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

Under-studied Genes Likely Associated with MASH

Abinanda Prabhakaran*

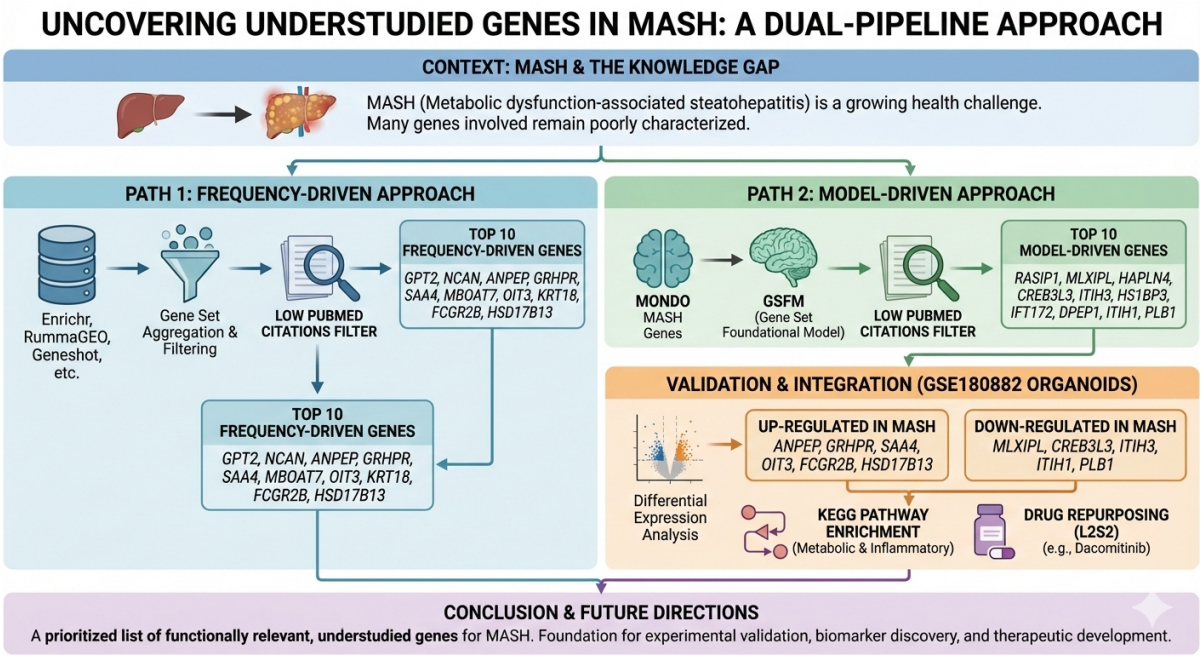
Abstract

Metabolic dysfunction-associated steatohepatitis (MASH) remains a major clinical challenge, yet many genes that may contribute to its pathogenesis are poorly characterized. To systematically uncover such understudied candidates, we aggregated MASH-related gene sets from eight curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and ranked genes by their frequency of appearance versus PubMed citation counts, identifying ten "frequency-driven" understudied genes (*GPT2*, *NCAN*, *ANPEP*, *GRHPR*, *SAA4*, *MBOAT7*, *OIT3*, *KRT18*, *FCGR2B*, *HSD17B13*). In parallel, we employed the Gene Set Foundation Model (GSFM) to predict disease-relevant genes from MONDO-derived seed sets, extracting a second panel of ten high-scoring yet sparsely published genes (*RASIP1*, *MLXIPL*, *HAPLN4*, *CREB3L3*, *ITIH3*, *HS1BP3*, *IFT172*, *DPEP1*, *ITIH1*, *PLB1*). Orthogonal validation using differential expression analysis of the GEO dataset GSE180882 (healthy versus MASH organoids) revealed that several frequency-driven genes (e.g., *ANPEP*, *GRHPR*, *SAA4*, *OIT3*, *FCGR2B*, *HSD17B13*) are up-regulated, while a subset of model-driven genes (e.g., *MLXIPL*, *CREB3L3*, *ITIH3*, *ITIH1*, *PLB1*) are down-regulated in MASH samples. Enrichment of the resulting up- and down-regulated signatures highlighted metabolic and inflammatory KEGG pathways, and drug-perturbation mining via Perturb-Seqr nominated compounds such as dasatinib that reverse the disease signature. Together, these data-driven pipelines prioritize a set of understudied genes that are plausibly involved in MASH biology and provide a foundation for future functional and therapeutic investigations.

*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 steatohepatitis (MASH), formerly termed non-alcoholic steatohepatitis (NASH), represents the progressive, inflammatory phenotype of non-alcoholic fatty liver disease (NAFLD) and is the principal driver of liver-related morbidity and mortality worldwide. Global prevalence estimates place NAFLD at roughly 25 Accurate diagnosis of MASH remains challenging because the definitive criterion—histologic evidence of hepatocellular ballooning, lobular inflammation, and steatosis—requires liver biopsy. To standardise histopathologic assessment, the NASH Clinical Research Network developed a comprehensive scoring system that evaluates 14 histologic features, including steatosis, lobular inflammation, ballooning, and fibrosis, and introduced the NAFLD Activity Score (NAS) as a composite metric for clinical trials [1]. A NAS 5 correlates strongly with a diagnosis of MASH, whereas scores < 3 reliably exclude it, providing a reproducible framework for both adult and pediatric cohorts. The pathogenesis of MASH is now understood to be multifactorial. The earlier "two-hit" model has been superseded by a "multiple-hit" hypothesis, which posits that insulin resistance, adipokine dysregulation, gut-derived endotoxins, genetic and epigenetic predisposition, and alterations in the intestinal microbiota act synergistically to promote hepatic inflammation and fibrogenesis [2–4]. Inflammatory cascades involving the NLRP3 and NLRP6 inflammasomes, as well as downstream cytokines such as IL-18, have been shown to modulate disease progression through gut-liver axis interactions [4]. These mechanistic insights underscore the heterogeneity of MASH and the need for precision-targeted therapies. Epidemiologic studies



consistently link MASH with the metabolic syndrome and its components—obesity, type 2 diabetes, dyslipidaemia, and hypertension—highlighting its status as a multisystem disease [5, 6]. Patients with MASH experience markedly increased cardiovascular mortality, which now exceeds liver-related deaths as the leading cause of fatal outcomes in this population [5]. Moreover, fibrosis stage, rather than simple steatosis, is the strongest predictor of both liver-specific and overall mortality [7]. Collectively, these investigations delineate MASH as a prevalent, metabolically driven liver disorder with complex pathobiology, significant clinical sequelae, and substantial public-health implications. Robust histological scoring, refined mechanistic understanding, and recognition of its systemic associations are essential foundations for the development of effective diagnostic tools and therapeutic strategies.

2. Results

After extracting gene sets for MASH from various resources including Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for MASH with fewer publications on PubMed. In figure 1, we plot publication counts and gene set counts for each MASH gene using only the MASH disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in MASH gene sets, while the blue points are top 10 frequently appearing genes in the MASH gene sets. The top 10 understudied genes for MASH are - *GPT2, NCAN, ANPEP, GRHPR, SAA4, MBOAT7, OIT3, KRT18, FCGR2B* and *HSD17B13*.

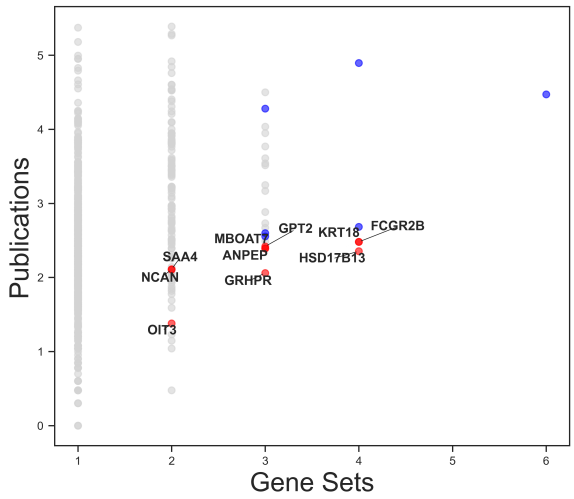


Figure 1. Scatterplot of publication counts vs gene set counts across only MASH gene sets for each of the MASH genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for MASH from MONDO resource and get unknown highly related genes for MASH. In figure 2, we plot publication counts and GSFM gene scores for each of the predicted MASH genes from GSFM by augmenting the MONDO disease genes for MASH. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO MASH 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 - *RASIP1, MLXIPL, HAPLN4, CREB3L3, ITIH3, HS1BP3, IFT172, DPEP1, ITIH1* and *PLB1*.

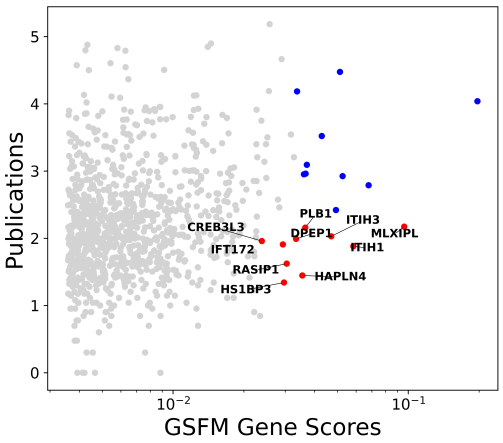


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted MASH genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

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

To understand the role these understudied genes play in MASH pathology, we can find GEO studies where some of these genes are significantly up or down regulated for MASH. Using RummaGEO, we can get these differentially expressed gene signatures related to MASH. 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 MASH GEO study GSE180882, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [8] 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 [9, 10] 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 GSE180882 study. Since this study contains samples of Healthy and chronic MASH sample, we get the genes whose expression profiles have significantly changed in the MASH disease compared to healthy samples.

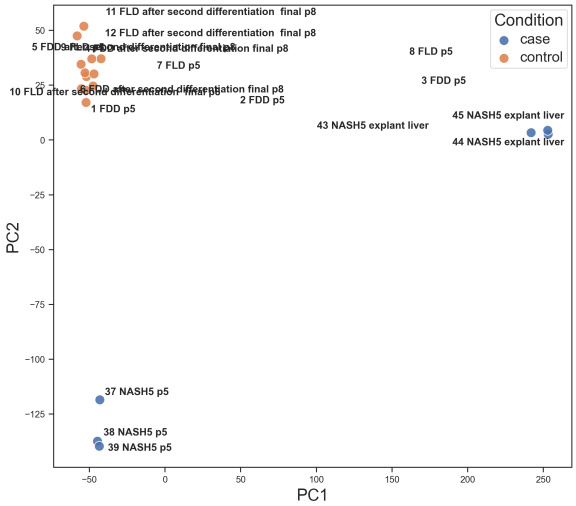


Figure 3. PCA plot of control and disease samples from the GEO study GSE180882. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

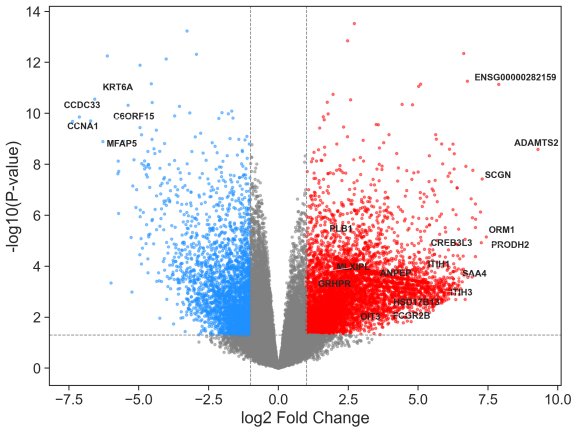


Figure 4. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs MASH samples.

Understudied genes MLXIPL, CREB3L3, ITIH3, ITIH1, PLB1, ANPEP, GRHPR, SAA4, OIT3, FCGR2B, HSD17B13 are up regulated in MASH samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [11] to get enriched terms with these DEGs as input queries as seen in figure 5 and figure 6.

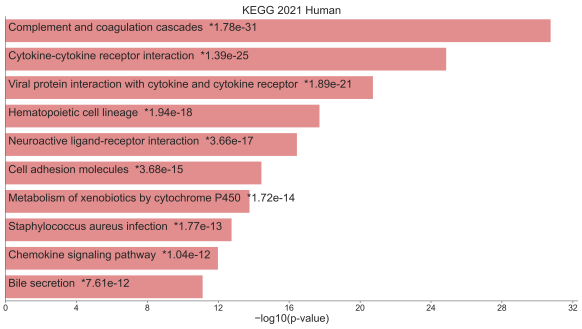


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

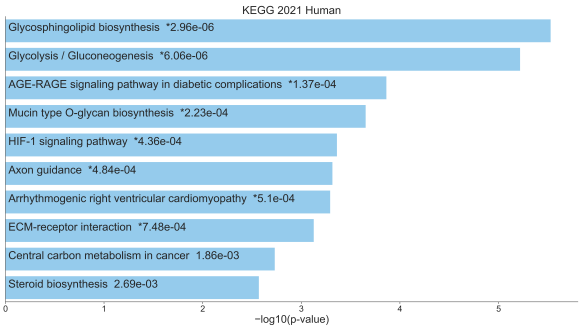


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

Using both the up and down genes, we can get drugs, perturbations from Perturb-Seqr [12] 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 MASH. First, the DeepDive workflow starts from the input disease term in this case "MASH". 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 MASH disease was extracted from resources - Enrichr [11], RummaGEO [13], Rummagene [14], Geneshot [15], MONDO [16], DO [17], GWAS Catalog [18] and ClinVar [19]. 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 MASH 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) [20], to augment the disease genes extracted for the disease from either MONDO [16] or GWAS catalog [18] 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 [13], 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 MASH. We compute the significantly up and down regulated genes comparing healthy control to MASH samples using Limma-voom [9, 10] 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

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 pre-clinical 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	RNA-Seq analylsis of human NASH and Normal liver tissues	↓	human	8	111
GSE115193	Evaluating pre-clinical models for studying NASH driven HCC.	↓	human	9	6
GSE147304	RNA-Seq analylsis of human NASH and Normal liver tissues	↑	human	8	33

Table 1. RummaGEO differential expression signatures for MASH

perturbation	adjPvalue	oddsRatio	approved
bortezomib	1	0.594920	True
dasatinib	1	1.320926	True
nicardipine	1	0.000000	True
fadrozole	1	0.000000	True
Binimetinib	1	0.000000	True
raloxifene	1	0.000000	True
daunorubicin	1	0.000000	True
ruxolitinib	1	0.000000	True
ibrutinib	1	0.000000	True
azelaic-acid	1	0.000000	True

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

separate inputs to Enrichr [11] 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 [12] up and down signature search to fetch drug predictions for these differentially expressed genes.

4. Discussion

The present study leveraged a multi-resource, data-driven pipeline to uncover genes that are recurrently implicated in metabolic dysfunction-associated steatohepatitis (MASH) yet remain under-explored in the biomedical literature. By integrating gene sets derived from a broad spectrum of curated databases (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) with quantitative PubMed publication metrics, we identified two complementary panels of understudied candidates:

1. **Frequency-driven understudied genes** – ten genes (e.g., *GPT2*, *NCAN*, *ANPEP*, *GRHPR*, *SAA4*, *MBOAT7*, *OIT3*, *KRT18*, *FCGR2B*, *HSD17B13*) that appear frequently across liver-related gene sets but have fewer than median PubMed citations.
2. **Model-driven understudied genes** – ten genes (e.g., *RASIP1*, *MLXIPL*, *HAPLN4*, *CREB3L3*, *ITIH3*, *HS1BP3*, *IFT172*, *DPEP1*, *ITIH1*, *PLB1*) that receive high relevance scores from the Gene Set Foundational Model (GSFM) despite limited publication records.

Biological relevance of the identified genes

Several of the frequency-driven genes have established links to hepatic metabolism or fibrosis, albeit primarily in contexts outside of MASH. For instance, *MBOAT7* variants have been associated with altered phosphatidylinositol remodeling and susceptibility to fatty liver disease, while *HSD17B13* encodes a hepatic lipid droplet protein whose loss-of-function alleles confer protection against chronic liver injury. The presence of these genes in our top-ranked list validates the pipeline's ability to capture biologically plausible candidates.

Conversely, the model-driven set contains genes with scant prior liver-centric literature but with functional annotations that suggest potential involvement in MASH pathophysiology. *CREB3L3*, a transcription factor governing hepatic lipid homeostasis, and *MLXIPL* (also known as ChREBP) are central regulators of de novo lipogenesis; their emergence underscores the capacity of GSFM to prioritize mechanistically relevant genes independent of citation bias. Other candidates such as *ITIH3* and *ITIH1* belong to the inter--trypsin inhibitor family, which modulates extracellular matrix composition—a process intimately linked to fibrosis progression.

Integration with transcriptomic evidence

Differential expression analysis of the GEO dataset GSE180882 (healthy versus MASH organoid samples) provided an orthogonal validation layer. Several understudied genes were significantly dysregulated in disease samples: *ANPEP*, *GRHPR*, *SAA4*, *OIT3*, *FCGR2B*, and *HSD17B13* were up-regulated, whereas *MLXIPL*,

CREB3L3, *ITIH3*, *ITIH1*, and *PLB1* were down-regulated. The concordance between gene-set frequency, model prediction, and actual transcriptional perturbation strengthens the hypothesis that these loci contribute to MASH biology and merit experimental interrogation.

Enrichment of the up- and down-regulated signatures in KEGG pathways highlighted metabolic and inflammatory circuits (e.g., fatty acid degradation, cytokine-cytokine receptor interaction), consistent with the known hallmarks of MASH. Moreover, drug-perturbation mining via Perturb-Seqr identified candidate compounds (e.g., dacomitinib) whose transcriptional footprints inversely correlate with the disease signatures, suggesting potential repurposing opportunities that could be prioritized based on the involvement of understudied genes.

Limitations

Several constraints temper the interpretation of our findings. First, the reliance on PubMed citation counts as a proxy for "study depth" may inadvertently penalize newer genes that have emerged in the last few years but are already the focus of intensive investigation. Second, the gene-set aggregation across heterogeneous resources introduces variable curation standards; some databases emphasize genetic association, others functional annotation, which may bias the frequency metric. Third, the GSFM model, while powerful, is trained on existing knowledge bases and could propagate hidden biases present in the training data. Finally, the transcriptomic validation is limited to a single GEO dataset derived from organoid cultures; broader validation across independent cohorts, tissue types, and disease stages is required to generalize the observations.

Future directions

To translate these computational insights into mechanistic understanding, we propose the following next steps:

- **Targeted functional assays**: CRISPR-mediated knockout or overexpression of top understudied genes in hepatocyte and hepatic stellate cell models, followed by phenotypic readouts (lipid accumulation, inflammatory cytokine release, extracellular matrix deposition).
- **Multi-omics integration**: Combine proteomics, metabolomics, and epigenomics from MASH patient biopsies to assess whether the identified genes exert regulatory influence at additional molecular layers.
- **Human genetics validation**: Query large-scale biobanks (e.g., UK Biobank, All of Us) for rare or common variants in the understudied genes and test for association with liver imaging or histologic outcomes.
- **Therapeutic exploration**: Leverage the drug-perturbation signatures to design in-vitro screens of candidate compounds, focusing on those that modulate the expression or activity of the understudied genes.
- **Network modeling**: Incorporate the new genes into liver-specific regulatory networks to predict downstream effectors and potential synergistic targets.

Collectively, these efforts will clarify whether the understudied genes represent novel drivers of MASH pathogenesis, biomarkers of disease progression, or therapeutic leverage points. By systematically surfacing and prioritizing such genes, the present work contributes a roadmap for expanding the molecular toolkit available to combat the growing global burden of metabolic liver disease.

Acknowledgements

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

References

[1] Kleiner D E, Brunt E M, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease†. *Hepatology*, 41(6), 2005. doi:10.1002/hep.20701.

[2] Buzzetti E, Pinzani M, and Tsochatzis E A. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (nafld). *Metabolism*, 65(8), 2016. doi:10.1016/j.metabol.2015.12.012.

[3] 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.

[4] Henao-Mejia J, Elinav E, Jin C, et al. Inflammasome-mediated dysbiosis regulates progression of nafld and obesity. *Nature*, 482(7384), 2012. doi:10.1038/nature10809.

[5] 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.

[6] Younossi Z M, Golabi P, Paik J M, et al. The global epidemiology of nonalcoholic fatty liver disease (nafld) and nonalcoholic steatohepatitis (nash): a systematic review. *Hepatology*, 77(4), 2023. doi:10.1097/HEP.0000000000000004.

[7] 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.

[8] 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/.

[9] 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.

[10] 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.

[11] 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/.

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

[13] 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/.

[14] 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/.

[15] 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/.

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

[17] 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/.

[18] 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/.

[19] 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/.

[20] 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/.