Integrative Analysis of GWAS and Single-Cell eQTL Data Identifies Immune Cell-Type-Specific Genetic Signals Co-occurring in Inflammatory Diseases and Lung Cancer Risk

Author: Yu Langxuan[1], Wen Zhe[2]

ABSTRACT

Background

Non-small cell lung cancer (NSCLC) frequently co-occurs with chronic inflammatory conditions such as COPD, RA, IPF, psoriasis, and dermatophytosis. While epidemiological links are established, the genetic basis of this co-occurrence in specific cell types remains poorly resolved.

Methods

We applied a signal-driven fine-mapping approach to NSCLC GWAS index variants ($p < 1 \times 10^{-}$ □), integrating single-cell eQTL (sc-eQTL) data across cell types within ±100 kb windows to identify stable, compact variant sets that jointly explain GWAS and sc-eQTL signals. Loci with variant sets that optimally account for both NSCLC and sc-eQTL signals were evaluated for directional consistency between GWAS and sc-eQTL effects. In a subset of these loci, we performed locus-specific Mendelian randomization to evaluate whether the same set of genetic variants that influence risk of inflammatory comorbidities also show directionally concordant effects on NSCLC risk.

Results

Hundreds of genomic loci with robust concordance between NSCLC risk and cell type–specific gene regulation were identified. Directional analysis revealed consistent regulatory effects in specific immune contexts, and Mendelian randomization supported shared genetic liability between multiple inflammatory diseases and NSCLC risk.

This integrative genetic analysis was conducted independently by the first author.
Correspondence regarding this work will be available upon formal publication. For inquiries, contact: yulangxuan@sjtu.edu.cn

Conclusion

Integrative analysis of GWAS and single-cell eQTL data enables the identification of shared regulatory architectures linking inflammatory diseases to lung cancer risk. This framework may support the role of immune cell-type-specific regulation in disease etiology and may inform strategies for risk stratification and mechanistic investigation.

Background

Non-Small Cell Lung Cancer (NSCLC) is the leading cause of cancer death worldwide, accounting for approximately 18% of all cancer-related mortality[1]. Despite advances in early detection and therapy, over 60% of patients present with advanced or metastatic disease, and 5-year survival remains below 20%[2]. Prevention through risk factor modification may offer a more effective strategy to reduce global burden.

Recent efforts have turned to single-cell expression quantitative trait locus (sc-eQTL) data to pinpoint the cell types and target genes through which genetic risk variants operate[3]. However, current sc-eQTL resources are inherently fragmented: public releases usually typically report only variant–gene pairs with nominal p values below 0.05, while the majority of tested associations are omitted[3–9]. This selective reporting results in incomplete and sparse regulatory maps across cell types. Consequently, conventional colocalization methods such as COLOC, which assume dense SNP coverage, consistent linkage disequilibrium (LD) structure, and homogeneous variant representation across datasets, often yield inaccurate or irreproducible fine-mapping results when applied to NSCLC GWAS in conjunction with uncomplete sc-eQTL data[10,11].

Besides, Chronic inflammatory diseases including IPF, dermatomyositis, COPD, RA, and psoriasis are epidemiologically linked to increased lung cancer risk[12–16]. Chronic inflammation promotes carcinogenesis through immune activation and tissue remodeling. Mendelian randomization studies suggest potential causal roles, implicating shared genetic liability[17–21]. However, these analyses operate at the trait level and do not resolve specific variants or cell-type-specific regulatory mechanisms. the coarse granularity of Mendelian randomization analysis and sc-eQTL data motivate a shift from epidemiological observation to cell-type-resolved molecular mechanism. Although sc-eQTL data theoretically enable such resolution, their incompleteness risks propagating bias rather than clarifying causality.

To address these challenges, we introduce an integrative framework engineered to operate under fragmented genomic architectures. Leveraging high-resolution sc-eQTL data despite their sparsity, our approach prioritizes shared risk variants with cell-type-specific regulatory potential between NSCLC and chronic inflammatory diseases. By combining signal-driven variant selection[22], cross-modal concordance quantification, directional effect assessment[23,24], and locus-specific Mendelian Randomization[25,26],

our method identifies high-confidence loci where genetic effects on gene expression align with both NSCLC risk and inflammatory disease susceptibility, thereby bridging the gap between association and mechanism in the context of realistic, incomplete genomic data.

Anchoring on GWAS variants with suggestive association ($p < 1 \times 10^{\square\square}$), we integrate local sc-eQTL signals within ±100 kb windows and apply a signal-driven variant set prioritization procedure. This approach identifies compact and stable sets of variants that collectively explain the maximum shared genetic signal across GWAS and sc-eQTL modalities, explicitly accounting for the sparse and heterogeneous structure typical of public sc-eQTL resources. To ensure prioritized loci are relevant to NSCLC–inflammation comorbidity, we further filter them based on significant enrichment of associations in independent inflammatory disease GWAS summary statistics, serving as a phenotype-informed relevance screen.

Method

1. Signal-Driven integration of GWAS and Fragmented sc-eQTL

Conventional colocalization methods (e.g., COLOC) assume complete SNP coverage, consistent linkage disequilibrium (LD) structure, and homogeneous variant representation across datasets, which is rarely satisfied in cross-study integration of single-cell eQTL (sc-eQTL) data[10,11]. To address this, we performed a signal-driven integrative variant prioritization framework designed to operate under the realistic constraints of fragmented single-cell eQTL data. For each candidate locus, defined as a ±100□kb genomic window centered on a GWAS lead SNP ($p \square < \square 1 \square \times \square 10^{\square\square}$),we include variants that satisfied all of the following criteria: (i) present in both GWAS and single-cell eQTL (sc-eQTL) summary statistics; (ii) reported as significant cis-sc-eQTLs (nominal $p \square < \square 0.05$ in the original sc-eQTL study); (iii) with LD estimates available from the 1000 Genomes Phase 3 European reference panel (1000G EUR). Overlapping loci were analyzed independently to prevent conflation of distinct regulatory architectures. All variants were harmonized to GRCh37 (hg19) using the UCSC liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We queried the GRCh37 reference via the Ensemble REST API (https://grch37.rest.ensembl.org) and only retained variants with alleles consistent across datasets. Only variants with consistent allele coding and forward-strand alignment were retained for downstream analysis.

To quantify the degree of shared genetic signal between GWAS and sc-eQTL associations, we adopted a symmetric, energy-based heuristic termed the Signal Overlap Proportion (SOP). SOP is defined as the average of two directional ability metrics: the proportion of total GWAS association signal that explained by the selected sc-eQTL-

derived causal variant set; and the proportion of total sc-eQTL signal explained by the GWAS-derived causal variant set.

Formally:

$$\mathrm{SOP} = \frac{1}{2}\left(\frac{E^{\mathrm{by}\,S_{\mathrm{QTL}}}_{\mathrm{GWAS,explained}}}{E_{\mathrm{GWAS,total}}} + \frac{E^{\mathrm{by}\,S_{\mathrm{GWAS}}}_{\mathrm{QTL,explained}}}{E_{\mathrm{QTL,total}}}\right)$$

where each $E_{explained}$ is computed as the cumulative signal energy accounted for by the selected variant set during the greedy conditional Z-selection procedure, and all energies are evaluated in the same spectrally truncated subspace to ensure comparability. We use SOP ≥ 0.7 as an empirical threshold to prioritize loci with substantial cross-modal signal overlap.

A complete description of the procedure is provided in the ***Supplementary Methods***. A theoretical justification of the proposed framework is provided in ***Supplementary Note 1***, including the statistical consistency of variance-normalized signal energy estimation, the validity of block-level LD definitions, the convergence of spectral regularization, and the stability under incremental SNP addition. By this design, our framework serves as an alternative in the presence of fragmented QTL data, thereby enabling inference in realistic genomic settings where traditional methods may lack reproducibility. Custom analysis scripts implementing the procedures described in this study are available at: https://github.com/SpaceRan/Signal-Driven-integration-of-GWAS-and-Fragmented-sc-eQTL. The analytical pipeline is illustrated in Figure 1

Fig 1a. Workflow of the analytical pipeline.

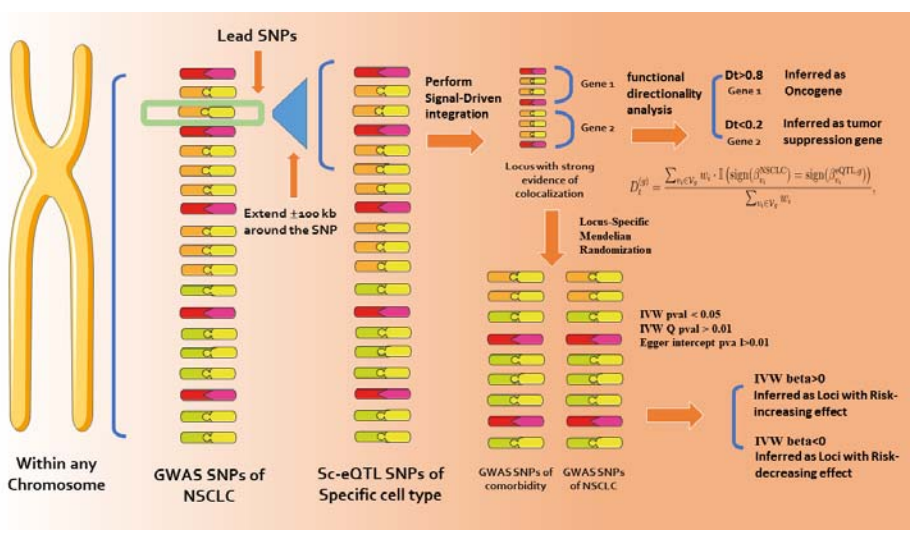
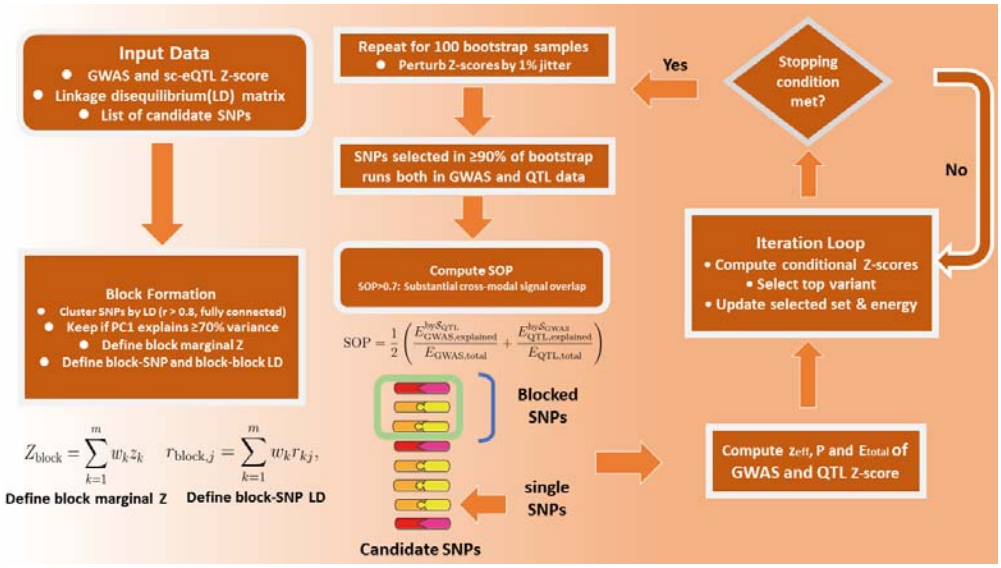


Fig 1b. Workflow of the variant prioritization.



"Substantial cross-modal signal overlap" indicates that the genetic associations from GWAS and sc-eQTL analyses show strong reciprocal support within a locus, suggesting a shared underlying causal variant or regulatory mechanism.

2. Directional Concordance Between GWAS and sc-eQTL Effects

For each genomic locus satisfying SOP ≥ 0.7, we further assessed the functional directionality of the shared genetic signal by computing a weighted direction agreement score Dt for every gene with significant sc-eQTL support within that locus[24]. Specifically, let Vg denote the set of variants in the locus that are cis-associated with gene g in our sc-eQTL data. For each such gene, we define:

$$D_t^{(g)} = \frac{\sum_{v_i \in \mathcal{V}_g} w_i \cdot \mathbb{I}\left(\text{sign}(\beta_{v_i}^{\text{NSCLC}}) = \text{sign}(\beta_{v_i}^{\text{eQTL},g})\right)}{\sum_{v_i \in \mathcal{V}_g} w_i},$$

where wi=| | reflects the marginal GWAS effect size of variant vi , and wi=|

| is the effect of vi on the expression of gene. The indicator function I(·) equals 1 when the GWAS and eQTL effects are directionally concordant (i.e., both increase or both decrease risk/expression), and 0 otherwise.

This gene-specific score quantifies, within a single high-confidence locus, whether NSCLC risk-increasing alleles tend to upregulate or downregulate each candidate target gene. We interpret Dt(g)≥0.8 as higher NSCLC risk is associated with increased expression of gene g , suggestive of a potential oncogenic role. Conversely, Dt(g)≤0.2 suggests that risk alleles are linked to reduced expression, consistent with tumor-

suppressive function. Genes with intermediate scores were considered ambiguous and excluded.

3. Locus-Specific Mendelian Randomization comorbidity and NSCLC Risk

To evaluate whether loci prioritized by SOP (SOP $\geq$ 0.7) harbor genetic architecture consistent with a directional biological link between specific comorbid diseases (idiopathic pulmonary fibrosis (IPF), dermatomyositis, chronic obstructive pulmonary disease (COPD), rheumatoid arthritis (RA), and psoriasis) and NSCLC risk, we performed locus-specific Mendelian randomization using the set of variants identified by our stability-filtered fine-mapping procedure. For compatibility with Mendelian randomization, we decomposed block-represented signals back to the constituent SNPs. All variants were harmonized to GRCh37 (hg19) and filtered for strand consistency as described in Section 1.1. Non-standard variants (indels, multi-allelic SNPs) were excluded to ensure unambiguous effect direction mapping. We then applied instrument strength filtering ($F > 10$) to derive a final set of candidate instruments per locus. We then applied Mendelian randomization to serve as a hypothesis-generating evaluation to evaluate consistency of genetic effects across diseases within specific loci[27].

For each SOP locus, we began with the set of bootstrap-stable fine-mapped variants and performed formal Mendelian randomization with them. Variants were harmonized to the forward strand, with alleles flipped where necessary to ensure consistent effect direction between exposure and outcome. All variants within each fine-mapped locus were retained for Mendelian randomization analysis, regardless of individual F-statistics, to preserve the complete posterior probability structure and avoid arbitrary exclusion of potentially causal variants. No LD-based clumping was applied, as SNPs within each locus were intentionally retained to jointly model the regional genetic architecture. For quality control and bias assessment, we calculated and reported both the mean and minimum F-statistic across SNPs within each locus. Loci with mean F-statistic < 10 were flagged for sensitivity analysis to evaluate robustness. We proceeded with causal effect estimation using the inverse-variance weighted method, supplemented by MR-Egger to evaluate robustness. Heterogeneity was assessed using Cochran's Q statistic, and directional pleiotropy was tested via the MR-Egger intercept. When fewer than three instruments were available, formal MR was not performed.

Result

1. Data resources

Summary-level GWAS data for all outcomes and exposures were selected from studies that included diverse ancestral groups; however, we restricted analyses to summary statistics derived exclusively from individuals of European ancestry, as reported in the original publications or metadata. To minimize potential sample overlap between exposure and outcome datasets, we excluded studies sharing overlapping cohorts based on study descriptions and available metadata. Linkage disequilibrium (LD) estimates for European populations were derived from the 1000 Genomes Phase 3 reference panel (EUR) [28].

In total, we incorporated **15** distinct GWAS datasets across six major phenotypes: **2** for chronic obstructive pulmonary disease (COPD), 1 for Idiopathic pulmonary fibrosis (IPF), 1 for dermatophytosis,**2** for psoriasis, 6 for rheumatoid arthritis, and **4** for Non-small-cell lung cancer. Comprehensive details on GWAS data sources, including sample sizes, phenotype definitions, and data access information, are provided in ***Table 1***. Single-cell eQTL data was drawn from **10** studies, covering **150** different cell types, including monocytes, macrophages (alveolar and interstitial), CD4+ T cells, CD8+ T cells, regulatory T cells, B cells, dendritic cells, NK cells, neutrophils, epithelial cells, and among others. All datasets underwent genomic coordinate conversion and strand harmonization as described in Methods. Single-cell QTL data were obtained from studies in biologically relevant cell types; full details are summarized in ***Supplementary Table 1*.**

2. Signal-Driven integration of GWAS and sc-eQTL data

We identified 4795 GWAS lead SNPs with p value less than 1×10^{-5} that overlapped with at least one significant cis sc-eQTL with nominal p value less than 0.05 across 10 single cell eQTL datasets and 4 NSCLC GWAS datasets. For each lead SNP we defined a ±100 kb candidate locus yielding 4795 genomic regions for analysis as described in Methods. LD information from the 1000 Genomes Phase 3 European panel was available for 100% of variants per locus.

After applying our signal driven prioritization framework we obtained stable variant sets in both GWAS and sc-eQTL data for 803 of the 4795 loci. These 803 loci each yielded a stable variant set in GWAS and a stable variant set in sc-eQTL through independent selection procedures. In these loci the Signal Overlap Proportion SOP which quantifies the degree of shared signal between the two modalities had a median value of 0.337 with interquartile range G to H. Notably 258 loci exhibited strong signal concordance with SOP greater than or equal to 0.7.

We next assessed directional concordance between GWAS and sc-eQTL effects in the 258 high-SOP loci. Overlapping loci were analyzed independently (Methods1.3), and many corresponded to the same gene or regulatory region. For each gene with significant cis-sc-eQTL support within a locus, we computed a weighted direction agreement score $Dt(g)$ (Methods, Section 1.3). After filtering for unambiguous directional relationships ($Dt(g) \geq 0.8$ or $Dt(g) \leq 0.2$), we identified 17 distinct cell type–gene pairs (***Table 2***). In each pair, all prioritized variants within the corresponding locus showed consistent effect directions, indicating that NSCLC risk-increasing alleles either uniformly upregulated or downregulated the target gene.

Table 2 lists cell type–gene pairs that show strong directional concordance (D□ ≥ 0.8 or ≤ 0.2) between NSCLC risk and sc-eQTL effects, together with the inferred functional role of each gene. For most genes, the inferred role is consistent across different cell types. ACTN1 is inferred as an oncogene in multiple T cell subsets. CTSH is uniformly classified as a tumor suppressor in both monocytes and dendritic cells. HYKK and MORF4L1 each appear in a single cell type but align with the broader pattern of unambiguous directional inference, with HYKK suggesting an oncogenic role and MORF4L1 a tumor-suppressive function. Similarly, PSMA4 is consistently classified as a tumor suppressor across several CD4+ T cell states, while IREB2 and CHRNA3 are uniformly inferred as oncogenes in their respective cellular contexts.

A notable exception is CHRNA5, which is inferred as a tumor suppressor in CD4 memory T cells after acquiring effector functions at 5 days, but as an oncogene in CD4 naive T cells cycling after the first cell division at 40 hours and in normal monocytes. This discrepancy suggests that the functional interpretation of CHRNA5 in NSCLC risk may depend on cellular context.

3. Locus-Specific Mendelian Randomization

We performed locus-specific Mendelian randomization using genomic loci with SOP score at least 0.7, a threshold indicating robust colocalization between non-small cell lung cancer genome-wide association signals and single-cell expression quantitative trait loci (sc-eQTL) in specific cell types. Only loci with inverse-variance weighted p-value less than 0.05, Cochran's Q p-value greater than 0.01, and MR-Egger intercept p-value greater than 0.01 were interpreted. Results are shown in Table 3.

For chronic obstructive pulmonary disease, rheumatoid arthritis, and psoriasis, most loci indicated a risk-increasing effect on non-small cell lung cancer. These loci colocalized with sc-eQTL in immune cell types including CD4+ T cell subsets and monocytes.

Several loci were shared across comorbid diseases. The rs7166764 locus colocalized with sc-eQTL in monocytes support a risk-increasing effect in both psoriasis and rheumatoid arthritis background on non-small cell lung cancer. The rs34138960 locus colocalized with sc-eQTL in CD4+ memory T cells after acquisition of effector functions similarly indicated increased risk effect from both rheumatoid arthritis and chronic

obstructive pulmonary disease background. In all shared loci, the direction of effect on non-small cell lung cancer risk was consistent across diseases and cell types.

Notably, the rs4902649 locus colocalized with sc-eQTL in CD4+ memory T cells after 5 days of activation and was associated with a reduced risk of non-small cell lung cancer for both dermatomyositis and rheumatoid arthritis, representing a rare protective pattern among the observed signals.

Although some loci associated with different diseases or identified in distinct cell types overlapped, they consistently showed the same direction of effect on non-small cell lung cancer risk. This directional concordance across cell types, diseases, and sc-eQTL datasets reinforces the robustness of these cell-type-specific regulatory relationships in mediating comorbidity-driven susceptibility to lung cancer.

Table 1 Summary of GWAS Data

Phenotype Definition	Consortium / Study	Sample Size (N)	Ancestry	Data Access
COPD, Early/Later Onset	IEU	361194	European	https://gwas.mrcieu.ac.uk/datasets/ukb-d-COPD_EARLYANDLATER/
COPD	FinnGenR12	433208	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_J10_COPD.gz
IPF	FinnGenR12	52283	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_IPF.gz
Dermatophytosis	FinnGenR12	494048	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_AB1_DERMATOPHYTOSIS.gz
Psoriasis	IEU	462933	European	https://gwas.mrcieu.ac.uk/datasets/ukb-b-10537/
Psoriasis	FinnGenR12	494941	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_L12_PSORIASIS.gz
Rheumatoid Arthritis	IEU	463010	European	https://gwas.mrcieu.ac.uk/datasets/ukb-b-11874/
Rheumatoid Arthritis	IEU	361194	European	https://gwas.mrcieu.ac.uk/datasets/ukb-d-M13_RHEUMA/
Rheumatoid Arthritis	IEU	337159	European	https://gwas.mrcieu.ac.uk/datasets/ukb-a-105/
Rheumatoid Arthritis	IEU	57284	European	https://gwas.mrcieu.ac.uk/datasets/ebi-a-GCST002318/
Rheumatoid Arthritis	FinnGenR12	331429	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_M13_RHEUMA.gz
NSCLC	IEU	374687	European	https://gwas.mrcieu.ac.uk/files/ieu-b-4954/ieu-b-4954.vcf.gz
NSCLC	IEU	212453	European	https://gwas.mrcieu.ac.uk/files/bbj-a-133/bbj-a-133.vcf.gz
NSCLC	IEU	337159	European	https://gwas.mrcieu.ac.uk/files/ukb-a-54/ukb-a-54.vcf.gz
NSCLC	FinnGenR12	385195	European	https://storage.googleapis.com/finngen-public-data-r12/summary_stats/release/finngen_R12_C3_LUNG_NONSMALL_EXALLC.gz

Table 2 Distinct cell type–gene pairs with directional concordance (D□ ≥ 0.8 or ≤ 0.2) between NSCLC risk and sc-eQTL effects

Gene	Inferred functional role	Cell type
ACTN1	tumor suppressor gene	CD4 Memory after acquiring effector functions (5 d)
CHRNA3	oncogene	Monocyte (Mono) Normal
CHRNA5	oncogene	CD4+ Naive T cycling after the first cell division (40h)
CHRNA5	oncogene	Monocyte (Mono) Normal
CHRNA5	tumor suppressor gene	CD4 Memory after acquiring effector functions (5 d)
CTSH	tumor suppressor gene	Monocyte (Mono)
HYKK	oncogene	CD4 Memory after acquiring effector functions (5 d)
IREB2	oncogene	CD4+ Naive T cycling after the first cell division (40h)
IREB2	oncogene	Monocyte (Mono)
IREB2	oncogene	Monocyte (Mono) Normal
MORF4L1	tumor suppressor gene	B Cell
PSMA4	tumor suppressor gene	CD4+ Central Memory T before dividing (16h)
PSMA4	tumor suppressor gene	CD4+ Naive T cycling after the first cell division (40h)
PSMA4	tumor suppressor gene	CD4+ Naive T harvested after 16 h stimulation
PSMA4	tumor suppressor gene	CD4 Memory after acquiring effector functions (5 d)
PSMA4	tumor suppressor gene	CD4 Memory before dividing (16h)

Table 3 Locus-specific Mendelian randomization supports causal links between comorbidity and NSCLC risk in specific cell type–

Comorbidity	MR effect direction	Cell type	Sc-eQTL source	Lead SNPs
COPD	Loci with Risk-increasing effect	CD4+ Central Memory T before dividing (16h)	Soskic2022-Natp	rs1394371
COPD	Loci with Risk-increasing effect	CD4+ Naive T harvested after 16 h stimulation	Soskic2022-Natp	rs1394371, rs1504550, rs17484235, rs2009746, rs72736802, rs72738704, rs72738732, rs72738736, rs951984, rs951985
COPD	Loci with Risk-increasing effect	CD4 Memory after acquiring effector functions (5 d)	Soskic2022-Natp	rs28498264, rs34138960, rs34664138, rs35212593, rs3813570, rs57064725, rs59133824, rs59683676
COPD	Loci with Risk-increasing effect	CD4 Memory before dividing (16h)	Soskic2022-Natp	rs2009746, rs72736802, rs72738704, rs72738732, rs951984, rs951985
COPD	Loci with Risk-increasing effect	Monocyte (Mono)	Schmiedel2022-Scip	rs7166764
COPD	Loci with Risk-increasing effect	Monocyte (Mono) Normal	Schmiedel2022-Scip	rs4887059
Dermatomyositis	Loci with Risk-**decreasing** effect	CD4 Memory after acquiring effector functions (5 d)	Soskic2022-Natp	rs4902649
Psoriasis	Loci with Risk-increasing effect	Monocyte (Mono)	Schmiedel2022-Scip	rs7166764

rheumatoid arthritis	Loci with Risk-**decreasing** effect	CD4 Memory after acquiring effector functions (5 d)	Soskic2022-Natp	rs4902649
rheumatoid arthritis	Loci with Risk-increasing effect	CD4 Memory after acquiring effector functions (5 d)	Soskic2022-Natp	rs34138960, rs34664138, rs35212593, rs3813570, rs57064725, rs59133824, rs59683676, rs28498264,
rheumatoid arthritis	Loci with Risk-increasing effect	Monocyte (Mono)	Schmiedel2022-Scip	rs7166764

Discussion

Lung cancer remains the leading cause of cancer related mortality worldwide, with a persistently low five-year survival rate largely attributable to late stage diagnosis[2]. While tobacco exposure is the predominant risk factor, a growing body of epidemiological evidence indicates that chronic inflammatory diseases, including chronic obstructive pulmonary disease, psoriasis, and rheumatoid arthritis, are independently associated with increased lung cancer incidence[12–16]. These comorbidities are clinically identifiable and potentially modifiable through immunomodulatory or anti inflammatory interventions. However, the molecular mechanisms underlying these epidemiological associations remain poorly characterized[17–21]. Specifically, it is not yet established which genetic variants, cell types, or biological pathways mediate the observed risk elevation. Without such resolution, efforts to repurpose existing therapies or design preventive strategies for high risk subpopulations remain empirically driven rather than mechanism informed.

Single cell expression quantitative trait locus data offer the potential to connect disease associated genetic variants to gene regulation at cell type resolution. In practice, however, integration of single cell eQTL with GWAS summary statistics is frequently hindered by incomplete or fragmented linkage disequilibrium information. This fragmentation arises from limited sample sizes in single cell studies, sparse coverage of specific cell populations, and technical heterogeneity across sequencing platforms. Conventional colocalization methods, such as HEIDI and COLOC, require a complete sc-QTL data and invertible LD matrix as input[10,29]. When sc-QTL data are missing, the matrix becomes rank deficient and non invertible, causing these methods to fail. Even when LD estimates are available, regions of strong LD, such as a ±100 kb region, often produce ill conditioned matrices with high condition numbers. This numerical instability renders methods that rely on matrix inversion or posterior computation, including SUSIE, unreliable or inapplicable.

We implemented a conditional Z score based iterative selection procedure, conceptually adapted from the COJO framework[22]. The approach operates within local genomic intervals ±100 kb[30]. The method is designed to identify a minimal set of variants that jointly explain the maximum proportion of shared signal between GWAS and single cell eQTL datasets, under conditions of incomplete sc-eQTL data. Candidate variants were first restricted to those overlapping between GWAS and sc-eQTL summary statistics as a prerequisite for cross-modal comparison. However, to ensure robust inference under fragmented LD and heterogeneous study designs, we required substantial shared genetic signal as quantified by the Signal Overlap Proportion (SOP $\geq$ 0.7).

Our signal-driven integration framework addressing the challenges posed by fragmented variant coverage in single-cell datasets identified 258 genomic loci with strong cross-modal concordance (SOP $\geq$ 0.7) between NSCLC GWAS signals and cell type–specific sc-eQTLs. Within these high-confidence loci, we established unambiguous

directional relationships between NSCLC risk-increasing alleles and target gene expression in 17 distinct cell type–gene pairs. Notably, ACTN1 was inferred as an oncogene in CD4 Memory cell, while *CTSH* was uniformly classified as a tumor suppressor in both monocytes and dendritic cells. Similarly, PSMA4 was robustly annotated as tumor-suppressive in $CD4^+$ T cells, and IREB2 and CHRNA3 were consistently assigned oncogenic roles. A particularly exceptional observation was the context-dependent functional role of *CHRNA5*. In effector-memory $CD4^+$ T cells, risk alleles were associated with reduced expression, supporting a tumor-suppressive interpretation; in contrast, the same alleles correlated with increased expression in naive cycling $CD4^+$ T cells and monocytes, consistent with an oncogenic function, which emphasizes the importance of cellular context in precision oncology.

Beyond gene regulation, locus-specific Mendelian randomization lent causal support to epidemiological links between inflammatory comorbidities and NSCLC. Genetic predisposition to chronic obstructive pulmonary disease, rheumatoid arthritis, and psoriasis consistently increased NSCLC risk, with colocalized signals enriched in immune cell types including monocytes and CD4□ memory T cells. At shared loci such as rs7166764 (monocytes) and rs34138960 (effector CD4□ T cells), the direction of effect on NSCLC risk was concordant across multiple inflammatory diseases, pointing to a shared immunogenetic pathway that may underlie comorbidity-driven cancer susceptibility. In contrast, the rs4902649 locus stood out as a rare protective signal, where genetic liability to both dermatomyositis and rheumatoid arthritis was associated with decreased NSCLC risk, which suggests that certain immune-mediated states may confer resilience rather than vulnerability. The remarkable consistency of effect direction across diverse cell types, diseases, and analytical frameworks strongly supports the biological relevance of these findings and highlights them as promising candidates for functional follow-up and therapeutic repurposing.

Our study has several important limitations. First, the validity of our signal-driven inference relies on the assumption that unobserved common variants are adequately captured through linkage disequilibrium with observed, nominally significant SNPs in dense, high-LD regions. While this may hold when the observed set is sufficiently large and well distributed, it is not guaranteed, particularly because accurate proxy representation requires not only correlation in genotype (LD) but also concordance in effect size direction and magnitude across traits. In sparse single-cell eQTL data, this dual requirement is difficult to verify, and missing variants with distinct regulatory effects could distort signal reconstruction. Thus, our approach provides a conditional approximation under idealized assumptions rather than a robust solution to data fragmentation. Second, the sc-eQTL datasets used in this study were derived from non-cancerous or non-lung contexts, which may not reflect the gene regulatory landscape relevant to NSCLC pathogenesis. Finally, our framework identifies statistical concordance between GWAS and sc-eQTL signals but does not establish causality. Even locus-specific Mendelian randomization serves here only as a consistency check across traits, not as proof of a causal chain. All findings should therefore be interpreted as correlational hypotheses requiring experimental validation[10].

Conclusion

Integrative analysis of GWAS and single-cell eQTL data enables the identification of shared regulatory architectures that link inflammatory comorbidities to lung cancer risk. This framework highlights the potential role of immune cell-type-specific gene regulation in disease etiology and may inform future strategies for risk stratification and mechanistic investigation.

Reference

[1] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer J Clinicians 2021;71:209–49. https://doi.org/10.3322/caac.21660.

[2] Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. CA A Cancer J Clinicians 2023;73:17–48. https://doi.org/10.3322/caac.21763.

[3] Yazar S, Alquicira-Hernandez J, Wing K, et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science 2022;376:eabf3041. https://doi.org/10.1126/science.abf3041.

[4] Natri HM, Del Azodi CB, Peter L, et al. Cell type-specific and disease-associated eQTL in the human lung 2023. https://doi.org/10.1101/2023.03.17.533161.

[5] Schmiedel BJ, Gonzalez-Colin C, Fajardo V, et al. Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type–dependent effects of disease-risk variants. Sci Immunol 2022;7:eabm2508. https://doi.org/10.1126/sciimmunol.abm2508.

[6] LifeLines Cohort Study, BIOS Consortium, Van Der Wijst MGP, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet 2018;50:493–7. https://doi.org/10.1038/s41588-018-0089-9.

[7] Soskic B, Cano-Gamez E, Smyth DJ, et al. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. Nat Genet 2022;54:817–26. https://doi.org/10.1038/s41588-022-01066-3.

[8] Resztak JA, Wei J, Zilioli S, et al. Genetic control of the dynamic transcriptional response to immune stimuli and glucocorticoids at single cell resolution 2021. https://doi.org/10.1101/2021.09.30.462672.

[9] Randolph HE, Fiege JK, Thielen BK, et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science 2021;374:1127–33. https://doi.org/10.1126/science.abg0928.

[10] 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. https://doi.org/10.1371/journal.pgen.1004383.

[11] Lu Z, Wang X, Carr M, et al. Improved multiancestry fine-mapping identifies cis-regulatory variants underlying molecular traits and disease risk. Nat Genet 2025;57:1881–9. https://doi.org/10.1038/s41588-025-02262-7.

[12] Carlisle JW, Steuer CE, Owonikoko TK, et al. An update on the immune landscape in lung and head and neck cancers. CA A Cancer J Clinicians 2020;70:505–17. https://doi.org/10.3322/caac.21630.

[13] Khanna U, Galimberti F, Li Y, et al. Dermatomyositis and malignancy: should all patients with dermatomyositis undergo malignancy screening? Ann Transl Med 2021;9:432–432. https://doi.org/10.21037/atm-20-5215.

[14] Uhlenbruch M, Polke M, Von Eiff D, et al. The Comorbidity of Lung Cancer and ILD: A Review. Pneumologie 2025:a-2512-8349. https://doi.org/10.1055/a-2512-8349.

[15] Park HY, Kang D, Shin SH, et al. Chronic obstructive pulmonary disease and lung cancer incidence in never smokers: a cohort study. Thorax 2020;75:506–9. https://doi.org/10.1136/thoraxjnl-2019-213732.

[16] Constantin A-A, Arghir MA, Avasilcăi D, et al. Beyond Comorbidity: Pulmonary Adenocarcinoma in a Patient with Rheumatoid Arthritis—A Case Report and Literature Review. Life 2025;15:1118. https://doi.org/10.3390/life15071118.

[17] Wang X, Wang X, Wang H, et al. Association between psoriasis and lung cancer: two-sample Mendelian randomization analyses. BMC Pulm Med 2023;23:4. https://doi.org/10.1186/s12890-022-02297-0.

[18] Zheng H, Nie D, Huang X. Association between idiopathic pulmonary fibrosis and risk of different pathological types of lung cancer: a Mendelian randomization study. J Cancer Res Clin Oncol 2023;149:7751–7. https://doi.org/10.1007/s00432-023-04727-w.

[19] Wang L, Dong Y, Yang Q, et al. Exploring the association between rheumatoid arthritis and non-small cell lung cancer risk: a transcriptomic and drug target-based analysis. Hereditas 2025;162:28. https://doi.org/10.1186/s41065-025-00396-6.

[20] Liao K, Yang D, Jin L, et al. Bidirectional mendelian randomization and single-cell sequencing reveal T cell-mediated causal links between COPD and lung adenocarcinoma. International Journal of Surgery 2025;111:4528–38. https://doi.org/10.1097/JS9.0000000000002448.

[21] 吴学婷. The Causal Relationship between Dermatomyositis and Lung Cancer Mediated by Immune Cells: A Mendelian Randomization Study with

Mediation Analysis. JCPM 2025;04:741–56. https://doi.org/10.12677/jcpm.2025.42239.

[22] Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Yang J, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369–75. https://doi.org/10.1038/ng.2213.

[23] Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet 2019;51:592–9. https://doi.org/10.1038/s41588-019-0385-z.

[24] Nica AC, Montgomery SB, Dimas AS, et al. Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet 2010;6:e1000895. https://doi.org/10.1371/journal.pgen.1000895.

[25] Yeung SLA, Luo S, Iwagami M, et al. Introduction to Mendelian randomization. ACE 2025;7:27–37. https://doi.org/10.37737/ace.25004.

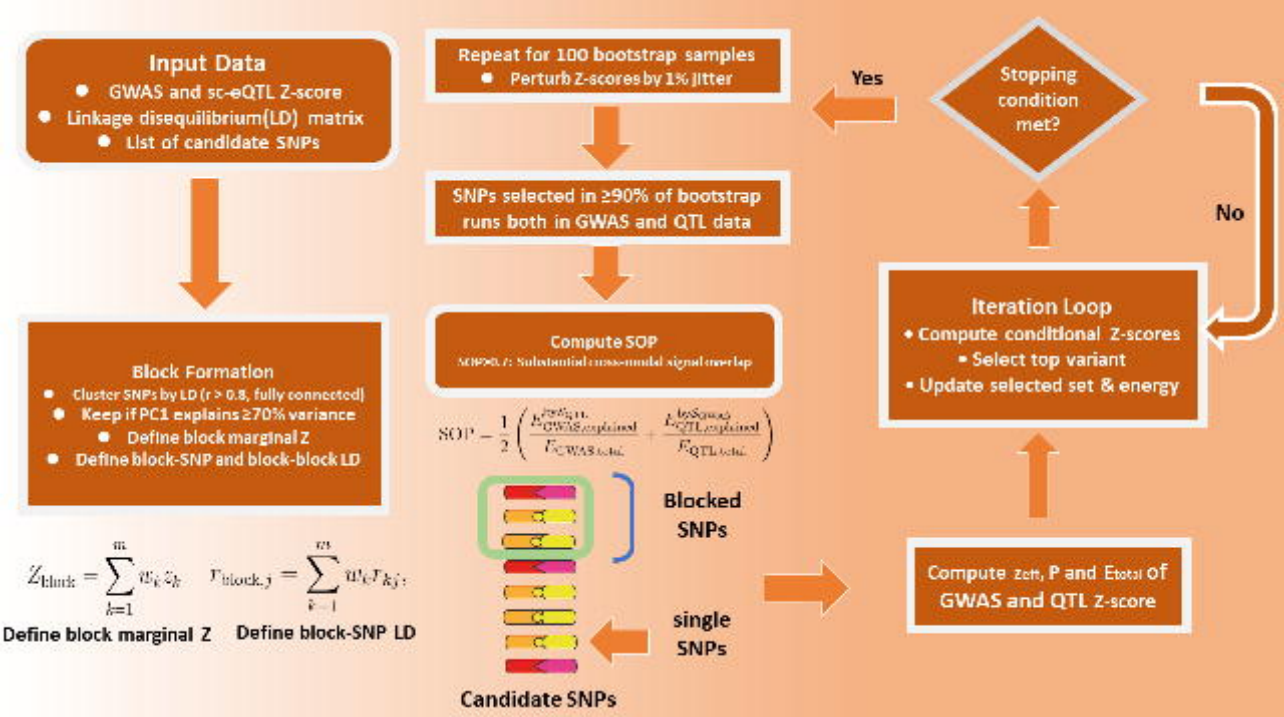
[26] Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun 2018;9:224. https://doi.org/10.1038/s41467-017-02317-2.

[27] Ray A, Alabarse P, Malik R, et al. Single-cell transcriptome-wide Mendelian randomization and colocalization analyses uncover cell-specific mechanisms in atherosclerotic cardiovascular disease 单细胞转录组范围内的孟德尔随机化与共定位分析揭示动脉粥样硬化性心血管疾病中的细胞特异性机制. The American Journal of Human Genetics 2025;112:1597–609. https://doi.org/10.1016/j.ajhg.2025.06.001.

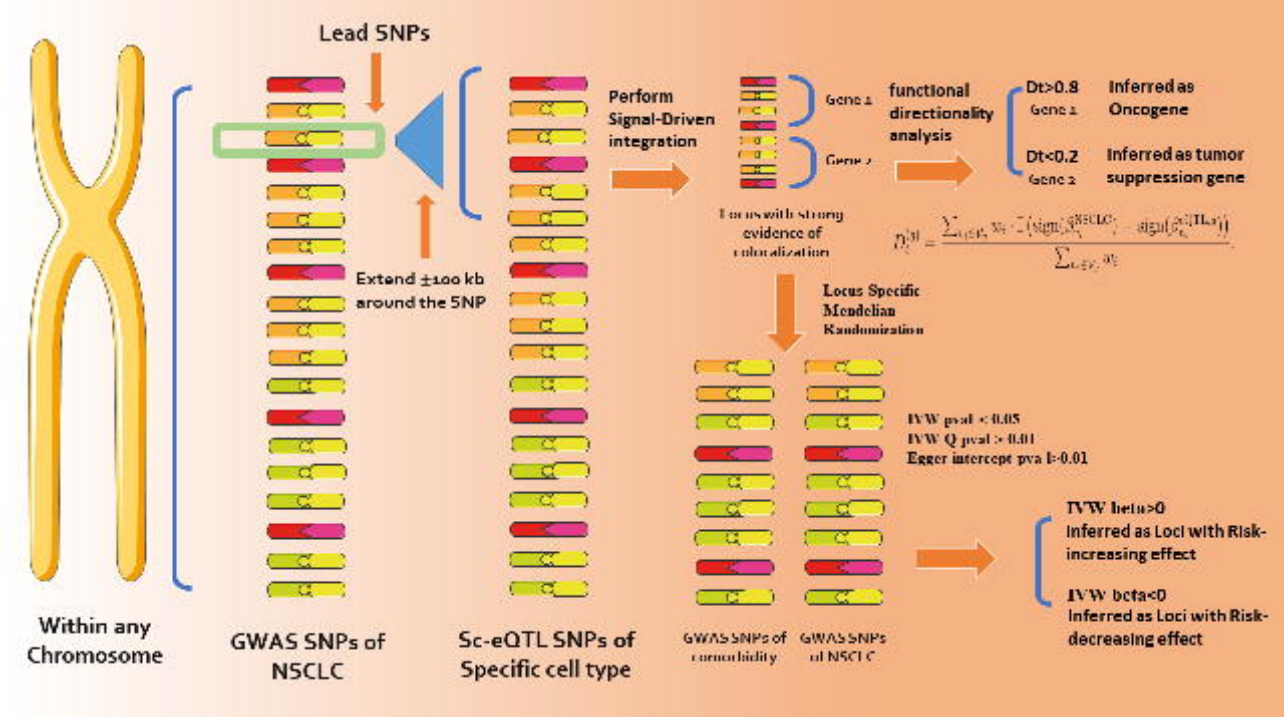
[28] 1000 Genomes Project Consortium, Auton A, Abecasis GR, et al. A global reference for human genetic variation. Nature 2015;526:68–74. https://doi.org/10.1038/nature15393.

[29] Hormozdiari F, van de Bunt M, Segrè AV, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. The American Journal of Human Genetics 2016;99:1245–60. https://doi.org/10.1016/j.ajhg.2016.10.003.

[30] Pickrell JK, Berisa T, Liu JZ, et al. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet 2016;48:709–17. https://doi.org/10.1038/ng.3570.



Input Data
GWAS and sc-eQTL Z-score
Linkage disequilibrium(LD) matrix
List of candidate SNPs
Block Formation
Cluster SNPs by LD (r > 0.8, fully connected)
Keep if PC1 explains ≥70% variance
Define block marginal Z
Define block-SNP and block-block LD
$Z_{block} = \sum_{k=1}^{m} w_k z_k$
$r_{block,j} = \sum_{k=1}^{m} w_k r_{kj}$
Define block marginal Z
Define block-SNP LD
Repeat for 100 bootstrap samples
Perturb Z-scores by 1% jitter
SNPs selected in ≥90% of bootstrap runs both in GWAS and QTL data
Compute SOP
SOP≥0.7: Substantial cross-modal signal overlap
$SOP = \frac{1}{2}\left(\frac{E^{logE_{QTL}}_{GWAS,explained}}{E_{GWAS,total}} + \frac{E^{logS_{GWAS}}_{QTL,explained}}{E_{QTL,total}}\right)$
Blocked SNPs
single SNPs
Candidate SNPs
Compute z_{eff}, P and E_{total} of GWAS and QTL Z-score
Iteration Loop
Compute conditional Z-scores
Select top variant
Update selected set & energy
Stopping condition met?
Yes
No


Lead SNPs
Extend ±100 kb around the SNP
Within any Chromosome
GWAS SNPs of NSCLC
Sc-eQTL SNPs of Specific cell type
Perform Signal-Driven integration
Gene 1
Gene 2
Loci with strong evidence of colocalization
functional directionality analysis
Dt>0.8
Gene 1
Inferred as Oncogene
Dt<0.2
Gene 2
Inferred as tumor suppression gene
Locus Specific Mendelian Randomization
IVW pval < 0.05
IVW Q pval > 0.01
Egger intercept pval > 0.01
GWAS SNPs of comorbidity
GWAS SNPs of NSCLC
IVW beta>0
Inferred as Loci with Risk-increasing effect
IVW beta<0
Inferred as Loci with Risk-decreasing effect