Dean&Francis
ISSN 2959-6157

Analysis of Genes Associated with Liver Cancer Development

Tianyi Zheng[1, *]

[1]Department of bioinformatics, Harbin Institute of Technology University, Beijing, 100000, China

*Corresponding author: shuiqiping@ldy.edu.rs

Abstract:

Liver cancer, particularly hepatocellular carcinoma (HCC), is a leading cause of cancer-related deaths worldwide, with a high mortality rate due to its aggressive nature and potential for metastasis. This study aimed to identify key genes associated with liver cancer using RNA-seq data from The Cancer Genome Atlas (TCGA) database. Through differential expression analysis, five genes-GABBR2, KRBA1, SEZ6, LRRC28, and ASB7-were identified as significantly associated with liver cancer, with GABBR2 and SEZ6 exhibiting the most prominent associations. The analysis involved data preprocessing, rigorous statistical screening (P value and logFC), and comprehensive validation through violin plots, heatmaps, and Gene Ontology (GO) enrichment analysis. The results highlighted the involvement of these genes in critical biological processes, such as signal transduction and enzyme regulation, which are pivotal in the pathogenesis of liver cancer. The significant enrichment of these genes in specific GO terms further supports their potential role in liver cancer progression. This study provides valuable insights into the genetic basis of liver cancer, offering potential biomarkers and therapeutic targets for future research and clinical applications.

Keywords: Differential expression analysis; functional enrichment analysis; significant association.

1. Introduction

Liver cancer, particularly hepatocellular carcinoma (HCC), is one of the most common and deadly cancers globally. It directly impairs liver functions, leading to severe complications such as ascites and hemorrhage, with a significant risk of metastasis, causing immense suffering for patients. According to the global cancer statistics for 2022 released by IRCA on April 4, 2024, liver cancer accounts for 7.8% of all cancer-related deaths, ranking third worldwide, following lung and colorectal cancer.

Recent research has significantly advanced the understanding of the genetic mutations and molecular mechanisms driving HCC development. Key findings include the critical role of P53, a tumor suppressor gene, in regulating cell cycle and apoptosis. Studies have highlighted the significant expression of P53-re-

ISSN 2959-6157

lated genes, such as IL1A and F2R, in various cancers, including gastric cancer, suggesting similar roles in liver cancer. The differential expression of these genes in tumor versus normal tissues indicates their involvement in cancer progression and tumorigenesis [1].

Previous studies on the influencing factors of liver cancer have predominantly focused on clinical perspectives. However, research on the relationship between cancer and genes has not specifically targeted liver cancer. This study will employ bioinformatics and statistical methods to analyze gene expression data related to liver cancer, utilizing comprehensive resources like the GEO (Gene Expression Omnibus) and TCGA (The Cancer Genome Atlas) databases [2, 3]. These databases provide extensive gene expression data from various liver cancer patients, encompassing a wide range of genetic mutations and expression variations. This wealth of data can provide substantial support for this research.

Research on genes associated with liver cancer enables the identification of specific genetic mutations and expressions, which is crucial for screening high-risk populations, early prevention of cancer, and the development of targeted therapies. Such research holds significant scientific and clinical importance. Gene-expression profiling, essential in understanding cancer biology, could be instrumental in identifying unique genetic signatures for HCC, aiding in diagnosis and treatment [4]. Integrating global cancer statistics, gene expression analysis, and profiling provides a comprehensive framework for studying the genetic basis of liver cancer [5].

Databases used for gene expression analysis in other cancers can also be used to identify critical genetic alterations in liver cancer [6]. Prognostic models based on death-associated genes have been developed to predict patient outcomes and guide treatment decisions in HCC, offering promising tools for early diagnosis and prognosis, though further clinical validation is required [7].

Advanced bioinformatics analysis methods, including differential gene expression analysis, gene enrichment analysis, and network analysis, have been validated in numerous cancer studies. These methods are effective in identifying critical genes and signaling pathways associated with liver cancer. By utilizing these techniques, this paper aims to establish predictive models that can further validate the reliability of the analysis results.

2. Methods

2.1 Data Source

The data for this study were sourced from the TCGA liver cancer RNAseq dataset. Due to the extensive size of the raw data, it was organized into a table summarizing the differential gene expression between cancer and normal samples, facilitating the smooth progression of subsequent research.

2.2 Screening of Relevant Genes

2.2.1 Data preprocessing

First, data processing is performed. The code reads RNA-seq count data from the file, extracts the sample type information, and categorizes the samples into cancer and normal groups. A grouping factor (cancer vs. normal) is then created based on sample type, which is used in subsequent differential expression analysis.

2.2.2 Differential expression analysis

An object containing sample data and grouping information is created for differential expression analysis. The data are normalized to detect differences in gene expression between normal and cancer samples, with the identified genes preliminarily considered to be associated with liver cancer. The dispersion of the data is estimated to account for the biological variability in gene expression [8]. Differential expression analysis is conducted using a generalized linear model (GLM) to detect significant differences in gene expression between the groups (Figure 1).

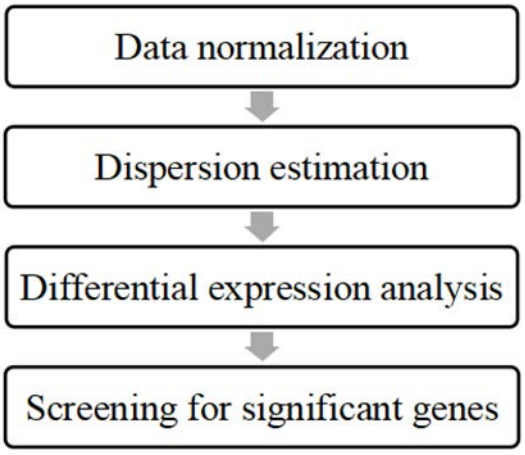

Fig. 1 Main points of differential expression analysis

The screening of genes is primarily based on two criteria: the p-value and logFC (log Fold Change). P-value is used to assess whether the observed gene expression differences are statistically significant; a p-value less than 0.05 typically indicates that the gene's expression difference between the groups is statistically significant, suggesting that the expression difference between cancer and normal samples are not due to random variation at a 95% confidence level.

logFC measures the magnitude of gene expression changes between the two groups. An absolute logFC greater than 1 usually indicates at least a twofold change in gene expression between the groups, filtering out genes that, despite statistical significance (low p-value), show minimal expression change. These genes may be biologically less important, allowing the focus to shift to genes with actual biological significance.

The overall goal of the code is to identify genes with significantly different expression levels between cancer and normal samples. By applying stringent statistical (p-value) and biological (logFC) filters, it is possible to identify genes with potential significance in disease research. These genes can be further used for biological validation, mechanistic studies, or as potential biomarkers [9, 10].

2.3 Verification of Result Accuracy

2.3.1 Violin plot

The violin plot visualizes data distribution and probability density by combining elements of boxplots and density plots. It aids in understanding the distribution of gene expression between tumor and normal groups. The symmetrical areas represent the density distribution, with wider regions indicating higher density. Black dots and vertical lines denote the median and interquartile range, while the Y-axis reflects gene expression levels. This plot provides an intuitive visualization of data distribution, enhancing the clarity of expression differences between groups and bolstering the credibility of the results.

2.3.2 Heatmap

A heatmap visually represents the expression patterns of significant genes across different samples, such as cancer and normal groups. By clustering expression data, it helps identify consistent expression differences between samples. The heatmap clearly shows these differences, confirming whether identified genes exhibit consistent patterns, such as high or low expression across groups. This facilitates a clear visualization of group differences, further strengthening the credibility of the results.

2.3.3 Functional enrichment analysis

Functional enrichment analysis evaluates whether differentially expressed genes have biological significance by assessing their enrichment in specific biological processes, molecular functions, or pathways. This analysis typically employs GO or KEGG pathway analysis to determine if the identified genes are enriched in relevant biological processes or pathways [9,10]. When genes are enriched in key biological pathways, it confirms the biological relevance of the analysis, underscoring their importance in disease progression. This method, when combined with boxplots and heatmaps, provides multi-level validation of the results, ensuring both statistical and biological accuracy.

3. Results and Discussion

Through step-by-step data screening, analysis, and verification, five genes associated with liver cancer were identified: GABBR2, KRBA1, SEZ6, LRRC28, and ASB7. Among them, GABBR2 and SEZ6 show the most significant association with liver cancer.

3.1 Analysis of Relevant Genes

The primary objective of this experiment, conducted through R code, is to identify genes with significantly differential expression between cancer and normal samples. Through stringent statistical screening (p-value) and biological screening (logFC), five genes associated with liver cancer were identified. The following table 1 presents the relevant genes that meet the specified criteria.

Table 1. Relevant genes

gene	logFC	p-value
GABBR2	2.466	0.006
KRBA1	2.242	0.015
SEZ6	3.627	0.019
LRRC28	2.173	0.042
ASB7	2.008	0.047

Based on the results, all five genes identified in the study are significantly associated with liver cancer. The study of these genes may provide novel targets for the early diagnosis and treatment of liver cancer in the future, offering hope for improved patient prognosis.

3.2 Accuracy Verification

3.2.1 Analysis of violin plot results

The violin plot reveals that gene expression significantly differs between the tumor and normal groups, with higher

ISSN 2959-6157

expression values observed in the tumor group and lower values in the normal group. These results are statistically significant. The distinct difference in expression patterns strongly supports the role of these genes in the disease mechanism. LIHC refers to Liver Hepatocellular Carcinoma, where T represents the number of tumor samples, and N represents the number of normal samples.

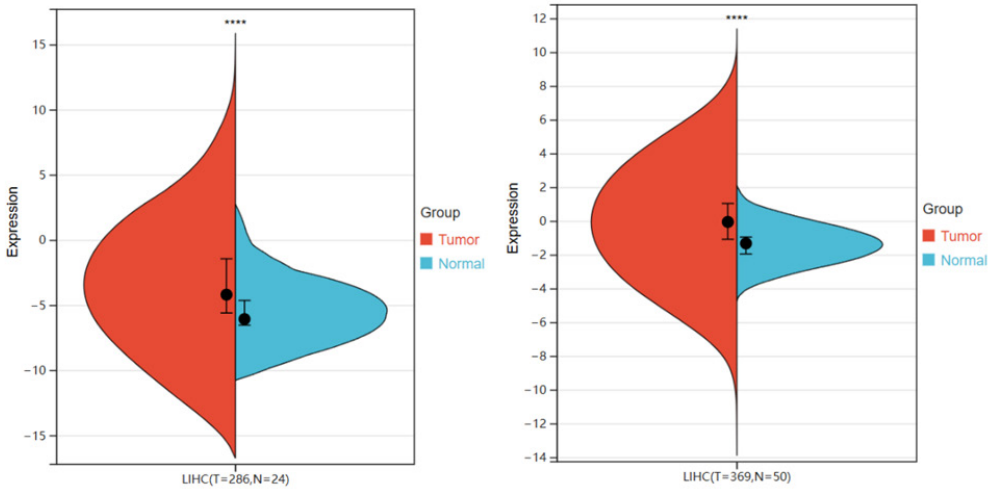

Fig. 2 GABBR2 and KRBA1

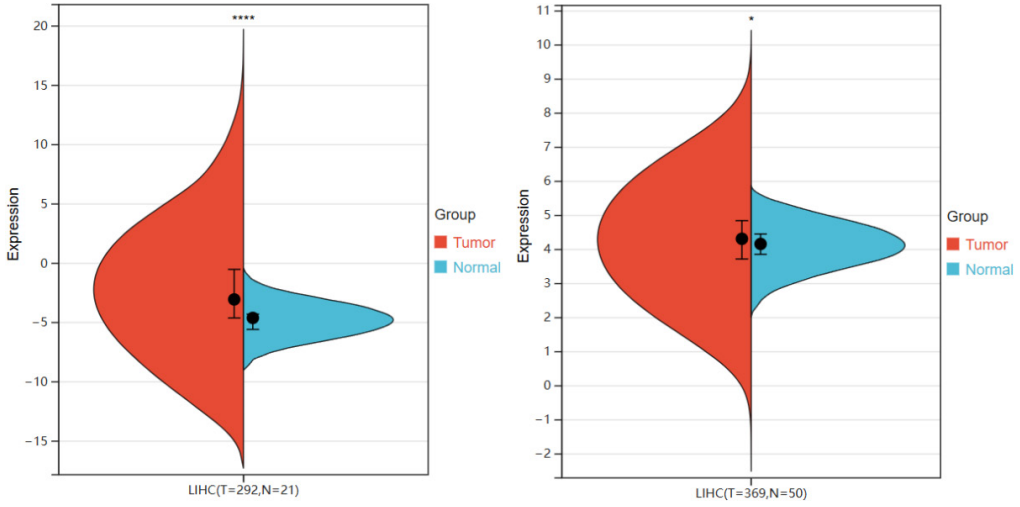

Fig. 3 SEZ6 and LRRC28

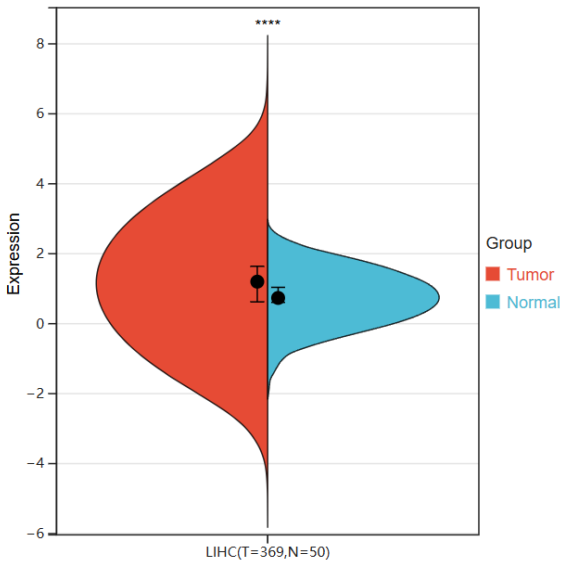

Fig. 4 ASB7

In the figure 2, 3 and 4, T represents the number of tumor samples, N represents the number of normal samples, red represents the tumor group, and blue represents the normal group. It can be seen that the expression of all screened genes in the tumor group is significantly higher than that in the normal group. This means that the expression level of this gene is higher in tumors, and this difference is statistically significant.

3.2.2 Analysis of heatmap results

The dendrograms at the top and left of the heatmap represent the hierarchical clustering of samples and genes. These dendrograms help identify the similarities and differences among samples and genes. The color gradient in the heatmap indicates the variation in gene expression levels. Each column corresponds to a different sample label, and each row represents a significantly differentially expressed gene. By analyzing the color gradient, the patterns of gene expression can be identified, and these distinct expression patterns support the significant differential expression of these genes under different phenotypes or conditions (Figure 5).

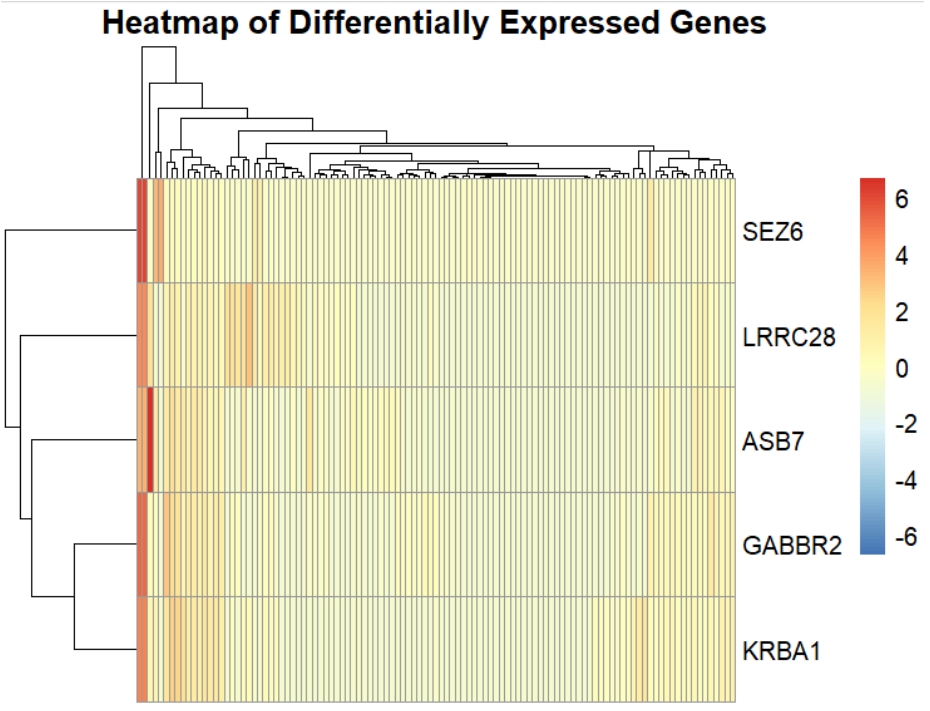

Fig. 5 Heatmap of different expressed genes

Through the analysis of the heatmap, if the samples are clearly clustered together according to their groups (tumor and normal) in the dendrogram, it indicates that these significantly differentially expressed genes exhibit consistent expression patterns across different sample groups. This consistency suggests that the expression of significant genes within cancer and normal samples is similar among individuals within each group, ruling out the possibility of coincidental findings, thereby further validating the reliability of the identified genes.

3.2.3 Analysis of enrichment results

Through Gene Ontology (GO) enrichment analysis, it is possible to understand the enrichment of significantly differentially expressed genes in specific biological processes (table 2, 3).

ISSN 2959-6157

Table 2. Gene Ontology (GO) enrichment results

gene	q-value	p-value	p.adjust
GABBR2	0.009	0.004	0.035
GABBR2	0.009	0.004	0.035
GABBR2	0.009	0.005	0.035
SEZ6	0.009	0.005	0.035
SEZ6	0.009	0.007	0.035
GABBR2	0.009	0.008	0.035

Table 3. Description of geneID

gene	Description
GABBR2	negative regulation of adenylate cyclase activity
GABBR2	negative regulation of cyclase activity
GABBR2	negative regulation of lyase activity
SEZ6	regulation of protein kinase C signaling
SEZ6	cerebellar Purkinje cell layer development
GABBR2	gamma-aminobutyric acid signaling pathway

These analysis results demonstrate the involvement of the identified significant genes (GABBR2 and SEZ6) in multiple key biological processes. Furthermore, these findings have undergone rigorous statistical correction, thereby ensuring a high level of credibility.

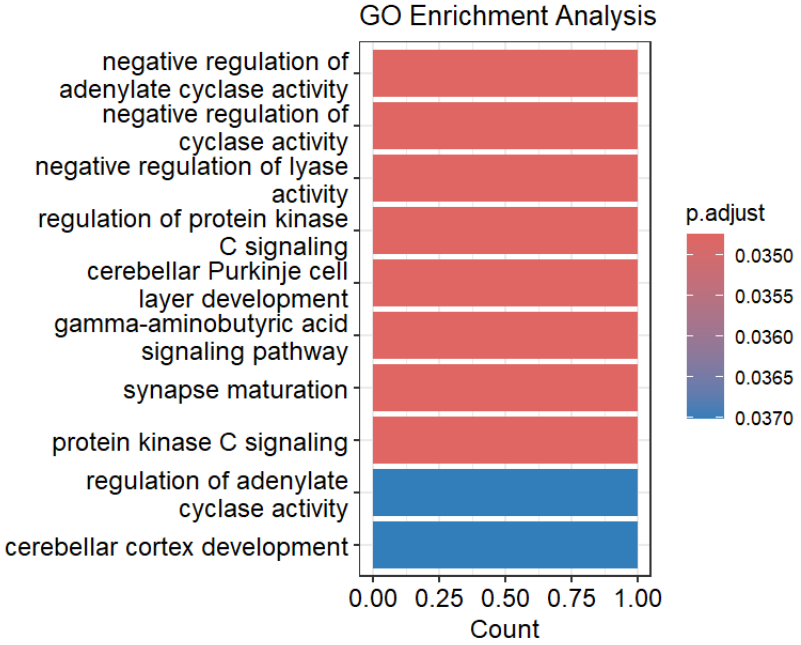

Fig. 6 GO enrichment analysis plot

The GO enrichment analysis plot was obtained, with the X-axis displaying the count ratio of significantly differentially expressed genes for each GO term, and the Y-axis listing the significantly enriched GO terms. Each term describes a specific biological process. The color represents the adjusted p-value (p.adjust), with a gradient from red to blue indicating lower to higher adjusted p-values. The length of the bars reflects the proportion of genes within each GO term relative to the entire gene set in the enrichment analysis.

The red portion in the color bar indicates that these terms are statistically highly significant (with a lower p.adjust), meaning that these biological processes are highly enriched in the gene set. This suggests a significant association of GABBR2 and SEZ6 with liver cancer (Figure 6).

4. Conclusion

This study analyzed RNA-seq data from the TCGA liver cancer dataset to identify genes significantly associated with liver cancer. The research process involved several key steps: data preprocessing, differential expression analysis, and validation of result accuracy using multiple methods. Five genes-GABBR2, KRBA1, SEZ6, LRRC28, and ASB7-were found to be associated with liver cancer, with GABBR2 and SEZ6 showing the most significant association.

By combining differential expression analysis with rigorous statistical and biological validation methods, this study provides a comprehensive approach to identifying liver cancer-related genes. The identified genes offer potential insights for further research into the mechanisms of liver cancer and may serve as biomarkers or therapeutic targets for future clinical studies.

References

[1] Wang Zhiyuan, et al. Expression and functional analysis of P53 related genes IL1A and F2R in gastric cancer. Journal of Oncology, 2024.

[2] Xing Long, Wu Shuangli, Wu Tiecheng, et al. Establishment and Validation of a Prognostic Model for Aging Related Genes in Lung Adenocarcinoma Based on TCGA Database. Hainan Medical, 2024, 35(10): 1374-1379.

[3] Jin Lei, Tang Xiaolei, Gu Junfei. TCGA database analysis of molecular characteristics, carcinogenic effects, and related immune and pharmacogenomic features of EGFR in colorectal cancer. Journal of Qiqihar Medical College, 2024, 45(17): 1606-1614

[4] Sotiriou C, Piccart M J. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care. Nat Rev Cancer, 2007, 7(7): 545553.

[5] Sung H, Ferlay J, Siegel R L, 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, 2021, 71(3): 209-249.

[6] Coe B P, Chari R, Lockwood W W, et al. Evolving strategies for global gene expression analysis of cancer. J Cell Physiol, 2008, 217(3): 590-597.

[7] Su Shuzhi, Zhang Kaiyu, Wang Ziying. A gene feature extraction method based on cross perspective similarity order preservation. Journal of Electronics and Information Technology, 2023, 45(1): 317-324.

[8] Gao Jingya, Yu Yang, Xie Yan, et al. Differential expression of miRNA between medullary thyroid carcinoma and papillary thyroid carcinoma using bioinformatics analysis. Prevention and treatment of endemic diseases in China, 2020, 35(02): 105-108.

[9] Jiang Miaomiao. Identification of differentially expressed genes and their functions related to hypoxic-ischemic encephalopathy through bioinformatics analysis. Tianjin Medical University, 2020.

[10] Chen Guo, Tian Limin, Li Xiaoping. Screening of prognostic related genes for stage III colon cancer patients based on bioinformatics. Journal of Ningxia Medical University, 2022, 44 (05): 493-498+504.