Journal of Translational Medicine

RESEARCH

Open Access

Genetic insights into idiopathic pulmonary fibrosis: a multi-omics approach to identify potential therapeutic targets

Zhuofeng Wen[1†], Weixuan Liang[2†], Ziyang Yang[3†], Junjie Liu[4†], Jing Yang[1], Runge Xu[2], Keye Lin[2], Jia Pan[2] and Zisheng Chen[1*]

Abstract

Objective To identify potential therapeutic targets and evaluate the safety profiles for Idiopathic Pulmonary Fibrosis (IPF) using a comprehensive multi-omics approach.

Method We integrated genomic and transcriptomic data to identify therapeutic targets for IPF. First, we conducted a transcriptome-wide association study (TWAS) using the Omnibus Transcriptome Test using Expression Reference Summary data (OTTERS) framework, combining plasma expression quantitative trait loci (eQTL) data with IPF Genome-Wide Association Studies (GWAS) summary statistics from the Global Biobank (discovery) and Finngen (duplication). We then applied Mendelian randomization (MR) to explore causal relationships. RNA-seq co-expression analysis (bulk, single-cell and spatial transcriptomics) was used to identify critical genes, followed by molecular docking to evaluate their druggability. Finally, phenome-wide MR (PheW-MR) using GWAS data from 679 diseases in the UK Biobank assessed the potential adverse effects of the identified genes.

Result We identified 696 genes associated with IPF in the discovery dataset and 986 genes in the duplication dataset, with 126 overlapping genes through TWAS. MR analysis revealed 29 causal genes in the discovery dataset, with 13 linked to increased and 16 to decreased IPF risk. Summary data-based MR (SMR) confirmed six essential genes: ANO9, BRCA1, CCDC200, EZH1, FAM13A, and SFR1. Bulk RNA-seq showed FAM13A upregulation and SFR1 and EZH1 downregulation in IPF. Single-cell RNA-seq revealed gene expression changes across cell types. Molecular docking identified binding solid affinities for essential genes with respiratory drugs, and PheW-MR highlighted potential side effects.

Conclusion We identified six key genes—ANO9, BRCA1, CCDC200, EZH1, FAM13A, and SFR1—as potential drug targets for IPF. Molecular docking revealed strong drug affinities, while PheW-MR analysis highlighted therapeutic potential and associated risks. These findings offer new insights for IPF treatment and further investigation of potential side effects.

Keywords Idiopathic pulmonary fibrosis, Multi-omics, Therapeutic targets, Genetic insights, Druggability

†Zhuofeng Wen, Weixuan Liang, Ziyang Yang and Junjie Liu contributed equally as co-first authors of this manuscript.

*Correspondence:
Zisheng Chen
502463784@qq.com
Full list of author information is available at the end of the article

Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic interstitial lung disease (ILD) characterized by scarring of lung tissue, which progressively impairs lung function, eventually leading to respiratory failure and death [1]. Epidemiological data indicate that the incidence of IPF ranges from 0.09 to 1.30 per 10,000 people, with prevalence between 0.33 and 4.51 per 10,000 people. As the global population ages, the incidence of IPF is expected to increase further [2]. Notably, IPF has a high mortality rate, with most patients surviving only 3 to 5 years after diagnosis [3]. This short survival time places a significant burden not only on patients but also on their families. Although Food and Drug Administration (FDA)-approved antifibrotic drugs like pirfenidone and nintedanib can slow the progression of IPF, they are not curative [4]. Moreover, the complex nature of IPF leads to considerable variability in how patients respond to treatment [5]. Therefore, there is an urgent need to identify new therapeutic targets and develop combination treatment strategies to improve patient outcomes and quality of life more effectively.

Randomized controlled trial (RCT) is the gold standard for IPF drug research, but patient heterogeneity, high costs, and their focus on single-target therapies limit their ability to address the complex, multifactorial nature of the disease, slowing drug development [4, 6, 7]. In contrast, omics-based studies have proven to be a powerful approach for identifying drug targets, with multi-omics integration providing robust evidence for target discovery and validation [8]. Transcriptome-wide association Studies (TWAS) help identify genes associated with complex traits by linking genetic regulation of gene expression to disease mechanisms, facilitating a better understanding of IPF-related genes[9, 10]. Genome-wide association Studies (GWAS) also enable researchers to identify gene variants and integrate data across multiple domains [11]. Furthermore, molecular docking supports drug target research by offering insights into the druggability of identified genes. At the same time, single-cell and bulk-sequencing analyses provide detailed information on gene expression and tissue localization, enhancing the precision of target identification [12, 13].

Therefore, our study integrates GWAS data from the Global Biobank Meta-Analysis Initiative (GBMI) and FinnGen Consortium and plasma expression quantitative trait loci (eQTL) data from the eQTLGen Consortium. Using TWAS within the Omnibus Transcriptome Test using the Expression Reference Summary data (OTTERS) framework, we identify genes associated with IPF. To validate these findings, we apply primary Mendelian Randomization (MR) and summary data-based MR (SMR) for causal inference, pinpointing genes causally linked to IPF. We then use scRNA-seq, bulk RNA-seq and spatial transcriptomics to examine differential gene expression between IPF patients and healthy controls across cell types. Finally, molecular docking evaluates the druggability of identified genes, and phenome-wide MR (PheW-MR) using UK Biobank data assesses associations across 679 common diseases. Figure 1 provides a detailed overview of the workflow employed in our study.

Methods

Data source

eQTL data

Our research derived the cis-eQTL (within ± 1 megabase of gene transcription start sites) summary-level data for 16,699 genes from eQTLGen Consortium for subsequent analyses, which features a meta-analysis of 37 cohorts with a total of 31,684 samples [14]. The dataset primarily consisted of blood samples (25,482 samples, accounting for 80.4%) and peripheral blood mononuclear cell samples (6202 samples, accounting for 19.6%). Detailed information can be found in Supplementary Table S1.

GWAS summary statistics for IPF

Our study included case–control GWAS summary statistics for IPF from the GBMI[15] as the discovery dataset, comprising 8006 cases and 1,246,742 controls. To enhance the robustness of our findings, we incorporated IPF GWAS data from the FinnGen consortium as a validation dataset, totaling 2401 cases and 448,636 controls [16]. For further details, please refer to Supplementary Table S1.

RNA-seq and spatial transform data of IPF

To investigate the mechanisms of genes associated with IPF, we collected two single-cell datasets from the Gene Expression Omnibus (GEO) database: GSE136831 [17] and GSE135893 [18]. These datasets include samples from 32 IPF patients and 28 control samples. We also selected the peripheral blood mononuclear cell (PBMC) dataset GSE28042 [19] from the GEO database, comprising 75 IPF samples and 19 control samples. Moreover, we selected the spatial transform data of IPF (GSM8087036, GSM8087037 and GSM8087038) and control (GSM8087031, GSM8087033, GSM8087035) from 10 × Visium Cytassist platform in dataset GSE248082 [20]. For more details, please refer to Supplementary Table S1.

TWAS

In this study, we employed a TWAS framework utilizing OTTERS to integrate eQTL intersections specifically for analysis with IPF GWAS summary statistics [21]. OTTERS leverages summary-level eQTL

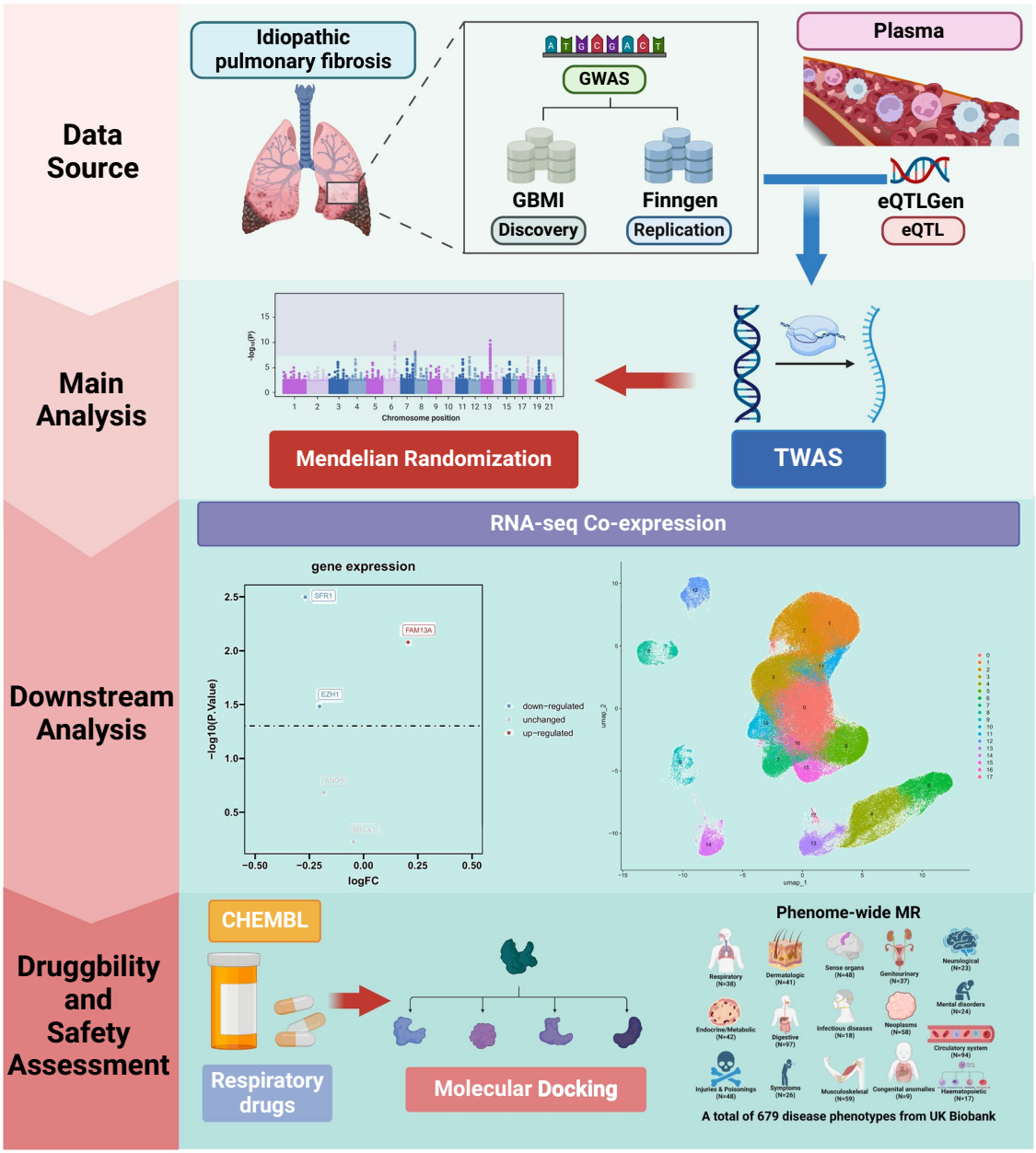

Fig. 1 Schematic representation of the study workflow (created with Biorender.com)

reference data to calculate eQTL weights and implement four advanced summary-data-based polygenic risk score (PRS) methods: P + T (p-value threshold with linkage disequilibrium (LD) clumping) [22], lassosum (a frequentist LASSO regression-based method) [23], SDPR (a nonparametric Bayesian Dirichlet Process Regression model-based method) [24], and PRS-CS (a Bayesian multivariable regression model with continuous shrinkage (CS) priors) [25].

OTTERS operates in two stages. The initial phase estimates cis-eQTL effect sizes using multivariate regression modeling on summary-level reference data, which is assumed to provide insights into the expression of individual genetic variants and genes. Univariate regression models are used to estimate effect sizes and p-values, after which each PRS method utilizes cis-eQTL summary data and an external LD reference panel from a similar ancestry to determine cis-eQTL weights. Specifically, this study employs plasma-derived cis-eQTL weights

previously established in the original OTTERS investigation[21]. Once the weights are established, each method interpolates gene expression (GReX) for each gene, enabling gene-based association analysis in the test GWAS dataset during the second stage. Utilizing the publicly available OTTERS python codework[21], we implemented parameter configurations (models = P0.001, P0.05, lassosum, SDPR, PRScs) to execute this second-stage analysis. This phase culminates in generating TWAS p-values using the ACAT-O method, which adopts a Cauchy distribution for inference, ultimately producing the final OTTERS *P-value*.

MR

MR is employed to discern causality between an exposure (e.g., gene) and an outcome (e.g., IPF)[26]. Cis-eQTLs significantly associated with IPF, obtained from the aforementioned TWAS, were used to select genetic instruments(IVs). IVs were selected by removing those in linkage disequilibrium (LD), with an R^2 threshold of < 0.001 and a clumping distance of 10,000 kb. To ensure selecting robust results, weak IVs were excluded based on F-statistics, with IV having an F-statistic < 10 being removed.

$$F = \frac{R^2(N-2)}{(1-R^2)}$$

where N and R^2 are the sample size and the variance explained by IVs, respectively.

MR analysis was performed using the "TwoSampleMR" package [27]. For any gene with a single instrument, the Wald ratio method was used to estimate the change in log odds of IPF risk for each standard deviation (SD) increase in plasma gene levels represented by the instrument. The inverse variance weighting (IVW) method was used for genes with multiple instruments to obtain MR effect estimates. Heterogeneity tests based on Q statistics were conducted to assess the heterogeneity of genetic instruments. Additional analyses, including simple mode, weighted mode, weighted median, and MR-Egger, were performed to account for horizontal pleiotropy [27]. MR-Egger results were only reported when the intercept indicated the presence of horizontal pleiotropy. A further MR analysis was conducted based on GWAS summary data from GBMI and FinnGen to replicate the identified genes. A *P-value* < 0.05 was defined as the threshold for statistical significance in duplication.

Additionally, SMR analysis was performed as a complementary method to validate the causal relationship between the identified genes and disease [28]. Multiple single nucleotide polymorphisms (SNPs) within the region were used for the Heterogeneity in Dependent Instruments (HEIDI) test to distinguish proteins associated with disease risk due to shared genetic variation rather than genetic linkage [28]. SMR and HEIDI tests were conducted using SMR software (SMR v1.03). A *P-value* < 0.05 was defined as the significance level for SMR analysis. A *P-value* > 0.01 from the HEIDI test indicated that linkage disequilibrium did not drive the association between the gene and IPF [29].

Bulk and single-cell RNA expression analysis

We first utilized the GSE28042 dataset from the GEO database for bulk transcriptomic analysis, comprising 75 IPF samples and 19 control samples, to investigate the mechanisms of crucial gene action. After importing the data using the "GEOquery" R package, we extracted the expression matrices for six genes: BRCA1, EZH1, FAM13A, SFR1, ANO9, and CCDC200. To correct for non-biological differences between samples, we used the "normalizeBetweenArrays" function from the "limma" package [30]. We then employed the "FactoMineR" packages to eliminate batch effects and perform PCA analysis [30, 31]. Subsequently, we conducted differential expression analysis for the selected genes between the IPF and control samples using the "lmFit" function,, based on a linear model [30]. To ensure the reliability of our findings, we applied the "eBayes" function for Bayesian testing to evaluate p-values and calculate the average Log2 fold change (LogFC). Additionally, we created volcano plots for enhanced result visualization, identifying genes with *P-value* greater than 0.05 as having no significant expression difference between the groups. In contrast, genes with *P*-values less than 0.05 and LogFC > 0 were classified as up-regulated in IPF, while those with LogFC < 0 were classified as down-regulated.

Following this, we expanded our analysis by utilizing the GSE136831 dataset from GEO, which includes 32 IPF samples and 28 control samples [32]. From this dataset, we extracted exonic data from the IPF and control samples for single-cell RNA analysis. Initially, we employed the "Seurat" R package for data entry and quality control. Cells were filtered based on specific feature thresholds, including a range of 300 to 4000 unique genes detected per cell, less than 20% mitochondrial gene expression, less than 3% hemoglobin gene expression, a total RNA count of less than 30,000, and less than 1% platelet gene expression. After ensuring the dataset's integrity, we used SCTransform for normalization [33], then performed multi-sample integration by finding anchors between samples. We used the JackStraw method to identify the most suitable dimensions for downstream analysis. Doublets cells were removed by estimating a 7.5% doublet probability, and environmental RNA contamination was eliminated using the decontX R package [34–36]. For cell

type annotation, we utilized the GPT-4o model from the "GPTCelltype" R package [37]. Finally, we analyzed differential gene expression using the "FindMarkers" function, with a minimum cell threshold set to 3. To evaluate the robustness of our findings, we performed Wilcoxon tests to calculate p-values and determine the average LogFC. We also created heatmaps for more precise visualization of our results. Genes with *P-values* less than 0.05 for each cell type were considered significantly differentially expressed.

Simulated knockout profile of key genes

To further investigate the potential impact of key genes on the pathogenesis of IPF, we conducted a simulated knockout analysis. scTenifoldKnk [38] is a method used for virtual knockout experiments to predict gene functions. First, scTenifoldKnk constructs a denoised single-cell gene regulatory network (scGRN) based on scRNA-seq data. The scGRN is then replicated, and the out-edge values of the target gene in the adjacency matrix of the replicated scGRN are set to zero, generating a pseudo-knockout scGRN. With two scGRNs—one representing the initial network and the other as the pseudo-knockout network—the regulatory significance of the target gene can be assessed through various comparison programs. The two scGRNs are mapped into the same low-dimensional space, and the distance between gene projections reveals the impact of gene knockout on the scGRN: larger disturbances in the low-dimensional space indicate greater importance of the target gene within the scGRN.

We used scRNA-seq data from the GSE136831 dataset, representing IPF, in scTenifoldKnk, and extracted expression matrices for six key genes (ANO9, BRCA1, CCDC200, EZH1, FAM13A, and SFR1) to perform the pseudo-knockout analysis. The list of disturbed genes was ranked based on the fold change in the distance between the gene projections of the two scGRNs, with p-values assigned using a chi-square distribution with one degree of freedom. The most disturbed genes should exhibit close connections with the target gene. Subsequently, we performed Gene Ontology (GO) pathway enrichment analysis on the top 10 genes significantly affected by the knockout of the six key genes.

Spatial transcriptomics

In the analysis of spatial transcriptomics, we utilized Scanpy [39] within a Python 3.10 environment to process data from the 10×Visium CytAssist platform, including samples from both IPF and control groups. Quality control steps were implemented with the following thresholds: mitochondrial gene percentage < 20%, hemoglobin gene expression < 3%, total RNA count < 30,000, and platelet gene expression < 1%. Data normalization was performed using the normalize_total function with the max_fraction parameter set to 0.05. Highly variable genes were identified for each sample using the highly_variable_genes function (flavor = "seurat", n_top_genes = 2000). Clustering was conducted via the Leiden algorithm with parameters resolution = 0.8 and n_iterations = 10. Cell type annotation was performed using the ScType Python package, employing a deconvolution algorithm with a lung tissue-specific reference model, followed by spatial visualization[40]. For gene exploration, expression profiles and annotated cell types were extracted and stratified by IPF and control groups. Violin plots were generated to visualize gene expression patterns across distinct cell types.

Molecular docking

To validate the potential roles of the top genes in current respiratory disease drugs, molecular docking was applied to construct medication–gene–disease pathways, guided by most corresponding literature and previous analyses. This method assesses the feasibility and binding effectiveness of drug interactions. Molecular docking is a widely utilized approach for predicting protein–ligand interactions. We predicted the binding free energy and inhibitory potency of approved respiratory disease drugs on proteins encoded by the essential genes. Protein receptors were sourced from the Protein Data Bank (PDB)[41] (https://www.rcsb.org/) using GetPDB software, while molecular ligands were obtained from ChEMBL[42]. In cases where the receptor's protein files were unavailable in the PDB, the predicted 3D structure from AlphaFold was utilized instead. Molecular docking was performed using Dockey software[43] and QuickVina-W[44]. We used OpenBabel[45] to convert unsupported formats to PDB format and extract molecular base information, including the number of atoms, the number of rotatable bonds, molecular weight, and the calculated octanol–water partition coefficient (logP).Binding-free energies ≤ − 7.5 kcal/mol indicated binding solid affinity, while those > − 7.5 and ≤ − 5 kcal/mol indicated moderate affinity and values > − 5 kcal/mol were deemed to suggest very weak or negligible binding affinity. Visualization was conducted using PyMOL[46] (https://www.pymol.org/).

Dockey provides calculations of various metrics to help users select and optimize candidate lead compounds in virtual screening. Ligand efficiency (LE) is the initially proposed and widely used metric for evaluating the quality of the interaction between a ligand and a receptor[47]. LE is calculated according to the following equation[48]:

$$LE = -\Delta G/N$$

where ΔG represents the binding free energy, and N denotes the number of heavy atoms (non-hydrogen

atoms) in the ligand. Subsequently, size-independent ligand efficiency (SILE) and fit quality (FQ) were introduced to overcome the size dependence of LE. SILE evolved from LE and is calculated as follows[49]:

$$SILE = -\Delta G/N^{0}.3$$

FQ is scaled LE and can be estimated using the following equation[50]:

$$FQ = LE/(0.0715 + 7.5328/N + 25.7079/N^{2} - 361.4722/N^{3})$$

In addition to molecular size, lipophilicity is another important factor in drug discovery. Lipophilic ligand efficiency (LLE) allow us to assess lipophilicity. LLE can be derived from the following equation[51]:

$$LLE = -logKi - logP$$

Ki is the estimated inhibition constant and *logP* represents the computed octanol–water partition coefficient.

Analysis of drug side effects on 679 disease traits

PheW-MR analysis was utilized to investigate the potential side effects of drug targets associated with IPF-related phenotypes. We first established a connection between specific proteins and IPF-related phenotypes, then assessed these proteins' implications on various diseases to uncover any unintended effects of drug intervention. Data regarding the SNPs of these drug targets were derived from eQTL studies and aligned with the datasets used in the TWAS analysis. A total of 679 traits were selected, each with over 500 cases, utilizing GWAS summary statistics from the extensive UK Biobank cohort (N ≤ 408,961) as provided by Zhou et al. (Supplementary Table S2) [52]. Diseases were classified using PheCodes, a system that categorizes international classification of diseases codes into phenotypic outcomes, facilitating a comprehensive genetic analysis of various disease traits. A side effect is defined as the impact on other diseases when a drug target reduces the risk of the primary disease by 10%. To estimate side effects, we employed a formula that accounts for the interactions between the drug target and various diseases.

$$\beta_{\text{effect}} = \frac{\beta_{679\ \text{diseases}}}{\beta_{\text{IPF phenotypes}}} \times \ln(0.9)$$

where β_{679} diseases and β_{IPF} phenotypes were from the PheW-MR and the aforementioned SMR discovery analysis of proteins related to IPF phenotypes, respectively. Additionally, the study calculated odds ratios (OR) per SD increase in protein levels. Proteins with an OR value greater than one were considered to have potentially harmful effects on the disease. The standard error (SE) was estimated using the bootstrap method.

Result

TWAS

We conducted TWAS analyses on the discovery dataset (GBMI) and the duplication dataset (FinnGen Consortium). For the discovery dataset, we identified 696 genes associated with IPF ($P < 0.05$) (Supplementary Table S3 and Fig. 2A). The top five genes most significantly associated with IPF were FKBPL ($P = 4.08 \times 10\text{^}{-16}$), VARS2 ($P = 5.19 \times 10\text{^}{-14}$), PFDN6 ($P = 2.88 \times 10\text{^}{-13}$), HLA-DOB ($P = 5.72 \times 10\text{^}{-13}$), and HLA-C ($P = 1.79 \times 10\text{^}{-12}$). Similarly, in the duplication dataset, we found 986 genes associated with IPF ($P < 0.05$) (Supplementary Table S4 and Fig. 2B). The top five genes most significantly associated with IPF were: KMT5A ($P = 3.61 \times 10\text{^}{-15}$), CD151 ($P = 5.57 \times 10\text{^}{-10}$), KRTAP5-1 ($P = 4.06 \times 10\text{^}{-09}$), MYNN ($P = 4.49 \times 10\text{^}{-09}$), and NEIL2 ($P = 6.41 \times 10\text{^}{-09}$). By intersecting the TWAS results from both datasets, we identified 126 overlapping genes (Fig. 2C).

MR, SMR and HEIDI

To further strengthen the causal relationship between the identified genes and IPF, we performed MR analysis on the TWAS results from both datasets. We identified 29 genes with a causal relationship to IPF for the discovery dataset. Among them, 13 genes were associated with an increased risk of IPF, including ARL17A ($P = 1.32 \times 10\text{^}{-55}$, OR = 1.11), ZNF675 ($P = 2.27 \times 10\text{^}{-09}$, OR = 1.61), and FAM13A ($P = 1.92 \times 10\text{^}{-06}$, OR = 1.33), while 16 genes were linked to a reduced risk of IPF, with BET1L ($P < 1.32 \times 10\text{^}{-55}$, OR = 0.86), GSTO1 ($P = 2.81 \times 10\text{^}{-53}$, OR = 0.86), and IL27RA ($P = 3.36 \times 10\text{^}{-17}$, OR = 0.89) among the most significant. For the duplication dataset, we found 31 genes with a causal relationship to IPF. Notably, GSTO1 ($P = 1.13 \times 10\text{^}{-83}$, OR = 1.07) and TSPAN4 ($P = 3.30 \times 10\text{^}{-10}$, OR = 1.62) were among 17 genes associated with a decreased risk of IPF, whereas 16 genes, including HRAS ($P = 2.88 \times 10\text{^}{-07}$, OR = 0.56), LRRC37A2 ($P = 3.04 \times 10\text{^}{-07}$, OR = 0.77), and PTDSS2 ($P = 1.46 \times 10\text{^}{-05}$, OR = 0.77), were associated with an increased risk of Infertile results of the MR analysis can be found in Supplementary Table S5.

We performed supplementary validation using SMR and HEIDI analyses to strengthen the credibility of our MR findings. In the discovery dataset, 21 genes were confirmed to have causal relationships, all initially identified through TWAS analysis (detailed in Supplementary Table S6). In the duplication dataset, 15 genes were validated as having causal associations with IPF

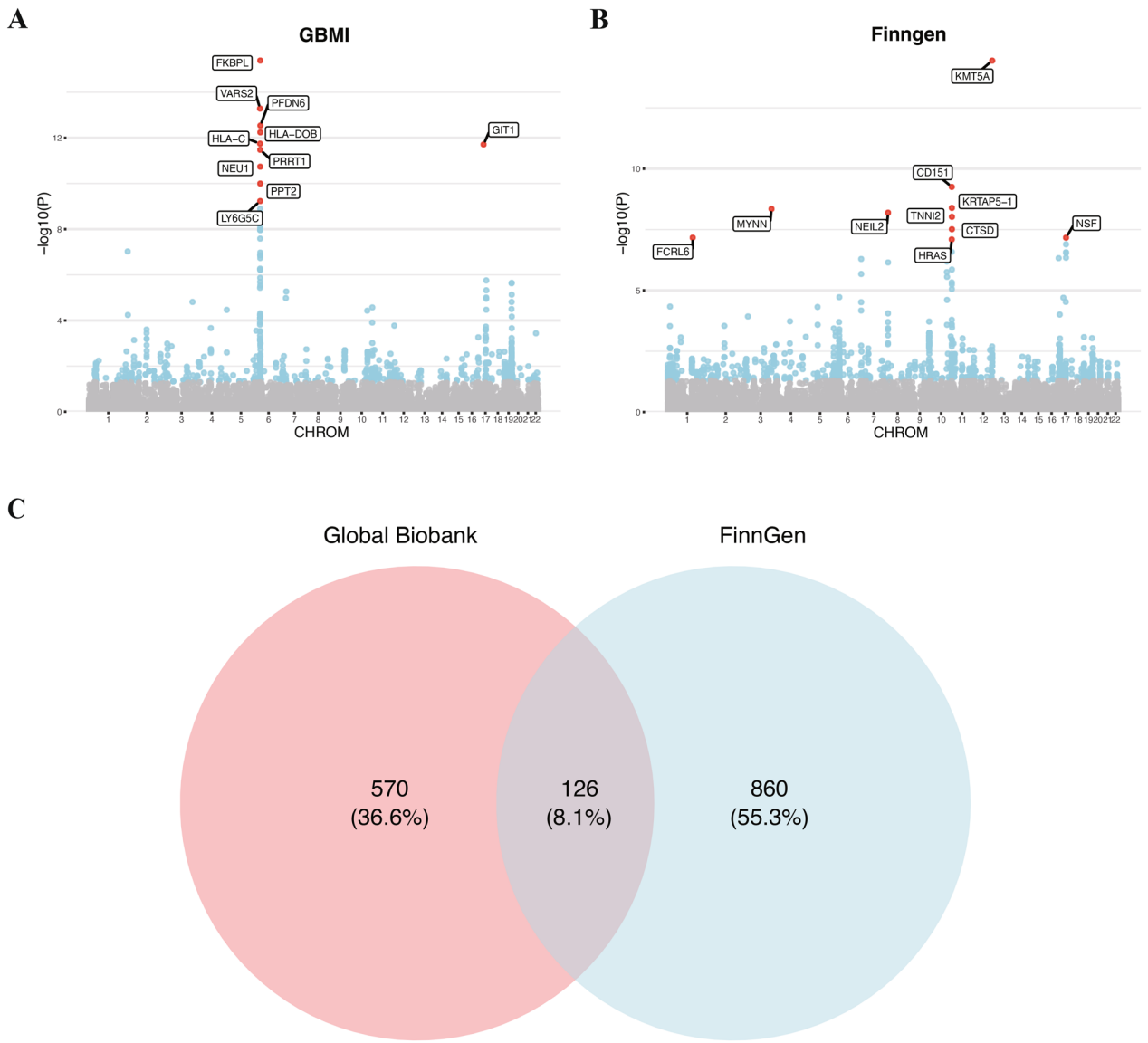

Fig. 2 Manhattan plots and Venn diagram showing TWAS results for significant gene associations. **A** Manhattan plot of TWAS results from the Global Biobank cohort. Grey points represent genes with P-values > 0.05, blue points indicate genes with P-values ≤ 0.05, and the top 10 most significant genes are highlighted in red. **B** Manhattan plot of TWAS results from the FinnGen. Grey points represent genes with P-values > 0.05, blue points indicate genes with P-values ≤ 0.05, and the top 10 most significant genes are marked in red. **C** Venn diagram showing the overlap of significant genes between the Global Biobank and FinnGen cohorts. The red circle represents the significant genes identified from the Global Biobank cohort, while the blue circle represents the genes identified from FinnGen

(Supplementary Table S7). To derive the most robust conclusions, we focused on genes validated by both SMR and MR analyses, which we considered the final results of our study (Fig. 3). This integrative approach identified six significant genes: ANO9, BRCA1, CCDC200, EZH1, FAM13A, and SFR1. Detailed statistical results for these six genes are presented in Table 1 and visualized in Fig. 4.

Bulk RNA-seq and enrichment analysis

To investigate the overall differential expression of key genes in IPF and their associated enriched pathways, we conducted differential gene expression and key gene pathway enrichment analysis using the GSE28042 dataset. We found that in IPF, FAM13A was upregulated (LogFC = 0.21, $P = 8.34 \times 10\text{^}-03$), while SFR1 (LogFC = −0.27, $P = 3.18 \times 10\text{^}-03$) and EZH1 LogFC = −0.20, $P = 3.29 \times 10\text{^}-02$) were downregulated.

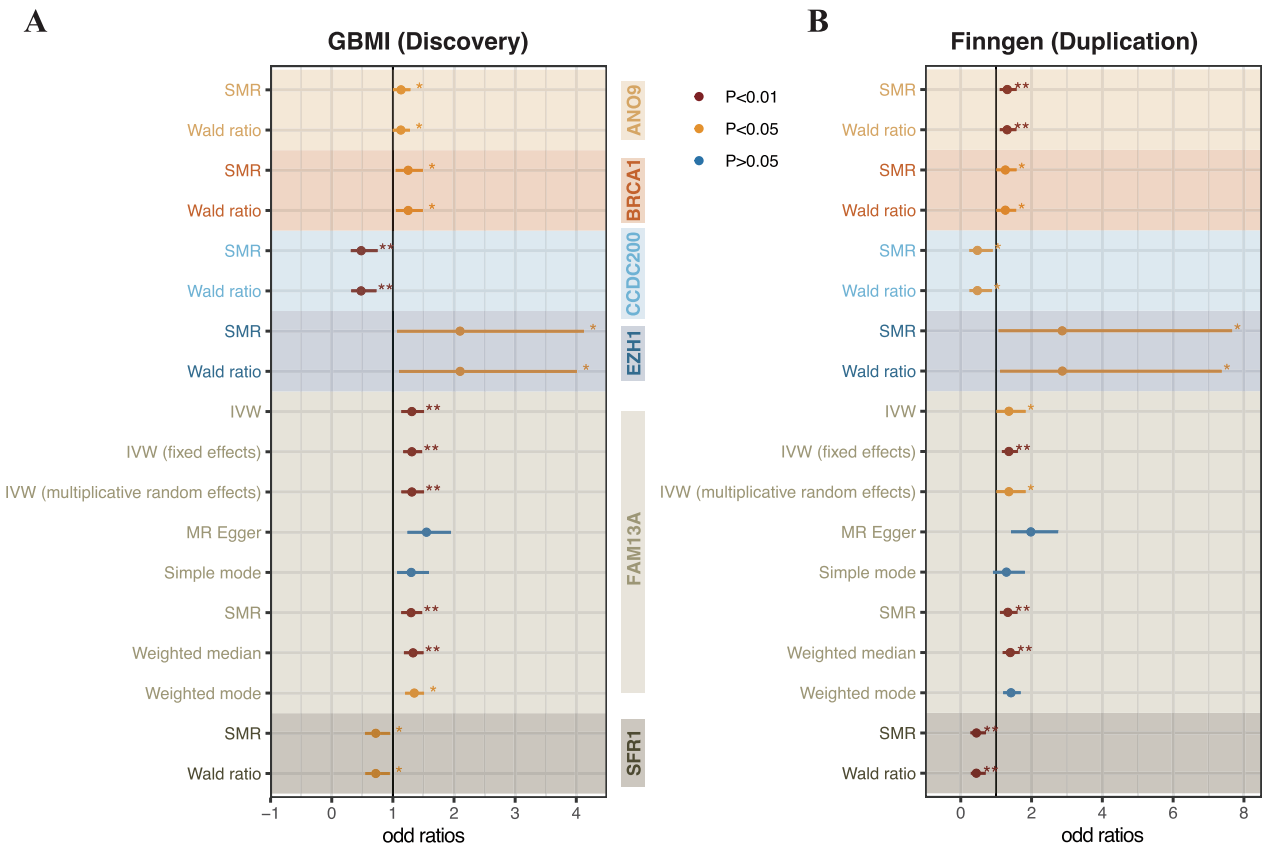

Fig. 3 The results of crucial gene analysis are shown across two datasets. **A** Forest plot displaying the MR and SMR analysis results for critical genes in the GBMI (discovery dataset). **B** Forest plot displaying the results of MR and SMR analysis for critical genes in the FinnGen (duplication dataset)

However, no significant changes were observed in the expression of ANO9 and BRCA1 (Fig. 5A and Supplementary Table S8). Additionally, enrichment analysis revealed that BRCA1 and EZH1 were enriched in pathways related to the negative regulation of gene expression and epigenetic and epigenetic regulation of gene expression. SFR1 was associated with pathways such as double-strand break repair via homologous recombination and recombinational repair, while ANO9 showed expression in the phospholipid scramblase activity pathway (Fig. 5B and Supplementary Table S9).

scRNA-seq

To systematically reveal the distribution of the six key genes in lung cells, we performed scRNA-seq analysis on 32 IPF samples and 28 normal control samples from the GSE136831 dataset. After quality control, we grouped 198,359 cells into 18 clusters (Fig. 6A) and used the ChatGPT-4o model to annotate 13 cell types: Macrophage, Langerhans Cell, Neutrophil, Macrophage (M2 Type), Dendritic Cell Progenitor, Eosinophil, Epithelial Cell, Cytotoxic T Cell, Ciliated Epithelial Cell, and Vascular Endothelial Cell (Fig. 6B). We then performed differential gene expression analysis on the six key genes across different cell types in both IPF and normal controls. Of the 78 gene-cell pairs analyzed, 30 showed significant results. Among these, 22 gene-cell pairs were upregulated, while eight were downregulated (Fig. 6C). The most significantly upregulated gene in IPF was CCDC200, with elevated expression in Cytotoxic T Cells (LogFC = 0.65, $P = 1.16 \times 10^{-24}$), Macrophages (LogFC = 0.84, $P = 7.36 \times 10^{-248}$), and Eosinophils (LogFC = 0.54, $P = 2.45 \times 10^{-28}$). Additionally, BRCA1 in Macrophages (LogFC = 0.29, $P = 1.29 \times 10^{-17}$) and SFR1 in Macrophages (LogFC = 0.33, $P = 2.98 \times 10^{-16}$) also showed notable upregulation. For the downregulated genes, CCDC200 again showed the most significant results, with decreased expression in Dendritic Cell Progenitors (LogFC = −1.02, $P = 2.65 \times 10^{-35}$) and Macrophages (M2 Type) (LogFC = −0.65, $P = 1.82 \times 10^{-20}$). The downregulation of EZH1 was also notable in Neutrophils (LogFC = −0.42, $P = 8.67 \times 10^{-05}$) and Macrophages (LogFC = −0.38, $P = 1.64 \times 10^{-06}$). Interestingly, no significant results were found in B Cells, Effector B Cells, or Fibroblasts. Detailed data can be found in **Supplementary Table S10.**

Table 1 Summary of six critical genes identified in the study

Symbol	P_TWAS_ Discovery	P_TWAS_ Duplication	P_MR_Discovery	OR_MR_ Discovery	P_MR_ Duplication	OR_MR_ Duplication	P_SMR_ Discovery	OR_SMR_ Discovery	P_SMR_ Duplication	OR_SMR_ Duplication
BRCA1	2.44E−04	1.58E−02	1.28E−02	1.25E + 00	4.04E−02	1.26E + 00	1.31E−02	1.25E + 00	4.09E−02	1.26E + 00
EZH1	5.87E−04	4.44E−02	2.48E−02	2.10E + 00	2.90E−02	2.87E + 00	3.15E−02	2.10E + 00	3.60E−02	2.87E + 00
FAM13A	2.10E−03	1.40E−02	1.39E−04	1.31E + 00	3.97E−02	1.36E + 00	1.03E−04	1.30E + 00	2.37E−03	1.33E + 00
SFR1	2.68E−03	4.50E−07	2.01E−02	7.20E−01	4.59E−04	4.44E−01	2.19E−02	7.20E−01	6.92E−04	4.44E−01
ANO9	3.73E−02	2.60E−07	4.93E−02	1.13E + 00	2.91E−03	1.31E + 00	4.98E−02	1.13E + 00	3.05E−03	1.31E + 00
CCDC200	2.43E−04	1.87E−02	5.35E−04	4.82E−01	1.88E−02	4.70E−01	1.17E−03	4.82E−01	2.27E−02	4.70E−01

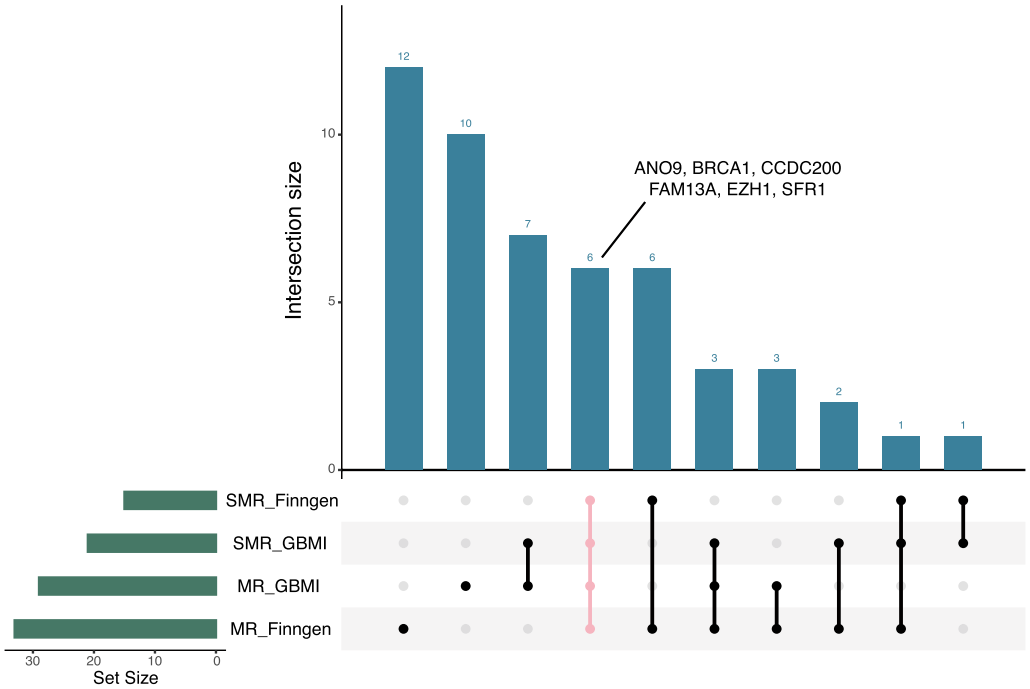

Fig. 4 Upset plot of candidate genes tested by different MR. The horizontal bar on the left represents several candidate genes obtained from different datasets and MR methods. Dots and lines represent subsets of genes. Vertical histogram represents number of genes in each subset. Genes tested by both MR and SMR were marked pink

Simulated knockout profile of key genes

We simulated the knockout of key genes in IPF using the scTenifoldKnk method. The final scTenifoldKnk analysis identified 3 genes affected by ANO9, 284 genes affected by BRCA1, 205 genes affected by CCDC200, 44 genes affected by EZH1, 286 genes affected by FAM13A, and 179 genes affected by SFR1, all with FDR < 0.05 (Supplementary Table S11). Pathway enrichment analysis revealed that the knockout of ANO9 primarily impacted immune cell migration functions, including pathways related to monocyte, dendritic cell, and leukocyte migration. This may disrupt the initiation and cellular localization of immune responses. The BRCA1 knockout predominantly affected cytoskeletal and ciliary-related pathways, playing a critical role in maintaining axonemal structure and the stability of cellular protrusions. The knockout of CCDC200 influenced viral invasion and host–pathogen interactions, suggesting that this gene may contribute to immune defense by regulating the viral lifecycle and host–pathogen interactions. EZH1 knockout significantly affected lamellipodium formation and the organization of adhesion complexes during cell migration, indicating its important role in cell polarity and motility. FAM13A and SFR1 knockouts affected host-symbiont interactions and viral invasion regulation, potentially modulating host immune responses to resist external invasions (Supplementary Table 12, Fig. 7).

Spatial transcriptomics

Spatial visualization (Fig. 8) revealed a significant increase in fibroblast proportion in IPF samples compared to controls. Immune system cells exhibited enhanced clustering and abundance in IPF, whereas their distribution in controls was sparse and minimal. Alveolar macrophages in IPF were dispersed around immune cell clusters and showed reduced proportions relative to controls. Gene exploration analysis identified six hub genes with differential expression between IPF and control groups across cell types. Violin plots demonstrated that EZH1 expression was globally downregulated in IPF, whereas BRCA1, FAM13A, and ANO9 exhibited elevated expression, which is consistent with the findings observed in the scRNA-seq analysis. CCDC200 and SFR1 displayed comparable expression levels between groups. Notably, all six genes showed upregulated expression in immune system cells of IPF compared to controls. Intriguingly, CCDC200 and SFR1 were nearly absent in immune cells of control samples. Detailed data can be found in (Fig. 9).

Molecular docking

We performed molecular docking for the six essential genes to assess their druggability (Fig. 10 and Supplementary Table S12). ANO9 exhibited the strongest binding affinity with FLUNISOLIDE at −13 kcal/mol, with an inhibition constant (Ki) of 295.72 pM. Additionally,

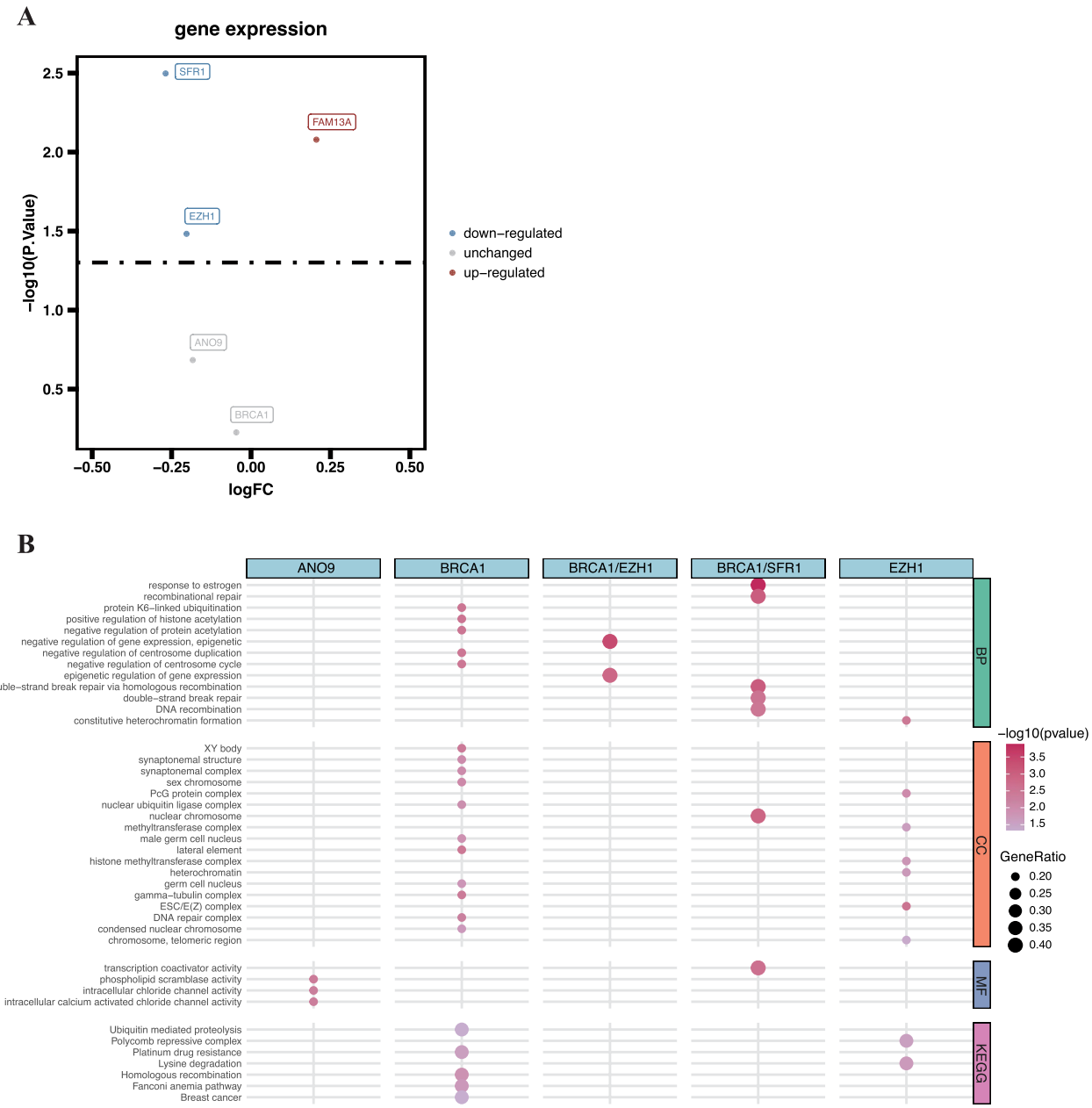

Fig. 5 **A** Volcano plot showing the differential expression of critical genes. The x-axis represents the logFC, and the y-axis represents the -log10(P-value). Genes are color-coded based on their expression changes: blue for downregulated genes, red for upregulated genes, and grey for unchanged genes. Notable upregulated genes include FAM13A, while SFR1 and EZH1 are among the downregulated genes. **B** Dot plot summarizing the functional enrichment analysis of critical genes across different categories, including Biological Process (BP), Cellular Component (CC), Molecular Function (MF), and Kyoto Encyclopedia of Genes and GenomesKEGG pathways. The dot color represents the −log10(P-value), and the size of the dots indicates the gene ratio. Significant pathways associated with BRCA1, BRCA1/SFR1, and EZH1 are highlighted

DEXAMETHASONE showed high binding affinity with both BRCA1 (Affinity = −11.79 kcal/mol, Ki = 2.28 nM) and EZH1 (Affinity = −14.94 kcal/mol, Ki = 11.19 pM). CCDC200 demonstrated the tightest binding with BETAMETHASONE (Affinity = −8.84 kcal/mol) at a Ki of 331.28 nM. Furthermore, AZATADINE and MECLIZINE exhibited strong binding affinities with FAM3A (Affinity = −13.78 kcal/mol, Ki = 79.27 nM) and SFR1 (Affinity = −9.836 kcal/mol, Ki = 61.68 nM), respectively. Notably, DEXAMETHASONE showed the highest binding affinity (Affinity ≤ −8 kcal/mol) across all six genes, indicating its potential strong interaction with these

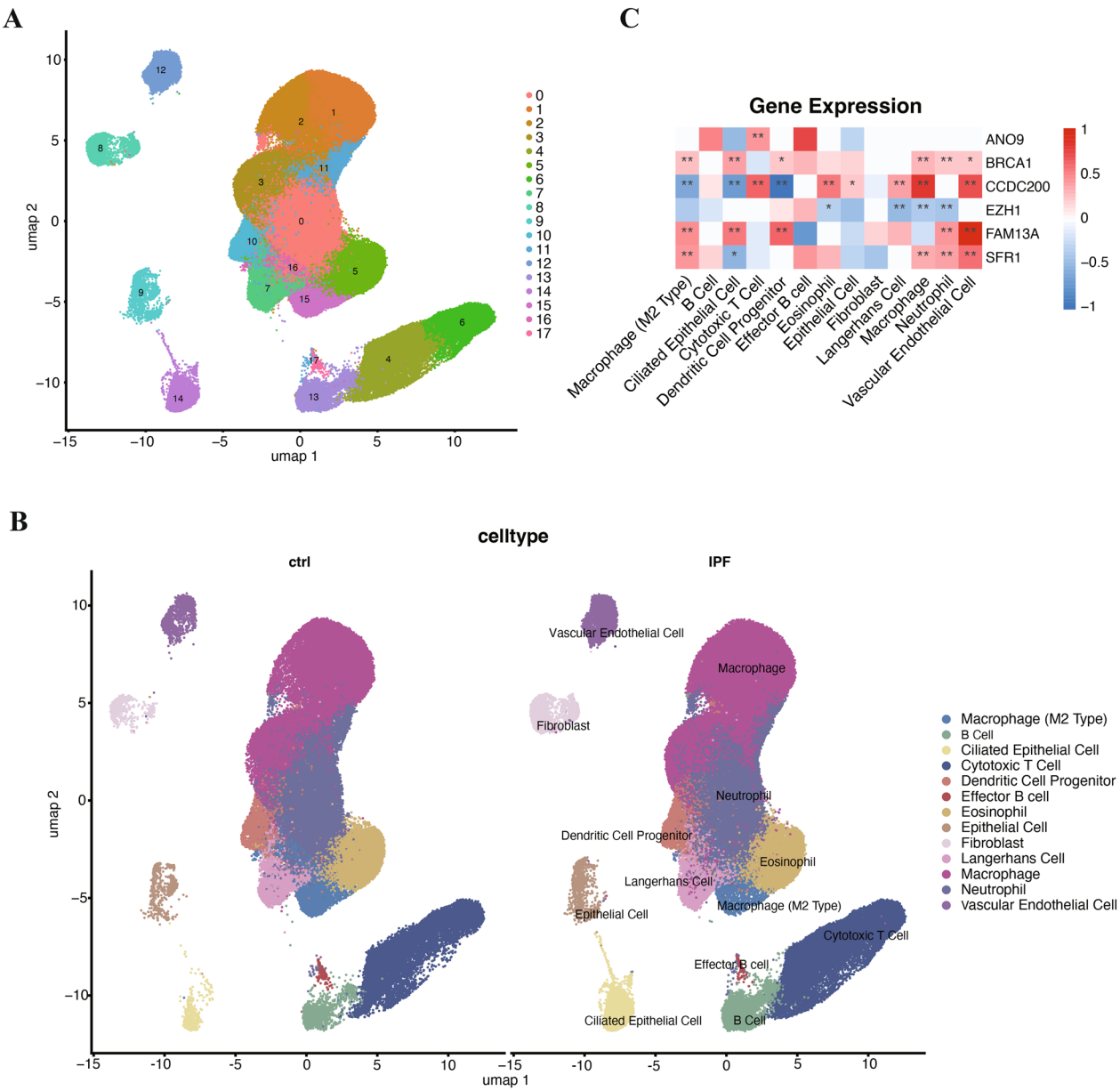

Fig. 6 **A** UMAP plot showing the clustering of cells into 18 distinct clusters based on single-cell RNA sequencing analysis. Each cluster is color-coded and labeled accordingly. **B** UMAP plot comparing cell types between control (ctrl) and IPF samples. Different cell types are indicated by color, showing the distribution of cells across conditions. **C** Heatmap displaying the gene expression levels of essential genes (ANO9, BRCA1, CCDC200, EZH1, FAM13A, SFR1) across various cell types. The color intensity represents the expression values, with red indicating upregulation and blue indicating downregulation. Asterisks indicate statistically significant differences in expression (**p < 0.01, *p < 0.05)

targets. The remaining LE, SILE, LLE, FQ, and hydrogen bond result information can be found in Supplementary Table S13.

Phenome-wide mendelian randomization (PheW-MR)

After determining the druggability of the critical genes, we conducted a PheW-MR analysis of these proteins against 679 disease traits to provide a more comprehensive description of potential side effects for each protein. We identified 47 significant associations, with 16 linked to harmful and 31 to beneficial side effects (Fig. 11 and Supplementary Table S14). Notably, ANO9 was consistently associated with an increased risk of several diseases, including Urethral stricture (not

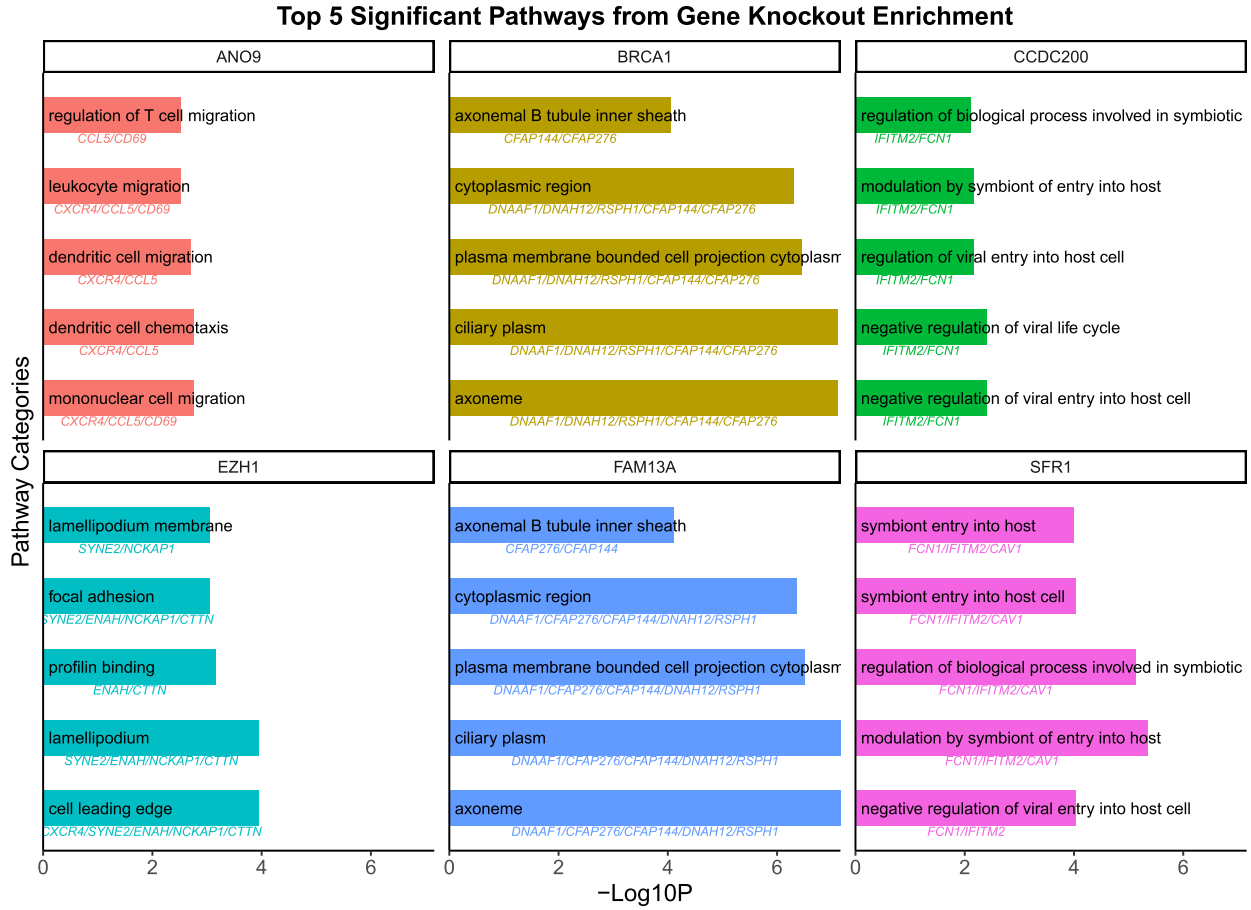

Fig. 7 Bar chart of the top 5 most significant pathways from gene enrichment analysis of 6 key gene knockouts. The pathways include biological processes such as mononuclear cell migration, dendritic cell chemotaxis and migration, leukocyte migration, regulation of T cell migration, and others. Statistical significance is indicated by − Log10P values

specified as infectious) ($P = 8.80 \times 10^{\wedge}-03$, OR = 1.20), Anal and rectal polyp ($P = 4.79 \times 10^{\wedge}-03$, OR = 1.14), and Abdominal pain ($P = 6.83 \times 10^{\wedge}-03$, OR = 1.06). In contrast, FAM3A and SFR1 were each associated with an increased risk of only one condition: Polyarteritis nodosa and allied conditions ($P = 1.92 \times 10^{\wedge}-03$, OR = 1.26) for FAM3A, and Burns ($P = 9.09 \times 10^{\wedge}-03$, OR = 1.45) for SFR1. Regarding beneficial side effects, both BRCA1 ($P = 2.43 \times 10^{\wedge}-05$, OR = 0.89) and CCDC200 ($P = 5.62 \times 10^{\wedge}-05$, OR = 0.90) were associated with a reduced risk of Varicose veins of the lower extremity. EZH1 ($P = 2.31 \times 10^{\wedge}-03$, OR = 0.63) was associated with a lower risk of Stomach cancer. FAM13A was linked to a reduced risk of other disorders of synovium, tendon, and bursa ($P = 2.65 \times 10^{\wedge}-04$, OR = 0.91), and SFR1 showed a significant association with a reduced risk of Other specified gastritis ($P = 4.45 \times 10^{\wedge}-03$, OR = 0.89).

Discussion

Our study employed a multi-omics approach to prioritize potential drug targets for IPF by effectively integrating results from TWAS and GWAS. Using advanced computational methods such as OTTERS and MR, we identified essential genes that have a causal association with IPF. These essential genes were further subjected to differential gene expression and enrichment analyses through scRNA-seq and bulk RNA-seq, may reveal the pathways through which these genes influence IPF progression. Our approach identified six essential genes with significant therapeutic potential for IPF: BRCA1, EZH1, FAM13A, SFR1, ANO9, and CCDC200. At the single-cell level, we revealed that these genes were predominantly expressed in macrophages. This observation was further validated by spatial transcriptomics analysis, which demonstrated similar spatial expression patterns. Additionally, we evaluated the drug safety profiles of these critical genes using PheW-MR and molecular docking, providing

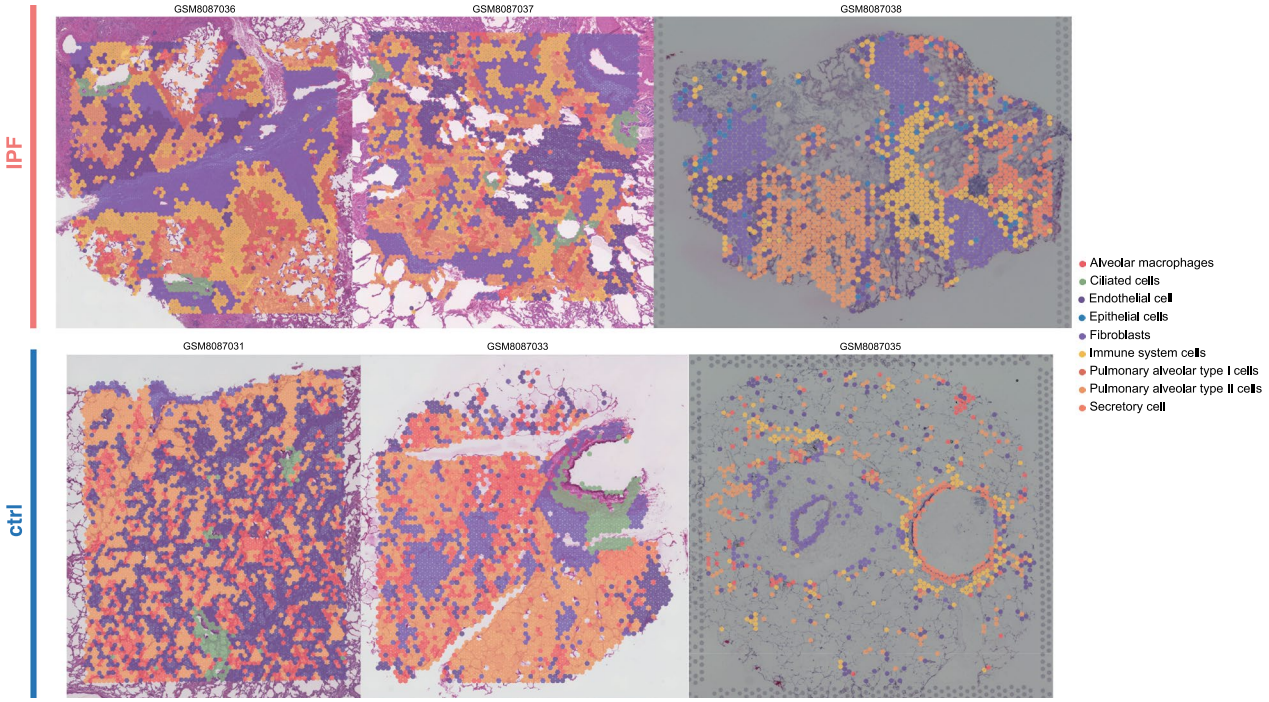

Fig. 8 Spatial plots visualize the cell types annotation across all samples

a comprehensive assessment of their potential clinical application. This validated their promise as drug targets for IPF treatment. This study provides new insights into the genetic underpinnings of IPF and highlights six promising candidate drug targets that could pave the way for developing effective therapeutic strategies for the disease.

IPF characteristic histopathological findings primarily include fibroblast foci, proliferative epithelial cells, and inflammatory responses[1]. The **BRCA1** gene encodes a multifunctional protein critical for DNA double-strand break repair through homologous recombination repair (HRR) and regulates gene expression, the cell cycle, and genome stability via histone modification and ubiquitination[53]. Our TWAS results indicate that **BRCA1** is associated with an increased risk of IPF, which was corroborated by subsequent GWAS findings. Enrichment and single-cell analysis revealed that **BRCA1** regulates gene expression during inflammatory responses and immune damage(Fig. 5–6). In IPF, prolonged environmental and inflammatory stress induces DNA damage in ciliated epithelial cells. BRCA1 upregulation promotes DNA repair (double-strand break/recombinational pathways), enabling survival of damaged cells that may accumulate mutations and secrete pro-fibrotic signals to exacerbate fibrosis[54, 55]. In macrophages (particularly M2-type), BRCA1 modulates gene regulation, epigenetic modifications, and ubiquitination pathways, potentially sustaining chronic inflammation. Dendritic cell progenitors exhibit BRCA1-mediated transcriptional/epigenetic dysregulation that may impair immune clearance. Epithelial BRCA1 drives EMT through ubiquitination pathways, increasing fibroblast production [56]. Moreover, **BRCA1** regulates fibrosis-related gene expression via DNA methylation and histone acetylation, contributing to abnormal extracellular matrix deposition, lung stiffening, and fibrosis[57]. In endothelial cells, BRCA1 upregulation induces aberrant angiogenesis through histone acetylation/DNA repair mechanisms, worsening tissue scarring and gas exchange impairment[58, 59].

The **EZH1** gene encodes a protein involved in maintaining stem cell pluripotency and regulating cell differentiation[60]. In IPF, **EZH1** downregulation impairs the Polycomb Repressive Complex pathway's ability to silence pro-inflammatory genes in neutrophils and macrophages, exacerbating lung inflammation and fibrosis progression[61]. Furthermore, reduced activity of the histone methyltransferase complex due to **EZH1** downregulation decreases H3K27 trimethylation, disrupting antigen presentation and promoting fibrosis-related gene activation[61, 62]. The downregulation of **EZH1** also weakens epigenetic silencing in eosinophils, leading to enhanced release of pro-fibrotic signals, such as TGF-β, promoting fibrosis[55]. These findings were confirmed by single-cell and enrichment analyses. As a gene strongly associated with

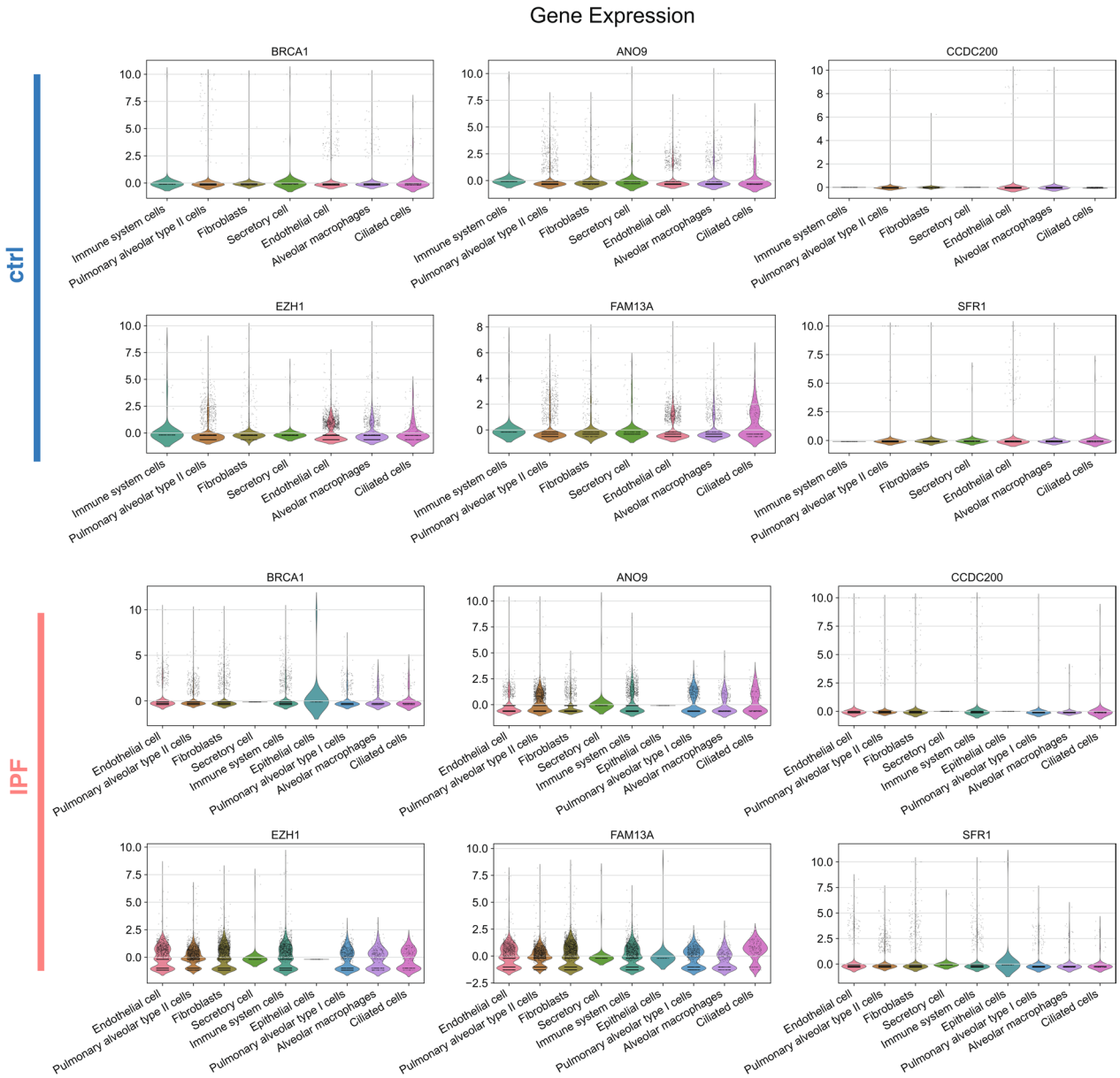

Fig. 9 Violin plot displaying the gene expression levels of essential genes (ANO9, BRCA1, CCDC200, EZH1, FAM13A, SFR1) across various cell types in Spatial Transcriptomics

IPF, **EZH1** was validated through both TWAS and GWAS, with its expression linked to an increased risk of IPF. **FAM13A** is significantly linked to IPF susceptibility, lung function, and prognosis[63]. Although not enriched in specific pathways, single-cell analysis indicated its upregulation in immune cells, ciliated epithelial cells, and endothelial cells during IPF. **FAM13A** is highly expressed in small airway epithelial cells and correlates with markers of EMT, suggesting its role in EMT during epithelial cell transformation[64]. Additionally, **FAM13A** upregulation in endothelial cells may lead to aberrant angiogenesis and increased vascular permeability, which could promote fibroblast migration and further fibrotic lesion expansion[65]. In neutrophils, M2 macrophages, and dendritic cell progenitors, **FAM13A** upregulation promotes IPF progression through immune dysregulation, increasing pro-inflammatory cytokine release and extracellular matrix deposition[66–68].

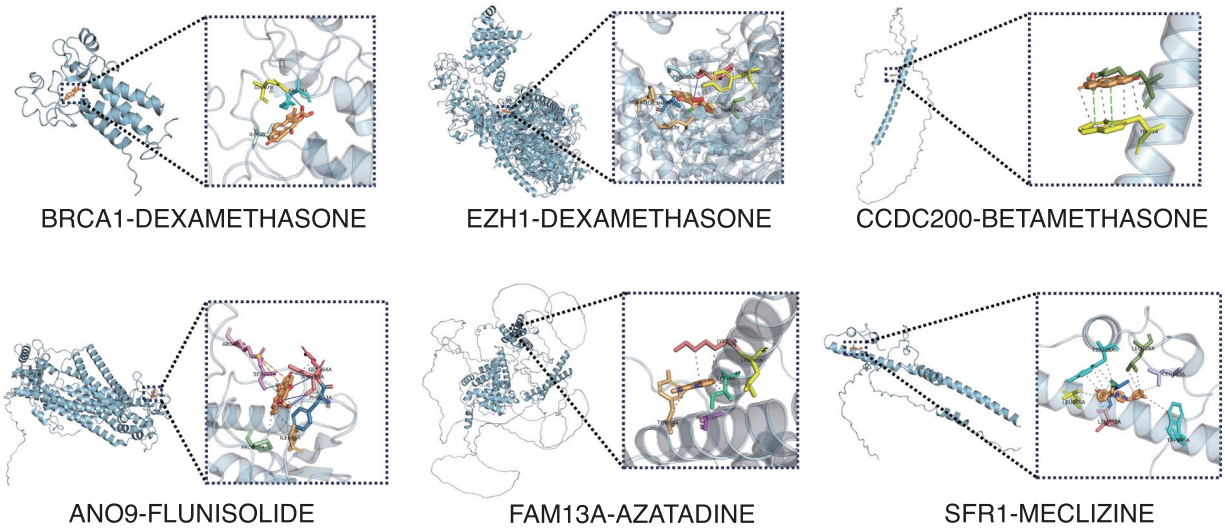

Fig. 10 Schematic representation of the docking interactions between critical genes and the most significant respiratory system drugs. Each panel highlights the binding interface between a critical gene and its corresponding drug

Although the direct impact of the other three candidate genes—**SFR1**, **ANO9**, and **CCDC200**—on IPF progression has not been explicitly confirmed, they were validated through TWAS and GWAS screening in our multi-omics analysis. In IPF, **ANO9** upregulation in cytotoxic T cells may contribute to fibrosis progression by affecting phospholipid scramblase activity, calcium-activated chloride channel activity, and intracellular chloride channel activity[69, 70]. This upregulation could impair apoptotic cell clearance and enhance immune dysregulation, further driving fibrosis[71]. Additionally, **ANO9** may promote T cell activation and migration, exacerbating local inflammation and fibrosis[72, 73]. For **SFR1**, our study revealed its enrichment in gene expression regulation pathways shared with **BRCA1**. **SFR1** is involved in DNA repair and gene expression regulation, enhancing the activity of recombinases like Rad51 and Dmc1[74], which could exacerbate immune dysregulation in **BRCA1**-associated pathways. **CCDC200**, a protein involved in regulating transcription factor expression and the cell cycle[75], may promote fibroblast senescence and facilitate IPF development [1]. However, further validation is needed.

However, the PheW-MR analysis also highlighted potential side effects associated with the modulation of these genes, which raises important concerns regarding their clinical application. Balancing the therapeutic benefits with these potential risks is crucial, as adverse effects may undermine the efficacy and safety of therapies targeting these genes. For example, while BRCA1 showed strong binding affinities with glucocorticoids, which are commonly used in treating acute exacerbations of IPF, the long-term impact of glucocorticoid use must be carefully considered due to potential side effects such as immune suppression and metabolic disturbances. Similarly, EZH1's involvement in inflammation regulation could lead to unintended consequences, such as exacerbating autoimmune conditions in some patients. Therefore, it is important to thoroughly assess these risks through clinical trials and individualized treatment plans. This validated their promise as drug targets for IPF treatment.

Compared to traditional approaches, our study offers several advantages in identifying drug targets for IPF treatment. First, OTTERS was utilized to conduct TWAS analysis on plasma proteins data related to IPF. Compared to other TWAS tools, OTTERS has a very clear advantage, namely the ability to handle both summary-level and individual-level data. OTTERS employs multiple PRS methods to estimate eQTL weights from aggregate data and conducts a comprehensive TWAS [22–24, 76]. Second, unlike conventional single-omics methods, our multi-omics analysis integrates genomic and transcriptomic data for gene screening and validation, providing a more comprehensive understanding of gene-disease relationships. Third, traditional genetic approaches often overlook the biological mechanisms underlying disease progression. In contrast, our study combined genetic data with single-cell sequencing and enrichment analysis to explore the mechanistic roles of genes in IPF progression. Moreover, we employed molecular docking and PheW-MR to evaluate the druggability

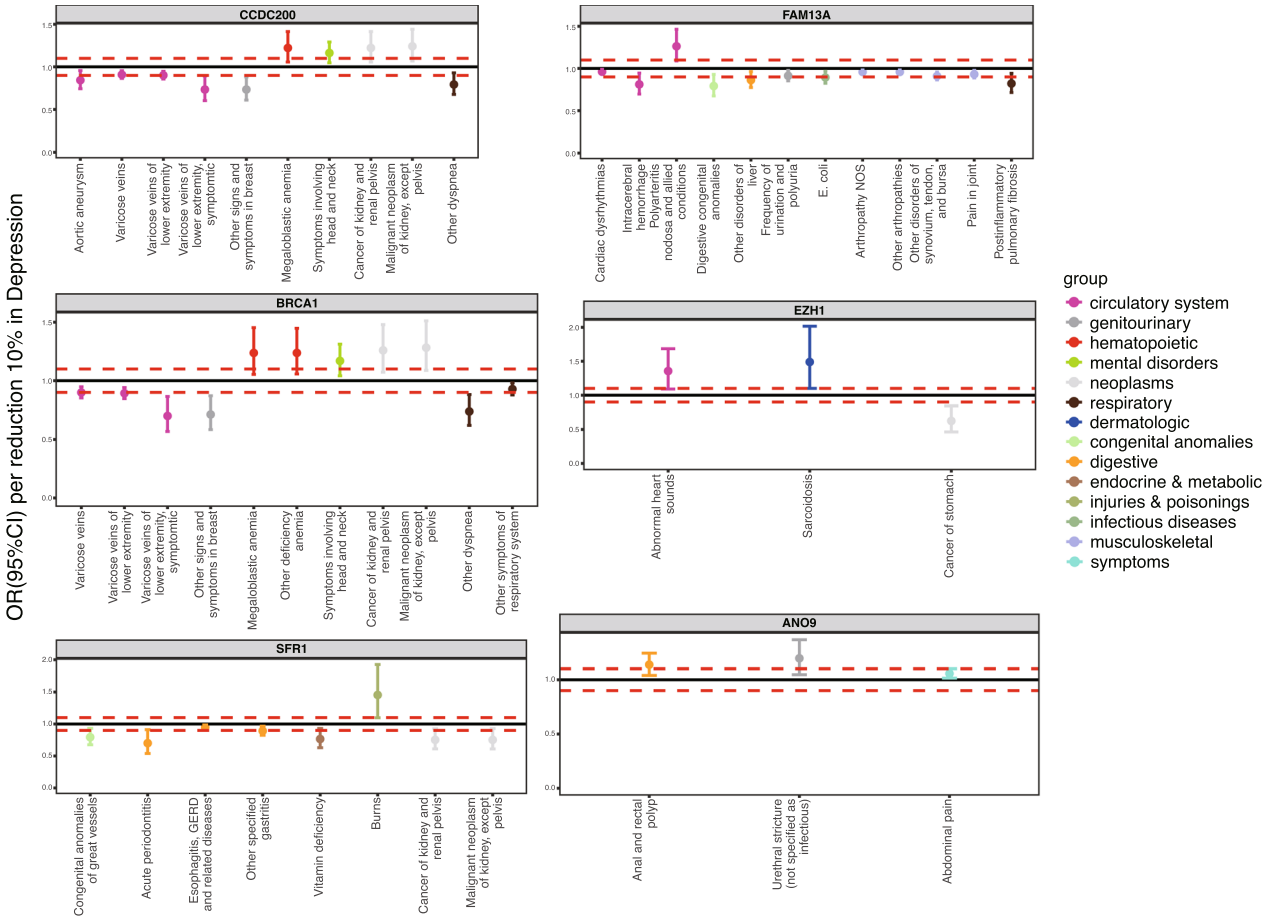

Fig. 11 Forest plots showing the OR and 95% CI for various disease categories associated with a 10% reduction in depression across six key genes: CCDC200, FAM13A, BRCA1, EZH1, SFR1, and ANO9. Each plot presents the OR (95% CI) for specific disease conditions, with the dotted red line representing the null effect (OR = 1). The color-coded dots correspond to different disease categories, as indicated by the legend on the right. Significant associations are highlighted where the confidence intervals do not cross the null line

and safety of candidate genes. Notably, our validated hub genes were highly expressed in immune cells and tissue repair-related cells, rather than fibroblasts, highlighting inflammation as a key factor in IPF development [77]. Current anti-inflammatory treatments for IPF remain empirical, with limited progress. However, studies suggest that early immune therapy interventions, particularly in early inflammatory-to-fibrotic transitions, may benefit IPF patients [78–80]. This suggests that early-stage IPF may respond to immunosuppressant therapies, warranting further investigation into optimal treatment strategies.

Our study acknowledges several limitations. The restriction to European populations may affect the generalizability of our findings, though existing evidence suggests that racial differences in IPF outcomes—such as age of onset, survival rates, and treatment access—are more likely driven by structural health disparities and socioeconomic factors rather than inherent genetic differences [81]. This supports the potential biological relevance of our findings across populations, despite potential variations in clinical manifestations due to environmental and social contexts. Besides, OTTERS is not without limitations. For instance, the methodology requires corresponding loci-specific training datasets to conduct its analysis and demands relatively high-performance computational equipment for implementation, which could potentially restrict its accessibility in resource-constrained research settings. The absence of IPF clinical cohort samples also precluded tissue-level validation and analysis of drug history and patient prognosis. Additionally, since conventional bleomycin-induced fibrosis mouse models fail to adequately recapitulate the gradual decline in forced vital capacity and other pulmonary function parameters characteristic of human IPF progression, we refrained from murine validation in this study [82]. Future investigations will employ genetically engineered mouse

models to address this pathophysiological recapitulation gap. These limitations highlight the need for further research incorporating multi-ancestry datasets, clinical data, and experimental models to address gene-environment interactions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06368-8.

Supplementary Material 1. S1: Summary of GWAS, eQTL, RNA-seq, scRNA-seq, and Drug Database Resources Used for IPF Research and Drug Target Discovery. S2: 679 Diseases from UK Biobank. S3: TWAS Results within the OTTERS Framework from the Discovery Dataset, Identifying Genes Associated with IPF. S4: TWAS Results within the OTTERS Framework from the Duplication Dataset, Identifying Genes Associated with IPF. S5: Identification of Genes Causally Linked to IPF through MR. S6: Genes Causally Associated with IPF Identified from the Discovery DatasetUsing SMR. S7: Genes Causally Associated with IPF Identified from the Duplication DatasetUsing SMR. S8: Expression levels of key genes in bulk RNA-seq datasets. S9: GO and KEGG Enrichment Results of key Genes. S10: Differential Expression of IPF-Related key Genes at the Single-Cell Level. S11: Genes Significantly Affected After Virtual Knockout. S12: Significantly Affected Pathways After Virtual Knockout of Key Genes. S13: Respiratory System Drugs Identified via Docking as Potential Interactors with key Genes. S14: Results of key genes intervention on-target side effects identified through PheW-MR analysis

Author contributions

All authors made significant contributions to this work and have approved the final manuscript. Concept and design: ZFW, WXL, ZYY, JJL and ZSC. Data curation: ZFW, WXL, ZYY, JJL and ZSC. Analysis and interpretation of data: ZFW, WXL, ZYY, JJL and ZSC. Computational resources and support: ZFW, WXL, ZYY, JJL and ZSC. Writing of the original draft and reviews: ZFW, WXL, JY, RGX, JP and KYL. Editing draft and reviews: WXL, ZFW and ZSC.

Funding

No funding.

Availability of data and materials

Publicly available datasets were analyzed in this study. This data can be found in Supplementary table S1 and Supplementary table S2.

Code availability

The code generated and analyzed during the current study are available from the corresponding author upon a reasonable request.

Declarations

Ethics approval and consent to participate

Each cohort included in this study was conducted using published studies and consortia, which provided publicly available summary statistics. All original studies have received ethical approval and agreed to participate, and summary-level data were provided for analysis.

Consent for publication

Not applicable.

Competing interests

The authors declare that there is no conflict of interest.

Author details

[1]1The Sixth School of Clinical Medicine, Department of Respiratory and Critical Care Medicine, the Affiliated Qingyuan Hospital (Qingyuan People's Hospital), Guangzhou Medical University, Qingyuan, China. [2]The First School of Clinical Medicine, Guangzhou Medical University, Guangzhou, China. [3]The Third School of Clinical Medicine, Guangzhou Medical University, Guangzhou, China. [4]The Second School of Clinical Medicine, Guangzhou Medical University, Guangzhou, China.

Received: 13 December 2024 Accepted: 7 March 2025
Published online: 16 March 2025

References

1. Moss BJ, Ryter SW, Rosas IO. Pathogenic mechanisms underlying idiopathic pulmonary fibrosis. Annu Rev Pathol. 2022;17:515–46.
2. Maher TM, Bendstrup E, Dron L, Langley J, Smith G, Khalid JM, Patel H, Kreuter M. Global incidence and prevalence of idiopathic pulmonary fibrosis. Respir Res. 2021;22:197.
3. Ge Z, Chen Y, Ma L, Hu F, Xie L. Macrophage polarization and its impact on idiopathic pulmonary fibrosis. Front Immunol. 2024;15:1444964.
4. Libra A, Sciacca E, Muscato G, Sambataro G, Spicuzza L, Vancheri C. Highlights on future treatments of IPF: clues and pitfalls. Int J Mol Sci. 2024;25:8392.
5. Bonella F, Spagnolo P, Ryerson C. Current and future treatment landscape for idiopathic pulmonary fibrosis. Drugs. 2023;83:1581–93.
6. Pitre T, Mah J, Helmeczi W, Khalid MF, Cui S, Zhang M, Husnudinov R, Su J, Banfield L, Guy B, et al. Medical treatments for idiopathic pulmonary fibrosis: a systematic review and network meta-analysis. Thorax. 2022;77:1243–50.
7. Guo H, Sun J, Zhang S, Nie Y, Zhou S, Zeng Y. Progress in understanding and treating idiopathic pulmonary fibrosis: recent insights and emerging therapies. Front Pharmacol. 2023;14:1205948.
8. Gao J, Liu M, Lu M, Zheng Y, Wang Y, Yang J, Xue X, Liu Y, Tang F, Wang S, et al. Integrative analysis of transcriptome, DNA methylome, and chromatin accessibility reveals candidate therapeutic targets in hypertrophic cardiomyopathy. Protein Cell. 2024;15:796–817.
9. Carlson JS, Marleau P, Zarkesh RA, Feng PL. Taking advantage of disorder: small-molecule organic glasses for radiation detection and particle discrimination. J Am Chem Soc. 2017;139:9621–6.
10. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Kote-Jarai Z, Eeles R, Freedman M, Haiman C, Pasaniuc B. consortium P: Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun. 2018;9:4079.
11. Spreafico R, Soriaga LB, Grosse J, Virgin HW, Telenti A. Advances in genomics for drug development. Genes (Basel). 2020;11:942.
12. Yang S, Guo J, Kong Z, Deng M, Da J, Lin X, Peng S, Fu J, Luo T, Ma J, et al. Causal effects of gut microbiota on sepsis and sepsis-related death: insights from genome-wide Mendelian randomization, single-cell RNA, bulk RNA sequencing, and network pharmacology. J Transl Med. 2024;22:10.
13. Xu F, Tong Y, Yang W, Cai Y, Yu M, Liu L, Meng Q. Identifying a survival-associated cell type based on multi-level transcriptome analysis in idiopathic pulmonary fibrosis. Respir Res. 2024;25:126.
14. Vosa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.
15. Zhou W, Kanai M, Wu KH, Rasheed H, Tsuo K, Hirbo JB, Wang Y, Bhattacharya A, Zhao H, Namba S, et al. Global biobank meta-analysis initiative: powering genetic discovery across human disease. Cell Genom. 2022;2:100192.
16. Kurki MI, Karjalainen J, Palta P, Sipila TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613:508–18.
17. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, Deluliis G, Januszyk M, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv. 2020;6:eaba1983.
18. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung MI, Taylor CJ, Jetter C, et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv. 2020;6:eaba1972.

19. Herazo-Maya JD, Noth I, Duncan SR, Kim S, Ma SF, Tseng GC, Feingold E, Juan-Guardela BM, Richards TJ, Lussier Y, et al. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. Sci Transl Med. 2013;5:205ra136.
20. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE248082
21. Dai Q, Zhou G, Zhao H, Vosa U, Franke L, Battle A, Teumer A, Lehtimaki T, Raitakari OT, Esko T, et al. OTTERS: a powerful TWAS framework leveraging summary-level reference data. Nat Commun. 2023;14:1271.
22. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
23. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41:469–80.
24. Zhou G, Zhao H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. 2021;17: e1009697.
25. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1776.
26. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23:R89-98.
27. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018. https://doi.org/10.7554/eLife.34408.
28. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun. 2018;9:918.
29. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.
31. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1–18.
32. Zhou X, Franklin RA, Adler M, Carter TS, Condiff E, Adams TS, Pope SD, Philip NH, Meizlish ML, Kaminski N, Medzhitov R. Microenvironmental sensing by fibroblasts controls macrophage population size. Proc Natl Acad Sci U S A. 2022;119: e2205360119.
33. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
34. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42:293–304.
35. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8(329–337): e324.
36. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, Campbell JD. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21:57.
37. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat Methods. 2024;21:1462–5.
38. Osorio D, Zhong Y, Li G, Xu Q, Yang Y, Tian Y, Chapkin RS, Huang JZ, Cai JJ. scTenifoldKnk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. Patterns (N Y). 2022;3: 100434.
39. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
40. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13:1246.
41. https://www.rcsb.org/
42. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52:D1180–92.
43. Du L, Geng C, Zeng Q, Huang T, Tang J, Chu Y, Zhao K. Dockey: a modern integrated tool for large-scale molecular docking and virtual screening. Brief Bioinform. 2023. https://doi.org/10.1093/bib/bbad047.
44. Hassan NM, Alhossary AA, Mu Y, Kwoh CK. Protein-ligand blind docking using quickvina-w with inter-process spatio-temporal integration. Sci Rep. 2017;7:15451.
45. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. open babel: an open chemical toolbox. J Cheminform. 2011;3:33.
46. https://www.pymol.org/
47. Hopkins AL, Keseru GM, Leeson PD, Rees DC, Reynolds CH. The role of ligand efficiency metrics in drug discovery. Nat Rev Drug Discov. 2014;13:105–21.
48. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. Drug Discov Today. 2004;9:430–1.
49. Nissink JW. Simple size-independent measure of ligand efficiency. J Chem Inf Model. 2009;49:1617–22.
50. Reynolds CH, Tounge BA, Bembenek SD. Ligand binding efficiency: trends, physical basis, and implications. J Med Chem. 2008;51:2432–8.
51. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov. 2007;6:881–90.
52. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50:1335–41.
53. Ismail T, Alzneika S, Riguene E, Al-Maraghi S, Alabdulrazzak A, Al-Khal N, Fetais S, Thanassoulas A, AlFarsi H, Nomikos M. BRCA1 and its vulnerable c-terminal brct domain: structure, function, genetic mutations and links to diagnosis and treatment of breast and ovarian cancer. Pharmaceuticals (Basel). 2024;17:333.
54. Vancheri C. Common pathways in idiopathic pulmonary fibrosis and cancer. Eur Respir Rev. 2013;22:265–72.
55. Pokhreal D, Crestani B, Helou DG. Macrophage implication in IPF: updates on immune, epigenetic, and metabolic pathways. Cells. 2023;12:2193.
56. Fu X, Tan W, Song Q, Pei H, Li J. BRCA1 and breast cancer: molecular mechanisms and therapeutic strategies. Front Cell Dev Biol. 2022;10: 813457.
57. Yang J, Liang C, Liu L, Wang L, Yu G. High-fat diet related lung fibrosis-epigenetic regulation matters. Biomolecules. 2023;13:558.
58. Patil RS, Maloney ME, Lucas R, Fulton DJR, Patel V, Bagi Z, Kovacs-Kasa A, Kovacs L, Su Y, Verin AD. Zinc-dependent histone deacetylases in lung endothelial pathobiology. Biomolecules. 2024;14:140.
59. Barnes PJ, Adcock IM, Ito K. Histone acetylation and deacetylation: importance in inflammatory lung diseases. Eur Respir J. 2005;25:552–63.
60. An R, Li YQ, Lin YL, Xu F, Li MM, Liu Z. EZH1/2 as targets for cancer therapy. Cancer Gene Ther. 2023;30:221–35.
61. Liu Y, Zhang Q, Ding Y, Li X, Zhao D, Zhao K, Guo Z, Cao X. Histone lysine methyltransferase Ezh1 promotes TLR-triggered inflammatory cytokine production by suppressing Tollip. J Immunol. 2015;194:2838–46.
62. Moll H, Neiß U, Demleitner K, Marx A, Mehlig M, Scheicher C, Reske K: Antigen presentation by Langerhans/dendritic cells. Immune Functions of Epidermal Langerhans Cells 1995:87–101.
63. Hirano C, Ohshimo S, Horimasu Y, Iwamoto H, Fujitaka K, Hamada H, Hattori N, Shime N, Bonella F, Guzman J, et al. FAM13A polymorphism as a prognostic factor in patients with idiopathic pulmonary fibrosis. Respir Med. 2017;123:105–9.
64. Zhu J, Wang F, Feng X, Li B, Ma L, Zhang J. Family with sequence similarity 13 member A mediates TGF-β1-induced EMT in small airway epithelium of patients with chronic obstructive pulmonary disease. Respir Res. 2021;22:192.
65. Hsu T, Nguyen-Tran HH, Trojanowska M. Active roles of dysfunctional vascular endothelium in fibrosis and cancer. J Biomed Sci. 2019;26:86.
66. Margraf A, Lowell CA, Zarbock A. Neutrophils in acute inflammation: current concepts and translational implications. Blood. 2022;139:2130–44.
67. Braga TT, Agudelo JS, Camara NO. Macrophages during the fibrotic process: M2 as friend and foe. Front Immunol. 2015;6:602.
68. Kishore A, Petrek M. Roles of macrophage polarization and macrophage-derived miRNAs in pulmonary fibrosis. Front Immunol. 2021;12: 678457.

69. Lowry AJ, Liang P, Song M, Wan Y, Pei ZM, Yang H, Zhang Y. TMEM16 and OSCA/TMEM63 proteins share a conserved potential to permeate ions and phospholipids. Elife. 2024. https://doi.org/10.7554/eLife.96957.3.
70. Le SC, Yang H. Structure-function of TMEM16 ion channels and lipid scramblases. Adv Exp Med Biol. 2021;1349:87–109.
71. Jun I, Park HS, Piao H, Han JW, An MJ, Yun BG, Zhang X, Cha YH, Shin YK, Yook JI, et al. ANO9/TMEM16J promotes tumourigenesis via EGFR and is a novel therapeutic target for pancreatic cancer. Br J Cancer. 2017;117:1798–809.
72. Behuria HG, Dash S, Sahu SK. Phospholipid scramblases: role in cancer progression and anticancer therapeutics. Front Genet. 2022;13: 875894.
73. Sala-Rabanal M, Yurtsever Z, Berry KN, McClenaghan C, Foy AJ, Hanson A, Steinberg DF, Greven JA, Kluender CE, Alexander-Brett JM, et al. Modulation of TMEM16B channel activity by the calcium-activated chloride channel regulator 4 (CLCA4) in human cells. J Biol Chem. 2024;300: 107432.
74. Feng Y, Singleton D, Guo C, Gardner A, Pakala S, Kumar R, Jensen E, Zhang J, Khan S. DNA homologous recombination factor SFR1 physically and functionally interacts with estrogen receptor alpha. PLoS ONE. 2013;8: e68075.
75. Priyanka PP, Yenugu S. Coiled-coil domain-containing (CCDC) proteins: functional roles in general and male reproductive physiology. Reprod Sci. 2021;28:2725–34.
76. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun. 2017;8:456.
77. Savin IA, Zenkova MA, Sen'kova AV. Pulmonary fibrosis as a result of acute lung inflammation: molecular mechanisms, relevant in vivo models, prognostic and therapeutic approaches. Int J Mol Sci. 2022;23:14959.
78. Spagnolo P, Distler O, Ryerson CJ, Tzouvelekis A, Lee JS, Bonella F, Bouros D, Hoffmann-Vold AM, Crestani B, Matteson EL. Mechanisms of progressive fibrosis in connective tissue disease (CTD)-associated interstitial lung diseases (ILDs). Ann Rheum Dis. 2021;80:143–50.
79. Hoffmann-Vold AM, Maher TM, Philpot EE, Ashrafzadeh A, Barake R, Barsotti S, Bruni C, Carducci P, Carreira PE, Castellví I, et al. The identification and management of interstitial lung disease in systemic sclerosis: evidence-based European consensus statements. Lancet Rheumatol. 2020;2:e71–83.
80. Seibold JR, Maher TM, Highland KB, Assassi S, Azuma A, Hummers LK, Costabel U, von Wangenheim U, Kohlbrenner V, Gahlemann M, et al. Safety and tolerability of nintedanib in patients with systemic sclerosis-associated interstitial lung disease: data from the SENSCIS trial. Ann Rheum Dis. 2020;79:1478–84.
81. Adegunsoye A, Vela M, Saunders M. racial disparities in pulmonary fibrosis and the impact on the black population. Arch Bronconeumol. 2022;58:590–2.
82. Borzone G, Moreno R, Urrea R, Meneses M, Oyarzun M, Lisboa C. Bleomycin-induced chronic lung damage does not resemble human idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2001;163:1648–53.

Publisher's Note