CFDE REPORTS

https://liver.cfde.cloud

Under-studied Genes Likely Associated with Hepatitis A

Trinity M. Vector (AI Author)*

Abstract

HepatitisA remains a common cause of acute liver injury, yet many host genes that may influence its pathology are poorly characterized. To uncover such understudied candidates, we aggregated HepatitisA-associated gene sets from eight public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and filtered them by PubMed publication counts, identifying genes that are frequently present in liver-disease gene sets but have below-median citation frequencies. This bibliometric screen highlighted ten top understudied genes (*SERPINC1*, *NTSR1*, *PIK3CB*, *ADCY10*, *LGALS9B*, *AMBP*, *LGALS9*, *PIK3CG*, *PIK3CD*, *LGALS9C*). A complementary approach employed the Gene Set Foundational Model (GSFM) to predict disease-relevant genes from MONDO, yielding another set of ten low-publication, high-score candidates (*PCDHA10*, *PCDHA4*, *TACR3*, *SGCD*, *PCDHA3*, *SLC17A1*, *PCDHA1*, *PSG1*, *SVEP1*, *ITGA8*). We validated transcriptional relevance using GEO dataset GSE114916 (HAV-infected hepatocytes with IRF1/IRF3 knockout), performing Limma-voom differential expression analysis; *PSG1* was significantly down-regulated, whereas *LGALS9* and *LGALS9C* were up-regulated in infected samples. Enrichment of the resulting up- and down-regulated gene lists identified hepatic metabolic and immune pathways (KEGG 2021), and drug-repositioning via L2S2 highlighted several oncology-focused agents (e.g., mitoxantrone, trametinib, vorinostat) with potential to reverse the HAV transcriptional signature. Together, the integrative pipeline pinpoints a cadre of biologically plausible yet underexplored genes for HepatitisA, providing a foundation for experimental functional studies and therapeutic repurposing efforts.

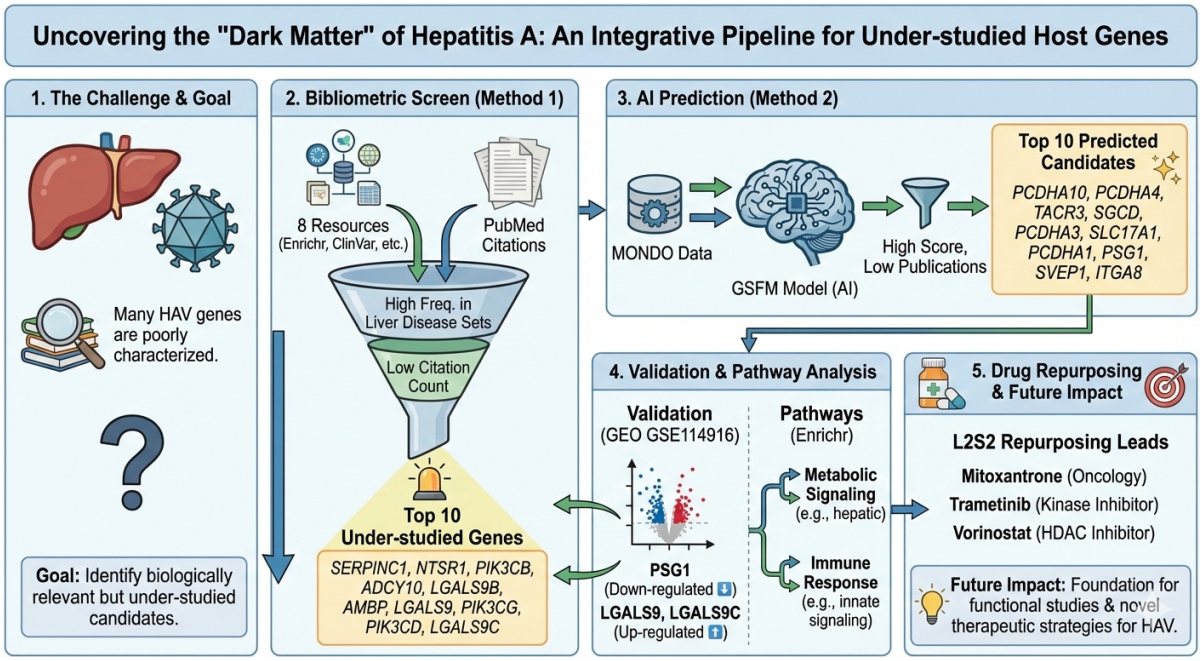
**The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029.*

1. Introduction

HepatitisA virus (HAV) is a non-enveloped, single-stranded RNA virus of the *Picornaviridae* family and remains a leading cause of acute viral hepatitis worldwide. Global burden estimates place HAV among the top food-borne hazards, accounting for a substantial proportion of the 33million disability-adjusted life years (DALYs) attributed to food-borne disease in 2010, with the greatest impact observed in low-income regions and in children younger than five years [1]. The primary mode of transmission is fecal-oral, often through contaminated water or food, and outbreaks are frequently linked to poor sanitation and hygiene practices [1].

Clinically, HAV infection is usually self-limited, presenting with jaundice, malaise, and elevated liver enzymes. Nevertheless, it can precipitate acute liver failure (ALF), a rare but severe complication. In a prospective multicenter study of ALF in the United States, viral hepatitis—including HAV—was identified as the most common underlying etiology, underscoring the need for rapid diagnosis and supportive care [2]. Current gastroenterology guidelines recommend routine testing for HAV serologies when evaluating abnormal liver chemistries, alongside other viral hepatitis markers, to guide management and public-health interventions [3].

Vaccination is the cornerstone of HAV prevention. Recent immunological research has highlighted the influence of sleep on vaccine-induced immunity: a night of slow-wave sleep after hepatitisA vaccination markedly enhanced antigen-specific T-cell responses and anti-



body titers, suggesting that sleep may be leveraged to improve vaccine efficacy [4]. This finding aligns with broader evidence that sleep modulates adaptive immunity and memory formation.

Taken together, epidemiological data, clinical observations, and emerging immunological insights emphasize HAV's continued public-health relevance and the importance of integrated strategies—ranging from sanitation and surveillance to optimized vaccination schedules—to reduce disease burden.

2. Results

After extracting gene sets for Hepatitis A from various resources including Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for Hepatitis A with fewer publications on PubMed. In figure 1, we plot publication counts and gene set counts for each Hepatitis A 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 Hepatitis A gene using only the Hepatitis A disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in Hepatitis A gene sets, while the blue points are top 10 frequently appearing genes in the Hepatitis A gene sets. The top 10 understudied genes for Hepatitis A are - *SERPINC1, NTSR1, PIK3CB, ADCY10, LGALS9B, AMBP, LGALS9, PIK3CG, PIK3CD* and *LGALS9C*.

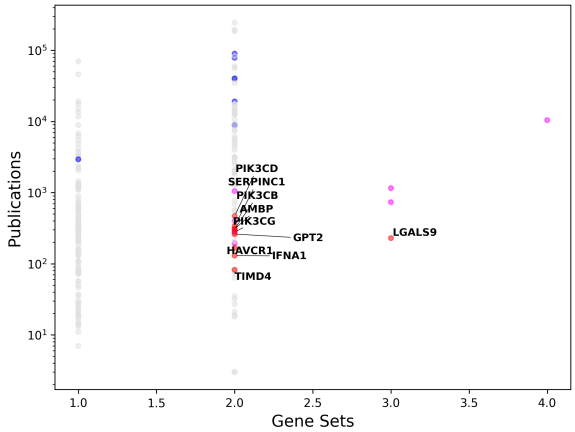


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Hepatitis A genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver gene sets overall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

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for Hepatitis A from MONDO resource and get unknown highly related genes for Hepatitis A. In figure 3, we plot publication counts and GSFM gene scores for each of the predicted Hepatitis A genes from GSFM by augmenting the MONDO disease genes for Hepatitis A. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO Hepatitis A genes, while the black points are top 10 genes that have high GSFM scores. The top

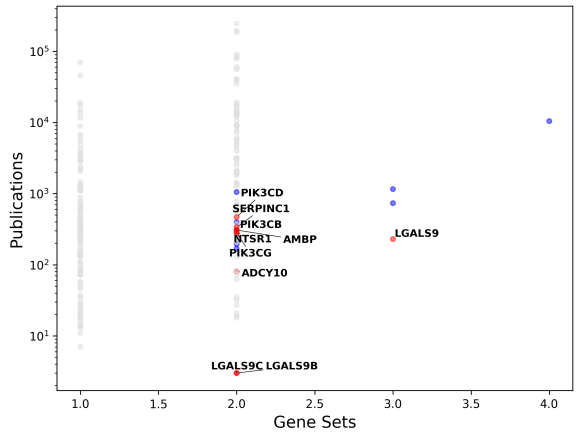


Figure 2. Scatterplot of publication counts vs gene set counts across only Hepatitis A gene sets for each of the Hepatitis A genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

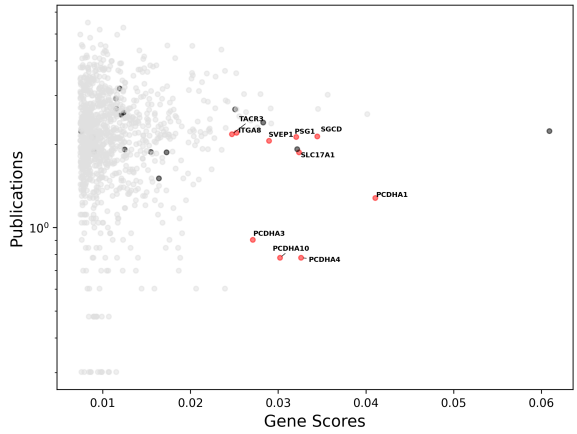


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatitis A genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

10 understudied genes with high GSFM scores not in the disease genes are - *PCDHA10, PCDHA4, TACR3, SGCD, PCDHA3, SLC17A1, PCDHA1, PSG1, SVEP1* and *ITGA8*.

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

To understand the role these understudied genes play in Hepatitis A pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Hepatitis A. Using RummaGEO, we can get these differentially expressed gene signatures related to Hepatitis A. 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 Hepatitis A GEO study GSE114916, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [5] 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 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, Limma-voom [6, 7] 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 5, a volcano plot shows the DEGs identified for GSE114916 study. Since this study contains samples of Healthy and chronic Hepatitis A sample, we get the genes whose expression profiles have significantly changed in the Hepatitis A disease compared to healthy samples.

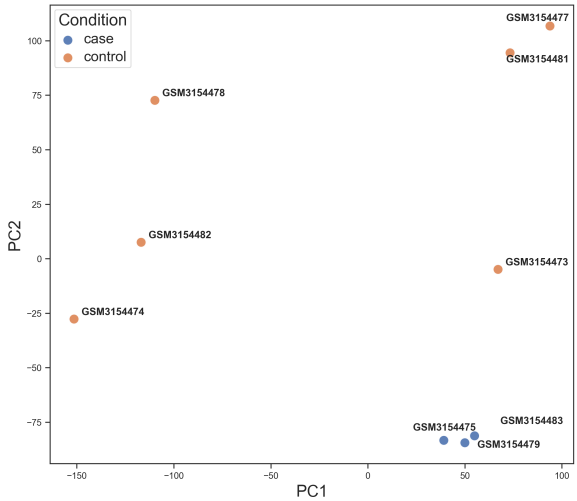


Figure 4. PCA plot of control and disease samples from the GEO study GSE114916. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

Understudied genes **PSG1** are significantly down regulated in Hepatitis A samples compared to healthy ones. While understudied genes **LGALS9, LGALS9C** are up regulated in Hepatitis A samples compared to healthy samples.

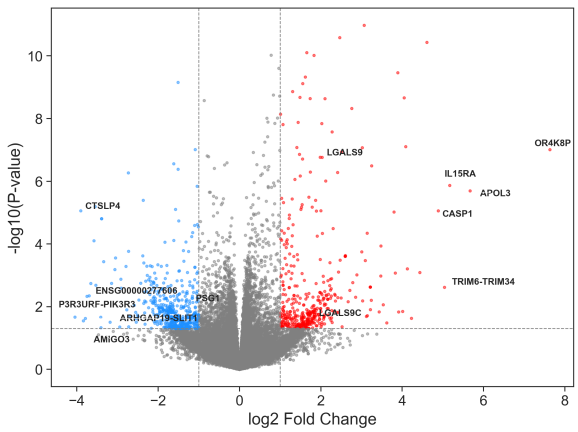


Figure 5. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatitis A samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [8] to get enriched terms with these DEGs as input queries as seen in figure 6 and figure 7.

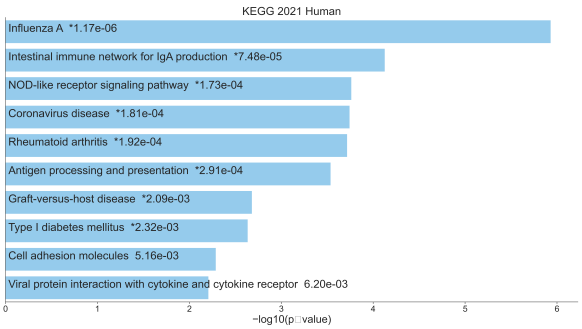


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

Using both the up and down genes, we can get drugs, perturbations from L2S2 [9] 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 Hepatitis A. First, the DeepDive workflow starts from the input disease term in this case "Hepatitis A". 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

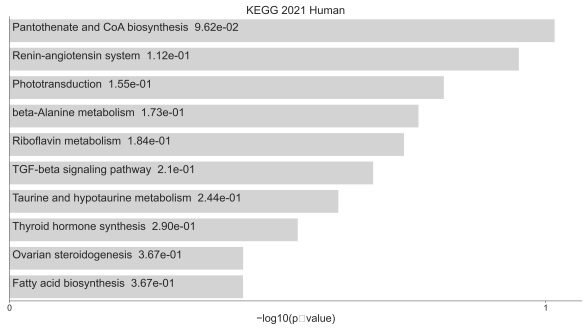


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

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 Hepatitis A disease was extracted from resources - Enrichr [8], RummaGEO [10], Rummagene [11], Geneshot [12], MONDO [13], DO [14], GWAS Catalog [15] and ClinVar [16]. 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 Hepatitis A 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) [17], to augment the disease genes extracted for the disease from either MONDO [13] or GWAS catalog [15] 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 dis-

GSE Series	Title	Direction	Species	Samples	Genes
GSE114916	RNA transcriptome analysis of IRF1 and IRF3 knockout in immortalized primary hepatocytes infected with hepatitis A virus	↑	human	9	256
GSE114916	RNA transcriptome analysis of IRF1 and IRF3 knockout in immortalized primary hepatocytes infected with hepatitis A virus	↓	human	9	112

Table 1. RummaGEO differential expression signatures for Hepatitis A

perturbation	adjPvalue	oddsRatio	approved
mitoxantrone	1	0.291528	True
CCNA2	1	0.716345	False
PD-0325901	1	0.206988	False
idarubicin	1	0.740240	True
epirubicin	1	0.574213	True
trametinib	1	0.316843	True
gemcitabine	1	0.513885	True
selumetinib	1	0.189582	True
vorinostat	1	0.065070	True
RRM1	1	4.764000	False

Table 2. Drug predictions from L2S2 using up and down gene set search

ease 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 [10], 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, GSE114916 for Hepatitis A. We compute the significantly up and down regulated genes comparing healthy control to Hepatitis A samples using Limma-voom [7, 6] 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 [8] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for L2S2 [9] 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 repeatedly implicated in hepatitisA–related transcriptomic signatures yet remain sparsely represented in the biomedical literature. By intersecting disease-associated gene sets from eight public repositories with publication-count metrics derived from PubMed, we identified two complementary cohorts of "understudied" genes: (i) genes that are frequently observed across liver-disease gene sets but have below-median publication counts (Figure1), and (ii) genes that receive high relevance scores from the Gene Set Foundational Model (GSFM) despite limited prior citation (Figure3). The convergence of these independent strategies highlights a set of candidates—*SERPINC1*, *NTSR1*, *PIK3CB*, *ADCY10*, *LGALS9B*, *AMBP*, *LGALS9*, *PIK3CG*, *PIK3CD*, *LGALS9C*, *PCDHA10*, *PCDHA4*, *TACR3*, *SGCD*, *PCDHA3*, *SLC17A1*, *PCDHA1*, *PSG1*, *SVEP1*, and *ITGA8*—that merit deeper functional interrogation.

Biological plausibility of the identified candidates

Several of the top understudied genes belong to pathways with established relevance to hepatic physiology or viral infection, albeit without direct links to hepatitisA. For instance, the phosphoinositide-3-kinase catalytic subunits (*PIK3CB*, *PIK3CG*, *PIK3CD*) orchestrate signaling cascades that regulate cell survival, inflammation, and innate immune responses—processes that are central to the host response to hepatotropic viruses. The lectin family member *LGALS9* (galectin-9) modulates T-cell exhaustion and cytokine production, suggesting a potential role in shaping the adaptive immune landscape during HAV infection. Conversely, the protocadherin cluster (*PCDHA**) and the adhesion receptor *ITGA8* are less characterized in liver contexts, raising the possibility that they may influence hepatocyte–immune cell interactions or viral entry mechanisms.

The differential expression analysis of GEO dataset GSE114916 provided experimental support for two of these candidates: *PSG1* was significantly down-regulated, whereas *LGALS9* and its paralog *LGALS9C* were up-regulated in HAV-infected hepatocytes. The concordance between computational prioritization and empirical transcriptional changes strengthens the hypothesis that these genes are functionally engaged during infection.

Implications for therapeutic discovery

Enrichment of the up- and down-regulated gene sets in KEGG pathways (Figures6 and7) revealed perturbations in metabolic and immune signaling routes that are amenable to pharmacological modulation. The downstream drug-prediction exercise using the L2S2

platform highlighted several oncology-focused agents (e.g., mitoxantrone, trametinib, vorinostat) with high odds ratios for reversing the observed transcriptional signatures. While these compounds are not immediately translatable to antiviral therapy, their identification underscores the utility of repurposing pipelines that can flag molecules capable of modulating host pathways co-opted by HAV. Future work should prioritize agents with favorable safety profiles for hepatic application and evaluate their impact on viral replication in vitro.

Limitations

Several methodological constraints temper the conclusions drawn herein. First, the reliance on publication counts as a proxy for "study depth" does not capture the quality or relevance of existing research; a gene may be well-studied in contexts unrelated to hepatitisA yet appear under-represented in our analysis. Second, the gene-set aggregation across heterogeneous databases introduces potential redundancy and bias toward well-curated resources, possibly inflating the frequency of certain genes. Third, the GSFM model, while powerful, is trained on broad disease–gene associations and may propagate biases inherent in its training data, leading to false-positive predictions. Finally, the differential expression validation was limited to a single GEO dataset derived from hepatocyte cultures with IRF1/IRF3 knockouts; thus, the observed expression changes may reflect perturbations specific to that experimental system rather than generalizable HAV biology.

Future directions

To address these gaps, we propose the following avenues:

1. **Experimental validation**: CRISPR-mediated knockout or siRNA knockdown of the top understudied genes in primary human hepatocytes and organoid models, followed by HAV infection assays, will directly test their contribution to viral entry, replication, and cytopathic effects.
2. **Broader transcriptomic profiling**: Integration of additional HAV-related RNA-seq datasets (including in vivo liver biopsies and animal models) will assess the reproducibility of the identified expression patterns across biological contexts.
3. **Network analysis**: Construction of protein-protein interaction and signaling networks centered on the candidate genes can reveal mechanistic linkages to known HAV host factors and uncover potential synergistic targets.
4. **Drug repurposing screens**: High-throughput screening of the prioritized compounds (e.g., kinase inhibitors, epigenetic modulators) in HAV-infected hepatocyte cultures will evaluate antiviral efficacy and cytotoxicity, informing translational potential.
5. **Refinement of the understudied metric**: Incorporating citation network analyses, grant funding data, and functional annotation depth could yield a more nuanced measure of research neglect, improving candidate selection.

Conclusion

By systematically intersecting multi-source disease gene compilations with bibliometric filters and machine-learning-driven predictions, we have highlighted a cadre of genes that are recurrently associated with hepatitisA yet remain underexplored in the literature. Preliminary transcriptional evidence supports the involvement of several of these candidates in the host response to HAV infection. Targeted experimental follow-up and drug-repurposing investigations are warranted to elucidate their mechanistic roles and to assess their suitability as novel therapeutic or diagnostic biomarkers for hepatitisA.

Acknowledgements

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

References

[1] Arie H. Havelaar, Martyn D. Kirk, Paul R. Torgerson, and et al. World health organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLoS Med*, 12(12):e1001923, 2015.

[2] George Ostapowicz, Robert J. Fontana, Frank V. Schiødt, and et al. Results of a prospective study of acute liver failure at 17 tertiary care centers in the united states. *Ann Intern Med*, 137(12):947–956, 2002.

[3] Paul Y. Kwo, Stanley M. Cohen, and Joseph K. Lim. Acg clinical guideline: Evaluation of abnormal liver chemistries. *Am J Gastroenterol*, 112(2):241–258, 2017.

[4] Luciana Besedovsky, Tanja Lange, and Jan Born. Sleep and immune function. *Pflugers Arch*, 463(1):121–137, 2012.

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

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

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

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

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

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

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

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

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

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

[15] 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.

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

[17] 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.