Genes & Diseases (2023) 10, 1806–1808

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

RAPID COMMUNICATION

Construction of a novel predictive model with seven metabolism-related genes for hepatocellular carcinoma by machine learning

Single-cell RNA sequencing (scRNA-seq) and machine learning technologies, developed rapidly in recent years, offer an unparalleled opportunity to study the mechanism of tumorigenesis. Changes in metabolism-related genes (MRGs) result in hepatocellular carcinoma (HCC) progression, and MRGs have the potential to be used as a clinical prognostic indicator. In this study, a machine learning model was built using LASSO regression to find prognostic genes and create a new MRG risk signature of HCC patients. The patients were divided into the MRG high- and low-risk groups based on their risk scores. Significant overall survival advantages were observed in the MRG low-risk group. In addition, risk score and cancer status were used to establish a nomogram with high accuracy. The calibration curves and decision curves further demonstrated the satisfactory agreement between the predicted and observed values in the probability of overall survival. Subsequently, we discovered the different characteristics of the MRG high- and low-risk group patients based on tumor mutation burden, gene set enrichment, and immune infiltration. Our study identifies a novel MRG signature for risk-stratification and accurate identification of HCC patients with poor subtypes, which could precisely predict the prognosis of HCC and may guide personalized treatment for HCC patients.

The gene signature from the high-risk group may be used to conduct a risk assessment, and to enhance early cancer screening, early diagnosis, and treatment.[1] Models for predicting risk may also be able to help in the decision-making process for the clinical management of HCC patients.[2] ScRNA-seq developed rapidly which allows the study of transcriptional activity within an individual cell and enables the discovery of the gene expression of small but clinically significant tumor subpopulations.[3,4] In this study, scRNA-seq data of HCC from GSE149614 were used.[5] A UMAP algorithm was employed to visualize scRNA-seq data (Fig. 1A). With a cutoff of $|\text{logFC}| > 0.5$ and adjusted P value < 0.05, genes that were differentially expressed between the primary and metastatic tumor samples were identified. MRGs were obtained from the Reactome Pathway Database (https://reactome.org/). Using Venn diagrams, 78 differentially expressed genes were found to be metabolism-related (Fig. 1B and Table S1). The volcano plot of MRGs in HCC samples is shown in Figure S1A. The heatmap shows the top 50 MRG (Fig. S1B).

Subsequently, 26 genes related to the overall survival (OS) of 78 MRGs were selected via uni-Cox regression analysis (Table S2). The LASSO results indicated that eight MRGs were important which were then used as the input genes in the multi-Cox regression analysis (Fig. S1C, D). Multi-Cox regression analysis showed that *CYP27A1*, *CYP2C9*, *HMGCS2*, *NQO1*, *GLB1*, *PLPP1*, and *PGAM1* were hub MRGs. Kaplan–Meier analysis showed that high expression of *CYP27A1*, *CYP2C9*, *HMGCS2*, and *PLPP1* was correlated with better OS outcomes, while high expression of *NQO1* was correlated with worse OS outcomes (Fig. S2). T-SNE results of the expression of *CYP27A1*, *CYP2C9*, *HMGCS2*, *NQO1*, *GLB1*, *PLPP1*, and *PGAM1* in single-cell data (GSE149614) were shown in Figure S3. The MRG signatures were calculated based on the relative coefficient and expression of each gene as follows: risk score (RS) $= -0.00140 \times \mathit{CYP27A1} - 0.00195 \times \mathit{CYP2C9} - 0.00072 \times \mathit{HMGCS2} + 0.00120 \times \mathit{NQO1} + 0.01625 \times \mathit{GLB1} - 0.01114 \times \mathit{PLPP1} + 0.03327 \times \mathit{PGAM1}$ (Table S3). Each patient in the TCGA-LIHC and ICGC-LIRI-JP cohorts was then classified into MRG high- and low-risk groups by median risk score; TCGA-LIHC is 0.957 and ICGC-LIRI-JP is 0.967 (Table S4). Next, the MRG signature was evaluated

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2022.12.014

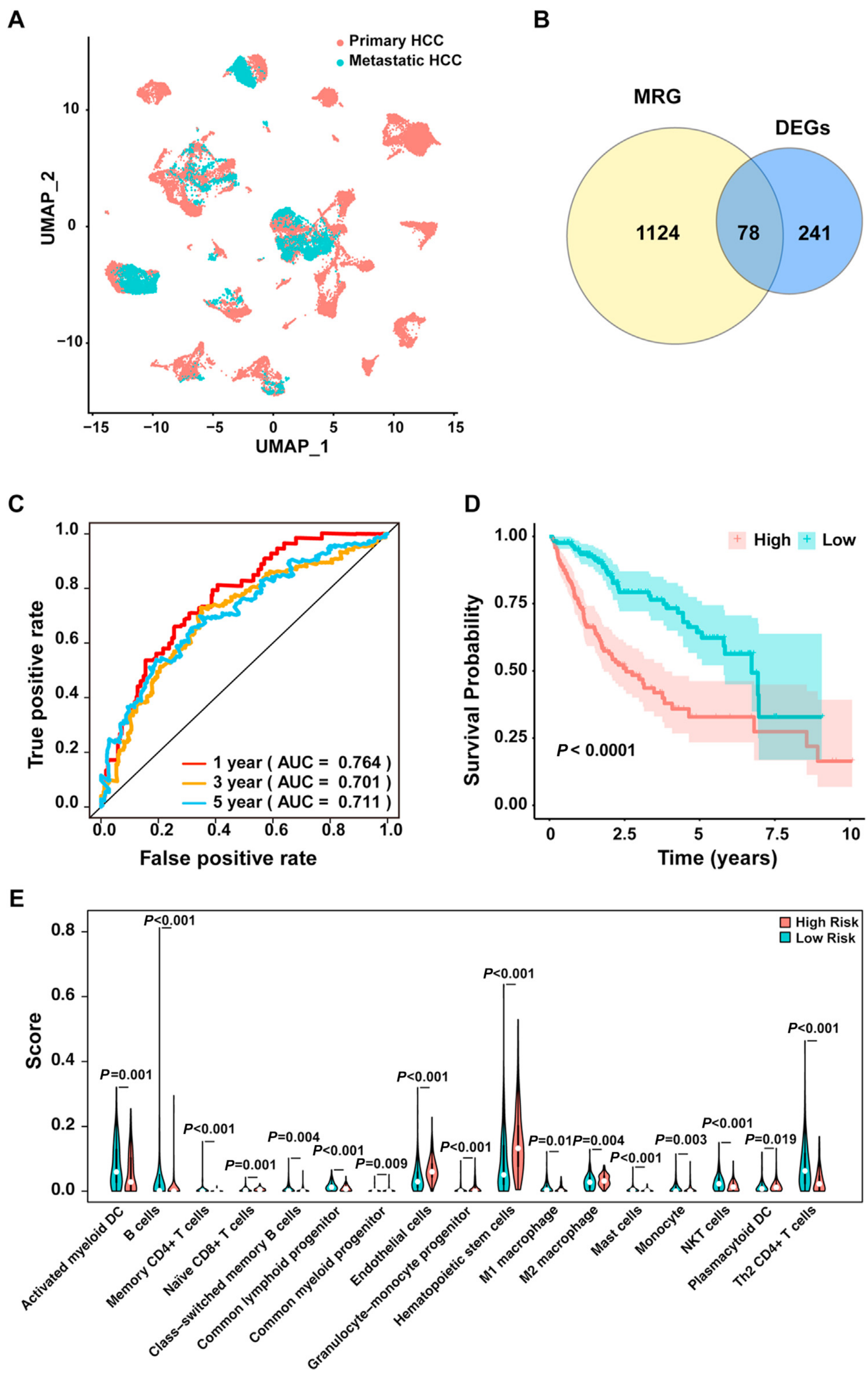

Figure 1 Development of a metabolism-related predictive model. **(A)** Characterization of single-cell RNA sequencing from primary and metastatic HCC cells and screening of marker genes. UMAP analyses were performed to analyze the single-cell RNA sequencing data GSE149614.[5] Group 1, metastatic HCC patients. Group 2, primary HCC patients. **(B)** Venn diagram showed the 78 metabolism-related genes (MRGs) which were differentially expressed between primary and metastatic HCC patients. **(C)** ROC analysis of seven core MRG genes, and their combined predictive efficiency in the TCGA-LIHC cohort. **(D)** Kaplan–Meier analysis of the possibility of overall survival in the TCGA-LIHC cohort with the MRG high- and low-risk scores. (**E**) Violin plot shows the difference in immune infiltration between the MRG high- and low-risk groups by xCell algorithm.

by ROC curves in the TCGA-LIHC cohort (Fig. 1C) and the ICGC-LIRI-JP cohort (Fig. S4A). In comparison to the patients in the MRG high-risk group, the OS rate of patients in the MRG low-risk group was significantly higher in both the TCGA-LIHC cohort (Fig. 1D) and the ICGC-LIRI-JP cohort (Fig. S4B). Survival analysis of the TCGA-LIHC cohort also indicated that the low-risk group had a higher OS value (Fig. S4C, D).

The nomogram prediction model was established by combining OS-related clinical parameters and RS. The OS-related clinical parameter was analyzed by univariate and multivariate Cox regression analyses (Fig. S5A, B). A clinically applicable nomogram forecasting individual survival probability at 1, 3, and 5 years is shown in Figure S5C. Next, calibration curves, ROC curves, and decision curve analyses were utilized for model validation. The calibration curves (Fig. S6A), ROC curves (Fig. S6B), and decision curve (Fig. S6C—E) analyses implied that the predictive power and clinical practicability of the nomogram prediction model were strong.

Furthermore, we discovered the different characteristics of the tumor mutation burden, gene set enrichment, and immune microenvironment between the MRG high- and low-risk group patients. As shown in Figure S7A and B, the genes with the most frequent mutations were *TP53* (42%) and *CTNNB1* (33%) in the MRG high- and low-risk groups, respectively. Comparing the waterfall plots of the MRG high-risk group and low-risk group, it was observed that *TP53*, *TTN*, and *MUC16* had higher frequencies in the MRG high-risk group (Fig. S7A, B). It can also be concluded that the MRG high-risk group suffered high tumor mutation burdens. The significant pathways in the different MRG risk groups were explored by Gene Set Enrichment Analysis. The top six pathways enriched in the MRG high- and low-risk groups were filtered (Fig. S8). The top six pathways that were significantly related to the progression of HCC in the MRG high-risk group were endocytosis, vasopressin-regulated water reabsorption, progesterone-mediated oocyte maturation, RNA degradation, oocyte meiosis, and pancreatic cancer (Fig. S8). In contrast, retinol metabolism, drug metabolism-cytochrome P450, fatty acid metabolism, primary bile acid biosynthesis, glycine, serine and threonine metabolism, and complement and coagulation cascades were primarily screened in the MRG low-risk group (Fig. S8 and Table S5). The immune microenvironment of the MRG high- and low-risk groups was analyzed by the xCell algorithm, and the results are shown in the violin plot in Figure 1E and Table S6. For patients in the MRG high-risk group, naïve $CD8^+$ T cells, common myeloid progenitors, endothelial cells, granulocyte-monocyte progenitors, hematopoietic stem cells, M2 macrophages, and plasmacytoid dendritic cells demonstrated high levels of infiltration (Fig. 1E). However, activated myeloid dendritic cells, B cells, memory $CD4^+$ T cells, class-switched memory B cells, common lymphoid progenitors, M1 macrophages, mast cells, monocytes, NKT cells, and Th2 $CD4^+$ T cells were positively correlated with a low RS (Fig. 1E).

In summary, through single-cell RNA sequencing and machine learning, we constructed a novel MRG prognostic model for HCC patients and evaluated its clinical application for the first time. Our prognostic model showed highly accurate risk stratification and accurate identification of HCC patients with poor subtypes, which could predict the prognosis of HCC and guide personalized treatment for HCC patients. We also discovered the different characteristics of tumor mutation burden, gene set enrichment, and immune infiltration in MRG high- and low-risk group patients. These results indicated that the underlying molecular mechanisms were markedly different, which provides a solid basis for the identification and treatment of high-risk HCC patients by the MRG signature.

Author contributions

Zuhui Pu and Lisha Mou designed the study and revised the manuscript. Lin Liu, Yumiao Qiu, and Yingying Liang performed the analysis and wrote the manuscript.

Conflict of interests

The authors declare no competing interests.

Funding

This study was supported by the Shenzhen High-level Hospital Construction Fund (2019) (China).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2022.12.014.

References

1. Cheng T, Zhan X. Pattern recognition for predictive, preventive, and personalized medicine in cancer. *EPMA J*. 2017;8(1):51—60.
2. Reig M, Forner A, Rimola J, et al. BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update. *J Hepatol*. 2022;76(3):681—693.
3. Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat Rev Genet*. 2021;22(1):3—18.
4. Zhang Q, He Y, Luo N, et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell*. 2019;179(4): 829—845.e20.
5. Lu Y, Yang A, Quan C, et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat Commun*. 2022;13(1):4594.

Lisha Mou, Lin Liu, Yumiao Qiu, Yingying Liang, Zuhui Pu *

Imaging Department, Shenzhen Institute of Translational Medicine, Health Science Center, The First Affiliated Hospital of Shenzhen University, Shenzhen Second People's Hospital, Shenzhen, Guangdong 518035, China

*Corresponding author. Shenzhen Second People's Hospital, NO.3002 Sungang Road, Futian District, Shenzhen, Guangdong 518035, China. Tel./fax: +(086) 0755 83366388 3230.
E-mail address: pupeter190@163.com (Z. Pu)

29 August 2022
Available online 23 January 2023