# Identification of tumor antigens and anoikis-based molecular subtypes in the hepatocellular carcinoma immune microenvironment: implications for mRNA vaccine development and precision treatment

Zhiyuan Zheng[1,2,3], Hantao Yang[4], Yang Shi[2], Feng Zhou[5], Lingxiao Liu[1,3], Zhiping Yan[1,3*] and Xiaolin Wang[1,3*]

*Correspondence:
yan.zhiping@zs-hospital.sh.cn;
fduwangxiaolin@outlook.com

[1] Department of Interventional
Radiology, Zhongshan Hospital
Fudan University and Shanghai
Institute of Medical Imaging,
Fudan University, 180 Fenglin
Road, Shanghai 200032, China
[2] Faculty of Medicine, Institute
for Experimental Molecular
Imaging, Uniklinik RWTH
Aachen and Helmholtz Institute
for Biomedical Engineering,
RWTH Aachen University,
52074 Aachen, Germany
[3] National Clinical Research
Center for Interventional
Medicine, Zhongshan Hospital,
Fudan University, 180 Fenglin
Road, Shanghai 200032, China
[4] Department of Neurosurgery,
Zhongshan Hospital Fudan
University, 180 Fenglin Road,
Shanghai 200032, China
[5] Department of Neurosurgery,
Baoji Central Hospital, 8 Jiangtan
Road, Baoji 721008, Shaanxi
Province, China

**Abstract**

Hepatocellular carcinoma (HCC) represents a formidable malignancy with a high lethality. Nonetheless, the development of vaccine and the establishment of prognostic models for precise and personalized treatment of HCC still encounter big challenges. Thus, the aim of this study was to develop HCC vaccines and explore anoikis-based prognostic models based on RNA sequencing data in GEO datasets (GSE10143, GSE76427) and the TCGA-LIHC cohort. Potential HCC antigens were identified using GEPIA2, cBioPortal, and TIMER2. Anoikis-related subtypes and gene clusters were defined by consensus clustering of 566 liver cancer samples based on 28 anoikis regulators, and we further analyzed their relationship with the immune microenvironment of HCC. A predictive model based on anoikis-related long noncoding RNAs (lncRNAs) was developed to accurately predict HCC prognosis. Seven overexpressed genes associated with HCC prognosis and tumor-infiltrating antigen-presenting cells were identified as potential tumor antigens for the development of HCC mRNA vaccines. Two subtypes based on anoikis-related genes (ARGs) and two gene clusters with different characteristics were identified and validated in defined cohorts. The tumor immune microenvironment between the two subtypes showed different cell infiltration and molecular characteristics. Furthermore, a prognostic score based on seven lncRNAs identified by LASSO regression was constructed, with the low-risk group having favorable prognosis, a "hot" immune microenvironment, and better response to immunotherapy. CCNB1, CDK1, DNASE1L3, KPNA2, PRC1, PTTG, and UBE2S were first identified as promising tumor antigens for mRNA vaccine development in HCC. Besides, we innovatively propose anoikis-based molecular subtypes, which not only enable personalized prognostic stratification of HCC patients but also provide a blueprint for identifying optimal candidates for tumor vaccines, enhancing immunotherapeutic strategies.

**Keywords:** HCC, Anoikis, Prognosis, Subtype, TCGA, GEO, Machine learning, Hepatocellular carcinoma, Anoikis, Tumor antigen, mRNA vaccine, Tumor immune microenvironment

Zheng *et al. Journal of Big Data*      (2023) 10:129

Page 2 of 32

## Introduction

Hepatocellular carcinoma (HCC), the most common primary liver malignancy, ranks as the seventh most prevalent cancer worldwide, with 905,677 newly diagnosed cases (4.7%) and 830,180 deaths (8.3%) [1]. Patients with early-stage HCC may qualify for curative treatments such as surgery and liver transplantation, while those with advanced HCC, comprising approximately 70% to 80% of patients, generally experience a survival rate of less than 12% over a 5-year period. This is primarily attributed to the scarcity of treatment options available and the suboptimal response to initial therapy [2, 3]. Therefore, identifying innovative approaches for earlier detection is essential for better prediction of therapy response and survival in patients with HCC [4]. Since 2017, immune checkpoint inhibitors (ICIs) and combinations based on ICIs have emerged as the most promising approach for improving the prognosis of patients with unresectable HCC. However, accumulating evidence has shown that the poor pharmacokinetic properties of antibodies can induce treatment-related adverse events, while off-target immunological effects on other organs can trigger immune-related adverse events, which can lead to treatment cessation and drug-related deaths, thereby limiting the application of this therapy [5]. Randomized trials assessing the efficacy of anti-PD-1 monotherapy in both the first-line and second-line setting showed no improvement in overall survival (OS). Consequently, the development of new combination therapy strategies aimed at activating the suppressive immune microenvironment in HCC is necessary [6].

Tumor-associated antigens (TAAs) are crucial targets for activating immunotherapy, as they are normal self-proteins that are overexpressed or re-expressed by tumors. Developing effective, minimally invasive, and long-lasting treatments targeting TAAs has been a long-standing goal in cancer therapy. Cancer vaccines aim to stimulate tumor-specific immune responses by targeting TAAs and increasing cytotoxic CD8+T cells. Additionally, effective immunotherapy responses have been associated with neoantigens, which are novel immunogenic sequences resulting from tumor mutations. Neoantigens are also being considered as potential targets for cancer immunotherapies, such as tumor vaccines, adoptive cell therapy (ACT), and antibody-based treatments, and as predictors for immune checkpoint blockers (ICBs). Despite the fact that HCC tumor cells can elicit significant immune responses against a range of mutant antigens, no clinically developed HCC vaccine targeting a tumor antigen currently exists. Therefore, there is an urgent need to identify novel antigens in HCC and develop HCC vaccines.

Apoptosis is a programmed cell death mechanism that helps maintain tissue homeostasis and prevent the development of cancer. When cells lose contact with neighboring cells or their extracellular matrix (ECM), they undergo a process called anoikis, which is a type of apoptosis that is triggered by detachment from the ECM [7]. Cancer cells detaching from primary sites acquire anoikis resistance and can colonize target organs or tissues via the circulatory and lymphatic systems [8]. Therefore, anoikis, a new hallmark of cancer metastasis, has attracted the attention of oncologists [9]. In HCC, strategies that reverse the anoikis resistance phenotype of HCC cells to inhibit cancer metastasis may be promising for the treatment of patients with local metastasis and potential vascular invasion. Some studies have elucidated the relationship between tumor resistance to anoikis and immunity, revealing that anoikis affects the tumor immune microenvironment and anoikis resistance of cancer cells [10, 11]. Zhang et al. discovered that the

Zheng *et al. Journal of Big Data* (2023) 10:129

Page 3 of 32

deactivation of IL1RAP promotes anoikis and suppresses Ewing sarcoma cell metastasis, revealing the promising potential of anoikis-targeted therapy in cancer [11]. Thus, exploring the role of anoikis in HCC will provide valuable insights into patient prognostic stratification and the development of novel therapeutic targets, which also has the potential to enhance combination therapy with cancer vaccines and immunotherapy in HCC.

LncRNAs have been discovered to modulate the expression of genes associated with anoikis, either by facilitating or impeding anoikis resistance in cancer. Previous studies have shown that silencing lncRNA ANRIL promotes caspase activity, leading to anoikis-induced cell death in glioma cells. Additionally, lncRNA-HOX antisense intergenic RNA (lncRNA-HOTAIR) is highly expressed in various cancers and has been shown to contribute to cell survival and epithelial–mesenchymal transition (EMT) [12]. However, the molecular role and clinical implications of anoikis in HCC remain unknown. Hence, deciphering the molecular underpinnings of the relationship between the anoikis phenotype and lncRNAs in HCC based on datasets from multiple clinical RNA sequencing sample cohorts may illuminate the intricate mechanisms underlying HCC progression and metastasis. Additionally, such investigations may uncover promising therapeutic targets, paving the way for effective interventions to impede the development of HCC.

In this study, we aimed to identify tumor antigens that could be utilized in the development of an mRNA vaccine for HCC. Furthermore, we constructed a prognostic risk score model based on anoikis-related lncRNAs, which has the potential to predict the prognosis of HCC patients as well as provide insights into the immunological tumor microenvironment (TME) of HCC. The risk model can be further employed to facilitate precision HCC treatment, such as chemoimmunotherapy, and the development of HCC vaccines.

## Methods

### Acquisition and preprocessing of HCC datasets

We initially searched the Gene-Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases for publicly available RNA-seq data that had complete clinical information. All adult patients and patients with missing survival information were excluded from further analysis. Finally, we enrolled a total of 566 patients from three HCC cohorts (GSE10143, GSE76427, and TCGA-LIHC) into a metacohort for further investigation. Microarray data (GSE10143 and GSE76427 cohorts) were downloaded from the GEO database at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10143 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76427, and RNA sequencing data (TCGA-LIHC cohort, FPKM values) were obtained from the Genomic Data Commons (GDC) at https://portal.gdc.cancer.gov/ [13–15]. The FPKM data were converted to transcripts per kilobase million (TPM) values, and batch effects due to nonbiological technical biases were adjusted using the "ComBat" method in the "sva" R package. Table 1 summarizes the baseline data for all relevant HCC datasets. Only genes with nonzero expression levels in over 50% of samples were retained. Additionally, mutation data were retrieved from the TCGA-LIHC database. A combined total of 132 anoikis-related genes were obtained from the GeneCard database (https://www.genecards.org/, Accessed on

**Table 1** Clinical baseline characteristics of HCC patients in TCGA and GEO cohorts

| Variable | TCGA cohorts | GEO cohorts | | |
|---|---|---|---|---|
| | n = 371 | GSE10143, n = 80 | GSE76427, n = 115 | |
| Age, year | 59.22 ± 13.42 | NA | 63.45 ± 12.68 | |
| Gender | | NA | NA | |
| Male | 119 | NA | 93 | |
| Female | 252 | NA | 44 | |
| Grade | | NA | BCLC stage | |
| G1 | 55 | NA | 0 | 4 |
| G2 | 180 | NA | A | 74 |
| G3 | 123 | NA | B | 28 |
| G4 | 13 | NA | C | 9 |
| Stage | | NA | | |
| I | 174 | NA | 55 | |
| II | 85 | NA | 35 | |
| III | 86 | NA | 21 | |
| IV | 5 | NA | 3 | |
| Unknown | 21 | NA | 1 | |
| AJCC-T | | NA | | |
| T1 | 184 | NA | NA | |
| T2 | 92 | NA | NA | |
| T3 | 81 | NA | NA | |
| T4 | 13 | NA | NA | |
| Unknown | 1 | NA | NA | |
| AJCC-N | | NA | | |
| N0 | 256 | NA | NA | |
| N1 | 4 | NA | NA | |
| Nx | 110 | NA | NA | |
| Unknown | 1 | NA | NA | |
| AJCC-M | | NA | | |
| M0 | 270 | NA | NA | |
| M1 | 4 | NA | NA | |
| Mx | 97 | NA | NA | |
| Overall survival | | NA | | |
| 1-year | 84.6% | 98.6% | 88.7% | |
| 3-year | 72.3% | 88.6% | 84.3% | |
| 5-year | 67.4% | 78.8% | 80.7% | |
| Survival status | | | | |
| Alive | 243 | 48 | 92 | |
| Death | 128 | 32 | 23 | |

Data are shown as n or n%

AJCC, American Joint Committee on Cancer; GEO, Gene Expression Omnibus; HCC, hepatocellular carcinoma; TCGA, The Cancer Genome Atlas; NA, not Applicable

25 December 2022) [16] and the Harmonizome portal (https://maayanlab.cloud/Harmonizome/, Accessed on 25 December 2022) [17]. After performing a regression analysis of the 132 ARGs (p < 0.05), we identified 28 anoikis-related regulators. The

Zheng *et al. Journal of Big Data*     (2023) 10:129

Page 5 of 32

annotation file was collected from the Ensembl database (http://asia.ensembl.org) for lncRNA selection.

### Identification of tumor antigens and Kaplan–Meier survival analysis

cBioPortal (https://www.cbioportal.org/) was used to merge unprocessed data from the TCGA cohort in this study [18, 19]. In addition, cBioPortal was utilized to visualize genetic mutations in HCC patients and to identify mutant genes. Differentially expressed genes (DEGs) in HCC patients were identified using ANOVA and the "Differential Genes" module of Gene Expression Profiling Interactive Analysis version 2 (GEPIA2, http://gepia2.cancer-pku.cn/#index), [20] with settings of |log2FC| value > 1 and q value < 0.01. Additionally, the OS and relapse-free survival (RFS) of HCC patients were assessed using the "survival analysis" module of GEPIA2 with a median cutoff of 50%, and the log-rank test was used for comparisons. The median value of identified tumor antigen expression was used to separate HCC patients into high and low expression groups, and the hazard ratio was calculated by the Cox proportional hazards regression model. The parameter settings were consistent throughout all analyses, and no p value adjustments were made. In various comparisons made between three or more groups, one-way ANOVA and Kruskal–Wallis tests were used, and p values less than 0.05 were considered statistically significant.

### Tumor immune microenvironment component analysis

The Tumor Immune Estimation Resource (TIMER) is a web server that provides comprehensive analysis of tumor-infiltrating immune cells, including dendritic cells (resting and activated) and macrophages (M0, M1, and M2) [21]. In this study, we used TIMER-based purity-adjusted partial Spearman's correlation analysis to investigate the relationship between tumor antigen expression and APC infiltration. We also utilized analytical modules for gene expression, somatic mutations, clinical outcomes, and somatic copy number changes to analyze and visualize the relationship between the number of tumor immune infiltrating cells (TIICs) and HCC tumor antigen-related genes. The cutoff value for statistical significance was $p < 0.05$.

### Unsupervised clustering and subtype analysis based on 28 ARGs

To uncover distinct anoikis regulatory patterns mediated by ARGs, we identified 28 regulators from two combined GEO datasets and TCGA cohorts. Based on the expression of these 28 ARGs, we used unsupervised clustering analysis to categorize patients into groups with different anoikis regulation patterns for further study. We used the consensus clustering algorithm [19] to estimate the number of clusters and their stability. We utilized the "ConsensusClusterPlus" packages to perform these processes, with 500 repetitions performed to ensure classification stability.

### Gene set variation analysis (GSVA) and functional annotation

We used the R package "GSVA" to conduct GSVA enrichment analysis to evaluate the differences in the biological process regulatory patterns of anoikis. GSVA is a nonparametric and unsupervised technique commonly used to predict the variation in pathway

Zheng *et al. Journal of Big Data*      (2023) 10:129

Page 6 of 32

and biological process activity in expression dataset samples [22]. We retrieved "c2. cp.kegg.v6.2.-symbols" gene sets from the MSigDB database for GSVA. A p value of less than 0.05 was considered statistically significant. We used the clusterProfiler R package to perform functional annotation for anoikis-related genes with an FDR threshold of 0.05 [23, 24].

### Identification and validation of anoikis subtypes

To classify patients into unique anoikis regulation subtypes, we utilized the expression of 28 ARGs and applied the empirical Bayesian method of the limma R package to identify DEGs between these types. We set a significance threshold for identifying DEGs at an adjusted p value of less than 0.001. Next, we quantified the anoikis regulatory patterns of each tumor sample and developed a scoring system, named the ARG score, to evaluate the anoikis pattern of each patient with HCC. To determine the ARG score, we normalized the DEGs with anoikis regulation pattern-related genes across all samples and identified the overlapping genes. We then used an unsupervised clustering approach to categorize the 566 patients into multiple groups for further investigation, and the consensus clustering approach was used to determine the optimal number of gene clusters and their stability. We conducted prognostic analysis for each gene in the signature using the univariate Cox regression model and isolated 28 genes with significant prognostic value for further analysis. We calculated the ARG score using principal component analysis (PCA), which centered the score on the set containing the largest block of highly correlated (or uncorrelated) genes and downweighed the contributions of genes that did not track with other set members.

### Further establishment of an anoikis-related lncRNA signature

After performing a differential expression analysis between the high and low ARG score groups, we identified anoikis-related lncRNAs using the R package "limma." Next, we used the Wilcoxon test to compare gene expression across 50 normal tissue samples and 374 HCC samples from TCGA, resulting in the identification of another set of 855 anoikis-related lncRNAs. Then, we investigated the association between overall survival (OS) and 104 lncRNAs that were differentially expressed and associated with anoikis in the TCGA training set using univariate Cox survival analysis. The R package "survival" was utilized to calculate the hazard ratio (HR) and p value, and 44 lncRNAs were chosen for further analysis using a significance threshold of $p < 0.001$ with the LASSO approach, aiming to identify hub lncRNAs associated with survival. We also plotted the partial likelihood deviation against log (lambda), with lambda being a tuning parameter. Multivariate Cox regression analysis was performed to assess the regression coefficients of significant prognostic lncRNAs associated with anoikis in the training set. These lncRNAs were selected based on their strong correlation with anoikis, and a signature of seven lncRNAs was ultimately generated using the coefficients. Based on their individual anoikis risk indices (ARIs), patients in the training set were divided into low- and high-risk groups using the median ARI value, which was performed by the package "survminer" in R. The Kaplan–Meier algorithm was used to compare the differences in OS between these groups. ROC curves were generated to evaluate the ARI formula's

Zheng *et al. Journal of Big Data*      (2023) 10:129

Page 7 of 32

predictive accuracy by examining its sensitivity and specificity. The formula was also applied to the validation cohorts to ensure its consistency. Since some clinicopathologic features, such as TNM stage, may affect prognosis in HCC patients, univariate and multivariate analyses were performed to determine whether the ARI was independent of these features.

Furthermore, ANOVA was conducted to explore the relationship between anoikis regulators and their associated molecular and cellular properties. The genes that were most susceptible to mutation were examined for significant changes using chi-square analysis. The functional characteristics of each gene module were annotated by analyzing gene ontologies by DAVID software [25, 26]. To determine the immunological enrichment score of each sample, which was defined as the number of genes that were upregulated or downregulated in a coordinated manner within the sample, single-sample gene set enrichment analysis (ssGSEA) was employed.

### Predicting the effective response to immunotherapy and drugs

The IPS of HCC patients was obtained from The Cancer Immunome Atlas (TCIA) website (https://tcia.at/home) and was objectively determined by considering four types of immunogenicity-determining genes, including genes related to effector cells, immunosuppressive cells, MHC molecules, and immune modulators. To compare the gene expression levels across the four distinct cell types, z scores were utilized. Higher IPS scores were found to be associated with increased immunogenicity [27]. In addition, the TIDE algorithm (http://tide.dfci.harvard.edu/) was employed to predict the response to immune checkpoint blockade (ICB) and to evaluate neoantigen potential [28, 29]. In addition, we used the R package "pRRophetic" to predict the sensitivity of chemotherapeutic agents based on a statistical model constructed based on gene expression and drug sensitivity data from a very large panel of cancer cell lines [30].

### Construction of a nomogram based on the ARI and comparison of its prediction abilities with clinicopathologic features

With the "rms" package in R, we were able to construct a nomogram that incorporates independent prognostic criteria to predict 1-, 3-, and 5-year OS rates. The AUC values were used to evaluate the nomogram's discriminatory power in predicting survival. The DCA curve supported the use of the nomogram for making useful and beneficial inferences. In addition to the factors already included in the nomogram, we generated a comprehensive nomogram indicator with the clinical and laboratory markers, which was a better predictor of overall survival.

### Statistical analysis

Data processing and analysis were performed using R software (version 4.1.0). For continuous variables with normal distribution and variance, an independent samples $t$ test was utilized, whereas for those without normal distribution and variance, the Wilcoxon rank-sum test was used. The Pearson correlation coefficient test was employed to analyze associations. A statistical significance threshold of $P < 0.05$ was considered meaningful. All the R packages and statistical methods were list in Additional file 2: Table S11.
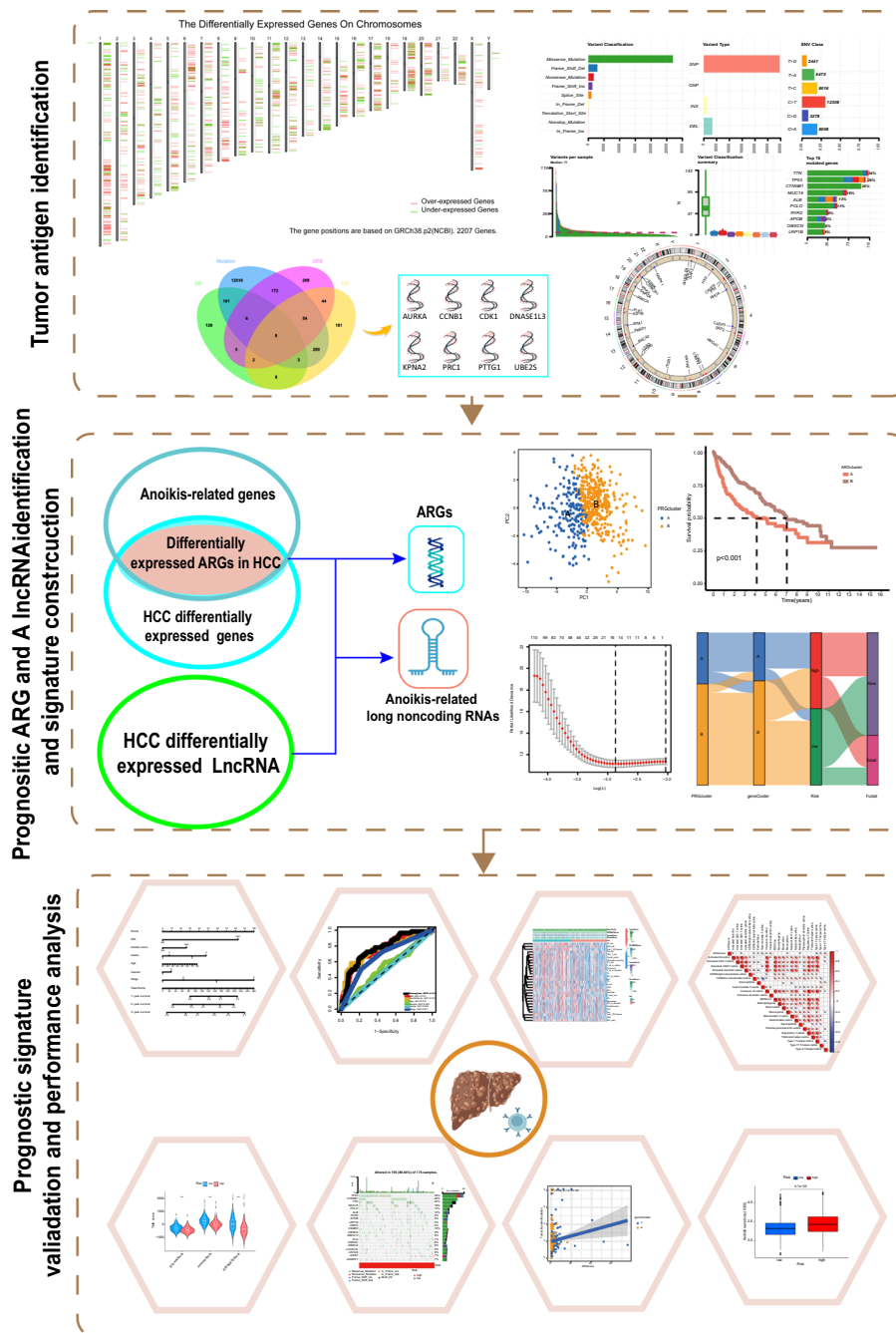
**Fig. 1** The flow diagram of the research process

## Results

### Identification of potential tumor antigens in HCC

The study flowchart is presented in Fig. 1. To identify potential PAAD antigens, we screened for aberrantly expressed genes, which resulted in the identification of 2,207 overexpressed genes (as determined by GEPIA) that potentially encode tumor-associated antigens (Fig. 2A). Next, using cBioPortal, we analyzed genomic changes and mutation counts in samples and identified a total of 13,054 mutated genes that

Zheng *et al. Journal of Big Data*     (2023) 10:129

Page 9 of 32
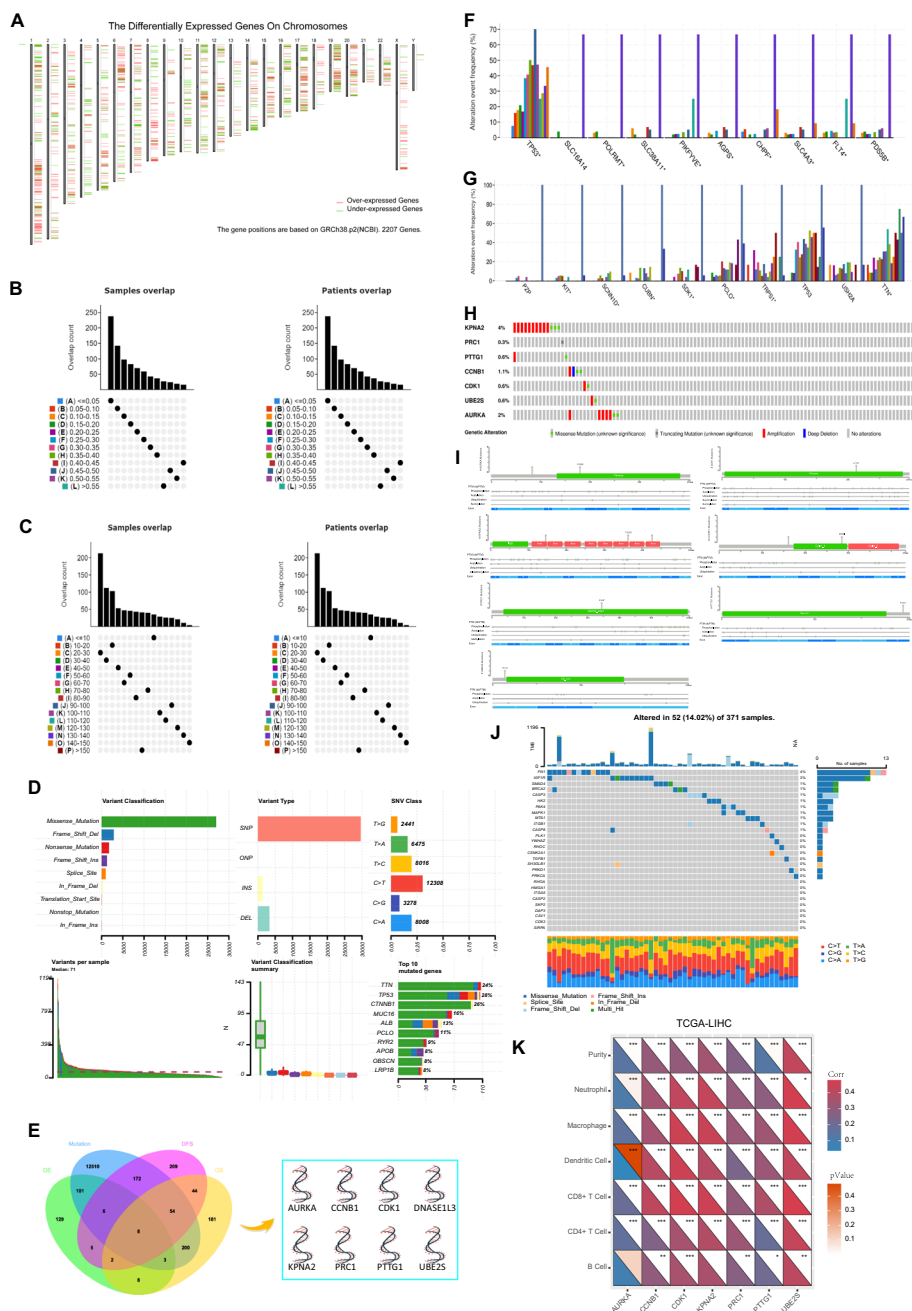


**Fig. 2** Recognition potential mRNA antigen associated tumor antigens and scenery of genetic and expression profiles of anoikis related regulators in hepatocellular carcinoma. **A** Chromosomal distribution of potential genes encoding liver cancer antigens. **B** HCC patients's samples overlapping in the altered genome fraction group **C** HCC patients's samples overlapping in the mutation count group. **D** Summary of Potentially Mutated Genes in Liver Cancer. **E** Venn diagram of liver cancer antigen gene identification, 8 antigen genes are highly expressed, mutated, and related to OS and RFS. **F** The HCC most frequently mutated genes in the altered genome fraction group. **G** The HCC most frequently mutated genes mutation count group. **H**, **I** The OncoPrint tab of genomic alterations in the TCGA dataset. **J** The frequency of the top 20 highly mutated anoikis related genes in HCC. **K** Heat map of the relationship between the tumor microenvironment and the identified potential antigen genes for liver cancer in the TCG database. *p < 0.05; **p < 0.01; ***p < 0.001

Zheng *et al. Journal of Big Data*      (2023) 10:129

Page 10 of 32

may encode tumor-specific antigens. Interestingly, most patients had low levels of genomic changes and mutation counts, indicating that HCC is seemingly not highly immunogenic (Fig. 2B, C). The top 10 most frequently altered genes in the fractional genomic alteration and mutation count groups are depicted in Fig. 2F and G, respectively. Notably, alterations to the tumor protein p53 gene were observed in a significant proportion of patients in both groups. Furthermore, analysis of hepatocellular carcinoma mutation data from the TCGA database revealed that deletion mutations were the most frequently observed type of mutation in liver cancer, with single-nucleotide polymorphisms being the most common type among these. Among single-nucleotide mutations, the most common substitution was C to T, followed by T to C, C to A, T to A, C to G, and T to G in decreasing order of frequency. After analyzing the mutated genome fraction and mutation count, the top three genes found to be mutated were TTN, TP53, and CTNNB1. In addition to these genes, MUC16, ALB, PLCO, RYR2, APOB, OBSCN, and LRP1B were also observed to have mutations in a significant proportion of hepatocellular carcinomas, with frequencies exceeding 8%. (Fig. 2D) These genes are likely to have significant research value for further mutation studies. Furthermore, in patients with overlapping mutations, genes such as titin, transcriptional repressor GATA binding 1, piccolo presynaptic cytomatrix protein, and usherin also showed a high frequency of mutations. In total, a set of 118 genes that were both overexpressed and mutated were identified by overlapping these two gene sets.

Through screening of 11 candidate genes closely related to overall survival (OS) of HCC, we identified eight genes that were significantly associated with recurrence-free survival (RFS) (Fig. 2E and Additional file 2: Table S1). These genes may play a crucial role in the progression of HCC and can be exploited in mRNA vaccine production. The OncoPrint tab presents a summary of the genomic changes in each mutated and overexpressed gene observed in the TCGA dataset (Fig. 2H, I). For example, overexpression of Aurora kinase A (AURKA) in tumor tissues was associated with a lower survival rate (Additional file 1: Fig. S1A). Ephrin B2 (EFNB2), cyclin B1 (CCNB1), cyclin-dependent kinase 1 (CDK1), karyopherin subunit alpha 2 (KPNA2), protein regulator of cytokinesis 1 (PRC1), PTTG1 and ubiquitin conjugating enzyme E2 S (UBE2S) all had high expression levels that were correlated with a poor prognosis (Additional file 1: Fig. S1B–T). One protective gene (AURKA) was found and further excluded.

The heatmap in Fig. 2K shows the correlation analysis of the infiltration of immune cells in the microenvironment and the seven genes in HCC. Heatmap analysis showed statistically significant correlations between seven of the genes (AURKA, CCNB1, CDK1, KPNA2, PRC1, PTTG1, and UBE2S) and tumor purity, neutrophils, macrophages, dendritic cells, and CD8-positive and CD4-positive T cells ($p < 0.05$). A significant correlation was found between the infiltration of B cells and the five genes, but this correlation was not found for AURKA and KPNA2. These findings suggest that these genes have important correlations with the aforementioned immune cells. A strong correlation was also observed between the expression of these seven candidate genes and B cell, macrophage, dendritic cell, and CD8 T-cell infiltration in HCC (Additional file 1: Fig. S2A–G). Since antigen-presenting cells (APCs) may directly

Zheng *et al. Journal of Big Data* (2023) 10:129

Page 11 of 32

process and display these seven tumor antigens to generate an immune response, they are targets that should be a focus of studies for the development of HCC mRNA vaccines. In summary, these seven potential genes that are both mutated and overexpressed are important in the progression and tumor microenvironment of HCC.

**Mutation and immune microenvironment analysis of the ARGs**

We combined two GEO datasets (GSE10143 and GSE76427) and the TCGA-LIHC cohort into a metacohort with OS data and clinical information (Table 1). Differential analysis and univariate regression analysis of the ARGs revealed that patients with high expression of the 28 ARGs had poorer overall survival (OS). (Additional file 1: Fig. S3 and Additional file 2: Table S2) The ARG network (Fig. 3A) revealed the landscape of the interactions, regulatory connections, and prognostic significance of the top 28 ARGs in HCC patients. We found a significant statistical relationship between these ARGs. From the network, it is apparent that all the genes functioned as risk factors for anoikis, with statistical evidence showing that each gene shared a strong positive regulatory link with other genes involved in anoikis, and these relationships were significant. High RHOC expression in tumors correlated with low SKP2 and MTA1 expression.

Based on the expression of these 28 ARGs, we were able to completely distinguish between HCC and normal samples. To investigate whether the mentioned genetic variants altered ARG expression in HCC patients, we examined the mRNA expression levels of anoikis regulators in normal and HCC samples. We found that changes in copy number variants (CNVs) may be the primary factor leading to ARG expression perturbations. All ARGs with increased CNVs showed significantly higher expression in HCC (Fig. 3B). In addition, our analysis of the frequency of CNVs showed that 28 ARGs had CNVs. Copy number variation analysis revealed that copy number gains (CNGs) accounted for the majority of copy number amplifications, whereas ICASP3, BRCA2, MTA1, and SMAD4 had a high frequency of deletion (decreased gene or sequence fragment copies in the genome) (Fig. 3C). Figure 3D depicts the location on the chromosomes where the CNVs of the ARGs were found. This evidence suggests that an imbalance in the expression of ARGs plays a critical role in the development of HCC, as our investigations revealed strikingly different changes in the genetic characteristics and gene expression of ARGs between normal and liver cancer samples.

Due to the higher frequency of mutations in the FN1 and IGF1R genes than in other genes, we analyzed how tumors with FN1 and IGF1R mutations and wild-type proteins expressed other ARGs. Compared to wild-type tumors, FN1-mutant tumors showed significant downregulation of snail family transcriptional repressor 2 (SNAI2), phosphatase and tensin homolog (PTEN), and fibronectin 1 (FN1), while tumor protein P53 and clusterin (CLU) were significantly upregulated and downregulated, respectively, in insulin-like growth factor 1 receptor (IGF1R)-mutant tumors (Fig. 3E–H). According to the findings, oncogenes were upregulated in the mutant group, while tumor suppressor genes were downregulated. These findings suggest that interactions between regulators of distinct ARGs likely contribute significantly to the development of tumor-specific patterns of anoikis regulation and TME cell infiltration.

We focused on FN1, a regulator, and discovered a highly favorable link between FN1 expression and a wide variety of immune cells that had infiltrated the TME. Subsequent
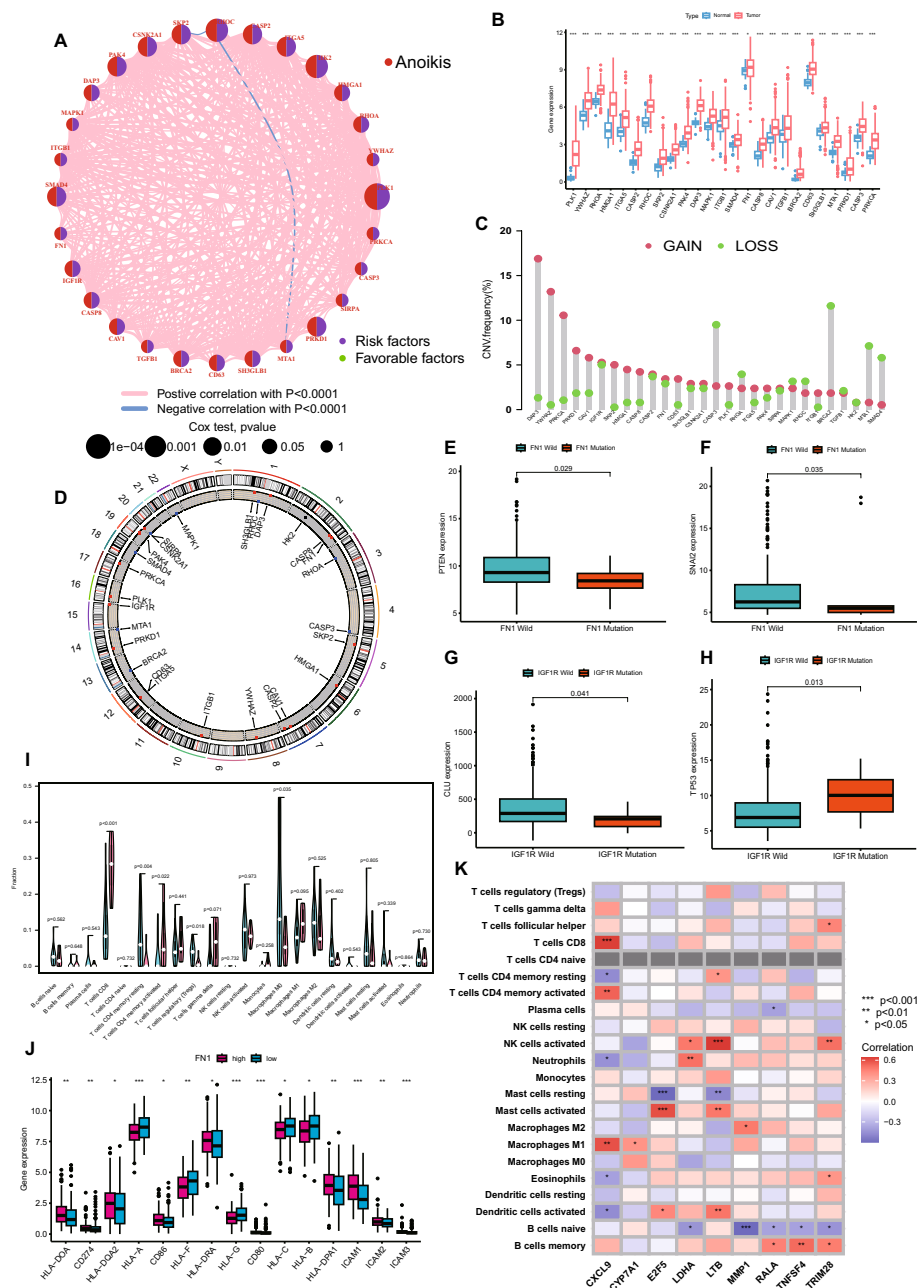
**Fig. 3** **A** Interaction relationship of anokis regulators. The size of each circle indicates the effect of each regulator on prognosis. **B** Expression of anoikis regulators in liver cancer and normal tissues. **C** Copy number variation analysis of 28 ARGs, red indicates the frequency of copy number gain, and green indicates the frequency of copy number decrease. **D** Chromosomal location map of ARGs with copy number variation. **E**, **F** Differentially expressed genes in FN1 mutant group and wild group. **G**, **H** Differentially expressed genes in IGF1R mutant group and wild group. **J** The expression of infiltrating immune cells in the tumor microenvironment in FN1-high expression and FN1-low expression groups. **K** The expression of MHC, costimulatory, and adhesion molecule expression in the tumor microenvironment in the FN1 high-expression and low-expression groups. **L** Correlation analysis heat map of ARGs and tumor infiltration-related cells

analyses revealed that FN1 was related to the tumor immunological microenvironment. To compare the total number of immune cells present in people with high and low FN1 expression, we used the ESTIMATE method. We analyzed the presence of twenty-two

immune cells in the TME in patients with high or low FN1 expression. Tumors with high FN1 expression showed significantly higher infiltration of CD8 T cells, CD4 memory activated T cells, and immune cells than those with low expression. (Fig. 3I) We also found that low expression of FN1 led to widespread increases in MHC, costimulatory, and adhesion molecule expression. Interestingly, the upregulation of the immune checkpoint molecule PD-L1 (CD274) in the high FN1 expression group occurred simultaneously with enhancement of immune-related pathways (Fig. 3J). To investigate further, we used Spearman's correlation analysis to examine the association between individual types of cells infiltrating the TME and each ARG (Fig. 3K).



**Fig. 4** **A**–**C** Unsupervised cluster analysis of 28 regulators of anoikis in a meta-cohort (TCGA + GEO). **D** Principal component analysis of two different anoikis-related gene regulation patterns. **E** Survival analysis of 2 modes of regulation of anoikis in a meta-cohort. **F** Analysis of the infiltration of immune cells in the tumor microenvironment of ARGs in two groups. **G** An unsupervised clustering heatmap of anoikis-related genes, with clinical shapes such as stage, gender, and age as annotations. **H** GSVA analysis heat map of two anoikis-related gene expression patterns. Red modules represent high expression gene patterns, blue represents low expression

**Identification of two anoikis regulation patterns mediated by 28 ARGs**

We performed unsupervised clustering based on the expression of 28 ARGs and were able to identify two distinctive patterns of anoikis regulation in the metacohort (Fig. 4A–C and Additional file 2: Table S3). Pattern A consisted of 190 cases, while pattern B had 376 cases. Using PCA to show the molecular patterns of different types in the two groups, we found that the ARG cluster was composed of two groups of samples with different characteristics (Fig. 4D). We also analyzed the prognosis of the two distinct anoikis patterns and found that the ARGscluster-B regulatory pattern conferred a greater advantage for survival than the ARGscluster-A regulatory pattern (Fig. 4E). To compare the differences in immune cell components between the two anoikis regulation modes, we used CIBERSORT, a deconvolution algorithm based on support vector regression for identifying immune cell types in malignancies. We found that ARGscluster-B had a disproportionately high number of immune cells, including gamma delta T cells, neutrophils, and type 17 T helper cells, while ARGscluster-A had an excessively high level of immune cells, including activated CD4 T cells, activated dendritic cells, CD56 natural killer cells, immature dendritic cells, MDSCs, natural killer T cells, natural killer cells, plasmacytoid dendritic cells, regulatory T cells, T follicular helper cells, and type 2 T helper cells (Fig. 5F). We also visualized the expression of the 28 regulators of anoikis in the metacohort using unsupervised clustering and a heatmap. Patient information such as ARG cluster, sex, tumor stage, project, and age were considered. The majority of ARGs had high expression in group A and low expression in group B, as illustrated by the heatmap (Fig. 5G). Using GSVA, we investigated the impact of different regulatory patterns on the biological pathways that are affected by anoikis regulation. GSVA indicated that ARGscluster-A was highly enriched in the cell cycle pathway, while ARGscluster-B was significantly enriched in metabolic pathways, including the arginine, proline, histidine, tyrosine, phenylalanine, glycine, serine, threonine, β alanine, tryptophan, propanoate, butanoate, fatty acid, linoleic acid, and retinol metabolism pathways and xenobiotics by cytochrome P450 pathway; furthermore, ARGscluster-B showed enrichment of glycolysis, gluconeogenesis, valine, leucine, and isoleucine degradation, complement and coagulation cascades, steroid hormone biosynthesis, and drug metabolism by cytochrome P450. (Fig. 4H) The findings from the GSVA indicated that the regulatory pattern found in cluster B was significantly associated with a high metabolic activity phenotype of HCC.

**Gene scoring and functional annotation for the anoikis phenotype**

We utilized the R package "limma" to identify 5588 DEGs associated with the anoikis phenotype to investigate the potential biological activity of each anoikis regulatory pattern. The DEGs were subjected to GO enrichment analysis with the "clusterProfiler" package, and Fig. 5A summarizes the significantly enriched biological processes (Additional file 2: Table S4). Interestingly, these genes were enriched in biological processes related to anoikis and immunometabolic pathways, providing further evidence for the crucial role of anoikis in the tumor microenvironment. KEGG analysis of the genes revealed the top seven pathways with activity counts greater than 40: the cell cycle, focal adhesion, regulation of actin cytoskeleton, PI3K-Akt signaling pathway, human papillomavirus infection, proteoglycans in cancer, and human T-cell leukemia
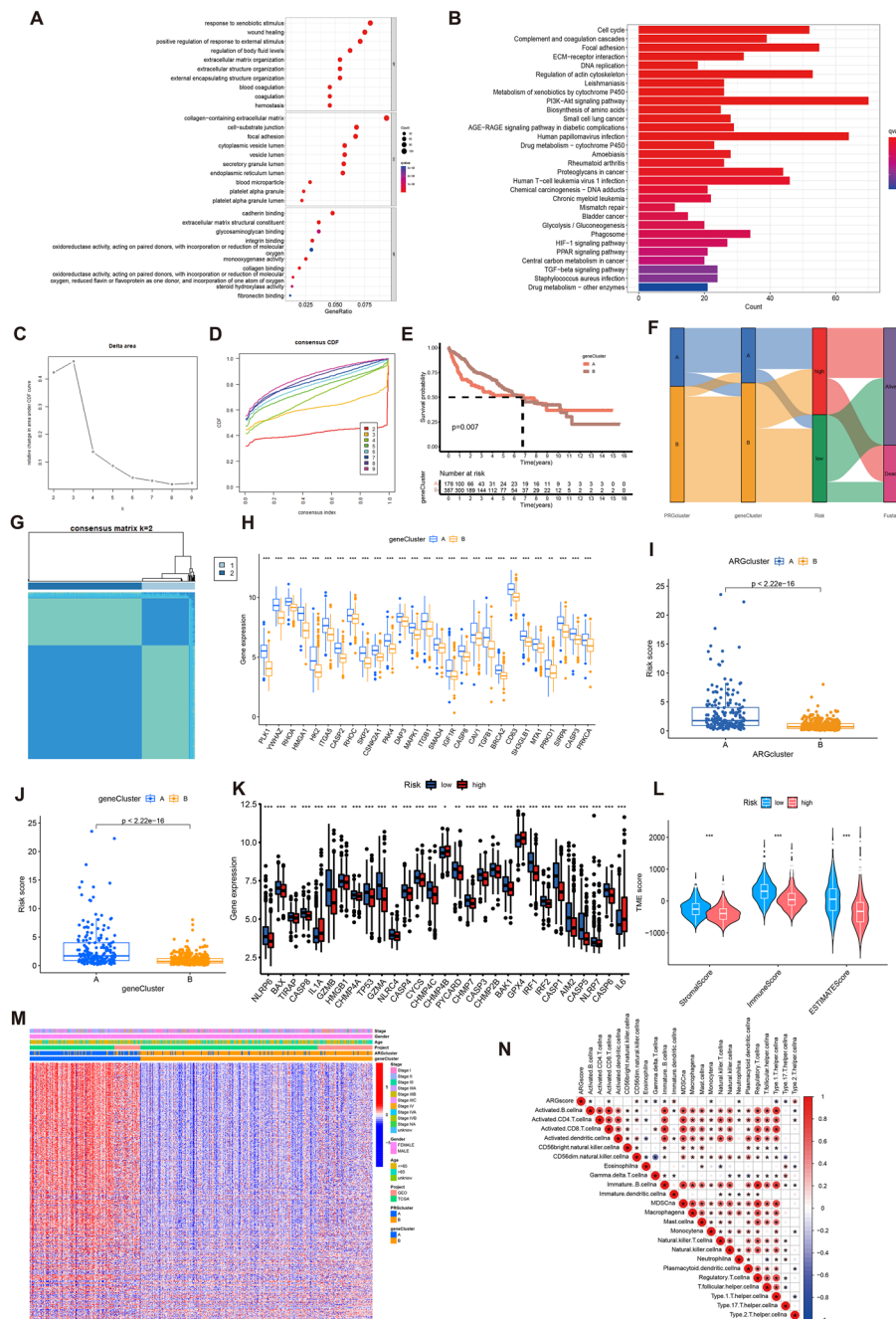
**Fig. 5 A** Gene ontology analysis of genes involved in anoikis (biological process, molecular function and cellular components). **B** KEGG pathway analysis of the intersection genes of two patterns of anoikis genes. **C**, **D** Consensus clustering of genes associated with anoikis phenotypes. **E** Survival analysis of genes associated with anoikis phenotype in two gene clusters. **F** Alluvial diagram of the relationship between the two modes of ARGs with the risk and survival status. **G** Unsupervised clustering matrix of anoikis clusters. **H** The gene expression of regulators of 28 anoikis genes was in two gene clusters. **I** Inter-group analysis of ARGs score between two ARGs clusters. **J** Inter-group analysis of ARGs gene cluster score between two clusters. **K** 28 anoikis-related genes expression in the high- and low-risk gene clustering group. **L** Tumor microenvironment scores difference of estimate score, immune score, stromal score in high-risk group and low-risk group. **M** Gene clustering analysis of two different anoikis gene clusters with stage, gender, age and project. **N** Correlation analysis between ARG score and tumor infiltrating immune cells

virus 1 infection. These pathways are closely linked to tumor cell proliferation, apoptosis, and metastasis, thus indicating the relevance of the model to the cancer-related anoikis phenotype (Fig. 5B and Additional file 2: Table S5). To verify this regulatory mechanism, we performed unsupervised clustering analysis of the 28 genes related to the anoikis phenotype and classified patients into distinct genomic subtypes. Unsupervised clustering identified two distinct anoikis genomic phenotypes, which we designated anoikis gene clusters A and B, consistent with the clusters identified based on anoikis regulatory patterns (Fig. 5C, D). Consistent with the predicted anoikis regulatory patterns, significant differences in the expression of ARGs were observed between the two ARG clusters (Fig. 5G and Additional file 2: Table S6).

The expression levels of anoikis gene cluster A were higher, while the expression levels of anoikis gene cluster B were lower, indicating two distinct phenotypic gene clusters related to anoikis. For example, differential analysis revealed significant differences in the expression of PLK1, YWHAZ, RHOA, HMGA1, HK2, ITGA5, CASP2, RHOC, SKP2, CSNK2A1, PAK4, DAP3, MAPK1, ITGB1, SMAD4, IGF1R, CSAP8, CAV1, TGFB1, BRCA2, CD63, SH3GLB1, MTA1, PRKD1, SIRPA, CASP3, and PRKCA between the two groups, with group A showing significantly higher expression levels than group B (Fig. 5H). Considering the role of these genes in the tumor microenvironment, it is speculated that group A has a higher level of tumor antigen activity, and the expression of these genes remodels the tumor microenvironment of HCC. Subsequently, we conducted a comparison between the ARGcluster and genecluster and found that the risk score of group A in both ARGcluster and genecluster was higher than that of group B, indicating that the patients in gene cluster A had a worse prognosis and higher risk (Fig. 5I, J). The expression levels of anoikis-related genes that were significantly different between the two groups were also observed to be higher in the high-risk group than in the low-risk group (Fig. 5K). The tumor microenvironment (TME) of HCC was evaluated using three scores: stromal, immune, and ESTIMATE scores. The results revealed that group A had higher stromal, immune, and ESTIMATE scores, demonstrating the intricate nature of the TME in this group. Notably, group A showed an increased level of immune activity, emphasizing the importance of considering the complexity of the TME for understanding the regulatory patterns of anoikis in HCC (Fig. 5L and Additional file 2: Table S7).

To assess the anoikis pattern of individual HCC patients, we developed a scoring system called the anoikis-related gene (ARG) score based on these phenotype-related genes. A Sankey diagram was used to depict the attribute changes of individual patients (Fig. 5F). Next, we aimed to determine the value of the ARG score in predicting the outcome of patients. Patients were classified as having a low or high ARG score based on the cutoff value, which was calculated using the "survminer" package in R. The results confirmed that patients with HCC could be classified into two groups with unique anoikis phenotype-related gene regulation modes. Patients with a favorable status (179 of 566 patients) for 7 years were mostly classified into anoikis gene cluster B, while those with a poor prognosis (387 patients) were clustered in anoikis gene cluster A. The anoikis gene clusters indicated that HCC has three unique anoikis phenotype-related gene regulation modes. The 5-year survival rate was found to be significantly higher among patients with a low ARG score than in those with a

high ARG score, as evidenced by the data presented in Fig. 5E. Additionally, the heatmap in Fig. 5M shows the gene expression differences between the ARGs and gene clusters. The results demonstrate that group A had a higher level of gene expression than group B, which is consistent with the previous findings.

Furthermore, the correlation analysis revealed a significant association between the ARG score and the proportions of immune cells present in HCC. For instance, the infiltration of activated B cells, CD8 T cells, CD56 natural killer cells, eosinophil γδ T cells, immature B cells, immature dendritic cells, monocytes, neutrophils, type 1 helper T cells, type 17 helper T cells, and type 2 helper T cells was lower when the ARG score was higher, as demonstrated in Fig. 5N.

### Development of a lncRNA signature associated with anoikis for diagnosis and treatment

The analysis of differences between 50 normal tissue samples and 374 HCC samples from TCGA was accomplished by "limma" packages in R, and a total of 16,773 lncRNAs were identified. Subsequently, the 370 HCC samples were randomly split into a "training" set comprising 185 samples and a "testing" set consisting of 185 samples for further analysis. Pearson's correlation analysis was utilized to determine the relationships between lncRNAs and ARGs, which led to the identification of 855 lncRNAs with a correlation coefficient greater than 0.4 and a p value less than 0.01, indicating their association with ARGs.

In the training group, the prognostic value of 28 ARG-related lncRNAs in liver cancer patients was then determined by using univariate Cox regression analysis ($p < 0.001$). To construct a prognostic anoikis-related lncRNA model, least absolute shrinkage and selection operator (LASSO) regression and cross validation were applied, resulting in the identification of prognostic lncRNAs (RNF216P1, AP002449.1, NRAV, AC100847.1, LINC01871, PSMB8-AS1, HMGN3-AS1, ZFPM2-AS1, TMEM220-AS1, AL590705.3, AC025176.1, AC027097.1, LINC01559, LINC01269A, and L117336.2) (Fig. 6A, B and Additional file 2: Table S8). Subsequently, multivariate Cox regression analysis was performed and revealed seven lncRNAs, which were included in the risk model score (AC100847.1, PSMB8-AS1, ZFPM2-AS1, AL590705.3, AC025176.1, LINC01559, and LINC01269), along with their corresponding hazard ratios (HRs) and 95% confidence intervals (CIs).

Each patient's anoikis risk index (ARI) was established using the following formula:

$$
\begin{aligned}
ARI = \; & (0.616675640597704 * AC100847.1) \\
& + (-0.797818424476765 * PSMB8 - AS1) \\
& + (0.441031959395331 * ZFPM2 - AS1) \\
& + (0.777545279336812 * AL590705.3) \\
& + (1.35551255903678 * AC025176.1) \\
& + (0.5062856764437 * LINC01559) \\
& + (0.665009661829924 * LINC01269)
\end{aligned}
$$

We performed survival analysis with ARI scores for HCC patients in TCGA. All the lncRNAs had positive associations with OS risk ($p < 0.0001$). Using the median value of
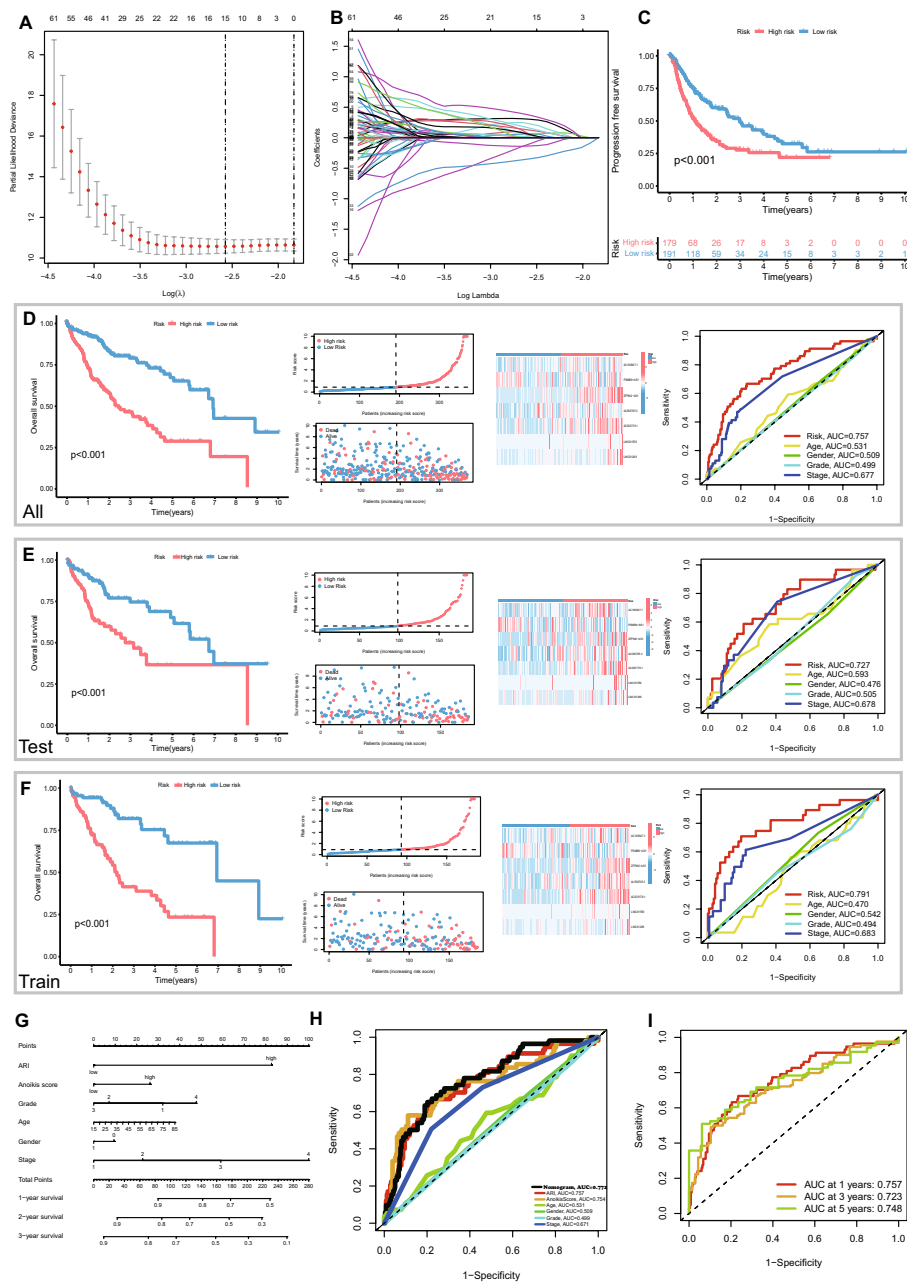
**Fig. 6** **A** Delta area curve. **B** Pyroptosis-associated lncRNAs identified using tenfold cross-validated lasso analysis. **C** Survival analysis of high and low risk groups of ARI score for progression-free survival. **D** The performance of anoikis-related lncRNA signatures in the combined group, including the OS survival curve, the risk map of survival time, the risk heat map of 7-lncRNAs, and the AUC curve of the model. **E** The performance of anoikis-related lncRNA signatures in the test group, including the OS survival curve, the risk map of survival time, the risk heat map of 7-lncRNAs, and the AUC curve of the model. **F** The performance of anoikis-related lncRNA signatures in the test group, including the OS survival curve, the risk map of survival time, the risk heat map of 7-lncRNAs, and the AUC curve of the model. **G** Dual-scoring nomogram with ARI and ARG score in TCGA cohort. (H) ROC curve of integrated nomogram. **I** AUC analysis of integrated nomogram

the risk score as the cutoff, all HCC patients were stratified into "low-risk" and "high-risk" groups. Patients with high ARI scores had lower disease-free survival and prognosis (Fig. 6C). Meanwhile, the results suggested that a higher risk score predicted a shorter OS of HCC patents in the training set, and similar results were confirmed in the validation set and metacohort. The AUC > 0.7 in the TCGA dataset suggested that the ARI signature exhibited excellent sensitivity and specificity for the prognostic prediction of HCC patients in the training and validation sets and the metacohort. Similarly, Kaplan–Meier analyses revealed that patients with a high ARI had a poor prognosis in the three cohorts (p < 0.01) (Fig. 6D–F). Furthermore, by applying principal component analysis (PCA) based on all the genes, ARGs, ARlncRNAs, and risk lncRNAs of the lncRNA signature, we were able to classify samples from HCC patients from the whole TCGA cohort into high-risk and low-risk categories for overall survival (OS) (Additional file 1: Fig. S4). Additionally, we included the ARI, age, sex, stage, and American Joint Committee on Cancer (AJCC) T stage, N stage, and M stage for univariate Cox regression analysis of the TCGA cohort to identify significant clinical variables for prognosis. Using univariate analysis, we found that the ARI signature score and stage were significantly correlated with OS (Additional file 1: Fig. S5A and Additional file 2: Table S9). Multivariate regression analysis was performed using the ARI and other clinical features to determine whether the ARI signature may be an independent prognostic factor for HCC patients. In three cohorts (test, training, and all cohorts), the 7-lncRNA signature was shown to be an independent and highly significant predictive factor. The results from the multivariate regression analysis showed that only stage and the signature remained significantly linked with OS in the whole TCGA cohort (Additional file 1: Fig. S5B and Additional file 2: Table S9).

To better demonstrate the predictive ability of the ARI score for liver cancer prognosis, a combined nomogram was developed (Fig. 6G) to predict OS at 1 year, 3 years, and 5 years for individual HCC patients based on multivariate regression analysis incorporating independent prognostic factors such as age, sex, grade, stage, anoikis score, and ARI. The nomogram had an AUC greater than 0.7, enabling it to distinguish between patients with better or worse prognoses (Fig. 6H). The discriminant curve analysis (DCA) curve and the calibration curve showed that the nomogram could provide useful and beneficial information for predicting HCC patient prognosis in all three cohorts (Fig. 6I). Notably, the comprehensive nomogram, which combined the anoikis signature, the anoikis-related lncRNA signature, and other clinicopathological characteristics, had an AUC of 0.772, indicating excellent predictive power that was superior to that of the two models alone. Survival analysis based on clinical characteristics showed that the group with high ARI scores had a poorer prognosis, with statistically significant differences in survival curves between the high-risk and low-risk ARI subgroups of patients across different clinical characteristics. (Additional file 1: Fig. S6A–P).

In summary, we constructed a prognostic signature and nomogram based on anoikis-related lncRNAs that could better predict the prognosis of patients with liver cancer with improved accuracy. Our study represents a significant step forward in the development of personalized prognostic tools for liver cancer patients. These findings highlight the potential of incorporating lncRNAs into clinical decision-making algorithms

and support the use of our signature and nomogram as valuable tools for guiding patient management.

### GESA and analysis of the association between ARI subtypes and immune modulators

The tumor's immunological state determines an mRNA vaccine's effectiveness. To investigate the immune cell components of ARI subgroups, we analyzed the ESTI-MATEScore, ImmuneScore, StromalScore and TumorPurity to explore the tumor microenvironment (TME) in the high and low ARI groups. The immune microenvironment analysis showed significant differences in the immune status between the two groups. Compared to the high ARI score group, the low ARI score group exhibited higher ESTI-MATEScores, ImmuneScores, and StromalScores and lower TumorPurity. (Additional file 1: Fig. S7A–D and Additional file 2: Table S10) Furthermore, single-sample GSEA (ssGSEA) was conducted to rank the 28 signature genes that were previously identified in the TCGA cohort analysis. As shown in Fig. 7A, the immune cells were classified into two clusters, and there was a noticeable difference in the distribution of immune cells between the two groups. Immune scores were significantly higher in the low-risk cluster than in the high-risk cluster. CD8 + T cells, inflammation-promoting cells, Th1 cells, TILs, HLA, Tfh cells, B cells, pDCs, neutrophils, T helper cells, DCs, and macrophages were higher in the low-risk cluster than in the high-risk cluster. These findings suggested that the ARI subtype reflects the immune status of HCC and may be used to select suitable individuals for mRNA immunization. Patients with low-risk scores, also considered to have "immunologically hot" tumors, may have more immune cell infiltration and better prognosis after receiving mRNA vaccinations containing these antigens. In the high-risk cluster, the percentages of NK cells, CD8 T cells, inflammation-promoting cells, TILs, B cells, T helper cells, and Treg cells and the levels of check-point molecules and T-cell costimulatory molecules were lower. We can speculate that the low-ARI subgroup had more tumor antigens and immune cell infiltration, which indicates that mRNA vaccination may be more beneficial and more effective in the low-ARI group than in the high-ARI group.

The results also showed that increased MHC, costimulatory, and adhesion molecule expression was found in the low-ARI group. (Fig. 7B) In further assessments of the association between HLA gene sets and the two ARI subtypes, patients in the ARI-low group had significantly higher expression levels than patients in the ARI-high group (Fig. 7C). In summary, ARI subtypes reflect the immune microenvironment and ARI score and thus can be used to select suitable individuals for mRNA immunization.

### Association of ARI with somatic mutations and immune checkpoints in HCC

To explore mutations in HCC patients, we compared the somatic mutation frequencies of the ARI-low and ARI-high subgroups in the TCGA cohorts by using the software "maftools". Figure 7D, E shows that the ARI-high group had a greater tumor mutation load than the low-risk score group. The mutant groups exhibited significantly higher ARIs than the wild-type groups for all tested altered genes. TP53 was expressed at a higher level in the mutant group than in the wild-type group (38% vs. 15%). These findings provide a new perspective for investigating the effects of anoikis modification on the TME, immune checkpoint blockade treatment, and somatic mutations in tumors.
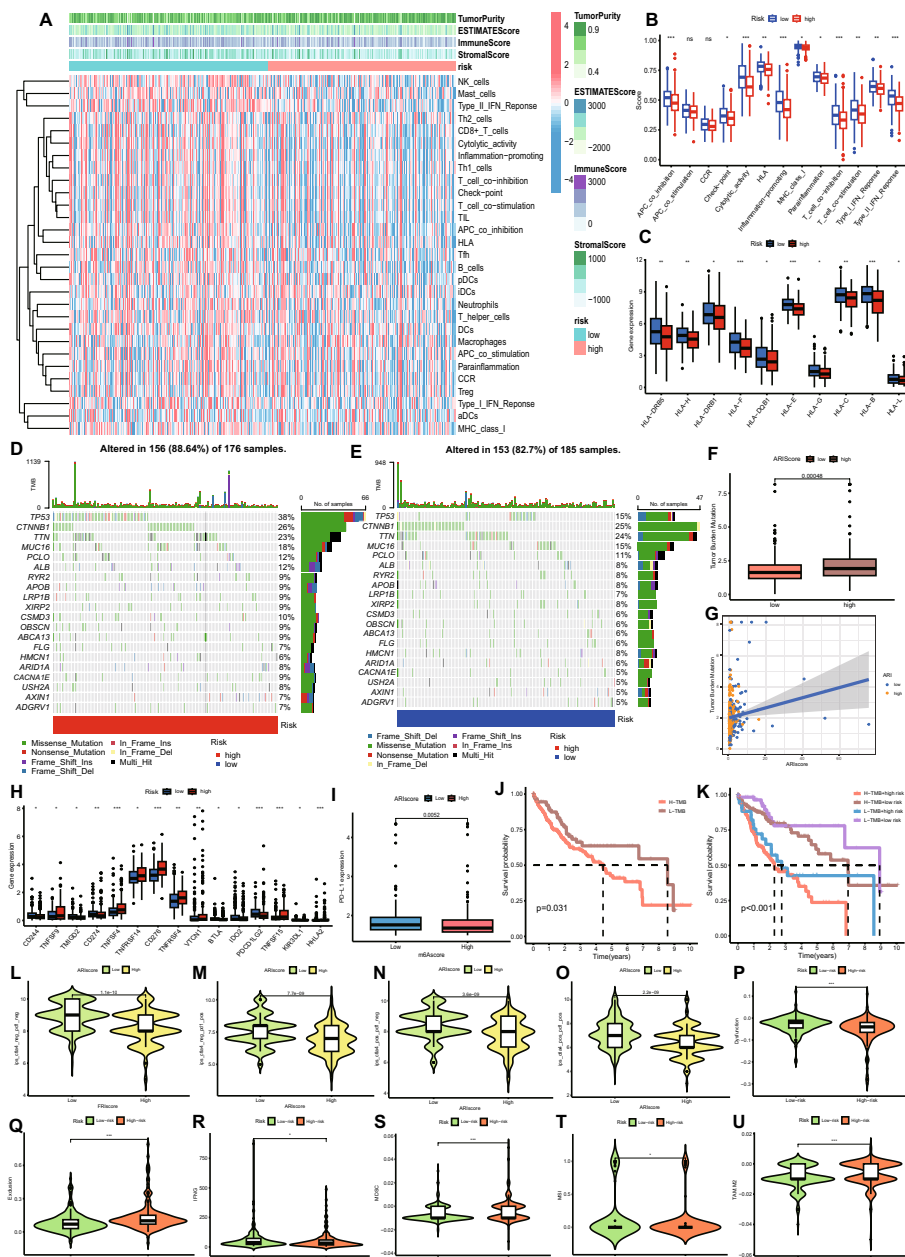
**Fig. 7** **A** Immune cell infiltration landscape map under ARI classification in the TCGA cohort. **B** Differences in immune-related pathways between ARI high-risk group and low-risk group. **C** Gene expression differences in HLA-associated gene sets within two apparently distinct ARI risk groups. **D** Waterfall plot of tumor somatic mutations in the ARI high-risk group. **E** Waterfall plot of tumor somatic mutations in the ARI low-risk group. **F** Comparison of tumor mutation burden between high and low risk groups. **G** Correlation between mutation burden and ARI score in TCGA database. **H** Expression of immune checkpoint molecules in ARI high- and low-risk groups. **I** Differential PD-L1 expression in ARI high- and low-risk groups. **J** Survival analysis in high and low tumor mutational burden in the TCGA database. **K** Survival analysis in TCGA database stratified by high and low tumor mutation burden combined with ARI high risk group. **L–U** Correlation analysis of ips, dysfunction, exclusion, IFNG, MDSC, MSI and TAM.M2 in ARI high and low risk groups in TCGA database

Furthermore, we observed a significant association between a higher ARI and higher TMB in our patient cohort. The tumors with a high-risk score were significantly correlated with higher TMB, as demonstrated by TMB quantification analysis (Fig. 7F and Additional file 2: Table S8). There was also a significant positive association between the ARI and TMB (Fig. 7G). Accumulating data have shown that patients with high TMB status have a longer-lasting clinical response to immune checkpoint inhibitor immunotherapy and anti-PD-L1 immunotherapy [26].

We analyzed the expression of immune checkpoint molecules in the low-risk and high-risk groups. The results indicated that CD244, TMIGD2, CD274, BRLA, IDO2, and PDCD1LG2 were highly expressed in the low-risk group, whereas TNFSF9, TMFSF9, TMFRSF4, VTCN1, TMFRSF15, and HHLA2 were highly expressed in the high-risk group, suggesting different expression levels of immune checkpoints in the ARI high- and low-risk groups (Fig. 7H). This finding suggests that the patient risk determined based on the ARI can be used to make appropriate immune checkpoint treatment decisions. Therefore, these findings suggest that variations in tumor anoikis regulatory patterns may play a critical role in mediating the clinical response to anti-PD-L1 immunotherapy.

Additionally, we compared the expression levels of PD-L1 in the high and low ARI groups. The results showed that patients with a low-risk score also had a high level of PD-L1 expression, indicating that patients with low ARI scores would likely benefit from anti-PD-L1 treatment (Fig. 7I). These findings indirectly confirmed the benefits of ARI in predicting the outcome of immunotherapeutic interventions. Furthermore, TMB has been suggested as a biomarker of ICI response [31]. Tumors are typically classified into two categories based on their somatic TMB: those with high TMB and a greater likelihood of responding to ICIs and those with low TMB and a lower likelihood of response. Previous studies have shown that patients with higher somatic TMB who are treated with anti-PD-L1 immune checkpoint blockade agents tend to experience improved responses, long-term survival, and long-lasting therapeutic benefits [31]. Furthermore, our study calculated the differences in the sensitivity to 16 representative chemotherapeutic agents between the high- and low-risk ARI groups to guide precise chemotherapy (Additional File 1: Fig. S8).

In addition, our research revealed that patients with a combination of a low ARI and high mutation load had a significant survival benefit. Patients with a low mutational load had better survival rates than those with a high mutational load (Fig. 7J). When the ARI score and mutation load were analyzed together, the results showed that patients with low tumor mutation load and low ARI score demonstrated the best survival prognosis, while patients with high tumor mutation load and high risk had the worst survival, and those in the high mutation load+low risk and low mutation load+high risk groups had intermediate survival. (Fig. 7K) These findings suggest that ARI combined with tumor mutation burden may be a more useful biomarker for predicting immunotherapeutic efficacy in patients with HCC. To gauge the relationship of ARI with immunogenicity, we further used IPS analysis. The ARI-low group demonstrated higher immune-related molecular biomarkers of IPS, IPS-CTLA4-PD1, dysfunction, exclusion, IFNG, MDSCs, MSI, and M2 TAMs. (Fig. 7I–U).

Overall, the HCC mutation analyses in this study revealed that the ARI high-risk group had a greater tumor mutation load than the ARI low-risk group and that variations in tumor anoikis regulatory patterns may play a critical role in mediating the clinical response to anti-PD-L1 immunotherapy, suggesting that ARI combined with tumor mutation burden

may be a more useful biomarker for predicting immunotherapeutic efficacy in patients with HCC. Additionally, the expression of immune checkpoint molecules and PD-L1 was found to be different in the ARI high- and low-risk groups, which suggests that the patient ARI risk score may be used to make appropriate decisions regarding immune checkpoint treatment.

## Discussion

mRNA vaccines had an impressive impact during the COVID-19 pandemic, leading to a surge in preclinical and clinical research in both oncology and infectious diseases [32]. Hepatocellular carcinoma (HCC) is a particularly lethal cancer with a complex molecular profile and limited treatment options once diagnosed. Although the development of mRNA vaccines has revolutionized the therapeutic treatment of HCC, the effects of mRNA cancer vaccines in HCC patients are still suboptimal and are not yet understood. Malignant cells produce both tumor-specific and non-tumor-specific antigens, making mRNA cancer vaccines a potential immunotherapy for the treatment of malignancies. Both types of antigens could be targeted with mRNA vaccines to induce tumor regression in preclinical models and humans [33]. However, most cancer vaccines under investigation are based on peptides representing only a single tumor-associated antigen, which could lead to the selection of T cells with low-affinity T-cell receptors (TCRs), or severe autoimmune toxicities can result when antigens are targeted with high-affinity engineered TCRs. On the other hand, tumor-specific neoantigens can arise from non-synonymous somatic mutations that result in the presentation of mutated peptides on the cell surface, where they can be recognized by T cells. Nevertheless, since neoantigen-specific T cells recognize peptides unique to tumors, they are not influenced by central tolerance and should not cause autoimmunity. Therefore, developing neoantigen strategies is a promising goal for immune cell-based cancer therapy.

The discovery of possible tumor-associated antigens (TAAs) and tumor-specific antigens (TSAs) is essential for the development of mRNA vaccines. Huang et al. examined prospective tumor antigens for the development of an mRNA vaccine by identifying genes that were mutated, amplified, or overexpressed in pancreatic adenocarcinoma and cholangiocarcinoma [34, 35]. Gui et al. identified potential antigens expressed in bladder cancer that could be used to develop mRNA vaccines for bladder cancer treatment and constructed a prognostic risk score model based on anoikis, which could be vital for developing personalized oncology therapeutic strategies, and these results enhanced the understanding of tumor mutation, tumor death mechanisms, and the tumor microenvironment [36]. In our study, we integrated the aforementioned methods to discover eight candidate tumor antigens (AURKA, CCNB1, CDK1, DNASE1L3, KPNA2, PRC1, PTTG1, and UBE2S) by identifying important genes that were amplified, mutated, and overexpressed in HCC. The oncogenesis and prognosis of HCC were also found to be intimately linked to these seven antigens.

It is worth mentioning that not only was there a significant correlation between high expression of tumor antigens and poor prognosis in HCC, but this expression could also cause the recruitment of APCs, which further supports their viability as mRNA vaccine antigen candidates. Du et al. cited significant evidence supporting the potential of AURKA as a therapeutic target for cancer [37]. The combination of

Zheng *et al. Journal of Big Data*     (2023) 10:129

Page 24 of 32

AURKA-targeting inhibitors and immunotherapy reduced Myc expression in tumor cells, enhanced HLDA1 and p53 protein levels, increased autophagy, caused apoptosis (neurofibroma), and significantly inhibited the growth of melanoma [38–40]. In a study conducted by another research team, alisertib promoted the development of an anticancer immune microenvironment by reducing the number of myeloid-derived suppressor cells and increasing the activity of CD8 + and CD4 + T lymphocytes, which showed alisertib combined with an anti-PD-L1 antibody exhibited strong synergistic effects [41]. According to a bioinformatics study by Si et al., CCNB1 is a marker for variant genes in hepatocellular carcinoma and is linked to variation in p53. The expression of CCNB1 is associated with a higher proportion of T effector cells, such as CD8 + T cells, which include cytotoxic T cells, natural killer cells, and other antitumor immune cells in the periphery [38, 42]. Reports have shown a correlation between CDK1 activity and patient prognosis, and CDK1 expression is increased in numerous human malignant tumor tissues, making it a crucial molecular target. Xiao and Li et al. demonstrated the impact of DNASE1L3 on HCC development, apoptosis, and glucose metabolism reprogramming. Targeting DNASE1L3 may be a viable therapeutic option for HCC, and detecting elevated levels of DNASE1L3 in the blood may help doctors diagnose hepatocellular carcinoma caused by the hepatitis B virus [43]. KPNA2 plays a critical role in HCC cells, and an inability of KPNA2 to import PLAG1 into the nucleus is a strong predictor of poor survival in HCC patients after hepatectomy [44]. Chen et al. discovered that PRC1 is markedly elevated in HCC tumors and highly related to early HCC recurrence [45]. In 46 HCC tumor samples, Huang et al. found that PTTG1 was commonly upregulated and strongly associated with PTTG3P [46]. This research revealed that UBE2S was significantly expressed in HCC, particularly in the nucleus, and was linked to HCC patients' clinical prognoses. Through its nuclear localization signal (NLS), UBE2S enters the nucleus, interacts with TRIM28, and promotes HCC development by ubiquitinating p27 [47]. Anoikis is an inherent protective mechanism in organisms that occurs when the connection between epithelial cells and the extracellular matrix (ECM) is broken, preventing the readhesion of dead or dying cells to an inappropriate location [48]. This mechanism is essential for the organism to respond to harmful intracellular or extracellular stimuli, such as viral infections, DNA damage, exposure to toxins, metabolic, oxidative, or hypoxic stressors, or loss of anchoring [49]. However, cancer cells may exhibit abnormal execution of anoikis, leading to tumor invasion, migration, distant organ metastases, and treatment resistance. In certain pathological circumstances, such as malignancies, cells may develop resistance to anoikis. Multiple studies have shown that tumor cell resistance to anoikis enables them to metastasize away from the primary tumor location via the lymphatic and circulatory systems, where they may continue to proliferate [50]. This resistance may also facilitate immune evasion and contribute to the development of treatment resistance. Although it has been shown that anoikis is involved in the invasion and metastasis of various solid tumors, few studies have thoroughly investigated the role of anoikis-related genes and lncRNAs in HCC. Additionally, the TME infiltration mediated by the integrated effects of distinct ARGs and anoikis-related lncRNAs has been underrecognized in the HCC TME. Determining the significance of specific anoikis regulatory patterns in TME cell invasion will enhance our

knowledge of the TME antitumor immune response and guide the development of more efficient immunotherapy approaches.

Using an unsupervised consistency clustering algorithm, patients with HCC were categorized, and 28 ARGs were selected to identify two distinct regulatory patterns of anoikis. The two clusters were then studied and assessed, revealing a difference in survival time between the two clusters of HCC patients. Differences in immune cell infiltration and immunological targets were examined between the subgroups, and the two groups with different patterns of anoikis and cellular infiltration into the tumor microenvironment displayed considerable differences. The results showed that ARG cluster A had a poor survival time, and anoikis-related genes were highly expressed in this cluster. In comparison to cluster B, cluster A exhibited a greater degree of immune cell infiltration into the tumor stroma, including infiltration of activated CD4 T cells, activated dendritic cells, CD56$^+$ natural killer cells, immature dendritic cells, MDSCs, natural killer T cells, natural killer cells, plasmacytoid dendritic cells, regulatory T cells, T follicular helper cells, and type 2 T helper cells.

Extensive research has shown that programmed cell death serves two roles in modulating the TME. In addition to its antitumor action, programmed cell death also indirectly facilitates immune escape by generating an immunosuppressive microenvironment [51]. Consistent with previous results, KEGG analysis showed higher enrichment of cell cycle pathways in group A (high expression of anoikis genes) than in group B, but the activity of metabolic pathways was lower in group A than group B, suggesting that group A had a worse prognosis. Although group A had a high level of immune cell infiltration, its survival was worse than that of group B, and we can reasonably speculate that the high expression of anoikis-resistance genes in cluster A created a special immunosuppressive microenvironment that made the infiltrating immune cells dysfunctional in the stroma. Previous research has shown that the immunosuppressive features of the tumor microenvironment facilitate cytotoxic immune cell exhaustion and death, promoting the development of protumoral immune cells such as Tregs, M2 macrophages, and myeloid-derived suppressor cells (MDSCs) [51]. This supports our findings, and it was therefore not surprising that a comprehensive exploration of the cellular infiltration patterns in the tumor microenvironment defined by the anoikis regulatory model revealed that cluster A had a higher level of active immune cells but poorer survival. This again demonstrates the complexity and heterogeneity of the role of types of apoptosis, including anoikis, in the tumor microenvironment, but further exploration of the role in the hepatocellular carcinoma microenvironment is needed.

Notably, recent research has focused on understanding how anoikis affects the activation of the immune microenvironment. There is now evidence that altered cancer metabolism is due, in part, to anoikis resistance. Alteration of kynurenine (Kyn, a key metabolic component) triggered a pathway that was increased by TNHCC in suspension culture. In general, Kyn can reduce immune surveillance, but in glioma, it reduces antitumor immune responses and enhances tumor cell survival and motility by acting in both a paracrine and autocrine mode. Kyn has been shown to act as an endogenous ligand to activate AhR in both immune cells and tumor cells [52]. Previous studies have suggested that HCC tumors with a hypermetabolic phenotype

generally have a favorable prognosis. Research conducted by Yang on the metabolic subtypes of HCC found that HCCs with high metabolic activity had distinct metabolic signatures and were similar to differentiated nonproliferative HCCs with low AFP expression and a favorable prognosis. The subtype with low metabolic activity, which was related to immunology profiles and high expression of immune checkpoint genes, showed pharmacological sensitivity to CTLA4 inhibitors and cabozantinib. This class played a minor role in the formation of metabolic signatures. The intermediate metabolic activity subtype, which was associated with a higher level of AFP and a poorer prognosis, had lower enrichment of metabolic signatures than the high metabolic activity subtype but higher enrichment than the low metabolic activity subtype [53]. Similar results were also found in the study by Huo et al. C1 exhibited the highest metabolic activity, the best prognosis, and the most eosinophils and natural killer cells. C2 had the lowest metabolic activity, the worst prognosis, the highest TP53 mutation rate, the highest immune checkpoint expression, and substantial regulatory T-cell infiltration. C3 displayed strong neutrophil and macrophage infiltration, moderate metabolic activity, and the highest LRP1B mutation rate [54].

Given that mRNA vaccines have shown benefits in only a relatively specialized group of cancer patients, we used the expression profiles of anoikis-lncRNAs to divide HCC patients into high-risk and low-risk subtypes to select the most promising candidates for the vaccine. Additionally, the ARG-based signature contained numerous genes, and there is a high degree of heterogeneity in tumor microenvironments, thus limiting the role of the signature in predicting liver cancer prognosis. Therefore, we constructed a prognostic model based on the ARI. Among the ARGs related to prognosis that we identified, accumulating evidence has shown that the ARGs used to build the ARI risk model play a crucial role in cancer. For instance, Zhang et al. found that the expression of PMSB8-AS1 was elevated in PC tissues and cell lines, and it was inversely correlated with survival in people with PC. Studies of the mechanism of PMSB8-AS1 found that the lncRNA promotes pancreatic cancer progression via STAT1 by sponging miR-382–3p and is involved in the regulation of PD-L1 [55]. As an upregulated lncRNA in liver cancer, ZFPM2-AS1 was identified as a competing endogenous RNA (ceRNA), competitively binding to miR-139 and regulating GDF10 expression. The authors suggested that enhancement of the malignant phenotype in liver cancer occurs via the ZFPM2-AS1/miR-139/GDF10 pathway, which has potential as a target for treatment in HCC [56]. A similar result has been found in gastric cancer: Kong et al. analyzed ZFPM2-AS1 and found that it controls the progression of gastric cancer and uncovered a new ZFPM2-AS1/MIF/p53 signaling pathway, providing insight into the molecular processes underlying the tumorigenicity of certain malignant gastric cells [57]. In gastric cancer, exosome-transferred LINC01559 stimulated the phosphatidylinositol 3-kinase/AKT serine/threonine kinase (PI3K/AKT) pathway by upregulating PGK1 and downregulating PTEN. Other studies have shown that LINC01559 recruits insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2) to stabilize ZEB1 mRNA in GC cells, thus upregulating ZEB1 [58, 59]. Other studies have shown that LINC01559 accelerates pancreatic cancer progression by relying on YAP [60]. Our bioinformatics analysis also identified these lncRNAs as anoikis-related prognostic markers; these results deepened the understanding of the

tumor microenvironment related to anoikis in HCC and supported the practical significance of the ARI model.

To verify the model, we conducted KM analysis, ROC curve analysis, random sampling validation, and subgroup analysis in the studied cohorts. These analyses demonstrated that the novel signature had strong prognostic power in predicting the outcome of HCC. Additionally, a statistically significant correlation between the ARI risk score and immune cell infiltration and immunotherapeutic efficacy was observed. The ARI risk score was a valid and practical factor that could be used to characterize the heterogeneity of tumor anoikis regulation patterns and to identify TME infiltration patterns and tumor immune phenotypes. Furthermore, integrated analysis confirmed that the risk score was an effective predictive indicator for liver cancer. Patients with a lower risk score had a longer overall survival rate, as assessed in the overall, test, and validation groups. The ROC curve analysis showed that the gene signature based on the lncRNAs had high sensitivity and specificity, demonstrating the excellent predictive efficiency of our model.

To further support the practical application and predictive validity of the model in clinical practice, we created a nomogram jointly considering the ARG and ARI scores as well as clinicopathological factors, including pathological stage, clinical stage, age, and sex. The combined ROC analysis demonstrated that the predictive validity of the combined nomogram was better than that of the two separate scoring systems alone. As a result, the combined nomogram provided a more accurate prediction of 1-, 2-, and 3-year survival in hepatocellular cancer, with an AUC of 0.723–0.757.

Next, we analyzed the expression of infiltrating immune cells in the ARI high-risk and low-risk groups. We found that the types and quantities of infiltrating immune cells in the low-risk group were significantly higher than those in the high-risk group, indicating that the low-risk group had an immunologically hot phenotype; the low-risk group was also associated with improved survival. The low-risk group was characterized by elevated levels of various immune-related factors, including APC coinhibitory molecules, checkpoint molecules, cytolytic molecules, HLA molecules, proinflammatory molecules, MHC class I molecules, parainflammatory markers, T-cell coinhibitory molecules, T-cell costimulatory molecules, and indicators of type II IFN responses. These findings further supported the hypothesis that the low-risk group possessed a robust immune profile. Moreover, we found high expression levels of HLA receptors in the low-risk group, further emphasizing the strong immune activation properties of this group. Finally, our results revealed a significant correlation between the ARI score and tumor mutation load: the low-risk ARI group had low TMB levels and high levels of immune cell infiltration, a characteristic of an immune hot phenotype, resulting in a more favorable survival prognosis than the high ARI score group. Tumor response to immune checkpoint therapy has been observed to vary based on the degree of immune cell infiltration and the mutational burden of the tumor. We investigated the relationship between the ARI and immune checkpoint signaling, and the results suggested that anoikis can influence the efficacy of immune checkpoint-based treatments. Generally, cancers classified as "hot", such as melanoma and lung cancer, exhibit greater responsiveness to therapy than "cold" tumors, such as pancreatic and prostate cancers. Tumors with a high mutational burden are believed to have more favorable outcomes with ICD therapy. However,

our work showed that patients in the low-ARI group had higher expression of PD1 and PD-L1, suggesting that ICD treatment may have better outcomes in this group.

Combining immune therapy with the exploration of tumor neoantigens to develop tumor antigen vaccines may enhance the efficacy of existing treatments in ARI high-risk populations if ICD is ineffective due to very immune-suppressive microenvironments or matrix barriers expressed in HCC. For instance, in a study using DCs pulsed with autologous tumor lysates in 31 treated patients, 12.9% had a partial response, 54.8% had stable disease, and the 1-year survival rate was 10.7% ($\pm 9.4$) compared to 63.3% ($\pm 12.0$) with maintenance therapy. In addition, the administration of monoclonal antibodies, adoptive T-cell therapy, or vaccinations combined with the administration of gene therapy vectors encoding monoclonal antibodies and/or immunostimulatory cytokines are effective methods for treating HCC. Combining existing tumor therapy models and ARI-based stratification in the development of tumor antigen vaccines may be a promising strategy for achieving a synergistic immunostimulatory effect, leading to significant clinical benefits of vaccine therapies.

In summary, this study analyzed the expression of antigens in liver cancer, providing a new perspective for developing liver cancer vaccines. Unsupervised clustering of ARGs related to anoikis and prognosis and construction of the ARI model revealed innovative ideas for enhancing the clinical response of patients to immunotherapy and identified distinct tumor anoikis phenotypes that can promote precision cancer immunotherapy. This study has some limitations; for example, the feasibility and effectiveness of antigen vaccine development has yet to be verified through experiments, and further follow-up studies are needed to validate its potential due to the absence of available RNA sequencing data of HCC immunotherapy cohorts and clinical data. In addition, machine learning and feature selection algorithms can be further employed to optimize models, enabling better development of liver cancer vaccines and more accurate predictions of the prognosis of liver cancer patients [61, 62].

## Conclusion

Our study identified anoikis-related regulators, including CCNB1, CDK1, DNASE1L3, KPNA2, PRC1, PTTG, and UBE2S, as promising tumor antigens for the development of mRNA vaccines for HCC. Moreover, we identified four distinct HCC subtypes related to anoikis and suggested that these subtypes could guide the selection of appropriate patients for immunotherapies. We also developed a novel lncRNA scoring model related to anoikis that could help estimate immunotherapy response, design mRNA vaccines for HCC therapy, and identify the most suitable patients for immunization. These findings have significant implications for the development of targeted immunotherapies for HCC patients.

**Abbreviations**

| | |
|---|---|
| HCC | Hepatocellular carcinoma |
| GEO | Gene Expression Omnibus |
| TCGA | The Cancer Genome Atlas |
| LASSO | Least absolute shrinkage and selection operator |
| HBV | Hepatitis B virus |
| HCV | Hepatitis C virus |
| ICIs | Immune checkpoint inhibitors |

| TAAs | Tumor-associated antigens |
|---|---|
| ACT | Adoptive cell therapy |
| ICBs | Immune-Checkpoint Blockade |
| mRNA | Messenger RNA |
| ECM | Extracellular matrix |
| lncRNAs | Long non-coding RNAs |
| EMT | Epithelial–mesenchymal transition |
| DEGs | Differential expressed genes |
| OS | Overall survival |
| RFS | Relapse-free survival |
| TTIMER | The tumor immune estimation resource |
| GSVA | Gene set variation analysis |
| ARGs | Anoikis related genes |
| PCA | Principal component analysis |
| ARI | : Anoikis risk indices |
| ssGSEA | Single sample gene set enrichment analysis |
| HR | Hazard ratio |
| KEGG | Kyoto Encyclopedia of genes and genomes |
| TSAs | Tumor specific antigens |
| ceRNAs | Competing endogenous RNAs |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-023-00803-7.

---

**Additional file 1: Figure S1.** Identification of tumor antigen genes associated with HCC prognosis in TCGA and meta cohorts. The KM curve of the overall survival for different group with different expression of (A) AURKA. (B) CCNB1. (C) CDK1. (D) KPNA2. (E) PRC1. (F) PTTG1. (G) UBE2S in TCGA cohorts. The KM curve of the disease free survival for different group with different expression of (H) AURKA. (I) CCNB1. (J) CDK1. (K) KPNA2. (L) PRC1. (M) PTTG1. (N) UBE2S in TCGA cohorts. The KM curve of the overall survival for different group with different expression of (O) AURKA (P) CCNB1 (Q) CDK1 (R) KPNA2 (S) PTTG1 (T) UBE2S in meta cohorts (TCGA + GEO). **Figure S2.** Antigten correlation Analysis of the Immune Microenvironment of Seven Potential Tumor Antigens, including tumor purity, B cell, CD8 + T cell, CD4 + T cell, Macrophage, Neutrophil, Dentritic cell. (A) AURKA. (B) CCNB1. (C) CDK1. (D) KPNA2. (E) PRC1. (F) PTTG1. (G) UBE2S. **Figure S3.** OS analysis for 28 Anoikis-related genes. **Figure S4.** PCA analysis for four cohorts. (A) PCA analysis of all genes in the TCGA cohort. (B) PCA analysis of anoikis-related genes in the TCGA cohort. (C) PCA analysis of anoikis-related lncRNAs in the TCGA cohort. (D) PCA analysis of anoikis-related risk lncR-NAs in the TCGA cohort. **Figure S5.** (A) Forest plot of univariate analysis with risk score. (B) Forest plot of multivariate analysis with risk score. **Figure S6.** Survival curves of different clinical feature subgroups under ARG score for (A) patients with age > 65 years (B) patients with age < = 65 (C) female patients subgroup (D) female patients subgroup (E) patients with G1-G2 subgroup (F) patients with G3-G4 subgroup (G) patients with M0 subgroup (H) patients with M1 subgroup (I) patients with Mx subgroup (J) patients with N0 subgroup (K) patients with Stage I subgroup (L) patients with Stage II subgroup (M) patients with Stage III-IV subgroup (N) patients with T1 subgroup (O) patients with T2 subgroup (P) patients with T3-T4 subgroup. **Figure S7.** The immune microenvironment analysis of high and low ARI score group in HCC. (A) ESTIMATEScores, (B) ImmuneScores, (C) StromalScores and (D) TumorPurity. **Figure S8.** Sensitivity for 16 representative chemotherapeutic agents based on ARI score.

**Additional file 2: Table S1.** Gene lists for Venn Plot. **Table S2.** Univariate analysis for 28 differential ARGs. **Table S3.** ARG cluster in meta cohorts. **Table S4.** GO analysis. **Table S5.** KEGG analysis. **Table S6.** ARG genes cluster. **Table S7.** Tumor microenvironment score between ARG cluster in meta cohorts: StromalScore, ImmuneScore, ESTIMATES-core and TumorPurity. **Table S8.** Coefficient of ARI signature models. **Table S9.** Univariate and multivariate analysis of ARI score and other clinical characteristics. **Table S10.** Tumor microenvironment score between ARI cluster in meta cohorts: StromalScore, ImmuneScore, ESTIMATEScore and TumorPurity. **Table S11.** Summary of Packages and statistical analysis.

---

### Availability of data and materials
The data discussed in this article can be accessed on the relevant web servers and can be used by any researcher for non-commercial purposes while maintaining the confidentiality of the participants. For more information, interested parties can contact the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declared no potential competing interests.

## References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Kudo M, Finn RS, Qin S, Han KH, Ikeda K, Piscaglia F, et al. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. Lancet. 2018;391(10126):1163–73.
3. Bruix J, Qin S, Merle P, Granito A, Huang YH, Bodoky G, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet. 2017;389(10064):56–66.
4. Tian Y, Xiao H, Yang Y, Zhang P, Yuan J, Zhang W, et al. Crosstalk between 5-methylcytosine and N(6)-methyladenosine machinery defines disease progression, therapeutic response and pharmacogenomic landscape in hepatocellular carcinoma. Mol Cancer. 2023;22(1):5.
5. Zhang Y, Wang M, Chen Q, Deng Y, Chen J, Dai Y, et al. Adverse events of immune checkpoint inhibitor-based therapies for unresectable hepatocellular carcinoma in prospective clinical trials: a systematic review and meta-analysis. Liver Cancer. 2023. https://doi.org/10.1159/000528698.
6. Cheng AL, Hsu C, Chan SL, Choo SP, Kudo M. Challenges of combination therapy with immune checkpoint inhibitors for hepatocellular carcinoma. J Hepatol. 2020;72(2):307–19.
7. Simpson CD, Anyiwe K, Schimmer AD. Anoikis resistance and tumor metastasis. Cancer Lett. 2008;272(2):177–85.
8. Meng Q, Lu YX, Wei C, Wang ZX, Lin JF, Liao K, et al. Arginine methylation of MTHFD1 by PRMT5 enhances anoikis resistance and cancer metastasis. Oncogene. 2022;41(32):3912–24.
9. Paoli P, Giannoni E, Chiarugi P. Anoikis molecular pathways and its role in cancer progression. Biochim Biophys Acta. 2013;1833(12):3481–98.
10. Kochetkova M, Kumar S, McColl SR. Chemokine receptors CXCR4 and CCR7 promote metastasis by preventing anoikis in cancer cells. Cell Death Differ. 2009;16(5):664–73.
11. Zhang HF, Hughes CS, Li W, He JZ, Surdez D, El-Naggar AM, et al. Proteomic screens for suppressors of anoikis identify IL1RAP as a promising surface target in ewing sarcoma. Cancer Discov. 2021;11(11):2884–903.
12. Shi T, Zhang C, Xia S. The potential roles and mechanisms of non-coding RNAs in cancer anoikis resistance. Mol Cell Biochem. 2022;477(5):1371–80.
13. Grinchuk OV, Yenamandra SP, Iyer R, Singh M, Lee HK, Lim KH, et al. Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. Mol Oncol. 2018;12(1):89–113.
14. Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. N Engl J Med. 2008;359(19):1995–2004.
15. Moeini A, Torrecilla S, Tovar V, Montironi C, Andreu-Oller C, Peix J, et al. An immune gene expression signature associated with development of human hepatocellular carcinoma identifies mice that respond to chemopreventive agents. Gastroenterology. 2019;157(5):1383-97.e11.
16. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13(4):163.
17. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016:baw100.
18. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
19. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
20. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. 2019;47(W1):W556–60.
21. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER20 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res. 2020;48(W1):W509–14.
22. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
23. Yu G, Wang L-G, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
24. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141.

25.  Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–21.
26.  da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
27.  Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017;18(1):248–62.
28.  Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, et al. Large-scale public data reuse to model immunotherapy response and resistance. Genome Med. 2020;12(1):21.
29.  Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018;24(10):1550–8.
30.  Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS ONE. 2014;9(9): e107468.
31.  Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol Cancer Ther. 2017;16(11):2598–608.
32.  Miao L, Zhang Y, Huang L. mRNA vaccine for cancer immunotherapy. Mol Cancer. 2021;20(1):41.
33.  Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines—a new era in vaccinology. Nat Rev Drug Discov. 2018;17(4):261–79.
34.  Huang X, Tang T, Zhang G, Liang T. Identification of tumor antigens and immune subtypes of cholangiocarcinoma for mRNA vaccine development. Mol Cancer. 2021;20(1):50.
35.  Huang X, Zhang G, Tang T, Liang T. Identification of tumor antigens and immune subtypes of pancreatic adenocarcinoma for mRNA vaccine development. Mol Cancer. 2021;20(1):44.
36.  Gui CP, Li JY, Fu LM, Luo CG, Zhang C, Tang YM, et al. Identification of mRNA vaccines and conserved ferroptosis related immune landscape for individual precision treatment in bladder cancer. J Big Data. 2022;9(1):88.
37.  Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in cancer: molecular mechanisms and opportunities for cancer therapy. Mol Cancer. 2021;20(1):15.
38.  Horwacik I, Durbas M, Boratyn E, Węgrzyn P, Rokita H. Targeting GD2 ganglioside and aurora A kinase as a dual strategy leading to cell death in cultures of human neuroblastoma cells. Cancer Lett. 2013;341(2):248–64.
39.  Durbas M, Pabisz P, Wawak K, Wiśniewska A, Boratyn E, Nowak I, et al. GD2 ganglioside-binding antibody 14G2a and specific aurora A kinase inhibitor MK-5108 induce autophagy in IMR-32 neuroblastoma cells. Apoptosis. 2018;23(9–10):492–511.
40.  Liu Y, Hawkins OE, Vilgelm AE, Pawlikowski JS, Ecsedy JA, Sosman JA, et al. Combining an aurora kinase inhibitor and a death receptor ligand/agonist antibody triggers apoptosis in melanoma cells and prevents tumor growth in preclinical mouse models. Clin Cancer Res. 2015;21(23):5338–48.
41.  Yin T, Zhao ZB, Guo J, Wang T, Yang JB, Wang C, et al. Aurora A inhibition eliminates myeloid cell-mediated immunosuppression and enhances the efficacy of anti-PD-L1 therapy in breast cancer. Cancer Res. 2019;79(13):3431–44.
42.  Si T, Huang Z, Jiang Y, Walker-Jacobs A, Gill S, Hegarty R, et al. Expression levels of three key genes CCNB1, CDC20, and CENPF in HCC are associated with antitumor immunity. Front Oncol. 2021;11: 738841.
43.  Li B, Ge YZ, Yan WW, Gong B, Cao K, Zhao R, et al. DNASE1L3 inhibits proliferation, invasion and metastasis of hepatocellular carcinoma by interacting with β-catenin to promote its ubiquitin degradation pathway. Cell Prolif. 2022;55(9): e13273.
44.  Zeng F, Luo L, Li D, Guo J, Guo M. KPNA2 interaction with CBX8 contributes to the development and progression of bladder cancer by mediating the PRDM1/c-FOS pathway. J Transl Med. 2021;19(1):112.
45.  Chen J, Rajasekaran M, Xia H, Zhang X, Kong SN, Sekar K, et al. The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/β-catenin signalling pathway. Gut. 2016;65(9):1522–34.
46.  Huang JL, Cao SW, Ou QS, Yang B, Zheng SH, Tang J, et al. The long non-coding RNA PTTG3P promotes cell growth and metastasis via up-regulating PTTG1 and activating PI3K/AKT signaling in hepatocellular carcinoma. Mol Cancer. 2018;17(1):93.
47.  Zhang RY, Liu ZK, Wei D, Yong YL, Lin P, Li H, et al. UBE2S interacting with TRIM28 in the nucleus accelerates cell cycle by ubiquitination of p27 to promote hepatocellular carcinoma development. Signal Transduct Target Ther. 2021;6(1):64.
48.  Kakavandi E, Shahbahrami R, Goudarzi H, Eslami G, Faghihloo E. Anoikis resistance and oncoviruses. J Cell Biochem. 2018;119(3):2484–91.
49.  Zhi Z, Ouyang Z, Ren Y, Cheng Y, Liu P, Wen Y, et al. Non-canonical phosphorylation of Bmf by p38 MAPK promotes its apoptotic activity in anoikis. Cell Death Differ. 2022;29(2):323–36.
50.  Zhang Z, Zhu Z, Fu J, Liu X, Mi Z, Tao H, et al. Anoikis patterns exhibit distinct prognostic and immune landscapes in osteosarcoma. Int Immunopharmacol. 2023;115: 109684.
51.  Liu J, Hong M, Li Y, Chen D, Wu Y, Hu Y. Programmed cell death tunes tumor immunity. Front Immunol. 2022;13: 847345.
52.  D'Amato NC, Rogers TJ, Gordon MA, Greene LI, Cochrane DR, Spoelstra NS, et al. A TDO2-AhR signaling axis facilitates anoikis resistance and metastasis in triple-negative breast cancer. Cancer Res. 2015;75(21):4651–64.
53.  Yang C, Huang X, Liu Z, Qin W, Wang C. Metabolism-associated molecular classification of hepatocellular carcinoma. Mol Oncol. 2020;14(4):896–913.
54.  Huo J, Cai J, Wu L. Comprehensive analysis of metabolic pathway activity subtypes derived prognostic signature in hepatocellular carcinoma. Cancer Med. 2023;12(1):898–912.
55.  Zhang H, Zhu C, He Z, Chen S, Li L, Sun C. LncRNA PSMB8-AS1 contributes to pancreatic cancer progression via modulating miR-382-3p/STAT1/PD-L1 axis. J Exp Clin Cancer Res. 2020;39(1):179.
56.  He H, Wang Y, Ye P, Yi D, Cheng Y, Tang H, et al. Long noncoding RNA ZFPM2-AS1 acts as a miRNA sponge and promotes cell invasion through regulation of miR-139/GDF10 in hepatocellular carcinoma. J Exp Clin Cancer Res. 2020;39(1):159.

Zheng *et al. Journal of Big Data*      (2023) 10:129

Page 32 of 32

57. Kong F, Deng X, Kong X, Du Y, Li L, Zhu H, et al. ZFPM2-AS1, a novel lncRNA, attenuates the p53 pathway and promotes gastric carcinogenesis by stabilizing MIF. Oncogene. 2018;37(45):5982–96.
58. Wang L, Bo X, Yi X, Xiao X, Zheng Q, Ma L, et al. Exosome-transferred LINC01559 promotes the progression of gastric cancer via PI3K/AKT signaling pathway. Cell Death Dis. 2020;11(9):723.
59. Shen H, Zhu H, Chen Y, Shen Z, Qiu W, Qian C, et al. ZEB1-induced LINC01559 expedites cell proliferation, migration and EMT process in gastric cancer through recruiting IGF2BP2 to stabilize ZEB1 expression. Cell Death Dis. 2021;12(4):349.
60. Lou C, Zhao J, Gu Y, Li Q, Tang S, Wu Y, et al. LINC01559 accelerates pancreatic cancer cell proliferation and migration through YAP-mediated pathway. J Cell Physiol. 2020;235(4):3928–38.
61. Dokeroglu T, Deniz A, Kiziloz HE. A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing. 2022;494:269–96.
62. Wang W, Feng S, Ye Z, Gao H, Lin J, Ouyang D. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. Acta Pharm Sin B. 2022;12(6):2950–62.

**Publisher's Note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.