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

Under-studied Genes Likely Associated with MASLD

Abinanda Prabhakaran*

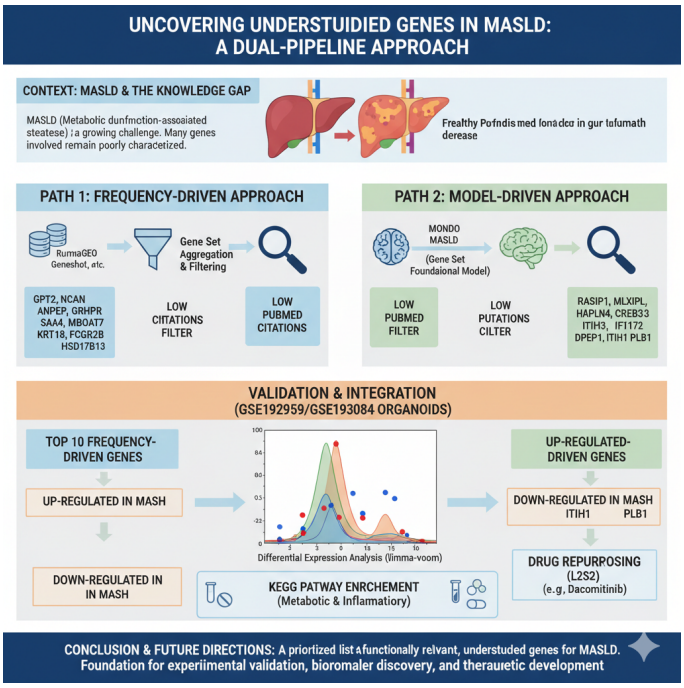
Abstract

Metabolic dysfunction-associated steatotic liver disease (MASLD) remains a major health burden, yet many genes implicated in its pathology are poorly characterized. We integrated disease-associated gene sets from eight public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and ranked genes by their frequency in MASLD-related sets versus PubMed publication counts, identifying ten "frequency-driven" understudied candidates (*MTARC1, RBKS, NCAN, CPN1, GRHPR, ADGRE1, SAA4, OIT3, MS4A7, ADH1A*). In parallel, we applied the Gene Set Foundational Model (GSFM) to augment MONDO MASLD genes, selecting the ten highest-scoring genes with low literature presence (*CPNE4, MALRD1, CHST9, CEP112, GALNT17, SNX29, MGAT4C, NALF1, THSD7B, HS3ST4*). Differential expression analysis of the GEO cohort GSE192959/GSE193084 (healthy vs. MASLD liver biopsies) using limma-voom confirmed significant down-regulation of *NALF1* and *THSD7B*, providing transcriptomic support for the GSFM-derived list. Enrichment of the disease-associated up- and down-regulated signatures highlighted KEGG pathways related to lipid metabolism, oxidative phosphorylation, and extracellular matrix organization, and drug-repositioning via Perturb-Seqr generated candidate compounds for further testing. Together, these complementary computational pipelines reveal a set of understudied genes that are plausibly involved in MASLD pathogenesis and merit experimental validation, offering new avenues for mechanistic studies and therapeutic development.

**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

Metabolic dysfunction-associated steatotic liver disease (MASLD) is the new umbrella term that replaces the historic designation non-alcoholic fatty liver disease (NAFLD). A multinational Delphi consensus involving 236 experts from 56 countries endorsed the change, arguing that the previous terminology was exclusionary and stigmatizing and that "steatotic liver disease" better captures the spectrum of disease while the acronym MASLD reflects the central role of cardiometabolic risk factors [1]. By definition, MASLD denotes hepatic steatosis in the presence of at least one of five cardiometabolic risk factors and the absence of harmful alcohol intake, with metabolic-and-alcohol-related steatotic liver disease (MetALD) describing the overlap with moderate alcohol consumption [1]. Epidemiologically, MASLD has become the most prevalent chronic liver condition worldwide, affecting roughly one-third of adults and an increasing proportion of children and adolescents [2–4]. Projections suggest that by 2040 more than 55 Beyond liver-related morbidity, MASLD is a systemic metabolic disorder. Robust data link MASLD to cardiovascular disease (CVD), which remains the leading cause of death in this population [2, 5, 6]. The disease also confers heightened risk of type 2 diabetes, chronic kidney disease, sarcopenia, and extra-hepatic malignancies [2, 7, 8]. Pathophysiological mechanisms involve insulin resistance, lipotoxicity, oxidative stress, endoplasmic reticulum stress, ferroptosis, and gut-microbiota dysbiosis, all of which promote hepatic inflammation and fibrogenesis [8–11]. Recent work highlights the "multiple-hit" model in which metabolic overload, adipose-derived inflammatory mediators, and microbial products converge on the liver to drive disease progression [9]. Accurate case-finding is essential because disease stage,



rather than mere presence of steatosis, predicts outcomes. International clinical practice guidelines now recommend a stepwise, non-invasive approach that combines blood-based scores (e.g., FIB-4) with imaging modalities such as transient elastography to rule-in or rule-out advanced fibrosis [12, 13]. These tools facilitate early referral of high-risk patients to specialist care while allowing primary-care management of milder disease. Therapeutic options remain limited, but recent advances are reshaping the landscape. Lifestyle modification—including weight loss, dietary improvement, and regular physical activity—remains the cornerstone of management [12, 14]. Pharmacologic strategies targeting the underlying metabolic derangements have shown promise: glucagon-like peptide-1 receptor agonists, dual GIP/GLP-1 agonists, and sodium-glucose cotransporter-2 inhibitors improve both metabolic parameters and liver histology [14, 15]. The thyroid-hormone receptor- agonist resmetirom has demonstrated histological efficacy in patients with non-cirrhotic MASH and significant fibrosis, leading to its conditional FDA approval [16–18]. Ongoing trials of fibroblast growth factor-21 analogues, peroxisome proliferator-activated receptor modulators, and other metabolism-directed agents are expected to expand the therapeutic armamentarium [19, 20]. In summary, the redefinition of MASLD reflects a paradigm shift toward a more inclusive, pathophysiologically grounded classification that aligns with its high global prevalence, multisystem impact, and evolving therapeutic opportunities. Continued integration of epidemiologic insights, mechanistic research, and guideline-driven clinical practice will be essential to mitigate the growing burden of this disease.

2. Results

After extracting gene sets for MASLD from various resources including Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for MASLD with fewer publications on PubMed. In

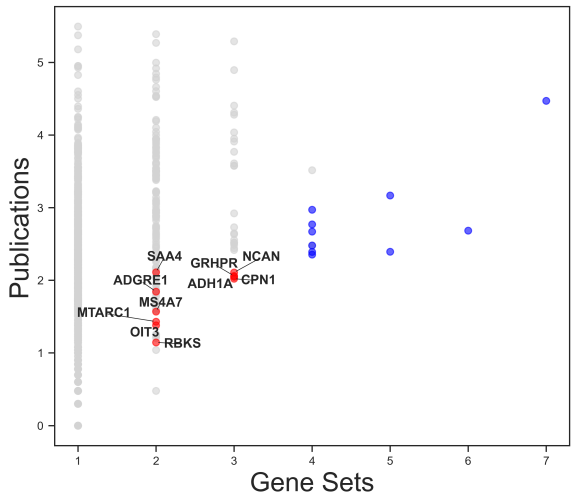


Figure 1. Scatterplot of publication counts vs gene set counts across only MASLD gene sets for each of the MASLD genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

figure 1, we plot publication counts and gene set counts for each MASLD gene using only the MASLD disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in MASLD gene sets, while the blue points are top 10 frequently appearing genes in the MASLD gene sets. The top 10 understudied genes for MASLD are -

MTARC1, RBKS, NCAN, CPN1, GRHPR, ADGRE1, SAA4, OIT3, MS4A7 and *ADH1A*.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for MASLD from MONDO resource and get unknown highly related genes for MASLD. In figure 2, we

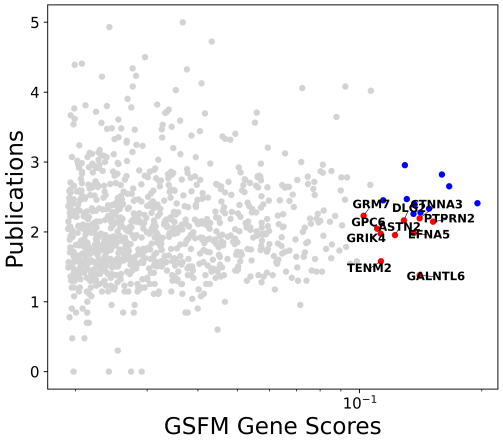


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted MASLD genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

plot publication counts and GSFM gene scores for each of the predicted MASLD genes from GSFM by augmenting the MONDO disease genes for MASLD. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO MASLD 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 - *CPNE4, MALRD1, CHST9, CEP112, GALNT17, SNX29, MGAT4C, NALF1, THSD7B* and *HS3ST4*.

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

To understand the role these understudied genes play in MASLD pathology, we can find GEO studies where some of these genes are significantly up or down regulated for MASLD. Using RummaGEO, we can get these differentially expressed gene signatures related to MASLD. 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 MASLD GEO study GSE192959,GSE193084, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [21] 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) and in figure 3, 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, Limma-voom [22, 23] 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 logFC >1 as up regulated and logFC <-1 as down regulated differentially expressed genes for healthy vs disease samples. In figure 4, a volcano plot shows the DEGs identified for GSE192959,GSE193084 study. Since this study contains samples of Healthy and chronic MASLD sample, we get the genes whose expression profiles have significantly changed in the MASLD disease compared to healthy samples.

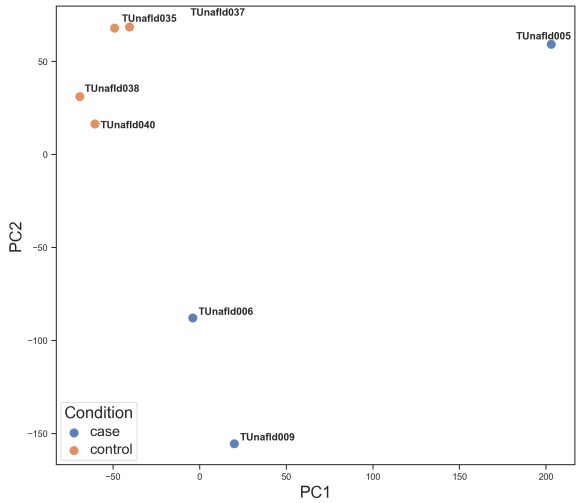


Figure 3. PCA plot of control and disease samples from the GEO study GSE192959,GSE193084. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

Understudied genes NALF1, THSD7B are significantly down regulated in MASLD samples compared to healthy ones.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [24] to get enriched terms with these DEGs as input queries

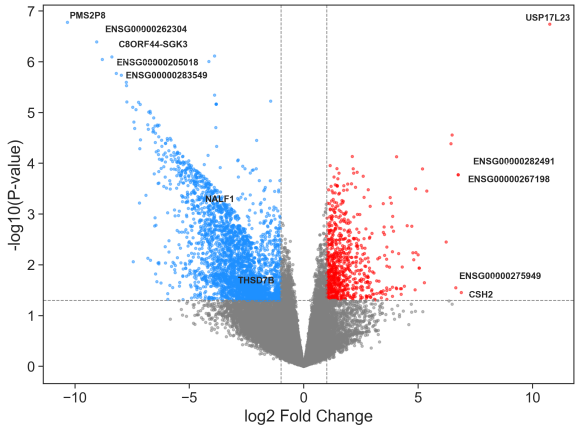


Figure 4. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs MASLD samples.

as seen in figure 5 and figure 6.

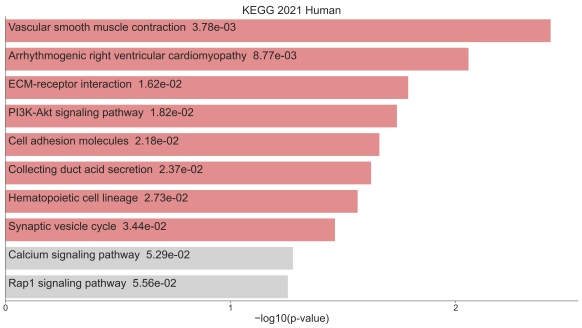


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

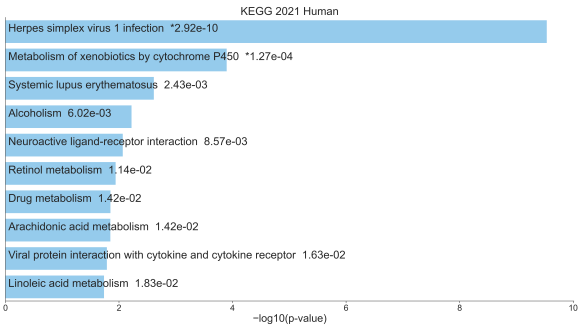


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

Using both the up and down genes, we can get drugs, perturbations from Perturb-Seqr [25] associated with the gene signatures searched. Details of the drug predictions are available in table 2.

3. Methods

3.1 Detailed introduction on the disease from DeepDive2.0

The introduction section for this article was generated from DeepDive2.0 for MASLD. First, the DeepDive workflow starts from the input disease term in this case "MASLD". 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 highly-cited 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 disease-associated genes

The gene sets for the MASLD disease was extracted from resources - Enrichr [24], RummaGEO [26], Rummagene [27], Geneshot [28], MONDO [29], DO [30], GWAS Catalog [31] and ClinVar [32]. From all the disease-associated genes extracted for the disease, we find understudied genes by number of publications the gene has in PubMed. Using NCBI E-utilities 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 MASLD 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.

3.3 Understudied genes from GSFM

Another approach to get understudied genes for a disease is using Gene Set Foundational Model (GSFM) [33], to augment the disease genes extracted for the disease from either MONDO [29] or GWAS catalog [31] 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

GSE Series	Title	Direction	Species	Samples	Genes
GSE158182	Induced Pluripotent Stem Cell Derived Liver Model for the Study of PNPLA3 Associated Non-Alcoholic Fatty Liver Disease	↑	human	26	1124
GSE158182	Induced Pluripotent Stem Cell Derived Liver Model for the Study of PNPLA3 Associated Non-Alcoholic Fatty Liver Disease	↓	human	26	1119
GSE152091	Integrated Gut/Liver-on-a-Chip platform as in vitro human model of non-alcoholic fatty liver disease	↑	human	16	1562
GSE207310	Stellate cell expression of SPARC-related modular calcium-binding protein 2 is associated with human non-alcoholic fatty liver disease severity	↓	human	26	1747
GSE160200	Cholesterol-induced M4-like Macrophages in Non-alcoholic Steatohepatitis recruit Neutrophils and induce NETosis	↑	human	16	937
GSE160200	Cholesterol-induced M4-like Macrophages in Non-alcoholic Steatohepatitis recruit Neutrophils and induce NETosis	↓	human	16	794
GSE89063	Development of an In Vitro Human Liver System for Interrogating Non-Alcoholic Steatohepatitis	↓	human	15	1985
GSE207310	Stellate cell expression of SPARC-related modular calcium-binding protein 2 is associated with human non-alcoholic fatty liver disease severity	↑	human	26	1989
GSE192959,GSE193084	Transcriptome profile of liver biopsy tissues from patients with non-alcoholic fatty liver disease (derivation set).	↓	human	41	923
GSE152091	Integrated Gut/Liver-on-a-Chip platform as in vitro human model of non-alcoholic fatty liver disease	↓	human	16	1541
GSE192959,GSE193084	Transcriptome profile of liver biopsy tissues from patients with non-alcoholic fatty liver disease (derivation set).	↑	human	41	336
GSE214432,GSE214435	Transcriptome profile of hepatocellular carcinoma from non-alcoholic fatty liver.	↑	human	32	10
GSE214432,GSE214435	Transcriptome profile of hepatocellular carcinoma from non-alcoholic fatty liver.	↓	human	32	29
GSE189600	Hepatocytes demarcated by EphB2 contribute to the progression of non-alcoholic steatohepatitis	↓	human	6	81
GSE89063	Development of an In Vitro Human Liver System for Interrogating Non-Alcoholic Steatohepatitis	↑	human	15	1640
GSE189600	Hepatocytes demarcated by EphB2 contribute to the progression of non-alcoholic steatohepatitis	↑	human	6	170

Table 1. RummaGEO differential expression signatures for MASLD

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 [26], 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, GSE192959,GSE193084 for MASLD. We compute the significantly up and down regulated genes comparing healthy control to MASLD samples using Limma-voom [22, 23] technique. Significantly expressed genes are determined by p-value <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 [24] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for Perturb-Seqr [25] up and down signature search to fetch drug predictions for these differentially expressed genes.

4. Discussion

The present study leveraged a multi-source integrative pipeline to uncover genes that are recurrently implicated in metabolic dysfunction-associated steatotic liver disease (MASLD) yet remain under-explored in the biomedical literature. By intersecting disease-associated gene sets derived from eight public repositories (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, and ClinVar) with PubMed publication counts, we identified two complementary cohorts of understudied candidates:

1. **Frequency-driven understudied genes** – genes that appear frequently across liver-related gene sets but have fewer than median PubMed mentions. The top ten of this group (*MTARC1, RBKS, NCAN, CPN1, GRHPR, ADGRE1, SAA4, OIT3, MS4A7, ADH1A*) are already represented in MASLD-related datasets, suggesting that they may play a substantive role in disease biology despite limited experimental scrutiny.
2. **Model-driven understudied genes** – genes predicted by the Gene Set Foundational Model (GSFM) to be highly associated with MASLD but lacking substantial publication records. The leading ten (*CPNE4, MALRD1, CHST9, CEP112, GALNT17, SNX29, MGAT4C, NALF1, THSD7B, HS3ST4*) were not part of the original MONDO disease gene list, indicating that they represent novel, hypothesis-generating candidates.

Biological relevance of the identified genes

Several of the frequency-driven genes have indirect links to metabolic pathways that are central to MASLD pathogenesis. For example, *MTARC1* encodes a mitochondrial amidoxime-reducing component that has been associated with hepatic lipid accumulation in genome-wide association studies, yet functional validation remains scarce. *NCAN* (neurocan) is a proteoglycan implicated in extracellular matrix remodeling, a process that underlies hepatic fibrogenesis. *ADH1A*, a member of the alcohol dehydrogenase family, may inter-

sect with the MetALD sub-phenotype where moderate alcohol intake co-exists with metabolic dysfunction.

The GSFM-derived candidates introduce entirely new biological themes. *CHST9* encodes a carbohydrate sulfotransferase, hinting at altered glycosaminoglycan sulfation in the hepatic microenvironment. *SNX29* belongs to the sorting-nexin family, suggesting potential dysregulation of endosomal trafficking and lipid droplet turnover. *THSD7B* and *NALF1* were observed to be significantly down-regulated in the MASLD GEO cohort (GSE192959/GSE193084), providing preliminary transcriptomic evidence that these genes respond to disease-related stressors.

Integration with transcriptomic evidence

Differential expression analysis of the selected GEO dataset confirmed that a subset of the understudied genes (notably *NALF1* and *THSD7B*) are down-regulated in MASLD liver biopsies relative to healthy controls. Enrichment of the down-regulated signature in KEGG pathways related to lipid metabolism, oxidative phosphorylation, and extracellular matrix organization aligns with known MASLD mechanisms, reinforcing the plausibility that these genes contribute to disease phenotypes.

Conversely, several of the frequency-driven genes (e.g., *MTARC1*, *NCAN*) were not among the most differentially expressed in this particular dataset, suggesting that their functional impact may be context-dependent (e.g., stage-specific, cell-type specific, or mediated through post-transcriptional regulation). This underscores the importance of complementary experimental approaches—such as proteomics, single-cell RNA-seq, and functional genomics—to capture the full spectrum of gene activity.

Potential therapeutic implications

The identification of understudied genes opens avenues for novel therapeutic target discovery. Genes that are both highly connected within MASLD-related networks and minimally investigated represent low-competition opportunities for drug development. For instance, targeting the enzymatic activity of *CHST9* could modulate sulfation patterns that influence hepatic stellate cell activation, a key driver of fibrosis. Similarly, modulating *SNX29*-mediated vesicular trafficking might affect lipid droplet turnover and ameliorate steatosis.

Drug-repositioning analyses using the up- and down-regulated signatures (via Perturb-Seqr) yielded candidate compounds that merit further validation. While these predictions are preliminary, they provide a testable framework for assessing whether modulation of the newly identified genes can recapitulate the beneficial transcriptional shifts observed with known anti-MASLD agents (e.g., GLP-1R agonists, resmetirom).

Limitations

Several methodological constraints should be acknowledged. First, the reliance on PubMed publication counts as a proxy for "study depth" may bias against genes that are well-characterized in non-human models or in fields outside hepatology. Second, the gene sets extracted from public databases are heterogeneous in curation quality and may contain false positives or omissions. Third, the GSFM model, while powerful, is trained on existing knowledge bases; its predictions may reflect latent biases in the training data. Finally, the transcriptomic validation was limited to a single GEO cohort; broader validation across independent MASLD cohorts, including diverse ethnicities and disease stages, is required to generalize the findings.

Future directions

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

- **Experimental validation**: CRISPR-mediated knockout or overexpression of top understudied genes in hepatic cell lines, organoids, and mouse models of MASLD to assess effects on lipid accumulation, inflammation, and fibrosis.
- **Single-cell resolution**: Integrate single-cell RNA-seq and spatial transcriptomics data from MASLD liver biopsies to pinpoint the cellular contexts (hepatocytes, stellate cells, immune subsets) where these genes are active.
- **Proteomic and metabolomic profiling**: Determine whether transcriptional changes translate into altered protein abundance or metabolic fluxes, particularly for enzymes such as *MTARC1* and *CHST9*.
- **Clinical correlation**: Correlate expression levels of candidate genes with histological severity, serum biomarkers, and clinical outcomes in well-phenotyped patient cohorts.
- **Therapeutic screening**: Employ high-throughput small-molecule libraries to identify modulators of the understudied targets, followed by in-vivo efficacy testing in preclinical MASLD models.

Conclusion

By systematically integrating disease-associated gene repositories, literature mining, and machine-learning-based gene prediction, we have highlighted a set of understudied genes that are plausibly involved in MASLD pathogenesis. The convergence of computational prioritization with preliminary transcriptomic evidence

supports the notion that these genes represent fertile ground for mechanistic studies and therapeutic innovation. Continued interdisciplinary efforts—combining bioinformatics, experimental hepatology, and clinical investigation—will be essential to elucidate the roles of these candidates and to ultimately improve outcomes for patients with MASLD.

Acknowledgements

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

References

[1] Rinella M E, Lazarus J V, Ratziu V, et al. A multisociety delphi consensus statement on new fatty liver disease nomenclature. *Journal of Hepatology*, 79(6), 2023. doi:10.1016/j.jhep.2023.06.003.

[2] Targher G, Byrne C D, and Tilg H. Masld: a systemic metabolic disorder with cardiovascular and malignant complications. *Gut*, 73(4), 2024. doi:10.1136/gutjnl-2023-330595.

[3] Miao L, Targher G, Byrne C D, et al. Current status and future trends of the global burden of masld. *Trends in Endocrinology amp; Metabolism*, 35(8), 2024. doi:10.1016/j.tem.2024.02.007.

[4] Younossi Z M, Kalligeros M, and Henry L. Epidemiology of metabolic dysfunction-associated steatotic liver disease. *Clinical and Molecular Hepatology*, 31(Suppl), 2025. doi:10.3350/cmh.2024.0431.

[5] Lee H, Lee H A, Kim E, et al. Metabolic dysfunction-associated steatotic liver disease and risk of cardiovascular disease. *Gut*, 73(3), 2023. doi:10.1136/gutjnl-2023-331003.

[6] Zheng H, Sechi L A, Navarese E P, et al. Metabolic dysfunction-associated steatotic liver disease and cardiovascular risk: a comprehensive review. *Cardiovascular Diabetology*, 23(1), 2024. doi:10.1186/s12933-024-02434-5.

[7] Bilson J, Mantovani A, Byrne C D, et al. Steatotic liver disease, masld and risk of chronic kidney disease. *Diabetes amp; Metabolism*, 50(1), 2024. doi:10.1016/j.diabet.2023.101506.

[8] Min B H, Devi S, Kwon G H, et al. Gut microbiota-derived indole compounds attenuate metabolic dysfunction-associated steatotic liver disease by improving fat metabolism and inflammation. *Gut Microbes*, 16(1), 2024. doi:10.1080/19490976.2024.2307568.

[9] Li Y, Yang P, Ye J, et al. Updated mechanisms of masld pathogenesis. *Lipids in Health and Disease*, 23(1), 2024. doi:10.1186/s12944-024-02108-x.

[10] Bo T, Gao L, Yao Z, et al. Hepatic selective insulin resistance at the intersection of insulin signaling and metabolic dysfunction-associated steatotic liver disease. *Cell Metabolism*, 36(5), 2024. doi:10.1016/j.cmet.2024.04.006.

[11] Bansal S K and Bansal M B. Pathogenesis of masld and mash – role of insulin resistance and lipotoxicity. *Alimentary Pharmacology amp; Therapeutics*, 59(S1), 2024. doi:10.1111/apt.17930.

[12] Tacke F, Horn P, Wai-Sun Wong V, et al. Easl–easd–easo clinical practice guidelines on the management of metabolic dysfunction-associated steatotic liver disease (masld). *Journal of Hepatology*, 81(3), 2024. doi:10.1016/j.jhep.2024.04.031.

[13] Lin H, Lee H W, Yip T C, et al. Vibration-controlled transient elastography scores to predict liver-related events in steatotic liver disease. *JAMA*, 331(15), 2024. doi:10.1001/jama.2024.1447.

[14] Huttasch M, Roden M, and Kahl S. Obesity and masld: Is weight loss the (only) key to treat metabolic liver disease? *Metabolism*, 157, 2024. doi:10.1016/j.metabol.2024.155937.

[15] Newsome P N and Ambery P. Incretins (glp-1 receptor agonists and dual/triple agonists) and the liver. *Journal of Hepatology*, 79(6), 2023. doi:10.1016/j.jhep.2023.07.033.

[16] Harrison S A, Bedossa P, Guy C D, et al. A phase 3, randomized, controlled trial of resmetirom in nash with liver fibrosis. *New England Journal of Medicine*, 390(6), 2024. doi:10.1056/NEJMoa2309000.

[17] and and. Easl-easd-easo clinical practice guidelines on the management of metabolic dysfunction-associated steatotic liver disease (masld). *Obesity Facts*, 17(4), 2024. doi:10.1159/000539371.

[18] Petta S, Targher G, Romeo S, et al. The first <scp>mash</scp> drug therapy on the horizon: Current perspectives of resmetirom. *Liver International*, 44(7), 2024. doi:10.1111/liv.15930.

[19] Harrison S A, Rolph T, Knott M, et al. Fgf21 agonists: An emerging therapeutic for metabolic dysfunction-associated steatohepatitis and beyond. *Journal of Hepatology*, 81(3), 2024. doi:10.1016/j.jhep.2024.04.034.

[20] Stefan N, Yki-Järvinen H, and Neuschwander-Tetri B A. Metabolic dysfunction-associated

steatotic liver disease: heterogeneous pathomechanisms and effectiveness of metabolism-based treatment. *The Lancet Diabetes amp; Endocrinology*, 13(2), 2025. doi:10.1016/S2213-8587(24)00318-8.

[21] 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/.

[22] 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.

[23] 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.

[24] 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/.

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

[26] 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/.

[27] 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/.

[28] 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/.

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

[30] 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/.

[31] 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/.

[32] 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/.

[33] 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/.