

Bioinformatic Analysis of Key Genes Related to Tumor Microenvironment in Hepatocellular Carcinoma

Jingxu Zhang

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

Hao Liu (✉ liuhaodoctor@hotmail.com)

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

<https://orcid.org/0000-0002-0182-287X>

Keyi Zhao

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

Zhiye Bao

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

Zhishuo Zhang

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

Ye Li

The First Hospital of China Medical University: The First Affiliated Hospital of China Medical University

Research

Keywords: TCGA, ESTIMATE, Hepatocellular carcinoma (HCC), Tumor microenvironment (TME), bioinformatic analysis, prognosis, OncoPrint

Posted Date: May 24th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-508119/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Tumor microenvironment (TME) plays important roles in the development of different types of cancer. However, the critical regulatory members of TME related to hepatocellular carcinoma (HCC) remain unclear. In this study, a bioinformatic analysis based on Cancer Genome Atlas (TCGA) and Estimation of STromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE) datasets was conducted to predict the key genes affecting TME in HCC.

Material and Methods: First, 340 patients and 20531 genes' expression data with ESTIMATE scores were filtered and combined to identify differentially expressed genes. Next, protein-protein interaction (PPI) network and functional enrichment analysis were conducted to find hub genes. Then, log-rank test and functional enrichment analysis were conducted on the consensus genes and hub genes. Finally, Kaplan-Meier curves of the hub genes were drawn. As verification, those genes were searched on Oncomine database.

Results: Among all differentially expressed genes, 916 genes were expressed in both the immune and stromal groups. The Gene Ontology (GO) terms they enriched were T cell activation, leukocyte migration, collagen-containing matrix, external side of plasma membrane, receptor ligand and activator activity. Cytokine-cytokine receptor interaction was the most significant Kyoto Encyclopedia of Genes and Genomes (KEGG) term. Furthermore, *cd3e*, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* hub genes were low expressed in 304 patients, participating in a variety of responses including immune response-activating cell surface receptor signaling, immune response-activating signal transduction, clathrin-coated vesicle membrane, immune receptor activity-peptide binding and amide binding pathways. Their low expression was also verified on Oncomine database.

Conclusion: *cd3e*, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* participated in many aspects related to TME, and their low expression constructs a signature, may predict a poor 5 years' survival in hepatocellular carcinoma.

Background

Liver cancer was the sixth most commonly diagnosed cancer and the third leading cause of cancer deaths worldwide in 2020 [1]. Hepatocellular carcinoma (HCC) comprises about 75-85% of liver cancer cases. Despite the developments in diagnostic methods, surgical techniques and novel drugs, late diagnosis and malignancy lead to poor long-term survival rate in HCC. The median survival following diagnosis is approximately 6-20 months [2]. Thus, discovering predictive biomarkers for HCC is vital for early diagnosis, effective treatment and better prognosis.

Apart from novel predictors like lncRNAs and miRNAs [3, 4], factors involved in immune system are also worth exploring in cancer. One such factor, the tumor microenvironment (TME) is defined as the cellular and physical environment surrounding the primary tumor, which has recently become a research hotspot [5]. It has cellular components like inflammatory and immune cells, stromal components like extracellular

matrix (ECM) proteins, soluble cellular factors, etc [6]. The interactions of tumor cells with these molecules and cells in the TME can influence the biological behavior of cancer cells and the prognosis of cancer patients [7]. For example, a selenoprotein known as glutathione peroxidase 3 (GPx3), has a dichotomous role in different tumor types, which means it could act as both tumor suppressor and tumor promoter [8]. Therapeutic strategies based on TME have been tested in clinical trials and showed promising results [9].

Meanwhile, as an excellent tool, The Cancer Genome Atlas (TCGA) database collects gene expression quantification data and clinical information of different types of cancers [10]. Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE) is a tool for predicting tumor purity, using gene expression data to detect the presence of infiltrating stromal/immune cells in tumor tissues [11]. These values, combined with the expression data have been successfully applied to predict the overall survival rate in different types of cancers, such as glioma and colorectal cancer [12, 13]. TME has also been used for survival analysis in other tumors like thyroid cancer [14]. However, there is insufficient evidence of TME in HCC bioinformatic analysis. This study aimed to find hub genes associated with HCC and TME using TCGA dataset combined with ESTIMATE immune score, therefore to construct a new model to predict the survival rate of HCC.

Materials And Methods

Data collection

The TCGA gene expression data and associated clinical characteristics of HCC patients were downloaded from cBioPortal (<http://www.cbioportal.org/datasets>) [15]. The immune score and stromal score of liver cancer were retrieved from ESTIMATE website (<http://bioinformatics.mdanderson.org/estimate/>) [14]. Samples with immune score and stromal score were separately studied.

Data analysis

All downloaded data were analyzed with R software (R version 4.0.2). Clinical data from 340 patients and 20531 genes' expression information were filtered from the raw data. ESTIMATE algorithm is based on single sample gene set enrichment analysis and generates three scores: 1) stromal score (that captures the presence of stroma in tumor tissue), 2) immune score (that represents the infiltration of immune cells in tumor tissue), and 3) estimate score (that indicates tumor purity). ESTIMATE score can reflect the purity of cancer, along with many biological behaviors and factors affecting tumor microenvironment. However, due to heterogeneity of different kinds of tumor, ESTIMATE score could show a variety form of results. It could not simply reflect patients' survival, nor the prognosis. Therefore, ESTIMATE score here were used as a filter, and had to combined with clinical research datasets like TCGA, only genes with a meaningful ESTIMATE score were considered participating in further analysis [16]. Immune/stromal score and mRNA expression information were merged based on sample codes. The ESTIMATE scores for each dataset were calculated and patients were divided into two equal groups of high or low ESTIMATE score by median split [17]. Limma package (ver 3.44.3) [18] was used to identify differentially expressed

genes between samples with high score and low score. Genes with fold change (FC) more than two times (i.e. $|\log_2 \text{FC}| > 1$) and p-value < 0.05 were considered as differentially expressed [19]. Subsequently, genes expressed differentially in both immune score and stromal score groups were considered meaningful in HCC's TME, and underwent further analysis.

Functional enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were constructed with clusterProfiler (ver 3.16.0) [20] and BiocManager (ver 1.30.10) package. They contained pathways of different aspects of biochemical functions. Only functions with p-values < 0.05 were considered to be significant.

Protein-protein interaction (PPI) network construction:

STRING database (<http://string-db.org/>, version 11.0) was used to construct the PPI network [21]. For higher confidence, gene-gene interactions with integrated scores > 0.90 were selected rather than 0.4 and visualized using Cytoscape (version 3.8.0) [22]. A total of 30 hub genes were filtered using plug-in cytoHubba (ver 0.1) to demonstrate interactions.

Survival analysis and Oncomine verification.

These genes underwent survival analysis with Kaplan-Meier method, and curves were generated. Moreover, GO and KEGG analyses were conducted on hub genes, and the difference between overall genes' function and hub genes' function was demonstrated.

Results

Differences in gene expression

We conducted gene expression analysis to identify the genes that were over-expressed or under-expressed in tumor tissues and stromal tissues. A total of 340 patients with immune score and stromal score were carefully screened from the raw data, whose clinical characteristics are shown in Table 1. Heat map showed first 50 gene expression data between high and low immune/stromal score. Volcano maps with all genes were also drawn (Figure 1-a, 1-b; Figure 2-a, 2-b). In the immune score group, 50 genes were over-expressed (marked as up), while 1224 genes were under-expressed (marked as down). In the stromal group, 66 genes were up, while 1863 genes were down. Besides, 916 common genes were matched from 1274 differentially expressed genes in the immune score group and 1929 genes in the stromal score group, which were shown in the Venn map (figure 3). Because key gene expression profile would be shown in survival analysis, we intersected the differently expressed genes only rather than providing up regulated and down regulated genes together in the Venn map.

Functional enrichment analysis

We performed functional enrichment analysis to further study the consensus genes. GO contained three groups: biological processes (BP), cellular components (CC), and molecular functions (MF), with 12 pathways in each group. T cell activation, collagen-containing ECM and receptor ligand binding were the most enriched GO terms, while cytokine-cytokine receptor interaction was the most significant KEGG term (Figure 4-a, 4-b).

Protein-protein interaction (PPI) network

We performed PPI network analysis to elucidate the interactions between consensus genes. The PPI network of differentially expressed genes contained 893 nodes, 3050 edges, and PPI enrichment p-value: $<1.0e-16$. Isolated nodes and nodes not connected to the main pathway were excluded. After processing with Cytoscape, 78 genes remained in the main stem of the PPI network, of which 30 hub genes were selected using plugin cytoHubba (Figure 5-a, 5-b) by Degree algorithm.

Functional enrichment analysis on hub genes:

Specifically, we conducted GO and KEGG pathway analyses within the hub genes to compare with overall consensus genes (Figure 6-a, 6-b). Those genes were expressed in 304 patients, with 234 alive and 70 dead. The immune response-activating cell surface receptor signaling pathway, immune response-activating signal transduction, clathrin-coated vesicle membrane, endocytic vesicle membrane, immune receptor activity, peptide binding and amine binding were enriched in the GO pathway, while Th1, Th2 and Th17 cell differentiation were enriched in the KEGG pathway.

Survival analysis:

We conducted survival analysis on the 30 hub genes. Among those, *cd3e*, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* were considered most significant. Kaplan-Meier curve showed that low expression of these genes was linked to poor overall survival of HCC patients within five years ($p<0.05$) (Figure 7).

Oncomine verification:

We searched *cd3e*, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* on Oncomine database to recheck their expression. The filters were genes, cancer vs normal group, hepatocellular carcinoma, under-expression gene rank and gene rank within top 10%. As a result, they were indeed low expressed in hepatocellular carcinoma in Roessler Liver group, a non TCGA dataset associated group, with all p value <0.05 . Their expression data were listed, making the signature more convincing (Figure 8).

Discussion

Hepatocellular carcinoma (HCC) is one of the most common malignancies worldwide, with an extremely poor prognosis [23]. One possible reason is the tumor microenvironment (TME) in HCC can shift from immunocompetent status to immunosuppressive status due to drugs or pathological condition [24]. Because HCC TME is potentially immunogenic and typically characterized by inflammation,

immunotherapy has been proposed as a potential novel therapeutic approach, which has prompted studies on various immuno-therapeutic strategies in advanced HCC patients. Therapies like chimeric antigen receptor T (CAR-T) cell therapy, checkpoint inhibitors, and onco-vaccines have shown promising results [25]. These studies were focused on a specific target provided by either experimental results or novel discovered genes from data analysis. The GO and KEGG pathways' results in this study showed that the immune system involved in HCC consists of response-activating cell surface receptor signaling pathway and immune response-activating signal transduction. Clathrin-coated vesicle, immune receptor activity, Th1, Th2 and Th17 cell differentiation are also involved. However, these concepts are insufficient to elucidate the mechanisms. Therefore, whether to specify the GO/KEGG method or develop a new analysis method should be examined in future research.

Through bioinformatic analysis, this study found several under-expressed hub genes related to HCC. Of these, *cd3e*, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* may participate in TME, and can be used as a signature. These genes were well documented in previous studies. *cd3e* and *cd3g* belong to the CD3 receptor group, which is crucial for T-cell development and regulation. Their defect can cause severe combined T and B-cell immunodeficiency, which increases the chances of malignancy [26]. In terms of *hla* genes, rs3077-C, an *hla-dp* allele, may affect hepatitis B virus surface antigen clearance [27], therefore, mutation or low expression of HLA-related genes may exacerbate HBV-related liver cancer. Since HLA is highly individualized, therapy can be adjusted accordingly [28]. Lymphocyte-specific proto-oncogene (*lck*) belongs to SRC family of protein tyrosine kinases (PTKs). The encoded protein is a key signaling molecule in the selection and maturation of T-cell development. It works in Lnc-Tim3 pathway, promoting T cell exhaustion, a phenotype that is correlated with compromised anti-tumor immunity. It may influence cancer therapies aimed at modulating the acquired immune system as consequence [17]. Mitogen-activated protein kinase kinase kinase kinase 1 (*map4k1*), also known as hematopoietic progenitor kinase 1 (*hpk1*) is a negative regulator of T-cell activation along with DUSP22 (JKAP), and DUSP14 [29, 30]. However, as a group, their interactions remain vague. They may participate in a common pathway, and need further study before they can be used for cancer diagnosis and survival prediction.

This study was based on bioinformatic analysis on TCGA and Oncomine database. Further analyses, such as cross comparison with GEO database [31], single or multiple factor Cox regression could also be conducted as supplements. Besides, new treatment strategies, new data can affect the prognosis and the analysis per se, therefore, researchers need to update data base timely, making studies based on data analysis more precise and convincing. Although bioinformatics is a well-accepted research method and provides prospective directions for future research, further analysis with datasets, in vitro experiments like PCR or gene knockout are still required to confirm gene expression and function.

Conclusions

cd3e, *cd3g*, *hla-dpa1*, *hla-dpb1*, *lck*, and *map4k1* participated in many processes related to TME, and their low expression constructs a signature, may predict a poor 5 year survival in hepatocellular carcinoma.

Abbreviations

BP Biological processes

CC Cellular components

ESTIMATE Estimation of STromal and Immune cells in Malignant Tumor tissues using Expression data

FC Fold change

GO Gene Ontology

GPx3 Glutathione peroxidase

HCC Hepatocellular carcinoma

KEGG Kyoto Encyclopedia of Genes and Genomes

MF Molecular functions

PPI Protein-protein interaction

ssGSEA Single sample Gene Set Enrichment analysis

TCGA The Cancer Genome Atlas

TME Tumor microenvironment

Declarations

Acknowledgements

Not applicable.

Funding

Not applicable.

Disclosure Statement

No competing financial interests exist.

Ethics approval and consent to participate

Not required

Consent for publication

All authors have acknowledged consent for publication

Availability of data and material

Datasets are open source

Competing interests

There are no conflict interests between authors.

Authors' contributions

Not available

Availability of supporting data

Not available

Authors' information:

Jingxu Zhang, 493387166@qq.com

Hao Liu, liuhaodoctor@hotmail.com

Keyi Zhao, 1669827301@qq.com

Zhiye Bao, nobledafeiyu1989@163.com

Zhishuo Zhang, 1178368836@qq.com

Ye Li, 623494039@qq.com

All authors are affiliated in The First Affiliated Hospital Of China Medical University, Shenyang, Liaoning, China

References

1. Sung, H., et al., *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J Clin, 2021.
2. *A new prognostic system for hepatocellular carcinoma: a retrospective study of 435 patients: the Cancer of the Liver Italian Program (CLIP) investigators*. Hepatology, 1998. **28**(3): p. 751-5.
3. Gu, Y., et al., *Prediction of circRNAs Based on the DNA Methylation-Mediated Feature Sponge Function in Breast Cancer*. Front Bioeng Biotechnol, 2019. **7**: p. 365.
4. He, S., et al., *Circular RNA hsa_circ_0000517 modulates hepatocellular carcinoma advancement via the miR-326/SMAD6 axis*. Cancer Cell Int, 2020. **20**: p. 360.

5. Jung, H.Y., L. Fattet, and J. Yang, *Molecular pathways: linking tumor microenvironment to epithelial-mesenchymal transition in metastasis*. Clin Cancer Res, 2015. **21**(5): p. 962-968.
6. Coperchini, F., et al., *Role of chemokine receptors in thyroid cancer and immunotherapy*. Endocr Relat Cancer, 2019. **26**(8): p. R465-r478.
7. Zou, Y., et al., *Application of the chemokine-chemokine receptor axis increases the tumor-targeted migration ability of cytokine-induced killer cells in patients with colorectal cancer*. Oncol Lett, 2020. **20**(1): p. 123-134.
8. Chang, C., et al., *Extracellular Glutathione Peroxidase GPx3 and Its Role in Cancer*. Cancers (Basel), 2020. **12**(8).
9. Li, S., et al., *Cisplatin promotes the expression level of PD-L1 in the microenvironment of hepatocellular carcinoma through YAP1*. Mol Cell Biochem, 2020.
10. Wang, Q., P. Li, and W. Wu, *A systematic analysis of immune genes and overall survival in cancer patients*. BMC Cancer, 2019. **19**(1): p. 1225.
11. Zhu, X., et al., *Potential Prognostic Value and Mechanism of Stromal-Immune Signature in Tumor Microenvironment for Stomach Adenocarcinoma*. Biomed Res Int, 2020. **2020**: p. 4673153.
12. Deng, X., et al., *Profiles of immune-related genes and immune cell infiltration in the tumor microenvironment of diffuse lower-grade gliomas*. J Cell Physiol, 2020. **235**(10): p. 7321-7331.
13. Liu, J., et al., *SCARA5 is a Novel Biomarker in Colorectal Cancer by Comprehensive Analysis*. Clin Lab, 2020. **66**(7).
14. Zhao, K., et al., *Identification of Key Genes in Thyroid Cancer Microenvironment*. Med Sci Monit, 2019. **25**: p. 9602-9608.
15. Lehrer, S. and P.H. Rheinstein, *Alzheimer Gene BIN1 may Simultaneously Influence Dementia Risk and Androgen Deprivation Therapy Dosage in Prostate Cancer*. Am J Clin Oncol, 2020.
16. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture from expression data*. Nat Commun, 2013. **4**: p. 2612.
17. Ji, J., et al., *Long non-coding RNA Lnc-Tim3 exacerbates CD8 T cell exhaustion via binding to Tim-3 and inducing nuclear translocation of Bat3 in HCC*. Cell Death Dis, 2018. **9**(5): p. 478.
18. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Res, 2015. **43**(7): p. e47.
19. Liu, X., X. Niu, and Z. Qiu, *A Five-Gene Signature Based on Stromal/Immune Scores in the Tumor Microenvironment and Its Clinical Implications for Liver Cancer*. DNA Cell Biol, 2020.
20. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters*. Omics, 2012. **16**(5): p. 284-7.
21. Dai, D., et al., *Weighted gene coexpression network analysis identifies hub genes related to KRAS mutant lung adenocarcinoma*. Medicine (Baltimore), 2020. **99**(32): p. e21478.
22. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res, 2003. **13**(11): p. 2498-504.

23. Zheng, Q., et al., *Immune checkpoint targeting TIGIT in hepatocellular carcinoma*. Am J Transl Res, 2020. **12**(7): p. 3212-3224.
24. Saito, A., et al., *Metformin changes the immune microenvironment of colorectal cancer in patients with type 2 diabetes mellitus*. Cancer Sci, 2020.
25. Park, R., et al., *Immuno-Oncotherapeutic Approaches in Advanced Hepatocellular Carcinoma*. Vaccines (Basel), 2020. **8**(3).
26. Lee, W.I., et al., *A Novel CD3G Mutation in a Taiwanese Patient With Normal T Regulatory Function Presenting With the COVID Phenotype Free of Autoimmunity-Analysis of all Genotypes and Phenotypes*. Front Immunol, 2019. **10**: p. 2833.
27. Creary, L.E., et al., *Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population*. Hum Immunol, 2019. **80**(10): p. 807-822.
28. Tang, S., et al., *Identification of C3 as a therapeutic target for diabetic nephropathy by bioinformatics analysis*. Sci Rep, 2020. **10**(1): p. 13468.
29. Chuang, H.C. and T.H. Tan, *MAP4K Family Kinases and DUSP Family Phosphatases in T-Cell Signaling and Systemic Lupus Erythematosus*. Cells, 2019. **8**(11).
30. Chuang, H.C. and T.H. Tan, *MAP4K3/GLK in autoimmune disease, cancer and aging*. J Biomed Sci, 2019. **26**(1): p. 82.
31. Liu, T., M. Zhang, and D. Sun, *Immune Cell Infiltration and Identifying Genes of Prognostic Value in the Papillary Renal Cell Carcinoma Microenvironment by Bioinformatics Analysis*. Biomed Res Int, 2020. **2020**: p. 5019746.

Tables

Table 1. The clinical information of 340 patients with hepatocellular carcinoma

Characteristics	Number	Percentage %
Age (years)		
<40	31	9.1
41-50	43	12.7
51-60	95	28
61-70	109	32
71-80	54	15.8
>80	8	2.4
Gender		
Male	109	32
Female	231	68
T stage		
T0	1	0.2
T1	171	50.2
T2	85	25
T3	73	21.7
T4	10	2.9
N stage		
N0	250	73.5
N1	3	1
NX	87	25.5
M stage		
M0	262	77
M1	3	1
MX	75	22
Stage		
stage 1	169	49.8
stage 2	83	24.4
stage 3	83	24.4

Figures

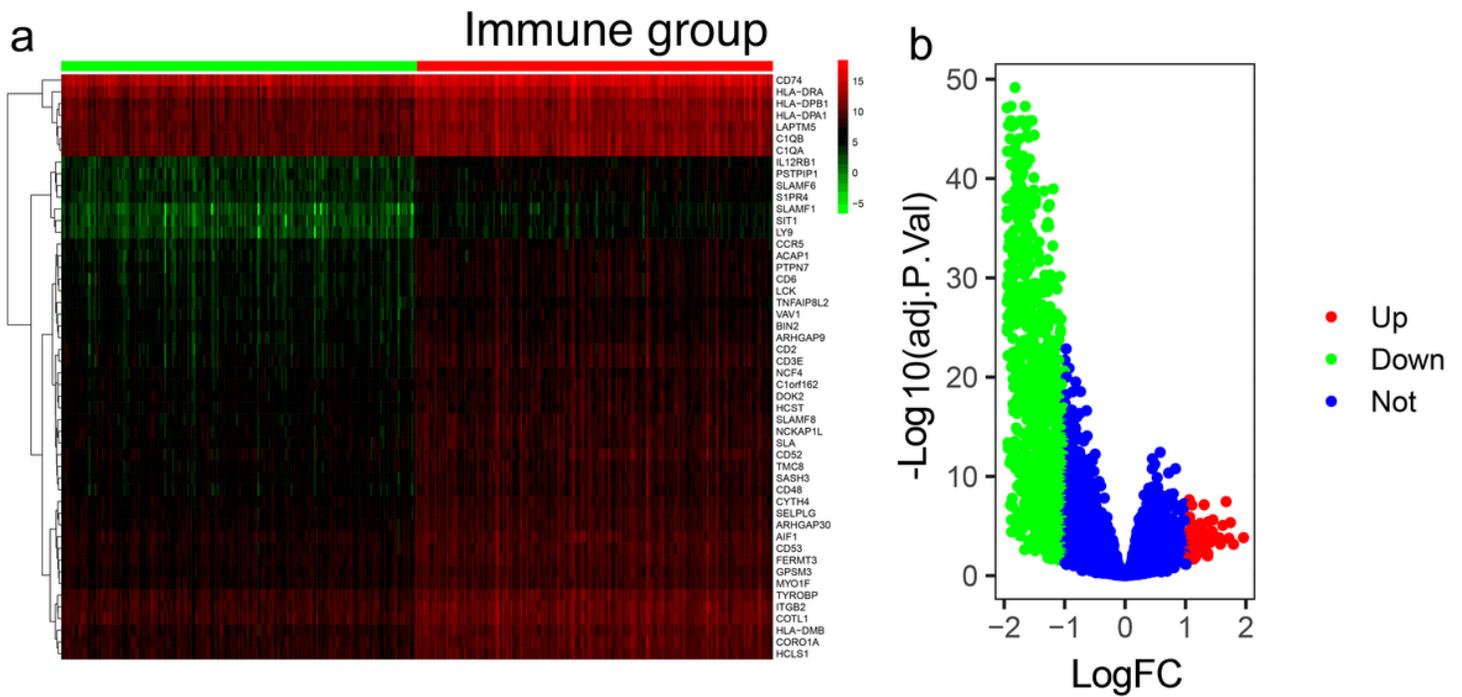


Figure 1

First 50 differentially expressed genes with immune score. Heat map (Figure 1-a) and volcano map (Figure 1-b) are shown. In the immune score group, 50 genes were over-expressed (marked as up), while 1224 genes were under-expressed (marked as down).

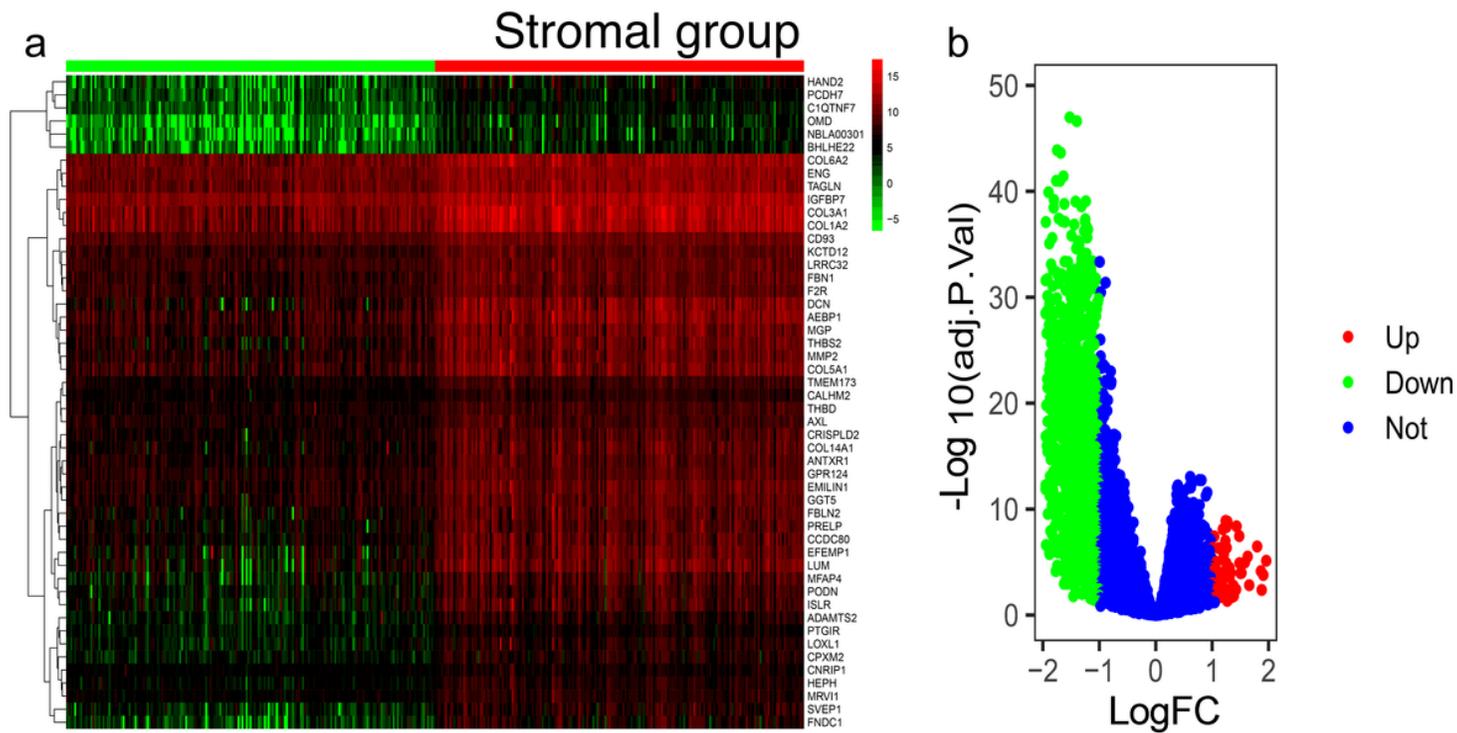


Figure 2

First 50 differentially expressed genes with stromal score. Heat map (Figure 2-a) and volcano map (Figure 2-b) are shown. In the stromal score group, 66 genes were over-expressed (marked as up), while 1863 genes were under-expressed (marked as down).

Differently expressed genes

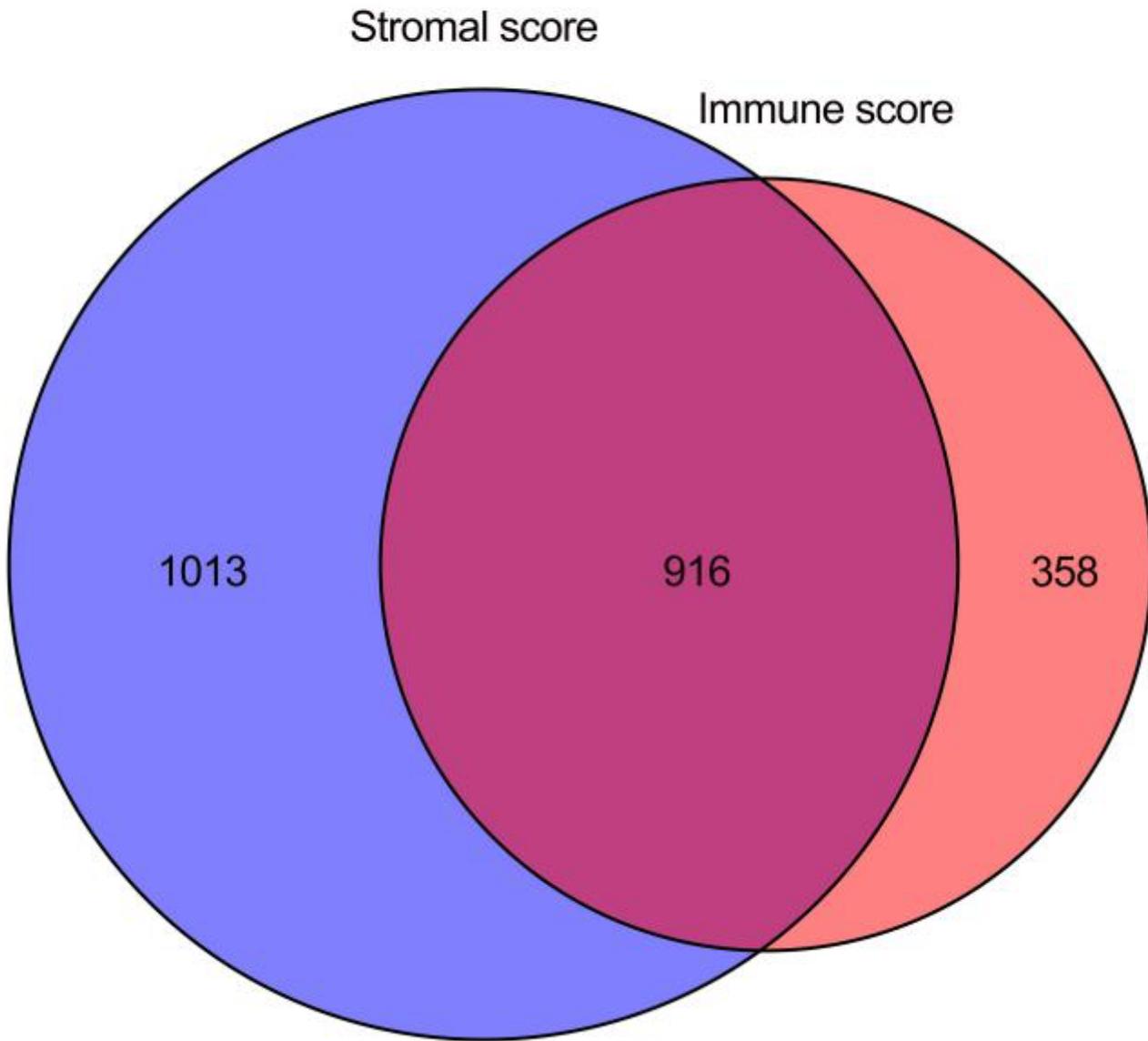


Figure 3

Venn plot of 1274 genes expressed differentially in the immune score group and 1929 genes in the stromal score, 916 consensus genes expressed in both immune score and stromal score.

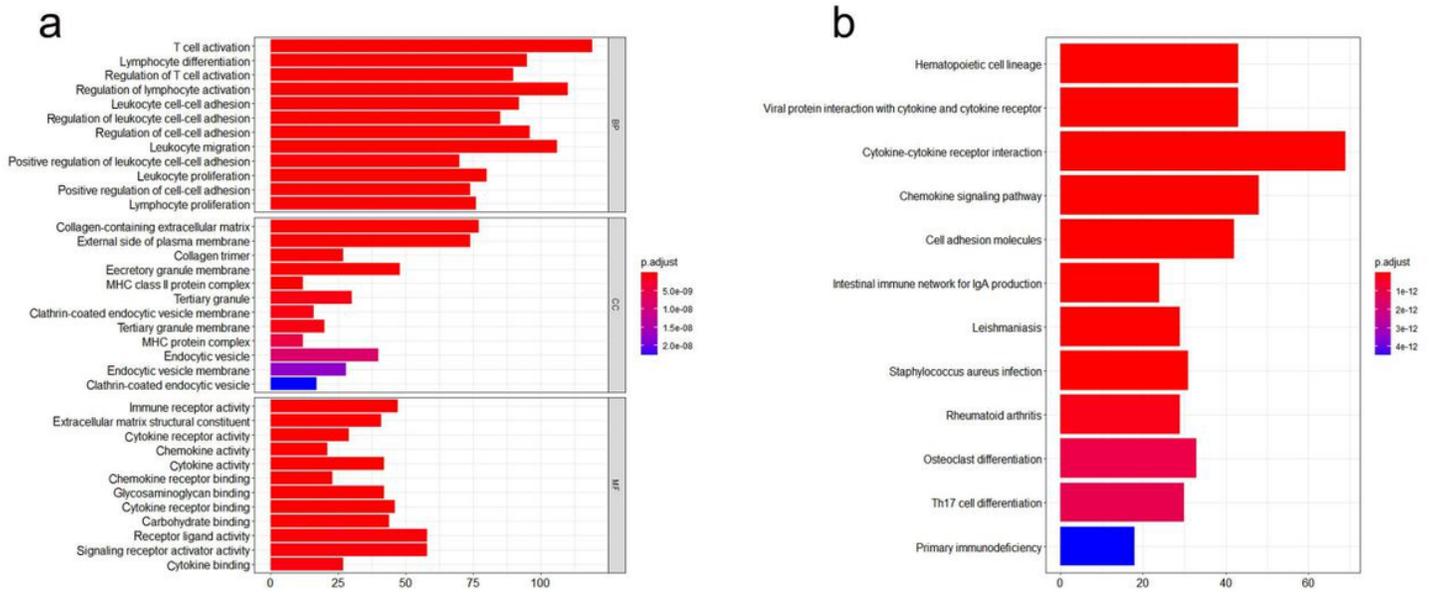


Figure 4

Functional enrichment analysis of differentially expressed genes. Figure 4-a, Top 12 biological processes (BP), cellular components (CC), and molecular functions (MF). Figure 4-b, Top 12 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

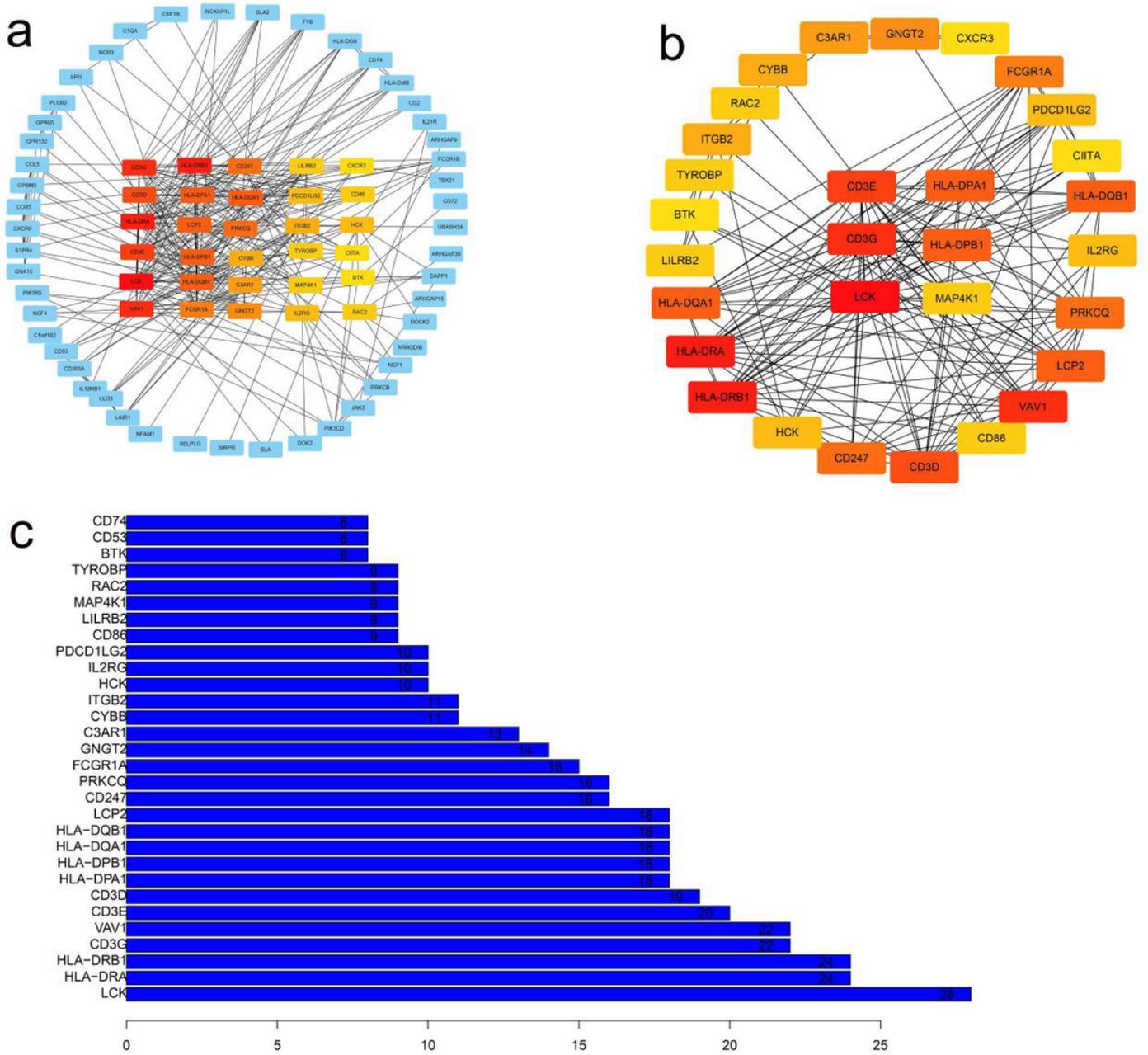


Figure 5

a, Protein-protein interaction (PPI) network of 78 differentially expressed genes and identification of 30 hub genes. Figure 5-b, PPI network of differentially expressed genes with integrated scores >0.9. Figure 5-c, PPI network of 30 hub genes with their degrees.

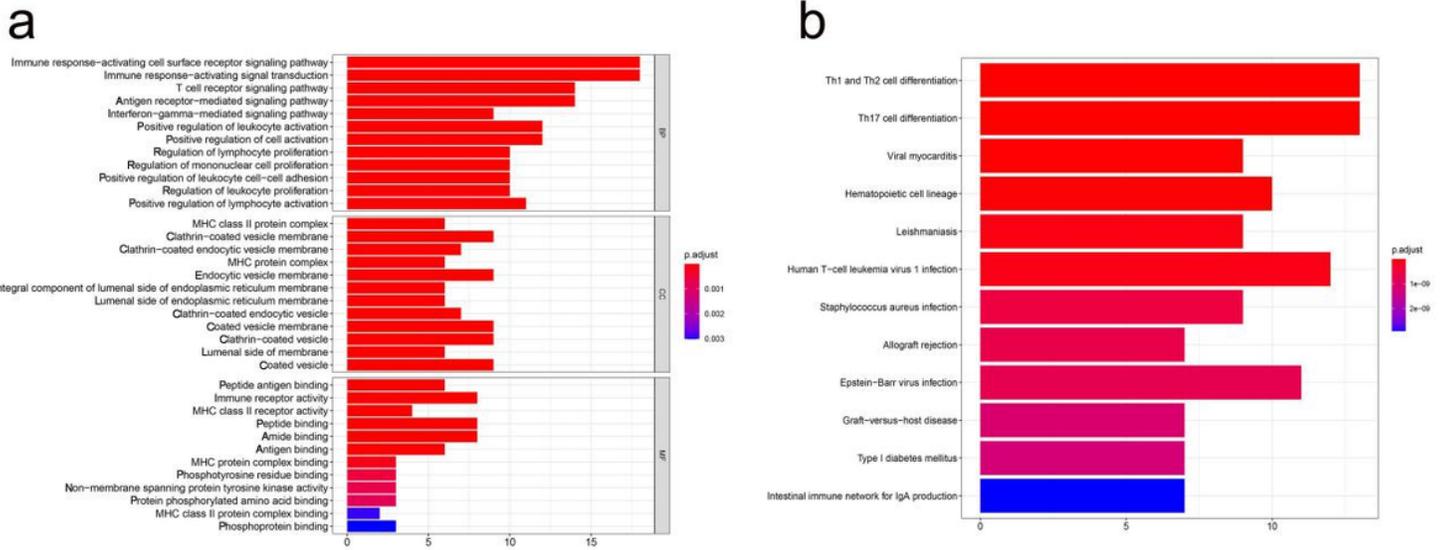


Figure 6

GO and KEGG analyses of hub genes. Figure 6-a, Top 12 biological processes (BP), cellular components (CC), and molecular functions (MF). Figure 6-b, Top 12 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

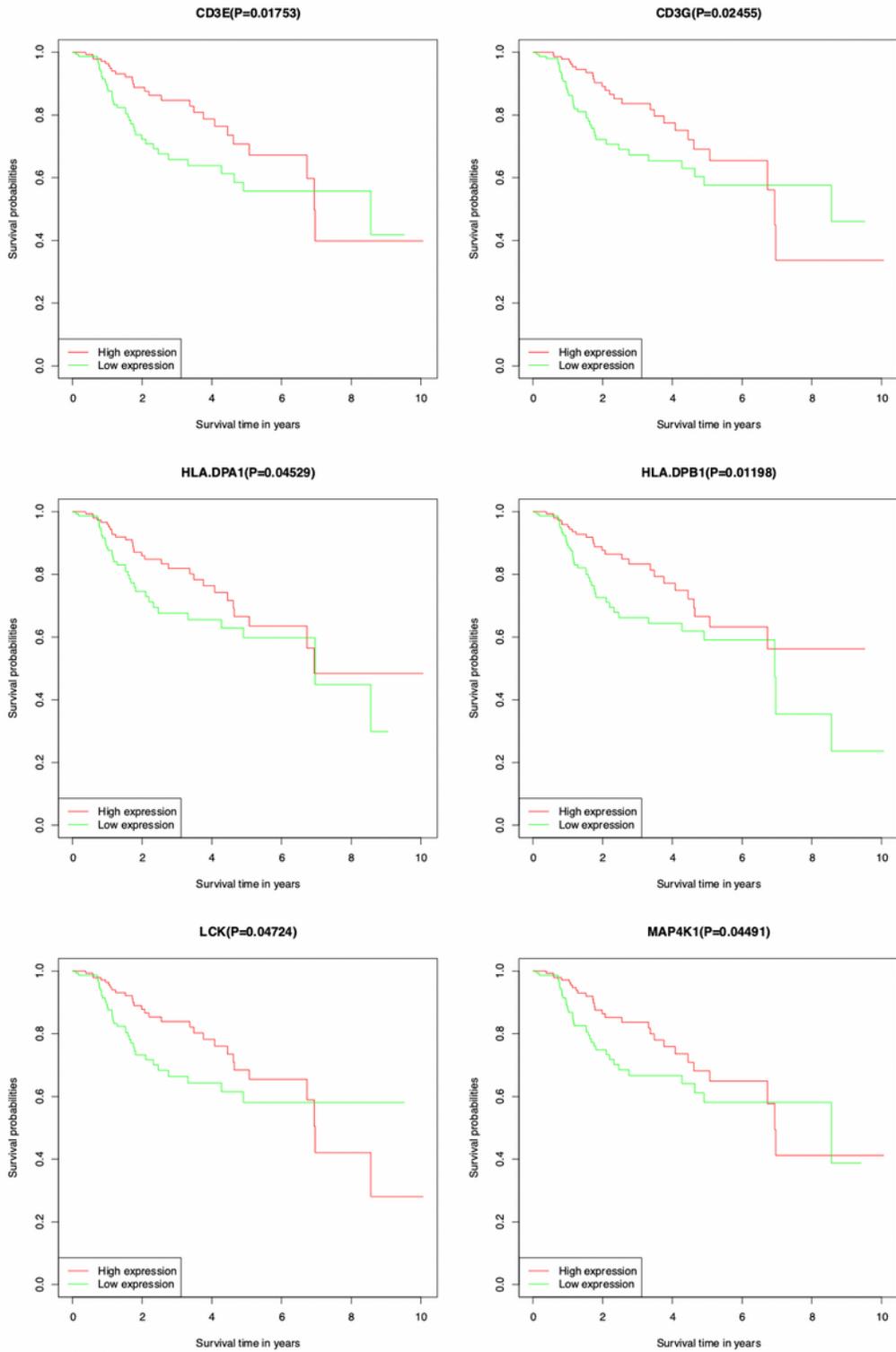


Figure 7

Kaplan-Meier curves of 6 most significant hub genes and 5-year survival. 304 patients, with 234 alive and 70 dead were analyzed. Low expressions of cd3e, cd3g, hla-dpa1, hla-dpb1, lck, and map4k1 correlated with poor overall survival within five years ($p < 0.05$).

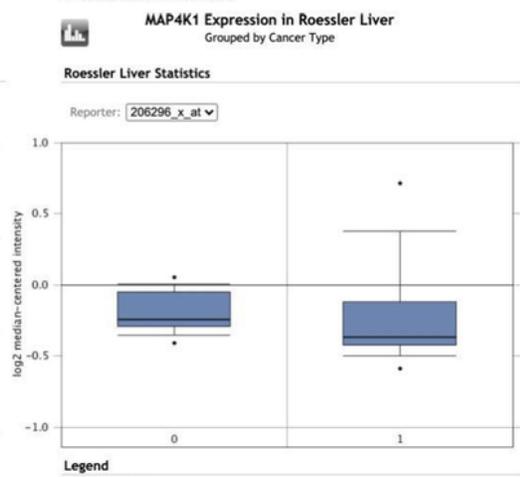
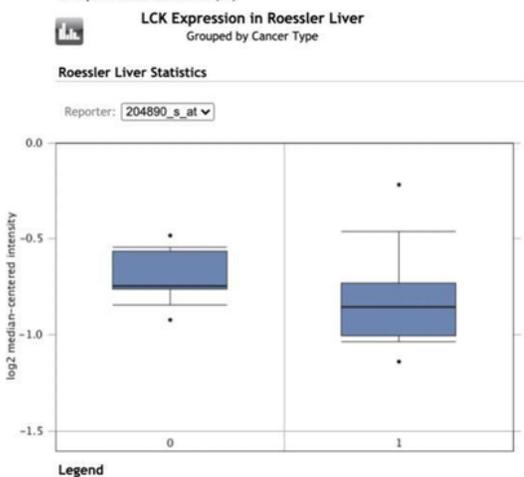
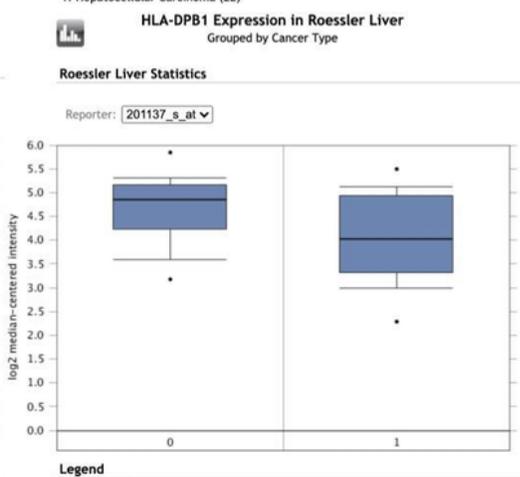
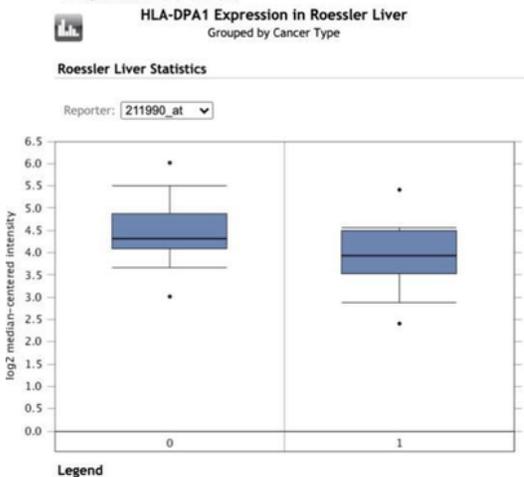
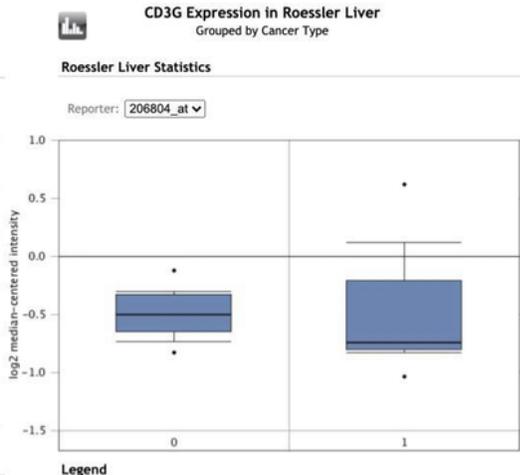
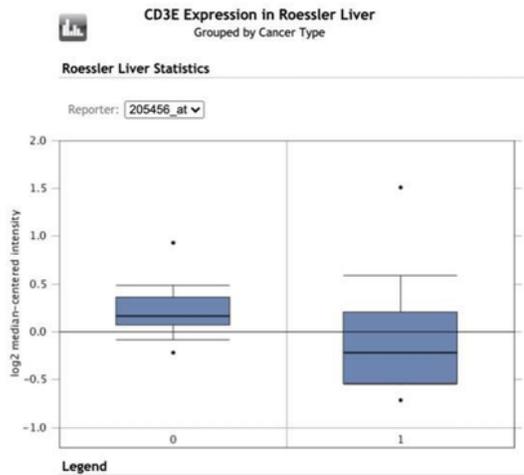


Figure 8

OncoPrint search results showed cd3e, cd3g, hla-dpa1, hla-dpb1, lck, and map4k1 were low expressed in Roessler Liver group, with all their p value<0.05.