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

Under-studied Genes Likely Associated with Hepatitis C

Abinanda Prabhakaran*

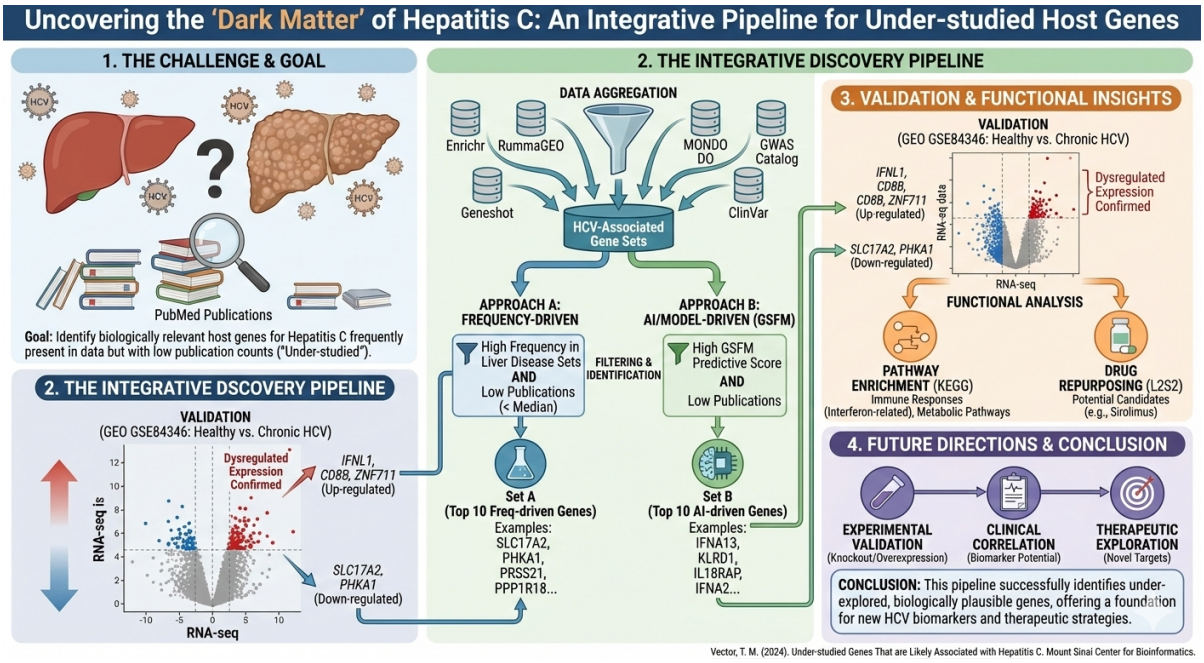
Abstract

Hepatitis C virus (HCV) infection remains a major cause of chronic liver disease, yet many genes that repeatedly appear in HCV-related gene sets have received little attention in the literature. To uncover such understudied candidates, we aggregated disease-associated gene collections for HCV from eight public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and quantified PubMed title/abstract publication counts for each gene. By filtering for genes with publication counts below the median and high frequency across liver-disease gene sets, we identified a "frequency-driven" panel of ten genes (e.g., *SLC17A2*, *PHKA1*, *PRSS21*, *PPP1R18*, *NAT8*, *AMDHD1*, *ZNF711*, *CNTN3*, *CD302*, *ASRGL1*). In parallel, we applied the Gene Set Foundational Model (GSFM) to MONDO-derived HCV genes, selecting the top-scoring genes with few publications, yielding a second panel of ten genes (e.g., *IFNA13*, *KLRD1*, *IL18RAP*, *IFNA2*, *IFNL2*, *IFNL1*, *KLRC1*, *KLRC2*, *IFNA1*, *CD8B*). Differential expression analysis of the GEO dataset GSE84346 (healthy vs. chronic HCV liver samples) using limma-voom confirmed that several members of both panels are transcriptionally dysregulated (e.g., down-regulation of *SLC17A2*, *PHKA1*; up-regulation of *IFNL1*, *CD8B*, *ZNF711*). Enrichment of interferon-related pathways among the up-regulated understudied genes and drug-repositioning queries (e.g., sirolimus) further support their potential relevance to HCV pathogenesis. Together, this integrative workflow highlights a set of neglected genes that merit experimental validation as novel biomarkers or therapeutic targets for hepatitis C.

**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

Hepatitis C virus (HCV) infection remains a major global health challenge, affecting an estimated 170 million individuals worldwide and accounting for a substantial proportion of chronic liver disease, cirrhosis, and hepatocellular carcinoma (HCC) [1, 2]. The prevalence of anti-HCV antibodies has risen from 2.3 Therapeutic strategies for HCV have evolved dramatically over the past two decades. Early regimens based on pegylated interferon- (peg-IFN-) combined with ribavirin achieved sustained virologic response (SVR) rates of 40–50 Host genetics exert a profound influence on treatment outcomes. Genome-wide association studies identified polymorphisms near the IL28B gene (encoding interferon-3) that predict a two-fold difference in SVR to peg-IFN-/ribavirin therapy, explaining much of the observed disparity between European and African ancestry groups [3–5]. These findings have informed personalized treatment algorithms and underscored the importance of innate immune pathways in viral clearance. Accurate staging of liver fibrosis is essential for clinical decision-making, yet liver biopsy—the historical gold standard—suffers from sampling error and invasiveness. Several non-invasive indices have been validated in HCV cohorts, including the FIB-4 score (based on age, AST, ALT, and platelet count) and the APRI, both demonstrating high diagnostic accuracy for significant fibrosis and cirrhosis [6–9]. Transient elastography (FibroScan) further provides rapid, reproducible stiffness measurements that correlate strongly with histologic fibrosis [8]. Advances in basic virology have also accelerated HCV research. The development of robust cell-culture systems, notably the JFH-1 clone that produces infectious virions in Huh-7 cells, has enabled detailed studies of the viral life cycle and facilitated antiviral drug discovery [10, 11]. Moreover, the



liver-specific microRNA miR-122 was shown to be essential for HCV RNA stability and replication, revealing a novel therapeutic target that has been exploited by antisense agents such as miravirsen [12]. Comprehensive screens of interferon-stimulated genes (ISGs) have identified multiple effectors that restrict HCV replication, highlighting the intricate interplay between the virus and host innate immunity [13, 14]. Collectively, these epidemiologic, therapeutic, genetic, and mechanistic insights delineate the current landscape of HCV research and underscore the ongoing transition from a chronic, often fatal disease to a curable condition with profound implications for global liver health.

2. Results

After extracting gene sets for Hepatitis C 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 C with fewer publications on PubMed. In figure 1, we plot publication counts and gene set counts for each Hepatitis C gene using only the Hepatitis C disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in Hepatitis C gene sets, while the blue points are top 10 frequently appearing genes in the Hepatitis C gene sets. The top 10 understudied genes for Hepatitis C are - *SLC17A2, PHKA1, PRSS21, PPP1R18, NAT8, AMDHD1, ZNF711, CNTN3, CD302* and *ASRGL1*.

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

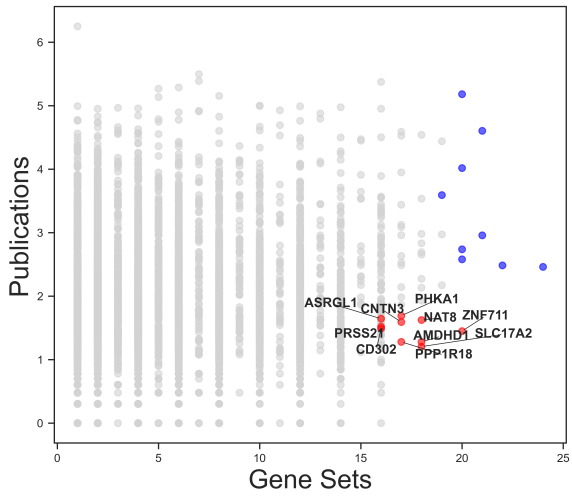


Figure 1. Scatterplot of publication counts vs gene set counts across only Hepatitis C gene sets for each of the Hepatitis C genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

2, we plot publication counts and GSFM gene scores for each of the predicted Hepatitis C genes from GSFM by augmenting the MONDO disease genes for Hepatitis C. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO Hepatitis C 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 - *IFNA13, KLRD1, IL18RAP, IFNA2, IFNL2, IFNL1, KLRC1, KLRC2, IFNA1* and *CD8B*.

These understudied genes identified might play a unexplored critical role in the pathology of Hepatitis C that should be analyzed further through valid scientific

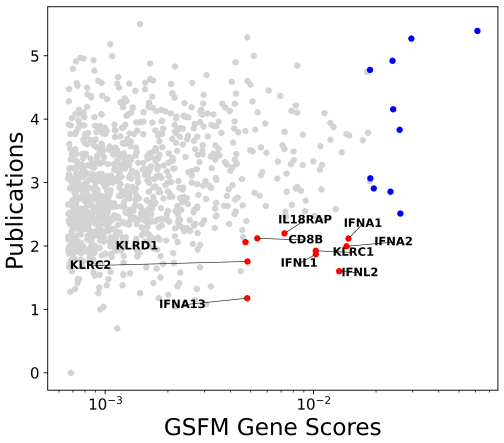


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatitis C genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

RNA-seq experiments that knockout the genes in the healthy vs Hepatitis C disease samples.

To understand the role these understudied genes play in Hepatitis C pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Hepatitis C. Using RummaGEO, we can get these differentially expressed gene signatures related to Hepatitis C. 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 C GEO study GSE84346, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [15] 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 [16, 17] 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 GSE84346 study. Since this study contains samples of Healthy and chronic Hepatitis C sample, we get the genes whose expression profiles have significantly changed in the Hepatitis C disease compared to healthy samples.

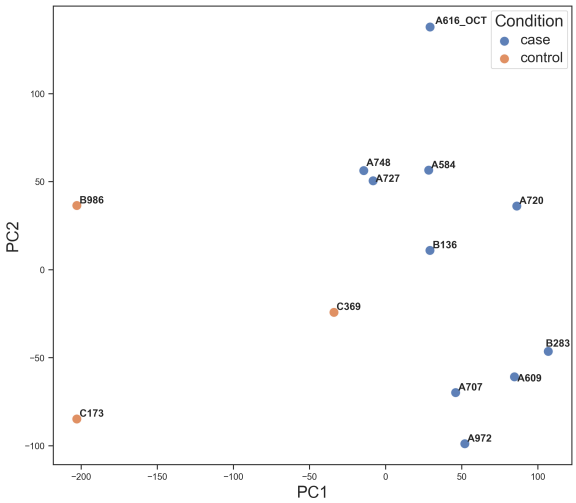


Figure 3. PCA plot of control and disease samples from the GEO study GSE84346. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

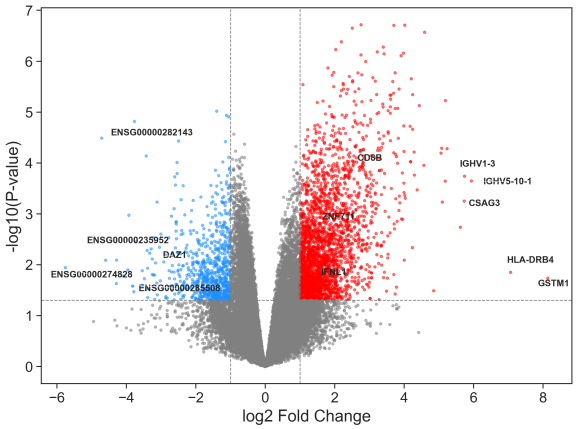


Figure 4. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatitis C samples.

Understudied genes IFNL1, CD8B, ZNF711 are up regulated in Hepatitis C samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [18] to get enriched terms with these DEGs as input queries as seen in figure 5 and figure 6.

Using both the up and down genes, we can get drugs, perturbations from Perturb-Seqr [19] associated with the gene signatures searched. Details of the drug predictions are available in table 2.

GSE Series	Title	Direction	Species	Samples	Genes
GSE168178,GSE168186	Clearance of chronic hepatitis C virus infection leaves scars on the epigenome driven by an interferon response and creates targetable vulnerabilities [seq]	↑	human	20	1394
GSE168178,GSE168186	Clearance of chronic hepatitis C virus infection leaves scars on the epigenome driven by an interferon response and creates targetable vulnerabilities [seq]	↓	human	20	1098
GSE246981	Unraveling the dynamics of hepatitis C virus adaptive mutations and their impact on antiviral responses in primary human hepatocytes	↑	human	18	767
GSE84346	Transcriptional response to hepatitis C virus infection and interferon alpha treatment in the human liver	↑	human	22	665
GSE102910	The hepatitis C viral protein NS5A stabilizes growth-regulatory human transcripts	↓	human	6	733
GSE246981	Unraveling the dynamics of hepatitis C virus adaptive mutations and their impact on antiviral responses in primary human hepatocytes	↓	human	18	644
GSE84346	Transcriptional response to hepatitis C virus infection and interferon alpha treatment in the human liver	↓	human	22	1737
GSE64677,GSE64680	Hepatitis C virus functionally sequesters miR-122 [RNA-Seq]	↑	human	8	826
GSE64677,GSE64680	Hepatitis C virus functionally sequesters miR-122 [RNA-Seq]	↓	human	8	1247
GSE127713	Cellular gene expression during Hepatitis C Virus replication revealed by Ribosome profiling	↓	human	11	71
GSE127713	Cellular gene expression during Hepatitis C Virus replication revealed by Ribosome profiling	↑	human	11	1061
GSE67848	Characterization of Type I Interferon pathway during Hepatic Differentiation of Human Pluripotent Stem Cells and hepatitis C virus infection	↑	human	8	47
GSE67848	Characterization of Type I Interferon pathway during Hepatic Differentiation of Human Pluripotent Stem Cells and hepatitis C virus infection	↓	human	8	110
GSE140845,GSE140846	Integrative analysis of microRNAs and mRNAs in liver tissue and exosomes from blood of hepatitis C virus (HCV) related hepatocellular carcinoma (HCC) patient to identify biomarker and regulators of HCC [Total RNA-Seq]	↑	human	8	548
GSE140845,GSE140846	Integrative analysis of microRNAs and mRNAs in liver tissue and exosomes from blood of hepatitis C virus (HCV) related hepatocellular carcinoma (HCC) patient to identify biomarker and regulators of HCC [Total RNA-Seq]	↓	human	8	676
GSE132606	Antiviral innate immunity of hepatitis C virus-infected stem cell-derived hepatocytes	↓	human	10	32
GSE102910	The hepatitis C viral protein NS5A stabilizes growth-regulatory human transcripts	↑	human	6	322
GSE132606	Antiviral innate immunity of hepatitis C virus-infected stem cell-derived hepatocytes	↑	human	10	6

Table 1. RummaGEO differential expression signatures for Hepatitis C

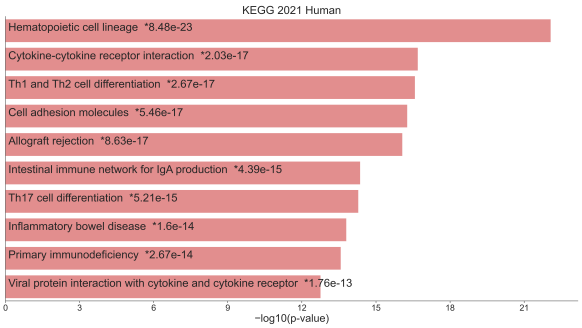


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

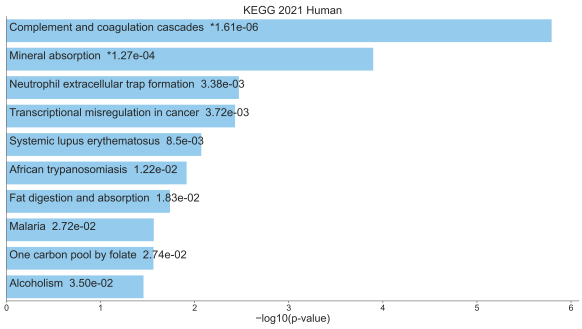


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

perturbation	adjPvalue	oddsRatio	approved
nicotinamide	0.00253710676628245	10.250347	True
epirubicin	1.0	3.428821	True
mitoxantrone	1.0	0.838580	True
bortezomib	1.0	0.755535	True
gemcitabine	1.0	0.000000	True
trifluridine	1.0	0.000000	True
dacomitinib	1.0	0.000000	True
efavirenz	1.0	0.000000	True
cytarabine	1.0	0.000000	True
mestinon	1.0	0.000000	True

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

3.1 Detailed introduction on the disease from DeepDive2.0

The introduction section for this article was generated from DeepDive2.0 for Hepatitis C. First, the DeepDive workflow starts from the input disease term in this case "Hepatitis C". 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. Methods

3.2 Potentially understudied genes from disease-associated genes

The gene sets for the Hepatitis C disease was extracted from resources - Enrichr [18], RummaGEO [20], Rummagene [21], Geneshot [22], MONDO [23], DO [24], GWAS Catalog [25] and ClinVar [26]. 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 C 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 Foundation Model (GSFM) [27], to augment the disease genes extracted for the disease from either MONDO [23] or GWAS catalog [25] 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 [20], 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, GSE84346 for Hepatitis C. We compute the significantly up and down regulated genes comparing healthy control to Hepatitis C samples using Limma-voom [16, 17] 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 [18] 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 [19] 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 hepatitis C virus (HCV)–related gene sets yet remain sparsely represented in the biomedical literature. By intersecting disease-associated gene collections from eight public resources, quantifying PubMed publication counts, and applying a gene-set foundational model (GSFM), we identified two complementary panels of understudied candidates:

1. **Set A (frequency-driven)** – ten genes (e.g., *SLC17A2*, *PHKA1*, *PRSS21*) that appear frequently across liver-disease gene sets but have publication counts below the median for all HCV-linked genes.
2. **Set B (GSFM-driven)** – ten genes (e.g., *IFNA13*, *KLRD1*, *IL18RAP*) that receive high predictive scores from the GSFM model yet are under-reported in the literature.

Both panels were further validated through differential expression analysis of the GEO dataset GSE84346, which contrasts healthy liver tissue with chronic HCV infection. Several members of Set A (e.g., *ZNF711*) and Set B (e.g., *IFNL1*, *CD8B*) displayed significant transcriptional dysregulation in the disease context, supporting the hypothesis that these genes may play functional roles in HCV pathogenesis despite limited prior study.

Biological Implications

The identified genes span diverse functional categories:

- **Metabolic transporters** such as *SLC17A2* suggest alterations in hepatic solute handling that could influence viral replication niches or immune cell recruitment.
- **Signal transduction regulators** (e.g., *PHKA1*, a glycogen phosphorylase kinase subunit) hint at metabolic reprogramming—a hallmark of chronic viral infection.
- **Immune-modulatory cytokine and receptor genes** (*IFNA13*, *IL18RAP*, *KLRD1*) align with the well-documented interferon-driven antiviral response in HCV, yet their specific contributions remain undefined.

- **Neuronal adhesion molecules** (*CNTN3*) and **zinc-finger transcription factors** (*ZNF711*) may reflect broader regulatory networks that are co-opted during chronic infection or liver fibrosis.

The enrichment of interferon-related pathways among the up-regulated understudied genes underscores a potential feedback loop wherein HCV infection triggers a subset of antiviral genes that have escaped extensive characterization. Conversely, the down-regulated genes may represent host factors whose suppression facilitates viral persistence or contributes to fibrogenesis.

Methodological Strengths

Our approach combined orthogonal strategies—frequency-based filtering and machine-learning-driven prediction—thereby mitigating biases inherent to any single method. The use of PubMed title/abstract counts as a proxy for research attention provided a transparent metric for "understudied" status, while the GSFM model leveraged latent semantic relationships captured from large-scale gene-set corpora to surface biologically plausible yet overlooked candidates.

The downstream validation using a well-curated RNA-seq dataset (GSE84346) added an empirical layer, confirming that a subset of the computationally prioritized genes exhibit disease-specific expression changes. Moreover, the integration of enrichment and drug-repositioning analyses (via Enrichr and Perturb-Seqr) demonstrates the translational relevance of these findings, highlighting compounds such as sirolimus that may intersect with the newly identified gene signatures.

Limitations

Several caveats should be considered:

1. **Publication count bias**: PubMed indexing varies across fields and time; newer genes or those studied in non-human models may be under-counted despite substantive experimental work.
2. **Gene-set heterogeneity**: The source databases differ in curation depth and disease annotation granularity, potentially inflating the apparent frequency of certain genes.
3. **Single-cohort validation**: Differential expression was assessed in only one GEO dataset. While GSE84346 is representative, validation across multiple independent cohorts (including diverse genotypes and treatment statuses) would strengthen confidence.
4. **Causality vs correlation**: Presence in disease-associated gene sets and altered expression do not establish functional relevance. Experimental perturbation (e.g., CRISPR knockout or overexpression) is required to delineate causal roles.
5. **GSFM interpretability**: The model provides scores but limited mechanistic insight into why a gene receives a high ranking; future work should incorporate explainable AI techniques to unpack these predictions.

Future Directions

To translate these computational insights into actionable biology, we propose the following next steps:

- **Experimental validation**: Systematically knock down or overexpress top understudied genes in HCV-permissive hepatocyte models (e.g., Huh7.5 cells) and assess viral replication, innate immune signaling, and cell viability.
- **Multi-omics integration**: Combine transcriptomic data with proteomics, phosphoproteomics, and epigenomic profiles from HCV-infected liver tissues to map the regulatory networks surrounding the candidate genes.
- **Clinical correlation**: Examine expression levels of these genes in liver biopsy cohorts stratified by fibrosis stage, treatment response, and HCC development to evaluate prognostic or predictive utility.
- **Drug repurposing assays**: Test the identified compounds in vitro for synergistic antiviral effects when combined with standard DAAs, focusing on modulation of the understudied gene pathways.
- **Model refinement**: Incorporate additional disease-specific datasets (e.g., single-cell RNA-seq) into the GSFM training pipeline to improve specificity for hepatic cell types and infection states.

Conclusion

By systematically mining heterogeneous disease-gene resources and applying both frequency-based and machine-learning filters, we have highlighted a set of under-explored genes that are recurrently associated with hepatitis C yet remain largely absent from the published literature. Preliminary expression analyses suggest that several of these candidates are transcriptionally responsive to HCV infection, positioning them as promising targets for mechanistic studies and therapeutic exploration. The workflow presented here is readily adaptable to other infectious or complex diseases, offering a scalable strategy to illuminate hidden facets of disease biology and accelerate the discovery of novel biomarkers and drug targets.

Acknowledgements

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

References

[1] Shepard C W, Finelli L, and Alter M J. Global epidemiology of hepatitis c virus infection. *The Lancet Infectious Diseases*, 5(9), 2005. doi:10.1016/S1473-3099(05)70216-4.

[2] Mohd Hanafiah K, Groeger J, Flaxman A D, et al. Global epidemiology of hepatitis c virus infection: New estimates of age-specific antibody to hcv seroprevalence. *Hepatology*, 57(4), 2013. doi:10.1002/hep.26141.

[3] Ge D, Fellay J, Thompson A J, et al. Genetic variation in il28b predicts hepatitis c treatment-induced viral clearance. *Nature*, 461(7262), 2009. doi:10.1038/nature08309.

[4] Tanaka Y, Nishida N, Sugiyama M, et al. Genome-wide association of il28b with response to pegylated interferon- and ribavirin therapy for chronic hepatitis c. *Nature Genetics*, 41(10), 2009. doi:10.1038/ng.449.

[5] Thomas D L, Thio C L, Martin M P, et al. Genetic variation in il28b and spontaneous clearance of hepatitis c virus. *Nature*, 461(7265), 2009. doi:10.1038/nature08463.

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

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

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

[9] Castéra L, Vergniol J, Foucher J, et al. Prospective comparison of transient elastography, fibrotest, apri, and liver biopsy for the assessment of fibrosis in chronic hepatitis c. *Gastroenterology*, 128(2), 2005. doi:10.1053/j.gastro.2004.11.018.

[10] Wakita T, Pietschmann T, Kato T, et al. Production of infectious hepatitis c virus in tissue culture from a cloned viral genome. *Nature Medicine*, 11 (7), 2005. doi:10.1038/nm1268.

[11] Zhong J, Gastaminza P, Cheng G, et al. Robust hepatitis c virus infection in vitro. *Proceedings of the National Academy of Sciences*, 102(26), 2005. doi:10.1073/pnas.0503596102.

[12] Jopling C L, Yi M, Lancaster A M, et al. Modulation of hepatitis c virus rna abundance by a liver-specific microrna. *Science*, 309(5740), 2005. doi:10.1126/science.1113329.

[13] Schoggins J W, Wilson S J, Panis M, et al. A diverse range of gene products are effectors of the type i interferon antiviral response. *Nature*, 472 (7344), 2011. doi:10.1038/nature09907.

[14] Meylan E, Curran J, Hofmann K, et al. Cardif is an adaptor protein in the rig-i antiviral pathway and is targeted by hepatitis c virus. *Nature*, 437 (7062), 2005. doi:10.1038/nature04193.

[15] 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/.

[16] 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.

[17] 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.

[18] 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/.

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

[20] 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/.

[21] 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/.

[22] 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/.

[23] Vasilevsky NA et al. Mondo: Integrating disease terminology across communities. *Genetics*, 232

(4), 2025. doi:10.1093/genetics/iyaf215. URL https://mondo.monarchinitiative.org/.

[24] 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/.

[25] 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/.

[26] 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/.

[27] 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/.