Journal of Clinical and Translational Hepatology **2026**
DOI: 10.14218/JCTH.2025.00676

Original Article

A Multi-omics and Machine Learning Framework Identifies Plasma SBDS as a Causal Biomarker and Therapeutic Target in Primary Sclerosing Cholangitis

Pengfei Cheng[1,2#], Yuanming Qiang[1,2#], Yibo Sun[1,2], Binwei Duan[1,2], Yabo Ouyang[1*] and Guangming Li[1,2*]

[1]*Department of General Surgery Center, Beijing YouAn Hospital, Capital Medical University, Beijing Institute of Hepatology, Beijing, China;* [2]*Clinical Center for Liver Cancer, Capital Medical University, Beijing, China*

Received: December 11, 2025 | **Revised:** February 25, 2026 | **Accepted:** March 02, 2026 | **Published online:** March 20, 2026

Abstract

Background and Aims: Primary sclerosing cholangitis (PSC) is an immune-mediated cholestatic liver disease. Its molecular etiology remains poorly defined, hindering the development of mechanism-based diagnostics and therapies. Therefore, this study aimed to identify key molecular drivers and causal biomarkers of PSC by integrating transcriptomics, machine learning, and genetic causal inference. ***Methods:*** We deployed an integrated computational framework combining transcriptomics, network biology, machine learning, and genetic causal inference. Peripheral blood transcriptomes from PSC patients and controls were analyzed to identify disease-associated modules. Candidate genes were refined via protein-protein interaction networks and a multi-algorithm machine learning screen. Causal inference was performed using two-sample Mendelian randomization, integrating plasma protein quantitative trait loci with PSC genome-wide association study summary statistics. ***Results:*** Transcriptomic analysis revealed a PSC-associated module enriched in ribosome biogenesis and protein homeostasis pathways. A machine learning-optimized nine-gene signature (including PTMA, SUMO1, Shwachman-Bodian-Diamond syndrome (SBDS), RPL7, EIF1AX, ANP32A, PCNA, FAM98A, and MPHOSPH6) achieved high diagnostic accuracy (mean AUC = 0.908) and was consistently downregulated in PSC. This signature was linked to a remodeled immune microenvironment characterized by myeloid skewing and specific transcriptional-immune covariation patterns. Mendelian randomization identified SBDS as a putatively causal protective factor, where genetically instrumented higher plasma SBDS protein levels were robustly associated with a lower PSC risk (IVW OR = 0.525, 95% CI: 0.356–0.773, P = 0.001). Sensitivity analyses supported the validity of the Mendelian randomization assumptions. ***Conclusions:*** Our study establishes disrupted ribosome homeostasis as a causal pathway in PSC and nominates plasma SBDS as a high-confidence diagnostic biomarker and therapeutic target. The integrative framework provides a generalizable strategy for discovering causal biomarkers in complex diseases.

Citation of this article: Cheng P, Qiang Y, Sun Y, Duan B, Ouyang Y, Li G. A Multi-omics and Machine Learning Framework Identifies Plasma SBDS as a Causal Biomarker and Therapeutic Target in Primary Sclerosing Cholangitis. J Clin Transl Hepatol 2026. doi: 10.14218/JCTH.2025.00676.

Introduction

Primary sclerosing cholangitis (PSC) is a progressive, immune-mediated cholestatic liver disease characterized by chronic inflammation and fibrosis of the intra- and extrahepatic bile ducts. This pathology frequently leads to biliary strictures and cirrhosis and confers a markedly elevated lifetime risk of cholangiocarcinoma.[1,2] A hallmark of PSC is its strong clinical association with inflammatory bowel disease, particularly ulcerative colitis, implicating shared dysregulation of mucosal immunity and gut–liver axis interactions in its pathogenesis.[3] This systemic dimension suggests that key molecular drivers of PSC may extend beyond the hepatic microenvironment and could be accessible in peripheral circulation.

The molecular etiology of PSC remains incompletely defined, a critical knowledge gap that hinders the development of mechanism-based diagnostics and targeted therapies.[4] Current diagnostic paradigms rely primarily on cholangiographic imaging and liver biochemistry, modalities that often identify advanced disease stages and offer limited prognostic insight.[5] Consequently, there is a pressing, unmet need to identify robust molecular biomarkers for early detection, risk stratification, and, crucially, for illuminating causal, druggable biological pathways.

Genomic studies have established a robust genetic architecture for PSC, with numerous risk loci identified through genome-wide association studies (GWAS), many of which are implicated in immune function.[6,7] However, a fundamental challenge lies in translating these statistical genetic associations into actionable biological mechanisms. While expression quantitative trait loci analyses connect genetic variants

Keywords: Primary sclerosing cholangitis; Causal biomarker; Protein quantitative trait locus; pQTL; Mendelian randomization; Immune infiltration.
[#]Contributed equally to this work.
[*]**Correspondence to:** Guangming Li and Yabo Ouyang, Department of General Surgery Center, Beijing YouAn Hospital, Capital Medical University, Beijing Institute of Hepatology, Beijing 100069, China. ORCID: https://orcid.org/0000-0003-3856-5667 (GL). Tel: +86-13501122244 (GL) and +86-18601914629 (YO), E-mail: liguangming@ccmu.edu.cn (GL) and yaboouyang@ccmu.edu.cn (YO)

to gene expression, proteins serve as the primary functional effectors of cellular processes and represent central targets for pharmacotherapy.[8] The recent advent of large-scale plasma protein quantitative trait locus (pQTL) maps provides a transformative resource, enabling the direct linkage of genetic variation to circulating protein levels, a proximal and clinically actionable molecular layer for causal inference.[9]

Mendelian randomization (MR) employs genetic variants as instrumental variables to infer causal relationships between an exposure and an outcome, largely circumventing confounding and reverse causation biases inherent in observational studies.[10] Integrating pQTL data with disease GWAS through MR methodologies offers a powerful and systematic framework for identifying circulating proteins with putative causal roles in disease pathogenesis.[11]

Here, we outline a multi-stage integrative framework designed to systematically discover and causally validate protein biomarkers for PSC. Our approach begins by leveraging bioinformatic and machine learning analyses of disease-relevant transcriptomic data to nominate a robust set of high-confidence candidate proteins. The central aim was then to rigorously test the causal role of these prioritized candidates in PSC pathogenesis. To achieve this, we employed two-sample MR, integrating large-scale plasma pQTL data with the latest PSC GWAS summary statistics. This comprehensive strategy was designed to bridge the gap between associative findings and causal mechanisms, with the ultimate goal of identifying high-confidence diagnostic biomarkers and revealing novel therapeutic targets for future validation.

Methods

Study design and analytical framework

We designed an integrated computational framework to systematically discover plasma protein biomarkers with a putative causal role in PSC. The analytical pipeline progressed through three sequential, hypothesis-driven phases to translate associative signals into causal evidence: (I) Transcriptomic discovery: Identification of the PSC-associated transcriptional landscape from public gene expression datasets. (II) Candidate prioritization: Refinement of candidates via network biology and machine learning to construct a minimal, robust diagnostic signature. (III) Causal inference: Integration of plasma pQTL with PSC GWAS data using MR to establish genetic evidence for causality. A schematic overview of the workflow is presented in Figure 1.

Acquisition and processing of omics data

Transcriptomic data: Gene expression datasets from individuals with PSC and matched healthy controls were obtained from the Gene Expression Omnibus. A primary training cohort was assembled by merging peripheral blood transcriptomic profiles from two independent studies (GSE119600 and GSE144521). For external validation, two distinct testing cohorts were used: one comprising liver tissue transcriptomes (GSE159676) and another based on urinary cell transcriptomes (derived from the urinary fraction of GSE144521). All expression data were $\log_2$-transformed and normalized across samples using the normalizeBetweenArrays function in the limma R package.[12] Potential batch effects arising from dataset merging were corrected using the ComBat algorithm from the sva R package.[13]

Genetic and proteomic data for causal inference: For the MR analysis, we used summary-level genetic data from two sources. The genetic association data for the outcome (PSC risk) were sourced from the largest publicly available PSC GWAS (OpenGWAS ID: ieu-a-1112). Genetic instruments for protein exposures were derived from a large-scale plasma proteome-wide association study (pQTL mapping) by Ferkingstad et al.[9]

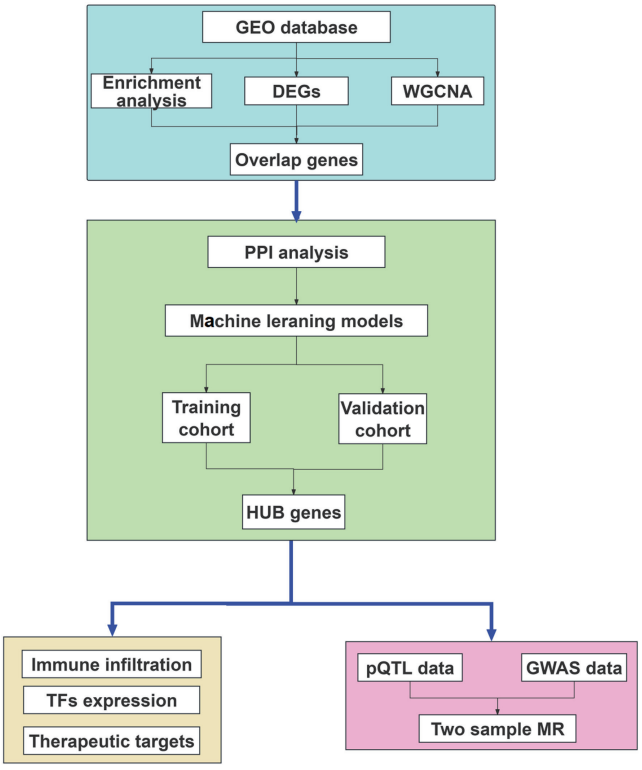

Fig. 1. Flowchart of the study. GEO, gene expression omnibus; DEGs, differentially expressed genes; WGCNA, weighted gene co-expression network analysis; PPI, protein-protein interaction; TFs, transcription factors; pQTL, plasma protein quantitative trait loci; GWAS, genome-wide association study; MR, Mendelian randomization.

Identification of PSC-associated transcriptional programs

Differential expression and co-expression network analysis: Differentially expressed genes (DEGs) between PSC and control samples in the training cohort were identified using the limma package, applying a significance threshold of absolute $\log_2$ fold change > 0.379 and a false discovery rate (FDR) < 0.05. Weighted gene co-expression network analysis (WGCNA) was performed on the DEGs to identify modules of highly correlated genes.[14] Module-trait relationships were assessed by correlating module eigengenes with the PSC phenotype. Modules with a significant correlation ($|r| > 0.5$, $P < 0.01$) were retained for downstream analysis.

Functional enrichment analysis: Genes from the PSC-significant WGCNA modules were subjected to functional annotation using Gene Ontology biological process pathway analyses, performed with the clusterProfiler R package.[15]

Prioritization of core biomarker candidates

Protein-protein interaction (PPI) network and hub gene identification: A PPI network for candidate genes from significant modules was constructed using the STRING database (minimum interaction confidence score > 0.7) and visualized in Cytoscape.[16,17] Network topology was analyzed

using the cytoHubba plugin. Hub genes were identified by ranking nodes based on the maximal clique centrality algorithm.[18]

Machine learning-based diagnostic signature development: The identified hub genes were further refined using a multi-algorithm machine learning pipeline to define a minimal, high-performance diagnostic signature.

Feature selection: Four distinct algorithms proficient in variable selection, namely LASSO regression, random forest (RF), stepwise generalized linear model, and gradient boosting with component-wise linear models, were applied to rank gene importance. The intersection and consensus from these methods produced a refined candidate gene subset.

Predictive modeling: The predictive power of this subset was evaluated using an extensive suite of twelve machine learning classifiers, including regularization methods (Ridge, ElasticNet), ensemble learners (GBM, XGBoost, RF), and others (SVM, LDA, Naïve Bayes). Models were built and hyperparameter-tuned using a stratified 10-fold cross-validation framework on the training cohort. This process generated 113 unique model-gene set combinations. Performance was evaluated by the area under the receiver operating characteristic curve (AUC) in both training and independent validation cohorts. The final optimal model and its constituent genes were selected based on achieving the highest mean AUC across cohorts.[19]

Model interpretation: To ensure interpretability, SHapley additive explanations (SHAP) analysis was employed on the optimal model.[20] SHAP provided both global feature importance rankings and local explanations for individual predictions, quantifying the contribution of each gene to the model's output and enhancing clinical translatability.

Biological characterization and validation

Diagnostic performance and correlation: The diagnostic utility of individual and combined signature genes was quantified by calculating the AUC using the pROC package.[21] Pairwise Spearman correlations among signature genes were assessed to infer potential co-regulation.

Immune microenvironment analysis: The relative proportions of 22 immune cell types in the liver tissue transcriptomes were estimated using CIBERSORTx with the LM22 signature matrix.[22] Associations between signature gene expression and immune cell infiltration levels were evaluated using Spearman correlation.

Transcriptional regulator inference: Potential upstream transcriptional regulators of the signature genes were predicted using the hTFtarget database (https://guolab.wchscu.cn/hTFtarget). Correlations between the expression levels of predicted transcription factors and their target signature genes were analyzed within the PSC cohort to suggest plausible in vivo regulatory relationships.[23]

MR for causal inference

A comprehensive two-sample MR framework was implemented to test for a causal effect of genetically predicted plasma protein levels on PSC risk.[24]

Instrument selection: Independent (linkage disequilibrium $r^2 < 0.001$ within a 10,000 kb window), genome-wide significant ($P < 5 \times 10^{-8}$) pQTLs for each protein were selected as instrumental variables.

Data harmonization: Effect alleles and estimates for exposure (pQTL) and outcome (PSC GWAS) datasets were aligned using the harmonise_data function from the TwoSampleMR R package.[25]

Primary causal estimation: The inverse variance weighted (IVW) method under a multiplicative random-effects model served as the primary analysis. Statistical significance was set at a two-sided $P < 0.05$. The effect size is reported as an odds ratio (OR) per one-standard-deviation increase in genetically predicted protein level.[26]

Complementary and sensitivity analyses: To ensure robustness, four supplementary MR methods (MR-Egger, weighted median, simple mode, weighted mode) were applied.[27–30] Sensitivity analyses included: 1) Cochran's Q test for heterogeneity ($P > 0.05$ preferred); 2) MR-Egger intercept test for directional pleiotropy ($P > 0.05$ preferred); 3) leave-one-out analysis to identify influential variants; and 4) MR-Steiger directionality test to confirm the assumed causal direction.[31]

Exploration of therapeutic potential

The druggability of the proteins identified as causally associated with PSC was explored by querying the DGIdb database, which catalogs known and predicted interactions between genes and drugs.[32]

Statistical implementation

All analyses were performed in R (version 4.5.2). Multiple testing corrections were applied using the Benjamini–Hochberg method to control FDR where appropriate. Statistical significance was defined as a two-sided $P < 0.05$ or an FDR-adjusted $q < 0.05$.

Results

Identification of DEGs and co-expression modules in PSC

Comparative transcriptomic analysis of the PSC training cohort versus healthy controls identified 514 DEGs ($|\log_2 FC| > 0.379$, FDR < 0.05) (Fig. 2A). Hierarchical clustering based on these DEGs demonstrated clear separation between PSC and control samples (Fig. 2B). WGCNA of these DEGs resolved several distinct co-expression modules (Fig. 2C). Module–trait relationship analysis identified multiple modules significantly correlated with the PSC phenotype, including MEpurple (r = 0.25, P = 0.009), MEbrown (r = 0.37, P = 8e-5), and MEblack (r = 0.42, P = 8e-6) (Fig. 2D). The MEblack module, demonstrating the strongest association, was selected for downstream analysis, comprising 159 high-confidence PSC-related candidate genes.

Functional enrichment implicates ribosome biogenesis and protein homeostasis

Functional enrichment analysis of the candidate genes from the MEblack module revealed significant terms related to ribosome biogenesis, cytoplasmic translation, and protein homeostasis (Fig. 3A and B). Gene Ontology analysis across biological process, cellular component, and molecular function categories highlighted prominent clusters including "ribosome biogenesis," "cytoplasmic translation," "structural constituent of ribosome," "protein stabilization," and "unfolded protein binding" (all FDR < 0.05) (Fig. 3C).

An integrated network and machine learning pipeline identifies a core diagnostic signature

The PPI network constructed from the candidate genes yielded 21 topologically central hub genes (Fig. 4A). These were refined using a multi-algorithm machine learning framework. Among 113 model combinations evaluated via stratified 10-fold cross-validation, an RF classifier achieved the highest and most generalizable performance, with a

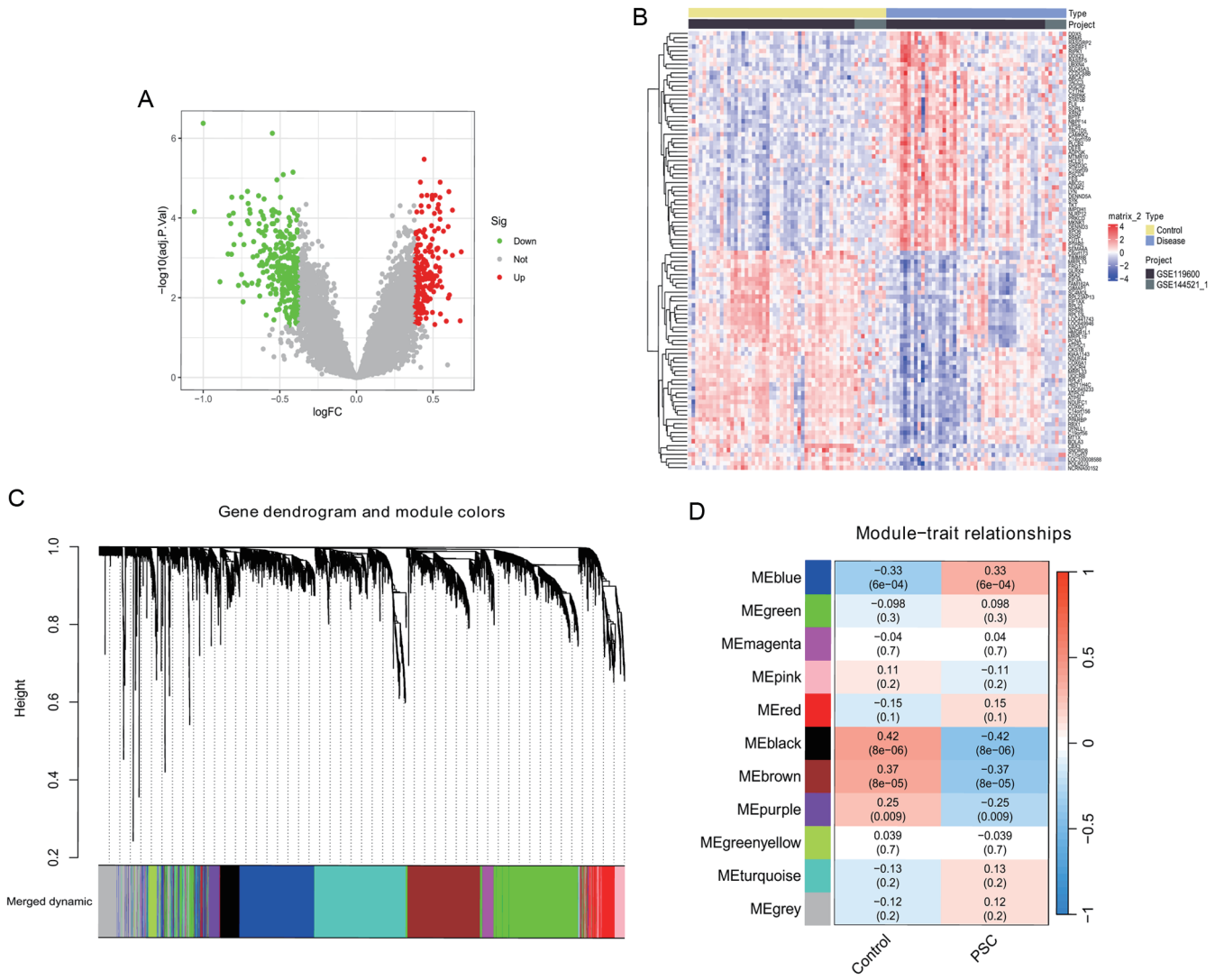

Fig. 2. DEGs and weighted gene co-expression network analysis (WGCNA) in PSC. (A) Volcano plot displaying DEGs between PSC samples and healthy controls; (B) Hierarchical clustering heatmap of the 514 DEGs; (C) Gene dendrogram (top) and module color assignment (bottom) from WGCNA; (D) Module–trait correlation heatmap. DEGs, differentially expressed genes; WGCNA, weighted gene co-expression network analysis; PSC, primary sclerosing cholangitis.

mean AUC of 0.908 across training and independent validation cohorts (Fig. 4B). In the training cohort, individual candidate genes showed diagnostic AUCs ranging from 0.728 to 0.792 (Fig. 4C).

Feature importance analysis within the optimal RF model identified a nine-gene diagnostic signature: PTMA, SUMO1, Shwachman-Bodian-Diamond syndrome (SBDS), RPL7, EIF1AX, ANP32A, PCNA, FAM98A, and MPHOSPH6. SHAP analysis quantified their contributions, identifying SUMO1 (mean |SHAP| = 0.0576) and PTMA (0.0519) as the most influential predictors (Fig. 4D and E). SHAP dependence plots illustrated the directional impact of each feature on the model's prediction of PSC risk (Fig. 4F). A detailed force analysis for a representative sample showed SUMO1 and SBDS as primary negative regulators, substantially lowering the prediction score below the baseline expectation (Fig. 4G).

Consistent with their diagnostic role, all nine signature genes were significantly downregulated in PSC samples versus controls (Wilcoxon test, $P < 0.05$) (Fig. 4H). Strong positive pairwise co-expression correlations among these genes (e.g., SUMO1 vs. SBDS: $r = 0.85$, $P < 0.001$) suggested functional co-regulation within the PSC state (Fig. 4I).

The PSC immune microenvironment exhibits myeloid skewing and altered transcriptional–immune coupling

Deconvolution of the peripheral immune cell composition revealed profound remodeling in PSC, characterized by a shift toward myeloid expansion and lymphoid contraction (Fig. 5A).

Quantitatively, we observed a significant reduction in $CD8^+$ T cells, γδ T cells, and activated NK cells ($P < 0.05$), alongside a marked expansion of neutrophils ($P < 0.01$) (Fig. 5B).

Integrated correlation analysis uncovered that expression of the nine-gene signature was selectively coupled to specific immune cell covariance modules, a pattern we term "transcriptional-immune circuit aliasing" (Fig. 5C). For instance, PCNA expression correlated positively with a $CD8^+$ T cell/

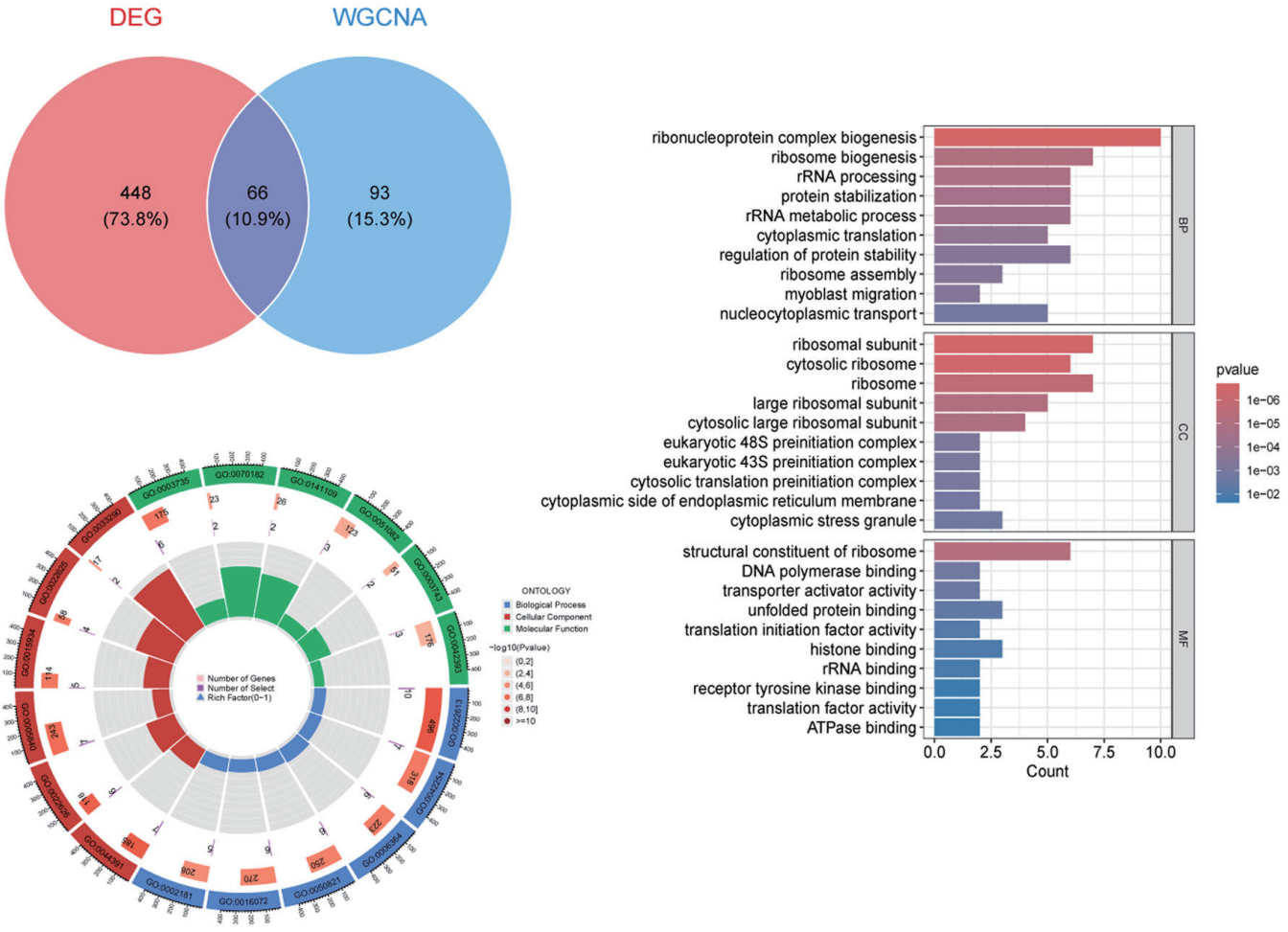

Fig. 3. Functional enrichment analysis of MEblack module candidate genes. (A) Venn diagram showing the overlap between DEGs and genes in the MEblack module (from WGCNA); (B) Bar plot of GO enrichment terms across three categories: BP, CC, and MF; (C) Circular plot summarizing GO enrichment results. ME, module eigengene; DEGs, differentially expressed genes; WGCNA, weighted gene co-expression network analysis; GO, gene ontology; BP, biological process; CC, cellular component; MF, molecular function.

resting mast cell module but negatively with a neutrophil/M0 macrophage module. PTMA aligned with a monocyte/resting mast cell axis while anti-correlating with a neutrophil/plasma cell axis. SBDS expression was inversely linked to a γδ T cell/M2 macrophage module.

Inference of a dysregulated transcriptional regulatory network

Analysis of predicted transcription factor–hub gene interactions within the PSC cohort revealed distinct regulatory patterns (Fig. 6A). POU2F1 emerged as a central negative regulator, showing strong inverse correlations with multiple hub genes including RPL7 ($r = -0.56$, $P < 0.001$), EIF1AX ($r = -0.52$, $P < 0.001$), PCNA ($r = -0.53$, $P < 0.001$), and SBDS ($r = -0.43$, $P < 0.01$). In contrast, LMO2 exhibited significant positive correlations with ANP32A ($r = 0.49$, $P < 0.001$), SUMO1 ($r = 0.48$, $P < 0.001$), and SBDS ($r = 0.29$, $P < 0.01$).

A reconstructed regulatory interaction network positioned ANP32A as a major hub, densely connected to several repressive TFs (POU2F1, IRF4, MYB, EBF1), suggesting it may be a focal point for concerted transcriptional dysregulation (Fig. 6B). The network also highlighted a specific regulatory axis for SBDS, potentially under the control of NFIC.

Drug repurposing analysis highlights potential therapeutic agents

Drug enrichment analysis targeting the hub genes identified several pharmacological agents with significant associations ($P < 0.01$), suggesting repurposing potential (Fig. 6C and D). Notable findings included the association of the topoisomerase I inhibitor topotecan and the DNA polymerase inhibitor aphidicolin with PCNA, implicating DNA replication pathways. Agents with anti-inflammatory or redox-modulatory properties, such as hesperetin (linked to PTMA and ANP32A) and dihydroergotamine (associated with EIF1AX and ANP32A), were also enriched. Compounds with hepatobiliary relevance, including epirubicin hydrochloride (SUMO1) and the tyrosine kinase inhibitor vandetanib (PTMA), were identified, alongside antibiotics and NSAIDs such as cefixime (SUMO1) and indomethacin (EIF1AX).

MR identified plasma SBDS as a causal protective factor in PSC

Two-sample MR analysis was performed to assess the causal relationship between genetically predicted plasma levels of

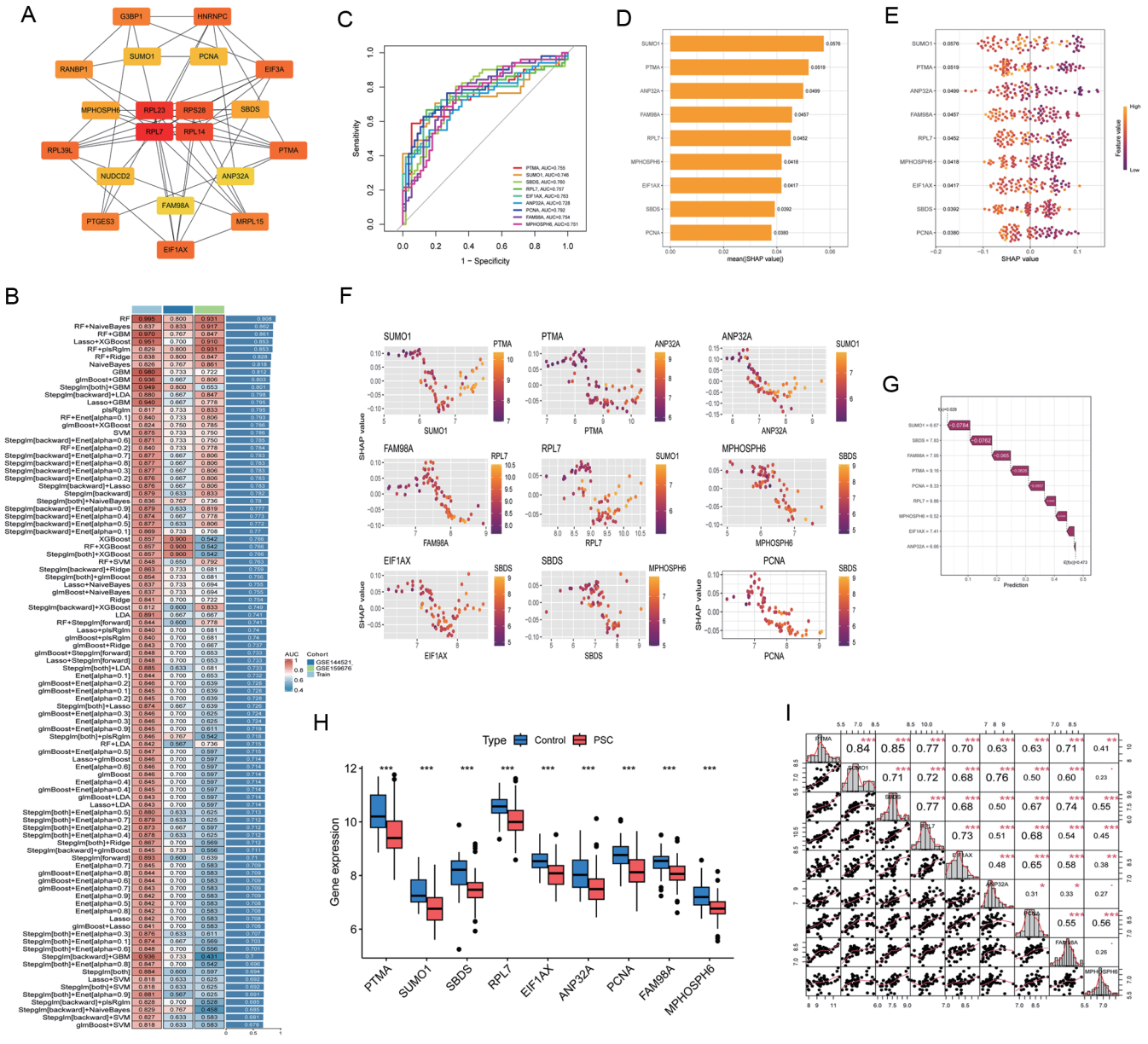

Fig. 4. Construction of a PSC diagnostic gene signature via network analysis and machine learning. (A) PPI network of MEblack module candidate genes; (B) Heatmap of AUC values for 113 machine learning model combinations (stratified 10-fold cross-validation); (C) ROC curves for individual candidate genes; (D) Feature importance (mean absolute SHAP value) of genes in the optimal RF model; (E) SHAP summary plot showing the impact of each signature gene on PSC risk prediction; (F) SHAP dependence plots illustrating the directional effect of each signature gene's expression on the model's prediction score (PSC risk); (G) SHAP force plot for a representative sample; (H) Boxplots of signature gene expression in control (blue) and PSC (red) samples; (I) Pairwise co-expression correlation matrix of the nine-gene signature. PSC, primary sclerosing cholangitis; PPI, protein-protein interaction; ME, module eigengene; AUC, area under the curve; ROC, receiver operating characteristic; SHAP, SHapley additive exPlanations; RF, random forest. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

the signature proteins and PSC risk. SBDS protein demonstrated a significant causal association with PSC, and four conditionally independent SNPs (rs1354034, rs138783789, rs74490291, rs79344818; pairwise $r^2 < 0.01$) served as instrumental variables.

Genetically predicted higher plasma SBDS levels demonstrated a protective causal effect on PSC risk, with an OR of 0.525 (95% CI: 0.356–0.773; P = 0.001) using the IVW method (Fig. 7A). A scatter plot of SNP-specific effects showed consistent directional estimates aligning with the IVW and weighted-median regression lines (Fig. 7B).

This causal inference was robust to sensitivity analyses. We found no significant heterogeneity among instrument-specific estimates (Cochran's Q = 1.063, P = 0.786) and no evidence of directional pleiotropy (MR-Egger intercept = 0.028, P = 0.616) (Fig. 7C and D). Steiger directionality testing confirmed the assumed causal direction from SBDS to PSC risk (P = 0.007). Leave-one-out analysis confirmed that the association was not driven by any single influential SNP (Fig. 7E). These results provide robust genetic evidence that elevated plasma SBDS is causally associated with a reduced risk of PSC.

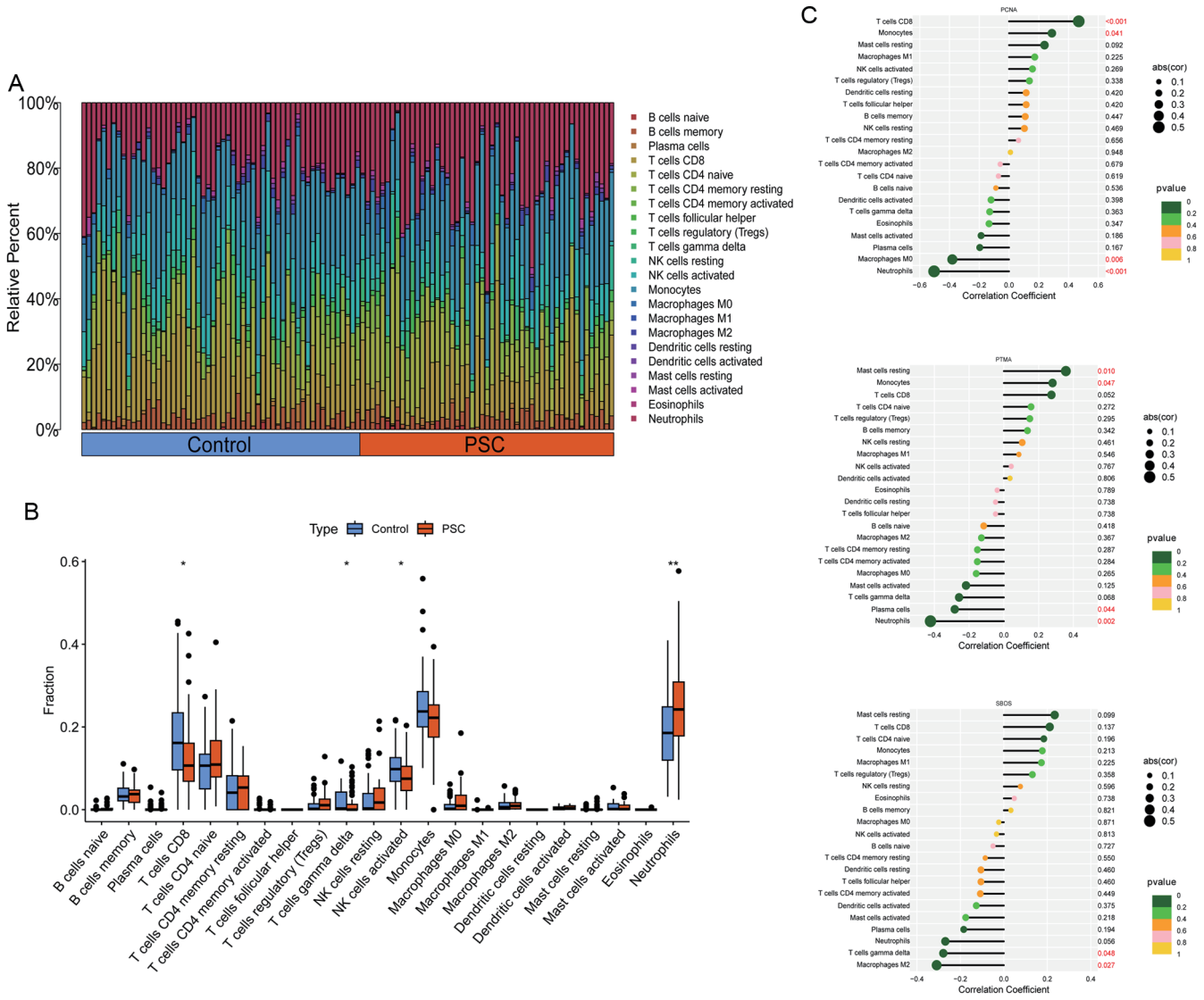

Fig. 5. Immune microenvironment remodeling and transcriptional–immune coupling in PSC. (A) Stacked bar plot of relative peripheral immune cell fractions in control (blue) and PSC (orange) samples; (B) Boxplots comparing immune cell fractions between control (blue) and PSC (red) samples; (C) Correlation plots of the nine-gene signature with immune cell covariance modules (grouped by cell type). PSC, primary sclerosing cholangitis. *P < 0.05, **P < 0.01.

Discussion

In this study, we developed and applied an integrated multi-omics framework to delineate the causal molecular architecture of PSC. This systematic strategy culminated in the identification of a robust, peripherally accessible gene expression signature for disease diagnosis. Furthermore, it provided genetic evidence nominating the protein SBDS as a putatively causal protective factor, wherein genetically predicted higher plasma SBDS levels were associated with reduced disease risk. Collectively, these findings advance a novel mechanistic understanding of PSC, centered on ribosome homeostasis, and establish a direct translational pathway for biomarker development and therapeutic target identification.

The nine-gene diagnostic signature (PTMA, SUMO1, SBDS, RPL7, EIF1AX, ANP32A, PCNA, FAM98A, MPHOSPH6), refined from a broader candidate set through a stringent machine learning pipeline, constitutes a functionally cohesive molecular network whose activities converge on ribosome biogenesis, translational control, and nucleic acid metabolism. These interconnected processes are fundamental to cellular homeostasis and the adaptive stress response, with their collective dysregulation forming a compelling pathogenic axis in chronic inflammatory and fibrotic diseases.[33,34]

The most salient functional theme uniting this signature is ribosome biology. Central to this is SBDS, which encodes a key assembly factor essential for the maturation of the 60S ribosomal subunit and the maintenance of translational fidelity; its deficiency is a known cause of ribosomopathies.[35,36] This core function is supported by RPL7, a canonical structural component of the 60S subunit,[37] and EIF1AX, a crucial translation initiation factor.[38] Their coordinated downregulation in PSC (Fig. 4H) strongly implies a systemic impairment of global protein synthesis capacity. Such a proteostatic deficit could underpin cellular dysfunction and heightened sensitivity to injury within the cholangiocyte microenvironment, a

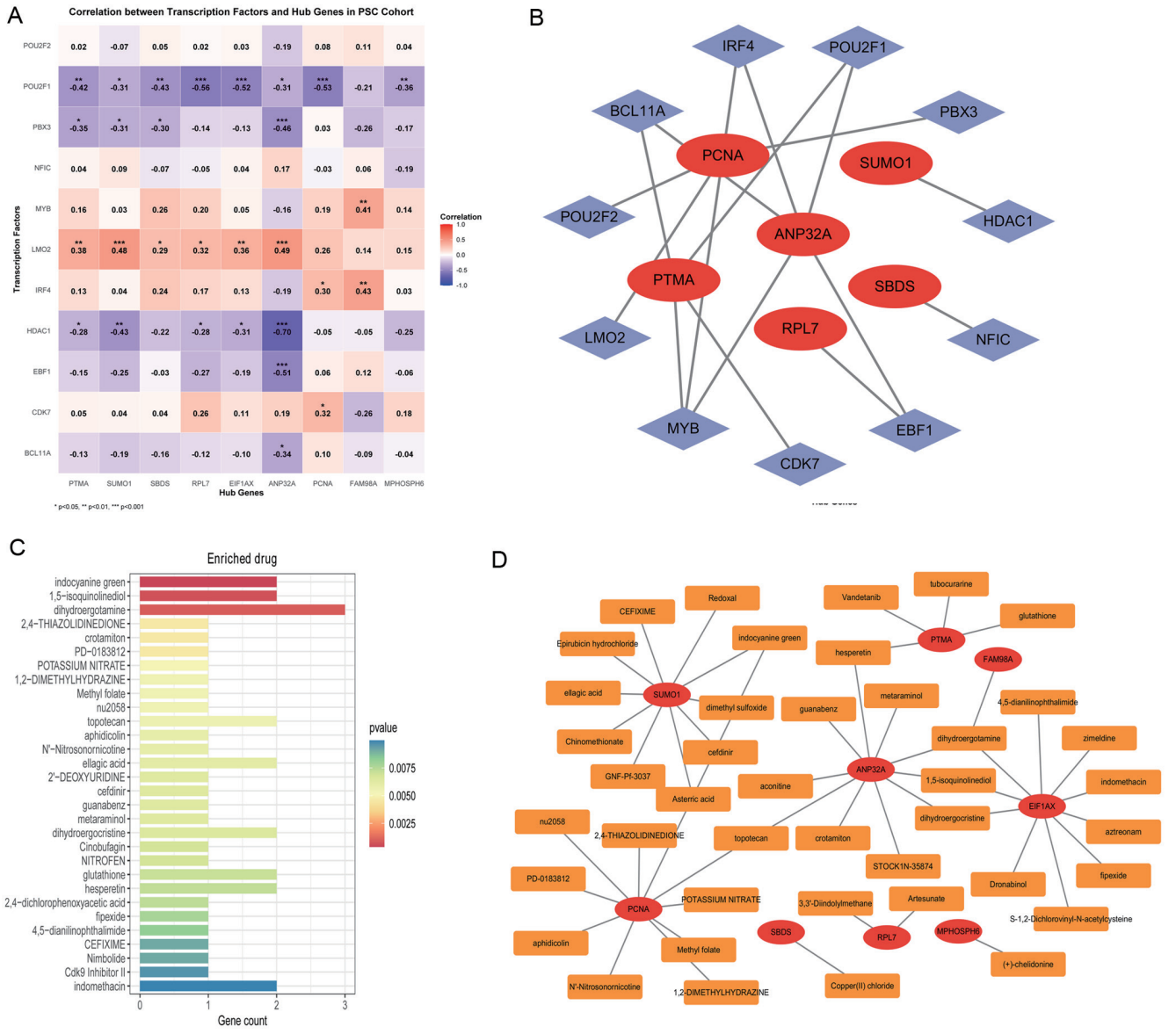

Fig. 6. Transcriptional regulatory network and drug repurposing in PSC. (A) Correlation matrix of TFs and hub genes in the PSC cohort; (B) Reconstructed transcriptional regulatory network; (C) Drug enrichment bar plot; (D) Drug–hub gene interaction network. PSC, primary sclerosing cholangitis; TFs, transcription factors.

mechanism implicated in other forms of liver injury.[34]

This perturbation in core translational machinery is further modulated by signature genes governing post-translational modification and nuclear chaperone function. SUMO1 mediates SUMOylation, a dynamic post-translational modification critical for regulating protein stability, localization, and activity, with particular importance for factors involved in DNA damage and stress responses.[39] Concurrently, ANP32A and PTMA function as histone chaperones and acidic nuclear phosphoproteins deeply involved in chromatin remodeling, transcription, and apoptosis.[40,41] The dysregulation of this suite of nuclear regulators points to a concomitant disruption of nuclear homeostasis and gene expression control, potentially locking cells into a maladaptive stress state that exacerbates inflammation and hampers repair.

Expanding beyond translational and nuclear control, the signature also encompasses direct modulators of cell cycle progression and proliferative capacity. PCNA acts as an essential processivity factor for DNA polymerases, serving as a central platform that coordinates DNA replication and repair.[42] MPHOSPH6 participates in mRNA export and cell cycle regulation, while FAM98A contributes to the assembly of protein complexes and methyltransferase activity.[43] The coordinated downregulation of these pro-growth and pro-synthesis factors within the signature may indicate a fundamental rewiring of cellular physiology in PSC. This could represent a state of proliferative arrest, which may be a consequence of sustained inflammatory damage, or it may alternatively reflect an adaptive cellular strategy to mitigate stress by reducing metabolic and biosynthetic demand.[34]

While the nine-gene signature as a whole provides a powerful diagnostic tool through the combined interpretation of its gene expression levels, our subsequent MR analysis allowed us to probe the individual causal roles of

A

exposure	nsnp	method	pval	OR(95% CI)
SBDS	4	MR Egger	0.129	0.454 (0.246 to 0.841)
	4	Weighted median	0.005	0.523 (0.332 to 0.823)
	4	Inverse variance weighted	0.001	0.525 (0.356 to 0.773)
	4	Simple mode	0.084	0.443 (0.237 to 0.829)
	4	Weighted mode	0.116	0.534 (0.305 to 0.935)

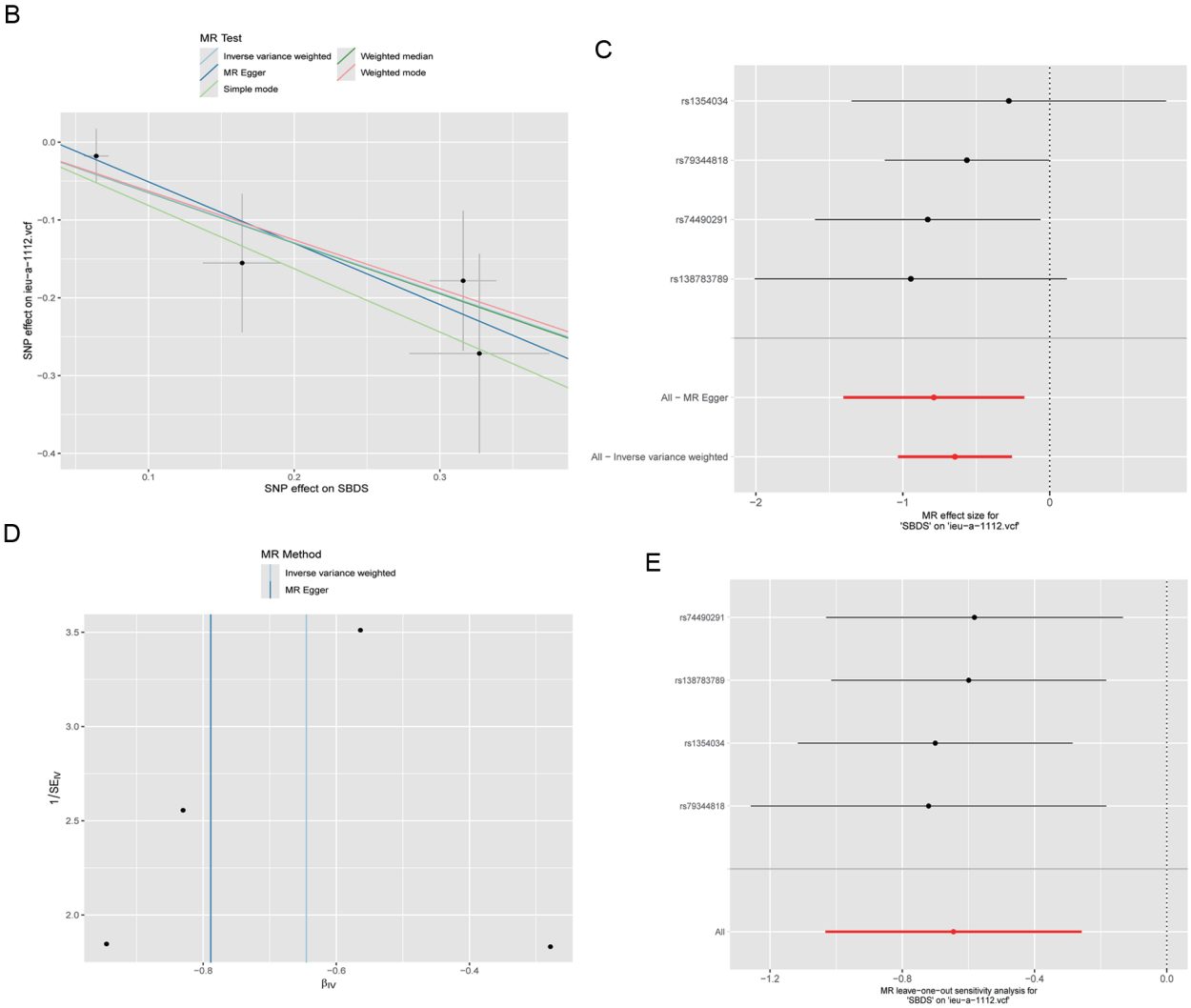

Fig. 7. Two-sample MR analysis of plasma SBDS and PSC risk. (A) MR effect estimates for the causal association between genetically predicted plasma SBDS levels and PSC risk; (B) Scatter plot of SNP-specific effects; (C) Heterogeneity analysis; (D) MR-Egger intercept plot; (E) Leave-one-out sensitivity analysis. MR, Mendelian randomization; PSC, primary sclerosing cholangitis; SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval; nsnp, number of single nucleotide polymorphisms.

its components at the protein level to uncover underlying disease mechanisms. It is crucial to distinguish the role of the multi-gene panel in achieving high diagnostic accuracy from the role of SBDS as a single, causally implicated factor for mechanistic understanding and therapeutic targeting.

Among these components, SBDS emerges as the central and causally validated hub. Our MR analysis, which focused on genetically predicted higher plasma levels of SBDS protein, provides the first genetic evidence that these higher levels are causally protective against PSC (IVW OR = 0.525, P =

0.001). This elevates SBDS from a merely correlative biomarker to a putative upstream regulator within the pathogenic network, highlighting its potential as a therapeutic target. We propose that a deficit in SBDS function, potentially arising from genetic predisposition, inflammatory inhibition, or other mechanisms, could initiate a molecular cascade. This cascade would be characterized by ribosomal stress, triggering the coordinated downregulation of associated translational machinery (RPL7, EIF1AX). This primary defect could subsequently dysregulate downstream processes vital for protein stabilization (SUMO1, ANP32A, PTMA) and cellular replication (PCNA). Consequently, the SBDS-centric signature likely delineates a critical pathogenic axis in PSC, wherein compromised ribosome homeostasis acts as a nexus, linking fundamental cellular stress to the overt inflammatory and fibrotic disease phenotypes.[36,44]

The immune microenvironment in PSC, marked by prominent neutrophil expansion and a relative contraction of $CD8^+$ T cells, is consistent with established reports of dominant innate immunity in this disease.[45] Our analysis uncovers a deeper layer of dysregulation, which we term "transcriptional-immune circuit aliasing." This phenomenon describes the selective coupling between the expression of key signature genes (e.g., PCNA, PTMA, and SBDS) and specific covariance modules of immune cells.[22] This finding suggests that the diagnostic signature is not merely a passive reflection but is functionally embedded within, and may actively participate in regulating, the dysregulated immune crosstalk characteristic of PSC. For instance, the inverse correlation observed between SBDS expression and a module comprising γδ T cells and M2 macrophages points to a potential role for this ribosome assembly factor in modulating pro-fibrotic or immunosuppressive immune axes, a hypothesis that merits dedicated experimental inquiry.[46]

The translational implications derived from our integrative analysis are significant and fall into two distinct yet complementary categories. First, for diagnostic purposes, the nine-gene expression signature as a whole presents a compelling candidate for a minimally invasive, blood-based molecular assay. The combined predictive power of these nine markers, reflecting a functionally cohesive molecular network, underpins its high diagnostic accuracy (AUC = 0.908) for distinguishing or stratifying PSC patients. Second, on a mechanistic and therapeutic level, our genetic causal inference provides robust evidence for SBDS's role. This elevates SBDS beyond its function as a component of the diagnostic panel to a high-confidence therapeutic target for intervention, wherein genetically predicted higher protein levels are causally protective. This distinction is vital: the signature offers a practical tool for diagnosis, while SBDS provides a specific, causally validated target for therapeutic intervention. Supporting this direction, our in silico drug repositioning analysis identified several existing compounds with favorable safety profiles that interact with signature gene products, indicating tangible repurposing avenues. Notable among these are the flavonoid hesperetin, associated with PTMA and ANP32A, and chemotherapeutic agents with known hepatobiliary distribution such as epirubicin, linked to SUMO1.[32] Consequently, a pivotal future direction involves delineating the precise mechanistic role of SBDS in cholangiocyte pathobiology and immune cell function, coupled with screening for pharmacological agents capable of augmenting its expression or activity. This approach aims to restore ribosome homeostasis, representing a novel and mechanistically grounded therapeutic strategy for PSC.[44]

Several limitations merit consideration. First, while MR strongly supports causality, functional validation of SBDS in relevant in vitro and in vivo models is essential to confirm its mechanistic role in bile duct inflammation and fibrosis. Second, the primary data sources (GWAS, pQTL) are predominantly from European ancestry populations, which limits the generalizability of our findings across diverse ethnic groups.[47] Future studies should incorporate multi-ancestry cohorts. Third, our analysis focused on plasma pQTLs; integrating protein QTL data from liver tissue or single-cell contexts could provide more tissue-specific mechanistic insights.[48] Finally, prospective validation in independent, clinically well-characterized cohorts is needed to assess the real-world utility of the diagnostic signature.

Conclusions

Our integrated multi-omics and machine learning framework makes two key, interconnected contributions. First, it establishes a robust nine-gene expression signature with high accuracy for PSC diagnosis and stratification, offering a readily accessible molecular tool. Second, it profoundly advances our mechanistic understanding by establishing disrupted ribosome homeostasis as a causal pathway in PSC, specifically nominating plasma SBDS protein as a key protective factor and a high-priority therapeutic target through genetic causal inference. This work therefore not only provides a powerful biomarker panel for clinical use but also moves beyond genetic associations to reveal a druggable, causal mechanism. The strategy outlined here provides a generalizable blueprint for translating complex disease genetics into causal biomarkers and mechanistic therapeutic hypotheses.

Acknowledgments

We express our deep thanks to the developers of the Figdraw mapping software.

Funding

This study was supported by the Noncommunicable Chronic Diseases–National Science and Technology Major Project (2023ZD0502400).

Conflict of interest

The authors have no conflict of interests related to this publication.

Author contributions

Provided direction and guidance throughout the preparation of this manuscript (GL, YO, BD), collected and interpreted the studies, and contributed to the writing, editing, and revision of the manuscript (PC, YQ, YS). All authors have read and approved the final version and publication of the article.

Ethical statement

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2024). All data used in this study were obtained from publicly available databases. Since the study did not involve human or animal subjects, approval from the Institutional Review Board and informed written consent were waived.

Data sharing statement

The original contributions presented in the study are included

in the article. Further inquiries can be directed to the corresponding authors.

References

[1] Karlsen TH, Folseraas T, Thorburn D, Vesterhus M. Primary sclerosing cholangitis - a comprehensive review. J Hepatol 2017;67(6):1298–1323. doi:10.1016/j.jhep.2017.07.022, PMID:28802875.

[2] Schrumpf E, Boberg KM. Epidemiology of primary sclerosing cholangitis. Best Pract Res Clin Gastroenterol 2001;15(4):553–562. doi:10.1053/bega.2001.0204, PMID:11492967.

[3] Tabibian JH, O'Hara SP, Larusso NF. Primary sclerosing cholangitis: the gut-liver axis. Clin Gastroenterol Hepatol 2012;10(7):819, author reply 819-819; author reply 820. doi:10.1016/j.cgh.2012.01.024, PMID:22343691.

[4] Björnsson ES, Kalaitzakis E. Recent advances in the treatment of primary sclerosing cholangitis. Expert Rev Gastroenterol Hepatol 2021;15(4):413–425. doi:10.1080/17474124.2021.1860751, PMID:33283566.

[5] European Association for the Study of the Liver. EASL Clinical Practice Guidelines on sclerosing cholangitis. J Hepatol 2022;77(3):761–806. doi:10.1016/j.jhep.2022.05.011, PMID:35738507.

[6] Ji SG, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, *et al*. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nat Genet 2017;49(2):269–273. doi:10.1038/ng.3745, PMID:27992413.

[7] Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, Doncheva NT, *et al*. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nat Genet 2013;45(6):670–675. doi:10.1038/ng.2616, PMID:23603763.

[8] Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, Wörheide MA, *et al*. Mapping the proteo-genomic convergence of human diseases. Science 2021;374(6569):eabj1541. doi:10.1126/science.abj1541, PMID:34648354.

[9] Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. Nat Genet 2021;53(12):1712–1721. doi:10.1038/s41588-021-00978-w, PMID:34857953.

[10] Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et *al*. Mendelian randomization. Nat Rev Methods Primers 2022;2:6. doi:10.1038/s43586-021-00092-5, PMID:37325194.

[11] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et *al*. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481–487. doi:10.1038/ng.3538, PMID:27019110.

[12] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et *al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43(7):e47. doi:10.1093/nar/gkv007, PMID:25605792.

[13] Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform 2020;2(3):lqaa078. doi:10.1093/nargab/lqaa078, PMID:33015620.

[14] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559. doi:10.1186/1471-2105-9-559, PMID:19114008.

[15] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et *al*. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141, PMID:34557778.

[16] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et *al*. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51(D1):D638–D646. doi:10.1093/nar/gkac1000, PMID:36370105.

[17] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et *al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–2504. doi:10.1101/gr.1239303, PMID:14597658.

[18] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014;8(Suppl 4):S11. doi:10.1186/1752-0509-8-S4-S11, PMID:25521941.

[19] Eley A, Hlaing TT, Breininger D, Helforoush Z, Kachouie NN. Monte Carlo Gradient Boosted Trees for Cancer Staging: A Machine Learning Approach. Cancers (Basel) 2025;17(15):2452. doi:10.3390/cancers17152452, PMID:40805154.

[20] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et *al*. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell 2020;2(1):56–67. doi:10.1038/s42256-019-0138-9, PMID:32607472.

[21] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et *al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. doi:10.1186/1471-2105-12-77, PMID:21414208.

[22] Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et *al*. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37(7):773–782. doi:10.1038/s41587-019-0114-2, PMID:31061481.

[23] Zhang Q, Liu W, Zhang HM, Xie GY, Miao YR, Xia M, *et al*. hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. Genomics Proteomics Bioinformatics 2020;18(2):120–128. doi:10.1016/j.gpb.2019.09.006, PMID:32858223.

[24] Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet 2014;23(R1):R89–R98. doi:10.1093/hmg/ddu328, PMID:25064373.

[25] Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, *et al*. The MR-Base platform supports systematic causal inference across the human phenome. Elife 2018;7:e34408. doi:10.7554/eLife.34408, PMID:29846171.

[26] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37(7):658–665. doi:10.1002/gepi.21758, PMID:24114802.

[27] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med 2017;36(11):1783–1802. doi:10.1002/sim.7221, PMID:28114746.

[28] Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32(5):377–389. doi:10.1007/s10654-017-0255-x, PMID:28527048.

[29] Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol 2017;46(6):1985–1998. doi:10.1093/ije/dyx102, PMID:29040600.

[30] Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol 2017;46(6):1734–1739. doi:10.1093/ije/dyx034, PMID:28398548.

[31] Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet 2017;13(11):e1007081. doi:10.1371/journal.pgen.1007081, PMID:29149188.

[32] Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, *et al*. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res 2021;49(D1):D1144–D1151. doi:10.1093/nar/gkaa1084, PMID:33237278.

[33] Jiao L, Liu Y, Yu XY, Pan X, Zhang Y, Tu J, *et al*. Ribosome biogenesis in disease: new players and therapeutic targets. Signal Transduct Target Ther 2023;8(1):15. doi:10.1038/s41392-022-01285-4, PMID:36617563.

[34] Malhi H, Kaufman RJ. Endoplasmic reticulum stress in liver disease. J Hepatol 2011;54(4):795–809. doi:10.1016/j.jhep.2010.11.005, PMID:21145844.

[35] Bezzerri V, Pegoraro A, Hristodor AM, Crane GM, Meneghelli I, Brignole C, *et al*. Ataluren improves hematopoietic and pancreatic disorders in Shwachman-Diamond syndrome patients: a compassionate program case-series. Nat Commun 2025;16(1):8189. doi:10.1038/s41467-025-63137-3, PMID:40897730.

[36] Zambetti NA, Bindels EM, Van Strien PM, Valkhof MG, Adisty MN, Hoogenboezem RM, *et al*. Deficiency of the ribosome biogenesis gene Sbds in hematopoietic stem and progenitor cells causes neutropenia in mice by attenuating lineage progression in myelocytes. Haematologica 2015;100(10):1285–1293. doi:10.3324/haematol.2015.131573, PMID:26185170.

[37] Jakovljevic J, Ohmayer U, Gamalinda M, Talkish J, Alexander L, Linnemann J, *et al*. Ribosomal proteins L7 and L8 function in concert with six A_3 assembly factors to propagate assembly of domains I and II of 25S rRNA in yeast 60S ribosomal subunits. RNA 2012;18(10):1805–1822. doi:10.1261/rna.032540.112, PMID:22893726.

[38] Ito H, Machida K, Fujino Y, Hasumi M, Sakamoto S, Nagai Y, *et al*. Canonical translation factors eIF1A and eIF5B modulate the initiation step of repeat-associated non-AUG translation. Nucleic Acids Res 2025;53(18):gkaf994. doi:10.1093/nar/gkaf994, PMID:41063344.

[39] Flotho A, Melchior F. Sumoylation: a regulatory protein modification in health and disease. Annu Rev Biochem 2013;82:357–385. doi:10.1146/annurev-biochem-061909-093311, PMID:23746258.

[40] Mandemaker IK, Fessler E, Corujo D, Kotthoff C, Wegerer A, Rouillon C, *et al*. The histone chaperone ANP32B regulates chromatin incorporation of the atypical human histone variant macroH2A. Cell Rep 2023;42(10):113300. doi:10.1016/j.celrep.2023.113300, PMID:37858472.

[41] Ioannou K, Samara P, Livaniou E, Derhovanessian E, Tsitsilonis OE. Prothymosin alpha: a ubiquitous polypeptide with potential use in cancer diagnosis and therapy. Cancer Immunol Immunother 2012;61(5):599–614. doi:10.1007/s00262-012-1222-8, PMID:22366887.

[42] Mailand N, Gibbs-Seymour I, Bekker-Jensen S. Regulation of PCNA-protein interactions for genome stability. Nat Rev Mol Cell Biol 2013;14(5):269–282. doi:10.1038/nrm3562, PMID:23594953.

[43] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2023;51(D1):D523–D531. doi:10.1093/nar/gkac1052, PMID:36408920.

[44] Ganapathi KA, Austin KM, Lee CS, Dias A, Malsch MM, Reed R, *et al*. The human Shwachman-Diamond syndrome protein, SBDS, associates with ribosomal RNA. Blood 2007;110(5):1458–1465. doi:10.1182/blood-2007-02-075184, PMID:17475909.

[45] Trinchieri G. Interleukin-12 and the regulation of innate resistance and adaptive immunity. Nat Rev Immunol 2003;3(2):133–146. doi:10.1038/nri1001, PMID:12563297.

[46] Koyama Y, Brenner DA. Liver inflammation and fibrosis. J Clin Invest 2017;127(1):55–64. doi:10.1172/jci88881, PMID:28045404.

[47] Suk FM, Wu CY, Fang CC, Chen TL, Liao YJ. β-HB treatment reverses sorafenib resistance by shifting glycolysis-lactate metabolism in HCC. Biomed Pharmacother 2023;166:115293. doi:10.1016/j.biopha.2023.115293, PMID:37567069.

[48] Battle A, Brown CD, Engelhardt BE, Montgomery SB, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhanc-

ing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts, Laboratory, Data Analysis &Coordinating Center (LDACC), NIH program management, Biospecimen collection, Pathology, eQTL manuscript working group. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204–213. doi:10.1038/nature24277, PMID:29022597.