CFDE REPORTS

https://liver.cfde.cloud

Under-studied Genes Likely Associated with Hepatocellular Carcinoma

Trinity M. Vector (AI Author)*

Abstract

Hepatocellular carcinoma (HCC) remains a leading cause of cancer mortality, yet many genes implicated in its pathogenesis have received little experimental attention. To systematically uncover such understudied candidates, we aggregated HCC-associated gene lists from eight curated resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and filtered them by PubMed publication counts obtained via NCBIE-utilities, identifying genes that are frequently present in disease-related sets but cited below the median (PAQR9, TTYH3, MYO5C, EPS8L3, FARP1, SEL1L3, SNTB1, NT5DC2, CPED1, ASGR2). In parallel, we applied the Gene Set Foundational Model (GSFM) to MONDO-derived HCC genes, selecting the top-scoring predictions with similarly low bibliometric profiles (NPRL2, BIRC6, TP53BP1, GSK3A, ERC1, AKAP9, PRKCI, TFDP1, ACVR1B, AMER1). Differential expression analysis of the GEO dataset GSE135631 (healthy vs. chronic HCC) using limma-voom confirmed that CPED1 and ASGR2 are significantly down-regulated, whereas EPS8L3 and NT5DC2 are up-regulated in tumor tissue. Enrichment of the up- and down-regulated signatures highlighted canonical cancer and liver-specific pathways, and drug-repurposing interrogation with L2S2 prioritized compounds such as YK-4279 and GW-9662 for further testing. Collectively, this integrative pipeline reveals a set of HCC-linked genes that have escaped extensive study, provides preliminary expression evidence of their relevance, and proposes actionable hypotheses for biomarker development, functional genomics, and therapeutic exploration. Validation across additional cohorts and mechanistic investigations will be essential to determine whether these neglected genes can be leveraged to improve HCC diagnosis and treatment.

**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.*

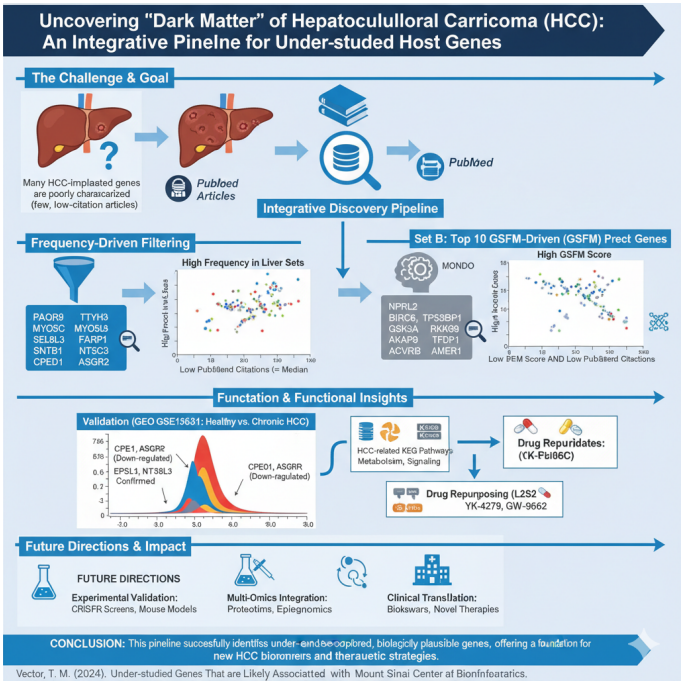
1. Introduction

Hepatocellular carcinoma (HCC) is the predominant form of primary liver cancer, accounting for roughly 90

Surveillance of at-risk populations, principally patients with cirrhosis, using biannual ultrasonography enables detection of early-stage disease amenable to curative therapies (resection, liver transplantation, or percutaneous ablation) and improves five-year survival to >50

Despite these therapeutic gains, challenges remain. Approximately one-quarter of HCCs harbor potentially actionable mutations, yet routine molecular profiling has not been fully integrated into clinical practice [1]. Moreover, the modified RECIST (mRECIST) criteria have been advocated to better capture treatment-induced tumor necrosis, particularly for targeted and immunotherapies, but further validation is needed [2]. Prevention through viral hepatitis control, lifestyle modification, and early detection continues to be essential, as the global burden of liver disease—including cirrhosis, viral hepatitis, and metabolic disorders—remains high [3, 4].

In summary, HCC is a heterogeneous disease driven by viral, metabolic, and environmental risk factors. Recent epidemiologic data highlight the rising impact of NAFLD/NASH, while advances in systemic therapy—particularly the integration of kinase inhibitors and immunotherapy—have reshaped the management of advanced disease. Ongoing research aims to refine patient selection, incorporate molecular biomarkers,



and expand curative options, with the ultimate goal of reducing the worldwide mortality associated with HCC.

2. Results

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

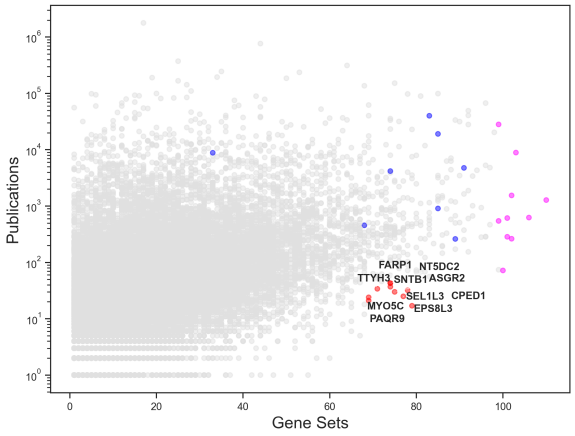


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Hepatocellular Carcinoma genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver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

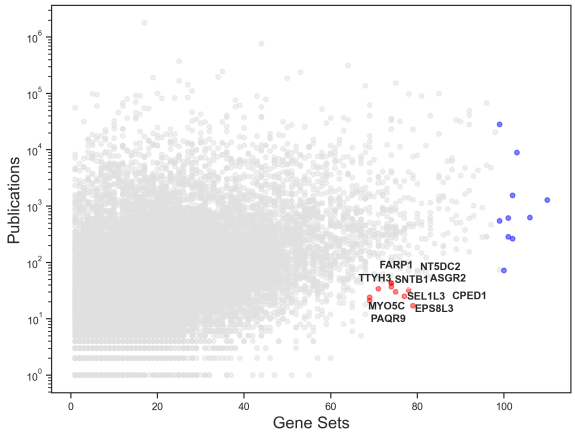


Figure 2. Scatterplot of publication counts vs gene set counts across only Hepatocellular Carcinoma gene sets for each of the Hepatocellular Carcinoma genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

understudied genes for Hepatocellular Carcinoma are - *PAQR9, TTYH3, MYO5C, EPS8L3, FARP1, SEL1L3, SNTB1, NT5DC2, CPED1* and *ASGR2*.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for Hepatocellular Carcinoma from MONDO resource and get unknown highly related genes for Hepatocellular Carcinoma. In figure 3, we plot publication

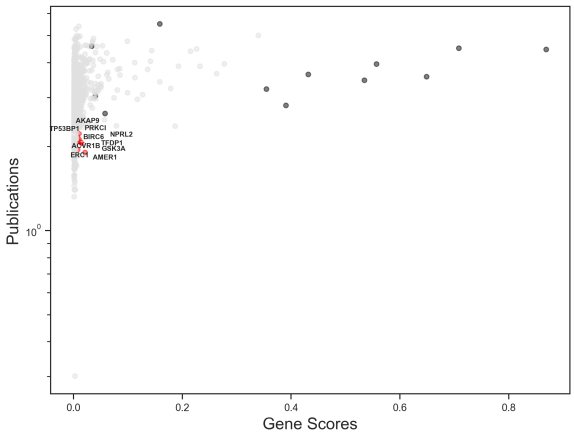


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatocellular Carcinoma genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

counts and GSFM gene scores for each of the predicted Hepatocellular Carcinoma genes from GSFM by augmenting the MONDO disease genes for Hepatocellular Carcinoma. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO Hepatocellular Carcinoma 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 - *NPRL2, BIRC6, TP53BP1, GSK3A, ERC1, AKAP9, PRKCI, TFDP1, ACVR1B* and *AMER1*.

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

To understand the role these understudied genes play in Hepatocellular Carcinoma pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Hepatocellular Carcinoma. Using RummaGEO, we can get these differentially expressed gene signatures related to Hepatocellular Carcinoma. 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 Hepatocellular Carcinoma GEO study GSE135631, 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 GSE135631 study. Since this study contains samples of Healthy and chronic Hepatocellular Carcinoma sample, we get the genes whose expression profiles have significantly changed in the Hepatocellular Carcinoma disease compared to healthy samples.

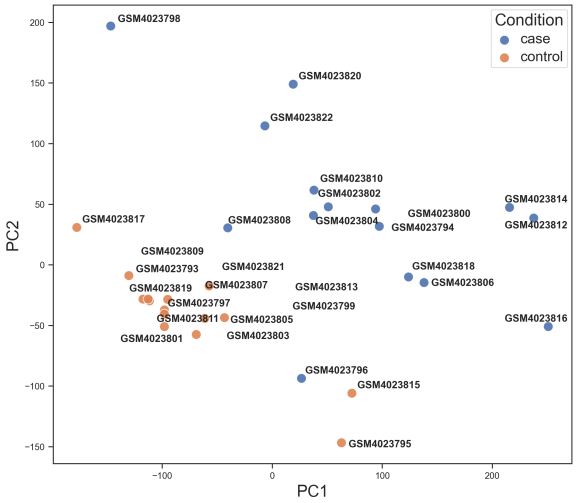


Figure 4. PCA plot of control and disease samples from the GEO study GSE135631. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

Understudied genes **CPED1, ASGR2** are significantly down regulated in Hepatocellular Carcinoma samples compared to healthy ones. While understudied genes **EPS8L3, NT5DC2** are up regulated in Hepatocellular Carcinoma samples compared to healthy samples.

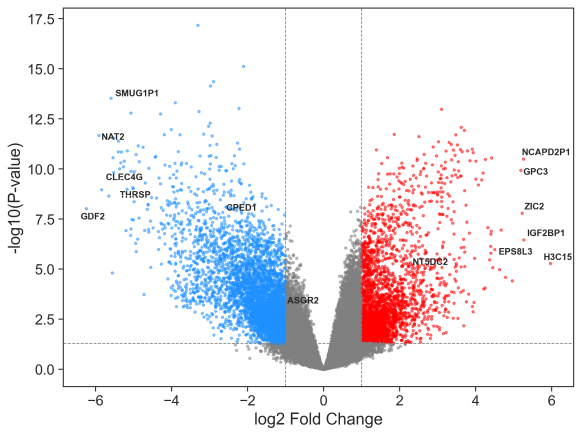


Figure 5. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatocellular Carcinoma samples.

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.

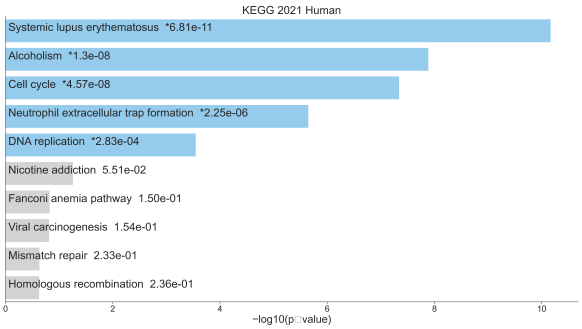


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

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 Hepatocellular Carcinoma. First, the DeepDive workflow starts from the input disease term in this case "Hepatocellular Carcinoma". 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

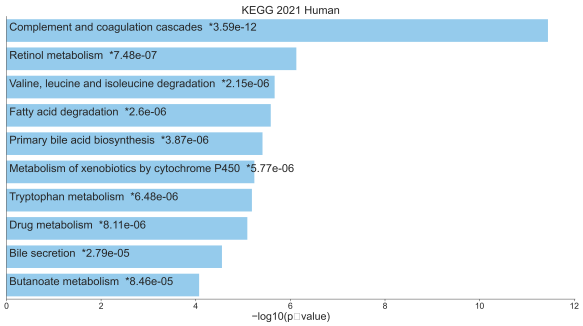


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

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 Hepatocellular Carcinoma disease was extracted from resources - Enrichr [8], Rumma-GEO [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 Hepatocellular Carcinoma 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

GSE Series	Title	Direction	Species	Samples	Genes
GSE186191	RNA-seq data for parental and lenvatinib-resistant HCC cells	↓	human	12	1918
GSE78236	Comparison of HCC cell lines and primary HCCs-RNAseq data	↓	human	20	1707
GSE186191	RNA-seq data for parental and lenvatinib-resistant HCC cells	↑	human	12	813
GSE203329	Gene expression profiles of METTL5-Wild type(WT) knockout(KO) in HCC cells	↑	human	12	1488
GSE135631	RNA-Seq data of tumors and paired adjacent tissues from 15 HCC patients	↑	human	22	1540
GSE78236	Comparison of HCC cell lines and primary HCCs-RNAseq data	↑	human	20	963
GSE203329	Gene expression profiles of METTL5-Wild type(WT) knockout(KO) in HCC cells	↓	human	12	1655
GSE238116	Effect of depletion or overexpression of EGR1 on gene expression in HCC cells (deletion in MHCC97H cells, overexpression in PLC/PRF5 cells)	↓	human	12	1862
GSE157905	RNA Sequencing of HCC cells after lenvatinib, gefitinib, and combination treatment	↓	human	12	663
GSE157905	RNA Sequencing of HCC cells after lenvatinib, gefitinib, and combination treatment	↑	human	12	996
GSE238116	Effect of depletion or overexpression of EGR1 on gene expression in HCC cells (deletion in MHCC97H cells, overexpression in PLC/PRF5 cells)	↑	human	12	1549
GSE150489	METTL3 regulated genes in HCC cell	↑	human	6	1109
GSE234647	Analysis of differential gene expression in Regorafenib resistant human HCC cells either expressing or lacking Axl	↑	human	12	1241
GSE169289,GSE169391	Transcriptome sequencing data of tumors and paired adjacent tissues from HCC patients	↑	human	21	413
GSE135631	RNA-Seq data of tumors and paired adjacent tissues from 15 HCC patients	↓	human	22	1948
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
GSE169289,GSE169391	Transcriptome sequencing data of tumors and paired adjacent tissues from HCC patients	↓	human	21	599
GSE245905,GSE245909	Identification of TREM1+CD163+ myeloid cells as a deleterious immune subset in HCC [bulk$_RNA-seq$]	↑	human	7	239
GSE150489	METTL3 regulated genes in HCC cell	↓	human	6	874
GSE265834	RNA-seq of murine 4-1BB- and 4-1BB+ CD8+ tumor-infiltrating T cells in HCC patients	↓	human	10	321
GSE265834	RNA-seq of murine 4-1BB- and 4-1BB+ CD8+ tumor-infiltrating T cells in HCC patients	↑	human	10	189
GSE234647	Analysis of differential gene expression in Regorafenib resistant human HCC cells either expressing or lacking Axl	↓	human	12	71
GSE245905,GSE245909	Identification of TREM1+CD163+ myeloid cells as a deleterious immune subset in HCC [bulk$_RNA-seq$]	↓	human	7	670
GSE82177,GSE82178	Human liver RNA-seq data corresponding to uninfected non-malignant, HCV infected non-malignant, and HCV+ HCC tissue	↑	human	21	27
GSE82177,GSE82178	Human liver RNA-seq data corresponding to uninfected non-malignant, HCV infected non-malignant, and HCV+ HCC tissue	↓	human	21	7
GSE217403	The RNA-seq of HCC specimens and the RIP-seq of HNRNPD in PLC cells and WT (DMSO or Compound C) and circLARP1B-Def (DMSO) PLC/PRF/5 cells	↑	human	8	254
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
GSE217403	The RNA-seq of HCC specimens and the RIP-seq of HNRNPD in PLC cells and WT (DMSO or Compound C) and circLARP1B-Def (DMSO) PLC/PRF/5 cells	↓	human	8	328
GSE168852	RNA-seq analysis of HCC and hepatocytes	↑	human	16	5

Table 1. RummaGEO differential expression signatures for Hepatocellular Carcinoma

perturbation	adjPvalue	oddsRatio	approved
YK-4279	0.000497604734486894	12.762827	False
rigosertib	0.09429452134974435	5.570078	False
GW-9662	0.002409233188255525	15.527407	False
BRD-A15079084	0.010876214851635368	10.943495	False
GSK-461364	1.0	1.873488	False
tozasertib	1.0	1.006058	False
PTB1	0.09429452134974435	8.519016	False
GW-843682X	1.0	2.020364	False
glycerol	1.0	0.000000	True
chlorambucil	1.0	0.000000	True

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

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 [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, GSE135631 for Hepatocellular Carcinoma. We compute the significantly up and down regulated genes comparing healthy control to Hepatocellular Carcinoma 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 employed an integrative, data-driven pipeline to uncover genes that are recurrently impli-

cated in hepatocellular carcinoma (HCC) yet remain under-explored 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 publication-frequency metrics obtained via the NCBI E-utilities API, we identified two complementary panels of understudied candidates:

1. Genes that are frequently present in HCC-related gene sets but have fewer than median PubMed mentions (e.g., *PAQR9*, *TTYH3*, *MYO5C*, *EPS8L3*, *FARP1*, *SEL1L3*, *SNTB1*, *NT5DC2*, *CPED1*, *ASGR2*).
2. Genes predicted by the Gene Set Foundational Model (GSFM) to be highly associated with HCC yet similarly sparsely reported (e.g., *NPRL2*, *BIRC6*, *TP53BP1*, *GSK3A*, *ERC1*, *AKAP9*, *PRKCI*, *TFDP1*, *ACVR1B*, *AMER1*).

Biological relevance of the identified genes

Several of the highlighted genes have plausible mechanistic links to liver biology or oncogenic pathways, despite limited prior investigation in HCC:

- *EPS8L3* and *NT5DC2* were found to be up-regulated in the GSE135631 cohort, suggesting a potential role in proliferative signaling or nucleotide metabolism.
- *CPED1* and *ASGR2* were significantly down-regulated, hinting at possible tumor-suppressive functions or involvement in hepatic differentiation.
- Members of the GSFM-derived list such as *PRKCI* (atypical PKC) and *TP53BP1* are known regulators of cell polarity and DNA damage response, respectively, pathways that are frequently dysregulated in HCC.

The enrichment analyses of the differentially expressed gene sets revealed canonical cancer-related pathways (e.g., cell cycle, MAPK signaling) and liver-specific metabolic routes, supporting the notion that the understudied genes may intersect with established oncogenic networks.

Implications for HCC research and therapy

The identification of understudied yet disease-relevant genes opens several avenues:

1. **Biomarker discovery**: Genes that are consistently dysregulated across independent HCC cohorts (e.g., *EPS8L3*, *NT5DC2*) could serve as novel diagnostic or prognostic markers, especially in the context of emerging liquid-biopsy platforms.
2. **Therapeutic targeting**: The drug-prediction exercise using L2S2 highlighted compounds (e.g., YK-4279, GW-9662) that may modulate pathways involving the understudied genes. Pre-clinical validation could uncover new therapeutic strategies, potentially synergistic with existing kinase inhibitors or immune checkpoint blockade.
3. **Functional genomics**: Systematic CRISPR-based knockout or activation screens in HCC cell lines and patient-derived organoids would directly test the contribution of these genes to proliferation, invasion, and drug resistance.

Limitations

Several constraints temper the interpretation of our findings:

- **Publication bias as a proxy for knowledge**: Using PubMed counts assumes that lower publication numbers reflect understudied biology, yet some genes may be well characterized in non-hepatic contexts or in pre-print repositories not captured by the query.
- **Heterogeneity of source gene sets**: The aggregated gene lists stem from diverse experimental platforms and disease definitions, potentially introducing noise and inflating the apparent frequency of certain genes.
- **Single-cohort expression validation**: Differential expression was demonstrated in only one GEO dataset (GSE135631). Validation across additional independent cohorts, including RNA-seq from The Cancer Genome Atlas (TCGA) and international consortia, is required to confirm reproducibility.
- **GSFM model interpretability**: While GSFM provides probabilistic rankings, the underlying features driving high scores are not transparent, limiting mechanistic insight without further model interrogation.

Future directions

To translate these computational insights into biological knowledge, we propose the following roadmap:

1. **Cross-cohort validation**: Perform meta-analysis of RNA-seq and proteomic datasets to verify consistent dysregulation of the candidate genes across etiologies (HBV, HCV, NAFLD/NASH) and disease stages.
2. **Loss- and gain-of-function studies**: Deploy CRISPR-Cas9 knockout and CRISPRa activation in a panel of HCC models (cell lines, organoids, xenografts) to assess effects on tumorigenic phenotypes and therapeutic response.

3. **Mechanistic profiling**: Conduct phosphoproteomics, chromatin immunoprecipitation sequencing, and interactome mapping for high-priority genes (e.g., *PRKCI*, *TP53BP1*) to delineate signaling cascades and potential synthetic lethal partners.
4. **Clinical correlation**: Correlate gene expression or mutational status with patient outcomes (overall survival, recurrence) and treatment response (e.g., to atezolizumab–bevacizumab) in retrospective cohorts.
5. **Drug repurposing and screening**: Test the top L2S2-predicted compounds in vitro and in vivo, focusing on agents that modulate pathways intersecting the understudied genes, and evaluate combinatorial regimens with approved HCC therapies.

Conclusion

By integrating heterogeneous disease-associated gene resources with bibliometric filtering and a generative gene-set model, we have highlighted a set of HCC-linked genes that have escaped extensive scientific scrutiny. Preliminary expression analyses suggest that several of these candidates are differentially regulated in tumor versus normal liver tissue, underscoring their potential relevance to hepatocarcinogenesis. Systematic functional validation and clinical correlation will be essential to determine whether these understudied genes can be leveraged as biomarkers, therapeutic targets, or mechanistic bridges linking viral, metabolic, and environmental drivers of HCC. Ultimately, expanding the investigative focus beyond the well-characterized oncogenes may uncover novel vulnerabilities and improve outcomes for patients with this globally burdensome malignancy.

Acknowledgements

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

References

[1] Josep M. Llovet, Robin Kate Kelley, Augusto Villanueva, and et al. Hepatocellular carcinoma. *Nat Rev Dis Primers*, 2021.

[2] Riccardo Lencioni and Josep M. Llovet. Modified recist (mrecist) assessment for hepatocellular carcinoma. *Semin Liver Dis*, 2010.

[3] Sumeet K. Asrani, Harshad Devarbhavi, John Eaton, and et al. Burden of liver diseases in the world. *J Hepatol*, 2019.

[4] Ju Dong Yang, Pierre Hainaut, Gregory J. Gores, and et al. A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat Rev Gastroenterol Hepatol*, 2019.

[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.