



OPEN ACCESS

EDITED BY

Jian-Guo Zhou,
University of Erlangen Nuremberg,
Germany

REVIEWED BY

Ankur Sharma,
Harry Perkins Institute of Medical
Research, Australia
Andrew Dellinger,
Elon University, United States

*CORRESPONDENCE

Wenzhe Gao,
gaowenzhexy@163.com
Jichun Sun,
sunjichun_1978@163.com
Hongwei Zhu,
zhw_0509@yeah.net

[†]These authors have contributed equally
to this work and share first authorship

SPECIALTY SECTION

This article was submitted to Cancer
Genetics and Oncogenomics,
a section of the journal
Frontiers in Genetics

RECEIVED 10 February 2022

ACCEPTED 12 August 2022

PUBLISHED 24 October 2022

CITATION

Li W, Liu J, Zhu W, Jin X, Yang Z, Gao W,
Sun J and Zhu H (2022), Identification of
biomarkers for hepatocellular
carcinoma based on single cell
sequencing and machine
learning algorithms.
Front. Genet. 13:873218.
doi: 10.3389/fgene.2022.873218

COPYRIGHT

© 2022 Li, Liu, Zhu, Jin, Yang, Gao, Sun
and Zhu. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Identification of biomarkers for hepatocellular carcinoma based on single cell sequencing and machine learning algorithms

Weimin Li^{1,2†}, Jixing Liu^{1†}, Wenjuan Zhu³, Xiaoxin Jin⁴, Zhi Yang⁵,
Wenzhe Gao^{1*}, Jichun Sun^{1*} and Hongwei Zhu^{1*}

¹Department of Hepatobiliary and Pancreatic Surgery, The Third Xiangya Hospital, Central South University, Changsha, Hunan, China, ²School of Information, Hunan University of Humanities, Science and Technology, Loudi, China, ³Division of Nephrology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong, China, ⁴The Second Xiangya Hospital, Central South University, Changsha, Hunan, China, ⁵Xiangya Hospital, Central South University, Changsha, Hunan, China

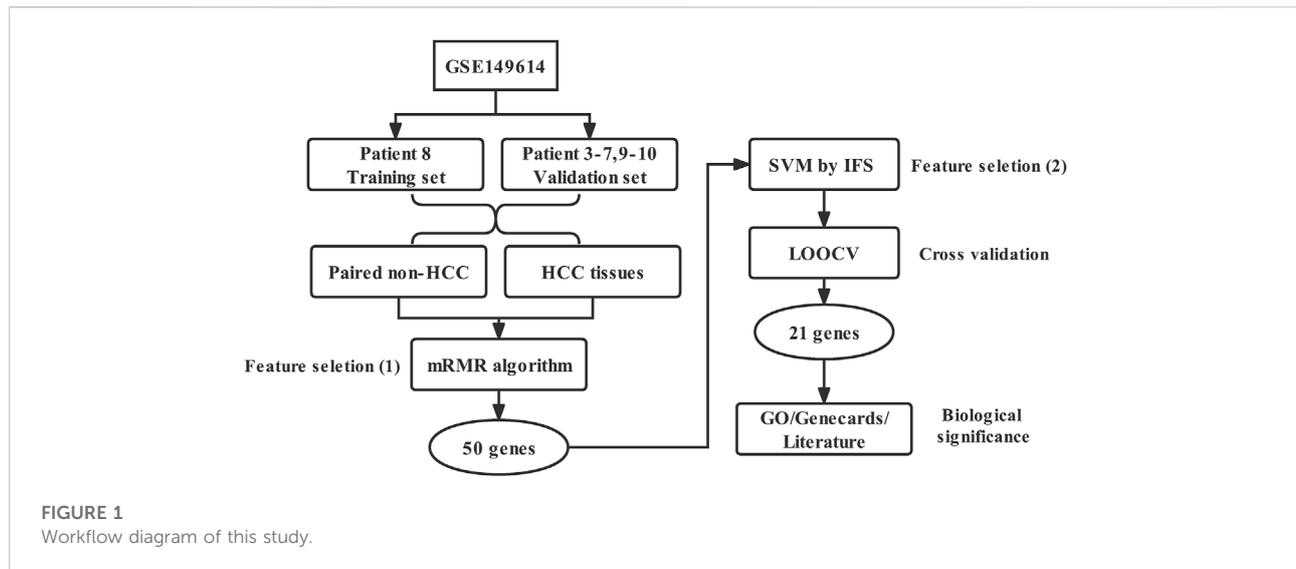
Hepatocellular carcinoma (HCC) remains one of the most lethal cancers around the world. Precision oncology will be crucial for further improving the prognosis of HCC patients. Compared with traditional bulk RNA-seq, single-cell RNA sequencing (scRNA-seq) enables the transcriptomes of a great deal of individual cells assayed in an unbiased manner, showing the potential to deeply reveal tumor heterogeneity. In this study, based on the scRNA-seq results of primary neoplastic cells and paired normal liver cells from eight HCC patients, a new strategy of machine learning algorithms was applied to screen core biomarkers that distinguished HCC tumor tissues from the adjacent normal liver. Expression profiles of HCC cells and normal liver cells were first analyzed by maximum relevance minimum redundancy (mRMR) to get a top 50 signature gene feature. For further analysis, the incremental feature selection (IFS) method and leave-one-out cross validation (LOOCV) were conducted to build an optimal classification model and to extract 21 potentially essential biomarkers for HCC cells. Our results provided new insights into HCC pathogenesis that might be valuable for HCC diagnosis and therapy.

KEYWORDS

hepatocellular carcinoma, ScRNA-seq, biomarkers, MRMR algorithm, support vector machine

Introduction

Hepatocellular carcinoma (HCC), with an annual incidence of 8.3 per 1,00,000 in population around the world, remains one of the most lethal malignancies in the digestive system. It is estimated that the 5-years survival rate for HCC patients is 18%, only a little bit higher than pancreatic cancer among all cancers, indicating that HCC is still one of the worst prognostic tumors worldwide (Siegel et al., 2020). Fortunately, with the development of modern cancer therapies, which integrate diverse neoadjuvant and adjuvant strategies with classic surgical resection, the survival rate of HCC has been



gradually improving in the past few decades (McGlynn et al., 2020). However, the highly heterogeneous nature of HCC determines that a large proportion of patients receiving standardized treatment will inevitably relapse (Petrowsky et al., 2020). Thus, precision oncology, including novel predictive and therapeutic oncogenetic markers, signals in tumor immune microenvironment and microbiome, etc. will be crucial for further improving the prognosis of HCC patients (Nault and Villanueva, 2020).

Tumor heterogeneity is the biggest obstacle to the development of precision treatment for HCC, which is decided by heterogeneous HCC cells, a changeful, complex microenvironment, and their involuted interaction. With the rapid development of profiling technology, bulk DNA and RNA sequencing have provided a lot of information about molecular phenotypes and evolutionary characteristics of HCC. In 2020, Amanda J. Craig and her colleagues reviewed the most important and common genetic alterations of HCC, including mutations in the TERT promoter, TP53, and CTNNB1, copy number variations in multiple genes, and aberrations in DNA methylation at the genome level (Craig et al., 2020). Rebouissou and Nault (2020) discovered signal pathways that are frequently altered in HCC patients included telomere maintenance, including Wnt/ β -Catenin, P53/cell cycle regulation, oxidative stress, epigenetic modifiers, AKT/mTOR signaling, and MAPK pathway. These findings have generated classification schemas of HCC molecular subtypes. However, these results still encounter many difficulties in real-world clinical applications. On the one hand, bulk sequencing could only detect the average condition of gene alterations or expression status, unable to distinguish the gene expression signature of diverse cells in cancer samples, or get a

categorized gene feature between tumors and normal tissues. On the other hand, limited by sample size and traditional methods for differential expression analysis (like Limma and EdgeR), results of bulk RNA-seq often lack stability and repeatability among different batches of experiments. These drawbacks reduce the potential practical value of bulk RNA-seq results.

The advent of single-cell RNA sequencing (scRNA-seq) is a revolutionary development in the field of profiling. Since proposed by Professor F Tang in 2009 (Tang et al., 2010), researchers have immediately realized the infinite value of this technology. It enables the transcriptomes profiling of a lot of individual cells assayed in an unbiased manner, allowing researchers to sort and study the specific characteristics of a single cell or a group of cells individually (Stegle et al., 2015). This technology perfectly fits the research on cancer that has innegligible internal heterogeneity. Thus, scRNA-seq for tumor research has sprung up in the past decade, it has been applied in multiple cancers including HCC. Zhang et al. (2019) detected HCC specific immune cells for five HCC patients using scRNA-seq and discovered that CD45 immune cells, LAMP3(+) dendritic cells, and tumor-associated macrophages were specific infiltrating immune cells in HCC and were associated with patients' poor prognosis. Ho et al. (2019) grouped HCC stem cells for two subgroups through scRNA-seq according to the expression of EPCAM; they also identified a CD24/CD44-enriched cell subpopulation within the EPCAM(+) cells which might indicate a novel stemness-related cell subclone of HCC. These studies illustrate the significance of scRNA-seq for deeply understanding the evolutionary differences among HCC patients, the heterogeneity between HCC tissues and normal livers, between HCC parenchymal cells and microenvironmental

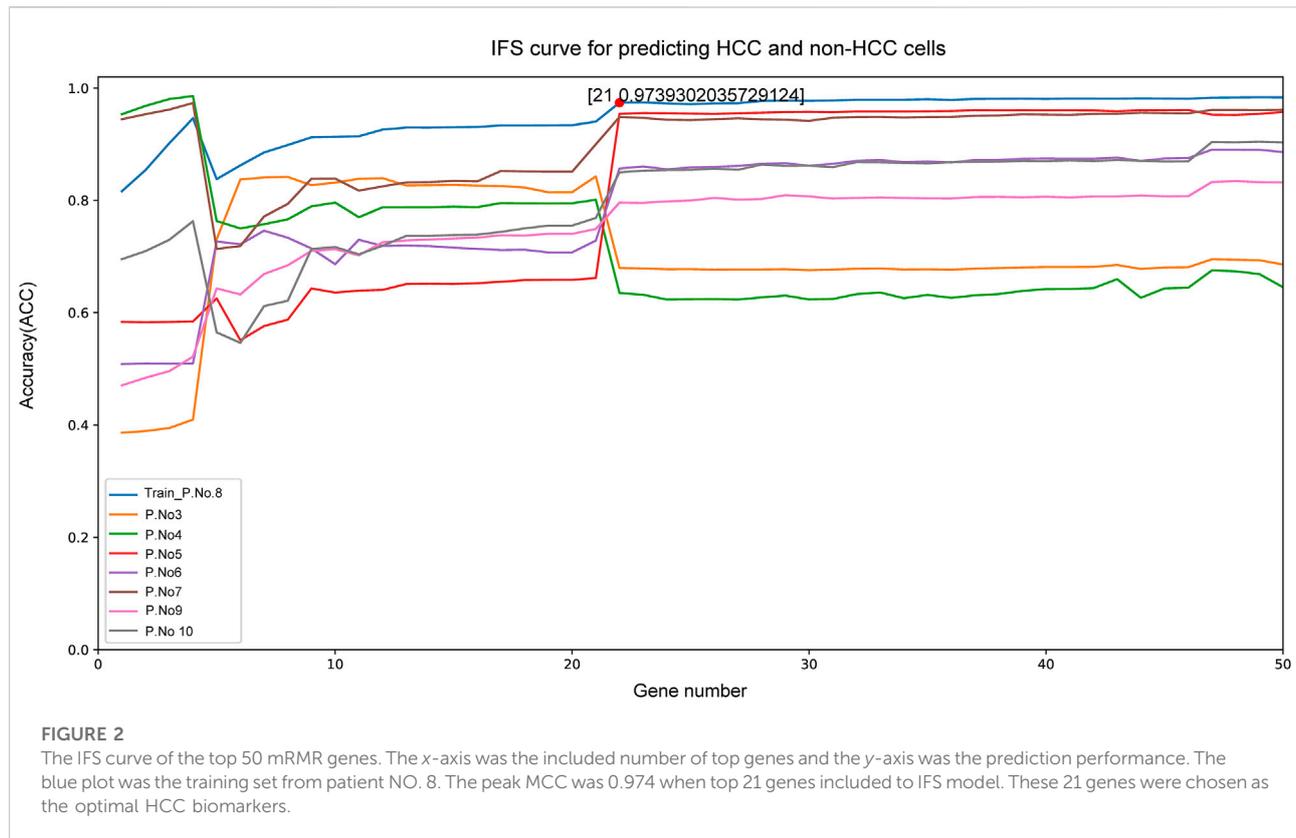


TABLE 1 The 21 optimal HCC biomarker genes got from IFS method.

| Rank | Gene | Rank | Gene | Rank | Gene |
|------|-------|------|------------|------|--------|
| 1 | SPP1 | 8 | GPR18 | 15 | SLAMF6 |
| 2 | FCN3 | 9 | AKNA | 16 | TRGC1 |
| 3 | FCRL6 | 10 | FCMR | 17 | STAT4 |
| 4 | S1PR5 | 11 | AC092580.4 | 18 | SCML4 |
| 5 | CD8A | 12 | AIM1 | 19 | HBB |
| 6 | SAA1 | 13 | GZMM | 20 | PLAC8 |
| 7 | CD160 | 14 | IGHA1 | 21 | APOA2 |

mesenchymal cells. Thus, the multi-dimensional interpretation of tumor heterogeneity by scRNA-seq will effectively solve the current clinical problem of chemoresistance and tumor recurrence and guide tumor immunotherapy and targeted therapy for HCC patients. However, limitations still exist in scRNA-seq research nowadays. First, it seems difficult for scRNA-seq to provide specific genetic markers to guide clinical diagnosis and treatment, which focus more on the alteration of organism and histology levels, making single-cell profiling too “microscopic.” Second, even though scRNA-seq provides a temporal map of the tumor microenvironment and

cell development and many new clusters of tumor progenitors and immune cells were identified, it is still hard to conduct molecular biological research on these new discoveries to elucidate the pathogenesis underlying the course of diseases. In a word, these limitations were mainly caused by insufficient mining of scRNA-seq big data. Innovative algorithmic strategies are demanded to provide new biological implications for scRNA-seq.

Max-Relevance and Min-Redundancy (mRMR) algorithm provides a highly robust feature selection scheme in machine learning and has been applied in multi-omics medical research in recent years. However, during the process of continuously adding features, mRMR only considers the local optimal solution. Thus, after obtaining the feature set with the importance ranking from high to low through this algorithm, a secondary feature selection is usually followed. For example, [Morgan et al. \(2021\)](#) applied mRMR along with an explainable boosting machine (EBM) classifier for CT radiomics to predict local failure following chemoradiation for head and neck cancer patients. [Gao et al. \(2020\)](#) chose an mRMR plus Random Forest model to find the lncRNA signature in bulk RNA-seq for immunophenotype prediction in Glioblastoma. In scRNA-seq, [Cheng et al. \(2020\)](#) applied an mRMR plus Support Vector Machines (SVM) to screen core biomarkers that distinguish the

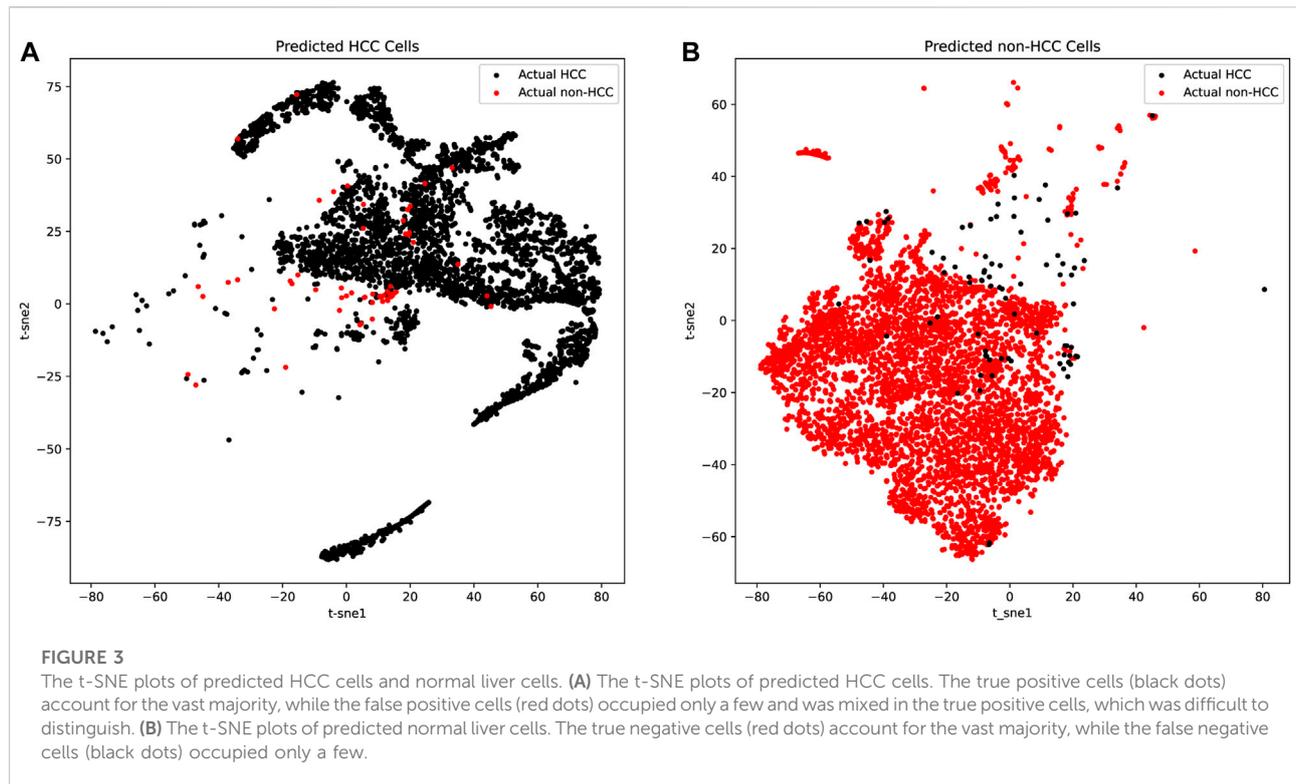


TABLE 2 Confused matrix of the 21 selected genes.

| | Predicted HCC | Predicted non-HCC |
|----------------|---------------|-------------------|
| Actual HCC | 4,713 | 82 |
| Actual non-HCC | 105 | 4,728 |

discrepancy between GBM tumor and pericarcinomatous environment. Based on the above research, we believed that the algorithmic scheme centered on mRMR might be used to screen biomarkers between cancer and non-cancer in scRNA-seq as well. This might have biological significance in assisting tumor diagnosis and tumor tissue identification during biopsy as well as providing novel parenchyma and stromal biomarkers for a certain cancer type.

In this study, based on paired scRNA-seq results of HCC and adjacent normal liver cells from eight patients, we designed a new computational strategy, consisting of machine learning algorithms, to screen core biomarkers that could distinguish the discrepancy between HCC and normal liver tissue. Gene expression profiles of tumor cells and paired hepatocytes were analyzed by maximum relevance minimum redundancy (mRMR) to get a 50-hub-gene feature. For further screening and classification of the 50-gene-feature, a support vector machine (SVM) algorithm was adopted. Results yielded a gene set with 21 genes that might be essential biomarkers for HCC tumor patients.

Materials and methods

Single cell gene expression profiles of HCC tumors and normal liver tissues

Single-cell gene expression profile data of HCC was obtained and downloaded from Gene Expression Omnibus (GEO) database in NCBI, the accession number was GSE149614. In this dataset, >70,000 single-cell transcriptomes for 10 HCC patients were sequenced and further measured using Illumina NovaSeq 6000 platform (GPL24676). Here we extracted expression profiles of patient 8 as our training set and patient three to seven and 9-10 as our validation set. Patient No. 1 and 2 were excluded from our study for a lack of sequencing data of paired normal liver tissues. The number of expressed genes was counted in paired HCC and normal liver samples of each patient. We utilized this dataset to further establish our gene feature for the purpose of discriminating HCC cells from normal liver cells.

mRMR ranking of discriminative genes

To achieve the goal of best discriminating the two types of tissues using the least number of genes, the Max-Relevance and Min-Redundancy (mRMR) algorithm was applied (Peng et al., 2005). This algorithm aimed to find a gene set that had the biggest correlation between the selected genes and samples (Max-

TABLE 3 GO (BP) enrichment results of the 21 selected genes.

| GO_Term | P | Adj.P | Genes |
|---|----------|----------|------------------------------|
| Lymphocyte mediated immunity | 2.49E-05 | 0.014115 | CD8A/CD160/GZMM/IGHA1/SLAMF6 |
| Leukocyte mediated cytotoxicity | 0.000165 | 0.030107 | CD160/GZMM/SLAMF6 |
| Positive regulation of natural killer cell mediated cytotoxicity | 0.000262 | 0.030107 | CD160/SLAMF6 |
| Positive regulation of natural killer cell mediated immunity | 0.000384 | 0.030107 | CD160/SLAMF6 |
| Alpha-beta T cell activation | 0.000402 | 0.030107 | CD160/GPR18/SLAMF6 |
| Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains | 0.000447 | 0.030107 | CD8A/GZMM/IGHA1/SLAMF6 |
| Cell killing | 0.000648 | 0.030107 | CD160/GZMM/SLAMF6 |

TABLE 4 Summary for markers related to the malignant phenotype or clinical prognosis of HCC.

| GeneName | Location | Function summaries | Related pathways | Reported functions in HCC |
|----------|--------------|--|---|--|
| SPP1 | 4q22.1 | 1. Forming an integral part of the mineralized matrix and is to cell-matrix interaction. 2. Acting as a cytokine enhancing IFN- γ and IL-12, reducing production of IL-10, essential in type I immunity. | 1. Cytokine activity 2. Integrin binding 3. Protein binding 4. Extracellular matrix binding | 1. Prognostic marker for HCC Zheng et al. (2018) , Ouyang et al. (2020) 2. Enhancer of cell growth Wang et al. (2019) |
| SAA1 | 11p15.1 | 1. A major acute phase protein that is highly expressed in response to inflammation 2. Major biomarker for diverse tumors | 1. Heparin binding and chemoattractant activity 2. Activated TLR4 signalling 3. Signaling by GPCR | Lowly expressed in HCC patients, indicating worse prognosis Zhang et al. (2020) |
| STAT4 | 2q32.2-q32.3 | A member of the STAT family of transcription factors activated by cytokines | 1. DNA-binding transcription factor activity 2. Sequence-specific DNA binding 3. JAK-STAT signaling pathway | Tumor suppressor in HCC that inhibit proliferation and promote apoptosis Li et al. (2016) |
| PLAC8 | 4q21.22 | A highly conservative protein, physiology function unknown | 1. Chromatin binding activity 2. Differentiation of white and brown adipocyte | Downregulated in HCC, indicating poor prognosis when lowly expressed by promoting cell proliferation Zou et al. (2016) |

Relevance), but the least correlation between genes inside this gene set (Min-Redundancy). The redundancy between genes was minimized as genes with similar expression characteristics were removed, except for the most representative genes remained. This method was confirmed effective in finding core biomarkers in sequencing analysis, especially in scRNA-seq with large and spare expression data ([Cheng et al., 2020](#)). It effectively overcame the shortcomings of traditional differential expression analysis in bulk RNA-seq, helping us to get a smaller number of biomarkers with the highest representation.

The mathematical model of this algorithm was shown as followed. First, we defined all genes, selected genes, and to be selected genes as Ω , Ω_s , and Ω_t , respectively. The relevance (D) of

gene g from Ω_t with cell type t can be measured with mutual information (I)

$$D = I(g, t)$$

And the redundancy R of the gene g with the selected genes in Ω_s are

$$R = \frac{1}{m} \left(\sum_{g_i \in \Omega_s} I(g, g_i) \right)$$

Now, our goal is to get the gene g_j from Ω_t so that D takes the maximum value (D_{\max}) and R takes the minimum value (R_{\min}), which can be expressed as the following function

TABLE 5 Summary for markers reported in other cancers.

| GeneName | Location | Function summaries | Related pathways | Reported functions in other cancers |
|------------|----------|---|--|---|
| FCN3 | 1p36.11 | 1. Calcium-independent lectin activity, found in all human serum. 2. Functioning in innate immunity 3. Related to innate immune. system and creation of C4 and C2 activators. | 1. Carbohydrate binding activity 2. Antigen binding activity 3. Complement pathway in association with MASPs and sMAP | Highly expressed in ovarian cancer and leukaemia patients Szala et al. (2013), Sokołowska et al. (2020). |
| APOA2 | 1q23.3 | 1. The second most abundant protein in high density lipoprotein (HDL) particles 2. Stabilizing HDL structure by its association with lipids | 1. Lipoprotein metabolism 2. Signaling by GPCR 3. Protein homodimerization 4. Activity of lipid binding | A minimally invasive biomarker for detecting pancreatic cancer, bladder cancer, and metastatic renal cell cancer patients Vermaat et al. (2012), Chen et al. (2015), Sato et al. (2020) |
| AC092580.4 | 2p25.1 | Affiliated with the lncRNA class | NA | 1. Highly expressed in relapse AML patients Feng et al. (2018) |
| S1PR5 | 19p13.2 | 1. Receptor for a bioactive lysophospholipid S1P 2. Both intracellular as a second messenger and extracellular as a receptor ligand | 1. G protein-coupled receptor activity 2. Sphingosine-1-phosphate receptor activity | 2. Highly expressed and promoting proliferation and invasion in clear cell carcinoma (ccRCC) and colon cancer patients Peng et al. (2020), Zhou et al. (2020) |
| AKNA | 9q32 | 1. Centrosomal protein that plays a key role in cell delamination 2. An epithelial-to-mesenchymal transition (EMT) regulator 3. A transcription factor that specifically activates the expression of the CD40 | 1. RNA polymerase II proximal promoter sequence-specific DNA binding activity 2. Proximal promoter DNA-binding transcription activator activity | 1. A tumor suppressor in gastric cancer by modulating EMT Wang et al. (2020) 2. An immune activator in cervical cancer Manzo-Merino et al. (2018) |
| AIM1 | 6q21 | NA | NA | 1. A classical tumor suppressor with high mutational frequency in melanoma Ray et al. (1997) 2. Suppressing tumor migration in prostate cancer Haffner et al. (2017) |

$$\max_{g_j \in \Omega_i} \left[I(g_j, t) - \frac{1}{m} \left(\sum_{g_i \in \Omega_i} I(g_j, g_i) \right) \right] \quad (j = 1, 2, \dots, n)$$

After n rounds of evaluation, all genes (Ω) will be ranked as a new gene list

$$S = \{g'_1, g'_2, \dots, g'_i, \dots, g'_N\}$$

The subscript i here reflects the trade-off between relevance with tissue type and redundancy with selected genes. The smaller index i is, the better discriminating power the gene has, and the higher of the corresponding gene g_i ranks.

Screening and optimization of single cell HCC biomarkers

We then constructed 50 support vector machines (SVM) classifiers and applied an incremental feature selection (IFS) method (Ye et al., 2017) using Top 50 mRMR genes to further screen optimized biomarker genes. The 50 gene sets are defined as

$$S_k = \{g'_1, g'_2, \dots, g'_k\} \quad k = (1, 2, \dots, 50)$$

Each candidate gene set includes the top k genes in the mRMR gene set S .

To prevent overfitting and evaluate the generalization ability of prediction performance for each SVM classifier, the leave-one-out cross validation (LOOCV) (Cheng et al., 2017) was then applied. Here we briefly described the procedure of LOOCV. Supposed that a dataset has N samples, in each round of LOOCV, there are $(N-1)$ samples adopted for training and the remaining one sample for testing. This process keeps running until all the N samples have been tested for one time after N rounds.

Since the positive and negative sample sizes are imbalance, the Matthews correlation coefficient (MCC) (Matthews, 1975), which considered both sensitivity and specificity, seems idealized for our IFS optimizing process. The calculation formula of MCC is shown as followed:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP , TN , FP , and FN are the abbreviation of true positive, true negative, false positive, and false negative, respectively.

TABLE 6 Summary for immune cell markers.

| GeneName | Location | Function summaries | Related pathways | Immune cell type | Reported in HCC |
|----------|--------------|--|--|--|-------------------------------|
| IGHA1 | 14q32.33 | Constant region of immunoglobulin heavy chains | <ol style="list-style-type: none"> 1. Cell surface interactions 2. Response to elevated platelet cytosolic Ca²⁺ 3. Antigen binding activity 4. Immunoglobulin receptor binding activity | B cells | Not Reported |
| FCRL6 | 1q23.2 | MHC class II receptor | protein phosphatase binding | NK cells and CTLs | Not Reported |
| CD8A | 2p11.2 | <ol style="list-style-type: none"> 1. A classic surface glycoprotein on most CTLs mediateing immune cell interactions 2. A coreceptor for MHC class I molecule: peptide complex | <ol style="list-style-type: none"> 1. Protein homodimerization activity 2. Coreceptor activity | CTLs | Reported Sangro et al. (2020) |
| CD160 | 1q21.1 | <ol style="list-style-type: none"> 1. A transmembrane on immune cells, mainly NK cells and activated T cells 2. Upon persistent antigen stimulation, it may contribute to CTL exhaustion | <ol style="list-style-type: none"> 1. Innate Lymphoid Cell Differentiation Pathways 2. Class I MHC mediated antigen processing and presentation 3. MHC class I receptor activity | NK cells and activated T cells | Reported Sun et al. (2018) |
| GPR18 | 13q32.3 | <ol style="list-style-type: none"> 1. A cannabinoid-activated orphan G protein-coupled receptor 2. Selected expressed on immune cells | <ol style="list-style-type: none"> 1. Peptide ligand-binding receptors 2. G protein-coupled receptor activity | Tumor-infiltrating B lymphocytes (TIL-Bs) and CD8 ⁺ T cells | Not Reported |
| GZMM | 19p13.3 | A member of granzymes, serine proteases activity | <ol style="list-style-type: none"> 1. Serine-type endopeptidase activity 2. Endopeptidase activity 3. Creation of C4 and C2 activators | NK cells | Not Reported |
| SLAMF6 | 1q23.2-q23.3 | A type I transmembrane protein, belonging to the CD2 subfamily of the immunoglobulin superfamily | <ol style="list-style-type: none"> 1. Class I MHC mediated antigen processing and presentation 2. Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell | Mainly in NK cells, also existing in T and B cells | Not Reported |

After all the above procedures, an IFS curve was finally formed. The x -axis of this curve denoted the number of genes in the SVM classifier (1–100) and the y -axis indicated the MCCs of it. Based on the IFS curve, an inflecting point that represented the usage of relatively few genes to get a relatively higher prediction accuracy was marked. The x -coordinate value of this inflecting point indicated number of genes in the final biomarker gene set and the y -coordinate value represented the prediction performance. This point was regarded as the most suitable SVM model for the final HCC biomarkers.

Biological significance analysis

For the gene signature predicted by the above algorithms, we then analyzed their expressed cell types, chromosomal location, and functions by GO, GENECARDS and literature reviewing. For GO, biological process (BP) was enriched and the p value as well as false discovery rate (FDR) based on hypergeometric distribution were calculated, FDR <0.05 was considered

significantly enriched. GENECARDS database was available in <https://www.genecards.org/>. Literature reviewing was applied using NCBI pubmed databases (www.ncbi.nlm.nih.gov/pubmed/) to search publications for every gene in recent 10 years.

The workflow diagram of this study was shown in Figure 1.

Results and discussion

Identifying the most discriminative feature by mRMR algorithm

After mRMR algorithm was applied, a feature, composed of top 50 most significant genes was established. This gene set was listed in Supplementary Table S1. Based on the principle of the mRMR algorithm, we believed that this feature was the most relevant one to distinguish HCC tumor cells from normal liver cells and had the least redundancy among the elements inside this gene set.

Further screening for the optimal HCC biomarker genes by IFS method

Given that our aim was to discriminate sample groups most significantly using biomarker genes as few as possible, the feature of 50 genes formed by mRMR algorithm was obviously too large to possess a practical value. Thus, we needed to choose an optimized group from these 50 genes as the final marker. To achieve this goal, IFS method was adopted. In the first round, only rank-first gene in mRMR was included as feature gene, then an SVM classifier was built to predict the group of each sample and validation was achieved by LOOCV and quantified by MCC value. In the second round, the rank-second gene was added into the previous 1-gene-feature and the above steps were repeated. This process kept repeated for 500 times until all 50 mRMR genes were included in the SVM model and an IFS curve was formed. As shown in Figure 2, the best peak MCC was 0.974 when 21 genes were included in SVM model in patient No.8 (train set), this peak MCC was also detected in other patients for validation (Figure 2). This peak was also validated effective in other patients except for patient No. 3 and No. 4, with MCC value no more than 0.7 but acceptable. Thus, these 21 genes were adopted as our final optimal HCC biomarkers (Table 1).

We further applied t-distributed stochastic neighbor embedding (t-SNE) for predicted HCC and non-HCC cells to detect both the tumor purity and the robustness of our classifier based on the 21 genes. As shown in Figures 3A,B and Table 2, there were only a few false positive (red dots in Figure 3A) and false negative dots (black dots in Figure 3B) mixed with true positive and true negative samples. However, the proportion of those false dots was extremely low with true dots and hard to classify. These t-SNE plots suggested that the HCC cells might contain non-HCC cells and vice versa, but most cells from the corresponding group were acceptable and the algorithms we applied could get the robust single cell biomarkers even when there were little tissue purity issues.

The biological functions of the 21 hub genes

The machine learning methods provided us with a new set of gene features for HCC to identify tumor from paired normal liver tissues. However, nothing was learned from the biological significance of this gene set. We first performed Gene Ontology (GO) enrichment analysis for biological process (BP) analysis of the selected 21 genes (Table 3). Results of GO enrichment showed that they were enriched in Natural Killer (NK) cells and T cells (Tc) related pathways, indicating that changes in immune microenvironment are the core difference between HCC tumors and normal liver tissues.

To explore the functions of these 21 genes in more depth from a biomedical perspective, we reviewed the location, basic functions, as well as related biological pathways and processes for each gene through *Genecards* database (<https://www.genecards.org/>). Literatures about the biological functions of these 21 genes in HCC and/or other malignant tumors were also thoroughly searched through Pubmed database. We finally divided these 21 genes into four categories (Siegel et al., 2020) Markers related to the malignant phenotype or clinical prognosis of HCC (Table 4); (McGlynn et al., 2020) Markers without reports in HCC but were related to the pathogenesis and/or prognosis of other malignant tumors (Table 5); (Petrowsky et al., 2020) Marker genes expressed by immune cells (Table 6); (Nault and Villanueva, 2020) Other genes that have not yet been studied in cancers, including *FCMR*, *TRGCI*, and *HBB*. Subsequent research was worth exploring the role of these markers in the pathogenesis of HCC and their applying prospects in HCC diagnosis, monitoring and treatment. Furthermore, it was worth mentioning the immune-specific genes in Table 6. These genes were all markers for Natural Killer (NK) cells and T cells, indicating that changes in cytotoxic effects might play a vital role in the HCC immunity. These immune markers might be promising targets for enhancing the efficacy of HCC immunotherapy.

In summary, using an integrated machine learning strategy, mainly composed of mRMR and IFS, we analyzed scRNA-seq data from eight paired HCC tissues and normal liver tissues. A 21-gene-feature consisted of both cancer markers and immune cell markers was established. This feature was regarded as the core to distinguish HCC from normal liver tissue. Considering that the tissue obtained from clinical needle biopsy is often a mixture of tumor parenchyma, stroma and normal liver tissue, this 21-gene-feature might help both the clinical diagnosis of HCC and the identification of biopsy-obtained tissue types. Besides, given that these 21 genes, most of which had not been fully explored in HCC, were expressed in different parenchymal and mesenchymal cells, the following research might focus on their biological function and molecular mechanism in distinct HCC-related cell cluster.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

JS and HZ designed the concept of this research. WL and JL conducted all the studies in this paper. WZ and WG validated and proofread data. WG wrote this paper. XJ and ZY revised this paper.

Funding

This article was supported by the following fundings 1) National Natural Science Foundation of China (No. 81803085); 2) Natural Science Foundation of Hunan Province, China (Nos 2019JJ40414, 2018JJ2612, 2019JJ50285, and 2020JJ4919); 3) Science and Technology Project of Changsha, Hunan, China (No. kq2004146); 4) Scientific Research Project for Distinguished Young Scientists of Provincial Education Department of Hunan, China (No. 18B460).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Chen, H., Chen, Y., Tsai, R., Chen, M., Chen, S., Xiao, M., et al. (2015). A sensitive and selective magnetic graphene composite-modified polycrystalline-silicon nanowire field-effect transistor for bladder cancer diagnosis. *Biosens. Bioelectron.* 66, 198–207. doi:10.1016/j.bios.2014.11.019
- Cheng, H., Garrick, D., and Fernando, R. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *J. Anim. Sci. Biotechnol.* 8, 38. doi:10.1186/s40104-017-0164-6
- Cheng, Q., Li, J., Fan, F., Cao, H., Dai, Z., Wang, Z., et al. (2020). Identification and analysis of glioblastoma biomarkers based on single cell sequencing. *Front. Bioeng. Biotechnol.* 8, 167. doi:10.3389/fbioe.2020.00167
- Craig, A., von Felden, J., Garcia-Lezana, T., Sarcognato, S., and Villanueva, A. (2020). Tumour evolution in hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* 17 (3), 139–152. doi:10.1038/s41575-019-0229-4
- Feng, Y., Shen, Y., Chen, H., Wang, X., Zhang, R., Peng, Y., et al. (2018). Expression profile analysis of long non-coding RNA in acute myeloid leukemia by microarray and bioinformatics. *Cancer Sci.* 109 (2), 340–353. doi:10.1111/cas.13465
- Gao, M., Wang, X., Han, D., Lu, E., Zhang, J., Zhang, C., et al. (2020). A six-lncRNA signature for immunophenotype prediction of glioblastoma multiforme. *Front. Genet.* 11, 604655. doi:10.3389/fgene.2020.604655
- Haffner, M., Esopi, D., Chau, A., Gürel, M., Ghosh, S., Vaghasia, A., et al. (2017). AIM1 is an actin-binding protein that suppresses cell migration and micrometastatic dissemination. *Nat. Commun.* 8 (1), 142. doi:10.1038/s41467-017-00084-8
- Ho, D., Tsui, Y., Sze, K., Chan, L., Cheung, T., Lee, E., et al. (2019). Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and stemness-related subpopulations in liver cancer. *Cancer Lett.* 459, 176–185. doi:10.1016/j.canlet.2019.06.002
- Li, J., Liang, L., Liu, Y., Luo, Y., Liang, X., Luo, D., et al. (2016). Clinicopathological significance of STAT4 in hepatocellular carcinoma and its effect on cell growth and apoptosis. *Onco. Targets. Ther.* 9, 1721–1734. doi:10.2147/OTT.S100040
- Manzo-Merino, J., Lagunas-Martínez, A., Contreras-Ochoa, C., Lizano, M., Castro-Muñoz, L., Calderón-Corona, C., et al. (2018). The human papillomavirus (HPV) E6 oncoprotein regulates CD40 expression via the AT-hook transcription factor AKNA. *Cancers* 10 (12), E521. doi:10.3390/cancers10120521
- Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405 (2), 442–451. doi:10.1016/0005-2795(75)90109-9
- McGlynn, K., Petrick, J., and El-Serag, H. (2020). *Epidemiology of hepatocellular carcinoma*. Baltimore, Md: Hepatology.
- Morgan, H., Wang, K., Dohopolski, M., Liang, X., Folkert, M., Sher, D., et al. (2021). Exploratory ensemble interpretable model for predicting local failure in head and neck cancer: The additive benefit of CT and intra-treatment cone-beam

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.873218/full#supplementary-material>

- computed tomography features. *Quant. Imaging Med. Surg.* 11 (12), 4781–4796. doi:10.21037/qims-21-274
- Nault, J., and Villanueva, A. (2020). *Biomarkers for hepatobiliary cancers*. Baltimore, Md: Hepatology.
- Ouyang, G., Yi, B., Pan, G., and Chen, X. (2020). A robust twelve-gene signature for prognosis prediction of hepatocellular carcinoma. *Cancer Cell. Int.* 20, 207. doi:10.1186/s12935-020-01294-9
- Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8), 1226–1238. doi:10.1109/TPAMI.2005.159
- Peng, R., Wang, Y., Mao, L., Fang, F., and Guan, H. (2020). Identification of core genes involved in the metastasis of clear cell renal cell carcinoma. *Cancer Manag. Res.* 12, 13437–13449. doi:10.2147/CMAR.S276818
- Petrowsky, H., Fritsch, R., Guckenberger, M., De Oliveira, M., Dutkowski, P., and Clavien, P. (2020). Modern therapeutic approaches for the treatment of malignant liver tumours. *Nat. Rev. Gastroenterol. Hepatol.* 17, 755–772. doi:10.1038/s41575-020-0314-8
- Ray, M., Wistow, G., Su, Y., Meltzer, P., and Trent, J. (1997). AIM1, a novel non-lens member of the betagamma-crystallin superfamily, is associated with the control of tumorigenicity in human malignant melanoma. *Proc. Natl. Acad. Sci. U. S. A.* 94 (7), 3229–3234. doi:10.1073/pnas.94.7.3229
- Rebouissou, S., and Nault, J. (2020). Advances in molecular classification and precision oncology in hepatocellular carcinoma. *J. Hepatol.* 72 (2), 215–229. doi:10.1016/j.jhep.2019.08.017
- Sangro, B., Melero, I., Wadhawan, S., Finn, R., Abou-Alfa, G., Cheng, A., et al. (2020). Association of inflammatory biomarkers with clinical outcomes in nivolumab-treated patients with advanced hepatocellular carcinoma. *J. Hepatol.* 73 (6), 1460–1469. doi:10.1016/j.jhep.2020.07.026
- Sato, Y., Kobayashi, T., Nishiumi, S., Okada, A., Fujita, T., Sanuki, T., et al. (2020). Prospective study using plasma apolipoprotein A2-isoforms to screen for high-risk status of pancreatic cancer. *Cancers* 12 (9), E2625. doi:10.3390/cancers12092625
- Siegel, R., Miller, K., and Jemal, A. (2020). Cancer statistics, 2020. *Ca. Cancer J. Clin.* 70 (1), 7–30. doi:10.3322/caac.21590
- Sokolowska, A., Świerzek, A., Gajek, G., Gołos, A., Michalski, M., Nowicki, M., et al. (2020). Associations of ficolins and mannose-binding lectin with acute myeloid leukaemia in adults. *Sci. Rep.* 10 (1), 10561. doi:10.1038/s41598-020-67516-2
- Stegle, O., Teichmann, S., and Marioni, J. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16 (3), 133–145. doi:10.1038/nrg3833
- Sun, H., Xu, J., Huang, Q., Huang, M., Li, K., Qu, K., et al. (2018). Reduced CD160 expression contributes to impaired NK-cell function and poor clinical outcomes in patients with HCC. *Cancer Res.* 78 (23), 6581–6593. doi:10.1158/0008-5472.CAN-18-1049

- Szala, A., Sawicki, S., Swierzko, A., Szemraj, J., Sniadecki, M., Michalski, M., et al. (2013). Ficolin-2 and ficolin-3 in women with malignant and benign ovarian tumours. *Cancer Immunol. Immunother.* 62 (8), 1411–1419. doi:10.1007/s00262-013-1445-3
- Tang, F., Barbacioru, C., Nordman, E., Li, B., Xu, N., Bashkurov, V., et al. (2010). RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* 5 (3), 516–535. doi:10.1038/nprot.2009.236
- Vermaat, J., Gerritse, F., van der Veldt, A., Roessingh, W., Niers, T., Oosting, S., et al. (2012). Validation of serum amyloid α as an independent biomarker for progression-free and overall survival in metastatic renal cell cancer patients. *Eur. Urol.* 62 (4), 685–695. doi:10.1016/j.eururo.2012.01.020
- Wang, G., Sun, D., Li, W., and Xin, Y. (2020). AKNA is a potential prognostic biomarker in gastric cancer and function as a tumor suppressor by modulating EMT-related pathways. *Biomed. Res. Int.* 2020, 6726759. doi:10.1155/2020/6726759
- Wang, J., Hao, F., Fei, X., and Chen, Y. (2019). SPP1 functions as an enhancer of cell growth in hepatocellular carcinoma targeted by miR-181c. *Am. J. Transl. Res.* 11 (11), 6924
- Ye, Y., Zhang, R., Zheng, W., Liu, S., and Zhou, F. (2017). Rifs: A randomly restarted incremental feature selection algorithm. *Sci. Rep.* 7 (1), 13013. doi:10.1038/s41598-017-13259-6
- Zhang, Q., He, Y., Luo, N., Patel, S., Han, Y., Gao, R., et al. (2019). Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell.* 179 (4), 829–845. e20. doi:10.1016/j.cell.2019.10.003
- Zhang, W., Kong, H., Gao, X., Dong, Z., Lu, Y., Huang, J., et al. (2020). Immune infiltration-associated serum amyloid A1 predicts favorable prognosis for hepatocellular carcinoma. *World J. Gastroenterol.* 26 (35), 5287–5301. doi:10.3748/wjg.v26.i35.5287
- Zheng, Y., Huang, Q., Ding, Z., Liu, T., Xue, C., Sang, X., et al. (2018). Genome-wide DNA methylation analysis identifies candidate epigenetic markers and drivers of hepatocellular carcinoma. *Brief. Bioinform.* 19 (1), 101–108. doi:10.1093/bib/bbw094
- Zhou, H., Yin, X., Bai, F., Liu, W., Jiang, S., and Zhao, J. (2020). The role and mechanism of S1PR5 in colon cancer. *Cancer Manag. Res.* 12, 4759–4775. doi:10.2147/CMAR.S239118
- Zou, L., Chai, J., Gao, Y., Guan, J., Liu, Q., and Du, J. (2016). Down-regulated PLAC8 promotes hepatocellular carcinoma cell proliferation by enhancing PI3K/Akt/GSK3 β /Wnt/ β -catenin signaling/ β -catenin signaling. *Biomed. Pharmacother. = Biomedicine Pharmacother.* 84, 139–146. doi:10.1016/j.biopha.2016.09.015