Cross-Tissue Biomarkers for Hepatitis B Virus-Related Hepatocellular Carcinoma: A Comparative Study of Blood and Liver Transcriptomics

Mayurakshi Mukherji[1], Shubham Thakur[2], Fenil Parmar[3], and Saket Choudhary [4]

[1]Koita Centre For Digital Health , 23d1628@iitb.ac.in
[2]Koita Centre For Digital Health , 24d1622@iitb.ac.in
[3]Koita Centre For Digital Health , 24d1624@iitb.ac.in
[4]Koita Centre For Digital Health , saketc@iitb.ac.in

ABSTRACT

Hepatitis B Virus (HBV) is a leading cause of hepatocellular carcinoma (HCC) worldwide, especially in regions with high HBV prevalence. The transition from HBV infection to HBV-related HCC (HBV-HCC) involves distinct molecular changes in liver tissue, and understanding these changes could provide valuable biomarkers for early, non-invasive detection. This study investigates differential gene expression in blood and liver samples from healthy individuals, HBV-infected patients, and HBV-HCC patients. By analyzing bulk and single-cell RNA sequencing data, we aim to identify cross-tissue biomarkers that indicate HBV progression to HCC. We employ a multi-omics approach integrating transcriptomics, epigenomics, and proteomics to validate identified biomarkers, focusing on gene expression patterns, pathway enrichment, and tissue-specific molecular signatures. This study seeks to contribute valuable insights into non-invasive HBV-HCC diagnostics and the broader understanding of HBV-associated carcinogenesis.

INTRODUCTION

Hepatocellular carcinoma (HCC) is a common form of liver cancer, and chronic HBV infection is a primary driver of HCC development globally, responsible for nearly half of HCC cases worldwide. The transition from HBV infection to HBV-related HCC (HBV-HCC) is complex, with genetic, epigenetic, and proteomic alterations in both liver and blood samples. Early detection of HBV-HCC, particularly through non-invasive methods, is crucial for effective disease management and prognosis improvement. While liver tissue remains the primary focus of HCC studies, identifying biomarkers in blood could enable more accessible screening, reducing the need for invasive procedures.

This study focuses on comparative transcriptomic analysis of blood and liver samples from various patient groups: healthy, HBV-infected, and HBV-HCC individuals. We aim to understand gene expression differences across these groups, identify unique HBV-related molecular signatures, and explore cross-tissue biomarkers that could serve as non-invasive indicators of disease progression. Additionally, we validate our findings, providing a multi-omics perspective on HBV-related HCC development.

METHODOLOGY

Data Collection and Preprocessing

The data collection process focused on gathering RNA-seq datasets from public repositories such as NCBI GEO and TCGA. These datasets represented three categories: healthy individuals, HBV-infected patients, and HBV-HCC patients, and included bulk RNA-seq data from liver and blood samples. For datasets with available raw counts, reads were aligned to the reference genome, and gene expression levels were quantified using featureCounts.

However, the third dataset lacked raw count files, presenting a unique challenge that required additional computational steps to extract usable gene expression data. To address this, Kallisto was used for pseudo-alignment, allowing us to quantify transcript abundances without a traditional alignment step. Reads from the third dataset were mapped against the human cDNA reference transcriptome, generating transcript-level abundance estimates. This approach, while computationally efficient, required further processing to convert transcript-level data into gene-level counts suitable for differential expression analysis. Once these gene-level counts were generated, they served as the foundation for downstream analyses, ensuring compatibility with the other datasets and enabling consistent interpretation across all sample categories.

Differential Gene Expression (DGE) Analysis

To identify differentially expressed genes (DEGs), DGE analysis was performed separately for bulk and single-cell RNA-seq data. Bulk RNA-seq data was analyzed using DESeq2, focusing on differential expression between the three key groups: healthy, HBV-infected, and HBV-HCC individuals. For single-cell RNA-seq data, Scanpy was utilized for data normalization, clustering, and DEG analysis. Statistical significance was determined through adjusted p-values using the Benjamini-Hochberg correction, with genes showing significant fold changes and adjusted p-values below 0.05 designated as DEGs.

Functional Enrichment Analysis

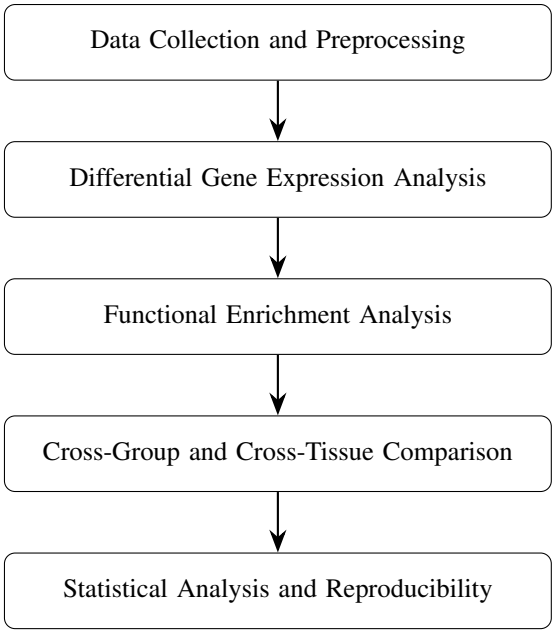
To interpret the biological significance of DEGs, functional enrichment analysis was performed. Using DAVID and ClusterProfiler, Gene Ontology (GO) enrichment analysis was conducted to identify overrepresented GO terms among DEGs. Additionally, pathway enrichment analysis through KEGG and Reactome pinpointed significant biological pathways affected by HBV progression to HCC. Visualization of these results was carried out with enrichment plots, bar charts, and network diagrams, illustrating relationships between genes and pathways for deeper biological insight.

Cross-Group and Cross-Tissue Comparison

Cross-group and cross-tissue comparisons were conducted to examine gene expression relationships across different sample types and conditions. Principal Component Analysis (PCA) and hierarchical clustering were applied to visualize sample similarities and differences, while heatmaps and volcano plots highlighted DEGs specific to HBV progression stages. Correlation analysis between liver and blood samples further identified cross-tissue biomarkers by examining gene expression patterns shared across tissues.

Statistical Analysis and Reproducibility

All statistical analyses were conducted using R and Python, utilizing packages such as DESeq2, edgeR, Scanpy, ClusterProfiler, and ggplot2 for robust statistical analysis and data visualization. The study workflow and associated code were documented thoroughly, ensuring transparency and reproducibility of results.



ANALYSIS

Analysis of First Dataset

The first dataset, GSE94660 dataset, contains RNA sequencing data from 21 HBV-associated hepatocellular carcinoma (HBV-HCC) patients, each with paired samples of tumor and adjacent non-neoplastic liver tissues. The study, conducted by researchers at the Icahn School of Medicine, focuses on exploring how HBV integration and other genomic changes may contribute to tumor recurrence risk based on the level of liver fibrosis. Using a robust pipeline developed to detect HBV integration sites, RNA from surgical specimens was sequenced on the Illumina HiSeq 2500 platform, with poly(A)-selected RNA used for library preparation. RNA-seq counts were already available on GEO and was used for differential gene expression analysis using DESeq2 library on R.

1. **Key Findings**

Differentially Expressed Genes (DEGs). The analysis identified a set of differentially expressed genes (DEGs) for each comparison. Genes like CENPF, PRC1, TOP2A, PLVAP, ASPM, and KIFC1 were significantly upregulated in HBV-associated hepatocellular carcinoma (HCC) tumor tissues compared to nearby non-cancerous liver tissues. These genes showed high expression levels (baseMean values from 226 to 1539) and positive log2 fold changes (ranging from 2.96 to 4.83), indicating notable increases in tumor tissue. The very low adjusted p-values (e.g., 3.53e-54 for CENPF and 2.16e-50 for KIFC1) reflect the strong statistical significance of these upregulations, suggesting these genes might play key roles in cancer development.

For downregulated genes, TTC36-AS1, KCNN2, LCAT, PVALB, SYT9, and LOC101927078 were significantly lower in HBV-HCC tumor tissues compared to non-cancerous tissues. These genes had large negative log2 fold changes (from -3.20 to -5.21), indicating a drop in expression in tumor tissues, with highly significant p-values (e.g., 2.73e-58 for TTC36-AS1). This downregulation may imply that these genes are involved in normal liver function and are suppressed during tumor development, possibly marking cellular processes lost in HBV-HCC.

Functional Enrichment. The functional enrichment analysis for up-regulated genes in HBV-HCC tumor tissue highlights significant involvement in processes associated with cell division and replication. Key enriched pathways include mitotic spindle organization, microtubule cytoskeleton organization involved in mitosis, sister chromatid segregation, DNA replication, and mitotic nuclear division, with the highest gene

counts observed in mitotic spindle organization. These processes are essential for cell cycle progression and division, reflecting the active proliferative state of tumor cells. The enrichment in DNA strand elongation and regulation of the G2/M phase transition further supports a heightened focus on cell cycle regulation, a characteristic often associated with tumor growth and cancer progression.

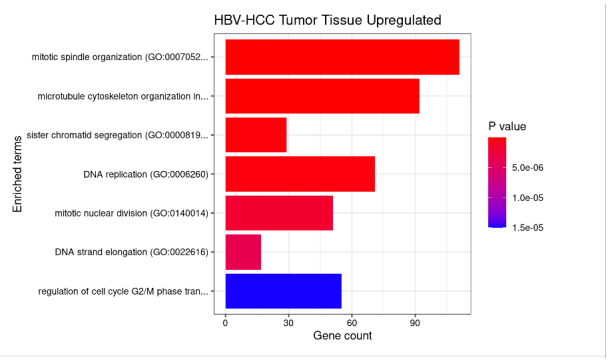


Figure 1. Enriched Pathway for Up-regulated Genes in HBV-HCC vs. HBV

In contrast, the downregulated genes in HBV-HCC tumor tissue are primarily associated with metabolic and immune-related functions. Notably enriched pathways include cellular amino acid catabolic processes, fatty acid catabolic processes, fatty acid beta-oxidation, and fatty acid oxidation, indicating a suppression of metabolic pathways critical for energy production and maintenance of normal liver function. Additionally, there is downregulation in processes related to immune regulation, such as complement activation and immune effector process regulation. This suppression may reflect an impaired immune response and altered metabolic state in HBV-HCC tumor tissues, potentially contributing to the tumor's evasion of immune surveillance and metabolic reprogramming associated with cancer progression.

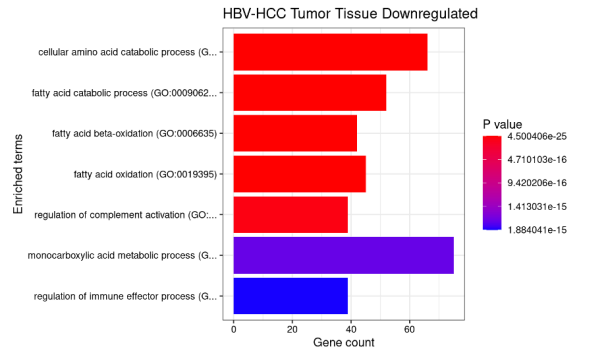


Figure 2. Enriched Pathway for Down-regulated Genes in HBV-HCC vs. HBV

2. **Pathways Identified**

Our study corroborate the findings present in the literature.

Cell Cycle and Mitosis-Related Pathways: The upregulation of pathways like mitotic spindle organization, sister chromatid separation, and DNA replication is common in cancers, including HBV-HCC. Enhanced cell cycle activity and mitotic errors, often triggered by HBV integration, lead to genomic instability and promote tumor growth through abnormal cell division, aligning with previous findings on cancer progression.

Metabolic Reprogramming: The downregulation of fatty acid oxidation and amino acid breakdown is consistent with metabolic reprogramming seen in HBV-HCC. Tumor cells shift from energy-efficient pathways to glycolysis (the Warburg effect) to meet increased energy and biosynthetic demands, aiding growth in low-oxygen environments and often linked to worse outcomes.

Immune Evasion: Suppressed immune pathways, including immune effector processes and complement activation, help HBV-HCC tumors escape immune detection. Chronic HBV infection causes immune exhaustion, which tumors exploit by further reducing immune activation, allowing them to evade immune attacks—especially through blocking the complement pathway, a strategy seen in several cancers.

Analysis of Second Dataset

The second dataset, released in April 2023, provides a detailed analysis of liver samples to examine inflammation across different phases of chronic hepatitis B (CHB). CHB is classified into four phases—immunotolerant (IT), immune-active (IA), inactive carrier (IC), and HBeAg-negative hepatitis (ENEG)—based on HBV DNA, HBeAg status, and ALT levels. This dataset offers a comparison of liver biopsies from CHB patients and healthy controls, using multiplex immunofluorescence and RNA sequencing. In total, 37 samples were analyzed for immune composition, and 78 were sequenced for gene expression profiling on the Illumina HiSeq 4000 platform. For our analysis, we focused on RNA-seq data from the IA phase and healthy controls to study differentially expressed genes between HBV and healthy liver tissues.

1. **Key Findings**

Differentially Expressed Genes (DEGs). In the analysis of HBV liver tissue compared to healthy liver tissue, several genes are significantly upregulated, including KMT2E-AS1, CHRNA2, HMOX1, LOC105369161,

CELF3, and HAPSTR1. These genes show high expression levels, with KMT2E-AS1 having the highest fold change (2.18) and very low p-values, indicating strong significance. Many of these genes may play roles in immune response and metabolism, helping the liver adapt to chronic HBV infection.

On the other hand, several genes are significantly downregulated, such as PRPF8, CREBBP, SNRNP200, ASMTL, LOC102724560, and HUWE1. These genes have moderate to high baseline expression but show notable decreases, with strong statistical significance. This downregulation likely affects genes involved in transcription regulation and splicing, which may impact protein synthesis and cell maintenance in HBV-infected liver tissue.

Functional Enrichment. In the immune-active (IA) phase of HBV, upregulated genes are linked to processes like fibroblast apoptosis regulation and tissue homeostasis. Key enriched pathways include protein neddylation and control of fibroblast cell death, suggesting stronger regulation of cell survival in liver tissue. Other upregulated processes, such as anion homeostasis and axis specification, indicate cellular adaptations to inflammation. Overall, these pathways reflect an immune-driven liver environment with heightened regulation of cell death and immune responses.

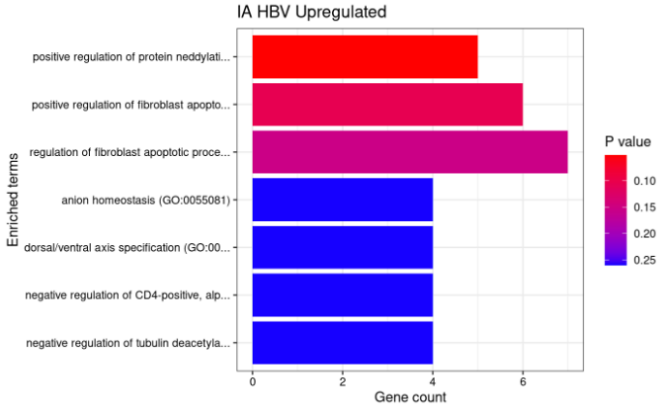


Figure 3. Enriched Pathway for Up-regulated Genes in HBV IA vs. HC

In the IA phase of HBV, downregulated genes are associated with basic cellular functions like protein synthesis, ncRNA processing, and ribosome formation. Pathways involved in gene expression and protein translation are suppressed, indicating a decrease in cellular maintenance activities. Processes like rRNA processing and protein targeting are also reduced. This downregulation suggests a shift from growth and maintenance to immune response and inflammation, as resources are directed away from basic functions to address HBV infection

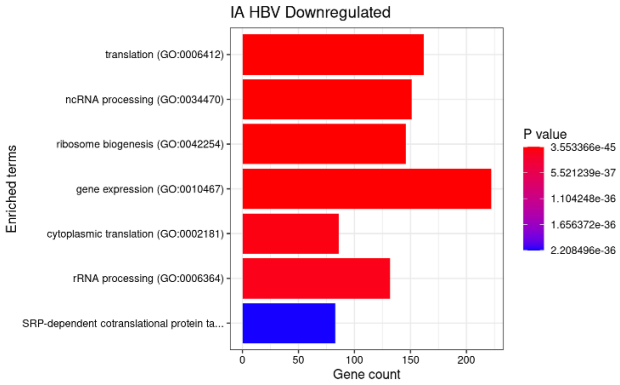


Figure 4. Enriched Pathway for Down-regulated Genes in HBV IA vs. HC

2. **Pathways Identified**

In this functional enrichment analysis for the immune-active (IA) phase of chronic HBV for DEG, we observe both familiar and potentially novel insights.

Upregulated pathways mainly involve regulation of apoptosis, especially in fibroblasts, and protein neddylation. These pathways are linked to immune activation and tissue remodeling in HBV infection, aligning with the immune-driven cell death and fibrosis observed in chronic HBV. Such processes help control liver damage and immune response by regulating fibroblast activity in the inflammatory environment.

Downregulated pathways focus on translation, ncRNA processing, ribosome formation, and protein synthesis. This supports findings that liver cells reduce metabolic activities during chronic HBV infection to conserve energy under viral stress. This shift aligns with decreased biosynthesis in chronic liver disease.

The specific upregulation of fibroblast apoptosis and protein neddylation suggests mechanisms that are not well-studied in HBV and may provide new insights. These pathways could be explored further as potential therapeutic targets in HBV progression.

Analysis of Third Dataset: HBV-HCC (PBMC)

The third dataset, focused on HBV-related hepatocellular carcinoma (HBV-HCC) in peripheral blood mononuclear cells (PBMCs), required extensive preprocessing using Kallisto for transcript quantification, as raw count files were not initially available for analysis. Kallisto was used to efficiently process and quantify transcript-level abundance from the raw RNA sequencing (RNA-seq) data. Once the transcript abundance data were generated, differential gene expression (DGE) analysis was conducted using the DESeq2 package. This analysis enabled the identification of genes with significant expression differences across different conditions. Comparisons were made between key conditions: Healthy vs. HBV, HBV vs. HBV-HCC, and HBV-HCC vs. Healthy, in order to identify molecular signatures , alterations and potential biomarkers of HBV-induced hepatocellular carcinoma

1. Key Findings

Differentially Expressed Genes (DEGs). The analysis identified a set of DEGs for each comparison. In the HBV vs. Healthy comparison, a moderate number of genes showed significant differential expression, with more upregulated genes than downregulated ones.

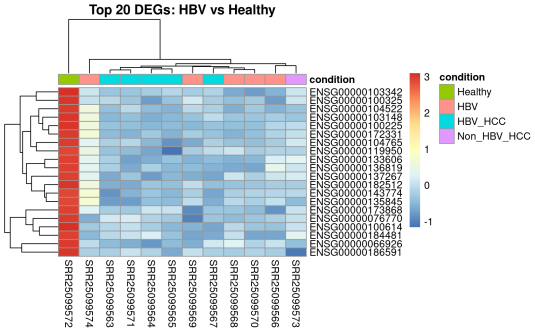


Figure 5. DEG For HBV Vs Healthy

The HBV-HCC vs. HBV comparison revealed genes associated with cancer progression, showing distinct upregulation and downregulation patterns reflective of the transition from infection to carcinoma.

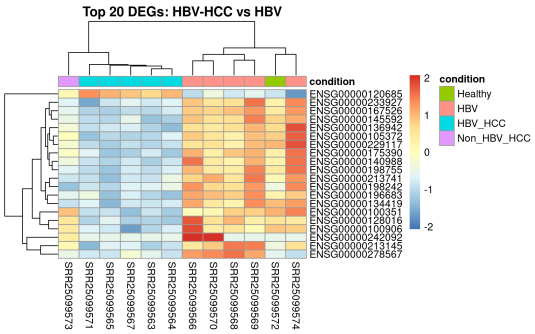


Figure 6. DEG For HBV-HCC vs HBV

In the HBV-HCC vs. Healthy comparison, several genes known to be involved in immune response and inflammation were significantly expressed, suggesting an immune-mediated response in HBV progression to HCC.

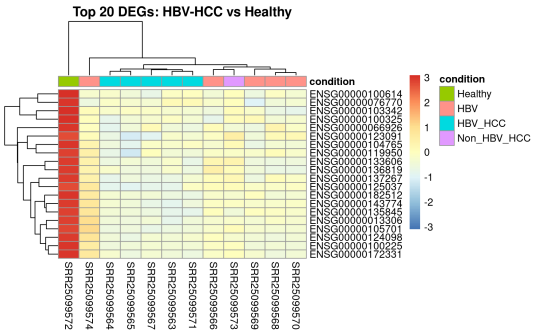


Figure 7. DEG For HBV-HCC Vs Healthy

Functional Enrichment. **GO Biological Processes:** GO enrichment analysis indicated enrichment in processes such as proteasome-mediated protein catabolism, autophagy, and immune response pathways, specifically in comparisons involving HBV-HCC samples.

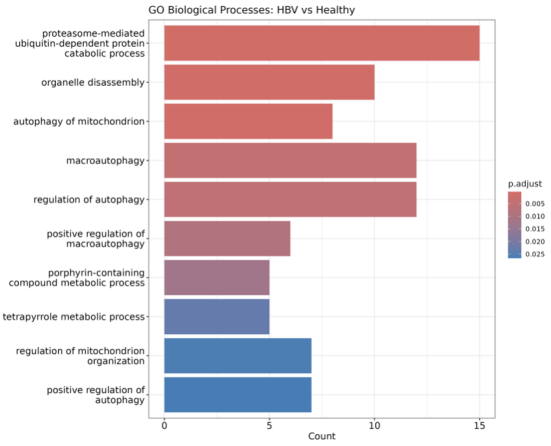


Figure 8. Enriched Pathway HBV Vs Healthy

KEGG Pathway Analysis: KEGG pathways significantly enriched in HBV-HCC included pathways related to mitophagy and Toxoplasmosis, reflecting cellular degradation processes and immune response in HBV-HCC samples.

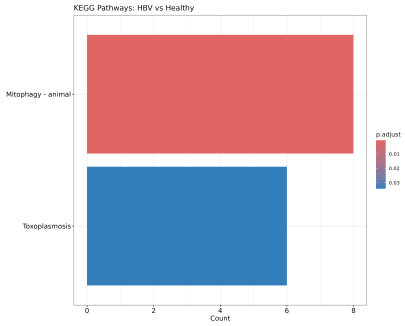


Figure 9. KEGG Pathway HBV Vs Healthy

Reactome Pathway Analysis: Reactome pathways were enriched in ubiquitination and NF-kappa-B signaling, pathways known to play roles in viral infection and immune regulation.

2. Comparison with Reference Study

In comparing our third dataset analysis with the findings from the study by Zhou et al., we observed both common and unique pathway enrichments. Zhou et al. emphasize immune-related pathways, such as immune activation and cell death regulation. In contrast, our analysis revealed additional pathways, particularly

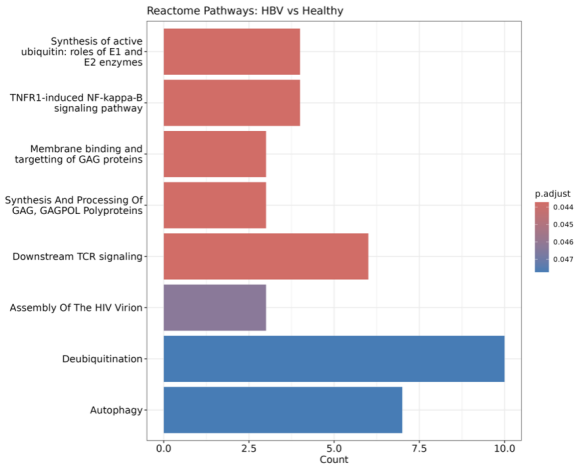


Figure 10. Reactome Pathway HBV vs Healthy

related to autophagy and mitochondrial processes, including the *autophagy of mitochondrion* and *regulation of mitochondrion organization*. These findings suggest a critical role for mitochondrial health and autophagy in HBV-related hepatocellular carcinoma (HBV-HCC).

3. **Novel Pathways Identified**

   Our study uniquely identified novel pathways, such as the *proteasome-mediated ubiquitin-dependent protein catabolic process* and *tetrapyrrole metabolic processes*. These pathways, associated with cellular stress and metabolic responses, provide new insights into the cellular mechanisms underlying HBV infection. Notably, these findings imply that mitochondrial regulation and protein degradation could be pivotal in understanding HBV pathogenesis and may serve as potential therapeutic targets aimed at restoring cellular balance and mitigating liver damage.

RESULTS

In this study, we aimed to explore the differential gene expression patterns linked to Hepatitis B Virus (HBV) infection and HBV-related hepatocellular carcinoma (HBV-HCC) across liver and blood samples. We analyzed three primary datasets: HBV-Liver, HBV-HCC-Liver, and HBV-Blood (PBMC). Below, we summarize the key findings from each dataset:

HBV-Liver Dataset

In the HBV-Liver dataset, we observed significant upregulation of genes involved in cell cycle regulation, such as *CENPF*, *PRC1*, and *TOP2A*, in HBV-HCC tumor tissues compared to non-cancerous liver tissues. Functional enrichment analysis revealed that these upregulated genes were enriched in pathways associated with mitosis, DNA replication, and microtubule organization. This suggests increased cell proliferation, a typical characteristic of cancer progression.

Conversely, downregulated genes in this dataset were primarily associated with metabolic processes like fatty acid oxidation and amino acid metabolism. This indicates a metabolic shift in liver tissues affected by HBV infection and HCC, as normal liver metabolic activity becomes altered during disease progression.

HBV-HCC-Liver Dataset

In the HBV-HCC-Liver dataset, a distinct pattern of gene expression was observed in HBV-HCC samples. Specifically, we found upregulation of immune-modulatory genes such as *KMT2E-AS1* and *HMOX1*, which were enriched in pathways regulating cellular stress responses and immune functions. This suggests an immune-driven alteration in liver tissues under chronic HBV infection.

Downregulated genes in this dataset were linked to pathways involved in ncRNA processing, protein synthesis, and ribosomal activities. The suppression of these pathways indicates that liver cells are redirecting their resources from normal cellular maintenance to focus on immune and stress response, particularly in the immune-active phase of HBV infection.

HBV-Blood (PBMC) Dataset

For the HBV-Blood (PBMC) dataset, which analyzed blood samples from HBV-related HCC patients, we observed upregulation of genes associated with immune responses, such as those involved in *NF-kappa-B* signaling, proteasome-mediated protein catabolism, and autophagy-related pathways. Enrichment of pathways related to autophagy and mitochondrial regulation reflects immune and cellular stress responses in the blood, indicative of HBV progression to HCC. These findings suggest that systemic immune responses, which mirror gene expression changes in liver tissues, are detectable in blood samples.

CROSS-DATASET INSIGHTS

Comparing the three datasets revealed common pathways and gene patterns that highlight consistent immune activation, metabolic reprogramming, and cell proliferation across both liver and blood samples. The upregulation of cell cycle-related genes and pathways in both liver datasets reinforces the notion that HBV-HCC is characterized by increased cellular proliferation. Additionally, the downregulation of metabolic pathways in liver tissues and activation of immune response pathways in blood samples indicate that HBV infection has a systemic impact, where metabolic changes in the liver coincide with detectable immune responses in blood.

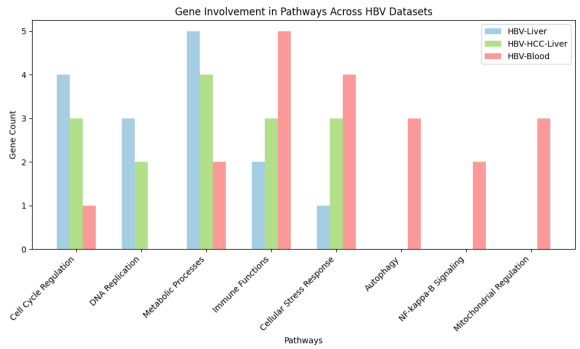


Figure 11. Common Pathways

VALIDATION OF EXPECTED OUTCOMES

This study successfully achieved several expected outcomes:

1. **Identification of Differentially Expressed Genes in HBV and HBV-HCC**: Genes such as *CENPF*, *TOP2A*, and *KMT2E-AS1* were found to be differentially expressed in a way that distinguishes HBV from HBV-HCC, particularly in pathways related to cell cycle regulation and immune responses.

2. **Potential for Blood-Based Biomarkers**: The consistent detection of immune response pathways and autophagy processes in both blood and liver samples supports the potential for blood-based biomarkers in HBV-HCC. This cross-tissue mirroring of expression profiles in immune-related genes provides a basis for developing blood-based RNA-seq assays for non-invasive disease detection.

3. **Correlation of Cross-Tissue Gene Expression**: The shared enrichment of pathways such as *NF-kappa-B* signaling and autophagy between liver and blood samples suggests that gene expression changes in HBV-HCC liver tissues are reflected in blood, validating our hypothesis of cross-tissue biomarker potential.

CONCLUSION

In conclusion, this study highlights significant overlaps in gene expression patterns between liver and blood samples, supporting the hypothesis that certain HBV-HCC-related biomarkers can be detected in blood. This finding corroborates the concept of non-invasive diagnostics for HBV-HCC, offering potential for early disease detection and monitoring.

The cross-tissue similarity in immune-related pathways and metabolic reprogramming underscores the translational potential of RNA-seq-based screening tools for HBV-related HCC. These findings not only validate blood as a source for biomarker identification but also set a foundation for future studies that could integrate multi-omics data to further refine and validate these biomarkers. Our results offer a comprehensive framework for developing blood-based diagnostic assays, providing a promising step towards non-invasive HBV-HCC screening and management.

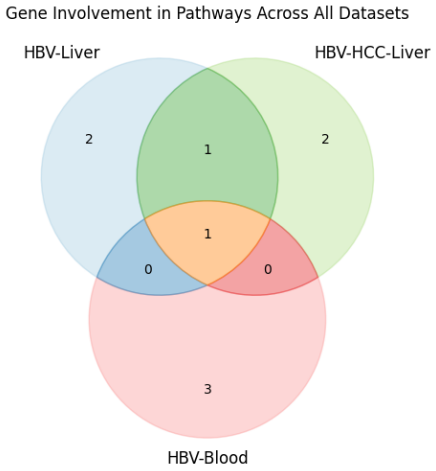


Figure 12. Common Pathways

LIMITATIONS

While this study provides meaningful insights into HBV-related HCC biomarkers, several limitations should be acknowledged. One primary limitation is the potential for batch effects, as this study integrates datasets from multiple sources, each with different experimental protocols and preprocessing steps. These batch effects, stemming from variations in sample preparation, sequencing platforms, and data processing, may introduce unwanted technical noise that could obscure or distort biological signals. Although statistical corrections for batch effects were applied, they may not fully eliminate these artifacts, particularly when combining bulk and single-cell RNA-seq data. Future research could benefit from using larger, homogeneously processed datasets or applying advanced batch effect correction methods to enhance the accuracy and generalizability of the results.

APPLICATIONS OF THIS PROJECT

The findings from this study on cross-tissue biomarkers for Hepatitis B Virus-related Hepatocellular Carcinoma (HBV-HCC) open several promising avenues for practical applications in healthcare and research. Below, we outline the key applications:

Non-Invasive Diagnostic Tools

The identification of gene expression changes in blood that mirror those in HBV-HCC-affected liver tissue suggests the possibility of developing blood-based diagnostic tests. Such tests could serve as a non-invasive alternative to liver biopsies, allowing for earlier and more accessible detection of liver cancer in at-risk populations. With further validation, these RNA-based blood tests could be integrated into routine screenings, especially in areas with high HBV prevalence.

Early Cancer Detection and Monitoring

This project's results could significantly impact early cancer detection and monitoring. Detecting HBV-HCC at an earlier stage through blood biomarkers could improve patient outcomes by enabling timely interventions. Additionally, these biomarkers could be used to monitor disease progression or treatment response, allowing healthcare providers to adjust therapies based on real-time molecular insights.

Personalized Medicine and Targeted Therapies

The identified biomarkers, which highlight immune and metabolic pathway alterations, may also aid in developing personalized treatment approaches for HBV-HCC patients. By understanding a patient's specific gene expression profile, treatments could be tailored to target these pathways, potentially increasing therapeutic efficacy. Moreover, the biomarkers could serve as targets for drug development, enabling the creation of therapies that directly interact with the pathways implicated in HBV-HCC.

Research into HBV-HCC Pathogenesis

The shared gene expression patterns across blood and liver tissues provide valuable insights into the mechanisms underlying HBV-HCC progression. Researchers can use this knowledge to investigate the biological processes involved in HBV infection and liver carcinogenesis further, which may reveal novel intervention points or therapeutic targets. Additionally, the cross-tissue approach could inspire similar studies in other diseases, fostering a broader understanding of systemic disease biomarkers.

Public Health Screening Programs

In regions with high HBV incidence, this study's findings could support public health initiatives by introducing blood-based screening programs aimed at early detection of HBV-HCC. Such programs could reduce the need for invasive procedures, encourage higher participation rates, and ultimately help reduce liver cancer morbidity and mortality. These screening programs could be particularly valuable in resource-limited settings where access to specialized medical procedures, like liver biopsies, is limited.

In summary, the applications of this project extend from clinical diagnostics and personalized medicine to public health and research. By developing a framework for non-invasive, RNA-based biomarker detection, this study lays the groundwork for innovative tools that could improve early detection, treatment, and understanding of HBV-HCC.

FUTURE DIRECTIONS

Building on the findings of this study, several future directions could extend and validate the identified biomarkers for HBV-related HCC. First, the application of these biomarkers to larger, independent patient cohorts would help determine their robustness and potential as non-invasive diagnostic tools. Conducting such studies across diverse populations could also help establish biomarker consistency and efficacy across different genetic backgrounds and environmental factors. Additionally, integrating other omics layers, such as proteomics and metabolomics, could offer a more comprehensive understanding of HBV-HCC progression by validating gene expression changes at the protein level and elucidating metabolic shifts associated with disease stages. Finally, longitudinal studies tracking biomarker profiles over time may provide insight into disease progression dynamics, potentially revealing key intervention points and improving early detection efforts.

AVAILABILITY OF DATASETS

The datasets used in this study are publicly available and can be accessed through the following links:

- **Dataset 1 (GSE94660):** Accessible on the NCBI Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94660
- **Dataset 2 (GSE230397):** Accessible on the NCBI Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE230397
- **Dataset 3 (GSE236281):** Accessible on the NCBI Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE236281

AVAILABILITY OF CODE

The code for this project is publicly available on GitHub. You can access the repository here. The repository includes all the necessary files and documentation for reproducing the results.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our mentors and advisors at the Koita Centre for Digital Health, IIT Bombay, for their invaluable guidance and support throughout this project. Special thanks to Professor Saket Choudhary for his expertise, encouragement, and insightful feedback that greatly improved the quality of our work.

We are also grateful to the teams that provided the publicly available datasets, including the NCBI Gene Expression Omnibus (GEO) database, which made this research possible. We acknowledge the use of open-source software and tools that were instrumental in our data analysis, including DESeq2, Scanpy, and Kallisto.

REFERENCES

1. Audi, I., Inoue, T., Tanaka, Y. (2020). Novel Biomarkers of Hepatitis B and Hepatocellular Carcinoma: Clinical Significance of HBcrAg and M2BPGi. *International Journal of Molecular Sciences*, 21(3), 949. https://doi.org/10.3390/ijms21030949
2. Hayashi, S., Nagaoka, K., Tanaka, Y. (2021). Blood-Based Biomarkers in Hepatitis B Virus-Related Hepatocellular Carcinoma, Including the Viral Genome and Glycosylated Proteins. *International Journal of Molecular Sciences*, 22(20), 11051. https://doi.org/10.3390/ijms222011051
3. Ho, D. W., Tsui, Y. M., Chan, L. K., Sze, K. M., Zhang, X., Cheu, J. W., ... Ng, I. O. (2021). Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma. *Nature Communications*, 12(1), 3684. https://doi.org/10.1038/s41467-021-24010-1
4. Jiang, Y., Han, Q., Zhao, H., Zhang, J. (2021). The Mechanisms of HBV-Induced Hepatocellular Carcinoma. *Journal of Hepatocellular Carcinoma*, 8, 435–450. https://doi.org/10.2147/JHC.S307962
5. Liu, Y., Al-Adra, D. P., Lan, R., Jung, G., Li, H., Yeh, M. M., Liu, Y. Z. (2022). RNA sequencing analysis of hepatocellular carcinoma identified oxidative phosphorylation as a major pathologic feature. *Hepatology Communications*, 6(8), 2170–2181. https://doi.org/10.1002/hep4.1945
6. Paslaru, L., Bindea, G., Nastase, A., Sorop, A., Zimbru, C., Herlea, V., ... Popescu, I. (2022). Comparative RNA-Sequencing Analysis Reveals High Complexity and Heterogeneity of Transcriptomic and Immune Profiles in Hepatocellular Carcinoma Tumors of Viral (HBV, HCV) and Non-Viral Etiology. *Medicina (Kaunas, Lithuania)*, 58(12), 1803. https://doi.org/10.3390/medicina58121803
7. Rizzo, G. E. M., Cabibbo, G., Craxì, A. (2022). Hepatitis B Virus-Associated Hepatocellular Carcinoma. *Viruses*, 14(5), 986. https://doi.org/10.3390/v14050986
8. Yeh, S. H., Li, C. L., Lin, Y. Y., Ho, M. C., Wang, Y. C., Tseng, S. T., Chen, P. J. (2023). Hepatitis B Virus DNA Integration Drives Carcinogenesis and Provides a New Biomarker for HBV-related HCC. *Cellular and Molecular Gastroenterology and Hepatology*, 15(4), 921–929. https://doi.org/10.1016/j.jcmgh.2023.01.001
9. Zheng, Q., Sun, Q., Yao, H., Shi, R., Wang, C., Ma, Z., ... Xia, H. (2024). Single-cell landscape identifies the immunophenotypes and microenvironments of HBV-positive and HBV-negative liver cancer. *Hepatology Communications*, 8(2), e0364. https://doi.org/10.1097/HC9.0000000000000364

SUPPLEMENTARY

Dataset-3 : HBV-HCC (PBMC)

Top GO Terms Comparison (Dataset-3 vs Zhou et al.)

Top GO Terms (Dataset-3)	Top GO Terms (Paper)
Proteasome-mediated ubiquitin-dependent protein catabolic process	Immune response
Organelle disassembly	Regulation of the immune system
Autophagy of mitochondrion (mitophagy)	Regulation of cell death
Macroautophagy	Haematopoiesis
Regulation of autophagy	Defense response
Positive regulation of macroautophagy	Regulation of kinase activity
Porphyrin-containing compound metabolic process	Immune cell activation
Tetrapyrrole metabolic process	
Regulation of mitochondrion organization	
Positive regulation of autophagy	

Table S1. Comparison of GO terms between my analysis and the findings from the paper.

Highlighting Common Themes

Autophagy and Mitochondrial Regulation:

- **Our Findings:** Pathways such as *autophagy of mitochondrion*, *macroautophagy*, and *regulation of mitochondrion organization* were enriched, indicating a key role of autophagy and mitochondrial processes in HBV infection.
- **Paper's Findings:** Zhou et al. focus on immune-related pathways like immune activation and cell death regulation, potentially influenced by mitochondrial health and autophagy.
- **Connection:** Disruptions in mitochondrial and autophagy processes may impact immune function, potentially contributing to the immune dysregulation observed in HBV-ACLF.

Proposing a Mechanistic Link

Dysregulated autophagy and mitochondrial processes could trigger inflammatory signaling, aligning with the immune activation pathways noted by Zhou et al. We also found enrichment in the *proteasome-mediated ubiquitin-dependent protein catabolic process*, essential for immune regulation, which may impact the immune disturbances seen in HBV-ACLF.

Kallisto Process for Transcript Quantification

In this project, we used Kallisto, a tool for efficient transcript quantification from RNA sequencing (RNA-seq) data, to process one of the datasets that lacked raw count files. Kallisto employs a method known as pseudo-alignment, which is faster than traditional alignment while maintaining high accuracy in transcript abundance estimation. Below is a description of the steps we followed using Kallisto for transcript quantification in our analysis:

1. **Indexing the Transcriptome**: First, we downloaded the reference transcriptome for the human genome, which was necessary for the pseudo-alignment process. We used Kallisto to generate an index from this reference transcriptome. The index serves as the basis for matching RNA-seq reads to known transcript sequences without aligning them to the genome.

2. **Pseudo-Alignment of RNA-seq Reads**: We then proceeded to process the RNA-seq reads from the third dataset using Kallisto's pseudo-alignment step. Unlike traditional alignment methods, Kallisto does not map reads to specific genome positions. Instead, it assigns the reads to the most likely transcripts from the reference transcriptome based on sequence similarity. This approach was computationally efficient and allowed us to process large volumes of data in a short time.

3. **Transcript Abundance Estimation**: Once the reads were pseudo-aligned, Kallisto estimated the abundance of each transcript using its built-in Expectation-Maximization (EM) algorithm. This algorithm iteratively calculates the probability of each read originating from specific transcripts, producing a final estimate of transcript abundances. The abundance values were provided in Transcripts Per Million (TPM) and raw counts, which we used for further downstream analysis.

4. **Transcript-to-Gene Aggregation**: After obtaining transcript-level abundances, we aggregated these values to the gene level, as this is the standard format for differential gene expression analysis. This step ensured that the transcript-level quantifications were suitable for integration with other datasets and compatible with tools such as DESeq2 for differential expression analysis.

5. **Output Files and Quality Control**: Kallisto produced several output files, including the transcript abundance estimates and a summary of the pseudo-alignment process. We reviewed the output to ensure the quality of the data and assessed the consistency of the abundance estimates. These files formed the basis for the differential gene expression (DGE) analysis and functional enrichment studies conducted later in the project.

Why Kallisto? Kallisto was an essential tool for our project because of its speed and accuracy. Given the large-scale data processing required for the RNA-seq datasets in our project, Kallisto provided a computationally efficient way to quantify transcript abundances, enabling us to process the data without significant delays. Additionally, Kallisto's pseudo-alignment approach minimizes biases introduced by traditional alignment tools, ensuring a more reliable estimate of transcript expression levels.

Considerations and Limitations: While Kallisto is highly effective for transcript quantification, we were aware of its limitations. For example, Kallisto may not capture novel transcripts that are not included in the reference transcriptome, which could lead to some underrepresentation of unannotated genes. Additionally, Kallisto's pseudo-alignment method assumes that the majority of reads correspond to known transcript sequences, which could be problematic in cases of highly polymorphic samples or extensive alternative splicing. However, for the scope of our study, these considerations were minimal, as the dataset provided robust transcript annotations.

This approach allowed us to efficiently process and analyze RNA-seq data, and the results formed the foundation for the downstream differential expression and functional enrichment analyses that were central to the findings of this project.