nature communications

Article

https://doi.org/10.1038/s41467-025-64571-z

Common genetic variation influencing the human lung imaging phenotypes

Received: 7 February 2025

Accepted: 22 September 2025

Published online: 29 October 2025

Check for updates

Meng Zhu[1,2,3,10], Lingbin Du[4,10], Lei Shi[5,10], Chen Ji[1,10], Chen Zhu[1,4,10], Ci Song[1,2,3,10], Lili Wu[6], Lingying Zhu[7], Jing Lu[8], Qun Zhang[8], Feiyun Wu[8], Chen Jin[1], Yuanlin Mou[1], Qiao Li[1], Jiahao Zhang[1], Mingxuan Zhu[1], Jiaying Cai[1], Caochen Zhang[1], Yating Fu[1], Linnan Gong[1], Dong Hang[1,2], Juncheng Dai[1,2], Yue Jiang[1,2], Guangfu Jin[1,2,3], Zhibin Hu[1,2], Hongxia Ma[1,2,9,11] ✉, Xiangdong Cheng[4,11] ✉ & Hongbing Shen[1,2,11] ✉

Lung structures are critical for gas exchange and contribute to the pathogenesis of respiratory diseases, exhibiting notable lobe-specific heterogeneity. To investigate their genetic basis, we apply a deep-learning AI system and *Pyradiomics* to define lobe-specific lung CT imaging phenotypes, conducting genome-wide analyses in 35,469 participants from the Lung Imaging Genomics Initiative in China. We identify 36 loci associated with voxel intensities and 138 loci linked to three-dimensional shape. Functional annotation reveals significant enrichment of relevant genes in pathways regulating early fetal lung development and loci enriched in fetal lung regulatory elements. Genetic correlations are identified between lung structures and chronic respiratory diseases as well as lung function, with a number of loci showing colocalization. Mendelian randomization analyses suggest a causal role of lung structures in chronic lung diseases and extrapulmonary traits. This study provides new insights into the genetic architecture of lung structures and their links to diverse clinical outcomes.

The lungs are the foundational organs of the respiratory system, primarily responsible for facilitating gas exchange between the environment and the bloodstream. Lung development begins with the formation of the conducting airways, followed by the expansion of the gas exchange area through alveolarization, continuing into young adulthood[1]. Lung diseases impact both zones and exhibit distinct characteristics on computed tomography (CT) images. Lung image-derived phenotypes (LIDPs) from CT reflect various aspects of lung structure, intensity, and texture, serving as indicators of lung condition and correlating closely with chronic respiratory diseases[1–4].

The right lung is divided into upper, central, and lower lobes by major and minor fissures, while the left lung is divided into upper and lower lobes by the major fissure. Each lobe is served by a second-

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China. [2]State Key Laboratory Cultivation Base of Biomarkers for Cancer Precision Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China. [3]Wuxi Medical Center, Nanjing Medical University, Nanjing, China. [4]Department of Cancer Prevention, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang, China. [5]Department of Radiology, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang, China. [6]Department of Cancer Prevention, Taizhou Cancer Hospital, Taizhou, China. [7]Department of Radiology, Taizhou Cancer Hospital, Taizhou, China. [8]Department of Health Promotion Center, the First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China. [9]The Second People's Hospital of Changzhou, the Third Affiliated Hospital of Nanjing Medical University, Changzhou Medical Center, Nanjing Medical University, Changzhou, China. [10]These authors contributed equally: Meng Zhu, Lingbin Du, Lei Shi, Chen Ji, Chen Zhu, Ci Song. [13]These authors jointly supervised this work: Hongxia Ma, Xiangdong Cheng, Hongbing Shen. ✉e-mail: hongxiama@njmu.edu.cn; chengxd@zjcc.org.cn; hbshen@njmu.edu.cn

order bronchus, which further divides into bronchopulmonary segments and eventually into respiratory bronchioles where alveoli appear. The anatomical and physiological heterogeneity of different lung lobes is associated with the selective distribution of lung diseases[5,6]. Therefore, a deep understanding the development and structure of each lobe is critical for addressing lung health and disease.

Genetic variations at the individual level have a lasting impact on lung structures and disease susceptibility. Although genes that disrupt lung development and structure have been extensively studied in model organisms, our current understanding of the genetic variations influencing human lung size and structure is largely limited to rare, highly penetrant variants[7–10]. Genome-wide association studies (GWAS) have identified tens to hundreds of genetic associations with common variants for chronic respiratory diseases, including chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), asthma, and lung cancer[11–15]. Additionally, over 1000 genetic variants have been linked to lung function, a comprehensive index of lung health[16]. Further investigation into genetic determinants on LIDPs will shed light on the causes for individual differences in lung structure and the mechanisms of chronic respiratory diseases.

Within the Lung Imaging Genomics Initiative (LIGI) in China, we developed a scheme for defining LIDPs of each lung lobe using a deep-learning AI system for segmentation[17] and *Pyradiomics* (v3.0) for radiomic features extraction[18]. To identify genetic loci associated with variations in human lung structure, we conducted the first GWAS of 160 LIDPs (32 × 5) among 35,469 Chinese Han participants. Additionally, we analyzed the distribution and function of loci associated with these LIDPs, explored genetic colocalizations between LIDPs and lung-related disorders and lung function, and performed a Mendelian randomization phenome-wide association study (MR-PheWAS) to clarify the associations between LIDPs and other diseases. An overview of the study design is shown in Fig. 1.

Results

Lung imaging acquisition and radiomic features extraction

In the LIGI, we recruited 36,551 participants aged 18–75 years from Zhejiang (N = 25,821) and Jiangsu (N = 10,730) and performed low-dose CT scans (120 kV and 40–60 mA). These CT images were first segmented into five lung lobes using a deep-learning AI system, followed by extraction of radiomic features using *Pyradiomics* (v3.0). Details regarding image processing and feature extraction can be found in the **Methods** section. Besides, we evaluated potential batch effects caused by differences in scanning protocols between centers using principal component analysis (PCA), and the results indicated minimal batch effects (Supplementary Fig. 1). Of the 36,551 CT images, we excluded 916 that failed manual checks by radiologists, did not pass automatic segmentation, or were identified as outliers through principal component analysis. We further excluded 166 samples that did not meet genomic data quality control standards. Consequently, 35,469 participants with qualified CT images and genotype data were included in the study (Supplementary Fig. 2 and Supplementary Data 1). This study primarily focused on 32 LIDPs of each lobe, including 18 first-order features describing the distribution of voxel intensities (density) and 14 shape features detailing the three-dimensional size and shape (**Methods**). Most of the LIDPs showed significant associations with sociodemographic factors (age, sex, smoking status, and BMI) suggesting their ability to capture the effects of demographic and environmental factors on lung structure (Supplementary Data 2). Considering the distinct clinical implications of LIDPs between first-order and shape features, further analysis was performed for these two dimensions separately.

Heritability and genetic correlation of lung imaging phenotypes

The LIDPs of first-order features showed an average heritability of 15.20% using BOLT-REML, ranging from 2.94% for minimum voxel intensities at the right upper lobe to 50% for TotalEnergy (sum of voxel values) at the right central lobe (Fig. 2a). Heritability estimates were

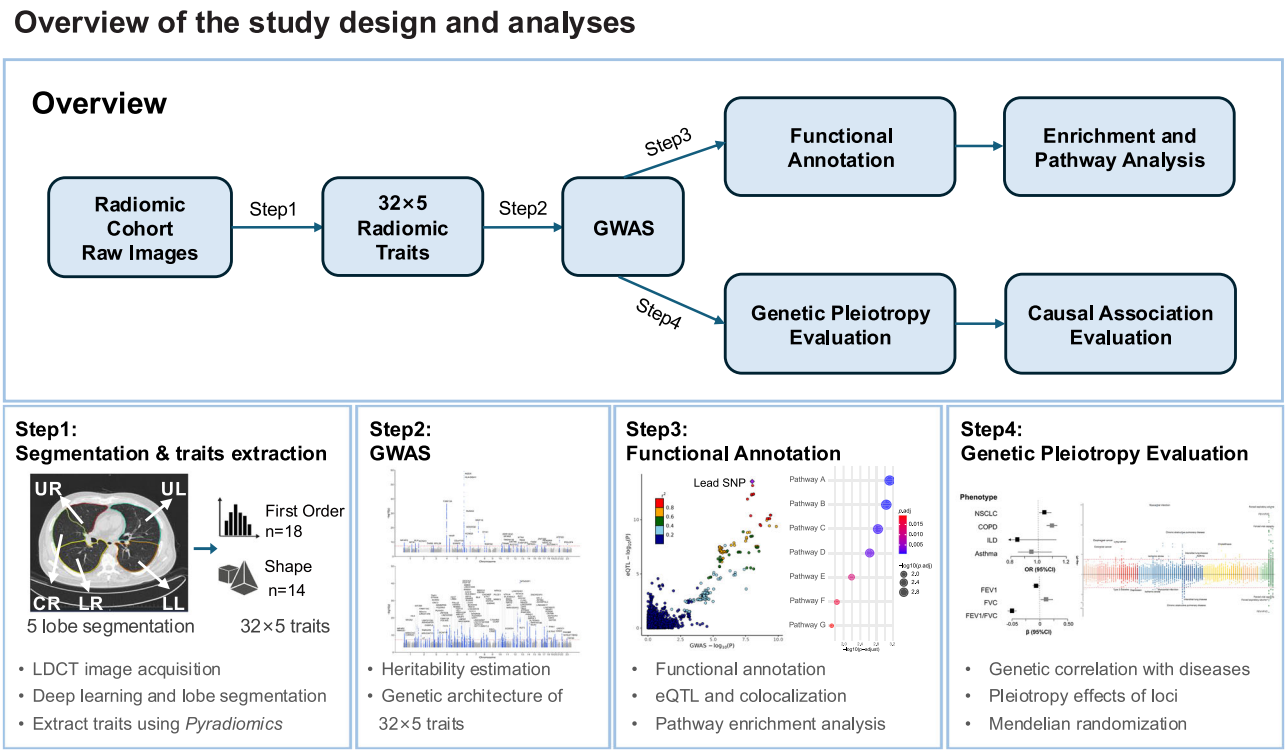


Fig. 1 | Overview of study design and analyses. We used lung radiomic traits as endophenotypes to explore the genetic architecture of lung structure. The upper workflow described the overall workflow, and the lower part described key analyses involved in each step. GWAS genome-wide association study, LDCT low-dose computed tomography, eQTL expression quantitative trait loci, UR right upper lobe, UL left upper lobe, CR right centre lobe, LR right lower lobe, LL left lower lobe.

https://doi.org/10.1038/s41467-025-64571-z

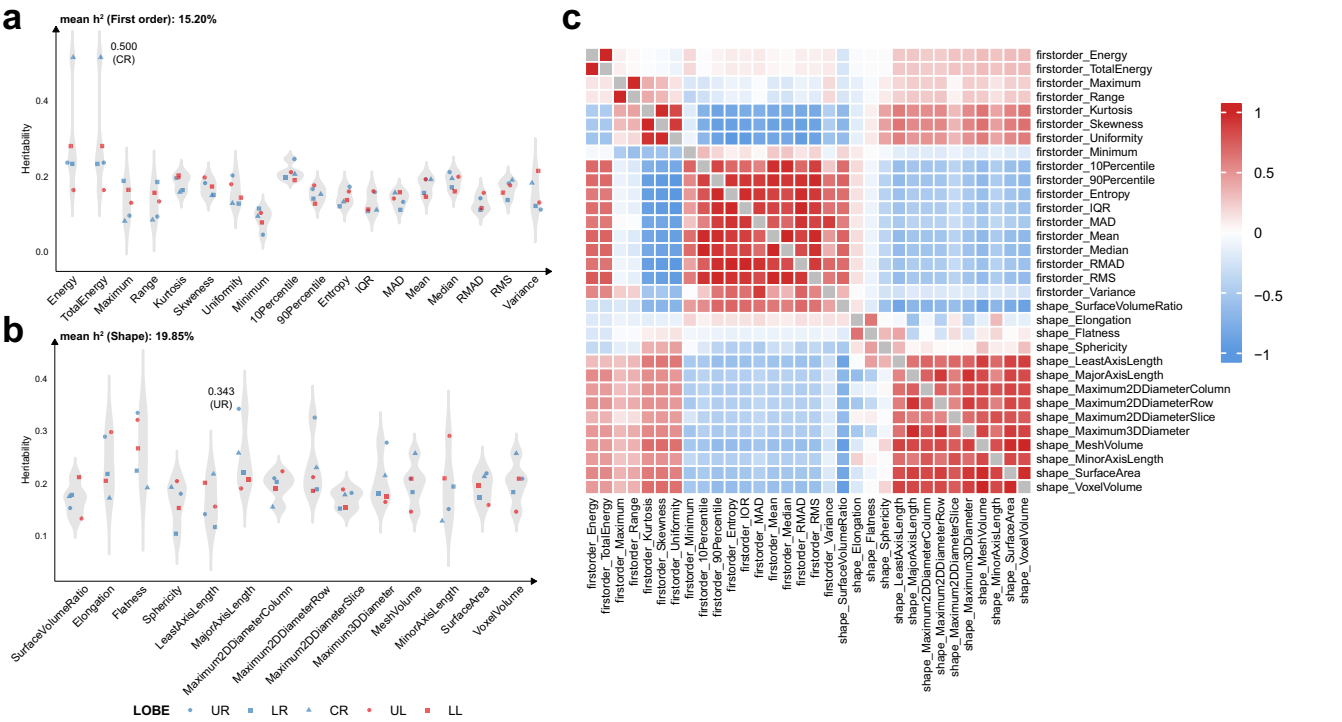


Fig. 2 | Overall phenotypic correlation and genetic correlation and results of heritability estimation. Plots show the heritability of all lung radiomic features (**a**, first-order; **b**, shape) using BOLT-REML. Different colors correspond to different lung lobes. Blue dots indicate the UR lobe, blue squares the LR lobe, blue triangles the CR lobe, red dots the UL lobe, and red squares the LL lobe. **c** Correlation heatmap shows the results of phenotypic correlations (upper right triangle) and genetic correlations (lower left triangle). Due to the low heritability of firstorder_Minimum in the right upper lung lobe estimated using LDSC, the mean heritability for firstorder_Minimum was calculated across the remaining four lung lobes only. The color gradient ranges from red, indicating correlations close to 1, to blue, indicating correlations close to -1. UR right upper lobe, UL left upper lobe, CR right centre lobe, LR right lower lobe, LL left lower lobe, IQR InterquartileRange, MAD MeanAbsoluteDeviation, RMAD RobustMeanAbsoluteDeviation, RMS RootMeanSquared.

significantly higher for LIDPs of shape features ($P = 6.42 \times 10^{-11}$), with an average of 19.85%, peaking at 34.31% for the major axis length of the right upper lobe (Fig. 2b). We further examined lobe-specific heritability and found no significant differences across lobes ($P > 0.05$ for first-order and shape features).

We found strong genetic correlations for the 32 LIDPs within each pair of the five lung lobes (average r_g: 0.549-0.864), except for the flatness (higher values imply sphere-like shape) (Supplementary Fig. 3). Within each lobe, generally high positive genetic correlations were observed for LIDPs within first-order and shape features, while negative correlations were observed between the two dimensions (Fig. 2c and Supplementary Fig. 4). However, first-order features reflecting the distribution of voxel intensities (kurtosis, skewness, and uniformity), extremum (maximum and range), and energy were positively associated with lung size. The surface volume ratio of shape features also demonstrated a positive correlation with lung voxel intensities. These findings suggest a mutual influence between first-order and shape LIDPs.

Furthermore, we observed high genetic correlations of LIDPs between the two recruitment centers ($r_g = 0.807$) and across different smoking statuses ($r_g = 0.764$), underscoring the robustness of our findings (Supplementary Data 3).

Common genetic variants associated with lung imaging phenotypes

In 35,469 LIGI participants, we conducted GWAS for the 160 LIDPs across 844,7934 variants with a minor allele frequency (MAF) ≥ 1% and an imputation information score (INFO) ≥ 0.3. The observed genomic inflation factors (λ_{gc}) in QQ plots (Supplementary Fig. 5) and the low linkage disequilibrium (LD) score regression (LDSC) intercepts, ranging from 0.99 to 1.04, indicated that this inflation was consistent with polygenicity rather than confounding (Supplementary Data 4).

Across the 90 LIDPs of first-order, we identified 346 independent variant-LIDP associations ($P < 5 \times 10^{-8}$), using PLINK clumping with a distance of 500 kb and LD r^2 of 0.1 (Supplementary Data 5). These associations were further grouped into 36 clusters by merging unique variants within 500 kb at LD r^2 of 0.1 across LIDPs (Fig. 3a), of which 27 clusters were associated with at least two LIDPs (Supplementary Data 6). Among these associations, we discovered 12 unique loci that satisfied a multiple testing significance threshold of $P < 5.56 \times 10^{-10}$ after Bonferroni correction, which consistently associated with multiple LIDPs. Notably, two single nucleotide polymorphisms (SNPs), rs4505789 (4q22.1, near *FAM13A*) and rs41268920 (6p21.32, near *AGER*), were associated with 45 and 50 first-order LIDPs, respectively (Supplementary Fig. 6a).

For the 70 shape LIDPs, we discovered 562 independent variant-LIDP associations at $P < 5 \times 10^{-8}$ (Supplementary Data 7). After cross-traits clumping, we observed 138 independent clusters (within 500 kb at $r^2 \leq 0.1$) (Fig. 3b and Supplementary Data 8), of which 60 remained significant when additionally corrected for multiple testing using the threshold of $P < 7.14 \times 10^{-10}$ after Bonferroni correction. More than half (57.2%, 79/138) of the identified variants were significant in at least two LIDPs, and the proportion was as high as 88.3% (53/60) in the identified loci after correction. Notably, we observed two independent SNPs near *HHIP* (rs72731582 and rs11100862, $r^2 = 0.03$) that were significantly associated with 23 and 31 shape LIDPs, respectively; the locus of *FAM13A* was also significantly associated with 15 shape LIDPs (Supplementary Fig. 6b).

To validate the robustness of the identified loci, we performed a subgroup analysis according to the recruitment center. Around 77.8%

https://doi.org/10.1038/s41467-025-64571-z

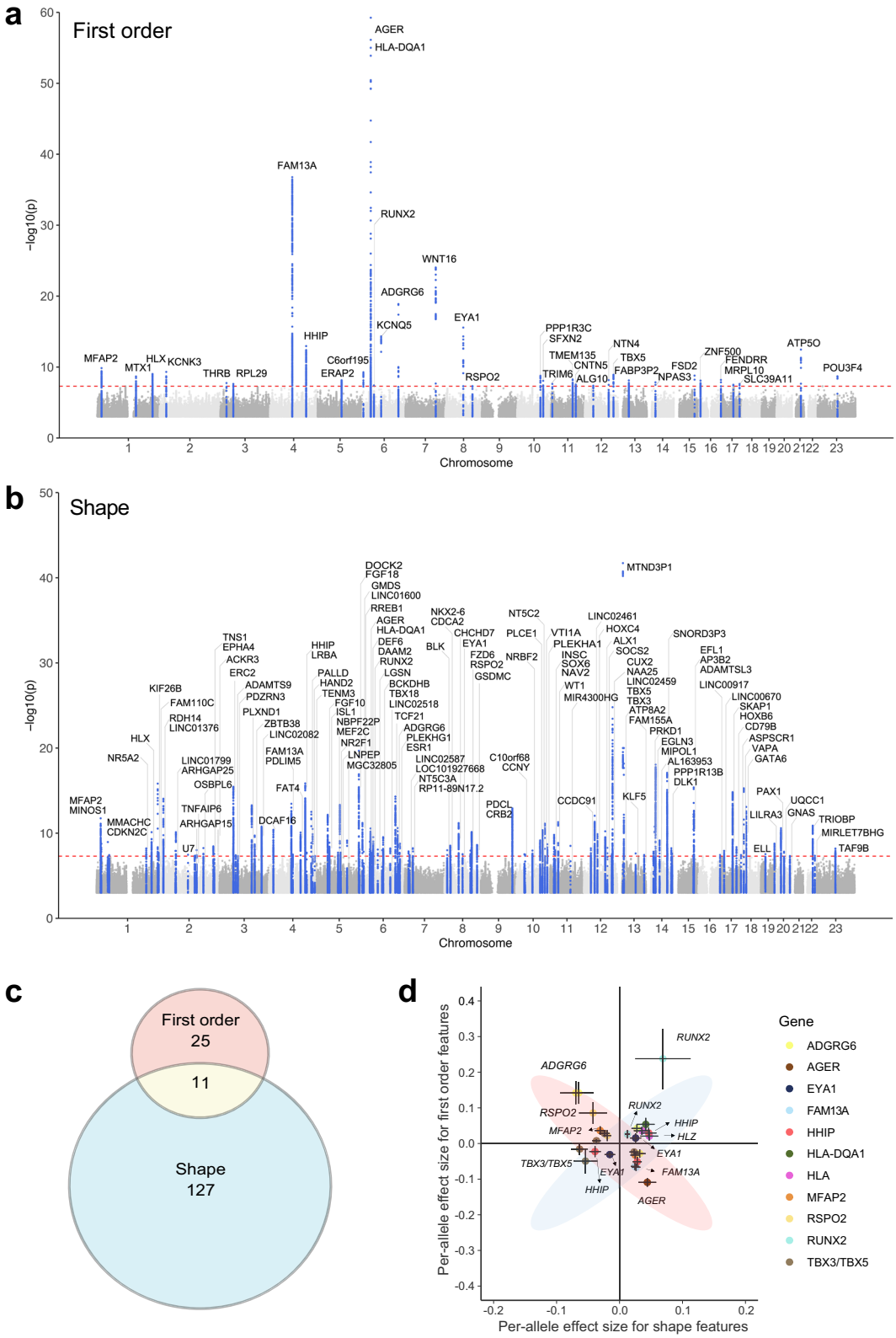


Fig. 3 | Genome-wide association study results for first-order and shape traits of lung imaging. Manhattan plots show the chromosomal position (x axis) and the strength of association (−log10 of the two-sided BOLT-LMM chi-squared test *P* value, y axis) for all first-order features (**a**) and shape features (**b**). The X chromosome is represented as '23'. Loci with $P < 5 \times 10^{-8}$ are shown in blue. For each locus, its *p*-value represents the smallest *P*-value of that locus across all first-order or shape features. **c** Venn diagram showing the number of loci uniquely associated at $P < 5 \times 10^{-8}$ with either the first-order related radiomic features or shape-related features, those in yellow are associated with both. **d** This figure illustrates the per-allele effect sizes (beta) and their 95% confidence intervals (CIs) of lead variants associated with traits of first-order and shape features. Error bars represent 95% confidence intervals (CIs) around the point estimates (beta). The lead SNPs were identified through GWAS analysis, where each SNP's effect size and standard error correspond to the most significant phenotype (smallest *P* value) within each feature category, based on GWAS results from 35,469 individuals using the BOLT-LMM Wald test (two-sided *P* values). No additional corrections beyond the GWAS framework were applied.

(28/36) of associations from first-order features and 89.1% (123/138) of associations from shape features were consistently significant in the two centers (Supplementary Data 9). To exclude the impact of smoking, we further tested these associations among 20,148 nonsmokers in our study and observed that all of the associations were consistently significant (Supplementary Data 9). Regional association plots of the 174 lead variants were verified by manual visual inspection.

Although first-order features and shape features reflect different aspects of lung structure, we identified 11 shared regions associated with both dimensions (Fig. 3c). We compared the per-allele effect sizes of lead variants associated with any LIDPs of first-order and shape features and observed two distinct distribution patterns (Fig. 3d). Genetic variants near transcription factors (TFs) involved in lung development, such as *HHIP*, *HLX*, *RUNX2*, *EYA1*, and *TBX3*/*TBX5*[19], showed relatively positive correlations between first-order and shape features. In contract, the effect sizes of SNPs near *AGER*, *FAM13A*, *MFAP2*, and *ADGRG6* showed negative correlations between first-order and shape features, and these SNPs were consistently associated with corresponding gene expression in our 338 Chinese lung samples and in 515 lung tissue samples from GTEx (Supplementary Fig. 7).

More than one-fifth (39 out of 174) of the identified SNPs were located near TFs, many of which play key roles in lung development[19]. Among the six TFs related to first-order features, rs12581666 near *TBX3*/*TXB5* was significantly associated with over half (60.0%, 54/90) of the LIDPs across the five lobes ($P < 0.05$); whereas rs4342857 near *HLX* was only associated with 10 LIDPs in some lobes (Supplementary Fig. 8 and Supplementary Data 10). We also identified 33 unique variants near 28 TFs were significantly associated with shape features (Supplementary Data 11 and Supplementary Fig. 8). Notably, independent SNPs ($r^2 < 0.1$) near *FGF10*, *TXB3*/*TXB5*, and *GATA6* exhibited highly heterogeneous associations with different lung lobes. These results suggest that TF regulation may vary selectively across different lung lobes, with surrounding genetic variants potentially playing distinct regulatory roles in lung development.

To assess the robustness of our GWAS findings, we re-evaluated 174 independent genome-wide significant loci (36 for first-order features, 138 for shape features) by adjusting for height (Supplementary Data 12-13). As expected, 96.55% (168/174) of the loci remained significant at $P < 1 \times 10^{-5}$, and 75.29% (131/174) surpassing the genome-wide significance threshold ($P < 5 \times 10^{-8}$). To further validate our GWAS results, we conducted a replication analysis using an independent dataset of ~7000 participants recruited in 2022 in Zhejiang province as part of the LIGI (Supplementary Data 14). Of the 174 lead SNPs identified in the discovery cohort, 145 (83.3%) remained nominally significant ($P < 0.05$) in this replication dataset. These findings provide additional evidence for the robustness of the identified associations.

Fine-mapping and functional characterization for candidate variants

To explore the secondary signals at the identified loci, we conducted a conditional and joint analysis (COJO) (**Methods**). The additional associations identified from these conditional analyses for each LIDP are reported in Supplementary Data 15. We then conducted a fine-mapping analysis using SuSiE to identify causal variants within 95% credible sets for each LIDP. As a result, we identified 5995 variants falling into 223 independent loci across first-order features (median 16, ranging from 1 to 201 per locus, Supplementary Data 16) and 19,924 variants falling into 493 independent loci across shape features (median 26, ranging from 1 to 577 per locus, Supplementary Data 17).

To functionally characterize the prioritized variants, we performed an integrated variant function annotation using SNPnexus (**Methods**). A total of eight unique nonsynonymous variants were identified across first-order features (Supplementary Data 18). Of those, rs17280293 (*ADGRG6*; p.S123G; combined annotation-dependent depletion (CADD) score = 24.9) was also associated with lung function[20] and diffusing capacity of the lung for carbon monoxide traits[21]. Similarly, we discovered 22 unique nonsynonymous variants among shape features (Supplementary Data 19). Notably, rs9379084 (*RREB1*; p.D1171N; CADD score = 29.7) had a posterior inclusion probability (PIP) of 1.00 for flatness of left low lobe. RREB1 is a zinc-finger TF that functions downstream of RAS signaling and is activated in lung lipofibroblasts, which play roles in lung development and regulation of epithelial cell migration[22].

In addition to the identified genes with nonsynonymous SNPs, we further mapped susceptibility genes underlying the GWAS associations based on our expression quantitative trait locus (eQTL) database, which was derived from 338 Chinese lung tissue samples. We examined the overlap between the GWAS-identified variants ($r^2 > 0.6$ with the identified lead SNPs) and significant eQTL variants (FDR < 0.05), identifying 48 unique eQTL genes for first-order features and 53 unique eQTL genes for shape features (Supplementary Data 20). To prioritize candidate susceptibility genes, Bayesian colocalization analyses were conducted using the eQTL database and the GWAS summary statistics of each LIDP, employing the coloc method[23]. A total of 207 gene-LIDP pairs were colocalized for first-order features (PPH4 > 0.7), including 17 unique eQTL genes (Fig. 4a and Supplementary Data 21). For shape features, we identified 244 gene-LIDP colocalizations, including 28 unique eQTL genes (Fig. 4b and Supplementary Data 22). Interestingly, we found that 12 genes (36.36% of the colocalization genes) were shared by both first-order and shape LIDPs, indicating that these genes (*ACY1*, *AGER*, *AP3B2*, *ERAP2*, *FAM13A*, *GBAP1*, *HHIP*, *HLA-DRB6*, *MFAP2*, *NBPF1*, *PLXND1*, and *UQCC1*) probably influence the lung's shape and density characteristics simultaneously. Additionally, we identified 767 gene-LIDP colocalizations (332 for first-order and 435 for shape), pointing to 269 unique eQTL genes outside of the genome-wide significant loci (Supplementary Data 23).

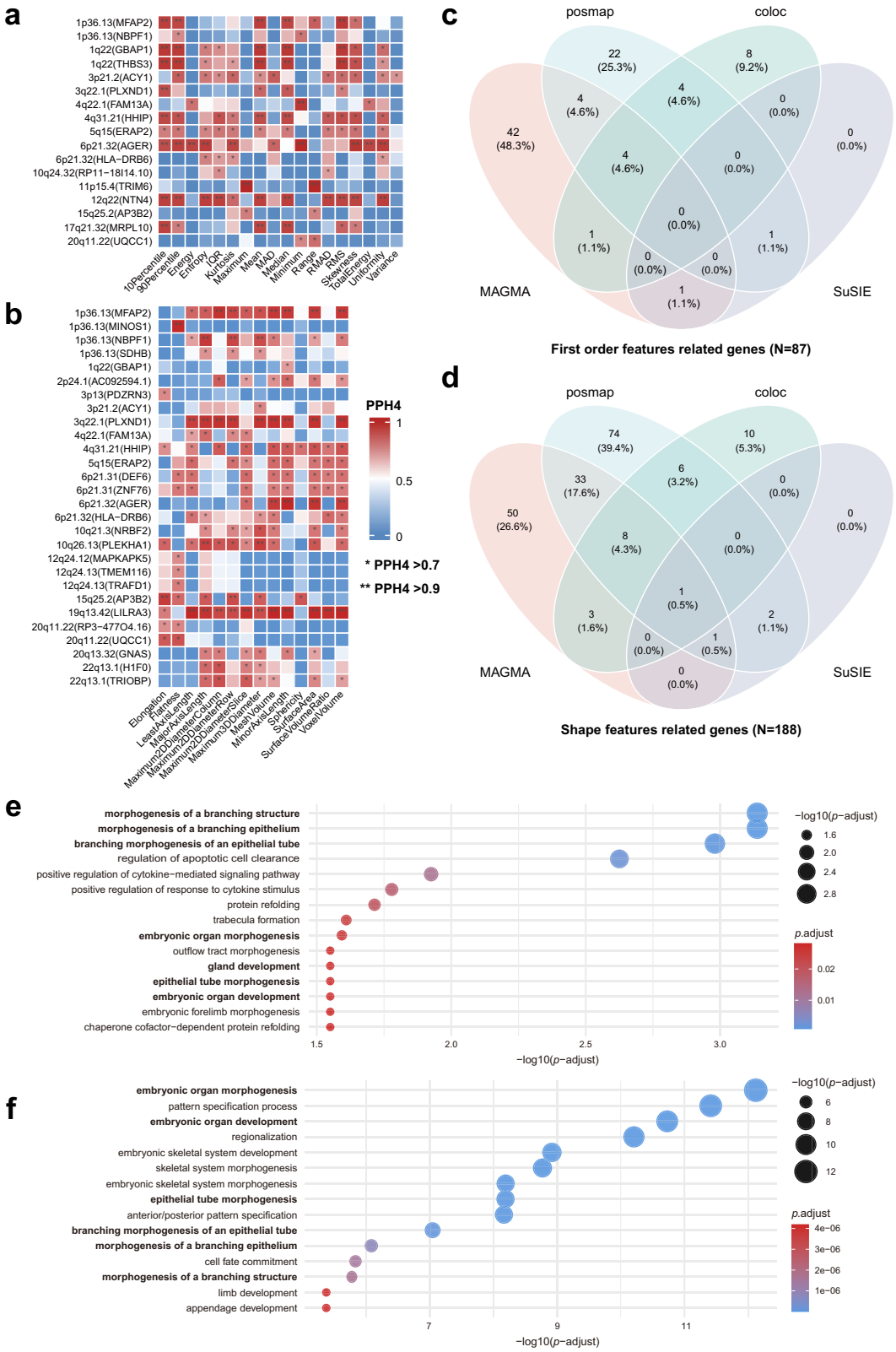
Using the Functional Mapping and Annotation (FUMA) platform, we identified 35 genes for first-order LIDPs and 125 genes for shape LIDPs according to positional mapping within 10 kb of lead SNPs ($P < 5 \times 10^{-8}$) (Supplementary Data 24). Gene-level analysis using MAGMA (**Methods**) detected 142 and 315 unique genes for first-order and shape LIDPs, respectively, satisfying $P < 2.7 \times 10^{-6}$ (0.05/18,517, Bonferroni's correction) (Supplementary Data 25).

Based on the prioritization of candidate genes from SuSIE, colocalizations analysis, positional mapping, and MAGMA (Fig. 4c, d), we additionally performed gene ontology and pathway analysis (**Methods**). A total of 40 and 199 pathways were significantly enriched for first-order and shape LIDPs, respectively, after Benjamini-Hochberg correction (Supplementary Data 26, 27). Notably, the candidate genes showed the most significant enrichment in pathways related to branching morphogenesis, epithelial tube morphogenesis, and embryonic organ morphogenesis and development (Fig. 4e, f). These pathways play a crucial role in the early development of fetal lung[24].

To characterize the functional features of non-coding SNPs identified in our study (~99%), we assessed their heritability enrichment within functional elements across 14 tissues (including fetal lung) by employing stratified LDSC[25] and GARFIELD[26] with derived genome-wide summary statistics for first-order and shape features (**Methods**). Significant genetic enrichments ($P < 0.05$ after Bonferroni correction) were observed in DNase, H3K4me1, and H3K9ac marks in the fetal lung, as well as H3K27ac marks in adult lung for first-order features (Fig. 5a). Similarly, significant enrichments were observed in the functional elements for shape features in both fetal and adult lung (Fig. 5b). Additionally, GARFIELD analysis confirmed significant enrichment of genetic associations in the DNase hypersensitivity regions of the fetal lung (Supplementary Fig. 9).

To further characterize specific cell types associated with first-order and shape features in the fetal lung, processed single-cell RNA (scRNA-seq) sequencing data of human fetal lungs (15–22 post-conception weeks) were obtained from the Lung Cell Atlas

https://doi.org/10.1038/s41467-025-64571-z



(https://lungcellatlas.org)[27]. A total of 26,568 genes for 14 cell types were included for cell-specificity analysis (Supplementary Fig. 10). We employed the single-cell disease relevance score (scDRS) to identify cell types with elevated expression of genes associated with first-order and shape features (**Methods**). Lung mesothelial cells exhibited strong associations with both first-order and shape features, aligning with the notion that the mesothelium serves as a source of progenitor cells for mesenchymal lineages during organogenesis (Fig. 5c, d and Supplementary Data 28, 29)[28]. Additionally, vascular endothelial cells and peripheral nervous system (PNS) cells showed specific associations with first-order features, emphasizing their roles in lung density[29]. In contrast, chondrocytes and fibroblasts were uniquely linked to shape features, highlighting their contributions to lung morphology[30]. Interestingly, megakaryocytic-erythroid cells in fetal lung were

https://doi.org/10.1038/s41467-025-64571-z

Fig. 4 | Colocalization analysis and pathway analysis. The heatmap shows the overall Bayesian colocalization ('coloc' R package) results between different first-order features (**a**) or shape features (**b**) and significant eQTL genes, which applies a Bayesian framework to estimate posterior probabilities for different hypotheses. For each radiomic feature-eQTL gene colocalization result, the colocalization posteriori probability of hypothesis 4 (PPH4) is indicated with shades of red (closer to 1) and blue (closer to 0). Results with PPH4 > 0.7 are marked with '*', and results with PPH4 > 0.9 are marked with '**'. For the colocalization results of a phenotype across the five lung lobes, the result with the highest PPH4 is shown. The venn diagram illustrates the overlap of candidate genes identified by the four methods (Near gene, SuSIE, coloc, MAGMA) for first-order features (**c**) and shape features (**d**). Pathway enrichment analyses of Gene Ontology terms were performed using 'clusterProfiler' R package. For each GO term, a hypergeometric test (two-sided) was used to assess over-representation of candidate genes (identified from SuSIE, colocalizations analysis, positional mapping, and MAGMA) for first-order features (**e**) and shape features (**f**). *P* values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR < 0.05). The enrichment of the top 15 pathways was visualized. posmap position mapping, SuSIE sum of single effects, coloc colocalization, MAGMA multi-marker analysis of genomic annotation.

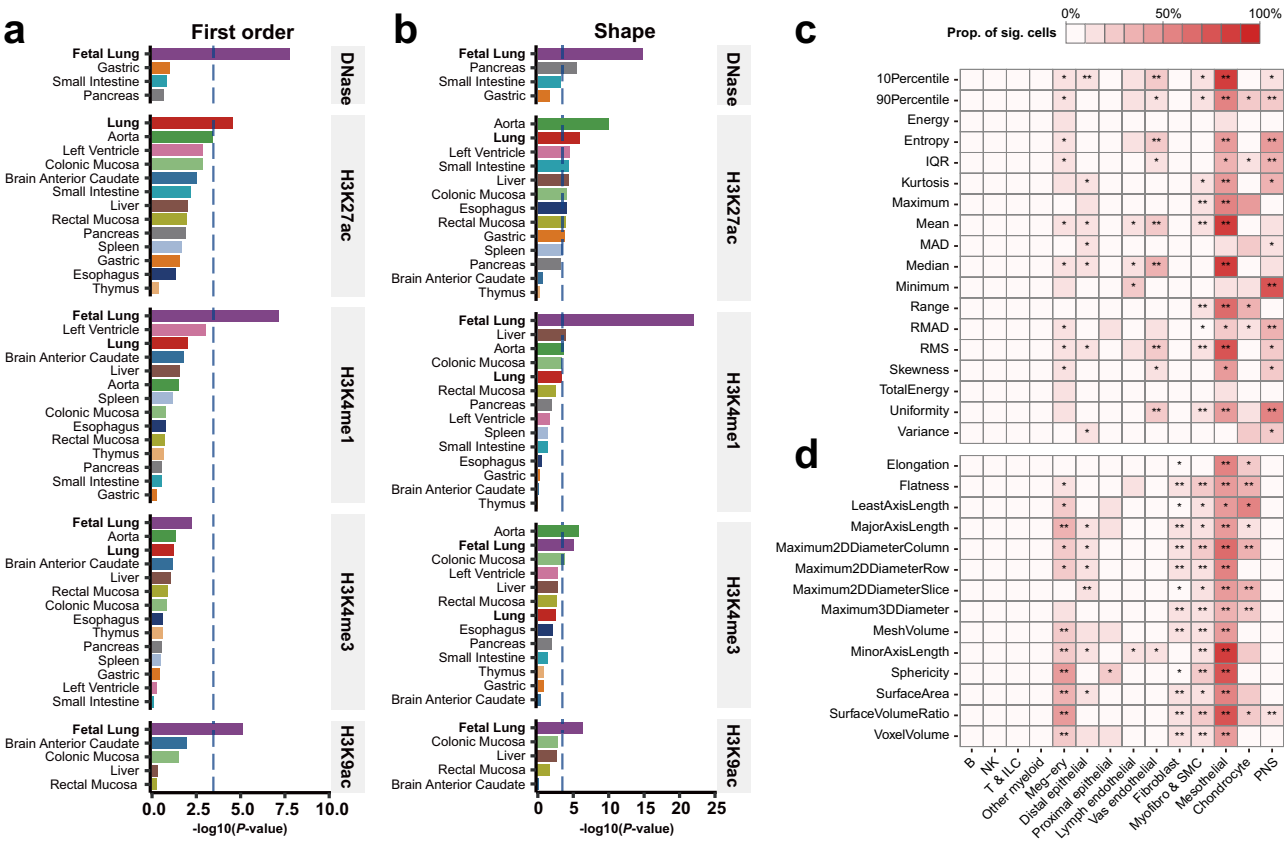


Fig. 5 | Tissue- and cell-type specificity functional annotation. The tissue-specificity functional annotation was performed using the processed GWAS summary data: the minimum BOLT-LMM *P* value for each SNP across all features within the first order or shape categories is extracted. a-b: Heritability enrichment of GWAS summary statistics in functional elements across 14 tissues using stratified LDSC for first-order features (**a**) and shape features (**b**). The x-axis represents the -log value of the enrichment fold *P* value, and the y-axis represents regulatory elements of different tissue types, each tissue category is labeled with a different color. The dashed blue line indicates the Bonferroni significance level. The heatmap illustrates the enrichment of different LIDPs (**c**: first order features; **d**: shape features) across various cell types in fetal lungs at 15 to 22 weeks, evaluated using the single-cell disease relevance score (scDRS). The heatmap colors represent the proportion of significantly associated cells for each cell type–phenotype pair ("Prop. of sig. cells" in the figure legend), with significance defined as $P < 0.05$ across all cells for a given phenotype. Cell-type–phenotype associations with $P < 0.05$ (MC test) are marked with '*', while those with $P_{FDR} < 0.05$ are marked with '**'. NK Natural Killer cells, T & ILC T cells and Innate Lymphoid Cells, Meg-ery megakaryocyte-erythroid progenitor cell, Myofibro & SMC myofibroblasts and smooth muscle cells, PNS peripheral nervous system cells.

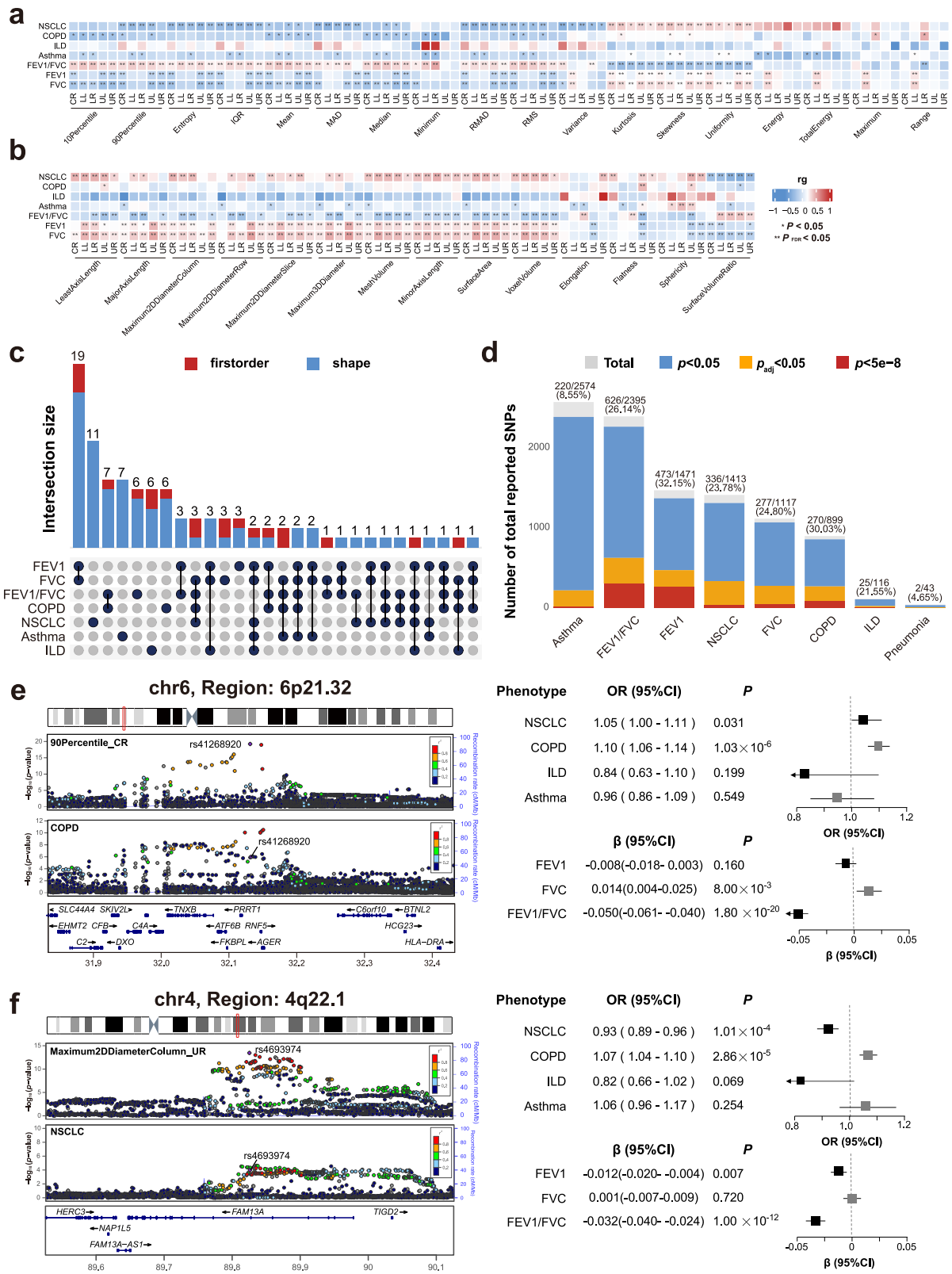
associated with both first-order and shape features, although the underlying mechanisms remain unclear[31].

Genetic correlations and pleiotropy of genetic variants with lung disorders and function

To explore the similarities in genetic architecture between LIDPs and lung disorders and function, we examined genetic correlations between 160 LIDPs and four chronic lung diseases, as well as three lung function traits, using cross-trait LDSC (**Methods**). For lung diseases, we focused on non-small cell lung cancer (NSCLC), chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), and asthma as they represent major chronic respiratory diseases widely studied in epidemiological research[32], covering a broad spectrum of respiratory conditions to explore potential shared genetic factors. The GWAS summary statistics of NSCLC, COPD, ILD, asthma, forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), and FEV1/FVC were derived from our previous study[10], the China Kadoorie Biobank (CKB)[33,34], and the Biobank Japan (BBJ) (**Methods**)[35].

At the FDR 5% level, we observed 196 significant genetic correlations between first-order LIDPs and NSCLC, FEV1/FVC, FEV1, and FVC (Fig. 6a and Supplementary Data 30). Overall, lung voxel intensities (i.e., mean and median voxel intensity) were positively associated with FEV1/FVC but negatively associated with NSCLC, FEV1, and FVC. In contrast, the three distribution characteristics (kurtosis, skewness, and uniformity) showed completely opposite genetic associations. For shape LIDPs, we observed 185 significant genetic correlations with the above lung

https://doi.org/10.1038/s41467-025-64571-z



disorders and lung function (Fig. 6b and Supplementary Data 31). On the whole, lung sizes were positively associated with FEV1, FVC, and NSCLC, while exhibiting a negative association with FEV1/FVC. On the contrary, inverse genetic association was observed for the surface volume ratio. Collectively, these results highlight the extensive genetic associations between lung imaging features and lung disorders, providing evidence for the potential of CT scans in the early assessment of lung health.

To identify the shared genetic effects between LIDPs and lung function and disorders, we further examined the associations of the identified lead SNPs with the four lung diseases and three lung function traits. More than half (53.4%, 93/174) of the identified SNPs were significantly associated with any of the seven traits (Fig. 6c, Supplementary Data 32, 33). Of the 8,051 SNPs recorded for lung function and disorders in the GWAS Catalog, 8.55% (asthma) ~32.15% (FEV1) were

Fig. 6 | Genetic correlations of lung imaging traits and chronic lung diseases and regional plot of representative genetic pleiotropy loci. We illustrated genetic correlations (calculated using LD Score Regression) between first-order features (**a**) or shape features (**b**) and complex traits and diseases. *P*-values were calculated using a Z-test and corrected for multiple comparisons using the Benjamini-Hochberg FDR procedure (two-sided). The genetic correlation is indicated with shades of red (closer to 1) and blue (closer to -1). A single asterisk indicates significance before correction. The double asterisks highlight genetic correlations that have passed multiple testing adjustments using the Benjamini-Hochberg procedure to control the FDR at the 5% level. **c** The Upset plot illustrates the significance (two-sided $P < 0.05$) of independent significant loci identified by two types of features across seven lung-related complex traits and diseases. Red represents loci from first-order features, while blue represents loci from shape features. The lower section of the Upset plot displays the combinations of different significant loci. **d** The bar plot illustrates the significance of loci reported in the GWAS catalog for seven lung-related complex traits and diseases in LIDPs. Gray represents the total number of reported loci, while other different colors indicate loci that are significant in at least one LIDP: blue for GWAS two-sided $P < 0.05$, orange for FDR-adjusted $P < 0.05$, and red for $P < 5 \times 10^{-8}$. The left panels show the genes within a 300 kb upstream and downstream region of the representative lead SNPs, rs41268920 (**e**) and rs4693974 (**f**), along with the regional plots and colocalization with lung diseases COPD (**e**) and NSCLC (**f**). In 6p21.32, we observed colocalization between the 90Percentile_CR and COPD (PPH4 = 0.993); In 4q22.1, we observed colocalization between the Maximum2DDiameterColumn_UR and NSCLC (PPH4 = 0.819). The right panels show forest plots of the effects of these SNPs across seven complex traits and diseases, including their corresponding GWAS two-sided *P*-values in the summary database. For continuous traits (FEV1, FVC, FEV1/FVC), point estimates (beta) with 95% confidence intervals are shown; for binary traits (COPD, ILD, NSCLC), odds ratios (ORs) with 95% confidence intervals are shown. The total sample sizes were: FEV1, FVC, and FEV1/FVC (100,285); COPD (meta-analysis: 10,060 cases/236,329 controls), ILD (meta-analysis: 1201 cases/252,729 controls), NSCLC (13,327 cases/13,328 controls). Detailed information is provided in Supplementary Data 45. NSCLC non-small cell lung cancer, COPD chronic obstructive pulmonary disease, ILD interstitial lung disease, FEV1 the first second of forced expiration, FVC forced vital capacity, UR right upper lobe, UL left upper lobe, CR right centre lobe, LR right lower lobe, LL left lower lobe, OR odds ratio, CI confidence interval.

significantly associated with any of the LIDPs compared with 4.65% for pneumonia after Benjamini-Hochberg correction (Fig. 6d).

We then conducted Bayesian colocalization analysis to evaluate whether LIDPs and chronic lung diseases share causal variants. For first-order LIDPs, we identified 215 colocalization signals, primarily in three loci: 4q22.1 (*FAM13A*), 6p21.32 (*AGER*), and 14q13.1 (*NPAS3*) (Supplementary Data 34). For example, at 6p21.32 (*AGER*), colocalization was observed between the 90th percentile of voxel intensities (90Percentile) of the right central lobe and COPD. The lead SNP, rs41268920, was also significantly associated with 50 first-order LIDPs and showed associations with NSCLC, FVC, and FEV1/FVC (Fig. 6e). For shape LIDPs, we identified 61 colocalization signals across 11 loci, including 4p15.31 (*DCAF16*), 4q22.1(*FAM13A*), 4q31.21 (*HHIP*), 5p12 (*FGF10*), 6p21.32 (*AGER*), 8q23.1 (*RSPO2*), 10q25.2 (*VTI1A*), 12q24.13 (*NAA25*), 15q25.2 (*AP3B2*), 16q24.1 (*LINC00917*), and 19q13.42 (*LILRA3*) (Supplementary Data 35). For instance, at 4q22.1 (*FAM13A*), colocalization was observed between the maximum 2D diameter column of the right upper lobe and NSCLC. The lead SNP, rs4693974, was also significantly associated with COPD, FEV1, and FEV1/FVC (Fig. 6f). Additionally, we assessed colocalization between genome-wide significant loci ($P < 5 \times 10^{-8}$) for the four chronic lung diseases and LIDPs, identifying three loci for NSCLC, four for COPD, and two for asthma that colocalized with at least one LIDP (Supplementary Data 36). These findings demonstrate that LIDPs share substantial genetic components with NSCLC, COPD, asthma, and ILD, highlighting potential connections between lung structural phenotypes and disorders.

Phenome-wide Mendelian-randomization analysis with lung-related traits and other diseases

In light of the widespread genetic correlations between LIDPs and lung-related traits, we examined the underlying causal genetic links between the 160 LIDPs and lung function and disorders using Mendelian randomization (MR). Additionally, we explored potential causal links with 90 additional diseases out of the lung in the BBJ **(Methods)**.

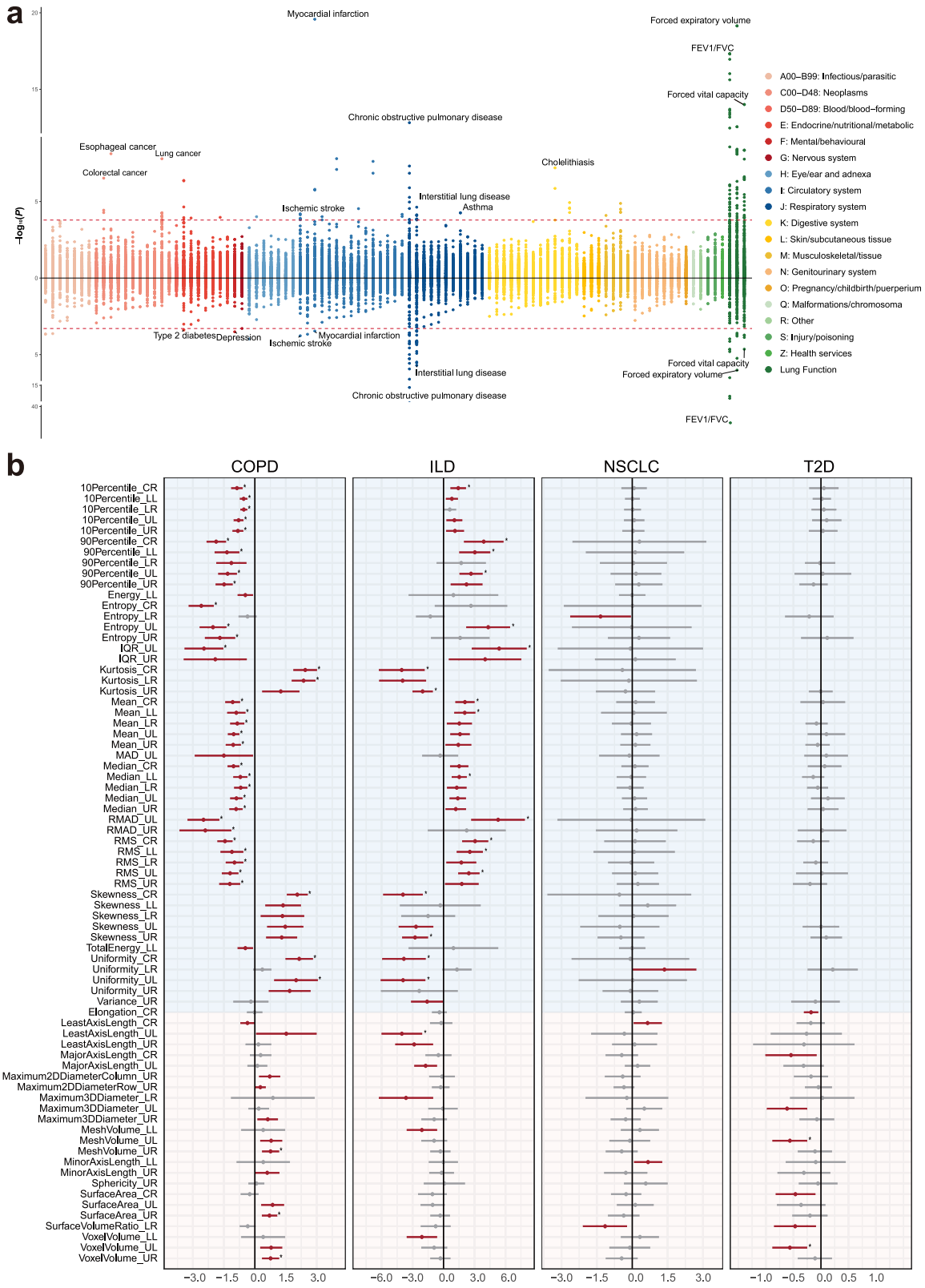
We identified causal genetic links underlying LIDPs and lung functions. Specifically, 36 first-order LIDPs to FEV1/FVC, three first-order and one shape LIDPs to FVC, and five first-order LIDPs to FEV1 were identified after Bonferroni's correction ($P < 3.88 \times 10^{-4}$, 0.05/129, considering 129 LIDPs analyzed in the MR analysis) (Fig. 7a and Supplementary Data 37). Meanwhile, we observed robust evidence of causal genetic links between LIDPs on COPD as well as ILD (Fig. 7b). Although several LIDPs (i.e., surface volume ratio of right lower lobe, odds ratio=0.32, $P = 0.016$) showed potential causal genetic effects on NSCLC, no significant associations were observed after Bonferroni's correction. Among the 103 significant associations identified by the IVW method, 65 had sufficient instrumental variables (IVs) for MR-Egger analysis, and 62 (95.4%) of them showed consistent effect directions across both methods, supporting the robustness of our findings (Supplementary Data 38). Besides, horizontal pleiotropy tests suggested that the reported associations are unlikely to be driven by confounding and are more likely to reflect true causal effects, further supporting the reliability of the results (Supplementary Data 39).

In addition to lung-related traits, multiple genetic causal effects of LIDPs on type 2 diabetes (T2D), myocardial infraction (MI), and ischemic stroke (IS) were identified after Bonferroni's correction ($P < 5.05 \times 10^{-4}$, 0.05/99, considering 99 LIDPs analyzed in the MR analysis) (Fig. 7a). A total of 17 loci associated with LIDPs were also significantly associated with T2D, MI, or IS (Supplementary Data 40) after Bonferroni's correction ($P < 1.62 \times 10^{-4}$, 0.05/308, considering 308 loci identified in BBJ database). Seven shape LIDPs showed potential genetic causal associations with T2D, with the volume of left upper lung lobe being significant after Bonferroni's correction (Fig. 7b). The association between lung function and T2D has been reported in several cohort studies and MR analyses[36,37]. Similarly, COPD and impaired lung function have been strongly associated with cardiovascular disease, particularly MI[38]. Our results provide new insights into the complex etiology of cross-organ interactions.

Implicated genes highlight druggable targets

The candidate genes were investigated for known gene-drug interactions using the Drug-Gene Interaction Database. Nearly 37.8% (45/119) of the first-order candidate genes and 43.7% (131/300) of the shape candidate genes were interacted with approved drugs or drugs in development (Supplementary Data 41 and Supplementary Data 42). Here, we highlight two examples of new genetic signals implicating targets for drugs utilization. One of our signals is an eQTL for *AGER*, which was colocalized with 27 LIDPs and COPD (Fig. 6e). *AGER* encodes advanced glycosylation end-product specific receptor, which is interacted with PYRIDOXAMINE (vitamin B6). AGER is selectively and specifically overexpressed in lung tissue (Supplementary Fig. 11). According to 46,807 participants with proteomics data from the UK Biobank, the lead SNP rs41268920 was significantly associated with reduced level of AGER in plasma, and high AGER in plasma were significantly associated with reduced risk of lung cancer, COPD, ILD, and asthma in smokers (Supplementary Fig. 12). Notably, inverse association between serum levels of vitamin B6 and lung cancer risk have been reported by the European Prospective Investigation into Cancer and Nutrition (EPIC)[39]. Another signal is *MAPKAPK5*, which encodes a kinase enzyme involved in key inflammatory pathways. *MAPKAPK5* can be inhibited by a small-molecule inhibitor, GLPG-0259, which is under investigation in a

https://doi.org/10.1038/s41467-025-64571-z



clinical trial (NCT01024517). The signal colocalized with *MAPKAPK5* expression in lung tissue and was linked to an increased risk of COPD and asthma, as well as reduced lung function (Supplementary Fig. 13).

Discussion

In this study, we analyzed the genetic architecture of the five lung lobes using CT scans from 35,469 Chinese Han participants in the LIGI. To the best of our knowledge, this is the first GWAS study of lung radiomics, and all GWAS summary statistics are freely available online (http://ccra.njmu.edu.cn/LIGI/). Our analysis revealed 36 genomic regions influencing the lung voxel intensities and 138 genomic regions affecting lung shape. These findings advanced our understanding of the genetic architecture underlying lung organization.

Fig. 7 | Genetic causal effects of lung imaging traits and lung diseases. a The scatter plots illustrate the correlation and causality between lung radiomic features and 97 complex phenotypes. The top part of the figure shows the two sided *P* values of identified lead SNPs in 97 complex traits, while the bottom part displays the Mendelian randomization *P* values (two sided, using the IVW technique as our major model). The horizontal dashed lines indicate the Bonferroni-corrected *P* value threshold (0.05/308 for upper part and 0.05/99 for lower part, considering 308 loci identified in the BBJ and 99 LIDPs analyzed in the Mendelian randomization in BBJ). Different categories of diseases or phenotypes are marked with different colors. **b** From left to right, the forest plots display the specific Mendelian randomization results for COPD, ILD, NSCLC, and T2D with lung radiomic features. Light blue areas indicate first-order features, and light red areas indicate shape features. Data are presented as Mendelian randomization IVW estimates (beta) ± 95% CIs. The dots represent the beta values from the Mendelian randomization IVW results. The horizontal bars represent 95% CIs, with color and an asterisk indicating significance levels: red denotes nominal significance ($P < 0.05$), an asterisk indicates significance after Bonferroni's correction ($P < 3.91 \times 10^{-4}$ for COPD and ILD, 0.05/128, considering 128 LIDPs analyzed in the MR analysis with COPD and ILD; $P < 3.88 \times 10^{-4}$ for NSCLC, 0.05/129, considering 129 LIDPs analyzed in the MR analysis with NSCLC; $P < 5.05 \times 10^{-4}$ for T2D, 0.05/99, considering 99 LIDPs analyzed in the MR analysis with T2D), and gray indicates nonsignificant correlations. The reported *P*-values were two-sided. The sample sizes were: COPD (meta-analysis: 10,060 cases/236,329 controls), ILD (meta-analysis: 1201 cases/252,729 controls), NSCLC (13,327 cases/13,328 controls), and T2D (45,383 cases/132,032 controls). Exact *P* values, effect sizes, and 95% CIs are provided in Supplementary Data 37-38. IVW inverse-variance weighted, COPD chronic obstructive pulmonary disease, ILD interstitial lung disease, NSCLC non-small cell lung cancer, T2D type 2 diabetes, UR right upper lobe, UL left upper lobe, CR right centre lobe, LR right lower lobe, LL left lower lobe, IQR InterquartileRange, MAD MeanAbsoluteDeviation, RMAD RobustMeanAbsoluteDeviation, RMS RootMeanSquared, CI confidence interval.

Using a deep-learning AI system for the segmentation of lung CT imaging, we were able to explore the genetic architecture of understudied lung radiomics in each of the five lobes, respectively. This improvement allowed for a more precise description of lung phenotypes and increased the discoverability of susceptibility loci. Of the 174 identified loci, 103 were significant in only one lobe, which might have been missed without segmentation. This finding aligns with the understanding that the lung is a fundamentally heterogeneous organ, with differences in perfusion-ventilation ratio, lymphatic flow, metabolism, mechanics, and the uneven distribution of lung diseases across lung lobes[40,41].

Our findings demonstrate that genetic variations affecting gene regulation during fetal lung development significantly influence adult lung structures. Heterozygous single-nucleotide variants (nonsense, frameshift) and copy number variant deletions in the *TBX4-FGF10-FGFR2* epithelial-mesenchymal signaling pathway, as well as other genes such as *TCF21*, have been observed in over 65% of newborns with lung dysplasia[42]. In this study, we observed that common noncoding genetic variants near *FGF10* and *TCF21*, along with their family genes such as *TBX3*/*TBX5*/*TBX18* and *FGF18*, can consistently affect lung structure, though to varying degrees[43,44]. In addition, SNPs near several TFs involved in branching morphogenesis (e.g., *GATA6*, *ISL1*)[45,46], proximal-distal patterning of epithelial cells (e.g., *SOX6*, *THRB*, *EYA1*, and *KLF5*)[47,48], and maintaining spatial specification of airway and vascular smooth muscle cells (e.g., *HHIP*, *MEF2C*, *WT1*) in fetal lung development[47] were also identified to be associated with adult lung structure in this study. Dysfunction of these TFs has been related to various abnormalities in airway development and regeneration in model organisms[49–51]. Although our identified SNPs were noncoding, they were significantly enriched in the regulation regions of the fetal lung. Taken together, our results provided genetic evidence of gene regulation during fetal lung development affects adult lung structure.

We demonstrate that genetic variations associated with lung structure also impact lung function and common chronic lung diseases. Notably, we identified three loci (linked to *FAM13A*, *HHIP*, and *AGER*) that affect multiple lung structure traits and colocalize with chronic lung diseases. In experimental animal models, *FAM13A*-deficient mice are protected from cigarette smoke-induced alveolar simplification through increased β-catenin signaling and subsequent increased epithelial cellular proliferation[52]. *AGER*-deficient mice showed increased albumin in the bronchoalveolar lavage fluid and are protected from alveolar damage and inflammation induced by hyperoxia[53]. *HHIP* in adult human lungs is mainly expressed by type 2 (AT2) alveolar epithelial cells, and *HHIP*-insufficient mice develop alveolar simplification when exposed to cigarette smoke, a representative symptom in COPD patients[54]. Mechanistic studies suggest that these genes may be involved in lung epithelial repair in response to injury[55]. In addition to the aforementioned chronic lung diseases, we also detected multiple loci related to pulmonary arterial hypertension (*KCNK3*)[56], acute lung injury (*KCNQ5*, *KIF26B*, *ACKR3*, *PLXND1*, and *ESR1*)[57–61], and severe COVID-19 (*DOCK2*, and *ERAP2*)[62,63]. Our analysis demonstrated several colocalization loci shared by lung structures and chronic lung diseases, perhaps suggesting that lung structure is a key phenotype along the causal pathway from genetic variation to the pathogenesis of lung diseases. These findings suggest that LIDPs may capture preclinical structural alterations, offering promising avenues for developing non-invasive tools for risk stratification, early diagnosis, and longitudinal monitoring of chronic lung diseases.

Although we identified multiple associations of LIDPs, this study has several potential limitations. First, pulmonary function tests were not performed in the LIGI. Considering the overall prevalence of spirometry-defined COPD is nearly 8.6% among individuals aged 40 years or order, our study might include partially preclinical patients[64]. Second, our findings may not be generalizable to individuals of European or other Asian ancestries. The Lung Cell Atlas database used for enrichment analysis is predominantly of European ancestry and does not match the LD patterns of the Chinese population, potentially reducing the power of our analysis. Currently, there are no published GWAS of similar radiomic features in European or other populations, which limits our ability to conduct direct cross-ancestry comparisons or identify ancestry-specific associations. Third, the publicly available GWAS summary statistics used in the colocalization analyses and lookup approaches were not checked for quality, which might introduce bias. Additionally, due to the limited number of incident cases of chronic lung diseases from CKB and BBJ, the GWAS summary statistics for chronic lung diseases used in our analysis are generally underpowered. Finally, the current study focused only on the first-order and shape features of CT images, which describe the size and density of the lung lobes. However, many high-order features, which capture the texture of the lungs by describing the spatial relationships among multiple voxels, will be studied in our future studies with larger sample sizes.

In summary, we identified several common genetic variants that contribute to variations in lung imaging phenotypes. These variants appear to influence lung development by regulating the gene expression and biological pathways active during fetal lung development. Our findings also underscore the potential connections between lung structures and chronic lung diseases, as well as other health conditions. The discovery of these common variants affecting lung structure has become feasible through collaborative analysis of CT data, paving the way for uncovering genetic mechanisms underlying lung development and disease.

Methods

Study participants

The research reported herein was conducted in compliance with all relevant ethical regulations and in accordance with the Declaration of

Helsinki. The Lung Imaging Genomics Initiative (LIGI) aims to reveal the genetic basis of lung radiomics in the Chinese population and establish the radiation-phenotype causal link of common respiratory diseases. All participants in this study were recruited from the LIGI in China, which collected genomic and lung imaging data from Zhejiang and Jiangsu provinces between April 2019 and November 2021. In Zhejiang, we recruited 28,185 participants who participated in the free lung cancer screening program organized by the local government. In Jiangsu, we recruited 12,007 participants from the routine physical examination population with pulmonary CT examination in the Health Promotion Center of the First Affiliated Hospital of Nanjing Medical University. Inclusion and exclusion criteria for the LIGI participants are provided in Supplementary Data 43. The LIGI study received approval from the Ethics Committee of Taizhou Cancer Hospital, the First Affiliated Hospital with Nanjing Medical University, and written informed consent was obtained from each participant. All participants were requested to complete an interviewer-administered electronic questionnaire covering lifestyle and other health-related information, as well as provide physical measurements and a blood sample at baseline. After excluding those who refused to undergo LDCT examination, declined to provide a blood sample, or had missing covariates, 36,551 eligible individuals were included. Further exclusions were made for participants who failed manual checks by radiologists ($N = 162$) or radiomic feature extraction ($N = 130$), were identified as extreme phenotypic outliers (±6 SD) through principal component analysis (PCA) ($N = 624$), or did not pass the quality control of genomic data ($N = 166$). As a result, Phase I of the LIGI study included 35,469 participants with qualified CT images and genotype data. The participant inclusion and exclusion process is detailed in Supplementary Fig. 1.

Imaging acquisition and radiomic feature extraction

In the LIGI, lung imaging data were acquired using a 16-slice spiral CT scanner in Zhejiang and a 128-slice spiral CT scanner in Jiangsu (120 kV and 40-60 mA). The detailed acquisition parameters were as follows: field of view 500 mm; collimation 16 × 0.6 mm; rotation time 0.5 s; and pitch 1.5. The scan matrix was no less than 512 × 512 pixels, and the images were reconstructed with a slice thickness ranging from 1.00 to 1.25 mm. The lung window settings were as follows: window width 1500 HU, window level −600 to −500 HU. Participants were instructed to take a deep breath and to hold their breath during scanning (5–8 s). The scanning range of the LDCT extended from the lung apex to the level of the posterior costophrenic angle (including the entire lungs and both chest walls, with full breast coverage for female participants).

To minimize inter-center variability introduced by differences in scanner models and imaging protocols, all LDCT images underwent a standardized preprocessing workflow prior to radiomic feature extraction. Specifically, images were resampled to isotropic 1 × 1 × 1 mm^3 voxel spacing, intensities were normalized within a standard lung window, and ComBat harmonization was applied to reduce batch effects. A deep-learning AI system (developed by Beijing Deepwise & League of PhD Technology Co. Ltd) was used to segment the five lung lobes[17]. The system was initially trained on a large dataset[65] and has been successfully employed in previous studies to segment lung lobes in COVID-19 patients[66,67]. Its segmentation accuracy has been independently validated, achieving an average Dice score of 0.95 across lung lobes as confirmed by experienced radiologists[17]. In our study, we randomly selected 800 images for re-evaluation by two radiologists, and the accuracy rate of automatic segmentation is 99.9% (799/800). Additionally, participants were excluded if fewer than five lung lobes were successfully segmented (i.e., zero to four lobes detected). Review of these excluded cases revealed that segmentation failures were primarily due to poor image quality (e.g., motion artifacts), incomplete lung coverage, or anatomical alterations such as atelectasis. After segmentation, radiomic features were extracted using *Pyradiomics* (v3.0) for each lobe[18]. A total of 1470 quantitative radiomics features were extracted from each lobe per patient, including 18 first order features, 14 shape features, 22 Gray Level Cooccurrence Matrix (GLCM) features, 16 Gray Level Run Length Matrix (GLRLM) features, 16 Gray Level Size Zone Matrix (GLSZM) features, 14 Gray Level Dependence Matrix (GLDM) features, and 5 Neighboring Gray Tone Difference Matrix (NGTDM) features, while the image type consisted of Original, Wavelet, Laplacian of Gaussian (LoG), Square, Square Root, Logarithm, Exponential, Gradient and LocalBinaryPattern2D. In this study, we focused on 18 first-order features describing the distribution of voxel intensities (density) and 14 shape features detailing the three-dimensional size and shape of each lobe. First-order features describe the distribution of voxel intensities within a lung lobe and are directly related to tissue density, while shape features describe the geometrical characteristics of each lung lobe. Detailed definitions of the 32 lung image-derived phenotypes (LIDPs) by the Imaging Biomarker Standardization Initiative (IBSI)[68] and representative interpretations are shown in Supplementary Data 44 and Supplementary Fig. 14.

Imaging data quality control and preprocessing

Radiographic images from participants eligible for the LIGI were manually checked by a team of thoracic radiologists from Zhejiang Cancer Hospital, each with over 8 years of experience in radiographic diagnosis. Participants were excluded based on the following criteria: (1) incomplete LDCT scans, (2) severe artifacts affecting feature extraction and (3) presence of pneumoconiosis. All manually validated images will undergo lung lobe segmentation and feature extraction (as mentioned above). The accuracy of segmentation by the AI system was verified manually by two radiologists with more than 8 years of diagnostic experience in chest imaging. We conducted a repeat assessment in 100 participants who underwent two rounds of LDCT screening in 2022 in Zhejiang. The consistency of radiomic features between the two assessments reached an average correlation coefficient of 0.853, indicating high reproducibility. The radiomic features were standardized using the scale function in R, and PCA was performed on the standardized radiomic features. We used the first 15 principal components (PCs), which explained over 85% of the variance, to calculate a principal component score for each sample. Participants with scores lying beyond six times the standard deviation (±6 SD) from the mean were considered outliers and excluded from further analysis.

Genotyping, imputation and genetic quality control

Participants were genotyped with Infinium Asian Screening Array BeadChip (Illumina, Inc., San Diego, CA, United States) with ~750,000 markers. Given the high racial homogeneity and substantial sample size of the LIGI, we conducted quality control following standard practices established in previous literatures[69,70]. Briefly, genetic quality control was carried out at both the variant and sample levels. At the variant level, we removed duplicated markers, excluded markers with high allele missing rates (>5%), significant deviations from Hardy-Weinberg equilibrium (HWE) ($P < 10^{-7}$), or a minor allele frequency (MAF) < 0.1%. At the sample level, we removed samples with sex mismatches and excluded those with high heterogeneity (>10 SD) or missing rates (>5%). Finally, a total of 582,663 SNPs from 35,469 participants were included in the subsequent analyses.

Imputation was performed based on a two-staged strategy using a combined reference panel of the 1000 Genomes Phase 3 reference panel ($N = 2504$) and the Nanjing Medical University Omics database ($N = 3020$)[71] with SHAPEIT4[70] and IMPUTE2[72]. After excluding SNPs with MAF < 1% and an imputation information score (INFO) < 0.3, a total of 8,447,934 SNPs were included in the final GWAS analysis. Variant positions were mapped to the GRCh37 human genome reference.

Heritability and GWAS association analyses

BOLT-REML (v2.3.4) was used to estimate the SNP heritability of the 160 LIDPs (32×5) based on the qualified imputed autosomal SNPs. Single-variant association analysis was performed using BOLT-LMM (v2.4), which accounts for cryptic population structure and sample relatedness[73]. In the genetic analysis, we adjusted for sex, age, BMI, pack years of smoking, region (Zhejiang/Jiangsu), and the first ten principal components[74]. Genetic variants on the X chromosome were analyzed using the generalized linear model (GLM) in PLINK-2.0[75], with the same covariates described above. Variants with an association $P < 5 \times 10^{-8}$ were considered genome-wide significant.

For each LIDP, we extracted all significant variants ($P < 5 \times 10^{-8}$) and then performed LD clumping using PLINK-1.9[75] with the East Asian (EAS) reference panel from the 1000 Genomes Project (the Phase III integrated variant set release, 504 East Asians). We outlined a 500 kb window (–clump-kb 500) and used a common LD threshold (–r^2 0.1) to identify independently significant SNPs for each LIDP (Supplementary Data 5 and Supplementary Data 7). Then, the independent lead SNPs of different LIDPs of first-order and shape, respectively, were combined to define a cluster by merging variants within 500 kb of each other. The variants with the smallest *P* value in each cluster was defined as significant lead SNPs for the first-order LIDPs and shape LIDPs, which were reported in Supplementary Data 6 and Supplementary Data 8. Subgroup analysis were further performed according to region (Zhejiang/Jiangsu) and smoking status. Locus plots of independently significant SNPs were produced with LocusZoom and verified by manual visual inspection[76].

Genetic correlation estimation

Genetic correlation across traits was assessed using cross-trait LDSC[77,78] within 160 LIDPs and 7 lung-related traits (FEV1, FVC, FEV1/FVC, COPD, ILD, asthma, and NSCLC). The genome-wide summary statistics of lung function were derived from the

China Kadoorie Biobank (CKB)[33]; the genome-wide summary statistics of COPD, ILD, and asthma were derived from the meta-analysis of CKB and the Biobank Japan (BBJ); and the genome-wide summary statistics of lung cancer were derived from our previous study[10]. For COPD, ILD, and asthma, we conducted genome-wide meta-analyses to enhance statistical power and robustness. The meta-analyses were performed using the METAL software[79], employing a fixed-effect inverse-variance weighting approach; more details can be found in Supplementary Data 45. We first prepared GWAS summary statistics for each trait, ensuring that they were aligned to the same reference allele. Only SNPs available in both datasets and passing quality control were retained. We employed the LD scores provided by the authors for the East Asian population, which were estimated from 1KG EAS individuals. Genetic correlation pairs at an FDR 5% level were considered statistically significant.

With LDSC, the genomic control factor (lambda GC) was partitioned into components reflecting polygenicity and inflation, using the software's defaults. The LDSC intercept was utilized to evaluate polygenicity and the genomic inflation factor. An intercept value closer to 1 suggests minimal population stratification and confounding due to cryptic relatedness or other biases.

Conditional analysis and Statistical fine-mapping

For each LIDP, we extracted all SNPs within ±1 Mb of each sentinel variant and employed the GCTA conditional and joint association analysis (GCTA-COJO)[80] to conduct a stepwise conditional analysis to select independent association signals ($P < 5.0 \times 10^{-8}$) for each locus.

Statistical fine-mapping was performed with SuSiE (v.0.11.92; https://github.com/stephenslab/susieR), allowing for up to five putative causal variants within each locus. The LD structure was referenced against the EAS dataset from the 1000 Genomes Project. Each region was defined by a ± 250 kb window centered on the lead variants. Variants in the 95% credible sets (representing a 95% likelihood of harboring at least one causal variant) were assessed for their predicted functional effects using the Variant Effect Predictor (VEP, https://grch37.ensembl.org/) and SNPnexus (https://www.snp-nexus.org/v4/)[81]. We annotated four scores, including fitcons, eigen, FATHMM and CADD, to assess the potential biological function of each variant. For missense variants in coding regions, we further used SIFT and PolyPhen to predict the pathogenicity.

Tissue-type-specific heritability enrichment analysis

Stratified LD score regression (S-LDSC, https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses) was employed for tissue-type-specific heritability enrichment estimation. We first generated two artificial genome-wide summary statistics for first-order features and shape features by retaining the statistical parameters of the corresponding LIDP with the smallest *P* value for each variant. Heritability enrichment was then performed for transcription regulation regions marked by histone medications (H3K4me1, H3K4me3, H3K9ac, and H3K24ac) and DNase hypersensitivity sites in 13 common adult tissues and fetal lung tissues. The partitioned LD scores of all annotations and the baseline model for EAS ancestry on HapMap3 SNPs were downloaded from the Broad Institute's repository (https://alkesgroup.broadinstitute.org/LDSCORE). Meanwhile, we also used the GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD, https://www.ebi.ac.uk/birney-srv/GARFIELD) to evaluate the enrichment of significant SNPs (5×10^{-5} and 5×10^{-5}) in regulatory regions from the ENCODE and Roadmap projects with default parameters.

Cell-type specificity analysis

To identify the cell types associated with different LIDPs, processed scRNA-seq datasets of human fetal lung were obtained from the Lung Cell Atlas (https://lungcellatlas.org). The single-cell disease relevance score (scDRS)[82] was used to assess polygenic disease enrichment in fetal lung cells at 15-22 weeks. Given the strong correlations among the 32 LIDPs across the five lung lobes (Supplementary Fig. 3), we combined results for each gene across the five lobes, retaining the most significant *P* values in MAGMA. The top 1000 genes for each phenotype, along with their weights (absolute values of ZSTAT), were used as inputs for scDRS. In the enrichment analysis, we adjusted for sex, total RNA count, and the number of detected genes. Using the downstream pipeline provided by scDRS, we also performed Monte Carlo (MC) testing to evaluate the association results for 14 broad cell types.

eQTL analysis of lung tissues

We have built an eQTL database involving 116 adjacent lung tissues in our previous study[71]. Here, we further collected 222 noncancerous lung tissues and matched blood samples from the Nanjing Chest Hospital. The DNA/RNA sample extraction, sequencing, and data processing were in line with our previous publication[71]. The genotyping of the 222 samples was performed with the Infinium Asian Screening Array BeadChip, and data processing was consistent with the above description. eQTL analysis was performed with FastQTL[83] according to the standard pipeline of GTEx[84]. The expression of each gene was normalized using an inverse normal transform. We adjusted for age, sex, smoking, sequencing batches, the top five principal components, and 45 Probabilistic Estimation of Expression Residuals (PEER) factors in a linear regression model.

Bayesian colocalization analysis

Bayesian colocalization analysis[85] was performed using the coloc package (version 5.2.2; https://chr1swallace.github.io/coloc) for each significant locus of the LIDPs and eQTL signals. Evidence of pairwise colocalization was defined as having a posterior probability of the

shared causal variant hypothesis (PPH4) > 0.7. EAS LD matrix from 1000 Genomes (phase 3) was incorporated into the LD-dependent approach. In addition, we also performed Bayesian colocalization analysis for LIDPs and the seven lung-related traits to evaluate whether two associated genetic signals were consistent with the shared causal variants.

Positional Mapping by FUMA

The web tool Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) has previously been described in detail (https://fuma.ctglab.nl/)[86]. We utilized positional mapping to associate genome-wide significant loci ($P < 5 \times 10^{-8}$ and $r^2 < 0.1$) with genes using FUMA's default settings and specialized datasets. Variants within a 10 kb window of known protein-coding genes in the human reference assembly (GRCh37/hg19) are mapped accordingly.

Gene-based analysis

The SNP-based *P* values were used for gene-based analysis using MAGMA[87] software (http://ctg.cncr.nl/software/magma) for gene-based analysis. MAGMA used a multiple regression approach to properly incorporate LD between markers and to detect multi-marker effects for a genome-wide gene association analysis, thereby reducing the potential inflation of association signals caused by correlated SNPs. We applied a stringent Bonferroni correction to account for multiple testing, considering associations with $P < 2.70 \times 10^{-6}$ (0.05/18,517) as statistically significant.

Gene Ontology (GO) enrichment analysis

We performed GO enrichment analysis on the candidate target genes identified through positional mapping, colocalized eQTL analysis, nonsynonymous mutations detected by SuSiE, and the genes identified by gene-based analysis. This analysis aimed to explore the enrichment of these genes in pathways defined in biological processes (BP), cellular components (CC), and molecular functions (MF). The GO analysis was performed using the clusterProfiler R package[88]. Pathways with a BH-adjusted *P* value < 0.05 were considered significant and retained for further investigation.

Phenome-wide Mendelian-randomization analysis

To investigate whether variants associated with LIDPs were also associated with other human complex traits in EAS, we obtained statistics for the non-overlapping lead variants across 90 disease traits with case≥500 from BBJ PheWeb (https://pheweb.jp). Variants were considered to exhibit pleiotropy if they met a significance threshold of $P < 3.13 \times 10^{-4}$ after multiple-test corrections.

We further evaluated the genetic causal relationships between the 160 LIDPs and seven lung-related complex traits, as well as 90 diseases from the BBJ using Mendelian randomization (MR) analysis. We preprocessed the GWAS summary statistics according to the standard MR preprocessing procedures. Specifically, in the exposure GWAS, the genetic variants were initially chosen at a significance level of 5×10^{-8}. To ensure that the genetic variants included in the downstream MR analysis were independent, we performed a LD-based clumping firstly with window size of 1 Mb and $r^2 < 0.01$ taking the EAS in 1KG as a reference panel. We used the harmonization procedure in the TwoSampleMR package (https://mrcieu.github.io/TwoSampleMR/) to infer the correct allele alignment. The inverse variance weighted (IVW) model was used as the major reported model, and significant results were prioritized based on a Bonferroni correction threshold of $P < 3.13 \times 10^{-4}$, which are reported in Supplementary Data 40.

Drug targets

Candidate genes with nonsynonymous variants in the credible sets of SuSiE, identified through positional mapping, colocalized eQTL analysis, and gene-based analysis, were cross-referenced with the gene-drug interactions table in the Drug-Gene Interactions Database (DGIdb, https://www.dgidb.org/). Mapped drugs were assigned corresponding CHEMBL IDs, and information regarding clinical trial data and indications for each drug was obtained from ChEMBL (https://www.ebi.ac.uk/chembl/).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genome-wide summary statistics for the 160 LIDPs generated in this study have been deposited in the Genome Variation Map (GVM) at National Genomics Data Center, China National Center for Bioinformation, under accession number GVP000047 (https://ngdc.cncb.ac.cn/gvm/). The summary statistics are available upon request, subject to participant privacy and ethical restrictions. Researchers may request access through the CNCB online application system by submitting a brief description of the intended scientific use. Requests are reviewed by the LIGI data access committee, and applicants will typically receive a response within 2–4 weeks. Access is granted solely for academic, non-commercial research purposes, and data use agreements prohibit attempts to re-identify participants or to use the data beyond the approved scope. Researchers can also query single-locus level results through the LIGI database (http://ccra.njmu.edu.cn/LIGI/). All variant positions are mapped to the GRCh37 human genome reference. Genome-wide summary statistics for lung function, chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), and asthma were obtained from the China Kadoorie Biobank (CKB) through the CKB data access system (https://www.ckbiobank.org/). Summary statistics for 90 complex traits from Biobank Japan (BBJ) were obtained from the BBJ PheWeb (https://pheweb.jp). The 1000 Genomes reference panel and East Asian (EAS) cell-type annotations for S-LDSC analyses were acquired from the Broad Institute's repository (https://alkesgroup.broadinstitute.org/LDSCORE). Regulatory region data for GARFIELD analyses were obtained from the GARFIELD website (https://www.ebi.ac.uk/birney-srv/GARFIELD). Single-cell RNA-seq data used in S-LDSC analyses were obtained from the Lung Cell Atlas (https://lungcellatlas.org/). Additional details on the data sources used in this study are provided in the main text and Supplementary Tables. Source data for all analyses are provided as a Source Data file. Source data are provided with this paper.

Code availability

Publicly available software and packages were used for bioinformatics analysis in the present study. The software and packages used in each analysis are described in Methods. The code used to perform the analyses in this study is publicly available at Zenodo (https://doi.org/10.5281/zenodo.17071649)[89]. Users are permitted to reuse, modify, and distribute the code in accordance with the terms of the license, with appropriate attribution to the original authors.

References

1. Schittny, J. C. Development of the lung. *Cell Tissue Res.* **367**, 427–444 (2017).
2. Elbehairy, A. F. et al. Advances in COPD imaging using CT and MRI: linkage with lung physiology and clinical outcomes. *Eur. Respir. J.* **63**, 2301010 (2024).
3. Zhou, T. H. et al. CT whole lung radiomic nomogram: a potential biomarker for lung function evaluation and identification of COPD. *Mil. Med Res.* **11**, 14 (2024).
4. Schniering, J. et al. Computed tomography-based radiomics decodes prognostic and molecular differences in interstitial lung disease related to systemic sclerosis. *Eur. Respir. J.* **59**, 2004503 (2022).

5. Shima, H. et al. Lobar distribution of non-emphysematous gas trapping and lung hyperinflation in chronic obstructive pulmonary disease. *Respir. Investig.* **58**, 246–254 (2020).
6. Lauria, M. et al. An analysis of the regional heterogeneity in tissue elasticity in lung cancer patients with COPD. *Front Med (Lausanne)* **10**, 1151867 (2023).
7. Thebaud, B. et al. Bronchopulmonary dysplasia. *Nat. Rev. Dis. Prim.* **5**, 78 (2019).
8. Galambos, C. et al. Phenotype characterisation of TBX4 mutation and deletion carriers with neonatal and paediatric pulmonary hypertension. *Eur. Respir. J.* **54**, 1801965 (2019).
9. Turcatel, G. et al. Lung mesenchymal expression of Sox9 plays a critical role in tracheal development. *BMC Biol.* **11**, 117 (2013).
10. Whitsett, J. A., Wert, S. E. & Trapnell, B. C. Genetic disorders influencing lung formation and function at birth. *Hum. Mol. Genet.* **13**, R207–R215 (2004).
11. Sakornsakolpat, P. et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* **51**, 494–505 (2019).
12. Allen, R. J. et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **201**, 564–574 (2020).
13. Tsuo, K. et al. Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *Cell Genom.* **2**, 100212 (2022).
14. Dai, J. et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir. Med.* **7**, 881–891 (2019).
15. Byun, J. et al. Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nat. Genet.* **54**, 1167–1177 (2022).
16. Shrine, N. et al. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat. Genet.* **55**, 410–422 (2023).
17. Xu, Q. et al. AI-based analysis of CT images for rapid triage of COVID-19 patients. *NPJ Digit Med.* **4**, 75 (2021).
18. van Griethuysen, J. J. M. et al. Computational Radiomics system to decode the radiographic phenotype. *Cancer Res* **77**, e104–e107 (2017).
19. Maeda, Y., Dave, V. & Whitsett, J. A. Transcriptional control of lung morphogenesis. *Physiol. Rev.* **87**, 219–244 (2007).
20. Shrine, N. et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).
21. Terzikhan, N. et al. Heritability and genome-wide association study of diffusing capacity of the lung. *Eur. Respir. J.* **52**, 1800647 (2018).
22. Zhang, S. et al. Single-cell transcriptome analysis reveals cellular heterogeneity and highlights Fstl1-regulated alveolar myofibroblasts in mouse lung at birth. *Genomics* **115**, 110677 (2023).
23. Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* **10**, 327–334 (2009).
24. Herriges, M. & Morrisey, E. E. Lung development: orchestrating the generation and regeneration of a complex organ. *Development* **141**, 502–513 (2014).
25. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
26. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* **51**, 343–353 (2019).
27. He, P. et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell* **185**, 4841–4860 e25 (2022).
28. Dixit, R., Ai, X. & Fine, A. Derivation of lung mesenchymal lineages from the fetal mesothelium requires hedgehog signaling for mesothelial cell entry. *Development* **140**, 4398–4406 (2013).
29. Rahaghi, F. N. et al. Pulmonary vascular density: comparison of findings on computed tomography imaging with histology. *Eur Respir J* **54**, 1900370 (2019).
30. Turcatel, G. et al. Cartilage rings contribute to the proper embryonic tracheal epithelial differentiation, metabolism, and expression of inflammatory genes. *Am. J. Physiol. Lung Cell Mol. Physiol.* **312**, L196–L207 (2017).
31. Livada, A. C., Pariser, D. N. & Morrell, C. N. Megakaryocytes in the lung: History and future perspectives. *Res. Pr. Thromb. Haemost.* **7**, 100053 (2023).
32. Collaborators, G. B. D. C. R. D. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir. Med.* **8**, 585–596 (2020).
33. Zhu, Z. et al. A large-scale genome-wide association analysis of lung function in the Chinese population identifies novel loci and highlights shared genetic aetiology with obesity. *Eur. Respir. J.* **58**, 2100199 (2021).
34. Walters, R. G. et al. Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.* **3**, 100361 (2023).
35. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
36. Zhu, J. et al. Genetic correlation and bidirectional causal association between Type 2 Diabetes and pulmonary function. *Front. Endocrinol* **12**, 777487 (2021).
37. Ford, E. S., Mannino, D. M. & National, H. Nutrition Examination Survey Epidemiologic Follow-up, S. Prospective association between lung function and the incidence of diabetes: findings from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *Diab. Care* **27**, 2966–2970 (2004).
38. Engstrom, G. et al. Lung function and cardiovascular risk: relationship with inflammation-sensitive plasma proteins. *Circulation* **106**, 2555–2560 (2002).
39. Johansson, M. et al. Serum B vitamin levels and risk of lung cancer. *JAMA* **303**, 2377–2385 (2010).
40. McWilliams, A. et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N. Engl. J. Med.* **369**, 910–919 (2013).
41. Nemec, S. F., Bankier, A. A. & Eisenberg, R. L. Upper lobe-predominant diseases of the lung. *AJR Am. J. Roentgenol.* **200**, W222–W237 (2013).
42. Vincent, M. et al. Clinical, histopathological, and molecular diagnostics in lethal lung developmental disorders. *Am. J. Respir. Crit. Care Med* **200**, 1093–1101 (2019).
43. Danopoulos, S. et al. FGF18 promotes human lung branching morphogenesis through regulating mesenchymal progenitor cells. *Am. J. Physiol. Lung Cell Mol. Physiol.* **324**, L433–L444 (2023).
44. Steimle, J. D. et al. Evolutionarily conserved Tbx5-Wnt2/2b pathway orchestrates cardiopulmonary development. *Proc. Natl. Acad. Sci. USA* **115**, E10615–E10624 (2018).
45. Keijzer, R. et al. The transcription factor GATA6 is essential for branching morphogenesis and epithelial cell differentiation during fetal pulmonary development. *Development* **128**, 503–511 (2001).
46. Kim, E. et al. Isl1 Regulation of Nkx2.1 in the Early Foregut Epithelium is required for Trachea-Esophageal separation and lung lobation. *Dev. Cell* **51**, 675–683 e4 (2019).
47. Cao, S. et al. Single-cell RNA sequencing reveals the developmental program underlying proximal-distal patterning of the human lung at the embryonic stage. *Cell Res* **33**, 421–433 (2023).
48. El-Hashash, A. H. et al. Eya1 controls cell polarity, spindle orientation, cell fate and Notch signaling in distal embryonic lung epithelium. *Development* **138**, 1395–1407 (2011).

49. Liao, C. M. et al. GATA6 suppression enhances lung specification from human pluripotent stem cells. *J. Clin. Invest.* **128**, 2944–2950 (2018).
50. Peng, T. et al. Coordination of heart and lung co-development by a multipotent cardiopulmonary progenitor. *Nature* **500**, 589–592 (2013).
51. Chuang, P. T., Kawcak, T. & McMahon, A. P. Feedback control of mammalian Hedgehog signaling by the Hedgehog-binding protein, Hip1, modulates Fgf signaling during branching morphogenesis of the lung. *Genes Dev.* **17**, 342–347 (2003).
52. Mantovani, G., Puddu, A., Leone, A. L., Tognella, S. & Del Giacco, G. S. Effect of conalbumin on phytomitogen stimulation and E-rosette formation of human peripheral lymphocytes in normal subjects. *Int. J. Tissue React.* **5**, 107–111 (1983).
53. Reynolds, P. R. et al. Receptors for advanced glycation end-products targeting protect against hyperoxia-induced lung injury in mice. *Am. J. Respir. Cell Mol. Biol.* **42**, 545–551 (2010).
54. Lao, T. et al. Hhip haploinsufficiency sensitizes mice to age-related emphysema. *Proc. Natl. Acad. Sci. USA* **113**, E4681–E4687 (2016).
55. Werder, R. B., Zhou, X., Cho, M. H. & Wilson, A. A. Breathing new life into the study of COPD with genes identified from genome-wide association studies. *Eur. Respir. Rev.* **33**, 240019 (2024).
56. Southgate, L., Machado, R. D., Graf, S. & Morrell, N. W. Molecular genetic framework underlying pulmonary arterial hypertension. *Nat. Rev. Cardiol.* **17**, 85–95 (2020).
57. Bein, K. et al. Genetic determinants of ammonia-induced acute lung injury in mice. *Am. J. Physiol. Lung Cell Mol. Physiol.* **320**, L41–L62 (2021).
58. Luo, H. et al. The regulation of circRNA_kif26b on alveolar epithelial cell senescence via miR-346-3p is involved in microplastics-induced lung injuries. *Sci. Total Environ.* **882**, 163512 (2023).
59. Van Loy, T. et al. Stimulation of the atypical chemokine receptor 3 (ACKR3) by a small-molecule agonist attenuates fibrosis in a pre-clinical liver but not lung injury model. *Cell Mol. Life Sci.* **79**, 293 (2022).
60. Leikauf, G. D. et al. Functional genomic assessment of phosgene-induced acute lung injury in mice. *Am. J. Respir. Cell Mol. Biol.* **49**, 368–383 (2013).
61. Koppelman, G. H. & Sayers, I. Evidence of a genetic contribution to lung function decline in asthma. *J. Allergy Clin. Immunol.* **128**, 479–484 (2011).
62. Namkoong, H. et al. DOCK2 is involved in the host genetics and biology of severe COVID-19. *Nature* **609**, 754–760 (2022).
63. D'Amico, S. et al. ERAP1 and ERAP2 Enzymes: A Protective Shield for RAS against COVID-19? *Int. J. Mol. Sci.* **22**, 1705 (2021).
64. Wang, C. et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. *Lancet* **391**, 1706–1717 (2018).
65. Ni, Q. et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. *Eur. Radio.* **30**, 6517–6527 (2020).
66. Huang, S. et al. Distribution Atlas of COVID-19 Pneumonia on Computed Tomography: A Deep Learning Based Description. *Phenomics* **1**, 62–72 (2021).
67. Hu, Z. J. et al. Lower circulating Interferon-Gamma is a risk factor for lung fibrosis in COVID-19 patients. *Front Immunol.* **11**, 585647 (2020).
68. Zwanenburg, A. et al. The Image Biomarker Standardization Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
69. McKay, J. D. et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
70. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
71. Wang, C. et al. Analyses of rare predisposing variants of lung cancer in 6004 whole genomes in Chinese. *Cancer Cell* **40**, 1223–1239 e6 (2022).
72. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
73. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet* **50**, 906–908 (2018).
74. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
75. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
76. Boughton, A. P. et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
77. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
78. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
79. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
80. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
81. Oscanoa, J. et al. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **48**, W185–W192 (2020).
82. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
83. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
84. Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
85. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
86. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
87. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
88. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
89. Ji, M. Common Genetic Variation Influencing the Human Lung Imaging Phenotypes. Zenodo. https://doi.org/10.5281/zenodo.17071649 (2025).

Acknowledgements

We are grateful to participants and researchers of the China Kadoorie Biobank (CKB) and the Biobank Japan (BBJ). This study was supported by the National Natural Science Foundation of China (82388102) for H.S; Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0524000; 2024ZD0520000; 2024ZD0520003), National Natural Science Foundation of China (82273724), Major Project of Changzhou Medical Center in Nanjing Medical University (CMCM202210) for H.M; the National Natural Science Foundation of China (82473708, 82273714), the Excellent Youth Foundation of Jiangsu

Province (BK20220100) for M.Z; the National Natural Science Foundation of China (82273700) for L.D.

Author contributions

M.Z. (Meng Zhu), L.D. and L.S. designed the study and wrote the manuscript. M.Z. (Meng Zhu) and C.J. (Chen Ji) analyzed data and drafted the manuscript. C.Z. (Chen Zhu) and C.S. contributed to data acquisition. H.M., X.C., and H.S. supervised this work. L.W., L.Z., J.L., Q.Z., F.W., C.J. (Chen Jin), Y.M., Q.L., J.Z., M.Z. (Mingxuan Zhu), J.C., C.Z. (Caochen Zhang), Y.F., L.G., D.H., J.D., Y.J., G.J., and Z.H. contributed to data interpretation. All authors critically reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64571-z.

Correspondence and requests for materials should be addressed to Hongxia Ma, Xiangdong Cheng or Hongbing Shen.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.