Integrative Analysis of Gut Microbiota and Host Transcriptomics in Hepatocellular Carcinoma

Jagoda Trzeciak

June 2025

Abstract

Hepatocellular carcinoma is a major global health burden, often diagnosed at advanced stages due to the lack of early clinical symptoms and reliable biomarkers. This study analyzes gut microbiome and liver tissue transcriptomic data from patients with HCC to explore potential molecular and microbial signatures of the disease. Microbiome analysis showed high variability across samples, while transcriptomic analysis identified numerous differentially expressed genes (DEGs) between tumor and non-tumor tissues. Machine learning highlighted several genes, including *CCT3* and *FLAD1*, as promising diagnostic candidates. Although a preliminary attempt was made to correlate microbial taxa with gene expression, the lack of sample-level metadata prevented direct integration.

1 Introduction

Hepatocellular carcinoma (HCC) is the most common form of primary liver cancer and a significant contributor to global cancer mortality. In 2020, it accounted for approximately 906,000 new cases and 830,000 deaths, ranking HCC as the sixth most frequently diagnosed type of cancer and the leading cause of cancer-related deaths worldwide [1]. The insidious nature of HCC, often progressing without symptoms in early stages, leads to late diagnosis with limited treatment options, thereby contributing to its poor clinical outcomes [2].

Hepatocellular carcinoma typically arises from chronic liver injury, which initiates a cascade of inflammation, fibrosis, and ultimately cirrhosis- conditions that markedly increase the risk of malignant transformation. This liver damage can result from a variety of etiological factors, including excessive alcohol consumption, metabolic syndrome-related disorders, such as non-alcoholic fatty liver disease (NAFLD) and its progressive form non-alcoholic steatohepatitis (NASH), and inherited metabolic conditions like hemochromatosis and alpha-1 antitrypsin deficiency. Among viral causes, chronic infection with hepatitis B virus (HBV) is particularly significant. It can lead to HCC both through cirrhosis-dependent and direct oncogenic mechanisms. HBV is estimated to account for approximately 50–55% of HCC cases globally, especially in endemic regions such as sub-Saharan Africa and East Asia [2, 3].

This project investigates the relationship between gut microbiome composition and host gene expression in HCC, drawing on data from the study by Huang et al. published in *Genome Medicine* [4].

2 Materials and Methods

2.1 Data Sources

The transcriptomic data were retrieved from the Gene Expression Omnibus under accession number GSE138485. This dataset comprises RNA-seq expression profiles from 32 patients diagnosed with hepatocellular carcinoma, each contributing paired samples of tumor and adjacent non-tumor liver tissue. For the purposes of this analysis, a pre-processed gene-level count matrix provided by NCBI was used, enabling direct downstream analysis without the need for raw read alignment or quantification.

Complementary microbiome data were obtained from the European Nucleotide Archive under accession number PRJEB8708. These data include raw 16S rRNA sequencing reads from fecal samples and were used to investigate gut microbial composition and diversity. While both datasets originate from the same study, no metadata were provided that would enable reliable linkage between the 16S rRNA and RNA-seq samples. As a result, it was not possible to determine whether the fecal samples correspond to cancer patients or healthy controls, nor to match them to specific individuals in the transcriptomic dataset.

2.2 Microbiome Analysis

The microbiome data were processed and analyzed using the `DADA2` and `phyloseq` packages in R. Raw paired-end sequencing reads were initially inspected to evaluate sequence quality across base positions. As expected, quality profiles showed a gradual decline toward the ends of the sequences, especially for reverse reads. Based on these profiles, forward and reverse reads were truncated at 270 and 160 nucleotides, respectively, and filtered using the `filterAndTrim()` function. Subsequently, the reads were dereplicated to collapse identical sequences within each sample, thereby reducing redundancy and computational burden. Error rates were modeled independently for forward and reverse reads using the `leanErrors()` function. Denoising was then performed with `dada()`, allowing for the inference of amplicon sequence variants (ASVs) with single-nucleotide resolution. Paired-end reads were merged, and chimeric sequences, frequently introduced during PCR amplification, were identified and removed using a consensus-based approach via `removeBimeraDenovo()`. An ASV table was constructed using the `makeSequenceTable()`. Taxonomic classification of ASVs was performed against the SILVA reference database (version 138.1) using the `assignTaxonomy()` function, enabling the assignment of microbial sequences to hierarchical taxonomic levels.

For downstream analysis, the ASV table and taxonomic assignments were incorporated into a `phyloseq` object to facilitate compositional profiling. Alpha diversity was quantified using the Shannon, Simpson, and Inverse Simpson indices, capturing both the richness and evenness of microbial communities within individual samples. Beta diversity was assessed using Bray-Curtis distance, and differences in microbial composition between samples were

visualized via non-metric multidimensional scaling (NMDS). Taxonomic composition was further explored by aggregating ASVs at the phylum, family, and genus levels using the `tax_glom()` function. Relative abundances were then visualized, providing a comparative overview of microbial community structure across samples.

2.3 Liver Tissue Data Analysis

The analysis of the RNA-seq liver tissue data was performed in R using `DESeq2` package. A dataset containing raw gene-level counts and sample condition metadata was prepared for differential expression analysis, which was performed using `DESeq()` function. Genes were considered as differentially expressed if their adjusted p-value was below 0.05 and their absolute $\log_2$ fold change was at least 0.8.

To explore expression patterns, histograms of $\log_2$ fold changes were generated using `ggplot2`, both for all genes and for the subset of significantly differentially expressed genes (DEGs). A heatmap of the top 40 DEGs was created using the `pheatmap` package. Gene annotations were obtained via `org.Hs.eg.db` and expression values were standardized using base R functions to enable comparison across samples. To evaluate the predictive potential of selected genes, a Random Forest classifier was trained using the `caret` package, and gene importance scores were extracted to identify those most relevant for distinguishing tumor from non-tumor samples.

An exploratory correlation analysis was also conducted to investigate potential associations between microbial taxa and host gene expression. Due to the absence of metadata linking individual microbiome samples to corresponding RNA-seq samples, identifiers were standardized and aligned manually based on sample order. Pearson correlation coefficients were computed between microbial abundances (OTUs) and expression levels of differentially expressed genes. The resulting correlation matrix provides an initial overview of potential microbiome–transcriptome interactions, though interpretations remain limited by the lack of confirmed sample-level matching.

3 Results

3.1 Gut Microbiome Composition and Diversity

Analysis of the 16S rRNA sequencing data revealed distinct microbial communities across fecal samples. After quality filtering, denoising and removing chimeras, a total of 194 amplicon sequence variants were retained across 15 samples. Taxonomic classification assigned these ASVs to a range of bacterial taxa, with the most abundant phyla being *Proteobacteria*. At the genus level, *Prevotella* was found to dominate several samples, consistent with previously reported gut microbial profiles in some liver disease contexts [5, 6].

Alpha diversity analysis indicated that most samples exhibited moderate microbial diversity. The Shannon index showed relatively consistent values across samples, suggesting moderate richness and evenness. The Simpson index was tightly clustered, reflecting its lower sensitivity to rare taxa. In contrast, the Inverse Simpson index displayed the greatest variability, indicating that some samples contained highly diverse and evenly distributed microbial communities (Figure 1).

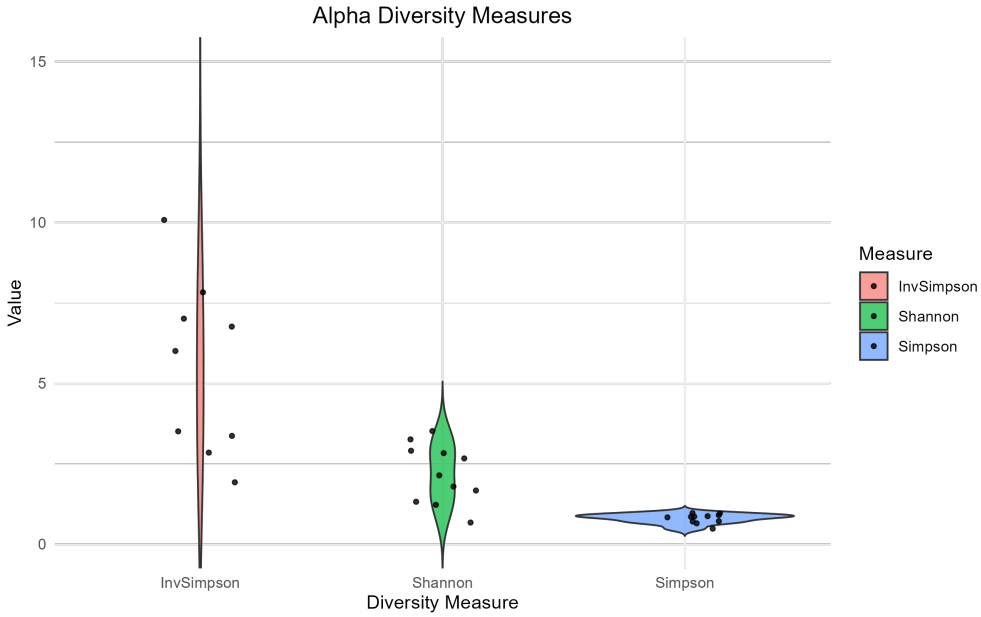

Figure 1: Alpha diversity across samples measured by Shannon, Simpson, and Inverse Simpson indices.

Beta diversity, assessed using Bray–Curtis dissimilarity and visualized through NMDS (Figure 2), revealed a broad spread of samples in the ordination space. This suggests substantial differences in taxonomic composition between samples. However, the absence of distinct clusters indicates that the samples do not group according to any shared characteristics-likely a consequence of missing metadata such as clinical status or other host factors.

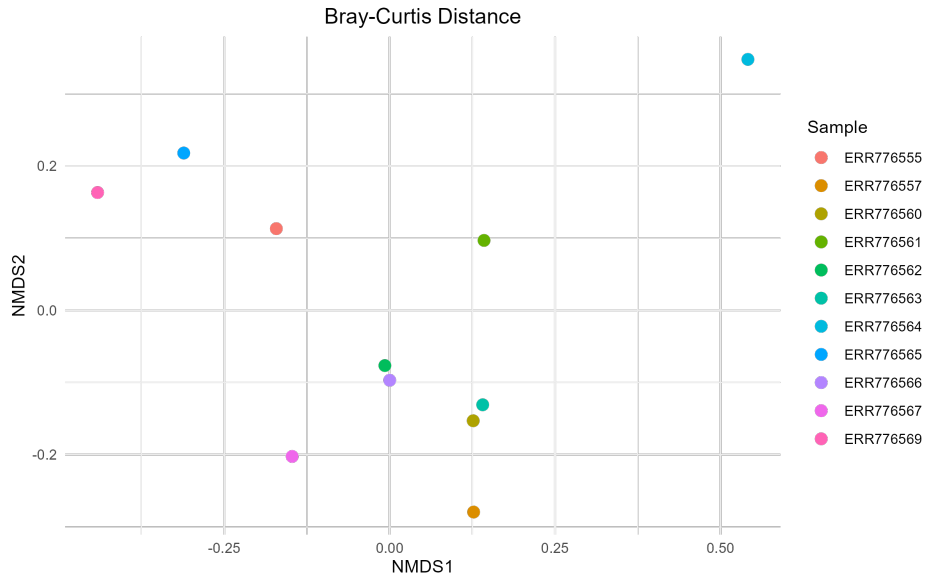

Figure 2: Beta diversity across different samples.

3.2 Differential Gene Expression in HCC

Differential expression analysis revealed widespread transcriptional alternations between tumor and adjacent non-tumor liver tissues (Figure 3). The overall distribution of $\log_2$ fold changes was tightly centered around zero, indicating that the majority of genes exhibited minimal expression differences. In contrast, the distribution of significant DEGs displayed a clear bimodal pattern, reflecting distinct set of genes that were either up- or downregulated in tumor tissue.

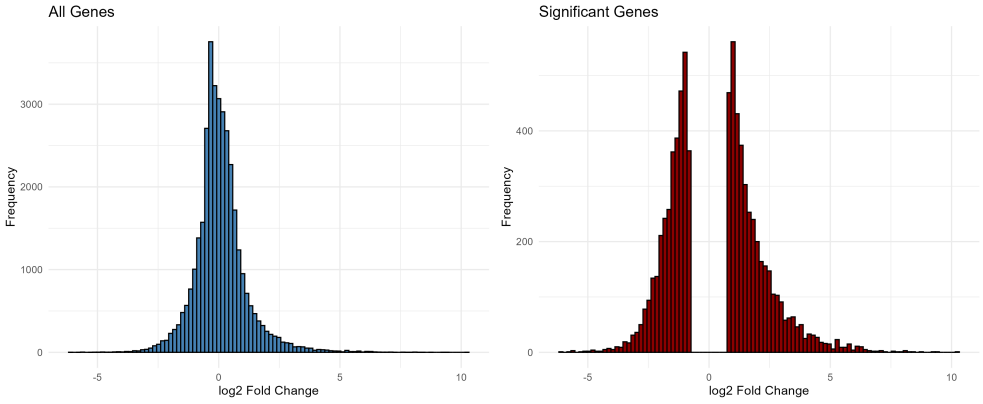

Figure 3: Alpha diversity across samples measured by Shannon, Simpson, and Inverse Simpson indices.

A heatmap of the top 40 most significantly differentially expressed genes (Figure 4) revealed a clear separation between tumor and non-tumor liver tissue samples, reflecting consistent transcriptional differences between these two conditions. Among the upregulated

genes, CCT3, CDC20, and CDC25C showed high expression in tumor tissue and have known roles in cell cycle regulation and tumor progression [7, 8]. In contrast, genes such as NCAM1 and CLEC4G showed higher expression levels in non-tumor tissue and have been implicated in normal hepatic function and tumor suppressive activity [9, 10].

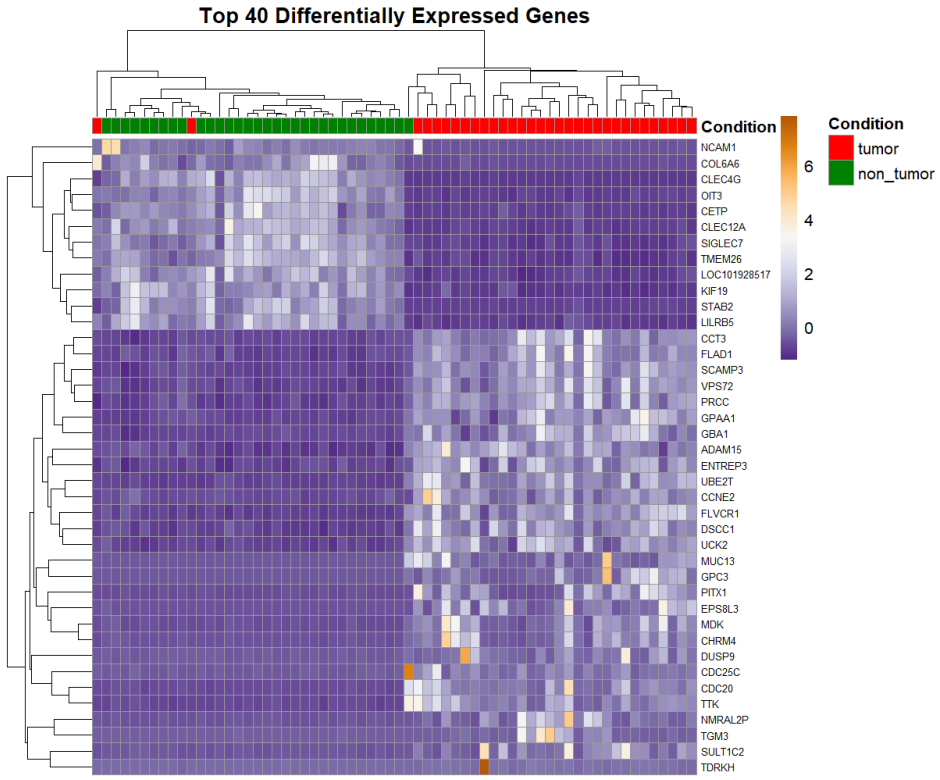

Figure 4: Heatmap of top 40 differentially expressed genes between tumor and non-tumor samples.

3.3 Biomarkers Prediction

The Random Forest classification model revealed several genes with strong discriminatory power between tumor and non-tumor samples (Figure 5). Among these, CCT3 ranked highest in importance, suggesting it may serve as a robust biomarker candidate for hepatocellular carcinoma. Known for its role in protein folding and cell proliferation, CCT3 has been previously implicated in multiple cancer types [7]. Notably, it demonstrated greater predictive power in this dataset than the well-established biomarker GPC3 [11, 12] and some studies also suggest that its diagnostic performance is comparable to that of AFP, the most commonly used biomarker for HCC [13].

Ranked just below CCT3 was FLAD1, a gene involved in mitochondrial metabolism and redox homeostasis. Overexpression of FLAD1 has been proposed as a potential biomarker in several malignancies, including breast cancer [14]. In hepatocellular carcinoma, the consistent upregulation of FLAD1 in tumor tissues may reflect underlying metabolic reprogramming and the heightened proliferative demands characteristic of cancer cells. Given

its functional role and strong predictive value in classification models, FLAD1 emerges as a promising candidate for further investigation as a potential biomarker for HCC.

Interestingly GPC3, widely used HCC biomarker, received a relatively low importance score in the model. This finding may appear unexpected, but closer inspection of the expression data provides a likely explanation. While GPC3 was indeed upregulated in many tumor samples, as observed in the heatmap, it displayed greater variability across the cohort compared to genes like CCT3 and CDC20. Random Forest models prioritize features that offer consistent discriminatory power across all samples, so the heterogeneous expression of GPC3 likely diminished its relative contribution to classification accuracy in this dataset.

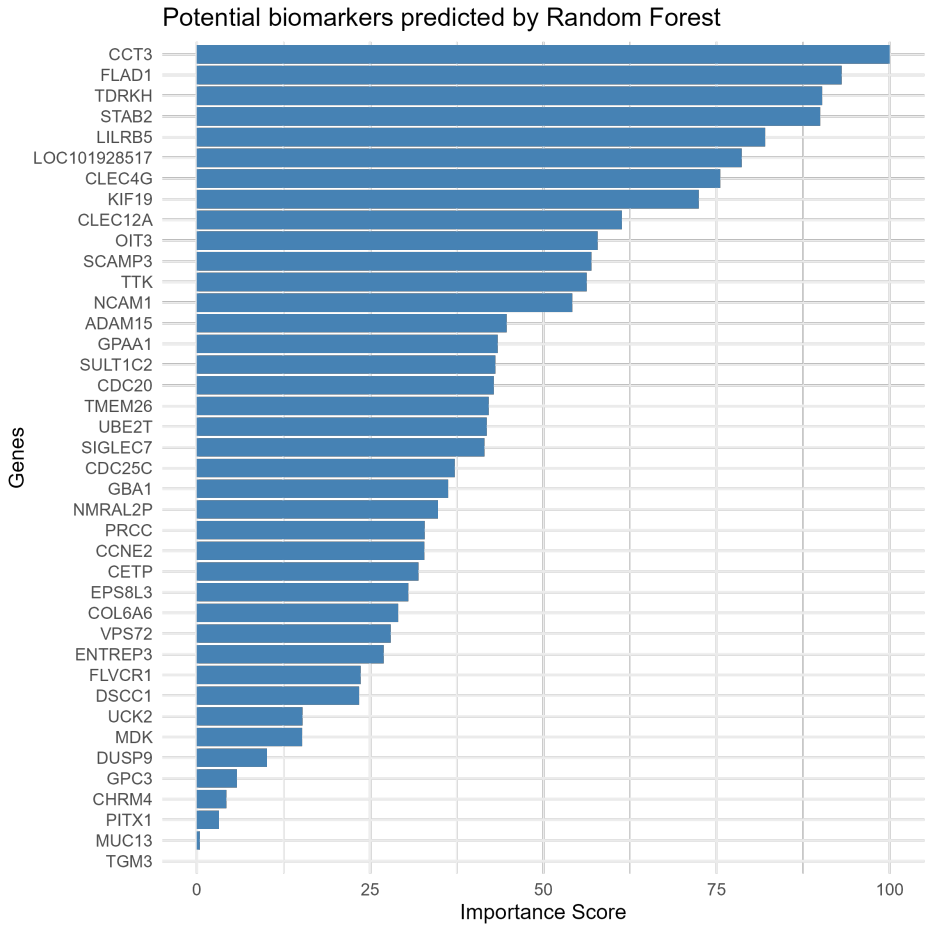

Figure 5: Potential biomarkers identified by Random Forest based on gene importance scores.

3.4 Exploratory Microbiome-Transcriptome Correlation Analysis

To investigate potential interactions between the gut microbiota and host gene expression, an exploratory correlation analysis was performed between microbial taxa and differentially expressed genes (DEGs). In the absence of metadata linking individual fecal and tissue samples, sample alignment was based on their corresponding order within the datasets. Pearson

correlation coefficients were then calculated to explore associations between microbial abundances and gene expression levels in tumor tissues.

While some moderate correlations were observed between specific OTUs and individual genes (Figure 6), the interpretability of these results is limited by the lack of verified sample-level pairing. As such, this component remains exploratory and primarily serves to illustrate the analytical framework for future integrative studies with better sample linkage.

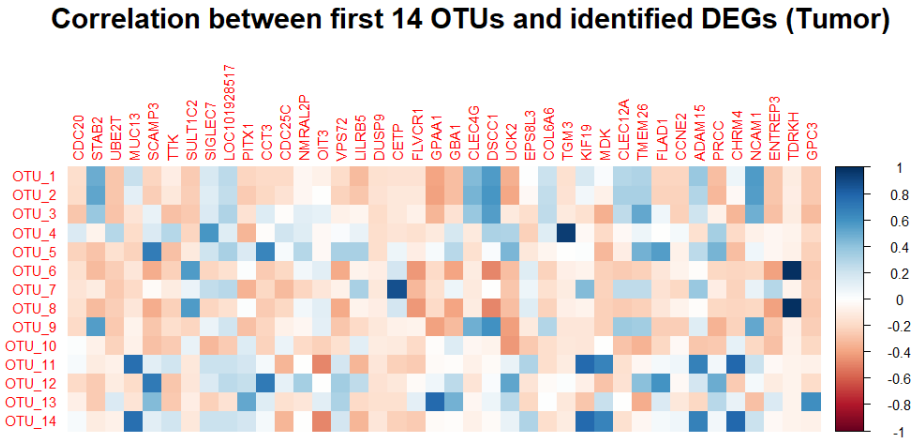

Figure 6: Correlation heatmap between the first 14 OTUs and top differentially expressed genes in tumor samples.

4 Discussion

This study integrated gut microbiome profiling and transcriptomic analysis of liver tissue to explore biological alterations associated with hepatocellular carcinoma. Although both data types were derived from the same publication, the lack of sample-level metadata linking fecal and tissue samples limited the ability to perform true multi-omic integration. Nevertheless, separate analyses provided valuable insights into microbial diversity, gene expression changes, and potential diagnostic biomarkers.

Transcriptomic analysis revealed hundreds of genes differentially expressed between tumor and adjacent non-tumor liver tissue, many of which are involved in cell cycle regulation, proliferation, and immune signaling. Notably, CCT3, CDC20, and CDC25C were consistently upregulated in tumors, consistent with their known roles in oncogenesis. Microbiome analysis showed moderate alpha diversity and high inter-sample variability, though no clear clustering emerged. Machine learning identified several DEGs, particularly CCT3 and FLAD1, with strong diagnostic potential, outperforming the conventional marker GPC3.

In summary, despite limitations related to sample matching, this study identifies distinct transcriptional and microbial signatures associated with hepatocellular carcinoma and underscores the utility of machine learning approaches in biomarkers prediction. These findings provide a valuable foundation for future integrative research exploring the tumor–microbiome axis in liver cancer.

References

[1] Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).

[2] Motola-Kuba, D. *et al.* Hepatocellular carcinoma: An overview. *Annals of Hepatology* **5**, 16–24 (2006).

[3] Llovet, J. M. *et al.* Hepatocellular carcinoma. *Nature Reviews Disease Primers* **7**, 6 (2021).

[4] Huang, H. *et al.* Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in hbv-related hepatocellular carcinoma. *Genome Medicine* **12**, 102 (2020).

[5] Abdelsalam, N. A. *et al.* The curious case of prevotella copri. *Gut Microbes* **15**, 2 (2023).

[6] Yeoh, Y. K. *et al.* Prevotella species in the human gut is primarily comprised of prevotella copri, prevotella stercorea and related lineages. *Scientific Reports* **12**, 9055 (2022).

[7] Liu, W. *et al.* Current understanding on the role of cct3 in cancer research. *Frontiers in Oncology* **12**, 961733 (2022).

[8] Yang, G. *et al.* Cdc20 promotes the progression of hepatocellular carcinoma by regulating epithelial–mesenchymal transition. *Molecular Medicine Reports* **24**, 483 (2021).

[9] Tsuchiya, A. *et al.* Hepatocellular carcinoma with progenitor cell features distinguishable by the hepatic stem/progenitor cell marker ncam. *Cancer Letters* **309**, 95–103 (2011).

[10] Zhang, Y. *et al.* Clec4s as potential therapeutic targets in hepatocellular carcinoma microenvironment. *Frontiers in Cell and Developmental Biology* **9**, 681372 (2021).

[11] Zhou, F. *et al.* Glypican-3: A promising biomarker for hepatocellular carcinoma diagnosis and treatment. *Medicinal Research Reviews* **38**, 741–767 (2018).

[12] Zheng, X. *et al.* Glypican-3: A novel and promising target for the treatment of hepatocellular carcinoma. *Frontiers in Oncology* **12**, 824208 (2022).

[13] Qian, E. *et al.* Expression and diagnostic value of cct3 and iqgap3 in hepatocellular carcinoma. *Cancer Cell International* **16**, 55 (2016).

[14] Mei, M. *et al.* Significant diagnostic and prognostic value of flad1 and related micrornas in breast cancer after a pan-cancer analysis. *Disease Markers* **2021**, 6962526 (2021).