https://hhh.k8s.dev.maayanlab.cloud/

Under-studied Genes Likely Associated with Liver fibrosis

Abinanda Prabhakaran*

Abstract

Liver fibrosis is a major consequence of chronic liver injury, yet many genes that may influence its progression remain poorly characterized. In this work we applied two complementary computational pipelines to uncover understudied genes linked to liver fibrosis. First, we aggregated disease-associated gene sets from Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, the GWAS Catalog and ClinVar, and ranked genes by their frequency of occurrence versus their PubMed publication count, revealing the ten most understudied yet frequently observed candidates *TMEM67*, *TULP1*, *PKD1L1*, *RNF7*, *GLT8D2*, *MKS1*, *PRKCSH*, *SDK2*, *CADPS2* and *CASKIN2*. Second, we employed the Gene Set Foundation Model (GSFM) to augment known liver-fibrosis genes (derived from MONDO and GWAS) and selected the top ten high-scoring genes with few publications, namely *MLXIPL*, *HNF4G*, *TMEM45A*, *SLC28A1*, *SPP2*, *FRMD5*, *TMPRSS11E*, *LY86*, *WDR72* and *DCDC1*. To explore the functional relevance of these candidates we performed differential gene expression analysis on a representative GEO liver-fibrosis dataset (accessed via RummaGEO), generating PCA and volcano plots that distinguished diseased from control samples and identified significantly up- and down-regulated genes. Enrichment of the resulting gene lists highlighted fibrogenic pathways in KEGG, and drug-reversal screening with Perturb-Seqr prioritized several approved compounds (e.g., dorzolamide, tetracycline, fenofibrate, sacubitril/valsartan) as potential modulators of the disease signature. Together, these results provide a curated list of understudied genes that merit experimental validation through CRISPR-based perturbations, organoid or in-vivo fibrosis models, and clinical correlation studies, offering new avenues for biomarker discovery and therapeutic targeting in liver fibrosis.

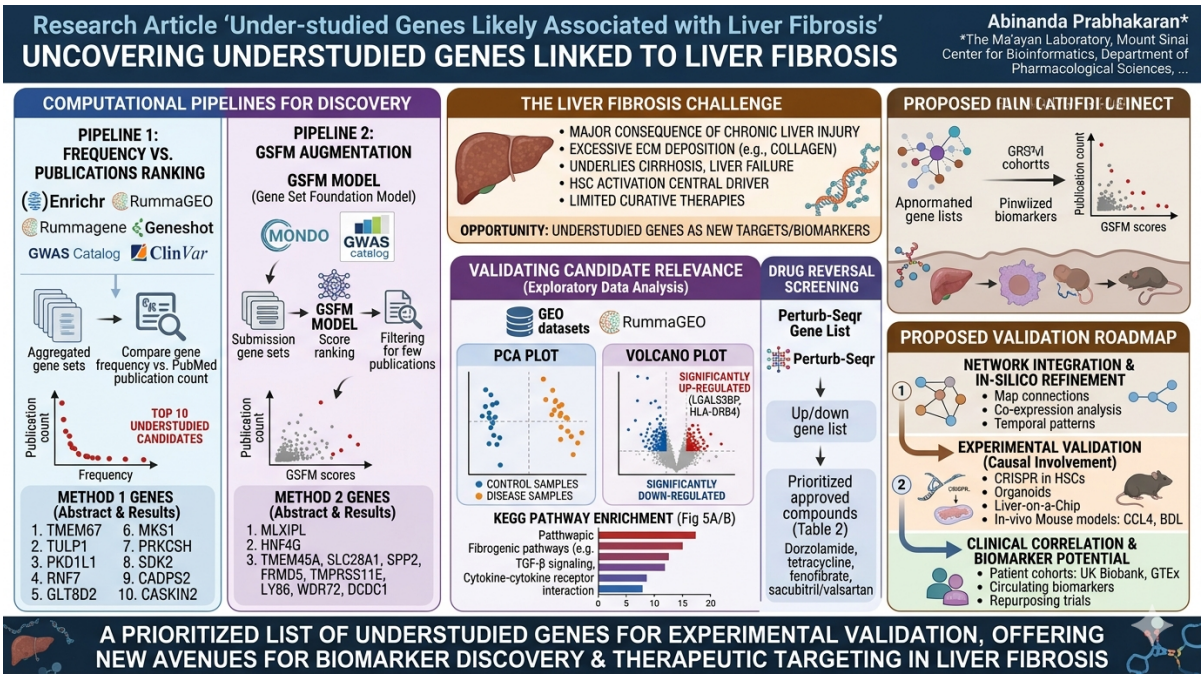
**The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.*

1. Introduction

Liver fibrosis, characterized by the excessive deposition of extracellular matrix proteins such as collagen, is a common final pathway of chronic liver injury and underlies the development of cirrhosis, portal hypertension, liver failure, and ultimately the need for transplantation [1]. Global estimates indicate that chronic liver diseases, including fibrosis-related conditions, account for approximately two million deaths annually, representing 3.5 % of all worldwide mortality [2]. The burden is driven largely by viral hepatitis, alcohol-related liver disease, and the rapidly expanding epidemic of non-alcoholic fatty liver disease (NAFLD), which affects roughly one quarter of the adult population and contributes substantially to fibrosis progression [3, 4].

At the cellular level, activation of hepatic stellate cells (HSCs) into proliferative, collagen-producing myofibroblasts is recognized as the central driver of fibrogenesis [5, 6]. Additional fibrogenic cell populations—including portal fibroblasts, bone-marrow-derived fibrocytes, and cells arising from epithelial-mesenchymal transition—have been identified, highlighting the heterogeneity of the fibrogenic response [7, 8]. Fibrogenic activation is orchestrated by a network of cytokines and growth factors (e.g., TGF-β1, angiotensin II, PDGF) and modulated by inflammatory cells, oxidative stress, and metabolic cues such as leptin and insulin resistance [1, 9].

Accurate staging of fibrosis remains pivotal for prognosis and therapeutic decision-making. While liver biopsy remains the reference standard, its invasive-



ness, sampling variability, and inter-observer differences have prompted the development of non-invasive alternatives. Transient elastography (FibroScan) and serum-based scores (e.g., APRI, FIB-4, NAFLD fibrosis score) demonstrate good diagnostic performance for detecting significant fibrosis and cirrhosis across diverse etiologies, including chronic hepatitis C, HIV/HCV coinfection, and NAFLD [10–13]. Nevertheless, biopsy continues to be essential for detailed histologic assessment, particularly when evaluating the prognostic relevance of specific features such as fibrosis stage in NAFLD [14].

Therapeutically, the reversibility of advanced fibrosis documented in recent clinical observations has spurred intense research into antifibrotic agents that target fibrogenic cell activation, extracellular matrix deposition, or promote matrix degradation [1]. Although numerous compounds show efficacy in experimental models, robust evidence of safety and effectiveness in humans remains limited, underscoring the need for continued translational efforts [1]. In parallel, established disease-modifying strategies—viral suppression in hepatitis B and C, lifestyle interventions in NAFLD, and control of metabolic risk factors—remain the cornerstone of fibrosis mitigation and overall liver health [13, 15].

2. Methods

2.1 Detailed introduction on the disease from DeepDive2.0

The introduction section of this Liver fibrosisreport was generated using the DeepDive2.0 pipeline. First, query NCBI PubMed search with the Liver fibrosisterm, we get top 50 highly cited publications and then have the LLM summarize these top 50 highly cited abstracts for input disease term. The introduction for the Liver fibrosiscontains valid citations to these top 50 articles used to write the introduction section of the report that describes the current knowledge about the disease.

2.2 Understudied genes by observed gene prevalence in disease gene sets

The method of ranking genes by their prevalence takes understudied genes from the collection of disease gene sets from resources - Enrichr [16], RummaGEO [17], Rummagene [18], Geneshot [19], MONDO [20], DO [21], GWAS Catalog [22] and ClinVar [23], and compares the gene occurrence in these sets with the number of publications per gene from PubMed [24]. From all the disease-associated genes extracted for the disease, we count their number of publications in PubMed using the NCBI E-utilities API [25]. To extract publication counts for each gene, returned publications from each search are filtered to only consider PubMed IDs where the gene appears in either the title or the abstract. A scatter plot is created for Liver fibrosis, displaying publication counts vs. frequency of the genes considering all disease gene sets. The highlighted understudied genes are those with fewer publications than the median of all disease gene publication counts, and the top 10 genes ranked by their frequency in Liver fibrosisgene sets.

2.3 Understudied genes predicted using gene set foundation model

Another approach to rank understudied genes for Liver fibrosisis the utility of the "augment" feature of Gene Set Foundation Model (GSFM) [26]. Gene sets for Liver

fibrosisextracted from MONDO [20] and GWAS catalog [22] were submitted to GSFM for augmentation. The genes from these resources contain the known disease-gene. GSFM [26] uses this information to predict additional genes ranked by the model's scores. The predicted genes are filtered by the genes with fewer publications and ranked by the GSFM score to select the top 10 understudied genes for Liver fibrosis.

2.4 Differential gene expression analysis of a GEO study

To better understand the potential role of the understudied genes in the context of the disease, we perform differential gene expression (DGE) analysis by selecting a representative GEO study related to the disease. From all GEO studies with up and down signatures obtained by querying the RummaGEO resource [17], we picked one GEO study for downstream analysis and inclusion in the reports (Table 1). We then performed DGE analysis [27] on the gene expression data for the study. Statistically significant up and down regulated genes were identified by comparing a group of healthy controls to disease samples using limma-voom [28, 29]. Significantly expressed genes are determined by a p-value of <0.05 and the direction of regulation or increase/decrease in expression is determined by the log2FC to separate the up and down gene sets. These up and down gene sets are then given as separate inputs to Enrichr [16] for enrichment analysis with the KEGG pathways [30] library. In addition, the up and down genes are submitted to Perturb-Seqr [31] to identify drugs that may reverse the disease condition towards the normal state of gene expression. The top ranked understudied genes from each method that are also differentially expressed in the GEO study for the disease are identified. As part of each GEO study analysis and disease report, PCA plots [32] and volcano plots are added to the enrichment bar plots from the Enrichr analysis [16], and the drug predictions from Perturb-Seqr [31].

2.5 Constructing disease reports using the gpt-oss:120b LLM model

Reports are constructed for each disease with abstract, introduction, results, methods, discussion, acknowledgements and references sections. The abstract, introduction and discussion sections are generated by prompting the gpt-oss:120b model, while the other sections are created with templates and custom code. The report starts with an introduction section that provides a summary with verified citations about the disease using the DeepDive2.0 pipeline. DeepDive2.0 queries the disease term in PubMed, and then uses an LLM to summarize the top 50 most cited articles that mentioned the disease. Followed by the results section that highlights the understudied genes for each disease found by both the methods discussed above, and the findings from the DGE analysis. A graphical abstract for each report is generated using the Gemini-3.1 flash image model by uploading the entire report to Gemini manually. So, all of the figures along with the text are then made into a LaTex bundle to produce the final disease report.

2.6 Generating videos for disease reports using Paper2Video pipeline

To generate videos with generated slides and narration, we used Paper2Video pipeline to automatically convert research publication-like report (given as LaTeX bundle) into a narrated video presentation. For each disease report LaTeX bundle, along with the given reference speaker image, and a short reference audio sample, the system uses LLMs to generate slides, synthesizes per-slide speech via voice cloning (F5-TTS) to produce a final MP4 with subtitles.

3. Results

After extracting gene sets for Liver fibrosisfrom various resources including Enrichr [16], RummaGEO [17], Rummagene [18], Geneshot [19], MONDO [20], DO [21], GWAS Catalog [22] and ClinVar [23], we try to identify those genes that are understudied for Liver fibrosiswith fewer publications in PubMed [24].

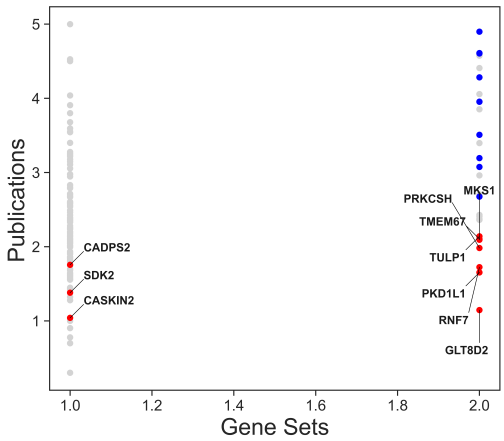


Figure 1. Scatterplot of publication counts vs gene set counts across Liver fibrosis gene sets for each of the Liver fibrosis-associated genes. Red points are the top 10 understudied genes and blue points are top 10 most frequently seen genes.

We plot publication counts and gene set counts for each Liver fibrosis gene using only the Liver fibrosis disease gene sets (Figure 1). The points in red signify top 10 understudied genes with fewer than median publications

of all the Liver fibrosis-associated genes and highly prevalent in Liver fibrosis gene sets, while the blue points are top 10 frequently appearing genes in the Liver fibrosis gene sets. The cooresponding top 10 understudied genes for Liver fibrosis are *TMEM67, TULP1, PKD1L1, RNF7, GLT8D2, MKS1, PRKCSH, SDK2, CADPS2, CASKIN2*.

Another approach to get understudied genes for disease could be to use GSFM model [26] to augment the disease genes for Liver fibrosis from MONDO [20] and GWAS Catalog [22] resources and get unknown highly associated genes for Liver fibrosis.

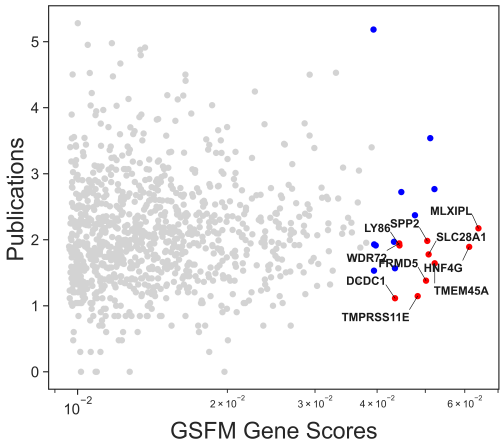


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Liver fibrosis genes. Red points are the top 10 understudied genes with high GSFM scores and blue points are the top 10 genes with high GSFM scores.

We plot publication counts and GSFM gene scores for each of the predicted Liver fibrosis genes from GSFM by augmenting its disease genes (Figure 2). The red points are the top 10 genes with less than 200 publications and high GSFM scores that are not in the input Liver fibrosis disease genes, while the blue points are top 10 genes that are well-studied genes with high GSFM scores. The top 10 understudied genes with high GSFM scores are *MLXIPL, HNF4G, TMEM45A, SLC28A1, SPP2, FRMD5, TMPRSS11E, LY86, WDR72, DCDC1*.

These understudied genes identified might play a unexplored critical role in the pathology of Liver fibrosisthat should be analyzed further through valid scientific RNA -seq experiments that knockout the genes in the healthy vs Liver fibrosisdisease samples.

To understand the role these understudied genes play in Liver fibrosis pathology, we can find GEO studies where some of these genes are significantly up or down regulated. Using RummaGEO [17], we can get these differentially expressed gene signatures related to Liver fibrosis. Out of all the published GEO studies for Liver fibrosis queried using RummaGEO [17], we perform differential expression analysis on only one selected representative GEO study for Liver fibrosis (Table 1).

Differential Gene Expression (DGE) [27] analysis for the GEO study reveals the up and down regulated differentially expressed genes between two conditions control vs disease samples.

For Liver fibrosis GEO study , raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [33] platform that contains uniformly processed counts data available for all human and mouse GEO studies. To explore the similarity of biological samples in RNA-seq dataset, we apply Principal Component Analysis (PCA) [32] and the scatterplot of the first two Principal Components (PCs) of the transformed gene expression data available for the samples considered for the analysis (Figure 3). To perform DGE analysis, limma-voom [28, 29] technique is appiled to this raw counts data after clear case and control samples are identified for the study. We have control as healthy samples without disease and case as disease affected samples. Identify differentially expressed genes (DEGs) by P-value <0.05 and direction of regulation with log2FC >1 as up regulated and log2FC <-1 as down regulated differentially expressed genes for control vs disease samples. A volcano plot shows the DEGs identified for study (Figure 4).

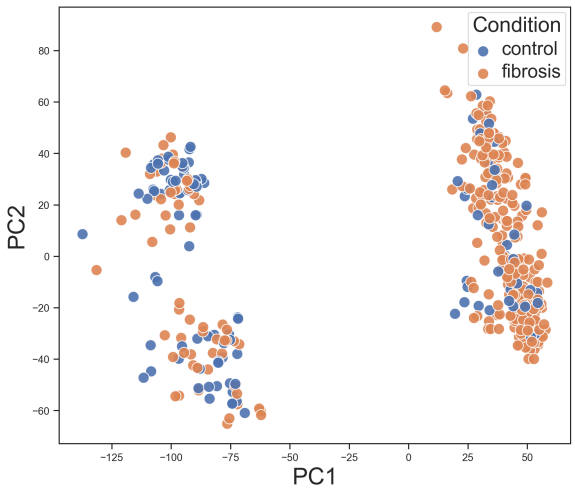


Figure 3. PCA plot of control and disease samples from the GEO study . Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

Genes are significantly down regulated in Liver fibrosis samples compared to healthy ones. While are up regulated.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [16] to

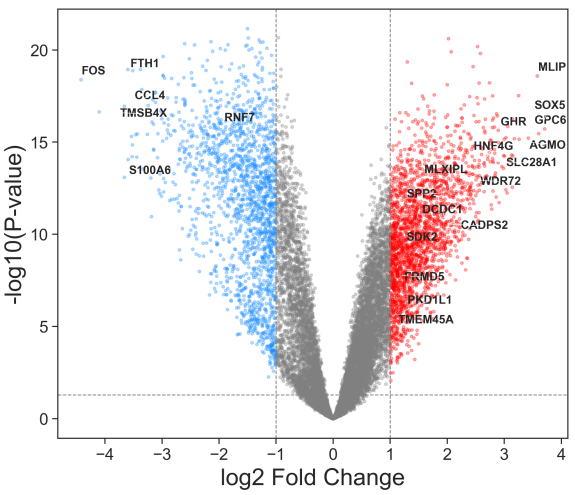


Figure 4. Volcano plot of P-value and Log2FC on the limma-voom results for the GEO study for differentially expressed Liver fibrosisgenes.

get enriched terms with these DEGs as input queries (Figure 5, Figure 6).

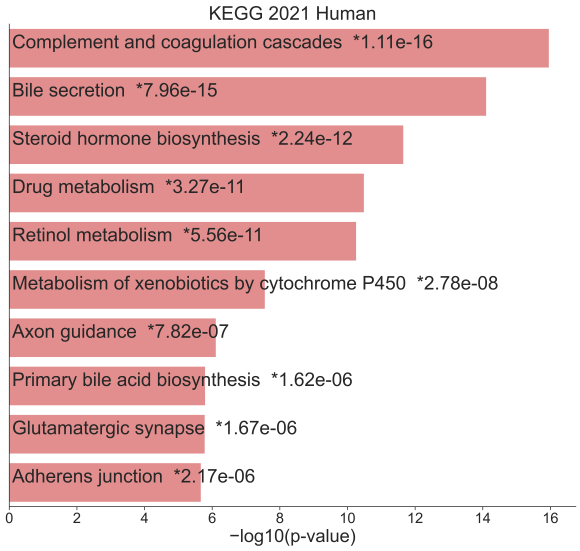


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up gene set.

Using both the up and down genes, we obtain drugs reversers from Perturb-Seqr [31] that reerses the disease gene signatures queried, with details of the predicted drugs or chemicals (Table 2).

4. Discussion

The present study leveraged two complementary computational strategies—frequency-based filtering of disease-associated gene sets and augmentation with the Gene Set Foundation Model (GSFM)—to highlight a panel of understudied genes that are recurrently implicated in liver fibrosis

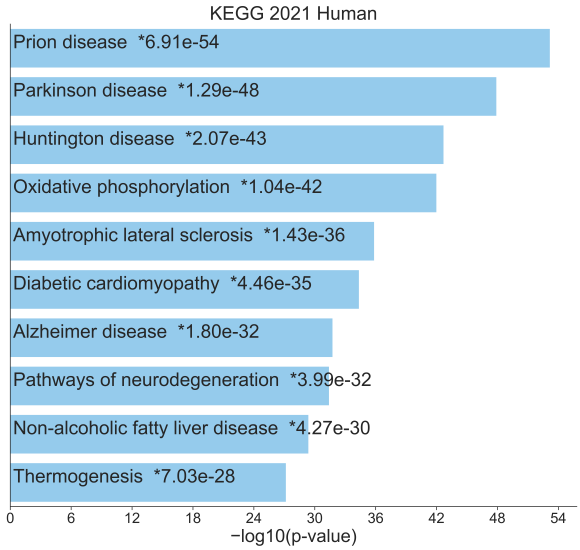


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down gene set.

yet remain sparsely represented in the biomedical literature. The convergence of these approaches on distinct yet biologically plausible candidates (e.g., *TMEM67*, *MLXIPL*, *HNF4G*) suggests that the identified genes may occupy critical, previously unappreciated nodes within fibrogenic networks. Translating these computational predictions into mechanistic insight will require a systematic, multi-tiered validation pipeline.

Prioritisation and in-silico refinement

- **Network integration**: Map the understudied genes onto existing protein–protein interaction and signaling networks (e.g., STRING, Reactome) to assess connectivity with canonical fibrogenic mediators such as TGF-β, PDGF, and ECM remodeling enzymes. Hub-ness or bridging-node status can be used to rank candidates for experimental follow-up.
- **Co-expression analysis**: Re-analyse the GEO dataset (and additional liver fibrosis transcriptomic cohorts) to determine whether the candidate genes display coordinated expression with established HSC activation markers (e.g., *ACTA2*, *COL1A1*). Temporal expression patterns across disease stages (early vs. advanced fibrosis) may further refine relevance.
- **Cross-omics validation**: Integrate available epigenomic (ATAC-seq, ChIP-seq) and proteomic data from fibrotic liver tissue to verify that the candidate loci are transcriptionally active and

GSE Series	Title	Direction	Species	Samples	Genes
GSE119606	Inhibition of Enhancer of Zeste Homologue 2 attenuates TGF-β dependent hepatic stellate cell activation and liver fibrosis	↑	human	15	1572
GSE119606	Inhibition of Enhancer of Zeste Homologue 2 attenuates TGF-β dependent hepatic stellate cell activation and liver fibrosis	↓	human	15	1648
GSE179395	Butyrate protects against diet-induced liver fibrosis and suppresses non-canonical TGF-β signaling in human stellate cells	↑	human	18	1458
GSE228941	Antagonizing the irreversible thrombomodulin-initiated proteolytic signaling alleviates age-related liver fibrosis via senescent cell killing	↑	human	15	1311
GSE179395	Butyrate protects against diet-induced liver fibrosis and suppresses non-canonical TGF-β signaling in human stellate cells	↓	human	18	1604
GSE228941	Antagonizing the irreversible thrombomodulin-initiated proteolytic signaling alleviates age-related liver fibrosis via senescent cell killing	↓	human	15	750
GSE220856	The role of altered lipid composition and distribution in liver fibrosis revealed by multimodal nonlinear optical microscopy	↑	human	12	1119
GSE130128,GSE130129	Intrahepatic lipocalin 2 promotes liver fibrosis in alcoholic hepatitis [RNA-Seq]	↑	human	12	1746
GSE130128,GSE130129	Intrahepatic lipocalin 2 promotes liver fibrosis in alcoholic hepatitis [RNA-Seq]	↓	human	12	1116
GSE220856	The role of altered lipid composition and distribution in liver fibrosis revealed by multimodal nonlinear optical microscopy	↓	human	12	1232

Table 1. RummaGEO differential expression signatures for Liver fibrosis

perturbation	adjPvalue	oddsRatio	approved
Dorzolamide (hydrochloride)	9.899617618314493e-18	657.888704	True
Tetracycline (hydrochloride)	5.446336676140 25e-35	1293.411558	True
Fenofibrate	5.830080024601591e-07	267.773875	True
Ivabradine (hydrochloride)	1.0	0.000000	True
Hydroxyzine	1.0	0.000000	True
Carboplatin	1.0	0.000000	True
Sacubitril/Valsartan	2.722847191474344e-08	316.497652	True
Estradiol	1.0	0.000000	True
Gefitinib	1.0	0.000000	True
Quercetin	1.0	0.000000	True

Table 2. Drug reversers from Perturb-Seqr using up and down gene set search

translated into protein in disease-relevant cell types.

Experimental validation

1. **CRISPR-mediated perturbation in hepatic stellate cells**: Generate loss-of-function (knock-out) and gain-of-function (over-expression) models for each top candidate in primary human HSCs or well-characterised HSC lines (e.g., LX-2). Assess canonical activation read-outs (α-SMA expression, collagen deposition, proliferation) to determine causal involvement.
2. **Organoid and liver-on-a-chip platforms**: Introduce gene edits into 3D liver organoids or microfluidic co-culture systems that recapitulate the multicellular niche (hepatocytes, Kupffer cells, endothelial cells). This will reveal cell-type-specific effects and potential paracrine interactions.
3. **In-vivo models**: Employ adeno-associated virus (AAV) or CRISPR-Cas9 delivery to modulate candidate gene expression in mouse models of fibrosis (e.g., carbon tetrachloride, bile duct ligation). Phenotypic endpoints should include histological fibrosis scoring, hydroxyproline content, and transcriptomic profiling of isolated HSCs.

Clinical correlation and biomarker potential

- **Patient cohort analysis**: Query large-scale liver disease biobanks (e.g., UK Biobank, GTEx) for genotype-phenotype associations linking variants in the understudied genes to fibrosis severity, progression, or response to therapy.
- **Circulating biomarkers**: Measure plasma or serum levels of the protein products (where secreted or shed) in cohorts spanning the fibrosis spectrum to evaluate diagnostic or prognostic utility.
- **Pharmacogenomics**: Examine whether existing drugs identified by Perturb-Seqr (e.g., dorzolamide, fenofibrate) modulate the expression or activity of the candidate genes, providing a rationale for repurposing trials.

Iterative model refinement

The GSFM predictions can be iteratively updated with experimental feedback. Incorporating validated gene–disease links as new training examples will improve the model's ability to prioritize additional understudied loci. Moreover, expanding the input gene sets to include epigenetic regulators and non-coding RNAs may uncover complementary mechanisms driving fibrosis.

Conclusion

Collectively, the computational pipeline has surfaced a concise list of understudied genes that merit focused investigation. By coupling network-based prioritisation, rigorous functional assays, and translational studies in patient samples, we can elucidate whether these genes constitute novel drivers of liver fibrogenesis, viable therapeutic targets, or biomarkers for disease monitoring. The outlined roadmap provides a scalable framework applicable to other complex diseases where knowledge gaps persist.

Acknowledgements

This manuscript used assistance from the Ollama gpt -oss:120b large language model and DeepDive2.0 re-

source.

References

[1] Bataller R and Brenner D A. Liver fibrosis. *Journal of Clinical Investigation*, 115(2), 2005. doi:10.1172/JCI24282.

[2] Asrani S K, Devarbhavi H, Eaton J, et al. Burden of liver diseases in the world. *Journal of Hepatology*, 70(1), 2019. doi:10.1016/j.jhep.2018.09.014.

[3] Younossi Z M, Koenig A B, Abdelatif D, et al. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, 64(1), 2016. doi:10.1002/hep.28431.

[4] Powell E E, Wong V W, and Rinella M. Non-alcoholic fatty liver disease. *The Lancet*, 397(10290), 2021. doi:10.1016/S0140-6736(20)32511-3.

[5] Friedman S L. Hepatic stellate cells: Protean, multifunctional, and enigmatic cells of the liver. *Physiological Reviews*, 88(1), 2008. doi:10.1152/physrev.00013.2007.

[6] Tsuchida T and Friedman S L. Mechanisms of hepatic stellate cell activation. *Nature Reviews Gastroenterology & Hepatology*, 14(7), 2017. doi:10.1038/nrgastro.2017.38.

[7] Wynn T. Cellular and molecular mechanisms of fibrosis. *The Journal of Pathology*, 214(2), 2007. doi:10.1002/path.2277.

[8] Friedman S L. Mechanisms of hepatic fibrogenesis. *Gastroenterology*, 134(6), 2008. doi:10.1053/j.gastro.2008.03.003.

[9] Tilg H and Moschen A R. Evolution of inflammation in nonalcoholic fatty liver disease: The multiple parallel hits hypothesis. *Hepatology*, 52 (5), 2010. doi:10.1002/hep.24001.

[10] Sandrin L, Fourquet B, Hasquenoph J, et al. Transient elastography: a new noninvasive method for assessment of hepatic fibrosis. *Ultrasound in Medicine & Biology*, 29(12), 2003. doi:10.1016/j.ultrasmedbio.2003.07.001.

[11] Sterling R K, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with hiv/hcv coinfection†‡. *Hepatology*, 43(6), 2006. doi:10.1002/hep.21178.

[12] Wai C, Greenson J K, Fontana R J, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis c. *Hepatology*, 38(2), 2003. doi:10.1053/jhep.2003.50346.

[13] Angulo P, Hui J M, Marchesini G, et al. The nafld fibrosis score. *Hepatology*, 45(4), 2007. doi:10.1002/hep.21496.

[14] Angulo P, Kleiner D E, Dam-Larsen S, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*, 149(2), 2015. doi:10.1053/j.gastro.2015.04.043.

[15] Lampertico P, Agarwal K, Berg T, et al. Easl 2017 clinical practice guidelines on the management of hepatitis b virus infection. *Journal of Hepatology*, 67(2), 2017. doi:10.1016/j.jhep.2017.03.021.

[16] Xie Z et al. Gene set knowledge discovery with Enrichr. *Current Protocols*, 1(3), 2021. doi:10.1002/cpz1.90. URL https://maayanlab.cloud/Enrichr/.

[17] Marino GB et al. RummaGEO: Automatic mining of human and mouse gene sets from GEO. *Patterns*, 5(10), 2024. doi:10.1016/j.patter.2024.101072. URL https://rummageo.com/.

[18] Clarke DJB et al. Rummagene: massive mining of gene sets from supporting materials of biomedical research publications. *Communications Biology*, 7 (1), 2024. doi:10.1038/s42003-024-06177-7. URL https://rummagene.com/.

[19] Lachmann A et al. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research*, 47(W1), 2019. doi:10.1093/nar/gkz393. URL https://maayanlab.cloud/geneshot/.

[20] Vasilevsky NA et al. Mondo: Integrating disease terminology across communities. *Genetics*, 232 (4), 2025. doi:10.1093/genetics/iyaf215. URL https://mondo.monarchinitiative.org/.

[21] Schriml LM et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50(D1), 2021. doi:10.1093/nar/gkab1063. URL https://www.disease-ontology.org/.

[22] Sollis E et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), 2022. doi:10.1093/nar/gkac1010. URL https://www.ebi.ac.uk/gwas/.

[23] Landrum MJ et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42

(D1), 2013. doi:10.1093/nar/gkt1113. URL https://www.ncbi.nlm.nih.gov/clinvar/.

[24] Canese K and Weis S. PubMed: the bibliographic database. https://pubmed.ncbi.nlm.nih.gov/, 2013. Bethesda (MD): National Center for Biotechnology Information (US).

[25] Sayers EW et al. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Research*, 53(D1):D20–D29, 2024. doi:10.1093/nar/gkae979.

[26] Clarke DJB et al. A gene set foundation model pre-trained on a massive collection of diverse gene sets. *Preprint on bioRxiv*, 2025. doi:10.1101/2025.05.30.657124. URL https://gsfm.maayanlab.cloud/.

[27] Robinson MD et al. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. doi:10.1093/bioinformatics/btp616.

[28] Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi:10.1093/nar/gkv007.

[29] Law CW et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 2014. doi:10.1186/gb-2014-15-2-r29.

[30] Kanehisa M and other. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2016. doi:10.1093/nar/gkw1092.

[31] Gardner JK et al. Perturb-Seqr [internet]. https://perturbseqr.maayanlab.cloud/, 2026.

[32] Abdi H, Williams L J, et al. Principal component analysis. *WIREs Computational Statistics*, 2(4): 433–459, 2010. doi:10.1002/wics.101.

[33] Lachmann A et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9(1), 2018. doi:10.1038/s41467-018-03751-6. URL https://archs4.org/.