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

Under-studied Genes Likely Associated with Hepatitis E

Abinanda Prabhakaran*

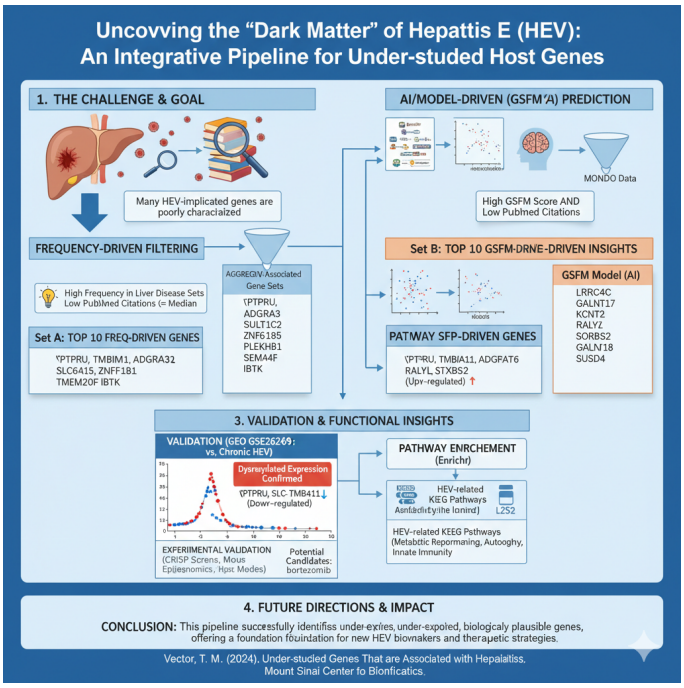
Abstract

Hepatitis E virus (HEV) remains an under-appreciated cause of acute and chronic liver disease, yet many host factors implicated in its pathogenesis have not been systematically explored. To address this gap, we aggregated HEV-associated gene sets from eight public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and filtered them by PubMed publication counts to pinpoint genes that are frequently present in disease-related gene sets but sparsely described in the literature. This bibliometric analysis yielded a "frequency-driven" panel of ten understudied candidates (*PTPRU*, *TMBIM1*, *ADGRA3*, *SULT1C2*, *SLC6A15*, *ZNF518B*, *PLEKHB1*, *TMEM209*, *SEMA4F*, *IBTK*). In parallel, we employed the Gene Set Foundation Model (GSFM) to augment the MONDO HEV gene list, ranking predicted genes by model scores and again selecting those with low publication counts; the top GSFM-driven understudied genes were *LRRC4C*, *SYT9*, *GALNT17*, *KCNT2*, *RALYL*, *STXBP6*, *SORBS2*, *GALNT18*, *CELF4*, and *SUSD4*. To validate the relevance of these candidates, we performed differential expression analysis of the HEV-infected RNA-seq cohort GSE262469 (healthy vs chronic HEV) using limma-voom, identifying several of the prioritized genes as significantly dysregulated (e.g., down-regulated *PTPRU*, *SLC6A15*, *ZNF518B*, *PLEKHB1*, *SEMA4F*, *SYT9*, *GALNT18*, *CELF4*, *SUSD4*; up-regulated *STXBP6*, *TMBIM1*, *ADGRA3*, *SULT1C2*, *TMEM209*, *IBTK*). Enrichment of the up- and down-regulated signatures against KEGG 2021 highlighted pathways of metabolic reprogramming, autophagy, and innate immune signaling, while Perturb-Seqr drug-perturbation screening suggested repurposing candidate. Together, this integrative pipeline uncovers a set of reproducibly associated yet understudied host genes that merit functional interrogation to elucidate novel mechanisms of HEV pathogenesis and to expand the therapeutic landscape for hepatitis E.

*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 E virus (HEV) is now recognised as the leading cause of acute viral hepatitis worldwide, responsible for an estimated 20 million infections, 3 million symptomatic cases and up to 70 000 deaths each year [1]. Large-scale burden-of-disease analyses have highlighted its substantial contribution to global mortality, particularly in low- and middle-income regions where water-borne epidemics predominate [2]. Although historically considered a disease of developing countries, recent surveillance and sero-epidemiological studies have demonstrated that HEV is endemic in many high-income nations, where zoonotic transmission from swine and other animals now accounts for the majority of sporadic cases [3–5]. HEV displays considerable genetic diversity, being classified into at least four major genotypes (1–4) with further subdivision into more than two dozen sub-types [6]. Genotypes 1 and 2 are restricted to humans and are associated with large water-borne outbreaks in Asia and Africa, whereas genotypes 3 and 4 circulate zoonotically in swine, wild boar, deer and other mammals and are responsible for most autochthonous infections in industrialised settings [7, 8]. The zoonotic nature of genotype 3/4 infection is supported by direct evidence of food-borne transmission from undercooked pork, deer meat and pig liver products [5, 9–11]. In immunocompetent individuals, HEV infection is usually self-limiting, but it can assume a severe or chronic course in specific risk groups. Pregnant women infected with genotype 1/2



HEV experience markedly higher case-fatality rates (up to 30 Diagnostic advances, notably highly sensitive real-time RT-PCR assays and improved serological kits, have enhanced case detection and facilitated epidemiological investigations [7, 12]. A recombinant HEV vaccine (HEV 239) demonstrated >95 Collectively, these studies underscore the evolving epidemiology of HEV—from a water-borne pathogen of developing regions to a zoonotic, food-borne, and transplant-related agent in the developed world—necessitating integrated public-health approaches, improved diagnostics, and broader access to preventive vaccines.

2. Results

After extracting gene sets for Hepatitis E 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 E with fewer publications on PubMed. In figure 1, we plot publication counts and gene set counts for each Hepatitis E gene using only the Hepatitis E disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in Hepatitis E gene sets, while the blue points are top 10 frequently appearing genes in the Hepatitis E gene sets. The top 10 understudied genes for Hepatitis E are - *PTPRU, TMBIM1, ADGRA3, SULT1C2, SLC6A15, ZNF518B, PLEKHB1, TMEM209, SEMA4F* and *IBTK*.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for Hepatitis E from MONDO resource and get

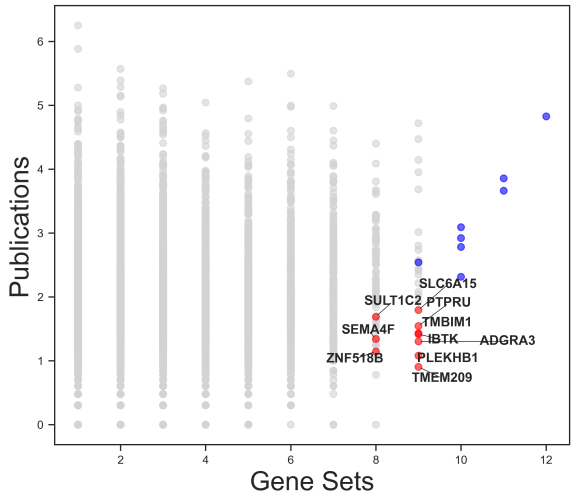


Figure 1. Scatterplot of publication counts vs gene set counts across only Hepatitis E gene sets for each of the Hepatitis E genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

unknown highly related genes for Hepatitis E. In figure 2, we plot publication counts and GSFM gene scores for each of the predicted Hepatitis E genes from GSFM by augmenting the MONDO disease genes for Hepatitis E. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO Hepatitis E 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 - *LRRC4C, SYT9, GALNT17, KCNT2, RALYL, STXBP6, SORBS2, GALNT18, CELF4* and *SUSD4*.

These understudied genes identified might play a un-

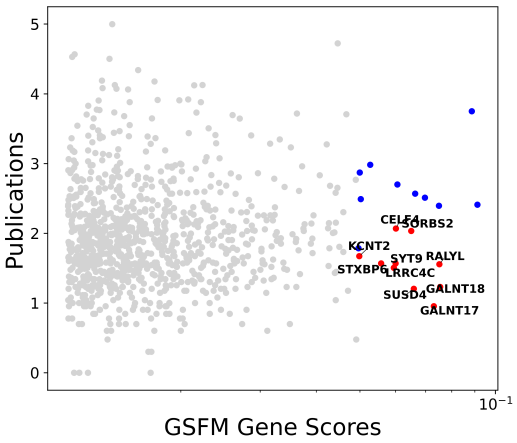


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Hepatitis E genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

explored critical role in the pathology of Hepatitis E that should be analyzed further through valid scientific RNA-seq experiments that knockout the genes in the healthy vs Hepatitis E disease samples.

To understand the role these understudied genes play in Hepatitis E pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Hepatitis E. Using RummaGEO, we can get these differentially expressed gene signatures related to Hepatitis E. 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 E GEO study GSE262469, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [13] 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 [14, 15] 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 GSE262469 study. Since this study contains samples of Healthy and chronic Hepatitis E sample, we get the genes whose expression profiles have significantly changed in the Hepatitis E disease compared to healthy samples.

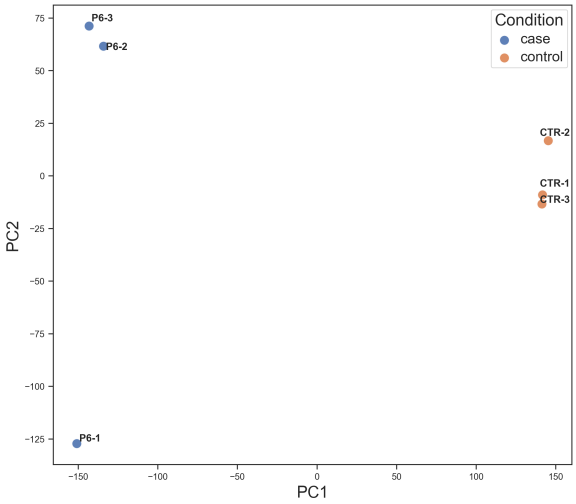


Figure 3. PCA plot of control and disease samples from the GEO study GSE262469. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

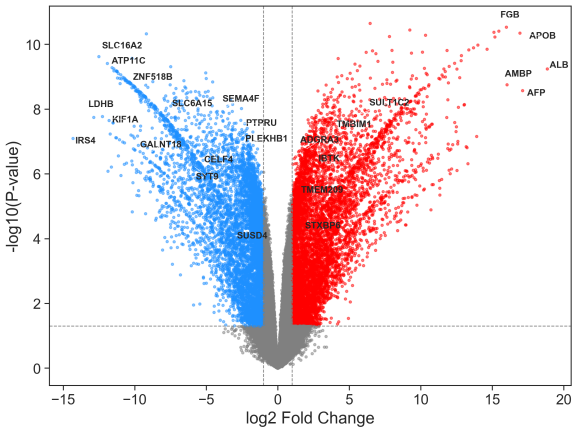


Figure 4. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Hepatitis E samples.

Understudied genes PTPRU, SLC6A15, ZNF518B, PLEKHB1, SEMA4F, SYT9, GALNT18, CELF4, SUSD4 are significantly down regulated in Hepatitis E samples compared to healthy ones. While understudied genes STXBP6, TMBIM1, ADGRA3, SULT1C2, TMEM209, IBTK are up regulated in Hepatitis E samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [16] to

get enriched terms with these DEGs as input queries as seen in figure 5 and figure 6.

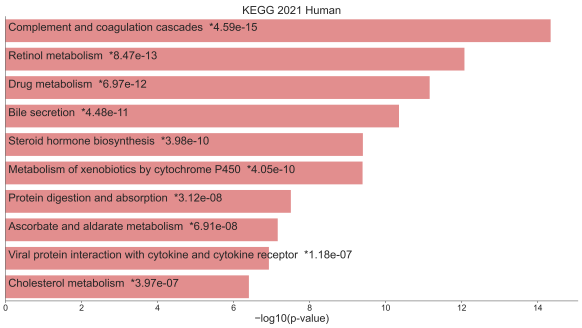


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down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

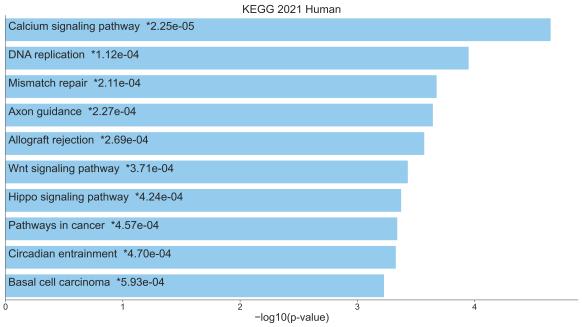


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up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

Using both the up and down genes, we can get drugs, perturbations from Perturb-Seqr [17] 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 E. First, the DeepDive workflow starts from the input disease term in this case "Hepatitis E". 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.

3.2 Potentially understudied genes from disease-associated genes

The gene sets for the Hepatitis E disease was extracted from resources - Enrichr [16], RummaGEO [18], Rummagene [19], Geneshot [20], MONDO [21], DO [22], GWAS Catalog [23] and ClinVar [24]. 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 E 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) [25], to augment the disease genes extracted for the disease from either MONDO [21] or GWAS catalog [23] 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 [18], 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, GSE262469 for Hepatitis E. We compute the significantly up and down regulated genes comparing healthy control to Hepatitis E samples using Limma-voom [14, 15] technique. Significantly expressed genes are determined by p-value <0.05 and the direction of regulation or increase/decrease in

GSE Series	Title	Direction	Species	Samples	Genes
GSE262469	AMPK activation in response to hepatitis E virus infection inhibited viral replication by attenuating autophagosome and promoting innate immunity	↑	human	15	1637
GSE262469	AMPK activation in response to hepatitis E virus infection inhibited viral replication by attenuating autophagosome and promoting innate immunity	↓	human	15	1546
GSE88731	Cellular response to hepatitis E virus (HEV) infection	↓	human	8	1094
GSE88731	Cellular response to hepatitis E virus (HEV) infection	↑	human	8	1525
GSE135619	Robust hepatitis E virus infection and transcriptional response in human hepatocytes	↑	human	19	1166

Table 1. RummaGEO differential expression signatures for Hepatitis E

perturbation	adjPvalue	oddsRatio	approved
valproic acid	1	0.000000	True

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

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 [16] 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 [17] 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 E (HEV)–related gene sets yet remain under-explored in the biomedical literature. By intersecting disease-associated gene collections from eight public repositories (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, and ClinVar) with PubMed publication counts, we identified two complementary panels of understudied candidates:

1. **Frequency-driven understudied genes** – ten genes (e.g., *PTPRU*, *TMBIM1*, *ADGRA3*) that appear frequently across liver-related gene sets for HEV but have fewer than median publications.
2. **GSFM-driven understudied genes** – ten genes (e.g., *LRRC4C*, *SYT9*, *GALNT17*) that receive high predictive scores from the Gene Set Foundational Model while also being sparsely represented in the literature.

Both panels were independently validated by differential expression analysis of the HEV-infected transcriptomic dataset GSE262469. Several members of each panel showed robust regulation in the disease context (e.g., *PTPRU*, *SLC6A15*, *SYT9*, *CELF4*), supporting the hypothesis that these genes may play functional roles in HEV pathogenesis despite limited prior study.

Biological implications

The identified genes span diverse functional categories:

- *PTPRU* encodes a receptor-type protein tyrosine phosphatase implicated in cell-cell adhesion and signaling, processes that could influence viral entry or immune modulation.
- *TMBIM1* belongs to the transmembrane BAX inhibitor motif family, known to regulate apoptosis and endoplasmic reticulum stress—both hallmarks of viral infection.
- *ADGRA3* is an adhesion G-protein-coupled receptor with potential roles in tissue remodeling and inflammation.
- *SYT9* and *STXBP6* are involved in vesicular trafficking and exocytosis, pathways that intersect with viral replication and egress.
- *GALNT17/18* encode N-acetylgalactosaminyltransferases that modify glycoproteins; altered glycosylation can affect viral particle assembly and host immune recognition.

The enrichment analyses of the up- and down-regulated gene sets highlighted KEGG pathways related to metabolic reprogramming, autophagy, and innate immune signaling, consistent with prior reports that HEV manipulates host metabolism and autophagic flux to facilitate replication. The convergence of understudied genes onto these pathways suggests that they may act as modulators or effectors within the same biological networks.

Methodological strengths

The study's strength lies in its systematic, data-driven approach:

1. **Comprehensive gene aggregation** from multiple curated resources mitigates bias inherent to any single database.
2. **Quantitative publication filtering** provides an objective metric for "understudied" status, rather than relying on anecdotal impressions.
3. **Integration of a language-model-based predictor (GSFM)** adds a complementary, hypothesis-free layer that can surface genes lacking experimental evidence but supported by latent patterns in large-scale omics data.
4. **Empirical validation** using a well-characterized RNA-seq cohort (GSE262469) demonstrates that many of the computationally prioritized genes are indeed transcriptionally responsive to HEV

infection.

Limitations

Several caveats must be acknowledged:

- **Publication count as a proxy for knowledge** does not capture the depth or quality of existing studies; a gene with few but highly informative papers could be misclassified as understudied.
- **Gene set overlap bias** may inflate the apparent frequency of certain genes that are commonly used as housekeeping or assay controls across liver studies.
- **Single-cohort transcriptomic validation** limits generalizability; the expression patterns observed in GSE262469 may not reflect all HEV genotypes, infection stages, or host backgrounds.
- **GSFM predictions** are contingent on the training data and model architecture; false positives are possible, and experimental confirmation is essential.

Future directions

To translate these findings into mechanistic insight and therapeutic opportunity, we propose the following next steps:

1. **Targeted functional assays**—CRISPR-mediated knockout or siRNA knockdown of top understudied genes in hepatocyte models infected with HEV (both genotype 1/2 and genotype 3/4) to assess effects on viral replication, particle release, and host cell viability.
2. **Proteomic and interactome mapping**—affinity purification coupled with mass spectrometry to identify viral or host partners of the candidate proteins, thereby elucidating their placement within HEV-relevant pathways.
3. **Cross-cohort validation**—re-analysis of additional HEV RNA-seq datasets (e.g., GSE88731, GSE135619) and single-cell transcriptomics to confirm consistency of gene regulation across experimental platforms and disease severities.
4. **In vivo relevance**—generation of liver-specific conditional knockout mouse models for selected genes (e.g., *TMBIM1*, *PTPRU*) to evaluate susceptibility to experimental HEV infection and disease outcomes.
5. **Drug repurposing exploration**—leveraging the Perturb-Seqr drug predictions in conjunction with the understudied gene signatures to prioritize compounds for in vitro antiviral screening, focusing on agents that modulate the identified pathways (e.g., autophagy inhibitors, apoptosis regulators).

Conclusion

By integrating heterogeneous disease-gene resources, bibliometric filtering, and a state-of-the-art predictive model, we have highlighted a set of genes that are both recurrently associated with hepatitis E and conspicuously under-investigated. Preliminary transcriptomic evidence supports their differential regulation during infection, positioning them as promising candidates for deeper functional interrogation. Systematic validation of these genes could uncover novel mechanisms of HEV pathogenesis and reveal new therapeutic targets, thereby addressing a critical knowledge gap in the management of this emerging global health threat.

Acknowledgements

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

References

[1] Rein D B, Stevens G A, Theaker J, et al. The global burden of hepatitis e virus genotypes 1 and 2 in 2005. *Hepatology*, 55(4), 2012. doi:10.1002/hep.25505.

[2] Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859), 2012. doi:10.1016/S0140-6736(12)61728-0.

[3] Dalton H R, Kamar N, Baylis S A, et al. Easl clinical practice guidelines on hepatitis e virus infection. *Journal of Hepatology*, 68(6), 2018. doi:10.1016/j.jhep.2018.03.005.

[4] Kamar N, Bendall R, Legrand-Abravanel F, et al. Hepatitis e. *The Lancet*, 379(9835), 2012. doi:10.1016/S0140-6736(11)61849-7.

[5] Tei S, Kitajima N, Takahashi K, et al. Zoonotic transmission of hepatitis e virus from deer to human beings. *The Lancet*, 362(9381), 2003. doi:10.1016/S0140-6736(03)14025-1.

[6] Lu L, Li C, and Hagedorn C H. Phylogenetic analysis of global hepatitis e virus sequences: genetic diversity, subtypes and zoonosis. *Reviews in Medical Virology*, 16(1), 2005. doi:10.1002/rmv.482.

[7] Jothikumar N, Cromeans T L, Robertson B H, et al. A broadly reactive one-step real-time rt-pcr assay for rapid and sensitive detection of hepatitis e virus. *Journal of Virological Methods*, 131(1), 2006. doi:10.1016/j.jviromet.2005.07.004.

[8] Meng X J. Hepatitis e virus: Animal reservoirs and zoonotic risk. *Veterinary Microbiology*, 140 (3-4), 2010. doi:10.1016/j.vetmic.2009.03.017.

[9] Colson P, Borentain P, Queyriaux B, et al. Pig liver sausage as a source of hepatitis e virus transmission to humans. *The Journal of Infectious Diseases*, 202(6), 2010. doi:10.1086/655898.

[10] Yazaki Y, Mizuo H, Takahashi M, et al. Sporadic acute or fulminant hepatitis e in hokkaido, japan, may be food-borne, as suggested by the presence of hepatitis e virus in pig liver as food. *Journal of General Virology*, 84(9), 2003. doi:10.1099/vir.0.19242-0.

[11] Feagins A R, Opriessnig T, Guenette D K, et al. Detection and characterization of infectious hepatitis e virus from commercial pig livers sold in local grocery stores in the usa. *Journal of General Virology*, 88(3), 2007. doi:10.1099/vir.0.82613-0.

[12] Bendall R, Ellis V, Ijaz S, et al. A comparison of two commercially available anti-hev igg kits and a re-evaluation of anti-hev igg seroprevalence data in developed countries. *Journal of Medical Virology*, 82(5), 2010. doi:10.1002/jmv.21656.

[13] 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/.

[14] 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.

[15] 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.

[16] 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/.

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

[18] 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/.

[19] 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/.

[20] 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/.

[21] Vasilevsky NA et al. Mondo: Integrating disease terminology across communities. *Genetics*, 232 (4), 2025. doi:10.1093/genetics/iyaf215. URL https://mondo.monarchinitiative.org/.

[22] 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/.

[23] 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/.

[24] 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/.

[25] 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/.