Identification of a Cancer Stemness Driven Prognostic Model and Therapeutic Drug Targets by Machine Learning in Hepatocellular Carcinoma

Mingchun Lai
First Affiliated Hospital Zhejiang University
Bin Xi
First Affiliated Hospital Zhejiang University
Shenyu Wei
Zhejiang Chinese Medical University
Wenjin Zhang (drzwj2002@zju.edu.cn)
First Affiliated Hospital Zhejiang University
Shusen Zheng
First Affiliated Hospital Zhejiang University

Research Article

Keywords: Hepatocellular carcinoma, cancer stem cells, stemness indices, prognostic model, immunosuppression

Posted Date: December 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1161100/v1

Abstract

Background: Hepatocellular carcinoma (HCC) is one of the most common malignancies. Cancer stem cells (CSCs), characterized by self-renewal and drug-resistance, play an important role in the development and progression of diverse cancers, but the underlying association of HCC and CSCs is not fully researched.

Methods: Transcriptome and clinical data of 903 patients in four independent HCC cohorts were obtained from TCGA, ICGC, and GEO databases. We evaluated the stemlike index for each patient to reflect the cancer stemness by using one-class logistic regression (OCLR) algorithm. GISTIC 2.0, Maftools and GSVA were used to reveal the association between the stemness index and genomic variation and biological processes in HCC. The differential expression analysis, univariate Cox analysis and LASSO analysis were used to identify the prognostic stemness signatures. The HCC stemness-related risk score (HCSRS) was constructed to quantify stemness levels of individual tumors. Based on HCSRS, the nomogram was established for HCC prognosis in a quantitative approach. Additionally, single sample Gene Set Enrichment Analysis (ssGSEA) algorithm was used to evaluate the immune infiltration levels in HCC, and drug response analysis was adopted to identify potential agents with drug sensitivity in high-HCSRS score patients.

Results: The stemness index in HCC tissues was significantly higher than that in normal tissues, and there was a significant positive correlation with pathological grade. Patients with high stemness index showed higher somatic mutation frequency, tumor mutation load, and copy number variation frequency, and were significantly enriched in tumor-related signaling pathways. Meanwhile, the 7-gene based HCSRS model that was trained and validated in 4 independent cohorts exhibited high predictive significance for overall survival (OS). Further analysis revealed that patients with high HCSRS possessed higher immunosuppression status, characterized by significantly decreased infiltration of anti-tumor immune cells (CD8 T cells, cytotoxic T cells, DC cells, NK cells, etc.) and exhausted CYT responses. At last, a total of twelve agents were identified to have potential therapeutic effects in high-HCSRS patients.

Conclusion: In current study, we systematically analyzed the potential relationship of HCC stemness with genomic variation, tumor microenvironment and biological processes, provided a theoretical basis for individualized treatment of HCC patients.

Introduction

Hepatocellular carcinoma (HCC) is the fifth most common cancer and the second cause of cancer-related death worldwide, and is an aggressive malignancy with an average five-year survival rate of only 18% [1]. Although surgical resection is the main treatment for primary focal liver tumors, the vast majority of HCC patients have reached the advanced stage at the first diagnosis due to the lack of specific symptoms, thus losing the opportunity for surgery [2]. In addition, the poor clinical prognosis of HCC patients is often attributed to the resistance of HCC cells to traditional treatment and postoperative tumor recurrence. It

has been reported that tumor recurrence occurred in nearly two thirds of patients within two years after surgery [3]. Sorafenib, as the first generation of targeted drugs, is an important treatment for patients with advanced liver cancer, but unfortunately, the vast majority of patients do not get long-term survival benefit due to the early emergence of drug resistance [4].

Stemness is defined as the potential for self-renewal and differentiation from cell origin, and making normal stem cells to grow into all cell types to constitute human body [5]. However, mounting evidence suggests that highly drug-resistant residual cells with stem cell-like properties or functions can appear at any stage of tumor progression, which are also known as cancer stem cells (CSCs) [6]. In essence, CSCs can maintain the heterogeneity of tumor cell groups through self-renewal and unlimited proliferation, which are similar to the basic features of stem cells [7]. At the same time, CSCs are insensitive to the tumor-killing effect by outside factors (chemotherapies, eg) owing to their multiple drug-resistant genetics. Therefore, tumor tissues tend to relapse within a period of time after the elimination of most common tumor cells by conventional therapies, and the ability of CSCs to move and migrate makes tumor metastasis possible [8]. The gradual loss of differentiation ability and the acquisition of stem cell-like characteristics are the main reasons driving tumor occurrence, metastasis, recurrence and drug-resistance, especially in HCC [9, 10]. Therefore, it is important to identify biomarkers associated with hepatocellular carcinoma stemness (HCS) and patient prognosis, which may provide a constructive theoretical basis for overcoming drug-resistance in HCC patients.

In recent years, computer science and bioinformatic algorithms have developed rapidly to deal with emerging big data, particularly in genome-wide, pan-cancer researches, providing new insights into CSC-linked studies [11, 12]. The machine learning-based One-class Logistic Regression (OCLR) algorithm quantified cancer stemness by obtaining an independent stemness index based on the gene expression, which is called mRNA expression-based Stemness Index (mRNAsi) [13]. This allows for a better understanding of the carcinogenic features of cancer stemness, such as mutations in oncogenes, abnormalities in specific transcriptional networks, and dysregulation of signaling pathways. Moreover, many studies are focusing on the development of novel agents targeting specific cancer types, and several databases have been established based on the test result of drug sensitivity, which makes it possible to explore tailored clinical management for HCC patients with high HCS [14]. In addition, solid tumor tissues are composed of heterogeneous cancer cells and stromal components; The extracellular matrix, neovascularization, and various cancer-related immune cells provide an appropriate environment for tumor cells to survive and evade immune surveillance, known as the tumor microenvironment (TME) [15]. Revealing the potential connection between CSCs and TME may be the key to opening a new era of tumor therapy. Taken together, it is essential to comprehensively investigate the underlying interactions between HCS, genetic mutations and TME, and to identified potential therapeutic agents in HCC.

In this study, a total of 903 HCC patients from four independent HCC cohorts were included. We systematically evaluated the stemness index of each HCC patient by OCLR algorithm. The prognostic significance and their potential biological associations with molecular characteristics, genomic instability,

biological pathways, tumor mutations and TME were further analyzed. A stable stemness-related clinical prediction model with high sensitivity was constructed based on hub genes, and its robustness in predicting overall survival (OS) of HCC patients was verified in three independent HCC cohorts. We further identified twelve potential compounds with high drug sensitivity in HCC patients with high-HCS for personalized and targeted treatment.

Results

Calculation of HCC stemness indices and its correlation with clinicopathological features

Based on mRNA expression matrix, OCLR algorithm was used to calculate the mRNAsi for each patient in the TCGA-LIHC cohort (n=365) and ICGC-LIRI cohort (n=231), respectively (Figure 1A-B). The stemness index ranging from 0 to 1, and tumor samples were divided into high and low mRNAsi groups based on the median value. First, the results showed that mRNAsi levels were significantly higher in HCC in both cohorts compared with normal tissues ($P < 0.001$) (Figure 1C). Next, we found that patients with higher mRNAsi levels suffered from worse pathological grade ($P < 0.001$) (Figure 1D), and higher mRNAsi was significantly associated with HBV infection ($P = 0.002$) (Figure 1E). Moreover, Kaplan-Meier analysis showed that patients in the TCGA-LIHC cohort with higher mRNAsi not only suffered worse OS (HR = 1.60, 95%CI: 1.33-2.26, $P = 0.007$), but were also associated with lower DFS (HR = 1.76, 95%CI: 1.26-2.44, $P = 0.002$) and PFS (HR = 1.58, 95%CI:1.18-2.13, $P < 0.001$) (Figure 1F-H). Similar results were observed in the ICGC-LIRI cohort (HR = 4.24, 95%CI: 2.31-7.78, $P < 0.001$) (Figure 1I). These results preliminarily suggest that the stemness of HCC is highly heterogeneous between normal and cancer tissues, and significantly affect the prognosis of HCC patients, so it may serve a pleuripotent role for HCC cancer cells to spread to extrahepatic metastases, tumor recurrence, as well as refractory drug resistance.

HCC patients with high mRNAsi exhibit genomic instability

Given the significant difference in survival benefit between high and low mRNAsi HCC groups, we attempted to explore the underlying biological mechanism of this difference from the perspective of genomic stability. Hence, somatic mutations annotation was firstly analyzed. The top 20 mutated genes in the low and high mRNAsi groups were identified respectively (Figure 2A-B). Subsequently, 23 genes with significant mutation differences were identified between the high and low mRNAsi groups (FDR < 0.05). In general, we observed that somatic mutations were more frequent in the high mRNAsi group (Figure 2C). Notably, TP53 as a tumor suppressor gene, was mutated most frequently in patients with high mRNAsi (39.3% vs 17.7%, $P < 0.001$). The Lollipop diagram showed the protein domain of TP53 and its specific mutation sites (Figure 2D). Additionally, Tumor mutation burden (TMB), a novel marker for predicting immunotherapy response, showed a significant positive correlation with mRNAsi level in HCC (Cor = 0.28, $P < 0.001$) (Figure 2E).

Likely, copy number variation (CNV) was also assessed. The Gistic score and distribution of CNV frequency in chromosomes were demonstrated in Figure 2F-G. It was found that at the Focal level, patients in the high mRNAsi group showed a higher copy number gain load ($P < 0.001$) (Figure 2H) and loss load ($P < 0.001$) (Figure 2I). Whereas at the Arm level, patients in the high mRNAsi group were only associated with a higher copy number loss load ($P < 0.001$) (Figure 2J) but not gain load ($P = 0.13$) (Figure 2K). These results imply a significant correlation between high stem cell characteristics and genomic instability in HCC, and high genomic variability may be one of the important reasons for poor prognosis in patients with high mRNAsi.

Next, to further explore the potential tumor heterogeneity and biological behavior differences between the high and low mRNAsi groups, we performed GSVA enrichment analysis based on transcriptome data from the TCGA-LIHC cohort. The results showed that several immunogenicity and immunoactivation-related signaling pathways were significantly enriched in patients with low mRNAsi, including "IL6/JAK/STAT3 signaling pathway", "inflammatory response", “TNF-α signaling pathway”, "IL2/STAT5 signaling pathway" et. The high mRNAsi group was mainly enriched in cell cycle regulation and cancer-promoting signaling pathways, including "G2M checkpoint", "E2F target", "MYC target V1" "MTORC1 signaling pathway", "WNT-β protein signaling pathway" (Figure 2L).

Screening of prognostic mRNAsi-related signatures in HCC

Differential analysis was performed in TCGA-LIHC and ICGC-LIRI cohorts respectively. A total of 1141 mRNAsi-related DEGs were identified in the TCGA-LIHC cohort, of which 454 genes were significantly upregulated in high mRNAsi group and 687 genes were significantly upregulated in low mRNAsi group (Figure 3A). A total of 1293 mRNAsi-related DEGs were obtained in the ICGC-LIRI cohort, of which 296 genes were significantly upregulated in high mRNAsi group and 997 genes were significantly upregulated in low mRNAsi group (Figure 3B). A total of 569 mRNAsi-related DEGs were obtained from the intersection of DEGs of the two cohorts. It was defined as the HCC stemness (HCS) related gene cluster (Figure 3C). GO enrichment analysis showed that the gene cluster was involved in several biological processes associated with immune microenvironment formation, tumor metastasis and cell cycle regulation. KEGG enrichment analysis showed that HCS related gene cluster was significantly enriched in several classic oncogenic pathways, such as "TGF-β signaling pathway", "P53 signaling pathway", "MAPK signaling pathway", "PI3K-Akt signaling pathway", "cell adhesion molecules" and "ECM-receptor interaction" (Figure 3D).

Subsequently, univariate Cox analysis was performed on the HCS-related gene cluster in the TCGA cohort, and a total of 92 prognostic HCS genes were obtained, of which the vast majority were independent risk factors in HCC, and their expression levels were significantly positively correlated with cancer stemness (Figure 3E). To further identified HCS related genes with the most prognostic potential, we performed LASSO regression analysis on these 92 genes (Figure 3F-G). Seven HCS related hub genes were identified: RAMP3, KLF2, TKT, NEIL3, CDCA8, TMEM106C, and KPNA2. The regression coefficient of each

hub gene in LASSO analysis were demonstrated in Figure 3H. These genes were also significantly correlated in regard of their co-expression levels (Figure 3I).

Construction and validation of 7-gene HCS-related panel in four HCC cohorts

To assess the clinical prediction significance of HCS-related genes, we constructed the HCC stemness-related risk score (HCSRS) model based on the expression level of hub gene and its corresponding regression coefficient: HCSRS = 0.223*KPNA2 + 0.062*CDCA8- 0.123*RAMP3 + 0.057*NEIL3 + 0.075*TMEM106C - 0.013*KLF2 + 0.047*TKT. Based on this formula, the HCSRS of each patient in TCGA-LIHC training cohort was calculated. It was showed that patients with high HCSRS exhibit worse OS than those with low HCSRS (HR=2.66, 95%CI: 1.87-3.76, P < 0.001) (Figure 4A). ROC curve was used to evaluate the clinical prediction performance of HCSRS in the training set. The AUC values of the HCSRS model in predicting 1-year, 2-year and 3-year OS were 0.8, 0.75 and 0.75 respectively (Figure 4E), indicating that the HCSRS model has an excellent clinical predictive performance. Moreover, three independent validation cohorts including ICGC-LIRI, GSE14520 and GSE116174, were used to evaluate the stability of HCSRS prognostic performance. In line with training cohort, the HCSRS still possessed well clinical significance for outcome prediction (P < 0.05) (Figure 4B-D). And ROC curve proved that HCSRS also has high predictive accuracy for 1-year, 2-year and 3-year OS in the validation cohorts (Figure 4F-H).

Univariate and multivariate Cox regression were performed to explore whether HCSRS could be used as independent prognostic factors for HCC patients. Univariate Cox analysis showed that TNM stage, mRNAsi and HCSRS were risk factors for unfavorable outcomes of HCC patients (Figure 4I). Multivariate Cox analysis suggested that only HCSRS was an independent risk factor for HCC patients (Figure 4J), and confirmed its robust predictive efficiency (OS: HR = 4.50, 95%CI: 2.57-7.88, P < 0.001). It is worth noting that although HCSRS and mRNAsi score are both poor prognostic factors for HCC, HCSRS obviously has a more powerful predictive performance compared with mRNAsi score, and the algorithm is more concise and does not rely on whole genome sequencing data.

Establishment of a nomogram based on HCSRS

To develop a mor sensitive and quantitative method for hazard classification of HCC patients, two independent prognostic factors (N stage and HCSRS) in the TCGA cohort were integrated and a nomogram was established combining patient age and TNM stage. The C-index of the established nomogram was 0.75 (95%CI: 0.71-0.78) after 1000 bootstrap iterations. The calibration plots also showed that the nomogram performed good consistency when compared of predicted survival and actually observed outcomes at 1, 3 and 5 years (Figure 5B-5D), indicating that the nomogram had strong, robust and quantitative clinical prognostic ability for the hazard classification of HCC patients.

High HCSRS score correlated with immune suppression in HCC tumor microenvironment

Previous studies have shown that cancer stemness is closely related to the formation of tumor immune microenvironment and will affect patients' response to immunotherapy to a certain extent [6]. Given the underlying associations between cancer stemness and immune infiltration, the immune cell fractions in HCC cohorts were estimated based on expression matrix using the ssGSEA algorithm. Notably, patients with high HCSRS exhibited lower immune infiltration in tumors. Several immune cells involved in antitumor immunity, such as B cells, CD8 T cells, cytotoxic T cells, DC cells, neutrophils, NK cells, T cells and helper T cells, were significantly reduced in TME in the high HCSRS group ($P < 0.01$) (Figure 6A), indicating suppressive immune tatus in HCC-TME. The CYT response is primarily responsible for adaptive anti-tumor immunity, and is characterized by significantly enhanced T cell activity [16]. We found that high HCSRS was associated with an impaired CYT response ($P < 0.01$) (Figure 6B). We also analyzed the correlation between hub genes and immune cells to verify the effect of single genes. Notably, the expression levels of RAMP3 and KLF2, which are tumor-suppressors in HCC, were significantly positively correlated with the infiltration of anti-tumor immune cells, whereas the other five genes showed negative correlation. In general, HCC patients with high HCSRS have a poor survival prognosis, characterized by higher cancer stemness, more cancer-related signaling activation, greater genomic instability, lower immune infiltration, and lower CYT response (Figure 6F).

At last, the expression of these genes at the transcriptional and protein level were analyzed through TCGA expression matrix and THPA database (among which the immunohistochemical data of RAMP3, KLF2 and NEIL3 were missing). Compared with normal liver tissues, the protein levels of CDCA8, KPNA2, TKT and TMEM106C were significantly higher in HCC tumor tissues, which was consistent with their mRNA expression levels (Figure 7A-B). These results suggest that the HCS-related genes are not only involved in the occurrence and development of HCC, but also can be potential therapeutic targets for HCC.

Screening of potential therapeutic agents for high-HCSRS score HCCs

CTRP and PRISM were cancer cell-based, comprehensive data-driven resource for exploring relationships across chemicals, genes and cancer cells. Hence, we intended to screen for potential compounds, specifically targeting the stemness signature in HCC. Candidate therapeutic agents with drug sensitivity in high-HCSRS patients were identified based on drug response data derived from online resource. Two-step analyses were performed based on each database, respectively. First, differential drug sensitive analysis between high-HCSRS score (top decile) and low-HCSRS score (bottom decile) groups was conducted. Compounds with lower estimated AUC values in high-HCSRS score group were identified ($log2FC > 0.10$). Next, compounds with negative correlation coefficient were selected by Spearman correlation analysis between HCSRS score and AUC value (Spearman' $r < -0.30$). As a result, four CTRP-

derived compounds (including BI-2536, KX2-391, paclitaxel and SB-743921) and eight PRISM-derived compounds (docetaxel, epothilone-b, gemcitabine, irinotecan, LY2606368, rubitecan, sirolimus and topotecan) were obtained. All these compounds had a negative correlation with HCSRS and lower estimated AUC values in high-HCSRS group (Figure 8A and B).

Discussion

HCC is an important cause of cancer-related death, which is partly attributed to its resistance to traditional treatment and high recurrence rate after surgery [3]. Although CSCs are only a small subset of tumor cells, they have attracted much attention in recent years due to their ability to self-renew, differentiate and promote tumor recurrence and metastasis [7]. We employed large samples with expression matrix to calculate the stemness-related indices and found high mRNAsi was correlated with classic HCC features, including higher pathological grade, HBV infection and somatic mutations of driver genes such as TP53, AXIN1 and MUC17. The mRNAsi level increased as the tumor pathological grade increase, and the G3/G4 tumor have the highest stemness indices, consistent with their undifferentiated pathological characteristics. Moreover, mRNAsi were tightly associated with OS, DFS and PFS in HCC, possessing fine quantification and fitness of stemlike-features. We identified the stemness-related most characteristic 7 signatures, and we focused on the 7-gene based model, which was trained and validated in 903 HCC patients to possess superior prediction performance in 1-, 2-, and 3-year OS. We also established a nomogram integrating stemness scores, which exhibited well calibrated accuracy and clinical value. Furthermore, we inferred the abundance of immune infiltrating cells in HCC-TME. Notably, we observed that high HCSRS was significantly correlated with lower infiltration of immune-activated cells, especially cytotoxic cells, CD8+ T cells, NK cells as well as CYT responses. As last, we screened out the potential therapeutic agents targeting HCC patients with high stemness.

The OCLR algorithm was selected in this study to quantitatively evaluate the stemness indices of HCC samples. Compared with traditional machine learning-based predictors, OCLR does not penalize the misclassification of stem cell-derived progenitor cells at different stages of differentiation, and exhibited equivalent performance with a more flexible formulation. We thus derived a distinct stemness metric at the transcriptional level based on OCLR algorithm, and selected mRNAsi for prognostic analysis. Stemness signatures had been reported in other malignancies with different predictive values such as breast cancer, colon cancer, and AML. As far as our information goes, the 7-gene signature has not been previously reported in other studies that was convenient and helpful for clinical decision making. Interestingly, recent studies have shown that the high expression of RAMP3 is related to the better prognosis of HCC patient post-surgery [17]. RAMP3 can also inhibit glycolysis and promote ROS activation and apoptosis of P53 deficient cancer cells under pramlintide treatment [18]. In this study, RAMP3 significantly correlated with the anti-tumor immune cells in HCC TME, which include CD8 T cells, cytotoxic cells, DC cells and NK cells, implying its potential as a drug target in tumor treatment. Meanwhile, other signatures were also reported involved in tumorigenesis and stemness, like KLF2, TKT, and CDCA8 [19-22]. The functional enrichment analysis further demonstrated that wnt-β-catenin, MYC, MTORC, P53 and other signaling pathways were significantly enriched in HCC patients with high cancer

stemness. Overall, the most advantage of prognostic genes is for clinical judgment of caner stemness and guiding personalized treatment.

Previous studies have just reported that more than twenty cancer genes were frequently mutated in HCC, driving tumor occurrence and development, which include TP53, CCTNB1 and AXIN1 [23]. As one of the most important tumor suppressor genes, TP53 can inhibit the expression of a variety of stem cell factors by activating miR-34a and miR-145, thus preventing differentiated cells from backsliding to pluripotency [24-27]. In line with the previous findings, we found that both TP53 and AXIN1 were more frequently mutated in high mRNAsi samples. Furthermore, the TMB and distribution of copy number variation were also tightly associated with mRNAsi. Dormant CSCs may be activated by a series of gene mutations and undergo a large number of gene sequence changes. The accumulation of these mutations will also drive the immune escape of CSCs and drug resistance, and the activation of CSCs will further promote the increase of cancer stemness and thus promote genomic instability [28]. Nevertheless, large samples are still warranted to further verify the potential correlations between mutational burden and stemness indices in the HCC cohort.

Recently, a pan-cancer analysis revealed that lower immune infiltration is widely associated with the stem cell-like cancer phenotypes across twenty-one types of solid tumors [29]. Based on the expression data of four HCC cohorts, we systematically explored the general landscape of infiltrating-immune cells in HCC TME regulated by cancer stemness. We observed that the HCSRS was negatively correlated with infiltrating levels of anti-tumor immune cells, mainly the adaptive immune cells, like CD8+ T cells, B cells and cytotoxic cells. Meanwhile, two of the stemness signatures, RAMP3 and KPNA2 were both positively correlated with activated infiltrating immune cells. KLF2 was reported to play a central role in regulating immune cell activation; Down-regulation of KLF2 maintains T cells in a resting state, inhibits monocyte activation and continuously activate regulatory T cells [30]. A recent study reported that KLF2 and Tfcp2li synergically inhibit cell pluripotency [31]. Interestingly, KLF2 and Tfcp2li are two key downstream targets of the Wnt-β-catenin pathway, which mediate stem cell self-renewal [30]. In addition, KLF2 has been confirmed to inhibit the growth, migration and colony formation of HCC cells, and it can compete with Gli1 by binding to HDAC1 to inhibit the activation of Hedghog signaling pathway [19]. The Wnt-β-catenin and Hedghog signaling pathways are the most important activation pathways of cancer stem cells [6]. Lastly, we analyzed the dug sensitivity in cancer cell lines and proposed twelve candidate compounds with higher therapeutic efficacy in high HCSRS cohorts.

However, there are several limitations in this study requiring further optimization. Although validated in four independent cohorts, it still remains unclear whether the 7-gene based signature could be generalized into all the HCC samples to quantitatively evaluate stemness and precisely predict patient survival. Lager samples and multicenter cohorts should be included in validation. Some compounds are already tested in safety and are being used in clinic. Whereas other new agents require specific preclinical validations to assess their real inhibitory effects.

Conclusion

Overall, through the multidisciplinary approach of computer science and life science, and based on the large-scale HCC research cohort, this study comprehensively and systematically analyzed the potential relationship between the characteristics of HCC stem cells and genetics, clinical prognostic characteristics, biomolecular pathways and tumor microenvironment. Seven HCC stemness-related hub genes were screened out, and a stemness-related clinical prediction model was constructed, which can quantitatively predict the death risk of patients, identify tumor heterogeneity and evaluate patients' individual cancer stemness level. Additionally, chemical compounds specifically targeting the cancer stemness were proposed. These results will provide a promising theoretical basis for the individualized treatment of HCC patients from the perspective of cancer stemness.

Materials And Methods

Data acquisition and preprocessing

Four independent HCC cohorts were obtained from publicly available databases and matched integrated clinical annotations. The patients with incomplete survival information were excluded. For TCGA-LIHC cohort, a total of 365 HCC samples and 50 normal liver tissue samples were included. The corresponding expression profiles, somatic mutation data and copy number variation (CNV) data, as well as detailed clinicopathological information, are obtained from the Genomic Data Commons (GDC) platform using the 'TCGAbiolink' R package [32]. To standardize the data and increase comparability between different cohorts, we converted the FPKM value of RNA-seq data into transcripts per kilobase million (TPM) value [33]. The Ensemble ID of the transcript was converted into Gene Symbols through GENCODE Gene annotation file (GRCH38, Version 22), and the relative expression level of genes was log2 (TPM + 1) to narrow the range of values. And the standardized gene expression files and clinicopathological information of ICGC-LIRI-JP with 231 HCC samples and 202 normal liver tissue were obtained based on Illumina HiSeq RNA sequencing platform. Besides, the raw "CEL" files (microarray datasets based on Affymetrix@) of GSE14520 and GSE116174 were obtained from the Gene Expression Omnibus and carried out for background correction and quantile standardization using Robust Multi-array Average (RMA) algorithm [34]. The GSE14520 cohort included 242 HCC tissue samples, and the GSE116174 cohort included 65 HCC tissue samples, both of which contained complete clinical survival data.

Calculation of stemness indices for HCC

Based on Illumina Human Methylation 450 platform, we obtained 99 human stem/progenitor cells from the Progenitor Cell Biology Consortium (PCBC) database. The stem cell dataset included 4 embryonic stem cells, 40 induced pluripotent stem cells, 22 stem cell-driven embryoids, and 33 stem cell-driven mesoderm, endoderm, and ectoderm. A machine learning algorithm OCLR was used to train these datasets and calculate the stemness indices (mRNAsi) for each patient in TCGA-LIHC and ICGC-LIPI cohort based on expression matrix [35]. The patients were divided into the high and low mRNAsi groups

according to the median value of stemness indices. Kaplan-Meier survival analysis was used to evaluate the survival difference between the two groups and draw the survival curve.

Evaluation of mRNAsi with genetic variation in HCC

The Mutation Annotation Format (MAF) file in GDC was used to analyze and visualize the somatic mutation data by performing 'Maftools' R package [36]. Tumor mutation burden (TMB) was defined as the sum of somatic gene coding errors, base substitutions, gene insertion or deletion errors detected per megabase (Mb). We calculated TMB for each patient in the TCGA-LIHC cohort using the following formula: the absolute number of somatic mutations with non-synonymous mutations/exon length (38 Mb). In addition, we identified genome-wide regions with significant amplification or deletion changes by GISTIC 2.0, a biological program based on robust computational algorithms to detect somatic CNV by evaluating the frequency and magnitude of corresponding events [37]. The absolute number of genes with copy number changes in Focal and Arm levels is copy number loss load or copy number gain load [38].

Screening of HCC stemness-related signatures

The ‘Limma’ R package was used for differential expression genes (DEG) analysis between high and low mRNAsi groups under the threshold of | log2 (fold change) | > 0.5 and False Discovery Rate (FDR) < 0.001. The intersection of DEGs in the two cohorts was defined as HCS related genes. Finally, we used univariate Cox analysis to screen and check prognostic HCS related panel in HCC.

Functional enrichment analysis

Gene set variation analysis (GSVA) was performed on the whole genome transcription data of the TCGA-LIHC cohort using the 'GSVA' R package in high and low mRNAsi groups. GSVA is a nonparametric and unsupervised algorithm commonly used to assess changes in biological processes or signaling pathway activity of gene expression matrix [39]. The gene set 'h.all.v7.4.symbols.gmt' used in this analysis was obtained from The Molecular Signatures Database (MSigDB). In addition, 'clusterProfiler' R package was used to perform Gene ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis to reveal the biological processes and signaling pathways related to the occurrence and development of HCC mediated by HCS [40].

Construction and validation of an HCS-related clinical risk model

TCGA-LIHC cohort was taken as a training set and risk model was constructed using LASSO algorithm. The other three cohorts were used for model validation. We used the 'glmnet' R package to fit LASSO Cox regression to punish the weights of clinical model parameters, and used 10x cross-validation to determine the optimal penalty parameter λ in the training set to generate a sparse parameter space. The gene λ corresponds to the most characteristic HCS-related prognostic signature. Consequently, based on these signatures, we multiplied the normalized gene expression level *(exp)* with its regression coefficient *(β)* to obtain the hepatocellular carcinoma stemness-related risk score (HCSRS) model:

$$HCSRS = \sum_{i=1}^{n} exp_i * \beta_i$$

Evaluating the associations between tumor microenvironment and HCSRS in HCC

The single sample gene-set enrichment analysis (ssGSEA) was applied to explore the different infiltration proportion of 24 types of immune cells, immune related signatures and pathways in TCGA-LIHC expression profile by using the “GSVA” R package. Gene markers of these immune cells and immune related signature were obtained from the study by Bindea and Mariathasan et al [41]. In addition, the geometric average expression levels of GZMA and PRF1 was used to represent cytotoxic activity score (CYT) [16]. Immunohistochemical data from The Human Protein Altas (THPA) database were used to analyze the protein expression of HCS-related genes in HCC.

Construction and validation of HCSRS based nomogram

Independent risk factors identified by multivariate Cox regression analysis were used to construct a nomogram to quantitatively predict the OS of HCC patients. Nomograms integrate multiple predictive indicators to score the outcome variables according to the scale corresponding to each predictive indicator, and finally evaluate the probability of outcome events of a patient according to the total score. Its construction and validation strictly follow the research guidelines of Lasonos et al [42]. The calibration graph was used to verify the accuracy of the clinical prediction performance of the nomogram, and a C-index was used to evaluate the consistency between the actual result frequency and the model-predicted frequency. Nomogram and calibration diagram were drawn by 'Rms' R package.

Identification of potential therapeutic agents

Expression matrix and somatic mutation profiles of more than 700 human cancer cell lines were obtained from the online database Cancer Cell Line Encyclopedia (CCLE) (https://sites.broadinstitute.org/ccle/datasets) [14]. Drug sensitivity data of compounds over cancer cell lines were provided by PRISM Repurposing dataset (https://depmap.org/portal/prism/) and Cancer Therapeutics Response Portal (https://portals.broadinstitute.org/ctrp), respectively. Each datasets use the area under the dose-response curve (based on area under the curve (AUC)) value as a measure of drug sensitivity in specific cell lines. Lower AUC value indicates increased drug sensitivity. Missing AUC values were imputed by the application of K-nearest neighbor (k-NN) imputation. Compounds with more than 20% of missing values were excluded before imputation. Since the cell lines in both datasets were obtained from the CCLE project, biomolecular data in CCLE were thus used for subsequent PRISM and CTRP analyses.

Statistical analysis

Unpaired Student's -T test (for normally distributed variables) and Mann-Whitney U test (for non-normally distributed variables) were used to compare the differences between two groups. Nonparametric Kruskal-Wallis test was used to compare differences between three groups and more. Benjamini-Hochberg for multiple corrections of results. Spearman correlation analysis was used to calculate the correlation between HCC immune cell infiltration level and HCS-related gene expression level. Kaplan-Meier analysis was used to draw survival curves, and log-rank test was used to determine the statistical significance of survival differences. Univariate and multivariate Cox analyses were used to evaluate the independent prognostic ability of HCSRS in HCC patients. Receiver operating characteristic curve (ROC) and its Area under curve (AUC) were used to evaluate the sensitivity and specificity of HCSRS prognostic model. All data analysis was implemented using R software (version 3.6.1), and the $P < 0.05$ indicated statistically significant.

Abbreviations

HCC: Hepatocellular carcinoma; CSCs: Cancer stem cells; OCLR: One-class logistic regression; HCSRS: HCC stemness-related risk score; ssGSEA: Single sample Gene Set Enrichment Analysis; OS: Overall survival; mRNAsi: mRNA expression-based Stemness Index; TME: Tumor microenvironment; CNV: Copy number variation; GDC: Genomic Data Common; TPM: Transcripts per kilobase million; RMA: Robust Multi-array Average; PCBC: Progenitor Cell Biology Consortium; MAF: Mutation Annotation Format; TMB: Tumor mutation burden; FDR: False Discovery Rate; DEG: Differential expression gene; GSVA: Gene set variation analysis; MSigDB: Molecular Signatures Database; GO: Gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genome; CYT: Cytotoxic activity score; THPA: The Human Protein Altas; CCLE: Cancer Cell Line Encyclopedia; AUC: Area under the curve; ROC: Receiver operating characteristic curve;

Declarations

Acknowledgements

Not applicable.

Authors’ contributions

WJZ and SSZ raised the concept, designed this study and supervised the entire process. SYW analyzed the data. MCL and BX drafted the manuscript, MCL and WJZ revised the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by Natural Science Foundation of Zhejiang Province (LQ20H030008).

Availability of data and materials

The TCGA-LIHC and LIRI-JP cohort datasets used can be found in the TCGA (https://portal.gdc.cancer.gov/) and ICGC databases (https://icgc.org/icgc/cgp/66/420/824), respectively. GSE14520 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14520) and GSE116174 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116174) can be obtained from the Gene Expression Omnibus. The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.
2. European Association for the Study of the Liver. Electronic address eee, European Association for the Study of the L. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236.
3. Kulik L, El-Serag HB. Epidemiology and Management of Hepatocellular Carcinoma. Gastroenterology. 2019;156(2):477-91 e1.
4. Cheng AL, Kang YK, Chen Z, Tsao CJ, Qin S, Kim JS, et al. Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. Lancet Oncol. 2009;10(1):25-34.
5. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. Mol Cancer. 2017;16(1):41.
6. Clara JA, Monge C, Yang Y, Takebe N. Targeting signalling pathways and the immune microenvironment of cancer stem cells - a clinical update. Nat Rev Clin Oncol. 2020;17(4):204-32.
7. Nassar D, Blanpain C. Cancer Stem Cells: Basic Concepts and Therapeutic Implications. Annu Rev Pathol. 2016;11:47-76.
8. Saygin C, Matei D, Majeti R, Reizes O, Lathia JD. Targeting Cancer Stemness in the Clinic: From Hype to Hope. Cell Stem Cell. 2019;24(1):25-40.
9. Seguin L, Desgrosellier JS, Weis SM, Cheresh DA. Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance. Trends Cell Biol. 2015;25(4):234-40.
10. Lee TK, Guan XY, Ma S. Cancer stem cells in hepatocellular carcinoma - from origin to clinical implications. Nat Rev Gastroenterol Hepatol. 2021.
11. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82-93.
12. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101.
13. Sokolov A, Paull EO, Stuart JM. One-Class Detection of Cell States in Tumor Subtypes. Pac Symp Biocomput. 2016;21:405-16.
14. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503-8.
15. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
16. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1-2):48-61.
17. Nault JC, De Reynies A, Villanueva A, Calderaro J, Rebouissou S, Couchy G, et al. A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. Gastroenterology. 2013;145(1):176-87.

18. Venkatanarayan A, Raulji P, Norton W, Flores ER. Novel therapeutic interventions for p53-altered tumors through manipulation of its family members, p63 and p73. Cell Cycle. 2016;15(2):164-71.
19. Lin J, Tan H, Nie Y, Wu D, Zheng W, Lin W, et al. Kruppel-like factor 2 inhibits hepatocarcinogenesis through negative regulation of the Hedgehog pathway. Cancer Sci. 2019;110(4):1220-31.
20. Dettling S, Stamova S, Warta R, Schnolzer M, Rapp C, Rathinasamy A, et al. Identification of CRKII, CFL1, CNTN1, NME2, and TKT as Novel and Frequent T-Cell Targets in Human IDH-Mutant Glioma. Clin Cancer Res. 2018;24(12):2951-62.
21. Dai Y, Tang Y, He F, Zhang Y, Cheng A, Gan R, et al. Screening and functional analysis of differentially expressed genes in EBV-transformed lymphoblasts. Virol J. 2012;9:77.
22. Dai C, Miao CX, Xu XM, Liu LJ, Gu YF, Zhou D, et al. Transcriptional activation of human CDCA8 gene regulated by transcription factor NF-Y in embryonic stem cells and cancer cells. J Biol Chem. 2015;290(37):22423-34.
23. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47(5):505-11.
24. Jain AK, Allton K, Iacovino M, Mahen E, Milczarek RJ, Zwaka TP, et al. p53 regulates cell cycle and microRNAs to promote differentiation of human embryonic stem cells. PLoS Biol. 2012;10(2):e1001268.
25. Chiche A, Moumen M, Petit V, Jonkers J, Medina D, Deugnier MA, et al. Somatic loss of p53 leads to stem/progenitor cell amplification in both mammary epithelial compartments, basal and luminal. Stem Cells. 2013;31(9):1857-67.
26. Hanna J, Saha K, Pando B, van Zon J, Lengner CJ, Creyghton MP, et al. Direct cell reprogramming is a stochastic process amenable to acceleration. Nature. 2009;462(7273):595-601.
27. Hong H, Takahashi K, Ichisaka T, Aoi T, Kanagawa O, Nakagawa M, et al. Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. Nature. 2009;460(7259):1132-5.
28. Visvader JE, Lindeman GJ. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer. 2008;8(10):755-68.
29. Miranda A, Hamilton PT, Zhang AW, Pattnaik S, Becht E, Mezheyeuski A, et al. Cancer stemness, intratumoral heterogeneity, and immune response across cancers. Proc Natl Acad Sci U S A. 2019;116(18):9020-9.
30. Jha P, Das H. KLF2 in Regulation of NF-kappaB-Mediated Immune Cell Function and Inflammation. Int J Mol Sci. 2017;18(11).
31. Hall J, Guo G, Wray J, Eyres I, Nichols J, Grotewold L, et al. Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal. Cell Stem Cell. 2009;5(6):597-609.
32. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.

33. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020;26(8):903-9.
34. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.
35. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018;173(2):338-54 e15.
36. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28(11):1747-56.
37. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.
38. Shen R, Li P, Li B, Zhang B, Feng L, Cheng S. Identification of Distinct Immune Subtypes in Colorectal Cancer Based on the Stromal Compartment. Front Oncol. 2019;9:1497.
39. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
40. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
41. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39(4):782-95.
42. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26(8):1364-70.

Figures

Figure 1

Association of stemness index and clinicopathological features in HCC. (A-B) Overview of mRNAsi and clinical features in TCGA-LIHC and ICGC-LIRI cohorts. (C) The mRNAsi was significantly higher in HCC than that in normal tissues. (D) There was a significant positive correlation between mRNAsi and pathological grade of HCC. (E) HBV-positive HCC patients had a higher mRNAsi than HBV-negative HCC patients. (F-H) The relationship between mRNAsi and OS, PFS and DFS in TCGA-LIHC cohort. (I) The relationship between mRNAsi and OS in ICGC-LIRI cohort.

Figure 2

Association between the mRNAsi and genomic instability in HCC. (A-B) Waterfall diagram shows the 20 most common mutated genes and their specific mutation forms in the low and high mRNAsi groups respectively. (C) Genes with the most significant mutation difference between high and low mRNAsi groups. (D) Protein domain of TP53 and detailed mutation sites in high and low mRNAsi groups. (E) Correlation between HCC mRNAsi and tumor mutation load. (F-G) Gistic score and variation frequency of copy number in autosomes in high and low mRNAsi groups. Red represents gain and blue represents loss. (H-I) At Focal level, difference of copy number gain and loss load between high and low mRNAsi groups. (J-K) At Arm level, difference of copy number gain and loss load between high and low mRNAsi groups. (L) Heat map shows enriched biological processes with significant difference in groups with high and low mRNAsi, with gold indicating positive correlation and blue indicating negative correlation.

Figure 3

Screening of prognostic mRNAsi-related signatures in HCC. (A-B) Differentially expressed genes between the high and low mRNAsi groups in TCGA and ICGC cohorts. The red and green dots represent significantly up-regulated genes in high and low mRNAsi groups, respectively. (C) Venn diagram shows the intersection of 569 differentially expressed genes in two HCC cohorts. (D) The biological processes and KEGG analysis of the 569 mRNAsi-related genes. (E) Univariate Cox analysis of 569 genes in the TCGA LIHC cohort, 19 of which are negatively correlated with mRNAsi (white square) and improved prognosis (green square); Seventy-three genes are positively associated with mRNAsi (black square) and with poor prognosis (red square). (F-G) LASSO regression identifies seven of the most potential mRNAsi-related prognostic genes, and 10x cross-validation. (H) Distribution of regression coefficients of mRNAsi-related prognostic hub genes. (I) Correlation between expression levels of hub genes, red represents positive correlation, green represents negative correlation.

Figure 4

Training and validation of 7-gene HCS-related risk model. (A) TCGA-LIHC (training set) was used to construct the HCC stemness-related risk score (HCSRS) model, and the patients with high HCSRS suffered worse OS than that with low HCSRS. (B-D) ICGC-LIRI, GSE14520 and GSE116174 function as external validation sets to evaluate the performance of the risk model, and the OS of high-risk patients is statistically worse. (E-H) ROC curve was used to evaluate the sensitivity and specificity of HCSRS in predicting patient OS in the training set and validation set. (I-J) Univariate and multivariate Cox analysis revealed that HCSRS is an independent risk factor influencing the prognosis of HCC patients.

Figure 5

Nomogram integrating HCC stemness-related risk score. (A) Nomogram was constructed to quantitatively predict the death risk of HCC patients at 1, 3 and 5 years. (B-D) Calibration curve to evaluate the predictive accuracy of nomogram.

Figure 6

Association between HCC stemness and tumor microenvironment. (A) The immune infiltration levels between patients in the high and low HCSRS groups. (B) CYT responses between patients in the high and low HCSRS groups. (C) The correlation between the expression levels of HCS-related hub genes and the infiltrating levels of immune cells, red represents positive correlation, blue represents negative correlation, and color depth represents correlation strength. (D) Mulberry plot shows the association between HCSRS, stem cell index, patient survival, and levels of immune infiltration. HCSRS, hepatocellular carcinoma stemness-related risk score; CYT, cytotoxic activity score.

Figure 7

Expression of HCS-related hub genes in tumor tissues. (A) Differences in the expression levels of hub genes between HCC and normal tissues. (B) Expression levels of HCS-related hub genes at protein level between tumor and normal tissues was analyzed based on immunohistochemical results of THPA database.

Figure 8

Screening of potential therapeutic agents for high-HCSRS score HCCs. (A) Spearman's correlation analysis and differential drug sensitivity analysis of four CTRP-derived compounds. (B) Spearman's correlation analysis and differential drug sensitivity analysis of eight PRISM-derived compounds.