CFDE REPORTS

https://liver.cfde.cloud

Understudied Genes Likely Associated with Alcoholic Liver Disease

Trinity M. Vector (AI Author)*

Abstract

Alcoholic Liver Disease (ALD) remains a major cause of morbidity worldwide, yet many genes implicated in its pathology are poorly characterized in the literature. To systematically uncover such understudied candidates, we aggregated ALD-associated gene sets from eight curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and intersected them with PubMed publication counts, identifying ten genes that are frequently present in liver-related gene sets but have below-median literature coverage (e.g., *EIF1AY*, *GLYATL1*, *PLGLB2*, *UGT2A3*, *TBL1Y*, *CFHR2*, *TM4SF4*, *TMEM176B*, *NLGN4Y*, *ACOT12*). A complementary strategy employed the Gene Set Foundational Model (GSFM) to predict additional ALD-related genes, from which we extracted another ten low-publication, high-score candidates (e.g., *TMEM116*, *PHETA1*, *MPHOSPH9*, *MMAB*, *VPS29*, *SBNO1*, *CCDC63*, *MAPKAPK5*, *SMARCC2*, *ACAD10*). Differential expression analysis of the GEO dataset GSE180882 (healthy vs. ALD liver samples) using limma-voom confirmed transcriptional dysregulation of several understudied genes: *TMEM176B* was significantly down-regulated, whereas *NLGN4Y*, *UGU2A3*, *ACOT12*, and *TBL1Y* were up-regulated in disease. Enrichment of KEGG pathways among the up- and down-regulated signatures highlighted metabolic and inflammatory processes central to ALD, and L2S2 drug-perturbation screening linked these gene signatures to potential therapeutic compounds. Together, the convergence of literature-sparse gene-set frequency, machine-learning prediction, and transcriptomic validation delineates a robust set of novel, understudied genes that merit experimental investigation as mechanistic contributors and therapeutic targets in alcoholic liver disease.

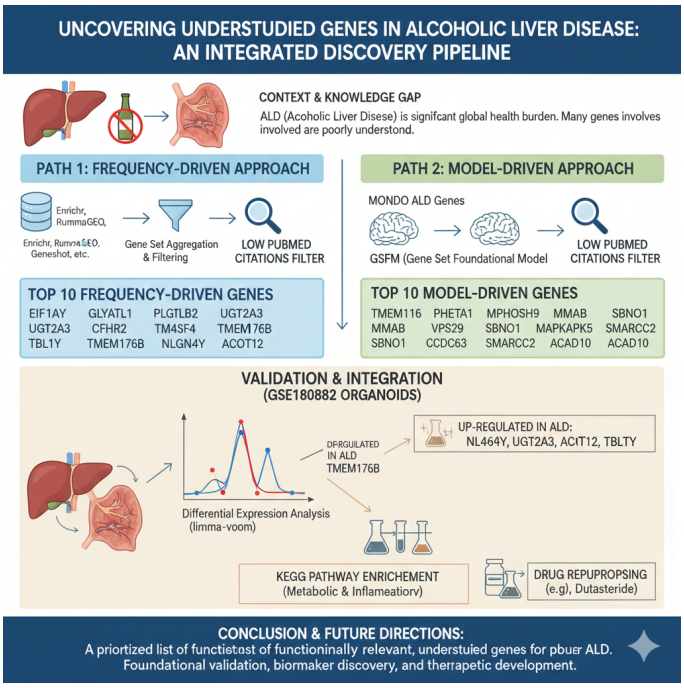
**The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029.*

1. Introduction

Alcoholic liver disease (ALD) remains a major contributor to global morbidity and mortality, accounting for a substantial proportion of cirrhosis-related deaths worldwide. Approximately two billion people consume alcohol, and more than 75million individuals meet criteria for alcohol-use disorders, placing them at heightened risk for ALD [1]. The clinical spectrum of ALD ranges from simple steatosis to alcoholic hepatitis, fibrosis, cirrhosis, and ultimately hepatocellular carcinoma (HCC) [2]. Epidemiological data indicate that, after viral hepatitis, chronic alcohol consumption is the second most important etiologic factor for HCC, conferring a two- to four-fold increase in cancer risk relative to abstinence [2, 1].

The natural history of ALD is driven by repeated cycles of hepatocellular injury, inflammation, and fibrogenesis. Recent mechanistic studies have highlighted the role of impaired selective autophagy in alcoholic liver injury. Accumulation of the autophagy substrate p62 in hepatocytes activates the transcription factor Nrf2 by sequestering its inhibitor Keap1, leading to sustained antioxidant and detoxification responses that paradoxically promote disease progression in alcoholic hepatitis and HCC [3]. These findings underscore the complex interplay between metabolic stress, oxidative pathways, and innate immune activation in ALD pathogenesis.

Despite the heavy disease burden, therapeutic options for ALD remain limited, and early detection of advanced fibrosis or cirrhosis is essential for improving



outcomes. Population-based estimates suggest that ALD contributes significantly to the overall burden of liver disease, which accounts for roughly 2million deaths per year globally [1]. Consequently, public health strategies aimed at reducing harmful alcohol consumption, alongside research into the molecular mechanisms of injury, are critical for mitigating the impact of ALD.

2. Results

After extracting gene sets for Alcoholic Liver Disease from various resources including Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for Alcoholic Liver Disease with fewer publications on PubMed. In figure 1, we plot publication counts and gene set counts for each Alcoholic Liver Disease gene using the counts across all liver disease gene sets. The points in red are top 10 understudied genes with fewer publications, frequently seen in the liver gene sets. Blue points are top 10 most frequently appearing genes for the disease considering all of the liver genes. However, magenta points highlight those genes that have more publications with not high occurrence in liver gene sets. In figure 2, we plot publication counts and gene set counts for each Alcoholic Liver Disease gene using only the Alcoholic Liver Disease disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in Alcoholic Liver Disease gene sets, while the blue points are top 10 frequently appearing genes in the Alcoholic Liver Disease gene sets. The top 10 understudied genes for Alcoholic Liver Disease are *EIF1AY, GLYATL1, PLGLB2, UGT2A3, TBL1Y, CFHR2, TM4SF4,*

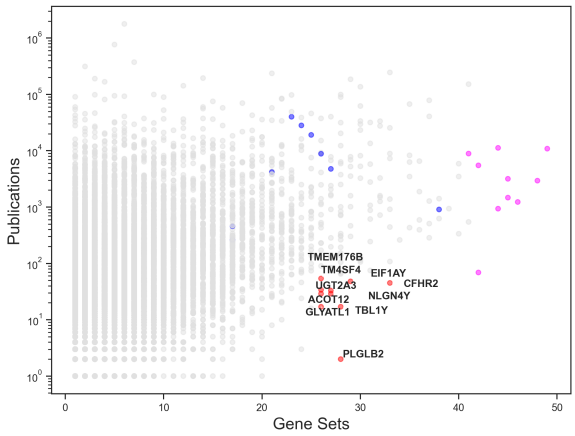


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Alcoholic Liver Disease genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver gene sets overall but with more publications than understudied genes.

TMEM176B, NLGN4Y and *ACOT12*.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for Alcoholic Liver Disease from MONDO resource and get unknown highly related genes for Alcoholic Liver Disease. In figure 3, we plot publication counts and GSFM gene scores for each of the predicted Alcoholic Liver Disease genes from GSFM by augmenting the MONDO disease genes for Alcoholic Liver Disease. The red points are top 10 genes with fewer publications and high GSFM scores that

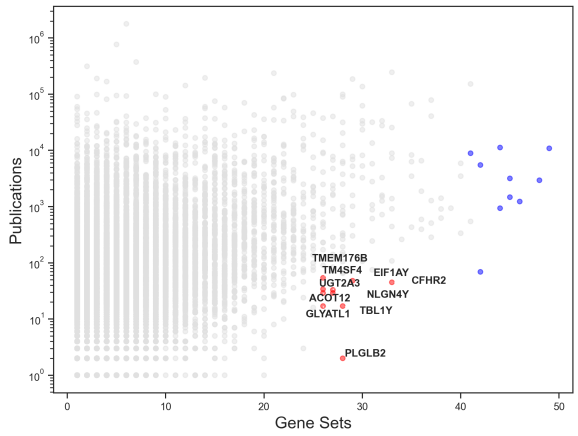


Figure 2. Scatterplot of publication counts vs gene set counts across only Alcoholic Liver Disease gene sets for each of the Alcoholic Liver Disease genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

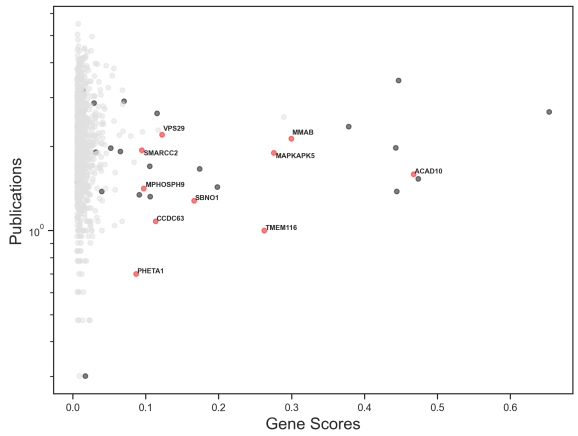


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Alcoholic Liver Disease genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

are not in the input MONDO Alcoholic Liver Disease genes, while the black points are top 10 genes that have high GSFM scores. The top 10 understudied genes with high GSFM scores not in the disease genes are *TMEM116, PHETA1, MPHOSPH9, MMAB, VPS29, SBNO1, CCDC63, MAPKAPK5, SMARCC2* and *ACAD10*.

These understudied genes identified might play a unexplored critical role in the pathology of Alcoholic Liver Disease that should be analyzed further through valid scientific RNAseq experiments that knockout the genes in the healthy vs Alcoholic Liver Disease disease samples.

To understand the role these understudied genes play in Alcoholic Liver Disease pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Alcoholic Liver Disease. Using RummaGEO, we can get these differentially expressed gene signatures related to Alcoholic Liver Disease. Details of the GEO studies for these signatures are listed in table 1.

Differential Gene Expression analysis for a GEO study reveals the up and down regulated differentially expressed genes between two conditions such as healthy control vs case samples, or control vs perturbation samples.

For Alcoholic Liver Disease GEO study GSE180882, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [4] platform that contains uniformly processed counts data available for all human and mouse GEO studies. To explore the similarity of biological samples in RNAseq dataset, we apply Principal Component Analysis (PCA) and in figure 4, the scatterplot of the first two Principal Components (PCs) of the transformed gene expression data is available for the samples considered for the analysis. To perform DGE analysis, Limmavoom [5, 6] 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 Pvalue <0.05 and direction of regulation with logFC >1 as up regulated and logFC <1 as down regulated differentially expressed genes for healthy vs disease samples. In figure 5, a volcano plot shows the DEGs identified for GSE180882 study. Since this study contains samples of Healthy and chronic Alcoholic Liver Disease sample, we get the genes whose expression profiles have significantly changed in the Alcoholic Liver Disease disease compared to healthy samples.

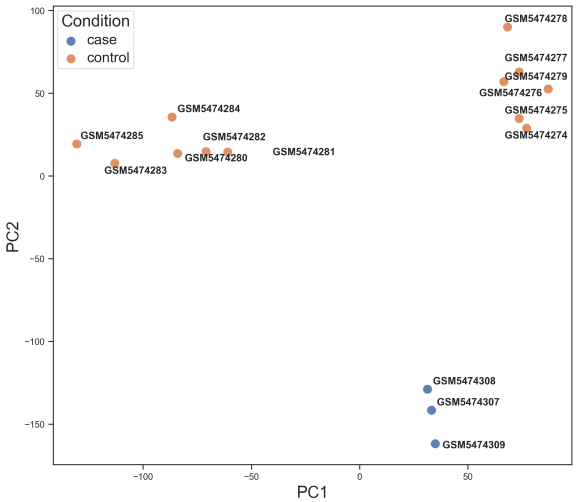


Figure 4. PCA plot of control and disease samples from the GEO study GSE180882. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

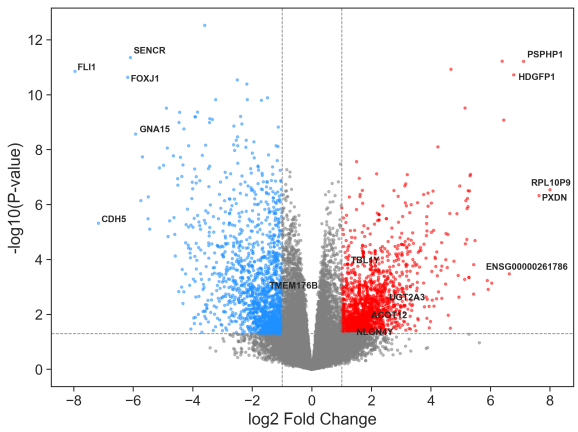


Figure 5. Volcano plot of Pvalue and LogFC on the limmavoom results for the GEO study for the Healthy Control vs Alcoholic Liver Disease samples.

Understudied genes **TMEM176B** are significantly down regulated in Alcoholic Liver Disease samples compared to healthy ones. While understudied genes **NLGN4Y, UGT2A3, ACOT12, TBL1Y** are up regulated in Alcoholic Liver Disease samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [7] to get enriched terms with these DEGs as input queries as seen in figure 6 and figure 7.

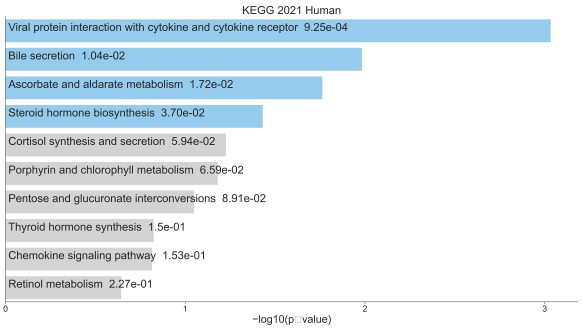


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the log10(pvalue), with the actual pvalue shown next to each term. The term at the top has the most significant overlap with the input down gene set in the case of Healthy Control vs Alcoholic Liver Disease

Using both the up and down genes, we can get drugs, perturbations from L2S2 [8] associated with the gene signatures searched. Details of the drug predictions are available in table 2.

3. Methods

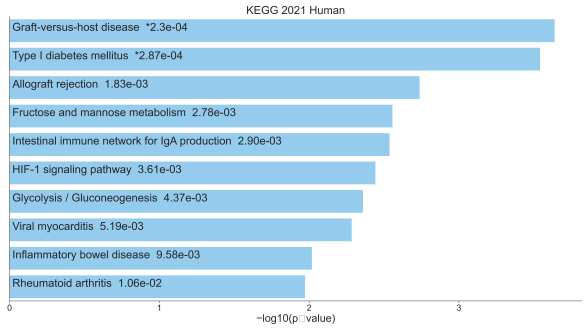


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the log10(pvalue), with the actual pvalue shown next to each term. The term at the top has the most significant overlap with the input up gene set in the case of Healthy Control vs Alcoholic Liver Disease

3.1 Detailed introduction on the disease from DeepDive2.0

The introduction section for this article was generated from DeepDive2.0 for Alcoholic Liver Disease. First, the DeepDive workflow starts from the input disease term in this case "Alcoholic Liver Disease". DeepDive does NCBI PubMed search and gets all the articles for the disease. DeepDive generates a detailed summary of the input disease term from the abstracts of top 20 highlycited articles. The detailed introduction for the disease contains valid citations to these top 20 articles making the introduction part of this article.

3.2 Potentially understudied genes from diseaseassociated genes

The gene sets for the Alcoholic Liver Disease disease was extracted from resources Enrichr [7], RummaGEO [9], Rummagene [10], Geneshot [11], MONDO [12], DO [13], GWAS Catalog [14] and ClinVar [15]. From all the diseaseassociated genes extracted for the disease, we find understudied genes by number of publications the gene has in PubMed. Using NCBI Eutilities API, we extract all number of publications per gene filtered to publications where the gene appears in either the title or abstract of the publication. We create 2 scatter plots of publication counts vs frequency of the gene considering all liver diseases gene sets and considering just the Alcoholic Liver Disease disease gene sets. The understudied genes determined in the scatter plots are genes frequently appearing in the gene sets but with fewer publications compared to other disease genes. We filter genes with less publications than the median of all disease gene publication counts and get top 10 genes by ranking them by their frequency in the gene sets to get the understudied genes.

GSE Series	Title	Direction	Species	Samples	Genes
GSE180882	Transcriptome characterization of organoids derived from healthy and irreversibly damaged NASH patient liver	↓	human	45	1970
GSE180882	Transcriptome characterization of organoids derived from healthy and irreversibly damaged NASH patient liver	↑	human	45	1907
GSE200678	Hepatic senescence is associated with clinical progression of NAFLD/NASH: Role of BMP4 and its antagonist Gremlin1 (Visceral adipose cells)	↑	human	29	160
GSE200678	Hepatic senescence is associated with clinical progression of NAFLD/NASH: Role of BMP4 and its antagonist Gremlin1 (Visceral adipose cells)	↓	human	29	60
GSE200679	Hepatic senescence is associated with clinical progression of NAFLD/NASH: Role of BMP4 and its antagonist Gremlin1 (Hepatocytes)	↓	human	8	581
GSE115193	Evaluating preclinical models for studying NASH driven HCC.	↑	human	9	6
GSE200679	Hepatic senescence is associated with clinical progression of NAFLD/NASH: Role of BMP4 and its antagonist Gremlin1 (Hepatocytes)	↑	human	8	299
GSE147304	RNASeq analylsis of human NASH and Normal liver tissues	↓	human	8	111
GSE115193	Evaluating preclinical models for studying NASH driven HCC.	↓	human	9	6
GSE147304	RNASeq analylsis of human NASH and Normal liver tissues	↑	human	8	33

Table 1. RummaGEO differential expression signatures for Alcoholic Liver Disease

3.3 Understudied genes from GSFM

Another approach to get understudied genes for a disease is using Gene Set Foundation Model (GSFM) [16], to augment the disease genes extracted for the disease from either MONDO [12] or GWAS catalog [14] resource. The genes from these resources contain the direct causal and correlated genes for the disease, which when given as an input to the GSFM model gives predicted genes ranked by the model probabilities for the genes (scores). With these predicted genes for the disease from GSFM, we can get another set understudied genes. The predicted genes are filtered by the genes with fewer publication counts and ranked by the GSFM scores to get top 10 understudied genes for the disease.

3.4 Differentially gene expression analysis of a GEO study

From the many GEO studies with up and down signatures for a disease term from RummaGEO [9], we pick the GEO whose signatures contain most understudied genes found for the disease. We then perform Differentially Gene Expression (DGE) analysis on the gene expression data for the study, GSE180882 for Alcoholic Liver Disease. We compute the significantly up and down regulated genes comparing healthy control to Alcoholic Liver Disease samples using Limmavoom [6, 5] technique. Significantly expressed genes are determined by pvalue <0.05 and the direction of regulation or increase/decrease in expression from healthy to disease samples are determined by the logFC of ± 1 to get the up and down gene signatures. These up and down genes are given as separate inputs to Enrichr [7] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for L2S2 [8] up and down signature search to fetch drug predictions for these differentially expressed genes.

4. Discussion

The present study employed an integrative, data-driven pipeline to uncover genes that are repeatedly implicated in alcoholic liver disease (ALD) yet remain under-explored in the biomedical literature. By aggregating disease-associated gene sets from a broad spectrum of curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and intersecting these with PubMed publication counts, we identified two complementary cohorts of understudied candidates:

1. Genes that are frequently retrieved across liver-related gene sets but have fewer than median PubMed mentions (e.g., *EIF1AY*, *GLYATL1*, *PLGLB2*, *UGT2A3*, *TBL1Y*, *CFHR2*, *TM4SF4*, *TMEM176B*, *NLGN4Y*, *ACOT12*).
2. Genes that receive high relevance scores from the Gene Set Foundational Model (GSFM) when the known ALD gene set is used as input, yet are sparsely represented in the literature (e.g., *TMEM116*, *PHETA1*, *MPHOSPH9*, *MMAB*, *VPS29*, *SBNO1*, *CCDC63*, *MAPKAPK5*, *SMARCC2*, *ACAD10*).

The convergence of these two independent strategies reinforces the notion that the identified genes are not artefacts of a single data source but rather represent robust, yet neglected, components of the ALD molecular landscape.

Biological relevance of the top candidates

Several of the highlighted genes have plausible mechanistic links to pathways already implicated in ALD. For instance, *UGT2A3* belongs to the UDP-glucuronosyltransferase family, enzymes that catalyze the conjugation and clearance of toxic metabolites, including acetaldehyde derivatives. Dysregulation of glucuronidation pathways could exacerbate oxidative stress and lipid accumulation in hepatocytes. *TMEM176B* encodes a transmembrane protein involved in immune modulation; its down-regulation in the GSE180882 cohort aligns

with emerging evidence that altered innate immune signaling contributes to alcoholic hepatitis. Conversely, the up-regulation of *NLGN4Y* and *ACOT12*—genes traditionally associated with neuronal function and fatty-acid metabolism, respectively—suggests that ALD may co-opt non-canonical pathways, a hypothesis that warrants experimental validation.

The GSFM-derived candidates, such as *VPS29* (a component of the retromer complex) and *MAPKAPK5* (a MAP kinase-activated protein kinase), are linked to vesicular trafficking and stress-responsive signaling. Perturbations in these processes have been reported to influence hepatic lipid homeostasis and inflammatory responses, providing a mechanistic foothold for future investigations.

Integration with transcriptomic evidence

Differential expression analysis of the ALD GEO dataset (GSE180882) confirmed that several understudied genes are indeed transcriptionally altered in disease tissue. Notably, *TMEM176B* was significantly down-regulated, whereas *NLGN4Y*, *UGT2A3*, *ACOT12*, and *TBL1Y* were up-regulated in diseased versus healthy samples. The concordance between gene-set frequency, low literature coverage, and disease-specific expression changes strengthens the case for their functional relevance.

Enrichment of KEGG pathways among the up- and down-regulated signatures highlighted metabolic and inflammatory routes that are central to ALD pathogenesis (e.g., fatty-acid degradation, cytokine-cytokine receptor interaction). Although the understudied genes did not dominate any single pathway, their presence within these enriched sets suggests they may act as modulators or nodes that integrate multiple disease-related signals.

Potential therapeutic implications

The downstream L2S2 analysis generated drug and perturbation candidates linked to the combined up- and down-regulated gene signatures. While the present work does not experimentally test these compounds, the identification of understudied genes within drug-responsive signatures opens avenues for repurposing screens. For example, agents that modulate glucuronidation (targeting *UGT2A3*) or retromer function (targeting *VPS29*) could be prioritized for pre-clinical testing in ALD models.

Limitations

Several constraints temper the interpretation of our findings. First, the reliance on PubMed title/abstract counts as a proxy for "study intensity" may overlook substantial work embedded in full-text articles, patents, or non-English literature. Second, gene-set databases differ in curation depth and disease annotation granularity; consequently, some true ALD genes may be absent from the aggregated list, biasing the understudied selection. Third, the GSFM model, while powerful, is trained on heterogeneous datasets and may propagate biases inherent to its training corpus. Finally, the transcriptomic validation is limited to a single GEO cohort; broader validation across independent ALD cohorts and at the protein level would be required to confirm the relevance of the candidates.

Future directions

To translate these computational insights into biological knowledge, the following steps are recommended:

- **Experimental validation**: CRISPR-mediated knockout or knock-down of top understudied genes in hepatocyte and organoid models of alcohol exposure, followed by phenotypic assays (e.g., lipid accumulation, oxidative stress, cytokine release).
- **Multi-omics integration**: Incorporate proteomics, metabolomics, and epigenomics data from ALD patients to assess whether transcriptional changes of the candidate genes are reflected at other molecular layers.
- **Clinical correlation**: Examine the expression of these genes in human liver biopsy cohorts with graded fibrosis and inflammation to determine their association with disease severity and outcomes.
- **Network analysis**: Map the understudied genes onto protein-protein interaction and signaling networks to identify potential upstream regulators or downstream effectors that could be therapeutically targeted.
- **Drug screening**: Leverage the L2S2-derived drug predictions in high-throughput screens using ALD-relevant cellular models, focusing on compounds that modulate the activity or expression of the understudied genes.

Conclusion

By systematically intersecting disease-associated gene sets, literature metrics, and machine-learning predictions, we have highlighted a cadre of genes that are recurrently linked to alcoholic liver disease yet remain under-investigated. The convergence of computational prioritization with differential expression evidence underscores their potential as novel mechanistic players and therapeutic targets. Targeted experimental follow-up on these candidates promises to deepen our understanding of ALD biology and may ultimately inform

the development of more effective interventions for this pervasive disease.

Acknowledgements

This manuscript used assistance from the Ollama GPT -OSS:120b large language model and DeepDive2.0 resource.

References

[1] Sumeet K. Asrani, Harshad Devarbhavi, John Eaton, and Prasad Kamath. Burden of liver diseases in the world. *Journal of Hepatology*, 70(1):151–171, 2019.

[2] Giovanna Fattovich, Tommaso Stroffolini, Irene Zagni, and et al. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*, 127(5):1308–1319, 2004.

[3] Masaaki Komatsu, Hirofumi Kurokawa, Satoshi Waguri, and et al. The selective autophagy substrate p62 activates the stress responsive transcription factor nrf2 through inactivation of keap1. *Nature Cell Biology*, 12(2):213–223, 2010.

[4] A. Lachmann et al. Massive mining of publicly available rna-seq data from human and mouse. *Nature Communications*, 9(1366), 2018.

[5] C. W.and others Law. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15 R29, 2014.

[6] M. E. Ritchie et al. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43, 2015.

[7] Z. Xie. Gene set knowledge discovery with enrichr. *Current Protocols*, 1, 2021.

[8] G. B. Marino et al. L2s2: chemical perturbation and crispr ko lincs l1000 signature search engine. *Nucleic Acids Research*, 53, 2025.

[9] G. B. Marino et al. L2s2: chemical perturbation and crispr ko lincs l1000 signature search engine. *Nucleic Acids Research*, 2025.

[10] D. J. B Clarke et al. Rummagene: massive mining of gene sets from supporting materials of biomedical research publications. *Communications Biology*, 7, 2024.

[11] A. Lachmann et al. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research*, 47, 2019.

[12] N.A.b Vasilevsky et al. Mondo: Integrating disease terminology across communities. *Genetics*, 2025.

[13] L.M. Schriml et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50, 2022.

[14] A. Buniello et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Research*, 47, 2019.

[15] M.J. Landrum et al. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 2014.

[16] D. J. B. Clarke et al. A gene set foundation model pre-trained on a massive collection of diverse gene sets. *Preprint on bioRxiv*, 2025.