	OR_lci	OR_uci
Suhre *et al.*	LUAD	*CD33*	Inverse variance weighted (multiplicative random effects)	3	−0.0165	0.0069	−0.0300	−0.0030	1.65E−02	0.9837	0.9705	0.9970
Sun *et al.*	LUAD	*CD33*	Wald ratio	1	0.0148	0.0186	−0.0217	0.0512	4.27E−01	1.0149	0.9786	1.0526
Võsa *et al.*	LUAD	*CD33*	Inverse variance weighted (multiplicative random effects)	3	−0.1304	0.0232	−0.1759	−0.0849	2.00E−08	0.8777	0.8387	0.9186
Emilsson *et al.*	LUSC	*CD33*	Wald ratio	1	−0.0019	0.0222	−0.0453	0.0415	9.31E−01	0.9981	0.9557	1.0424
Suhre *et al.*	LUSC	*CD33*	Inverse variance weighted (multiplicative random effects)	3	−0.0099	0.0074	−0.0244	0.0046	1.80E−01	0.9901	0.9758	1.0046
Sun *et al.*	LUSC	*CD33*	Wald ratio	1	0.0182	0.0216	−0.0241	0.0604	3.99E−01	1.0183	0.9762	1.0623
Võsa *et al.*	LUSC	*CD33*	Inverse variance weighted (multiplicative random effects)	3	−0.0893	0.0333	−0.1545	−0.0241	7.27E−03	0.9146	0.8568	0.9762
Emilsson *et al.*	LUAD	*ENG*	Wald ratio	1	0.1065	0.1062	−0.1016	0.3147	3.16E−01	1.1124	0.9034	1.3698
Suhre *et al.*	LUAD	*ENG*	Inverse variance weighted (multiplicative random effects)	2	−0.0895	0.0047	−0.0986	−0.0804	1.79E−82	0.9144	0.9061	0.9228
Võsa *et al.*	LUAD	*ENG*	Inverse variance weighted (multiplicative random effects)	2	−0.1349	0.0259	−0.1858	−0.0841	1.97E−07	0.8738	0.8304	0.9193
Emilsson *et al.*	LUSC	*ENG*	Wald ratio	1	−0.0115	0.1200	−0.2468	0.2238	9.24E−01	0.9885	0.7813	1.2508
Suhre *et al.*	LUSC	*ENG*	Inverse variance weighted (multiplicative random effects)	2	−0.0717	0.0144	−0.0999	−0.0435	6.05E−07	0.9308	0.9049	0.9574
Võsa *et al.*	LUSC	*ENG*	Inverse variance weighted (multiplicative random effects)	2	−0.0114	0.0637	−0.1363	0.1134	8.57E−01	0.9886	0.8726	1.1201
Suhre *et al.*	LUAD	*FSTL1*	Wald ratio	1	0.0119	0.0662	−0.1178	0.1415	8.58E−01	1.0119	0.8889	1.1520
Sun *et al.*	LUAD	*FSTL1*	Wald ratio	1	0.0156	0.0867	−0.1544	0.1856	8.58E−01	1.0157	0.8569	1.2039
Võsa *et al.*	LUAD	*FSTL1*	Wald ratio	1	0.0115	0.0641	−0.1142	0.1372	8.58E−01	1.0116	0.8921	1.1470
Emilsson *et al.*	LUSC	*FSTL1*	Wald ratio	1	−0.2379	0.0774	−0.3895	−0.0862	2.11E−03	0.7883	0.6774	0.9174
Sun *et al.*	LUSC	*FSTL1*	Wald ratio	1	−0.3118	0.1014	−0.5106	−0.1130	2.11E−03	0.7321	0.6001	0.8931
Võsa *et al.*	LUSC	*FSTL1*	Inverse variance weighted (multiplicative random effects)	2	−0.2323	0.0080	−0.2480	−0.2167	2.09E−186	0.7927	0.7804	0.8052
Emilsson *et al.*	LUAD	*ICOSLG*	Inverse variance weighted (multiplicative random effects)	2	−0.0075	0.0031	−0.0135	−0.0014	1.53E−02	0.9926	0.9866	0.9986
Sun *et al.*	LUAD	*ICOSLG*	Wald ratio	1	−0.0095	0.0313	−0.0708	0.0518	7.61E−01	0.9905	0.9316	1.0532
Võsa *et al.*	LUAD	*ICOSLG*	Inverse variance weighted (multiplicative random effects)	2	0.0484	0.0042	0.0401	0.0567	1.39E−30	1.0496	1.0410	1.0583
Emilsson *et al.*	LUSC	*ICOSLG*	Inverse variance weighted (multiplicative random effects)	2	0.0293	0.0480	−0.0646	0.1233	5.40E−01	1.0298	0.9374	1.1313
Sun *et al.*	LUSC	*ICOSLG*	Wald ratio	1	0.0293	0.0364	−0.0421	0.1007	4.21E−01	1.0297	0.9588	1.1059
Võsa *et al.*	LUSC	*ICOSLG*	Inverse variance weighted (multiplicative random effects)	2	−0.0662	0.0428	−0.1502	0.0177	1.22E−01	0.9359	0.8606	1.0179
Emilsson *et al.*	LUAD	*IL18R1*	Wald ratio	1	0.0332	0.0262	−0.0182	0.0846	2.06E−01	1.0338	0.9820	1.0883
Suhre *et al.*	LUAD	*IL18R1*	Inverse variance weighted (multiplicative random effects)	4	0.0272	0.0103	0.0071	0.0474	8.13E−03	1.0276	1.0071	1.0485
Sun *et al.*	LUAD	*IL18R1*	Inverse variance weighted (multiplicative random effects)	2	−0.0269	0.0169	−0.0601	0.0063	1.13E−01	0.9735	0.9417	1.0064
Võsa *et al.*	LUAD	*IL18R1*	Inverse variaSuhre *et al.*e weighted (multiplicative random effects)	2	0.0776	0.0032	0.0713	0.0839	1.90E−128	1.0807	1.0739	1.0875
Emilsson *et al.*	LUSC	*IL18R1*	Wald ratio	1	0.0358	0.0304	−0.0237	0.0954	2.38E−01	1.0365	0.9766	1.1000
Suhre *et al.*	LUSC	*IL18R1*	Inverse variance weighted (multiplicative random effects)	4	0.0174	0.0130	−0.0081	0.0429	1.81E−01	1.0176	0.9919	1.0439
Võsa *et al.*	LUSC	*IL18R1*	Inverse variance weighted (multiplicative random effects)	2	0.0141	0.0416	−0.0676	0.0957	7.36E−01	1.0142	0.9347	1.1004
Emilsson *et al.*	LUAD	*TIMP2*	Wald ratio	1	−0.0011	0.0607	−0.1200	0.1178	9.86E−01	0.9989	0.8870	1.1250
Võsa *et al.*	LUAD	*TIMP2*	Inverse variance weighted (multiplicative random effects)	3	−0.0504	0.0423	−0.1333	0.0324	2.33E−01	0.9508	0.8752	1.0330
Emilsson *et al.*	LUSC	*TIMP2*	Wald ratio	1	−0.1895	0.0701	−0.3269	−0.0520	6.89E−03	0.8274	0.7212	0.9493
Võsa *et al.*	LUSC	*TIMP2*	Inverse variance weighted (multiplicative random effects)	2	0.2370	0.0528	0.1335	0.3406	7.27E−06	1.2675	1.1428	1.4058
Emilsson *et al.*	LUAD	*VSIR*	Wald ratio	1	−0.0741	0.0892	−0.2489	0.1007	4.06E−01	0.9286	0.7797	1.1059
Sun *et al.*	LUAD	*VSIR*	Wald ratio	1	−0.0752	0.0848	−0.2413	0.0910	3.75E−01	0.9276	0.7856	1.0952
Võsa *et al.*	LUAD	*VSIR*	Inverse variance weighted (multiplicative random effects)	2	0.0955	0.0093	0.0773	0.1137	8.45E−25	1.1002	1.0803	1.1204
Emilsson *et al.*	LUSC	*VSIR*	Wald ratio	1	−0.1851	0.1031	−0.3873	0.0170	7.26E−02	0.8310	0.6789	1.0171
Sun *et al.*	LUSC	*VSIR*	Wald ratio	1	−0.1932	0.0984	−0.3860	−0.0004	4.95E−02	0.8243	0.6798	0.9996
Võsa *et al.*	LUSC	*VSIR*	Inverse variance weighted (multiplicative random effects)	2	−0.1279	0.0121	−0.1515	−0.1042	3.35E−26	0.8800	0.8594	0.9010

LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; eQTL, expression quantitative trait locus; pQTL, protein quantitative trait locus; OR, odds ratio.

https://dx.doi.org/10.21037/tlcr-24-65

Table S4 Sensitivity analyses of the genes in MR estimates

Exposure category	Outcome	Gene	Q	Q_pval	Isq	MR-Egger_ intercept_beta	MR-Egger_ intercept_se	pval
eQTL	LUAD	*CD33*	0.4340	0.8049	0.6112	0.0064	0.0381	0.8933
		ENG	0.2678	0.6048	0.4862			
		ICOSLG	0.0039	0.9502	0.0000			
		IL18R1	0.0047	0.9455	0.5243			
	LUSC	*VSIR*	0.0062	0.9370	0.0000			
		FSTL1	0.0119	0.9131	0.8162			
		TIMP2	0.0673	0.7953	0.2907			
pQTL	LUAD	*CD33*	0.4938	0.7812	0.0000	−0.0123	0.0212	0.6661
		ENG	0.0262	0.8713	0.8836			
		ICOSLG	0.0187	0.8913	0.0000			
		IL18R1	1.2326	0.7452	0.2550	−0.0034	0.0221	0.8920

eQTL, expression quantitative trait locus; pQTL, protein quantitative trait locus; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MR, Mendelian randomization.

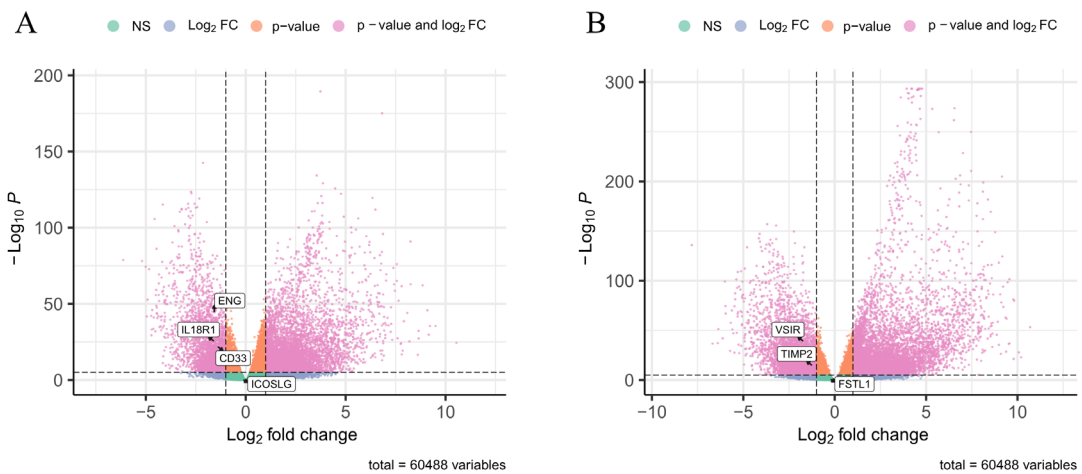

Figure S1 Volcano plot presenting the $-\log_{10}$(P) and $\log_2$(FC) of differentially expressed genes identified in LUAD and LUSC. Comparison was made between tumor tissue and normal tissue. ENG1, CD33 and IL18R1 were significantly down-regulated in tumor tissue. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; FC, fold change; NS, not significant.

https://dx.doi.org/10.21037/tlcr-24-65

Table S5 DESeq2 differential expressed genes analysis result between LUAD and LUSC tumor tissue and normal tissue

Category	EnsemblID	Symbol	BaseMean	Log_2FoldChange	LfcSE	Stat	p.value	p.adj
LUAD	ENSG00000105383	CD33	321.9144	−1.4210	0.1408	−10.0910	6.05E−24	7.38E−23
	ENSG00000106991	ENG	8396.8961	−1.5823	0.1102	−14.3548	9.95E−47	4.36E−45
	ENSG00000115604	IL18R1	282.2041	−1.5773	0.1460	−10.8042	3.29E−27	4.90E−26
	ENSG00000160223	ICOSLG	41.6021	−0.1383	0.1940	−0.7129	4.76E−01	5.58E−01
LUSC	ENSG00000035862	TIMP2	14860.8450	−1.2214	0.1517	−8.0525	8.11E−16	4.13E−15
	ENSG00000107738	VSIR	3306.6772	−1.7052	0.1275	−13.3699	9.07E−41	1.55E−39
	ENSG00000163430	FSTL1	14987.9301	−0.2431	0.1464	−1.6608	9.68E−02	1.35E−01

Only screened druggable genes were shown. $|log_2FoldChange|>1$ and adjusted P value <0.05 were considered as statistically significant. Genes with $log_2FoldChange >0$ were considered as highly expressed in tumor tissue while $log_2FoldChange <0$ were considered as lowly expressed in tumor tissue. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

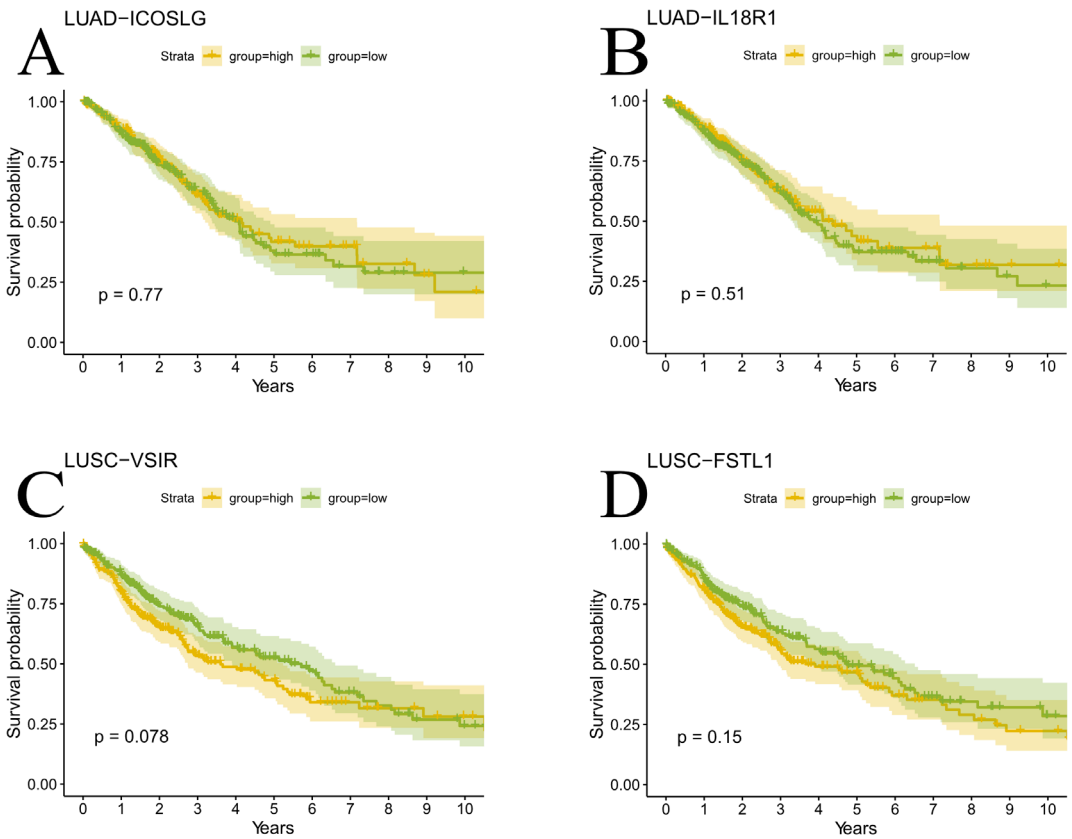

Figure S2 Overall survival curves comparing groups with high and low expression of druggable genes. (A) ICOSLG; (B) IL18R1; (C) VSIR; (D) FSTL1. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

https://dx.doi.org/10.21037/tlcr-24-65

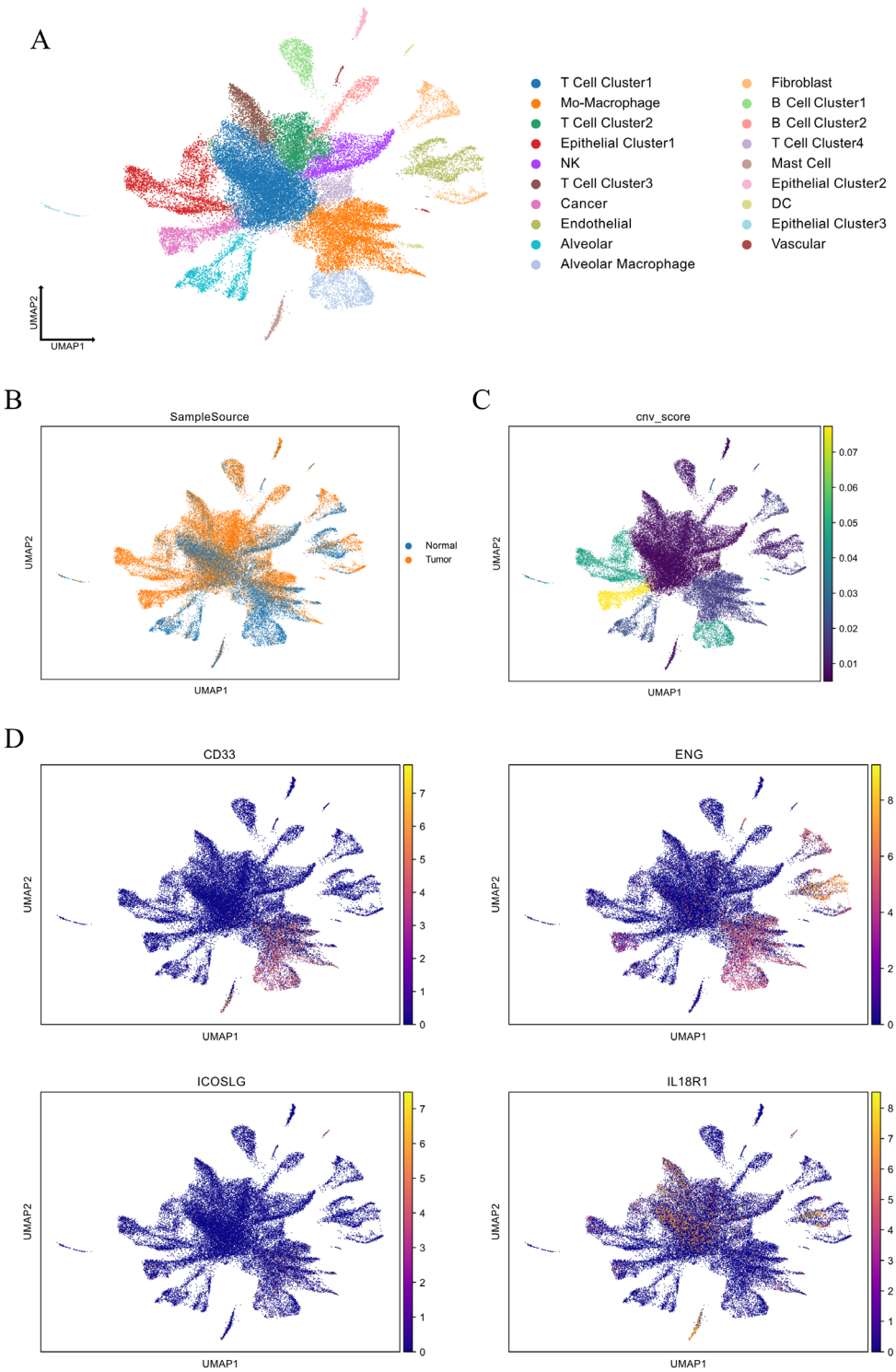

Figure S3 UMAP representations of single-cell transcriptome in LUAD. (A) Cell clusters; (B) sample came from tumor or non-tumor tissue; (C) CNV scores generated by inferCNV analysis, with high CNV scores indicating malignant cells; (D) expression levels of screened druggable genes. UMAP, Uniform Manifold Approximation and Projection; LUAD, lung adenocarcinoma; NK, natural killer cells; DCs, dendritic cells; pDCs, plasmacytoid dendritic cells; CNV, copy number variations.

https://dx.doi.org/10.21037/tlcr-24-65

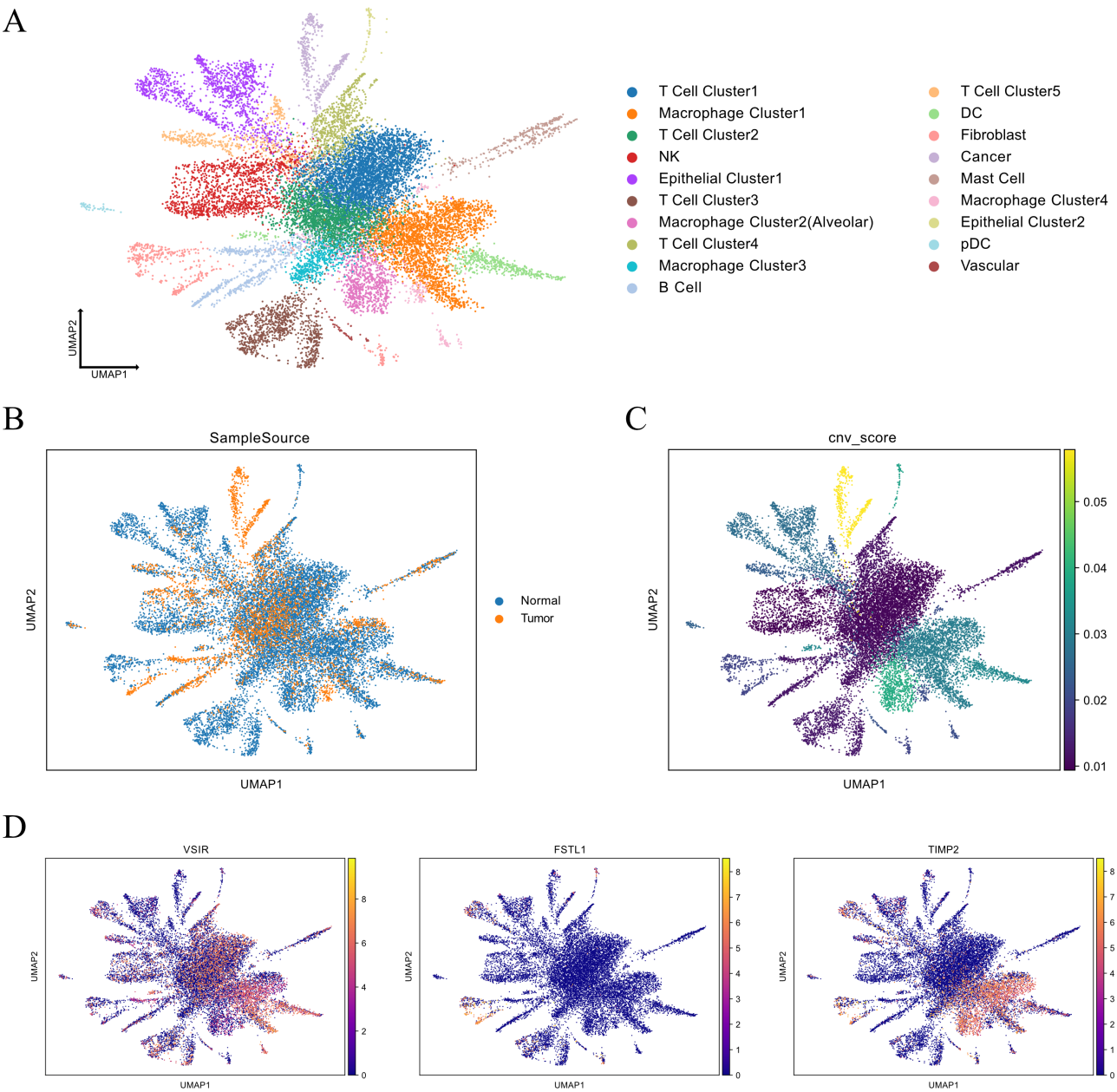

Figure S4 UMAP representations of single-cell transcriptome in LUSC. (A) Cell clusters; (B) sample came from tumor or non-tumor tissue; (C) CNV scores generated by inferCNV analysis, with high CNV scores indicating malignant cells; (D) expression levels of screened druggable genes. UMAP, Uniform Manifold Approximation and Projection; LUSC, lung squamous cell carcinoma; CNV, copy number variations.

https://dx.doi.org/10.21037/tlcr-24-65

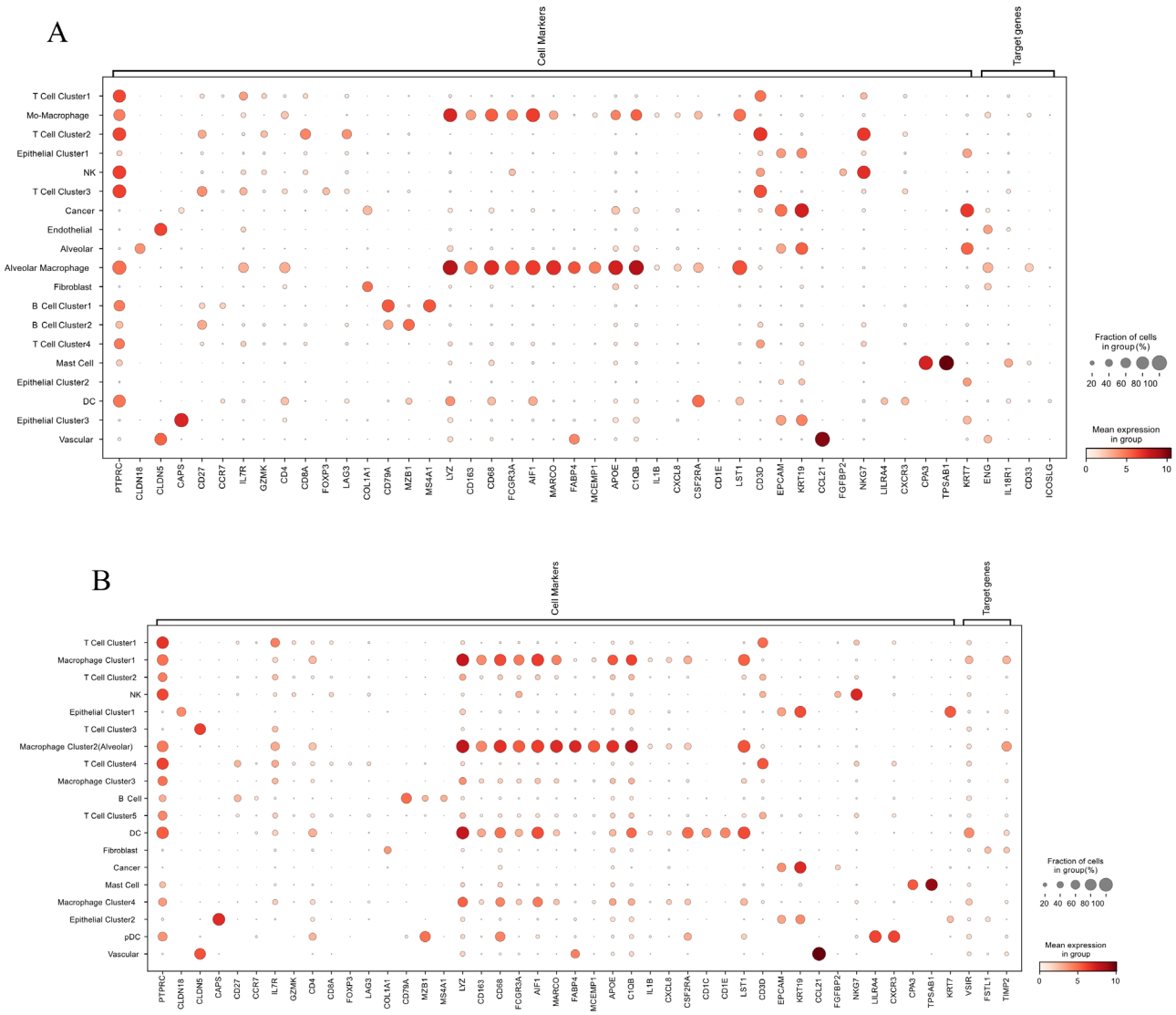

Figure S5 Dot plots showing the expression levels of selected marker genes and druggable genes in various cell subpopulations in LUAD (A) and LUSC (B). The left column presents the cell subtypes identified based on Leiden clustering method. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

https://dx.doi.org/10.21037/tlcr-24-65

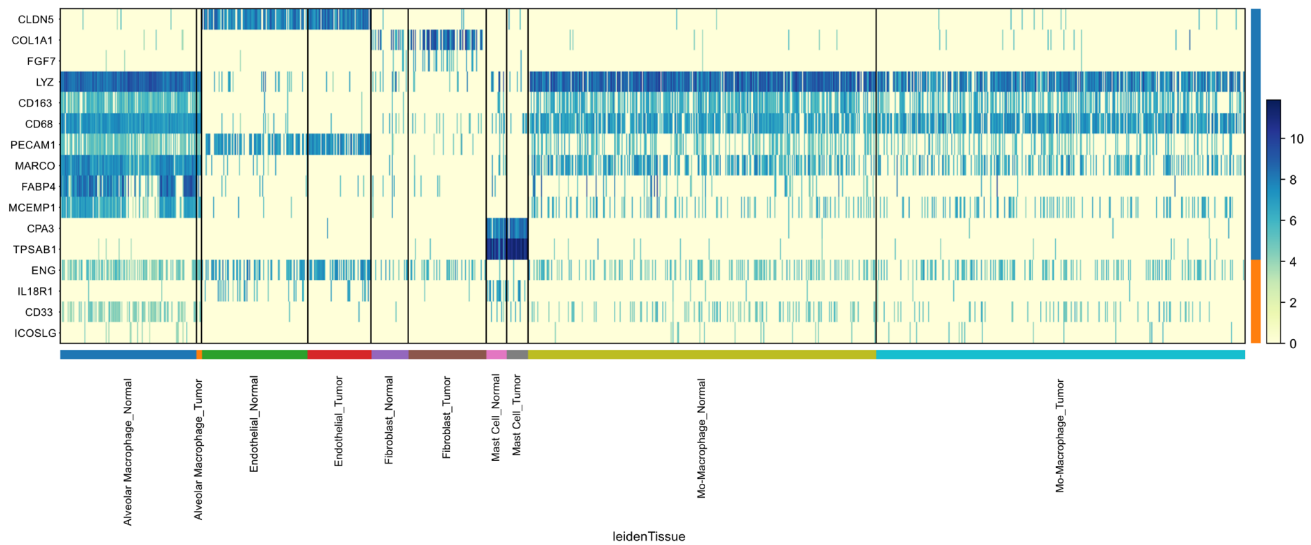

Figure S6 Heatmap showing the expression levels of marker genes and druggable genes in specific cell cluster of tumor or normal tissue in LUAD. Only cell clusters with high expression levels of druggable genes are displayed. LUAD, lung adenocarcinoma.

Figure S7 Heatmap showing the expression levels of marker genes and the druggable genes in specific cell cluster of tumor or normal tissue in LUSC. Only cell clusters with high expression levels of druggable genes are displayed. LUSC, lung squamous cell carcinoma.

https://dx.doi.org/10.21037/tlcr-24-65