ANNALS OF MEDICINE
2025, VOL. 57, NO. 1, 2503645
https://doi.org/10.1080/07853890.2025.2503645

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Discovery of novel diagnostic biomarkers of hepatocellular carcinoma associated with immune infiltration

Qiang Liu[a], Hua Zhang[a], Heng Xiao[a], Ao Ren[a], Ying Cai[b], Rui Liao[a], Huarong Yu[c], Zhongjun Wu[a] and Zuotian Huang[a,d]

[a]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China; [b]School of Nursing, Chongqing Medical University, Chongqing, China; [c]Chongqing Medical University, Chongqing, China; [d]Chongqing University Cancer Hospital, Chongqing, China

ABSTRACT

Objective: Diagnosis of hepatocellular carcinoma (HCC) remains challenging for clinicians. Machine learning approaches and big data analyses are viable strategies for identifying HCC diagnostic markers.

Materials and methods: In this study, we downloaded mRNA expression profiles of HCC from the GEO database and used random forest and machine learning algorithms, such as least absolute shrinkage and selection operator, to screen for reliable diagnostic genes. Disease Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Set Enrichment Analysis enrichment analyses were performed to explore differential gene functions and disease pathways. CIBERSORT was performed to calculate the immune cell infiltration of HCC and the correlation between diagnostic genes and immune cells. Cell experiments were performed to evaluate the function of R-spondin 3 (RSPO3) in HCC cells. Immunohistochemical staining was used to evaluate the protein expression of CD138, CD206 and iNOS.

Results: The results indicated that extracellular matrix protein 1 (ECM1), Niemann-Pick C1-Like 1 (NPC1L1) and RSPO3 were down-regulated in HCC compared with the normal group ($p < 0.05$), which was validated in clinical tissue samples. Moreover, ECM1, NPC1L1 and RSPO3 had high diagnostic values (AUC > 0.75) for HCC in both training and test groups. Immuno-infiltration analysis revealed that ECM1 and RSPO3 were highly positively correlated with neutrophil and macrophage M2 levels, whereas they were negatively correlated with Tregs. RSPO3-si affected cell proliferation and apoptosis in HCC. Furthermore, RSPO3 exhibited a positive correlation with tumour progression, the proportion of plasma cells and M2 macrophages in mice, while showing a negative association with M1 macrophages.

Conclusion: The present study identified ECM1, NPC1L1 and RSPO3 as new diagnostic biomarkers for HCC based on normal and diseased samples from HCC, meanwhile the pro-oncogenic function of RSPO3 and its regulation on immune infiltration have been confirmed.

ARTICLE HISTORY
Received 5 August 2024
Revised 12 November 2024
Accepted 12 December 2024

KEYWORDS
Hepatocellular carcinoma; diagnostic markers; immune cell infiltration; machine learning; GEO database

Introduction

Hepatocellular carcinoma (HCC) is the fifth most common type of cancer and the third leading cause of cancer-related deaths worldwide [1]. Due to the lack of early symptoms, a large number of patients with HCC are diagnosed at an advanced stage, resulting in a low survival rate [2]. Although computed tomography, magnetic resonance imaging, ultrasonography, positron emission tomography and angiography, biopsy and serologic analysis are the most widely used tools for hepatocellular carcinoma diagnosis, they are still limited by practitioner expertise, price and low sensitivity of serologic indicators [3]. More candidate markers for the early diagnosis of HCC need to be identified.

Transcriptome analysis is now widely used to stratify patients with HCC and determine its diagnosis and prognosis [4–6]. Alpha-fetoprotein is the gold standard diagnostic marker for HCC. However, its sensitivity and specificity are low and its expression may be

CONTACT Zhongjun Wu wzjtcy@126.com; Zuotian Huang 1351619201@qq.com Department of Hepatobiliary Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China

influenced by multiple non-HCC-related factors [7]. Several novel sensitive biomarkers, including APEX1 [8], CDCA8 [9] and H2AFY [10], have been identified through big data analysis to identify HCC early and to improve clinical outcomes. However, the early and specific diagnosis of HCC remains challenging.

Machine learning is a powerful tool that can help diagnose or predict molecules or groups of molecules for diseases or disorders based on gene expression data. Mathematical methods are applied to train a model that learns from data for a specific task, such as classification or feature selection. These techniques can help identify informative genes that can distinguish between different groups of samples, such as normal or tumour tissues [11,12], thereby facilitating the identification of potential disease-associated genes. Binder et al. used machine learning prediction and tau-based screening to identify potential Alzheimer's disease genes associated with immunity [13], and PRKAR2B and TGFBI were identified as diagnostic biomarkers for glomerular injury in diabetic nephropathy based on machine learning algorithms [14].

In this study, we identified and characterized three major diagnostic markers: extracellular matrix protein 1 (ECM1), Niemann-Pick C1-Like 1 (NPC1L1) and RSPO3 (R-spondin 3). ECM1 is a secreted protein that plays a crucial role in the extracellular matrix, influencing cellular behaviour, tissue remodelling and immune response regulation. It has been implicated in various cancers, including HCC, where it may contribute to tumour progression and metastasis. NPC1L1 is a membrane protein that facilitates cholesterol absorption in the intestines and is also involved in lipid metabolism. In the context of cancer, NPC1L1 has been linked to altered lipid metabolism, which can impact tumour growth and immune evasion. RSPO3 is a member of the R-spondin family, known for its role in enhancing Wnt signalling pathways, which are essential for cell proliferation and differentiation. Abnormal RSPO3 expression has been associated with several cancers, including its potential to modulate the tumour immune microenvironment. In this study, we analysed their relationships with the immune microenvironment, providing insights into their potential as diagnostic biomarkers for HCC.

Materials and methods

Microarray data acquisition

Gene expression data were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/), where we accessed multiple datasets derived from HCC and normal tissues. To ensure the integrity of the data, we combined these datasets and processed them using the 'sva' package in R to remove batch effects and normalize the data [15], thus obtaining a reliable training cohort for subsequent analyses. GSE62232 was employed as a validation cohort to validate differential expression and diagnostic performance.

This database study was exempted from full ethical approval by the First Affiliated Clinical Research Ethics Review Committee of Chongqing Medical University, in accordance with the relevant provisions of Article 32 of the Ethical Review Measures for Life Sciences and Medical Research Involving Human Subjects. The exemption was granted because the data utilized were publicly accessible and did not include any personally identifiable information.

Screening and functional enrichment analysis of differentially expressed genes

To identify differentially expressed genes (DEGs) between samples, we employed the limma package [16] for statistical analysis of gene expression data. This methodology involved the generation of *p*-values to evaluate the significance of observed expression differences. We controlled the false discovery rate (FDR) to establish a threshold for significance, with corrected *p*-values indicating the adjusted level of statistical significance. The error detection rate (controlled) determined the *p*-value threshold, and the corrected *p*-value was the adjusted *p*-value. The screening criteria were $|\log 2| > 1$ and *p*-value of 0.05 FDR. Create bubble plots to display DEG's Disease Ontology (DO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment using the R software 'clusterProfiler' package [17]. Gene Set Enrichment Analysis (GSEA) [18] was applied to assess trends in the distribution of preset sets of genes in the gene table to establish their contribution to the phenotype.

CIBERSORT estimation

Calculate immune cells for the input gene expression matrix using the CIBERSORT [19] portal website (http://cibersort.stanford.edu/). In the output data, the relative contents of 22 immune cells were estimated for each sample. Cases with $p < 0.05$ were deemed to have correct inference scores and were qualified for further investigation.

HCC cancer sample collection and quantitative polymerase chain reaction assay

We selected 10 patients who were clinically diagnosed with HCC at the Department of Hepatobiliary Surgery of the First Affiliated Hospital of Chongqing Medical University (Chongqing, China) and collected HCC and normal tissues adjacent to the cancer. All experiments were performed in accordance with the Declaration of Helsinki of the World Medical Association and were approved by the ethics committee of the First Affiliated Hospital of Chongqing Medical University (16 October 2023/No. K2023-440). All the participants provided written informed consent. No chemoradiotherapy or immunotherapy was administered before surgery, and no other malignant tumours were found. Total RNA was extracted from HCC tissues and cells using TRIzol reagent (Invitrogen, Carlsbad, CA, USA). cDNA was synthesized using the SYBR Premix Ex Taq kit (Takara, Japan), according to the manufacturer's instructions. Quantitative polymerase chain reaction (qPCR) was performed according to the manufacturer's instructions. Relative mRNA expression (normalized to GAPDH) was assessed using the $2^{-\Delta\Delta Ct}$ ($\Delta\Delta C_t = \Delta C_{t[\text{target gene}]} - \Delta C_{t[\text{GAPDH}]}$) method. Primers used in this study are listed in Table S1.

Culture of human hepatocarcinoma cells

The Hep3B and SMMC7721 cell lines were purchased from Wuhan Proxel Life Science and Technology Co., Ltd. Hep3B cell lines were cultured in complete medium containing MEM (Gibco, USA), and SMMC7721 cell lines were cultured in complete medium containing Roswell Park Memorial Institute (Gibco, USA), 10% foetal bovine serum (FBS, Gibco, USA) and 1% penicillin-streptomycin (Sigma, USA) at 37 °C and 5% CO_2. Cells were passaged when they reached 80% confluence in T25 cell culture flasks.

Cell viability assays

Cell viability was assessed using a cell counting kit-8 (CCK-8). Three 96-well plates were prepared, each seeded at a density of 2000 cells per well, containing 100 µL of complete medium per well. After 24 h of incubation on one of the plates, 10 µL of CCK-8 solution was added to each well and incubated for an additional 2 h. Finally, the optical density (OD) of each well was measured using a microplate reader at a wavelength of 450 nm, and cell viability was plotted based on the OD value. The other two plates were tested by the same method.

Flow cytometry apoptosis analysis

Human hepatoma cells were seeded in six-well plates (1.0×10^4 cells per well) and incubated in an incubator for 24 h, after which they were incubated with calcein-AM/PI for additional 30 min. The labelled cells were washed twice with PBS, collected in a small centrifuge tube, detected by flow cytometry and finally exported.

Animals and experimental protocol

Ten female C57BL/6 mice (6 weeks old) purchased from the Experimental Animal Center of Chongqing Medical University (Chongqing, China) were housed in a pathogen-free environment and used for Hepa1-6 hepatoma cell xenografts. For stably knockdown of RSPO3, Hepa1-6 cells were transfected with lentiviral (LV) vector (control) and LV-shRSPO3. Subsequently, xenografts were prepared by the subcutaneous injection of 1×10^6 Hepa1-6 cells. Tumour volume was measured every 3 days and was calculated as previously described. The subcutaneous tumours were harvested and tumour weight was calculated on day 15.

All experimental procedures were conducted in accordance with the ARRIVE (Animal Research: Reporting In Vivo Experiments) guidelines. Animals received humane care in accordance with the guidelines provided by the National Institutes of Health for the use of animals in laboratory experiments. The animal protocols used in this work were evaluated and approved by the Animal Use and Ethic Committee of 1st Affiliated Hospital of Chongqing Medical University (Protocol K2023-440; Chongqing, China).

Statistical analysis

All statistical analyses were performed using R version 4.2.0, and *p*-values less than 0.05 were considered significant. The Spearman correlation approach was used to investigate associations between immune cell expression. The Wilcoxon test was employed to assess the differences among the groups. The most relevant genes that could be used to distinguish tumours from normal tissues were identified using both random forest analysis and least absolute shrinkage and selection operator (LASSO) analysis. The final overlapping markers between the two methods were chosen. Subject operating characteristic (ROC) curves and time-dependent ROC curves were used to examine and quantify the sensitivity and specificity of the diagnostic prediction models. ROC curve analysis was conducted using the pROC package in R. A logistic regression model was subsequently trained on the expression data, with binary labels indicating tumour (1) and normal (0) samples. To enhance the model's robustness and mitigate the risk of overfitting, we employed fivefold cross-validation using the caret package. During this process, the model was trained and validated on different subsets of the data. The predicted probabilities for the positive class were then utilized to construct the ROC curve. Finally, the area under the curve (AUC) was calculated to quantify the model's ability to discriminate between tumour and normal samples. AUC values range from 0 to 1, with values closer to 1 indicating superior diagnostic performance.

Result

DEG identification and functional enrichment analyses

We identified 367 DEGs in the training set using the Limma program, of which 120 were up-regulated and 247 were down-regulated (Figure 1A). The heat map shows the expression of the top 50 genes in terms of up-regulation and down-regulation (Figure 1B). To explore the main functions of these genes, we performed DO and KEGG enrichment analyses, which revealed that DEGs were mainly associated with tumours, including biliary tract cancer, renal cell carcinoma and benign tumours of the organ system (Figure 2A). In terms of function, DEGs were mainly involved in tumour-related metabolic and progression pathways, for example, drug metabolism-cytochrome P450, cell cycle, p53 signalling pathway, ECM–receptor interaction and arachidonic acid metabolism (Figure 2B). DEG-based enrichment analysis may lose much of the information that non-differential genes confer on phenotypes. Therefore, we performed GSEA to validate the enrichment analysis results. As shown in Figure 2C,D, the down-regulation of metabolism-related pathways in tumours, as well as the increase in cellular recycling, DNA replication and ribosomal processes, were consistent with the KEGG results.

Identification of ECM1, NPC1L1, RSPO3 as the diagnostic biomarker for HCC

We screened key biomarkers for diagnostic use using a random forest approach and minimum absolute shrinkage and selection operator model. A total of 367 differential genes were screened for 14 core diagnostic genes and 33 core diagnostic genes (Figure 3A,B). We selected three overlapping genes, ECM1, NPC1L1 and RSPO3, as candidate biomarkers for subsequent analysis (Figure 3C). All three candidate genes were differentially expressed in the validation cohort, and ECM1, NPC1L1 and RSPO3 were significantly down-regulated in tumours ($p < 0.05$; Figure 4A–C). To further determine the expression of the three core genes, we collected tissue specimens from HCC patients and corresponding paracancerous tissues from the clinic for qPCR assays, which further confirmed that ECM1, NPC1L1 and RSPO3 were significantly down-regulated in HCC ($p < 0.05$; Figure 4D–F).

Diagnostic validity of candidate markers

To determine the diagnostic validity of the candidate markers, we had to validate them jointly in the training and test cohorts. The diagnostic performance of ECM1, NPC1L1 and RSPO3 was determined by plotting ROC curves and evaluating the AUC values. The AUC values of ECM1, NPC1L1 and RSPO3 in the training cohort were 0.994, 0.782 and 0.992, respectively (Figure

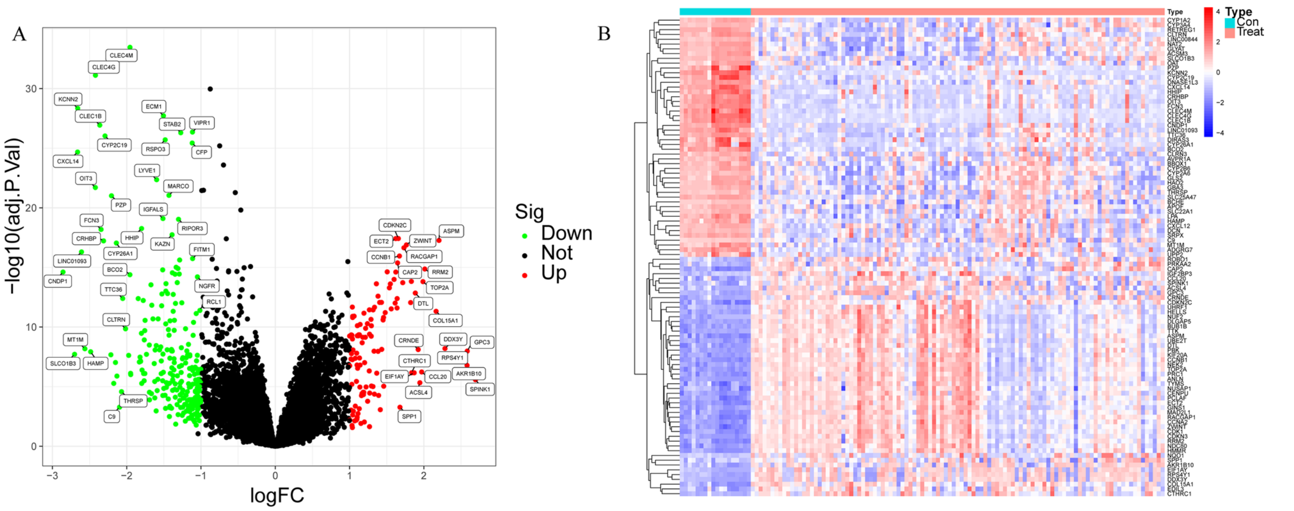

Figure 1. Screening of differential genes. (A) Volcano plot versus (B) heat map corresponding to genes differentially expressed in HCC samples compared to normal controls in the training cohort. Red: up-regulated expression; green: down-regulated expression.

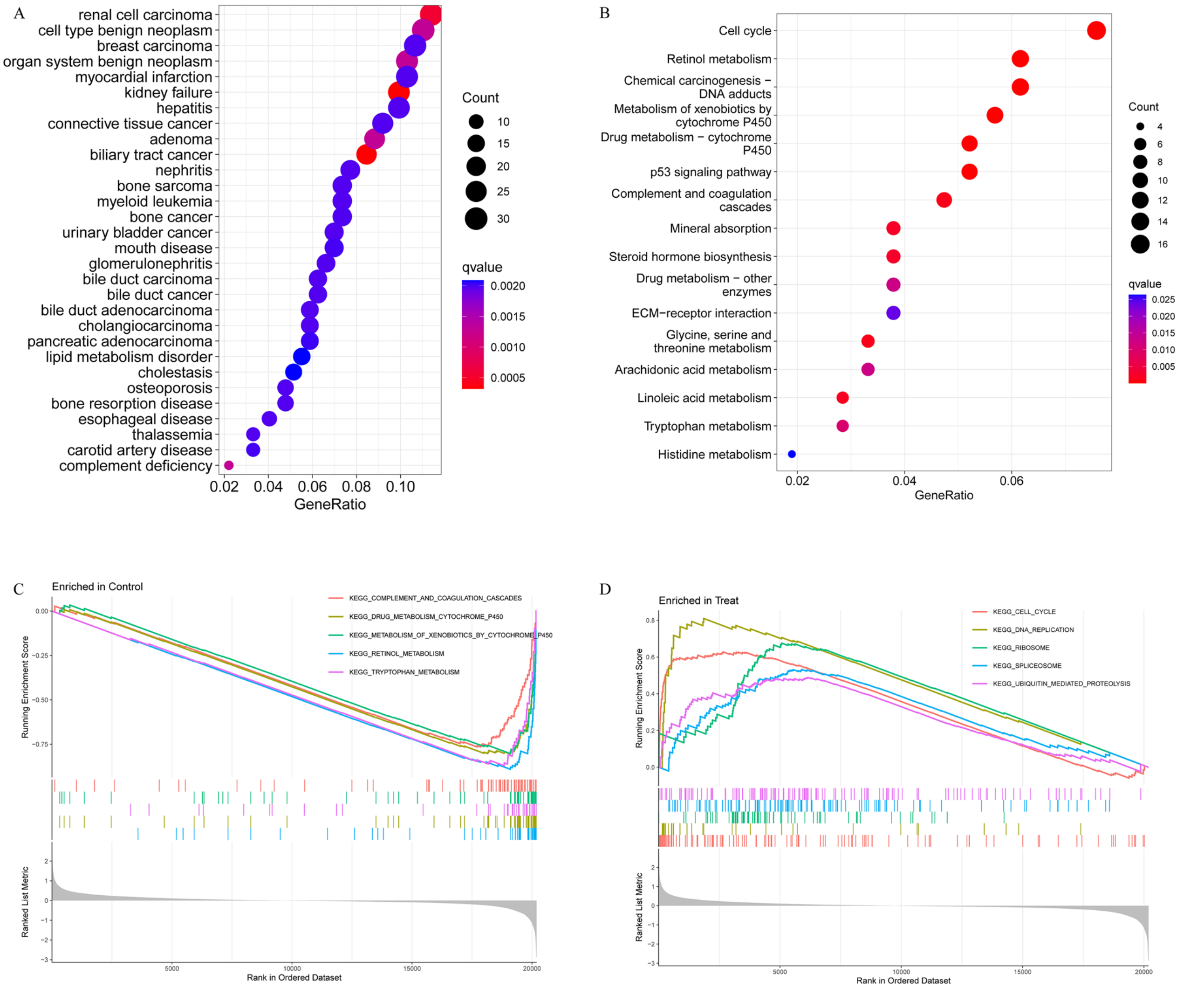

Figure 2. Enrichment analysis results of differential genes. Bubble plot of (A) DO and (B) KEGG enrichment analysis of differential genes. Results of enrichment analysis of related gene sets for (C) up- and (D) down-regulated pathways in HCC.

5A–C), indicating that all three candidate markers were good diagnosticians for the development of HCC. In the test cohort, the AUC value of all three candidate markers was 1 (Figure 5D–F), indicating the universality of their diagnostic performance.

Immune cell infiltration in HCC and its correlation with ECM1, NPC1L1, RSPO3

By applying CIBERSORT, we first examined the makeup of immune cells in HCC and normal liver tissues. As shown in Figure 6A, the proportion of Tregs, macrophages M0 and macrophages M1 in HCC was always greater than that in normal tissues, whereas the number of Monocytes, macrophages M2 and neutrophils decreased in HCC ($p < 0.05$). Figure 6B depicts the relationship between immune cells, showing a strong negative connection between CD8+T cells and CD4 memory resting cells ($R = -0.65$) and a substantial positive association between CD4 memory activating cells and T cell gamma delta ($R = 0.68$). Furthermore, macrophage M2 expression was negatively correlated with M1 and M0 macrophage expression ($R < -0.2$). Spearman correlation analysis further analysed the association between immune cells and diagnostic genes, and we found that ECM1 was highly positively correlated with macrophage M2 (Figure 7B; $R = 0.45$) and neutrophils and highly negatively correlated with Tregs and macrophage M0 (Figure 7A,C; $R = -0.38$; all $p < 0.001$). NPC1L1 showed the greatest positive correlation with mast cell activation (Figure 7E; $R = 0.23$) and, in positive contrast, the greatest negative correlation with mast cell inhibition (Figure 7D,F; $R = -0.24$; all $p < 0.05$). RSPO3 was positively correlated with plasma cells (Figure 7H; $R = 0.32$) and neutrophils, whereas it was negatively correlated with Tregs (Figure 7G,I;

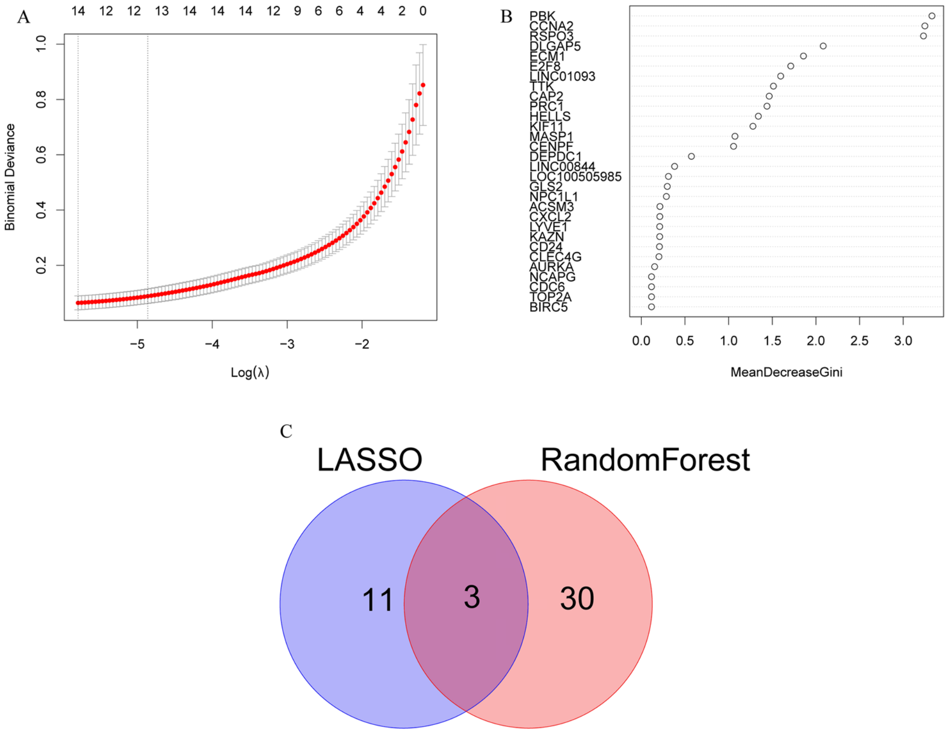

Figure 3. Identification of genetic biomarkers for HCC diagnosis. (A) Fourteen genes with minimal λ values were identified using LASSO analysis. The upper horizontal coordinate indicates the number of modelled genes corresponding to different λ values. (B) Random forest algorithm feature gene selection plot. (C) Venn diagram of the same biomarker screened by LASSO and random forest algorithm.

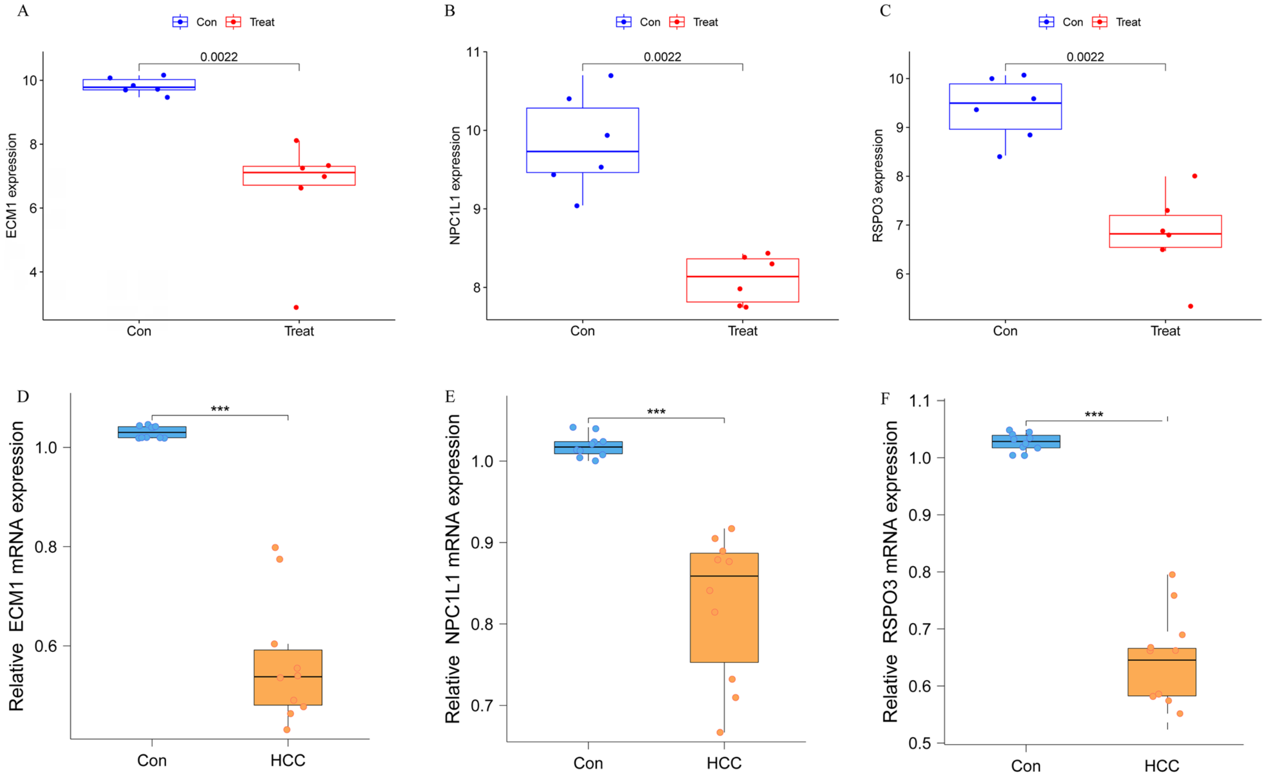

Figure 4. Altered expression of three core diagnostic genes in the validation cohort, including (A) ECM1, (B) NPC1L1 and (C) RSPO3. qPCR analysis showed that (D) ECM1, (E) NPC1L1 and (F) RSPO3 were significantly down-regulated in the HCC group compared to the paracancer tissue.

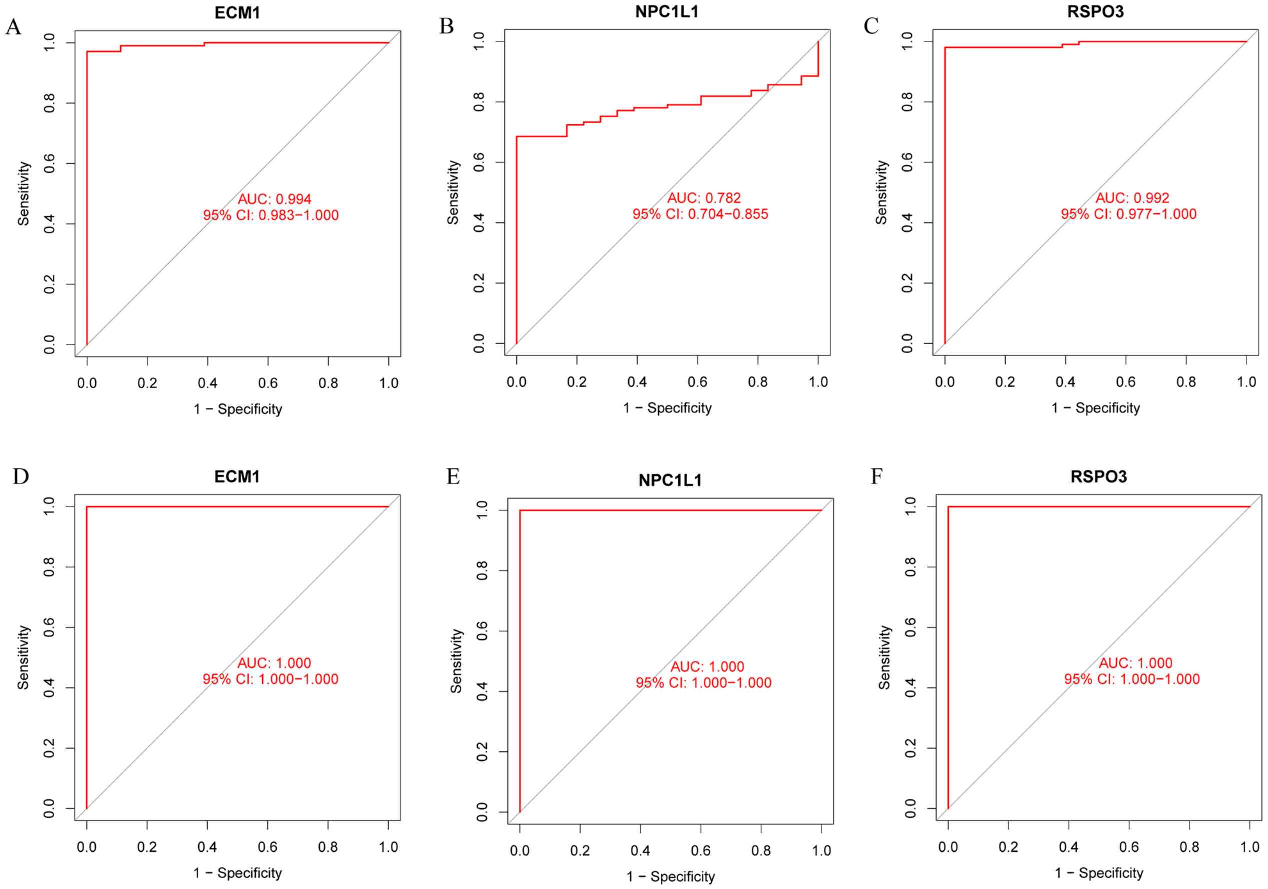

Figure 5. Verification of diagnostic performance. ROC curves of three core diagnostic genes in the training cohort, including (A) ECM1, (B) NPC1L1 and (C) RSPO3. ROC curves of three core diagnostic genes in the testing cohort, including (D) ECM1, (E) NPC1L1 and (F) RSPO3.

$R = -0.48$; all $p < 0.001$). In conclusion, these results demonstrate that the three newly identified diagnostic genes are associated with immune infiltration in HCC.

Expression and viability of the biomarker gene RSPO3-si in HCC

The expression of RSPO3 in Hep3B and SMMC7721 cells was evaluated using qRT-PCR. The results showed that all three RSPO3-si were significantly lower than the control group in both HCC cell lines (Figure 8A). For the Hep3B cell line, RSPO3-si1 and RSPO3-si2 were selected for subsequent experiments. For the SMMC7721 cell lines, RSPO3-si1 and RSPO3-si3 were selected for subsequent functional experiments. Subsequently, we confirmed that the knockdown of RSPO3 inhibited the proliferation of Hep3B and SMMC7721 cells (Figure 8B,C). We confirmed that knockdown of RSPO3 promoted apoptosis in Hep3B (Figure 8D–G) and SMMC7721 cells (Figure 8H–K).

Effect of RSPO3 on tumour growth and immune infiltration in mice model

We confirmed the effect of RSPO3 *in vitro*. Furthermore, to detect if RSPO3 knockdown could enhance the antitumour activity of macrophages in mice. It was revealed that sh-RSPO3 mice had smaller subcutaneous tumours than the control group (Figure 9A–C). To determine whether sh-RSPO3 had a polarization effect on subcutaneous macrophages, we detected plasma cells with CD138, M1 macrophages with iNOS, and M2 macrophages with CD206 (Figure 9D). It was showed that the RSPO3 knockdown was associated with declined plasma cells, M2 subtype macrophages, while linked with increased M1 subtype macrophages in mice model. These results indicated that RSPO3 was positively correlated with the infiltration of plasma cells, M2 macrophages and negatively correlated with the infiltration of M1 macrophages.

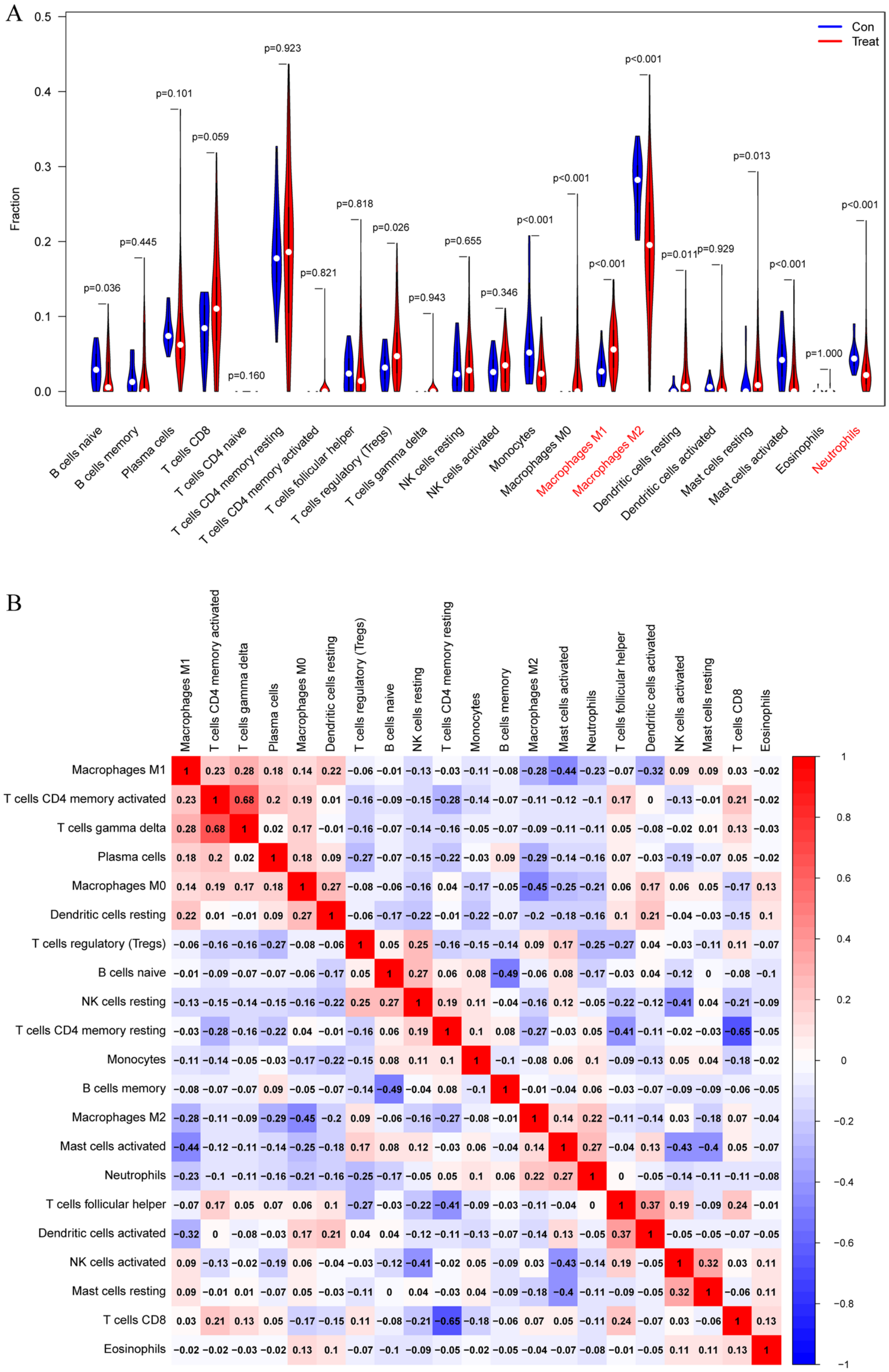

Figure 6. Analysis of immune cell infiltration in patients with HCC. (A) Comparison of the expression levels of 22 immune cells in normal tissue and cancer tissue samples from HCC patients. Blue and red represent healthy control (Con) and HCC samples, respectively. (B) Correlation matrix of the 22 immune cell subtypes. Red colour indicates positive correlation, while blue colour indicates negative correlation.

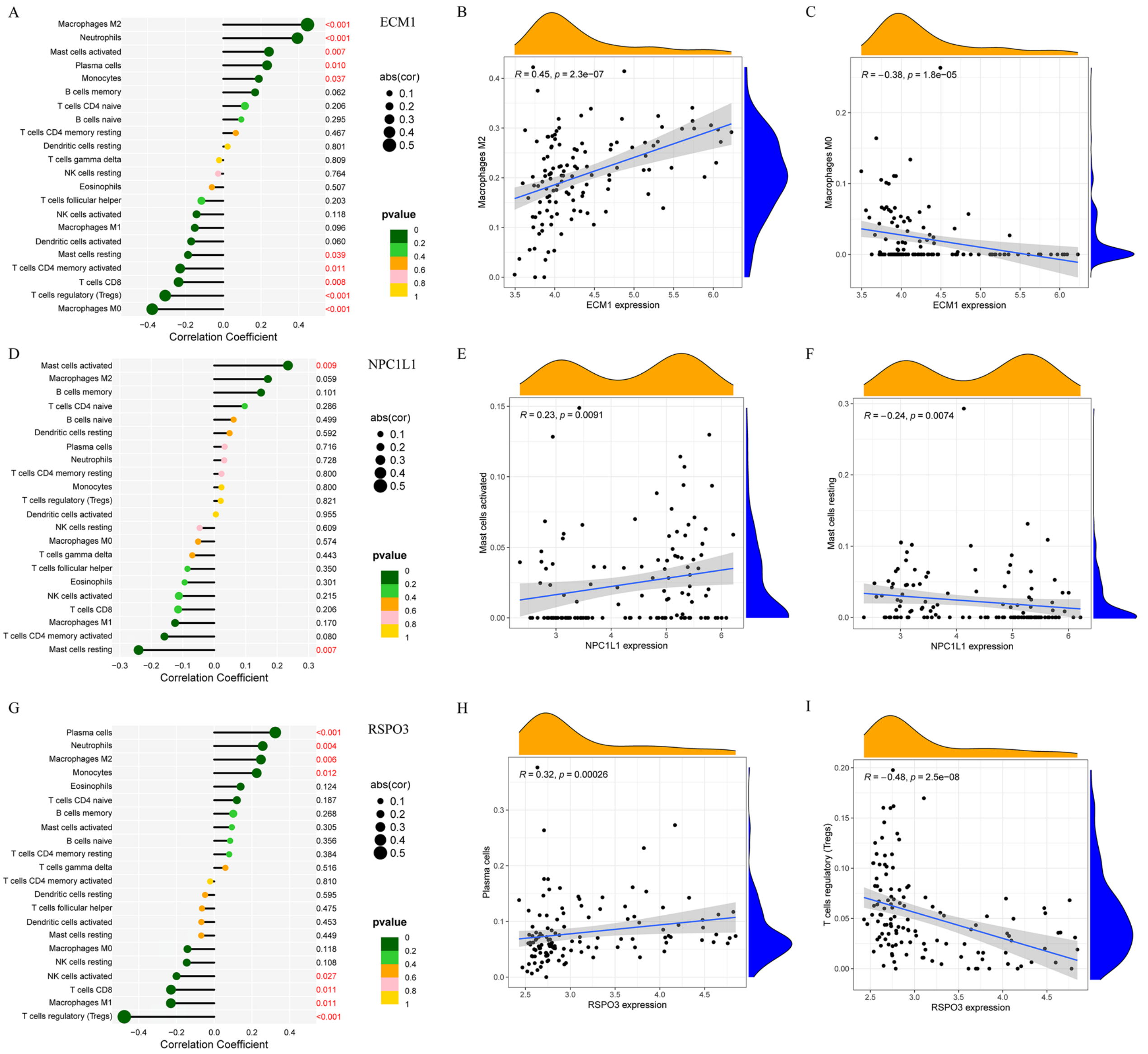

Figure 7. Correlation analysis of three diagnostic genes and immune cells. (A) Lollipop plot of correlation of ECM1 expression with immune cells in HCC, and correlation plots with (B) macrophage M2 and (C) macrophage M0. (D) Lollipop plot of correlation of NPC1L1 expression with immune cells in HCC, and correlation plots with (E) mast cell activation and (F) mast cell suppression. (G) Lollipop plot of correlation of RSPO3 expression with immune cells in HCC, and correlation plots with (H) plasma and (I) Tregs.

Discussion

According to U.S. cancer statistics for 2019, liver cancer ranks 14th in new cases diagnosed in a year and 5th in cancer-related mortality [20]. The early diagnosis of cancer can be of great benefit to patient treatment prognosis. In this study, three genes (ECM1, NPC1L1 and RSPO3) were found to be significantly reduced in HCC tumour samples compared to normal tissues, and ROC analysis showed very high sensitivity and specificity for the diagnosis of HCC. These data support the use of these genes as potential diagnostic biomarkers for HCC.

ECM1 is a glycoprotein that was identified as a candidate biomarker with a high diagnostic value but was not associated with overall survival, as in the study by Ge et al. [21]. The current study found that ECM1 down-regulates E-cadherin expression, up-regulates vimentin expression [22] and promotes migration and invasion of HCC cells by inducing epithelial–mesenchymal transition, while knockdown of ECM1 inhibited HCC cell function [23]. Although our results showed down-regulation of ECM1 expression in HCC tissues compared to normal tissues, this does not affect the phenotype in which high expression of ECM1 still promotes cancer cell proliferation. ECM1 appears to be a factor that can influence macrophage polarization to M1 or M2 phenotypes. ECM1 can promote M1 macrophage polarization in response to LPS stimulation by

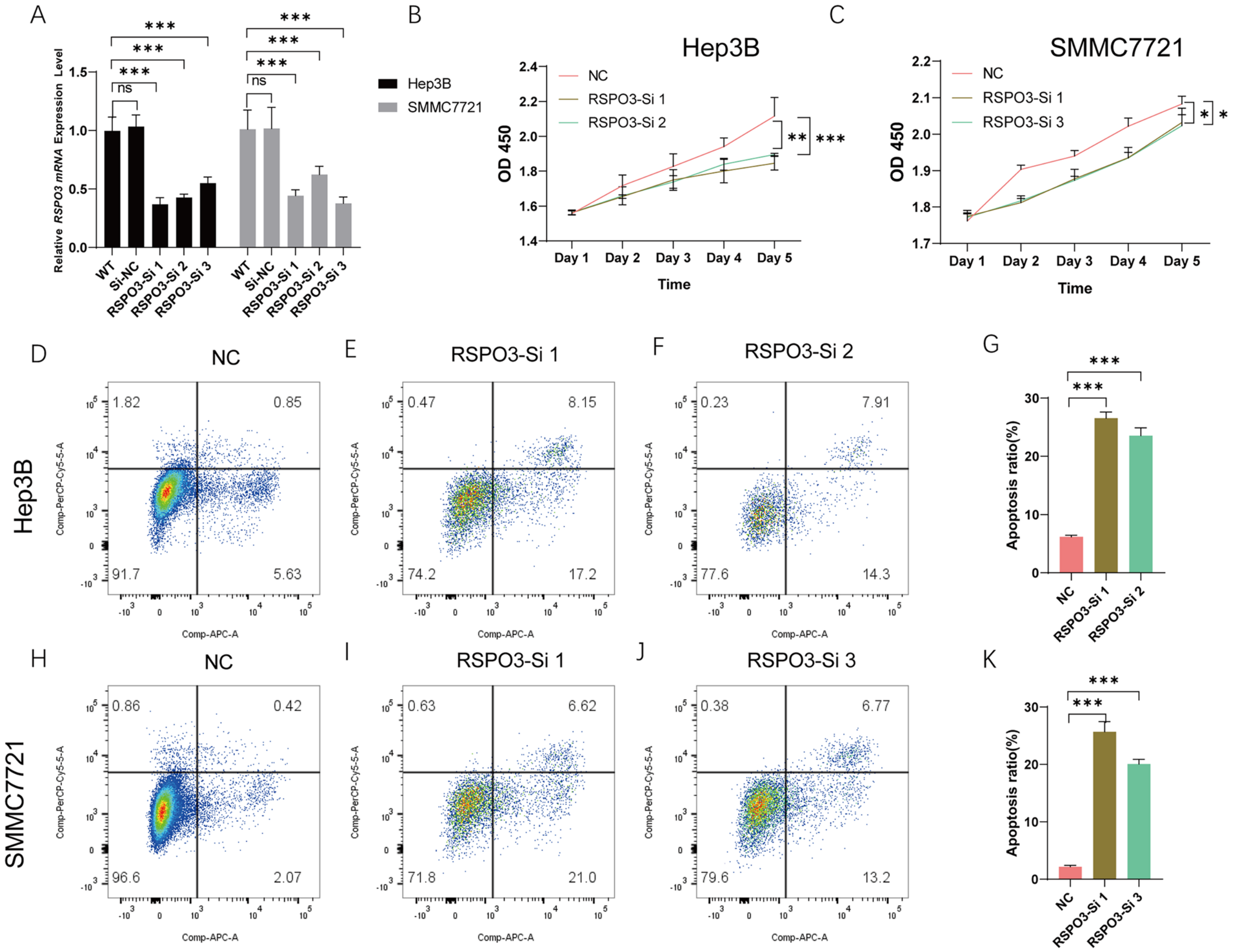

Figure 8. Expression and function validation of RSPO3-si in two HCC cell lines. (A) Knockdown efficiency of RSPO3-si in Hep3B and SMMC7721 cell lines. (B,C) The effect of RSPO3-si on Hep3B and SMMC7721 cells via CCK8. (D–G) Detection of the apoptosis rate of RSPO3-si in Hep3B cells. (H–K) Detection of the apoptosis rate of RSPO3-si in SMMC7721 cells.

activating the NF-κB pathway and enhancing the production of pro-inflammatory cytokines such as TNF-α and IL-6 [24]. Our results revealed that ECM1 was positively correlated with M2 macrophages and strongly negatively correlated with M0 macrophages. Thus, ECM1 may regulate the balance between M1 and M2 macrophages and influence their immune response.

NPC1L1 is a sterol transporter protein involved in lipid homeostasis in the small intestine and the liver. NPC1L1 has been implicated in the development and progression of liver cancer by modulating cholesterol metabolism, inflammation and cell proliferation [25]. Chen et al. found that NPC1L was weakly expressed in HCC, and a low expression level of NPC1L1 was significantly associated with poorer OS [26]. NPC1L1 may exert antitumour effects by inhibiting cell growth and inducing apoptosis in HCC cells. In addition, the GM-CSF/STAT5 pathway, inhibited by NPC1L1, controls macrophage polarization to the M2 phenotype, which inhibits antitumour immunity and promotes tissue repair [25]. Our work found that NPC1L could be used as a diagnostic marker for HCC associated with mast cell activation or inhibition, but more studies are needed to confirm this hypothesis.

RSPO3 is one of four R-spondin proteins that can activate the Wnt/β-catenin signalling pathway, which is involved in stem cell regulation and cancer development [27]. The clinical potential of RSPO3 as a novel therapeutic target has been established in a clinical trial to test the safety and efficacy of the neutralizing monoclonal anti-RSPO3 antibody OMP131-R10 (rosmantuzumab) in patients with advanced solid tumours and metastatic CRC [27]. Although discontinued owing to insufficient remission rates, it still serves as a milestone event to commemorate some initial efforts to disrupt the over-activation of RSPO3 in cancer. In addition, studies have indicated that macrophages secrete RSPO3 and stimulate Wnt/β-catenin signalling in hepatocytes [28]. This is consistent with our findings. This study has several limitations that should be

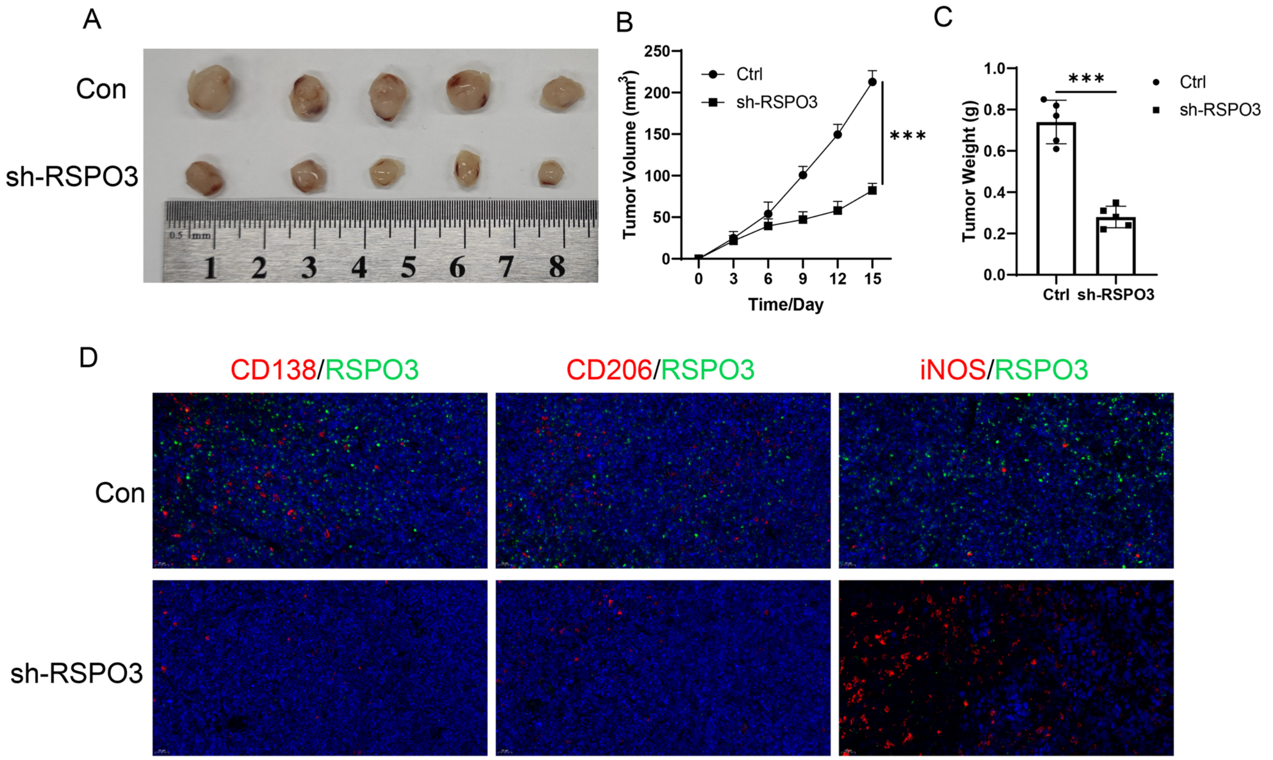

Figure 9. Infection of sh-RSPO3 on mice. (A,B) The volumes of tumour in sh-RSPO3 group and control group. (C) The weight of tumour in sh-RSPO3 group and control group. (D) The RSPO3, CD138, CD206 and iNOS expression in sh-RSPO3 group and control group.

acknowledged. Firstly, the use of non-HCC cell lines for experimental validation may limit the generalizability of our findings to actual HCC conditions, as these cell lines may not accurately replicate the tumour microenvironment. Secondly, the lack of *in vivo* validation restricts our ability to confirm the clinical relevance of the identified biomarkers. Future studies should incorporate *in vivo* models to better assess the diagnostic utility and functional roles of ECM1, NPC1L1 and RSPO3 in HCC. Addressing these limitations will provide a more balanced interpretation of our findings.

In summary, we utilized machine learning to identify and validate three potential diagnostic markers that could aid in early tumour screening. Additionally, the pro-oncogenic involvement of RSPO3 and its potential regulatory effects on plasma cell proportions and the induction of macrophage polarization have been validated.

Acknowledgements

Parts of this work have been previously published as a preprint (https://doi.org/10.21203/rs.3.rs-3259945/v1), and we appreciate the opportunity to share our findings with the scientific community prior to formal publication.

Authors contributions

Conceptualization, Q.L., H.Z., H.Y. and R.L. Methodology, Q.L., H.X. and A.R. Software, Q.L., H.X. and A.R. Validation, Z.W. and Z.H. Formal analysis, Q.L. and H.Z. Investigation, Q.L., H.Z. and R.L. Resources, Q.L., R.L. and Z.H. Data curation, Q.L. and Y.C. Writing—original draft preparation, Q.L., H.Z. and H.X. Writing—review and editing, H.Y. and Z.W. Visualization, H.X. and Y.C. Supervision, Q.L., Z.W. and Z.H. Project administration, R.L., Z.W. and Z.H. Funding acquisition, Q.L. and Y.C. All authors have read and agreed to the published version of the manuscript.

Consent for publication

Not applicable.

Disclosure statement

The authors declare that they have no competing interests.

Funding

This study was sponsored by Natural Science Foundation of Chongqing, China (cstc2021jcyj-msxmX0292) and Chongqing Postdoctoral Science Foundation Project (cstc2021jcyj-bshX0118).

Data availability statement

The datasets generated and/or analysed during the current study are available in the [GEO] repository, [https://www.ncbi.nlm.nih.gov/geo/. Accession number: GSE46408, GSE62232, GSE101685]. The data that support the findings of this study are available from the corresponding author Z.H., upon reasonable request.

References

[1] Yang YM, Kim SY, Seki E. Inflammation and liver cancer: molecular mechanisms and therapeutic targets. Semin Liver Dis. 2019;39(1):26–42. doi: 10.1055/s-0038-1676806.

[2] Forner A, Gilabert M, Bruix J, et al. Treatment of intermediate-stage hepatocellular carcinoma. Nat Rev Clin Oncol. 2014;11(9):525–535. doi: 10.1038/nrclinonc.2014.122.

[3] Department of Medical Administration NH, Health Commission of the People's Republic of C. [Guidelines for diagnosis and treatment of primary liver cancer in China (2019 edition)]. Zhonghua Gan Zang Bing Za Zhi. 2020;28(2):112–128.

[4] He Q, Yang J, Jin Y. Immune infiltration and clinical significance analyses of the coagulation-related genes in hepatocellular carcinoma. Brief Bioinform. 2022;23(4),1-21. doi: 10.1093/bib/bbac291.

[5] He D, Zhang X, Tu J. Diagnostic significance and carcinogenic mechanism of pan-cancer gene POU5F1 in liver hepatocellular carcinoma. Cancer Med. 2020;9(23):8782–8800. doi: 10.1002/cam4.3486.

[6] Liu J, Sun G, Pan S, et al. The Cancer Genome Atlas (TCGA) based m(6)A methylation-related genes predict prognosis in hepatocellular carcinoma. Bioengineered. 2020;11(1):759–768. doi: 10.1080/21655979.2020.1787764.

[7] Duan H, Zhao G, Xu B, et al. Maternal serum PLGF, PAPPA, beta-hCG and AFP levels in early second trimester as predictors of preeclampsia. Clin Lab. 2017;63(5):921–925. doi: 10.7754/Clin.Lab.2016.161103.

[8] Cao L, Cheng H, Jiang Q, et al. APEX1 is a novel diagnostic and prognostic biomarker for hepatocellular carcinoma. Aging. 2020;12(5):4573–4591. doi: 10.18632/aging.102913.

[9] Cui XH, Peng QJ, Li RZ, et al. Cell division cycle associated 8: a novel diagnostic and prognostic biomarker for hepatocellular carcinoma. J Cell Mol Med. 2021;25(24):11097–11112. doi: 10.1111/jcmm.17032.

[10] Ma X, Ding Y, Zeng L. The diagnostic and prognostic value of H2AFY in hepatocellular carcinoma. BMC Cancer. 2021;21(1):418. doi: 10.1186/s12885-021-08161-4.

[11] Shi K, Lin W, Zhao XM. Identifying molecular biomarkers for diseases with machine learning based on integrative omics. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(6):2514–2525. doi: 10.1109/TCBB.2020.2986387.

[12] Hou Q, Bing Z-T, Hu C, et al. RankProd combined with genetic algorithm optimized artificial neural network establishes a diagnostic and prognostic prediction model that revealed C1QTNF3 as a biomarker for prostate cancer. EBioMedicine. 2018;32:234–244. doi: 10.1016/j.ebiom.2018.05.010.

[13] Binder J, Ursu O, Bologa C, et al. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Commun Biol. 2022;5(1):125. doi: 10.1038/s42003-022-03068-7.

[14] Han H, Chen Y, Yang H, et al. Identification and verification of diagnostic biomarkers for glomerular injury in diabetic nephropathy based on machine learning algorithms. Front Endocrinol. 2022;13:876960. doi: 10.3389/fendo.2022.876960.

[15] Parker HS, Corrada Bravo H, Leek JT. Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ. 2014;2:e561. doi: 10.7717/peerj.561.

[16] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi: 10.1093/nar/gkv007.

[17] Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2(3):100141. doi: 10.1016/j.xinn.2021.100141.

[18] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–15550. doi: 10.1073/pnas.0506580102.

[19] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–457. doi: 10.1038/nmeth.3337.

[20] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7–34. doi: 10.3322/caac.21551.

[21] Ge S, Xu CR, Li YM, et al. Identification of the diagnostic biomarker VIPR1 in hepatocellular carcinoma based on machine learning algorithm. J Oncol. 2022;2022:2469592. doi: 10.1155/2022/2469592.

[22] Chen H, Jia W, Li J. ECM1 promotes migration and invasion of hepatocellular carcinoma by inducing epithelial-mesenchymal transition. World J Surg Oncol. 2016;14(1):195. doi: 10.1186/s12957-016-0952-z.

[23] Gao F, Xia Y, Wang J, et al. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. Genome Biol. 2014;15(12):533. doi: 10.1186/s13059-014-0533-9.

[24] Zhang Y, Li X, Luo Z, et al. ECM1 is an essential factor for the determination of M1 macrophage polarization in IBD in response to LPS stimulation. Proc Natl Acad Sci U S A. 2020;117(6):3083–3092. doi: 10.1073/pnas.1912774117.

[25] Zhang R, Zeng J, Liu W, et al. The role of NPC1L1 in cancer. Front Pharmacol. 2022;13:956619. doi: 10.3389/fphar.2022.956619.

[26] Chen KJ, Jin RM, Shi CC, et al. The prognostic value of Niemann-Pick C1-like protein 1 and Niemann-Pick disease type C2 in hepatocellular carcinoma. J Cancer. 2018;9(3):556–563. doi: 10.7150/jca.19996.

[27] Ter Steege EJ, Bakker ERM. The role of R-spondin proteins in cancer biology. Oncogene. 2021;40(47):6469–6478. doi: 10.1038/s41388-021-02059-y.

[28] Zhang M, Haughey M, Wang N-Y, et al. Targeting the Wnt signaling pathway through R-spondin 3 identifies an anti-fibrosis treatment strategy for multiple organs. PLoS One. 2020;15(3):e0229445. doi: 10.1371/journal.pone.0229445.