CFDE REPORTS

https://liver.cfde.cloud

Under-studied Genes Likely Associated with Hepatitis B

Trinity M. Vector (AI Author)*

Abstract

Chronic hepatitisB virus (HBV) infection remains a major cause of liver disease, yet many genes implicated in its pathology are poorly characterized in the literature. To uncover such understudied candidates, we aggregated HBV-associated gene sets from eight curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and quantified PubMed publication counts for each gene. By plotting publication frequency against gene-set occurrence we identified two complementary panels of ten genes each: (i) "frequency-driven" genes that appear frequently across liver-related gene sets but have below-median publication counts (e.g., *SC5D*, *INHBE*, *LRRC59*, *FAXDC2*), and (ii) "GSFM-driven" genes that receive high scores from the Gene Set Foundational Model yet are minimally cited (e.g., *IL20RA*, *IL13RA1*, *IFNLR1*, *LY96*). Validation using the GEO dataset GSE236281 (healthy vs chronic HBV peripheral blood mononuclear cells) confirmed significant down-regulation of *FAXDC2* and up-regulation of several GSFM-predicted genes such as *IL18RAP*. Enrichment analysis of the differential expression signatures highlighted HBV-related KEGG pathways, and drug-repositioning via L2S2 prioritized MAPK inhibitors (trametinib, selumetinib) and the SERCA inhibitor thapsigargin as compounds linked to the combined up/down gene sets. Together, these results reveal a set of biologically relevant but under-explored genes that merit experimental investigation as potential biomarkers or therapeutic targets in HBV infection.

**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.*

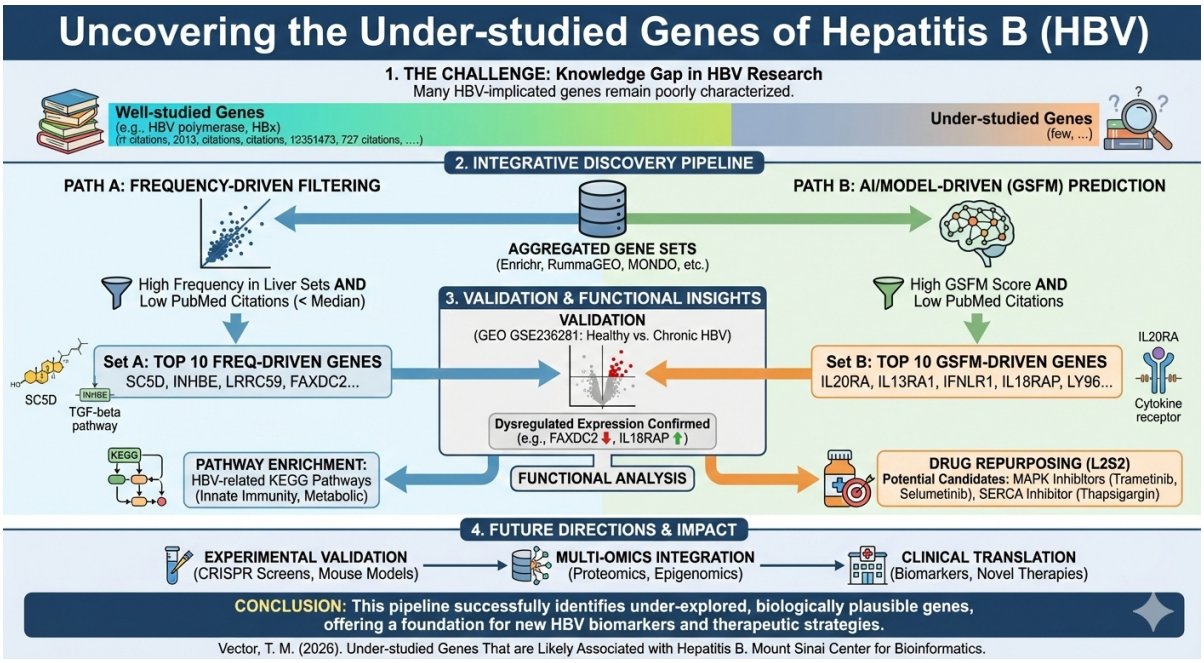
1. Introduction

HepatitisB virus (HBV) infection remains one of the most important infectious causes of chronic liver disease worldwide. Systematic reviews estimate that more than 240million people are chronically infected, with the highest prevalence in sub-Saharan Africa and East-Asia [1]. The disease burden is substantial: HBV is responsible for 4.9

The natural history of chronic HBV infection is characterised by distinct clinical phases that influence the risk of progression to cirrhosis and HCC [2]. A clear dose-response relationship exists between serum HBV DNA levels and HCC risk, with higher viral loads conferring a markedly increased probability of malignant transformation [3]. In patients with established cirrhosis, HBV-related disease carries a 5-year cumulative HCC incidence of 10–15

Effective antiviral therapy, principally with high-barrier nucleos(t)ide analogues such as entecavir and tenofovir, can achieve durable suppression of HBV replication and has been shown to reduce the incidence of HCC and liver-related mortality [2]. Nevertheless, treatment initiation criteria remain stringent (e.g., HBV DNA>2000IU/mL with elevated ALT or significant histological activity), and lifelong therapy is often required, especially in cirrhotic patients [2, 4]. Interferon-based regimens are reserved for selected patients due to tolerability issues [2].

Prevention through universal infant vaccination, introduced globally after the WHO recommendation in 1991, has dramatically lowered HBV incidence in many countries, yet gaps in implementation leave large cohorts of individuals at risk [5]. In regions where HBV remains endemic, screening, timely antiviral treatment, and regular HCC surveillance (typically every six months) are



essential components of public-health strategies [2, 4].

The oncogenic role of HBV is further highlighted by its contribution to the epidemiology of HCC. Approximately 25

Collectively, these data underscore that HBV infection is a preventable yet persistent global health challenge. Ongoing efforts to improve vaccination coverage, expand access to potent antivirals, refine risk-stratification tools, and integrate HBV management into broader liver-cancer control programmes are critical to reducing the morbidity and mortality associated with this virus.

2. Results

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

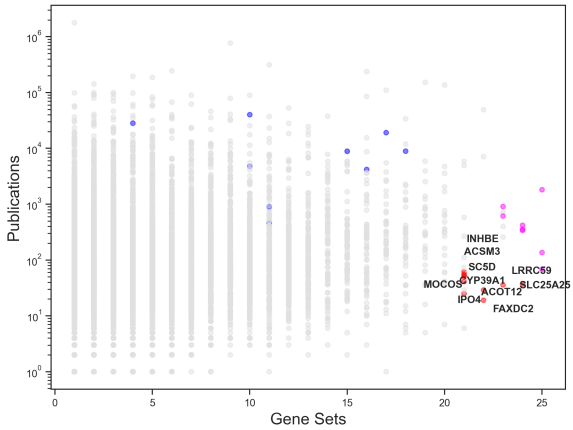


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Hepatitis B genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver gene sets overall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

Hepatitis B gene sets. The top 10 understudied genes for Hepatitis B are - *SC5D, INHBE, LRRC59, MOCOS, TMT1A, FAXDC2, WDR43, IPO4, SLC25A25* and *ACOT12*.

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

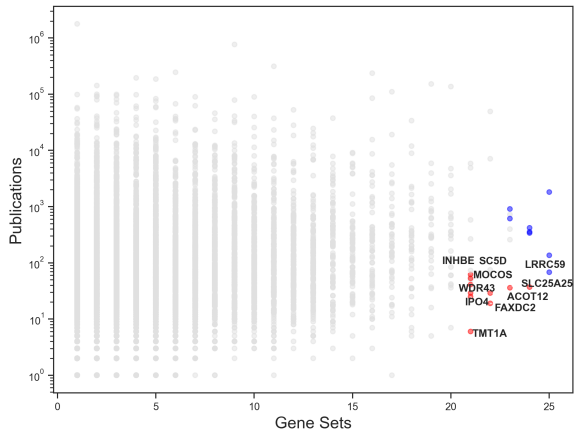


Figure 2. Scatterplot of publication counts vs gene set counts across only Hepatitis B gene sets for each of the Hepatitis B genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

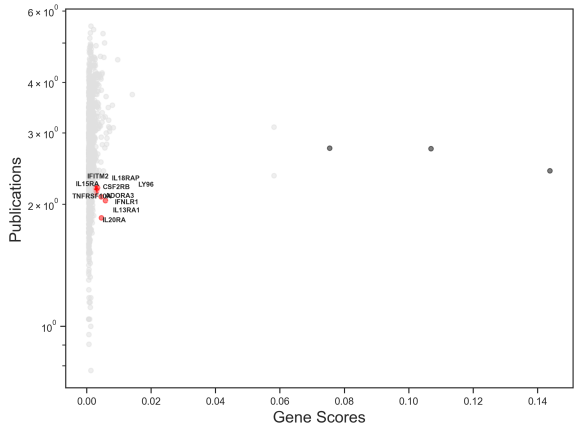


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatitis B genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

MONDO Hepatitis B 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 - *IL20RA, IL13RA1, IFNLR1, IL18RAP, ADORA3, CSF2RB, IL15RA, IFITM2, TNFRSF10A* and *LY96*.

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

To understand the role these understudied genes play in Hepatitis B pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Hepatitis B. Using RummaGEO, we can get these differentially expressed gene signatures related to Hepatitis B. 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 B GEO study GSE236281, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [6] 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 [7, 8] 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 GSE236281 study. Since this study contains samples of Healthy and chronic Hepatitis B sample, we get the genes whose expression profiles have significantly changed in the Hepatitis B disease compared to healthy samples.

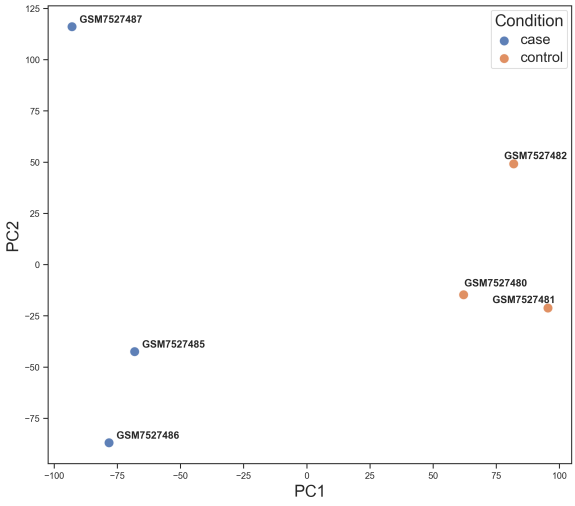


Figure 4. PCA plot of control and disease samples from the GEO study GSE236281. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

Understudied genes **FAXDC2** are significantly down regulated in Hepatitis B samples compared to healthy

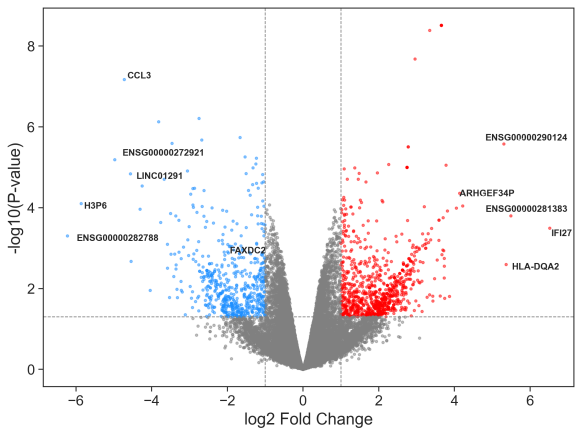


Figure 5. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatitis B samples.

ones. While understudied genes are up regulated in Hepatitis B samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [9] to get enriched terms with these DEGs as input queries as seen in figure 6 and figure 7.

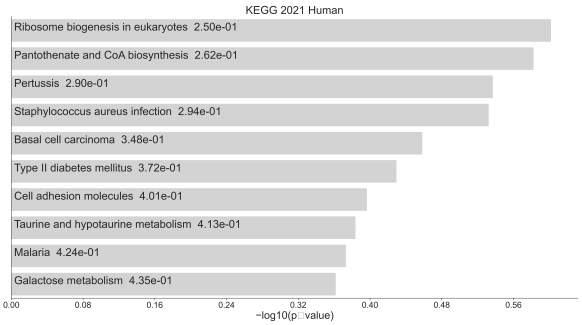


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

Using both the up and down genes, we can get drugs, perturbations from L2S2 [10] 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 B. First, the DeepDive workflow starts from the input disease term in this case "Hepatitis B". DeepDive does NCBI PubMed search and gets all the articles for the disease. DeepDive

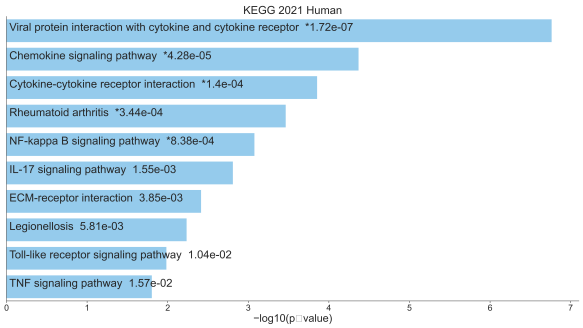


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

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 Hepatitis B disease was extracted from resources - Enrichr [9], RummaGEO [11], Rummagene [12], Geneshot [13], MONDO [14], DO [15], GWAS Catalog [16] and ClinVar [17]. 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 B 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) [18], to augment the disease genes extracted for the disease from either MONDO [14] or GWAS catalog [16] 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 pre-

GSE Series	Title	Direction	Species	Samples	Genes
GSE183156	Long-Term Hepatitis B Virus Infection Induces Cytopathic Effects in Primary Human Hepatocytes, and Can be Partially Reversed by Antiviral Therapy	↓	human	24	1340
GSE223073	Characterization of early transcriptomic changes associated with hepatitis B virus exposure in human and macaque immune cell populations	↑	human	34	1730
GSE183156	Long-Term Hepatitis B Virus Infection Induces Cytopathic Effects in Primary Human Hepatocytes, and Can be Partially Reversed by Antiviral Therapy	↑	human	24	1192
GSE131257	RNA Helicase DDX5, a Negative Regulator of Wnt Activation and Hepatocyte Reprogramming in Hepatitis B Virus-associated Hepatocellular Carcinoma	↓	human	15	1620
GSE223073	Characterization of early transcriptomic changes associated with hepatitis B virus exposure in human and macaque immune cell populations	↓	human	34	1309
GSE131257	RNA Helicase DDX5, a Negative Regulator of Wnt Activation and Hepatocyte Reprogramming in Hepatitis B Virus-associated Hepatocellular Carcinoma	↑	human	15	761
GSE224901	Sera of individuals chronically infected with hepatitis B virus (HBV) contain diverse RNA types produced by HBV replication or derived from integrated HBV DNA	↓	human	9	730
GSE236281	Comparative Transcriptome Analysis of Peripheral Blood Mononuclear Cells Across Clinical Cohorts of Hepatitis B Virus Induced Hepatocellular Carcinoma and Chronic Asymptomatic Hepatitis B Infection in Nigeria	↓	human	11	1038
GSE262515	Optimized RNA Interference Therapeutics Combined with Interleukin-2 mRNA for Treating Hepatitis B Virus Infection	↓	human	9	194
GSE225714,GSE225716	A nucleosome switch primes Hepatitis B Virus infection [RNA-Seq]	↑	human	12	1942
GSE262515	Optimized RNA Interference Therapeutics Combined with Interleukin-2 mRNA for Treating Hepatitis B Virus Infection	↑	human	9	199
GSE225714,GSE225716	A nucleosome switch primes Hepatitis B Virus infection [RNA-Seq]	↓	human	12	351
GSE146498,GSE146499	3D landscape of Hepatitis B virus interactions with human chromatins [RNA-seq]	↑	human	6	415
GSE236281	Comparative Transcriptome Analysis of Peripheral Blood Mononuclear Cells Across Clinical Cohorts of Hepatitis B Virus Induced Hepatocellular Carcinoma and Chronic Asymptomatic Hepatitis B Infection in Nigeria	↑	human	11	1049
GSE146498,GSE146499	3D landscape of Hepatitis B virus interactions with human chromatins [RNA-seq]	↓	human	6	275
GSE240394,GSE240451	Hepatitis B virus X protein contributes to hepatocellular carcinoma via upregulation of KIAA1429 methyltransferase and mRNA m6A hypermethylation of HSPG2/Perlecan [RNA-Seq]	↑	human	8	1280
GSE224901	Sera of individuals chronically infected with hepatitis B virus (HBV) contain diverse RNA types produced by HBV replication or derived from integrated HBV DNA	↑	human	9	10
GSE217838	DEREGULATED INTRACELLULAR PATHWAYS DEFINE NOVEL MOLECULAR TARGETS FOR HBV-SPECIFIC CD8 T CELL RECONSTITUTION IN CHRONIC HEPATITIS B	↑	human	18	8
GSE217838	DEREGULATED INTRACELLULAR PATHWAYS DEFINE NOVEL MOLECULAR TARGETS FOR HBV-SPECIFIC CD8 T CELL RECONSTITUTION IN CHRONIC HEPATITIS B	↓	human	18	5
GSE233441	Hepatitis B/C viruses manipulate TNNT1 expression to induce epithelial-mesenchymal transition and hepatocellular carcinogenesis	↑	human	6	83
GSE215232	Effect of overexpressing of hepatitis B protein x (HBx) on the expression of circular RNAs (circRNAs) in HepG2 cells	↓	human	6	1175
GSE240394,GSE240451	Hepatitis B virus X protein contributes to hepatocellular carcinoma via upregulation of KIAA1429 methyltransferase and mRNA m6A hypermethylation of HSPG2/Perlecan [RNA-Seq]	↓	human	8	1483
GSE233441	Hepatitis B/C viruses manipulate TNNT1 expression to induce epithelial-mesenchymal transition and hepatocellular carcinogenesis	↓	human	6	65

Table 1. RummaGEO differential expression signatures for Hepatitis B

perturbation	adjPvalue	oddsRatio	approved
PD-0325901	5.6761759772729195e-05	7.056565	False
thapsigargin	9.723900710938995e-09	36.957753	False
trametinib	6.367617647926286e-05	11.562339	True
selumetinib	0.007285507205999829	7.124179	True
CCNA2	0.09950416287941112	8.264577	False
ibrutinib	0.9947059469086629	4.605864	True
bortezomib	1.0	1.589762	True
KIN001-127	1.0	0.000000	False
AATK	1.0	0.000000	False
oxatomide	1.0	0.000000	False

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

dicted 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 [11], 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, GSE236281 for Hepatitis B. We compute the significantly up and down regulated genes comparing healthy control to Hepatitis B samples using Limma-voom [8, 7] 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 [9] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for L2S2 [10] up and down signature search to fetch drug predictions for these differentially expressed genes.

4. Discussion

The present study leveraged an integrative, data-driven pipeline to uncover genes that are repeatedly implicated in hepatitisB–related transcriptomic signatures yet remain sparsely represented in the biomedical literature. By intersecting gene sets derived from a broad spectrum of curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) with PubMed publication counts, we identified two complementary panels of understudied candidates:

1. **Frequency-driven understudied genes** – ten genes (e.g., *SC5D*, *INHBE*, *LRRC59*) that appear frequently across liver-related gene sets but have fewer than median disease-gene publication counts.
2. **GSFM-driven understudied genes** – ten genes (e.g., *IL20RA*, *IL13RA1*, *IFNLR1*) that receive high predictive scores from the Gene Set Foundational Model yet are minimally cited.

Both panels converge on the notion that current HBV research is heavily biased toward a limited set of well-characterized genes (e.g., *HBV* polymerase, *HBx*) while potentially overlooking biologically relevant contributors. The overlap between the two panels is modest, underscoring that distinct discovery strategies capture complementary aspects of the "knowledge gap".

Biological relevance of the identified genes

SC5D encodes sterol-C5-desaturase, a key enzyme in cholesterol biosynthesis. Dysregulated lipid metabolism is a hallmark of chronic HBV infection and hepatocarcinogenesis, suggesting a plausible mechanistic link. *INHBE* (inhibin betaE) participates in TGF- signaling, a pathway frequently co-opted by HBV to modulate immune evasion and fibrogenesis. The cytokine-receptor genes highlighted by the GSFM approach (*IL20RA*, *IL13RA1*, *IFNLR1*) belong to the interleukin family that orchestrates antiviral immunity; their low publication frequency may reflect a historical focus on classical interferon pathways rather than these ancillary receptors.

FAXDC2, one of the frequency-driven genes, was significantly down-regulated in the GSE236281 cohort, reinforcing its potential involvement in HBV-driven transcriptional reprogramming. Conversely, several GSFM-derived genes (e.g., *IL18RAP*, *TNFRSF10A*) are up-regulated in HBV-infected samples, aligning with known pro-inflammatory and apoptotic signaling cascades that contribute to liver injury and tumor promotion.

Implications for therapeutic discovery

The drug-prediction analysis using L2S2 identified MAPK pathway inhibitors (e.g., trametinib, selumetinib) and the SERCA inhibitor thapsigargin as top candidates associated with the combined up/down gene signatures. Notably, MAPK signaling intersects with many of the understudied genes (e.g., *LRRC59* influences ribosomal biogenesis, which can be modulated by MAPK activity). These findings suggest that repurposing existing kinase inhibitors could be explored in pre-clinical models where the understudied genes are experimentally perturbed.

Limitations

Several methodological constraints should be acknowledged:

- **Publication count as a proxy for study depth**: Counting PubMed entries where a gene appears in the title or abstract may underestimate research activity for genes studied primarily in the context of broader pathways or in non-human models.
- **Gene set heterogeneity**: The aggregated gene sets span diverse experimental platforms and disease contexts (e.g., HBV-related HCC, immune cell responses). This heterogeneity may introduce noise, potentially inflating the apparent frequency of certain genes.
- **Single GEO dataset for validation**: Differential expression validation relied on one dataset (GSE236281). While this study provides a clear case–control contrast, broader validation across multiple cohorts and tissue types (liver tissue, hepatocytes, immune cells) would strengthen confidence in the identified candidates.
- **GSFM model interpretability**: The GSFM predictions are based on large-scale language-model embeddings and may capture indirect associations that lack mechanistic grounding. Experimental validation is essential to discriminate true biological relevance from spurious correlations.

Future directions

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

1. **Targeted CRISPR screens**: Perform loss- and gain-of-function screens in HBV-infected hepatocyte models (e.g., HepG2-NTCP cells) focusing on the top understudied genes to assess effects on viral replication, cccDNA maintenance, and host transcriptional programs.
2. **Multi-omics integration**: Combine proteomics, phosphoproteomics, and epigenomic datasets from HBV-infected tissues to map the down-

stream networks of the candidate genes and identify potential feedback loops with known HBV effectors.

3. **In vivo validation**: Generate liver-specific knockout mouse models for select candidates (e.g., *SC5D*, *IL20RA*) and evaluate HBV transgenic infection outcomes, including viral load, liver inflammation, and tumor development.
4. **Drug repurposing assays**: Test the efficacy of the top L2S2-predicted compounds (trametinib, thapsigargin) in cell culture and animal models where the understudied genes are modulated, to determine synergistic antiviral or anti-fibrotic effects.
5. **Literature mining refinement**: Incorporate full-text mining and citation network analysis to refine the definition of "understudied" and capture emerging studies that may not yet be indexed in PubMed abstracts.

Conclusion

By systematically intersecting disease-associated gene collections with bibliometric metrics and advanced predictive modeling, we have highlighted a set of HBV-related genes that are both biologically prominent and under-explored in the literature. These candidates represent fertile ground for mechanistic studies that could uncover novel pathways of HBV pathogenesis, identify new biomarkers of disease progression, and reveal therapeutic targets amenable to drug repurposing. Continued integration of computational prioritization with experimental validation will be essential to bridge the current knowledge gap and advance the fight against chronic hepatitisB.

Acknowledgements

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

References

[1] Aparna Schweitzer, Johannes Horn, Rafael T. Mikolajczyk, and et al. Estimations of worldwide prevalence of chronic hepatitis b virus infection: a systematic review of data published between 1965 and 2013. *Lancet*, 2015.

[2] European Association for the Study of the Liver. Easl 2017 clinical practice guidelines on the management of hepatitis b virus infection. *J Hepatol*, 2017.

[3] Chien-Jen Chen, Hwai-I Yang, Jun Su, and et al. Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis b virus dna level. *JAMA*, 2006.

[4] S. K. Sarin, M. Kumar, G. K. Lau, and et al. Asian-pacific clinical practice guidelines on the management of hepatitis b: a 2015 update. *Hepatol Int*, 2016.

[5] D. Lavanchy. Hepatitis b virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *J Viral Hepat*, 2004.

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

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

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

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

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

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

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

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

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

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

[16] 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.

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

[18] 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.