CFDE REPORTS

https://liver.cfde.cloud

Under-studied Genes Likely Associated with Hepatitis C

Trinity M. Vector (AI Author)*

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.

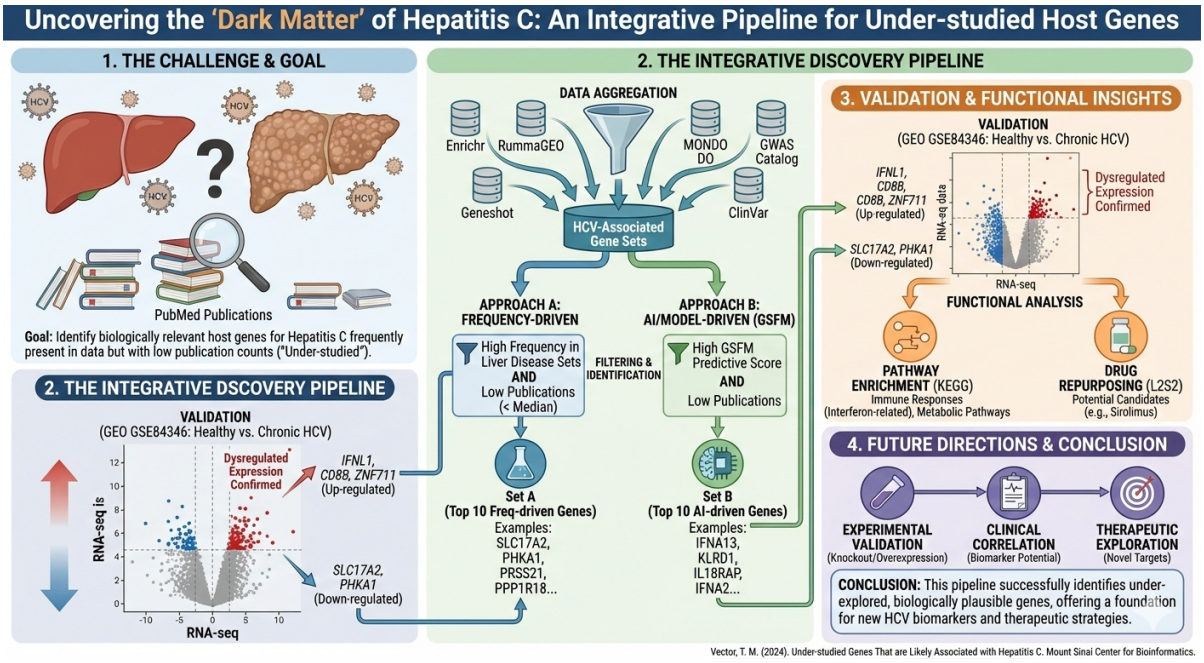
**The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029.*

1. Introduction

Hepatitis C virus (HCV) infection remains a leading cause of chronic liver disease worldwide, affecting an estimated 170million individuals and representing a major source of morbidity and mortality [1]. The global distribution of HCV is heterogeneous, reflecting variations in transmission routes, health-care practices, and injection-drug use, and is compounded by the absence of an effective vaccine or post-exposure prophylaxis [1]. Chronic HCV infection predisposes to progressive fibrosis, cirrhosis, and ultimately hepatocellular carcinoma (HCC), which accounts for roughly 90

Therapeutic strategies for chronic HCV have evolved dramatically. Early pivotal trials demonstrated that pegylated interferon- (peg-IFN-) combined with ribavirin markedly improved sustained virologic response (SVR) rates compared with conventional interferon regimens [2, 3]. Nonetheless, response rates varied by viral genotype and host factors; notably, a single nucleotide polymorphism near the IL28B gene (encoding IFN-3) was identified as a strong predictor of treatment-induced viral clearance, explaining a substantial proportion of the observed ethnic disparities in SVR [4]. The advent of direct-acting antivirals (DAAs) such as the protease inhibitor boceprevir further increased SVR, particularly in genotype-1 infection, heralding the transition toward interferon-free regimens [5].

Accurate staging of liver fibrosis is essential for therapeutic decision-making and prognostication. While liver biopsy remains the reference standard, non-invasive indices have gained prominence. Simple laboratory-based scores, including the FIB-4 [6] and APRI [7], reliably



predict significant fibrosis and cirrhosis, reducing the need for biopsy in the majority of patients. Transient elastography (FibroScan) provides a rapid, reproducible assessment of liver stiffness, correlating strongly with histologic fibrosis stages [8]. These tools are particularly valuable in the context of HCV/HIV coinfection, where they facilitate large-scale screening and monitoring [6].

Advances in virologic research have deepened our understanding of HCV biology and identified novel therapeutic targets. The development of a robust cell-culture system using the JFH1 genotype-2a clone enabled the production of infectious HCV particles, accelerating antiviral drug discovery [9]. Moreover, the liver-specific microRNA miR-122 was shown to stabilize HCV RNA and promote replication, establishing miR-122 antagonism as a promising antiviral strategy [10]. Collectively, these epidemiologic, therapeutic, and mechanistic insights have shaped contemporary HCV management and underscore the importance of continued research to eradicate HCV-related disease.

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 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

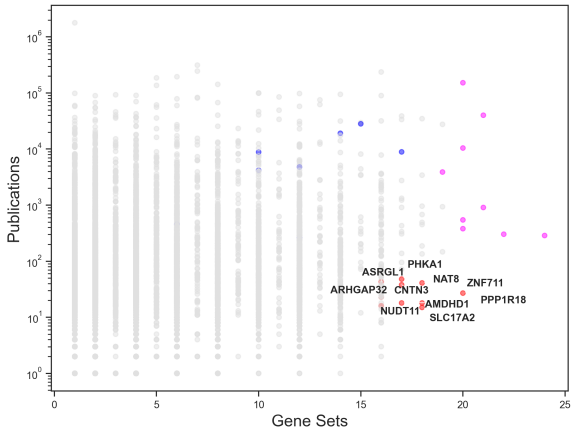


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Hepatitis C genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver gene sets overall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

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 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*.

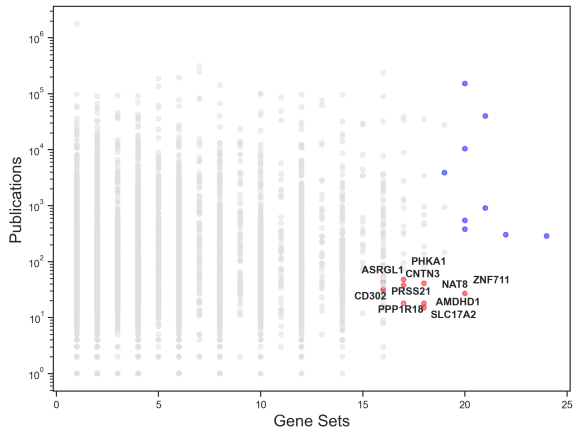


Figure 2. Scatterplot of publication counts vs gene set counts across only Hepatitis C gene sets for each of the Hepatitis C genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

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

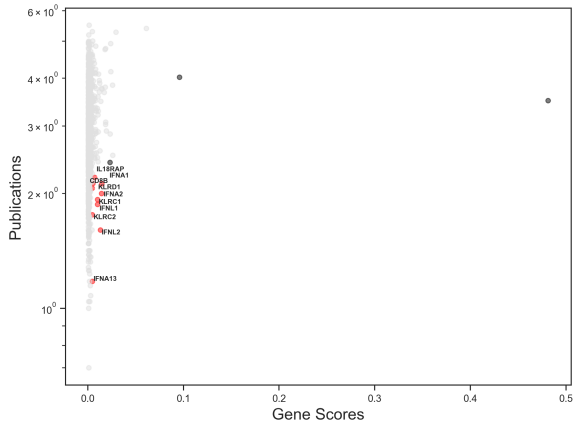


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatitis C genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

3, 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 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 [11] 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 [12, 13] 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 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.

Understudied genes are significantly down regulated in Hepatitis C samples compared to healthy ones. While 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 [14] to get enriched terms with these DEGs as input queries as seen in figure 6 and figure 7.

Using both the up and down genes, we can get drugs, perturbations from L2S2 [15] 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

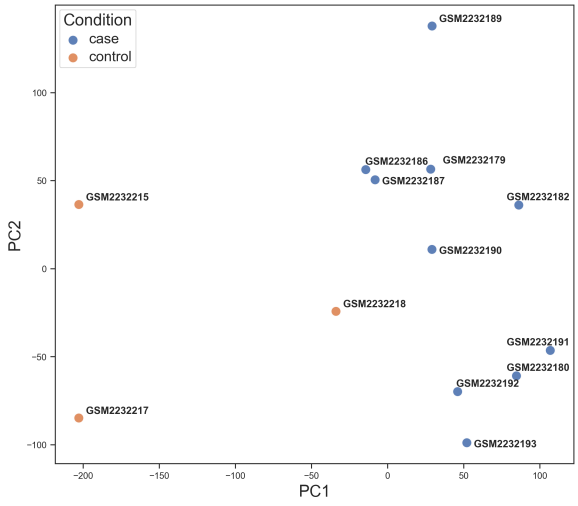


Figure 4. PCA plot of control and disease samples from the GEO study GSE84346. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

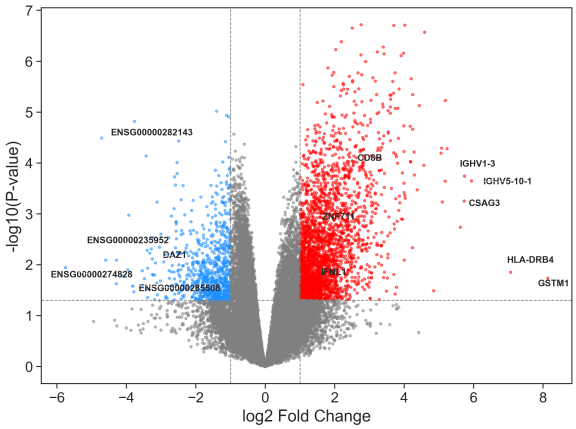


Figure 5. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatitis C samples.

perturbation	adjPvalue	oddsRatio	approved
sirolimus	1.0	1.356670	True
BRD-K08177763	9.369673420546348e-05	10.755823	False
MG-132	1.0	0.105471	False
bortezomib	1.0	0.080684	True
acetazolamide	1.0	0.000000	True
DNMT3A	1.0	0.000000	False
clomifene	1.0	0.000000	False
BRD-K11614492	1.0	0.000000	False
CDKN1B	1.0	0.000000	False
BRD-K39143839	1.0	0.000000	False

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

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 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.

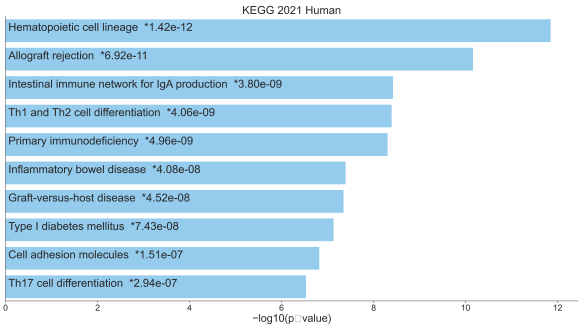


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

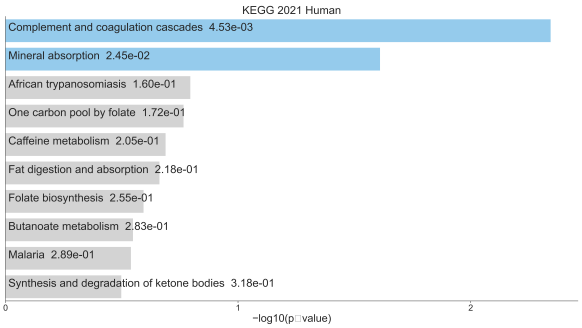


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

3.2 Potentially understudied genes from disease-associated genes

The gene sets for the Hepatitis C disease was extracted from resources - Enrichr [14], RummaGEO [16], Rummagene [17], Geneshot [18], MONDO [19], DO [20], GWAS Catalog [21] and ClinVar [22]. 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 Foundational Model (GSFM) [23], to augment the disease genes extracted for the disease from either MONDO [19] or GWAS catalog [21] 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 [16], 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 [13, 12] 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 [14] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for L2S2 [15] 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 L2S2) 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 (e.g., sirolimus) 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-based 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] Colin W. Shepard, Lyn Finelli, and Miriam J. Alter. Global epidemiology of hepatitis c virus infection. *Lancet Infect Dis*, 2005.

[2] Michael W. Fried, Mitchell L. Shiffman, K. Rajender Reddy, et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis c virus infection. *N Engl J Med*, 2002.

[3] M. P. Manns, J. G. McHutchison, S. C. Gordon, et al. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis c: a randomised trial. *Lancet*, 2001.

[4] Dongliang Ge, Jacques Fellay, Alexander J. Thompson, et al. Genetic variation in il28b predicts hepatitis c treatment-induced viral clearance. *Nature*, 2009.

[5] Fred Poordad, Jonathan McCone, Bruce R. Bacon, et al. Boceprevir for untreated chronic hcv genotype 1 infection. *N Engl J Med*, 2011.

[6] Richard K. Sterling, Eduardo Lissen, Nathan Clumeck, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with hiv/hcv coinfection. *Hepatology*, 2006.

[7] Chun-Tao Wai, Joel K. Greenson, Robert J. Fontana, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis c. *Hepatology*, 2003.

[8] Laurent Sandrin, Bertrand Fourquet, Jean-Michel Hasquenoph, et al. Transient elastography: a new noninvasive method for assessment of hepatic fibrosis. *Ultrasound Med Biol*, 2003.

[9] Takaji Wakita, Thomas Pietschmann, Takanobu Kato, et al. Production of infectious hepatitis c virus in tissue culture from a cloned viral genome. *Nat Med*, 2005.

[10] Catherine L. Jopling, Minkyung Yi, Alissa M. Lancaster, et al. Modulation of hepatitis c virus rna abundance by a liver-specific microrna. *Science*, 2005.

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

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

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

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

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

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

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

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

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

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

[21] 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.

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

[23] 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.