Identification of Multi-omics Biomarkers and Construction of the Novel Prognostic Model for Hepatocellular Carcinoma

Xiaokai Yan (YXK11011@163.com)
Zunyi Medical University https://orcid.org/0000-0002-8507-3473
Chiying Xiao
Zunyi Medical University
Kunyan Yue
Zunyi Medical University
Min Chen
Zunyi Medical University
Hang Zhou
Zunyi Medical University

Research

Keywords: hepatocellular carcinoma, prognostic model, multi-omics, TCGA, ICGC

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-452644/v1

Abstract

Background: Change in the genome plays a crucial role in cancerogenesis and many biomarkers can be used as effective prognostic indicators in diverse tumors. Currently, although many studies have constructed some predictive models for hepatocellular carcinoma (HCC) based on molecular signatures, the performance of which is unsatisfactory. To fill this shortcoming, we hope to construct a novel and accurate prognostic model with multi-omics data to guide prognostic assessments of HCC.

Methods: The TCGA training set was used to identify crucial biomarkers and construct single-omic prognostic models through difference analysis, univariate Cox, and LASSO/stepwise Cox analysis. Then the performances of single-omic models were evaluated and validated through survival analysis, Harrell's concordance index (C-index), and receiver operating characteristic (ROC) curve, in the TCGA test set and external cohorts. Besides, a comprehensive model based on multi-omics data was constructed via multiple Cox analysis, and the performance of which was evaluated in the TCGA training set and TCGA test set.

Results: We identified 16 key mRNAs, 20 key lncRNAs, 5 key miRNAs, 5 key CNV genes, and 7 key SNPs which were significantly associated with the prognosis of HCC, and constructed 5 single-omic models which showed relatively good performance in prognostic prediction with c-index ranged from 0.63 to 0.75 in the TCGA training set and test set. Besides, we validated the mRNA model and the SNP model in two independent external datasets respectively, and good discriminating abilities were observed through survival analysis ($P < 0.05$). Moreover, the multi-omics model based on mRNA, lncRNA, miRNA, CNV, and SNP information presented a quite strong predictive ability with c-index over 0.80 and all AUC values at 1,3,5-years more than 0.84.

Conclusion: In this study, we identified many biomarkers that may help study underlying carcinogenesis mechanisms in HCC, and constructed five single-omic models and an integrated multi-omics model that may provide effective and reliable guides for prognosis assessment and treatment decision-making.

Background

Liver cancer is one of the most common and prevalent human malignancies in the world, seriously threatening people's lives and health [1]. Hepatocellular carcinoma (HCC) is the predominant liver cancers, accounts for 70–85% of cases [2]. The development of HCC is an extremely complicated multi-step process, and the prognosis of HCC is affected by many factors, such as heredity, environment, and infection making the prognosis prediction very challenging[3]. Therefore, it is necessary and urgent to formulate a reasonable and effective prognostic evaluation model for HCC.

A lot of research has showed that change in the genome plays a key role in cancerogenesis and many biomarkers can be used as prognostic indicators [4, 5]. For example, LMO1 was identified as an important oncogene that promotes neuroblastoma initiation, progression, and widespread metastatic dissemination [6]. LncRNA SNHG10 was associated with poor overall survival of HCC, and mediated HCC cell

proliferation, invasion, and migration and facilitated the cell cycle and epithelial-mesenchymal transition [7]. miR-487a could enhance the proliferation and metastasis of HCC cells by directly binding to sprouty-related EVH1 domain containing 2 (SPRED2) or phosphoinositide-3-Kinase regulatory subunit 1 (PIK3R1) and can be used as a potential prognostic marker[8]. Vladimir Bezrookove et al. have proved the important role of PHIP copy-number elevation as a prognostic and progression marker for cutaneous melanoma [9]. SNP in 3'UTR of RAS-related proteins (RAP1A) was significantly associated with esophageal squamous cell carcinoma risk and metastasis [10]. On the basis of these studies, many prognostic models based on genetic changes have been constructed. However, most of these studies build models with single-omic data, such as gene expression, DNA methylation [11, 12]. Lack of information usually leads to decrease in the predictive ability of models. Therefore, the prognostic model build by multi-omics data should be more reasonable and efficient.

At present, many studies have been devoted to constructing a prognostic evaluation model for HCC [13-19]. To our knowledge, only Chaudhary K et al. [20] constructed a predictive model with multi-omics data. However, the predictive ability of this model is not so good (C-index = 0.70). Therefore, we hope to construct novel and accurate prognostic models that are more applicable in guiding prognostic assessments and treatment decision-making of HCCs.

Materials And Methods

Data acquisition

The genome Cancer atlas (TCGA)

The transcriptome profiling (including RNA-Seq and miRNA-Seq), the masked copy number segment profiles (CNV, copy number variation), and the single nucleotide polymorphism (SNP) data detected with VarScan2 were obtained from TCGA (https://cancergenome.nih.gov/). The corresponding HCC prognostic information was obtained from cBioPotal for cancer genomics (https://www.cbioportal.org). For RNA-Seq data, the raw count data were normalized with TPM (Transcripts per million) method. Patients died within 3 months and non-HCC patients were removed.

LIRI-JP dataset (Liver Cancer-RIKEN, JP project)

The mRNA expression data and corresponding prognostic information of LIRI-JP dataset were downloaded from ICGC data portal (The International Cancer Genome Consortium, https://icgc.org/). The mRNA expression data were normalized with TPM method. Patients died within 3 months and non-HCC patients were removed.

LICA-FR dataset (Liver Cancer-FR project)

The SNP data and corresponding prognostic information of LICA-FR dataset were downloaded from ICGC data portal too. Patients died within 3 months and non-HCC patients were removed.

Construction and validation of prognostic model

Model based on mRNA expression

In the TCGA dataset, the whole HCC samples were used as a training set. mRNAs expressed in over 95% of samples were retained, and the zero values in the expression matrix were replaced with the minimum non-zero value of the corresponding gene. Then the expression data were log2 transformed. Differentially expressed mRNAs (DE-mRNAs) between HCC and non-tumor samples were identified via 'limma' package [21], P value < 0.0001 and |logFC (log fold change) | > 3 were set as the cut-off criterion. Univariate Cox regression analysis was performed to identify mRNAs significantly associated with OS (Overall survival) among DE-mRNAs, P value < 0.05 was considered statistically significant. mRNAs with HR (hazard rate ratio) > 1 and up-regulated in HCC, and mRNAs with HR < 1 and down-regulated in HCC were used to perform LASSO (least absolute shrinkage and selection operator) COX analysis for key mRNAs selection. 10-fold cross validation with minimum criteria was applied to identify the optimal parameter λ in LASSO Cox analysis [22, 23], depending on which the regression coefficients of most features were shrunk towards zero and the key mRNAs can be identified. The key mRNAs were used to build a predictive model for HCC. The mRNA risk score for each patient was computed according to the summation of mRNA expression value multiplied the corresponding coefficient from the LASSO Cox analysis. The performance of the mRNA model in predicting OS was evaluated through survival analysis, Harrell's concordance index (C-index) [24], receiver operating characteristic (ROC) curve.

50% HCC samples in TCGA were randomly selected as a test set. Survival analysis, C-index, and ROC analysis were performed to validate the predictive ability of the mRNA model.

The LIRI-JP dataset was used as an independent, external cohort to assess the generalizability and accuracy of the mRNA model through survival analysis.

Model based on lncRNA expression

The methods to construct, evaluate and validate lncRNA mode are similar to that we used in the mRNA model above. But, it is needed to pay attention that, to get enough differently expressed lncRNAs (DE-lncRNAs) to establish a stable model, P value < 0.0001 and |logFC| > 1.5 were set as the cut-off criterion. Meanwhile, due to the lack of an external dataset with complete lncRNA expression and corresponding prognostic information, the external verification of lncRNA mode cannot be approached.

Model based on miRNA expression

In the TCGA training set, miRNAs expressed in over 80% of samples were retained, the zero values were processed in the same way described above. 'limma' package was performed to identify differentially expressed miRNAs (DE-miRNAs) with P value < 0.01 and |logFC| > 1.5. Univariate Cox regression analysis with P value < 0.05 was used to identify miRNAs significantly associated with OS among DE-miRNAs. Due to few miRNAs significantly associated with OS were obtained, and LASSO Cox is suitable for analyzing high-dimensional data [25], backward stepwise Cox proportional hazard analysis[26] was used

to screen key miRNAs. Then the same methods used in the lncRNA model were performed to construct, evaluate, and validate the miRNA model, and the external verification of miRNA mode cannot be approached for the same reason.

Model based on CNV

In the TCGA training set, the segment mean value is used to reflect the CNV of DNA fragments. A segment is called a gain or loss if the segment mean value is greater or less than zero. According to the GENCODE v34 annotation file (downloaded from https://www.gencodegenes.org) and segment mean value of DNA fragments, genes with copy number variation (CNV genes) were identified in each sample. Chi-square analysis was used to compare the statistical difference of CNV genes between HCC and non-tumor samples, and then the univariate Cox regression analysis was performed to identify CNV genes significantly associated with OS. The LASSO Cox analysis was used to screen key CNV genes and construct the CNV model. The CNV risk score for each patient was computed according to the summation of CNV gene status (non-CNV = 0; CNV = 1) multiplied the corresponding coefficient from the LASSO Cox analysis. The evaluation and validation methods of the CNV model are the same as those in the lncRNA model. Similarly, for the same reason, we could not perform external validation for the CNV model.

Model based on SNP

In the TCGA training set, the high-frequency SNPs (not including synonymous mutation) in HCC samples were selected to perform univariate Cox analysis. Backward stepwise Cox proportional hazard analysis was performed to identify key SNPs and build the SNP model. The SNP risk score for each patient was computed according to the summation of SNP status (wild = 0; mutation = 1) multiplied the corresponding coefficient from stepwise Cox analysis. Then the same methods we used in the mRNA model were performed to evaluate and validate the SNP model. The LICA-FR dataset was used as an independent, external cohort to assess the generalizability and accuracy of the SNP model through survival analysis.

Model based on multi-omics

A model based on multi-omics was established via mRNA, lncRNA, miRNA, SNV, and SNP risk score with multiple Cox regression analysis. Nomogram was used for visualization of the prediction model. Survival analysis, calibration plot, C-index, ROC and decision curve analysis (DCA) were performed to evaluate and compare the predictive ability of the multi-omics model with single-omic model. The same methods we used in mRNA model were performed to validate the multi-omics model. The entire workflow is shown in Figure 1.

Statistical analysis

All data processing and statistical analysis were performed with R (https://www.r-project.org/, v 3.6.0). Chi-square or Fisher’s exact test was used to assess differences in categorical variables. Student t-test or

non-parametric Mann-Whitney test was used to detect differences in continuous variables. Volcano, box, and histogram plots were performed with the R package “ggplot2”. Heatmap was plotted with the R package “gplots”. The survival analysis and Cox proportional hazard regression analysis were carried out on the R package “survival”. The C-index, stepwise Cox analysis, and nomogram were performed with the R package “rms”. LASSO Cox analysis was performed using the R package “glmnet”. The ROC curve was plotted using the R package “qROC”. The DAC analysis was performed using the R package “stdca.R”. The summarized mutation plots were constructed using the R package “GenVisR”. Circus graph was drawn using the “RCircos” package.

Results

Construction and validation of mRNA model

320 HCC samples with complete mRNA expression profiling and survival information were kept as a training set. 267 DE-mRNAs (including 184 up-regulated and 83 down-regulated mRNAs in HCC) were identified (Figure S1A, 1B, and Table S1) to perform univariate Cox regression analysis. Among those DE-mRNAs, 124 mRNAs were significantly associated with OS (Table S2). Then 91 mRNAs with HR > 1 and up-regulated in HCC, and 31 mRNAs with HR < 1 and down-regulated in HCC were selected to perform LASSO Cox analysis for key mRNAs selection (Figure 2A). Parameter log (λ) = -3.154 (λ = 0.04269) chosen by 10-fold cross validation method with minimum criteria was regarded as the best value (Figure S1C). 16 key mRNAs with nonzero coefficients (Figure 2B), associated with HCC OS (Figure S1D), and significantly changed in HCC samples (Figure S1E1, S1E2) were used to build the mRNA model (Figure S1F). The mRNA risk score for each patient was computed as follows: mRNA risk score = $\sum \beta i \times$ exp-mRNA, where exp-mRNA is the expression level of key mRNA and β is the regression coefficient derived from the LASSO COX analysis (Table S9). The mRNA model was evaluated with C-index, ROC curve, and survival analysis (Figure 2C, 2D), which showed a relatively good predictive ability (C-index = 0.752). 160 HCC samples were randomly selected as a test set to validate the mRNA model, which still had a good performance (C-index = 0.744) (Figure 2E, 2F). Besides, the mRNA model was externally verified in the LIRI-JP dataset with survival analysis, which also had a decent performance (P value = 0.00016) (Figure 2G).

Construction and validation of lncRNA model

320 HCC samples with complete lncRNA expression profiling and survival information were retained as a training set. 540 up-regulated and 27 down-regulated lncRNAs in HCC were identified (Figure S2A, 2B, and Table S3) to perform univariate Cox regression analysis (Table S4).137 lncRNAs with HR > 1 and up-regulated in HCC, and 9 lncRNAs with HR < 1 and down-regulated in HCC were selected to perform LASSO COX analysis for key lncRNAs selection (Figure 3A). Parameter log (λ) = -2.99 (λ = 0.05027) chosen by 10-fold cross validation method with minimum criteria was regarded as the best value (Figure S2C). 20 key lncRNAs with nonzero coefficients (Figure 3B), associated with OS (Figure S2D), and significantly

changed in HCC samples (Figure S2E1, S2E2) were used to build the lncRNA model (Figure S2F). The lncRNA risk score for each patient was computed as follows: lncRNA risk score = $\sum$βi × exp-lncRNA, where exp-lncRNA is the expression level of key lncRNA and β is the regression coefficient derived from the LASSO Cox analysis (Table S9). In the training set, the AUC of the lncRNA model at 1, 3, and 5 year OS was 0.854, 0.796, and 0.806 respectively, while the C-index was 0.753(Figure 3C). In the test set, the AUC at 1, 3, and 5 year OS was 0.807, 0.785, and 0.767 respectively, while the C-index was 0.727(Figure 3E). In addition, the log-rank analysis revealed that scoring using the lncRNA risk score could discriminate the risk groups in the training set and test set (P value < 0.0001) (Figure 3D, 3F).

Construction and validation of miRNA model

321 HCC samples with complete miRNA and survival information were retained as a training set. 16 up-regulated and 70 down-regulated miRNAs in HCC were identified (Figure S3A, 3B, and Table S5). 5 miRNAs with HR > 1 and up-regulated in HCC, and 5 miRNAs with HR < 1 and down-regulated in HCC were selected via univariate Cox regression analysis (Figure 4A, Table S6). Through the stepwise Cox analysis, 5 key miRNAs were selected to build the miRNA model (Figure 4B, S3C, and S3D). The miRNA risk score for each patient was computed as follows: miRNA risk score = $\sum$βi × exp-miRNA, where exp-miRNA is the expression level of key miRNA and β is the regression coefficient derived from the stepwise Cox analysis (Table S9). Survival analysis showed that the high risk group has poor outcome in the training set (P = 0.00037) and test set (P = 0.027) (Figure 4D, 4F). Besides, whether in the training set or the test set, the AUC values of the miRNA model at 1,3, and 5 year point were all more than 0.68, and the C-index values were over 0.65(Figure 4C, 4E).

Construction and validation of CNV model

324 HCC samples with complete CNV and survival information were retained as a training set. 5006 genes with different copy number alteration between HCC and non-tumor samples (Figure 5A, Table S7) were selected to perform univariate Cox regression analysis, and 357 CNV genes significantly associated with OS were identified(Table S8). Then LASSO COX analysis was performed for key CNV genes selection. Parameter log (λ) = -2.634269 (λ = 0.07177142) chosen by 10-fold cross validation method with minimum criteria was regarded as the best value (Figure S4A). 5 key CNV genes with nonzero coefficients (Figure 5B), significantly different in HCC samples (Figure S4B) and associated with OS (Figure S4C) were used to build CNV model (Figure S4D). The CNV risk score for each patient was computed as follows: CNV risk score = $\sum$βi × CNV gene status, where β is the regression coefficient derived from the LASSO Cox analysis (Table S9). The CNV model was evaluated with survival analysis in the training set (Figure 5D) and test set (Figure 5F), which showed a worse prognosis in the high risk group (at least one key CNV gene with copy number alteration). Moreover, the AUC values of the CNV model at 1,3, and 5 year OS were all over 0.65, and the C-index values were more than 0.63(Figure 5C, 5E).

Construction and validation of SNP model

313 HCC samples with complete SNP and survival information were retained as a training set. 85 high-frequency SNPs (Figure 6A, Figure S5A) in HCC were selected to perform univariate Cox analysis, and 10 high-frequency SNPs significantly associated with OS were identified (Figure S5B). Through stepwise Cox analysis, 7 key SNPs were used to build the SNP model (Figure S5C). The SNP risk score for each patient was computed as follows: SNP risk score = $\sum \beta i \times$ SNP status, where β is the regression coefficient derived from the stepwise Cox analysis (Table S9). In the training set, the AUC of the SNP model at 1, 3, and 5 year OS was 0.799, 0.703, and 0.745 respectively, while the C-index was 0.709(Figure 6B). In the test set, the AUC at 1, 3, and 5 year OS was 0.745, 0.660, and 0.737 respectively, while the C-index was 0.683(Figure 6D). In addition, survival analysis showed that the high risk group (at least one key SNP with non-synonymous mutation) has poor prognosis in the training set ($P < 0.0001$), test set ($P < 0.0001$), and external validation set ($P = 0.029$) (Figure 6C, 6E, and 6F).

Construction and validation of multi-omics model

302 HCC samples with complete mRNA, lncRNA, miRNA, CNV, SNP, and survival information were retained as a training set. Through multiple Cox regression analysis, the 5 single-omic models were integrated to construct a multi-omics model and visualized as a nomogram (Figure 7A). A fairly good agreement was observed between the expected and observed outcomes for 1, 3, and 5 year OS in the calibration curves of the multi-omics model (Figure 7B). Whether in the training set or the test set, the AUC values of the multi-omics model at 1,3, and 5 year point were all over 0.840, and the C-index values were more than 0.800 (Figure 7C, 7H), which were significantly greater than those of the 5 single-omic models (all P values are less than 0.05)(Figure 7G). DCA analysis showed that the multi-omics model has better performance in predicting prognosis than the 5 single-omic models (Figure 7E, 7F). In addition, we stratified patients into low, medium, and high risk groups based on the total points of nomogram (cut-off points were selected at each tertile point), and found that scoring using the nomogram effectively discriminated the risk groups in the training set and test set ($P < 0.0001$) (Figure 7D, 7I).

Discussion

With the development of molecular biology techniques, the therapeutic, diagnostic, and predictive value of molecular targets in cancer is gradually becoming evident [27]. At present, although multiple literatures have constructed prognostic evaluation models for HCC based on molecular signatures, the vast majority of which only utilized single-omic data, and lack of information usually indicates the decline of predictive ability. In addition, while a predictive model for HCC based on the multi-omics data has been established by Chaudhary K et al., the predictive ability of which is not too good with a C-index of 0.7 in the training set. Therefore, a more accurate, multi-omics prognostic model is needed for guiding prognostic assessments and treatment decision-making of HCCs.

In this research, 16 key mRNAs, 20 key lncRNAs, 5 key miRNAs, 5 key CNV genes, and 7 key SNPs significantly associated with the prognosis of HCC were identified, and the previous researches have demonstrated that some of these key molecules play important roles in the occurrence, development, metastasis, and prognosis of HCC. Chen H et al. [28] have found that APLN could activate phosphatidylinositol 3-kinase/protein kinase B (PI3K/Akt) pathway via APLN receptor to promote HCC growth, and the overexpression of which is significantly associated with poorer survival of patients, that is in line with our analysis (Figure S1D). MYCN was overexpressed in clinical samples from HCC patients, and the knockdown of which can inhibit HCC cell growth and invasion[29]. Also, our analysis further proved that overexpression of MYCN predicted poor outcome in HCC (Figure S1D). lncRNA LINC01554 was identified as a novel tumor suppressor that could suppress tumorigenicity in HCC via Akt/mTOR signaling pathway[30], and the down-regulation of LINC01554 significantly predicted worse survival[31] (Figure S2D). Chen S et al. have demonstrated that knockdown of PCAT6 could inhibit cell proliferation and migration in HCC [32], and our analysis further found that PCAT6 was a significant prognostic risk factor in HCC(Figure S2D). miR-452-5p could be sponged by circular RNA-ABCB10, elevating expressions of neuropilin-1 (NRP1) and ABL related gene (ABL2) and leading inhibition of the HCC progression[33], which might act as an oncogene and was associated with poor survival(Figure 4B). PCDH9 could inhibit HCC cell proliferation via inducing cell cycle arrest at the G0/G1 phase, and the frequent deletion of which was observed in Lv J et al.'s [34] and our study (Figure 5A). Meanwhile, survival analysis in the current study further proved the tumor-suppressor function of PCDH9 in HCC (Figure S4C). TP53 mutation is one of the most common mutation in human tumors[35], and our research also clarified this point in HCC (Figure 6A). Moreover, the prognostic impact of TP53 mutation in HCC has been proved by the previous and current researches [36] (Figure S5B). All these findings above greatly enhanced the reliability of our analysis results. However, the roles of many key molecules (e.g., GNAZ, PZP, ZNNT1, LINC00862, CNV of CCNA1, ARID1A mutation, etc.) we identified in HCC are still unclear, and further study of their underlying mechanism is warranted.

In this study, we have constructed multiple predictive models with single-omic data (mRNA, lncRNA, miRNA, CNV, or SNP), and an integrated multi-omics model. The performance of each single-omic model in prognostic prediction was not bad with c-index ranged from 0.63 to 0.75 in training and test set. Meanwhile, through survival analysis, the mRNA and SNP risk score were also demonstrated to be significant prognostic factors in the independent external validation set (Figure 2G, 6F), which greatly increased the credibility and universality of our analysis results. Nevertheless, compared with the single-omic models, the integrated model based on 5-omics data showed more robust performance in prognostic prediction. To the best of our knowledge, to date, the multi-omics model we established has the strongest predictive ability in HCC, among all published models based on molecular markers, with c-index more than 0.80 and all AUC values at 1, 3, 5-years more than 0.84 (Figure 7C, 7H). Although an independent external dataset with complete mRNA, lncRNA, miRNA, CNV, SNP, and corresponding clinical information is missing to verify the multi-omics model, most of the key molecules were demonstrated to play important roles in HCC in previous studies, and some single-omic models we constructed were proved to be effective in external datasets. Therefore, our multi-omics model should be a reasonable and

effective prognostic evaluation model and has potential value in guiding prognostic assessments and treatment decision-making of HCCs in the clinic. Certainly, further studies are necessary to validate the biological roles of many key molecules in HCC, and independent external datasets with complete multi-omics and clinical information are required to enhance the reliability of our results, which will be the focus of our follow-up research.

Conclusions

The current study identified 16 key mRNAs, 20 key lncRNAs, 5 key miRNAs, 5 key CNV genes, and 7 key SNPs that are significantly associated with HCC prognosis, which may be helpful in studying underlying carcinogenesis mechanisms in HCC. The five single-omic models and an integrated multi-omics model we constructed showed potential prognostic values, which may better guide clinicians to make prognosis assessment and treatment decision-making for HCC patients.

Abbreviations

HCC: hepatocellular carcinoma; TCGA: The Genome Cancer Atlas; ICGC: International Cancer Genome Consortium; C-index: Harrell's concordance index; ROC: Receiver operating characteristic; AUC: Area under the curve; LASSO: Least absolute shrinkage and selection operator; logFC: Log fold change; TPM: Transcripts per million; DCA: Decision curve analysis; OS: Overall survival; HR: Hazard rate ratio; DE-mRNAs: Differentially expressed mRNAs; DE-lncRNAs: Differently expressed lncRNAs; DE-miRNAs: Differentially expressed miRNA; CNV: Copy number variation; SNP: Single nucleotide polymorphism.

Declarations

Acknowledgments

We thank the TCGA and ICGC working groups for generating public data.

Funding

This work was supported by the Natural Science Foundation of Guizhou Province of China (Grant No. QKHCG [2019] 4433, Grant No. QKHJC-ZK [2021] YB468 and Grant No. QKPTRC [2019]-036).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

XY and HZ designed and supervised the project. CX collected the data used in this study. CX and KY completed the data analysis and interpretation. XY and MC wrote the initial paper. All authors read and approved the final manuscript.

Availability of data and materials

The datasets analyzed during the current study are available in the TCGA (https://cancergenome.nih.gov/) and ICGC data portal (https://icgc.org/).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2020. CA Cancer J Clin. 2020;70:7–30.doi:10.3322/caac.21590.
2. Bakiri L, Hamacher R, Graña O, Guío-Carrión A, Campos-Olivas R, Martinez L, et al. Liver carcinogenesis by FOS-dependent inflammation and cholesterol dysregulation. J Exp Med. 2017;214:1387–409. doi:10.1084/jem.20160935.
3. Marrero JA, Kudo M, Bronowicki JP. The challenge of prognosis and staging for hepatocellular carcinoma. Oncologist. 2010;15(Suppl 4):23–33. doi:10.1634/theoncologist.2010-S4-23.
4. Yan X, Wan H, Hao X, Lan T, Li W, Xu L, et al. Importance of gene expression signatures in pancreatic cancer prognosis and the establishment of a prediction model. Cancer Manag Res. 2019;11:273–83. doi:10.2147/cmar.s185205.
5. Cao W, Lee H. Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. 2020;11:3675.doi:10.1038/s41467-020-17227-z.
6. Zhu S, Zhang X, Weichert-Leahey N, Dong Z, Zhang C, Lopez G, et al. LMO1 Synergizes with MYCN to Promote Neuroblastoma Initiation and Metastasis. Cancer Cell. 2017;32:310–23.e315.doi:10.1016/j.ccell.2017.08.002.
7. Lan T, Li H, Zhang D, Xu L, Liu H, Hao X, et al. KIAA1429 contributes to liver cancer progression through N6-methyladenosine-dependent post-transcriptional modification of GATA3. Mol Cancer. 2019;18:186. doi:10.1186/s12943-019-1106-z.

8. Chang RM, Xiao S, Lei X, Yang H, Fang F, Yang LY. miRNA-487a Promotes Proliferation and Metastasis in Hepatocellular Carcinoma. Clin Cancer Res. 2017;23:2593–604. doi:10.1158/1078-0432.ccr-16-0851.
9. Bezrookove V, Nosrati M, Miller JR 3rd, De Semir D, Dar AA, Vosoughi E, et al. Role of Elevated PHIP Copy Number as a Prognostic and Progression Marker for Cutaneous Melanoma. Clin Cancer Res. 2018;24:4119–25. doi:10.1158/1078-0432.ccr-18-0791.
10. Wang K, Li J, Guo H, Xu X, Xiong G, Guan X, et al. MiR-196a binding-site SNP regulates RAP1A expression contributing to esophageal squamous cell carcinoma risk and metastasis. Carcinogenesis. 2012;33:2147–54. doi:10.1093/carcin/bgs259.
11. Zhou JG, Zhao HT, Jin SH, Tian X, Ma H. Identification of a RNA-seq-based signature to improve prognostics for uterine sarcoma. Gynecol Oncol. 2019;155:499–507. doi:10.1016/j.ygyno.2019.08.033.
12. Nassiri F, Mamatjan Y, Suppiah S, Badhiwala JH, Mansouri S, Karimi S, et al. DNA methylation profiling to predict recurrence risk in meningioma: development and validation of a nomogram to optimize clinical management. Neuro Oncol. 2019;21:901–10. doi:10.1093/neuonc/noz061.
13. Liu GM, Xie WX, Zhang CY, Xu JW. Identification of a four-gene metabolic signature predicting overall survival for hepatocellular carcinoma. J Cell Physiol. 2020;235:1624–36. doi:10.1002/jcp.29081.
14. Liu GM, Zeng HD, Zhang CY, Xu JW. Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. Cancer Cell Int. 2019;19:138. doi:10.1186/s12935-019-0858-2.
15. Long J, Chen P, Lin J, Bai Y, Yang X, Bian J, et al. DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. Theranostics. 2019;9:7251–67. doi:10.7150/thno.31155.
16. Long J, Wang A, Bai Y, Lin J, Yang X, Wang D, et al. Development and validation of a TP53-associated immune prognostic model for hepatocellular carcinoma. EBioMedicine. 2019;42:363–74. doi:10.1016/j.ebiom.2019.03.022.
17. Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, et al. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. J Cell Mol Med. 2018;22:5928–38. doi:10.1111/jcmm.13863.
18. Wang X, Liao X, Yang C, Huang K, Yu T, Yu L, et al. Identification of prognostic biomarkers for patients with hepatocellular carcinoma after hepatectomy. Oncol Rep. 2019;41:1586–602. doi:10.3892/or.2019.6953.
19. Wang Z, Zhu J, Liu Y, Liu C, Wang W, Chen F, et al. Development and validation of a novel immune-related prognostic model in hepatocellular carcinoma. J Transl Med. 2020;18:67. doi:10.1186/s12967-020-02255-6.
20. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018;24:1248–59. doi:10.1158/1078-0432.ccr-17-0853.

21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. doi:10.1093/nar/gkv007.
22. Lao J, Chen Y, Li ZC. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. 2017;7:10353.doi:10.1038/s41598-017-10649-8.
23. Lohavanichbutr P, Mendez E, Holsinger FC, Rue TC, Zhang Y, Houck J, et al. A 13-gene signature prognostic of HPV-negative OSCC: discovery and external validation. Clin Cancer Res. 2013;19:1197–203. doi:10.1158/1078-0432.ccr-12-2647.
24. Huitzil-Melendez F, Capanu M, O'Reilly E, Duffy A, Gansukh B, Saltz L, et al. Advanced hepatocellular carcinoma: which staging systems best predict prognosis? J Clin Oncol. 2010;28:2889–95.doi.
25. Wei JH, Haddad A, Wu KJ, Zhao HW, Kapur P, Zhang ZL, et al. A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. Nat Commun. 2015;6:8699. doi:10.1038/ncomms9699.
26. Wu J, Zhou L, Huang L, Gu J, Li S, Liu B, et al. Nomogram integrating gene expression signatures with clinicopathological features to predict survival in operable NSCLC: a pooled analysis of 2164 patients. J Exp Clin Cancer Res. 2017;36:4.doi.
27. Connor AA, Denroche RE, Jang GH, Lemire M, Zhang A, Chan-Seng-Yue M, et al. Integration of Genomic and Transcriptional Features in Pancreatic Cancer Reveals Increased Cell Cycle Progression in Metastases. Cancer Cell. 2019;35:267–82. .e267.doi:10.1016/j.ccell.2018.12.010.
28. Chen H, Wong CC, Liu D, Go MYY, Wu B, Peng S, et al. APLN promotes hepatocellular carcinoma through activating PI3K/Akt pathway and is a druggable target. Theranostics. 2019;9:5246–60. doi:10.7150/thno.34713.
29. Yasukawa K, Liew LC, Hagiwara K, Hironaka-Mitsuhashi A, Qin XY, Furutani Y, et al. MicroRNA-493-5p-mediated repression of the MYCN oncogene inhibits hepatic cancer cell growth and invasion. 2020;111:869–880.doi:10.1111/cas.14292.
30. Zheng YL, Li L, Jia YX, Zhang BZ, Li JC, Zhu YH, et al. LINC01554-Mediated Glucose Metabolism Reprogramming Suppresses Tumorigenicity in Hepatocellular Carcinoma via Downregulating PKM2 Expression and Inhibiting Akt/mTOR Signaling Pathway. Theranostics. 2019;9:796–810. doi:10.7150/thno.28992.
31. Ding Y, Sun Z, Zhang S, Chen Y, Zhou B, Li G, et al. Down-regulation of Long Non-coding RNA LINC01554 in Hepatocellular Cancer and its Clinical Significance. J Cancer. 2020;11:3369–74. doi:10.7150/jca.40512.
32. Chen S, Chen Y, Qian Q, Wang X, Chang Y, Ju S, et al. Gene amplification derived a cancer-testis long noncoding RNA PCAT6 regulates cell proliferation and migration in hepatocellular carcinoma. 2019;8:3017–3025.doi:10.1002/cam4.2141.
33. Yang W, Ju HY, Tian XF. Circular. RNA-ABCB10 suppresses hepatocellular carcinoma progression through upregulating NRP1/ABL2 via sponging miR-340-5p/miR-452-5p. Eur Rev Med Pharmacol Sci. 2020;24:2347–57. doi:10.26355/eurrev_202003_20501.

34. Lv J, Zhu P, Zhang X, Zhang L, Chen X, Lu F, et al. PCDH9 acts as a tumor suppressor inducing tumor cell arrest at G0/G1 phase and is frequently methylated in hepatocellular carcinoma. Mol Med Rep. 2017;16:4475–82. doi:10.3892/mmr.2017.7193.
35. Toufektchan E, Lejour V, Durand R, Giri N, Draskovic I, Bardot B, et al. Germline mutation of MDM4, a major p53 regulator, in a familial syndrome of defective telomere maintenance. Sci Adv. 2020;6:eaay3511. doi:10.1126/sciadv.aay3511.
36. Shen T, Li SF, Wang JL, Zhang T, Zhang S, Chen HT, et al. TP53 R249S mutation detected in circulating tumour DNA is associated with Prognosis of hepatocellular carcinoma patients with or without hepatectomy. Liver Int. 2020;10.1111/liv.14581.

Figures

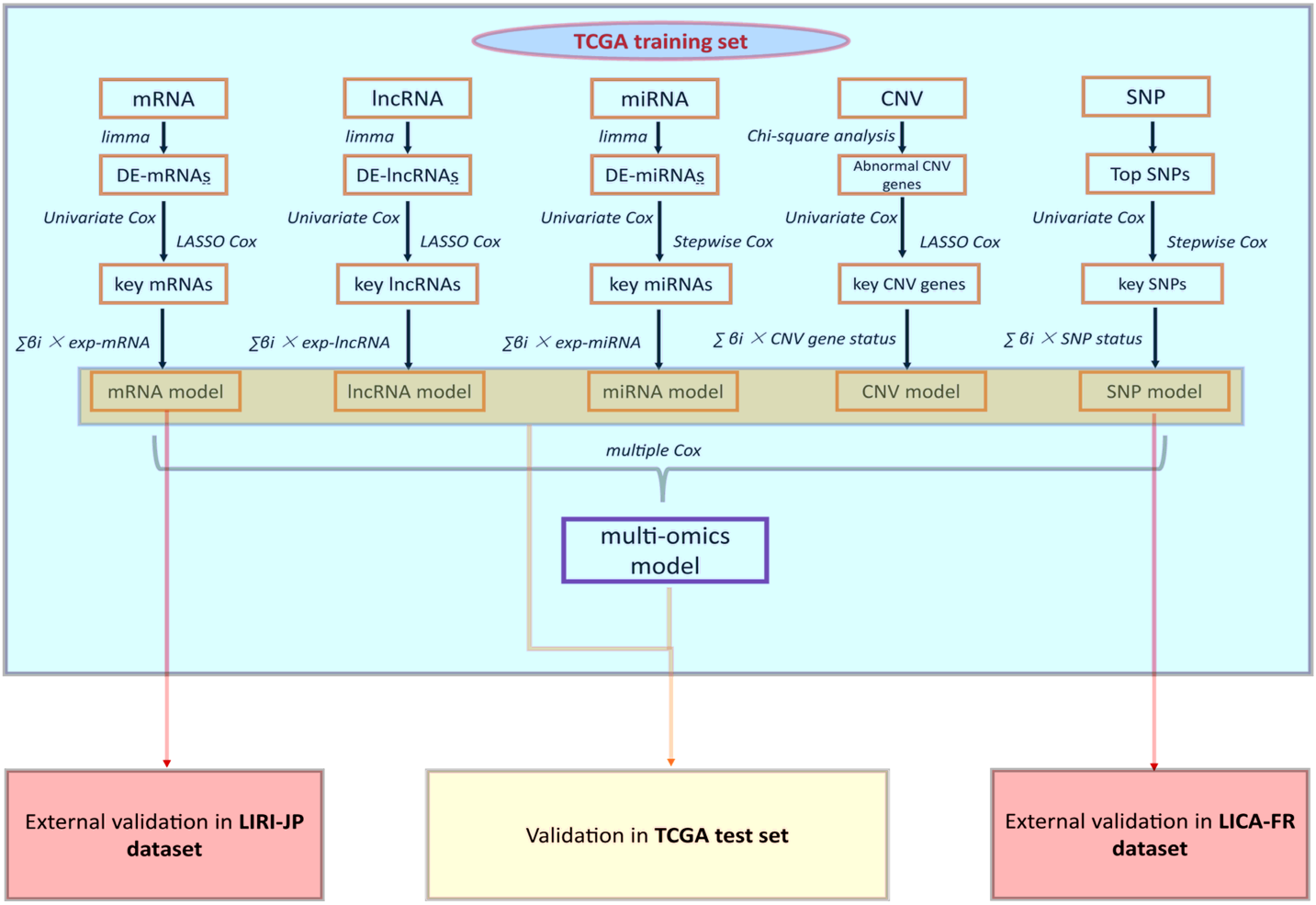


Figure 1

Overall workflow. The whole HCCs in TCGA was used as a training set and 50% HCCs were randomly selected as a test set. In the training set, the limma analysis was performed to identify DE-mRNAs, DE-lncRNAs, and DE-miRNAs. Chi-square analysis was used to screen abnormal CNV genes with the statistical difference between HCC and non-tumor samples, and the high-frequency SNPs (Top SNPs) in

HCC were selected for further analysis. The univariate Cox regression analysis, LASSO Cox analysis, and backward stepwise Cox proportional hazard analysis were used to identify key mRNAs, key lncRNAs, key miRNAs, key CNV genes, and key SNPs significantly associated with OS. Single-omic models (mRNA, lncRNA, miRNA, CNV, and SNP model) were constructed via the summation of key features (Expression level or status) multiplied by the corresponding coefficient from LASSO Cox analysis or stepwise Cox. Multi-omics model was constructed on the basis of 5 single-omic models through multiple Cox regression analysis. Besides, the 5 single-omic models and the multi-omics model were evaluated and verified in the training set and test set respectively, via survival analysis, C-index, ROC curve, and DCA analysis. Moreover, the mRNA and SNP model were externally validated in the LIRI-JP dataset and LICA-FR dataset, respectively.

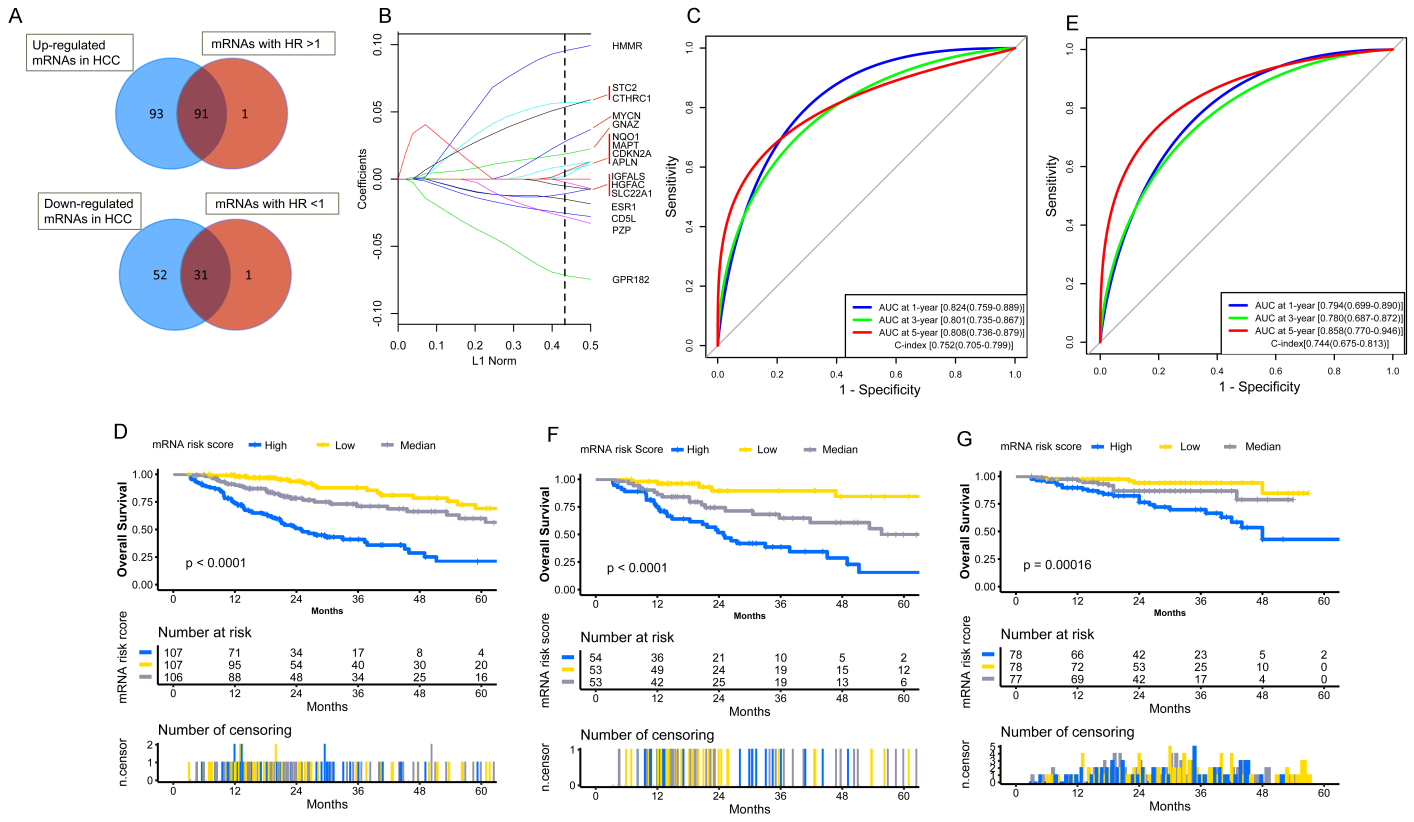


Figure 2

Construction and validation of the mRNA model. (A) Selection of mRNAs with HR > 1 and up-regulation, and mRNAs with HR < 1 and down-regulation in HCC. (B) LASSO coefficients of the 16 key mRNAs. The dotted vertical line is drawn at the λ value chosen by minimum criteria. L1 Norm represents the summation of absolute nonzero coefficients at each λ. Y-axis represents the values of nonzero coefficients at each λ. (C) The evaluation of the mRNA model via the ROC curve and C-index in the TCGA training set. (D) Kaplan-Meier survival analysis of the different risk groups stratified with the trisection of

the mRNA risk score in the TCGA training set. (E) The verification of the mRNA model via the ROC curve and C-index in the TCGA test set. (F) The verification of the mRNA model with Kaplan-Meier survival analysis in the TCGA test set. (G) The external validation of the mRNA model with Kaplan-Meier survival analysis in the LIRI-JP dataset.

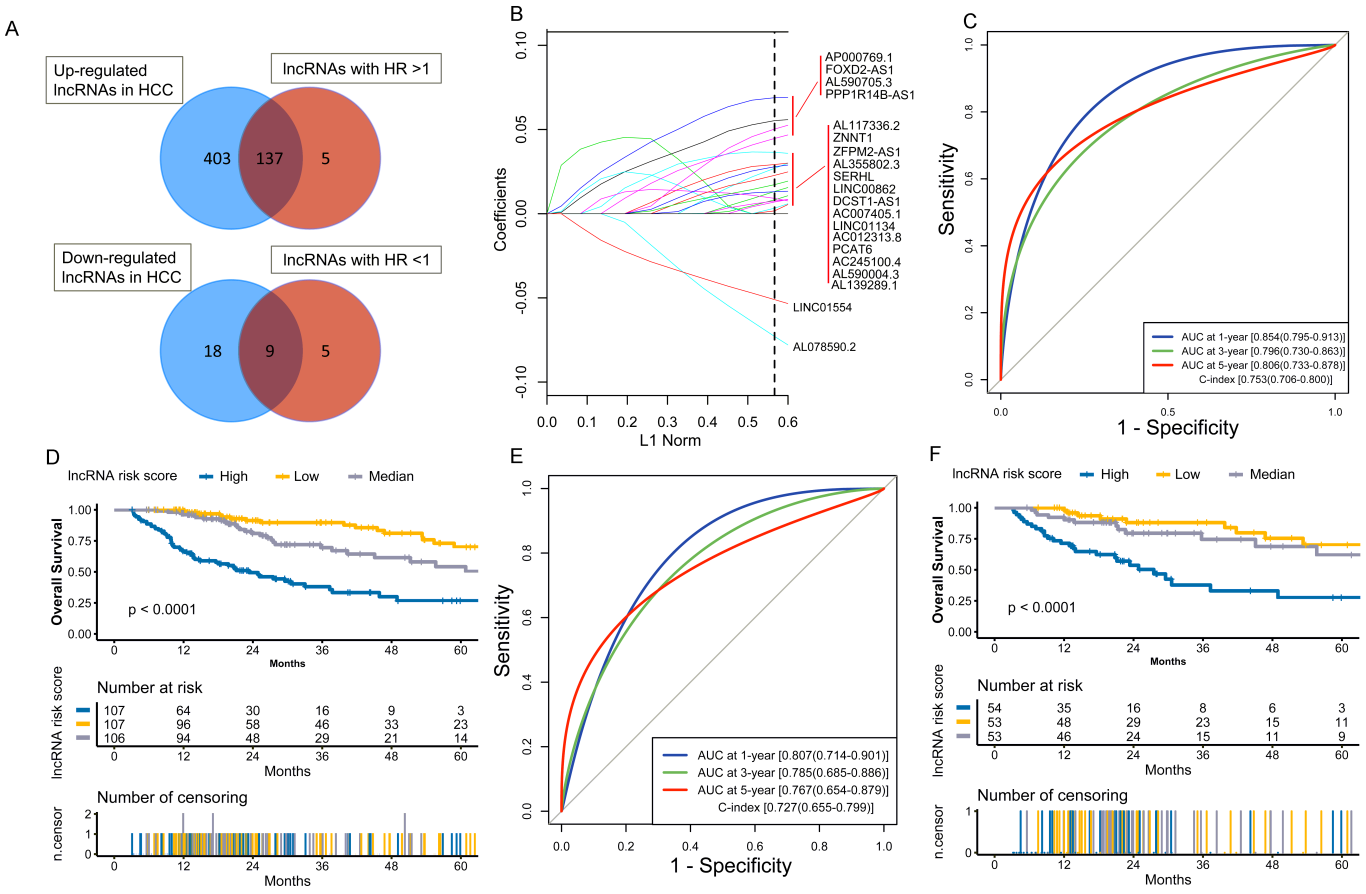


Figure 3

Construction and validation of the lncRNA model. (A) Selection of lncRNAs with HR > 1 and up-regulation, and lncRNAs with HR < 1 and down-regulation in HCC. (B) LASSO coefficients of the 20 key lncRNAs. (C) The evaluation of the lncRNA model via the ROC curve and C-index in the TCGA training set. (D) Kaplan-Meier survival analysis of the different risk groups stratified with the trisection of the lncRNA risk score in the TCGA training set. (E) The verification of the lncRNA model via the ROC curve and C-index in the TCGA test set. (F) The validation of the lncRNA model with Kaplan-Meier survival analysis in the TCGA test set.

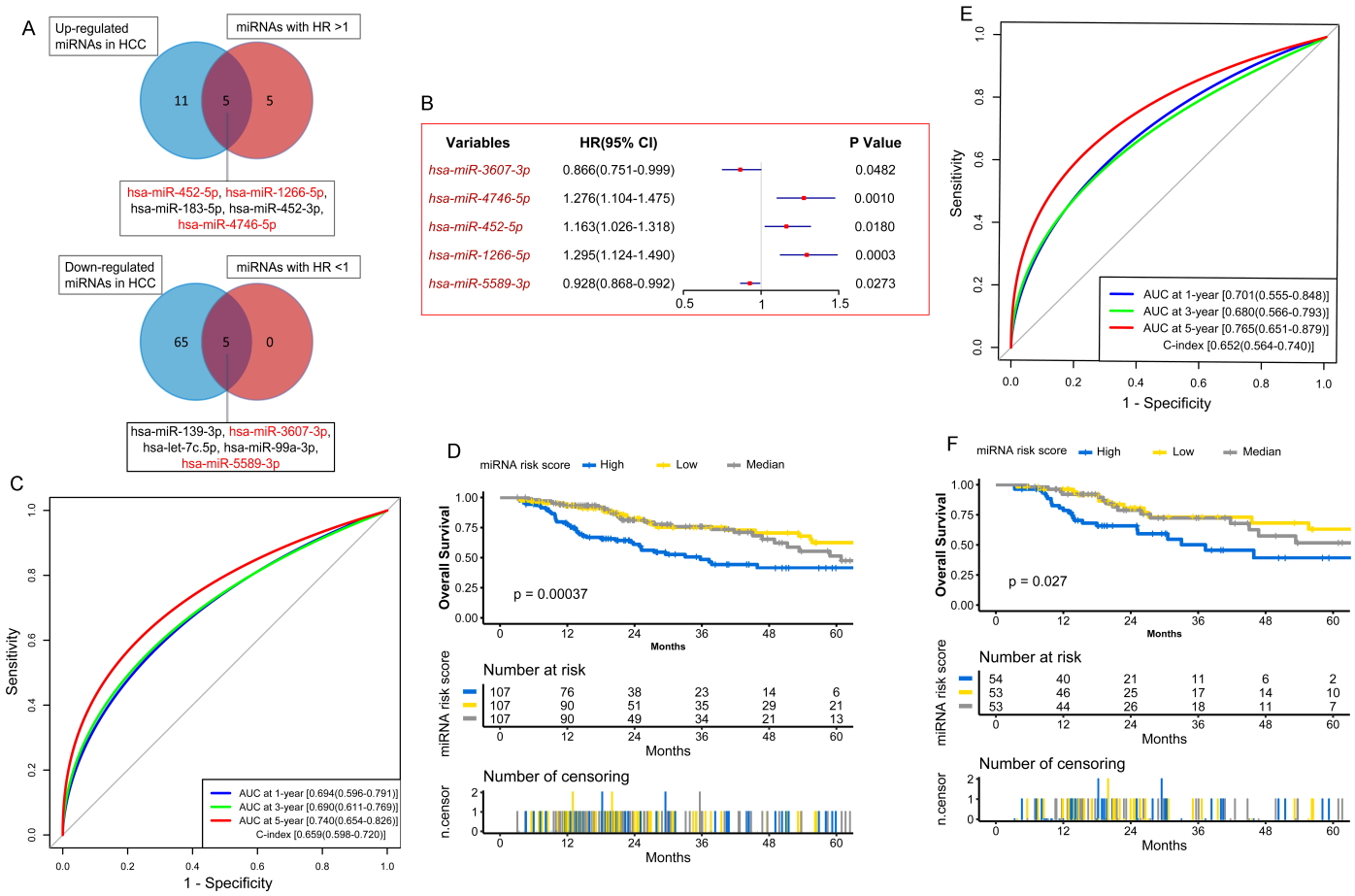


Figure 4

Construction and validation of the miRNA model. (A) Selection of miRNAs with HR > 1 and up-regulation, and miRNAs with HR < 1 and down-regulation in HCC. (B) Univariate Cox regression analysis of the 5 key miRNAs. (C) The evaluation of the miRNA model via the ROC curve and C-index in the TCGA training set. (D) Kaplan-Meier survival analysis of the different risk groups stratified with the trisection of the miRNA risk score in the TCGA training set. (E) The verification of the miRNA model via the ROC curve and C-index in the TCGA test set. (F) The validation of the miRNA model with Kaplan-Meier survival analysis in the TCGA test set.

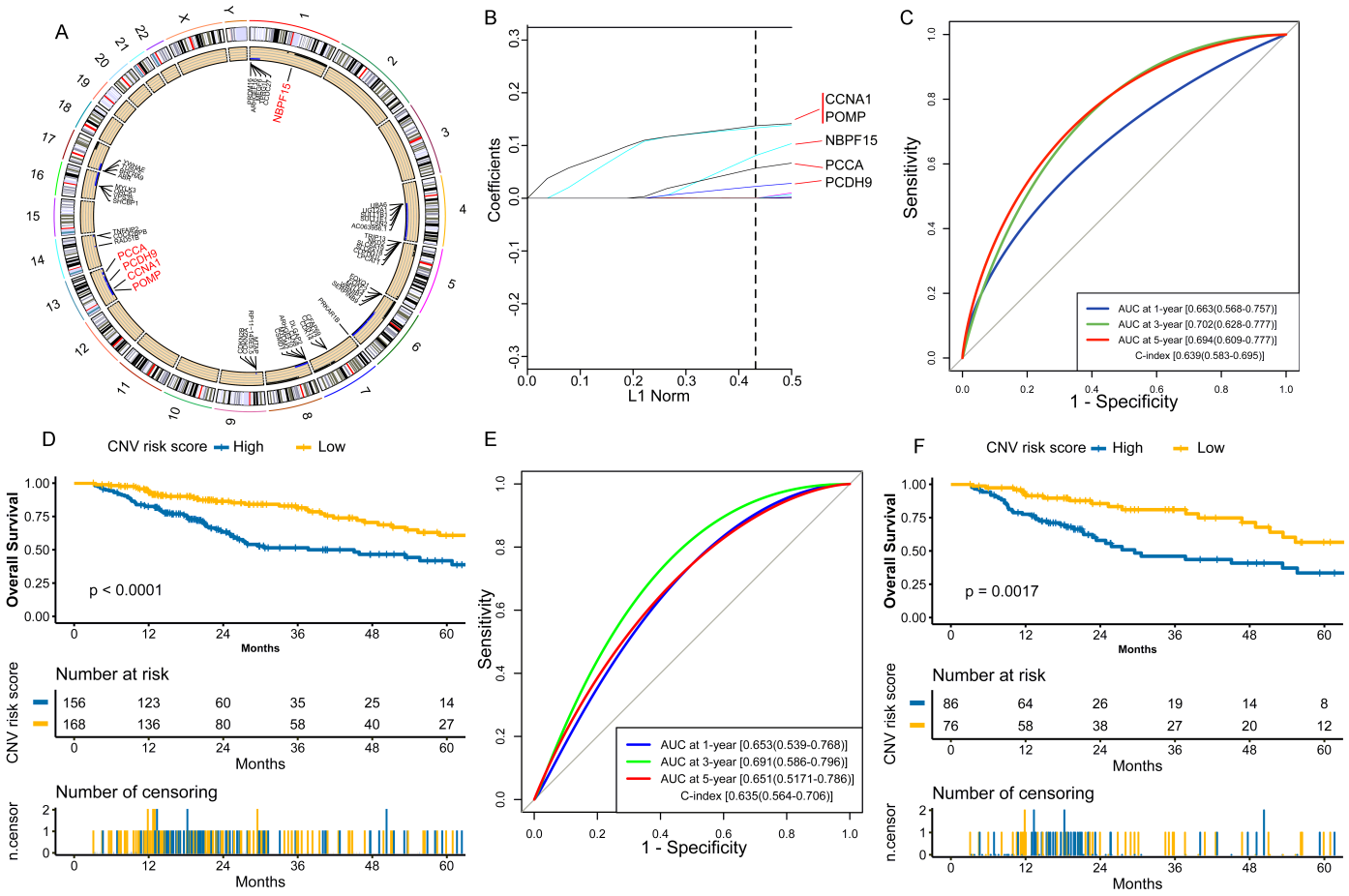


Figure 5

Construction and validation of the CNV model. (A) Circos plot shows a part of genes with different copy number alteration between HCC and non-tumor samples. The blue dots represent genes with copy number loss, and the black dots represent genes with copy number gain. (B) LASSO coefficients of the 5 key CNV genes. (C) The evaluation of the CNV model via ROC curve and C-index in the TCGA training set. (D) Kaplan-Meier survival analysis of the different risk groups stratified with the CNV risk score in the TCGA training set. Patients with no copy number alteration of the 5 key CNV genes were attributed to the low risk group, and the others were attributed to the high risk group. (E) The verification of the CNV model via the ROC curve and C-index in the TCGA test set. (F) The validation of the CVN model with Kaplan-Meier survival analysis in the TCGA test set.

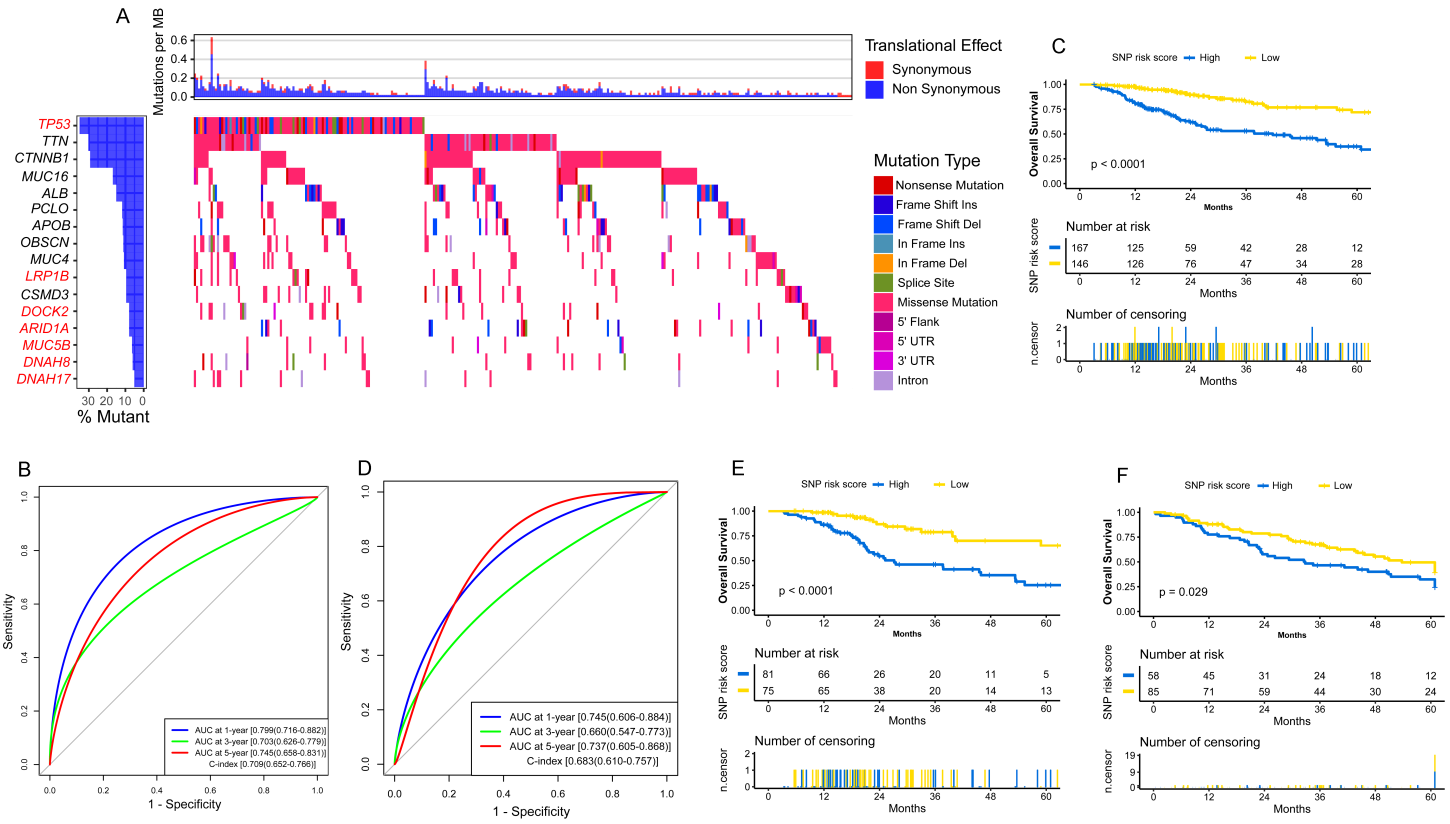
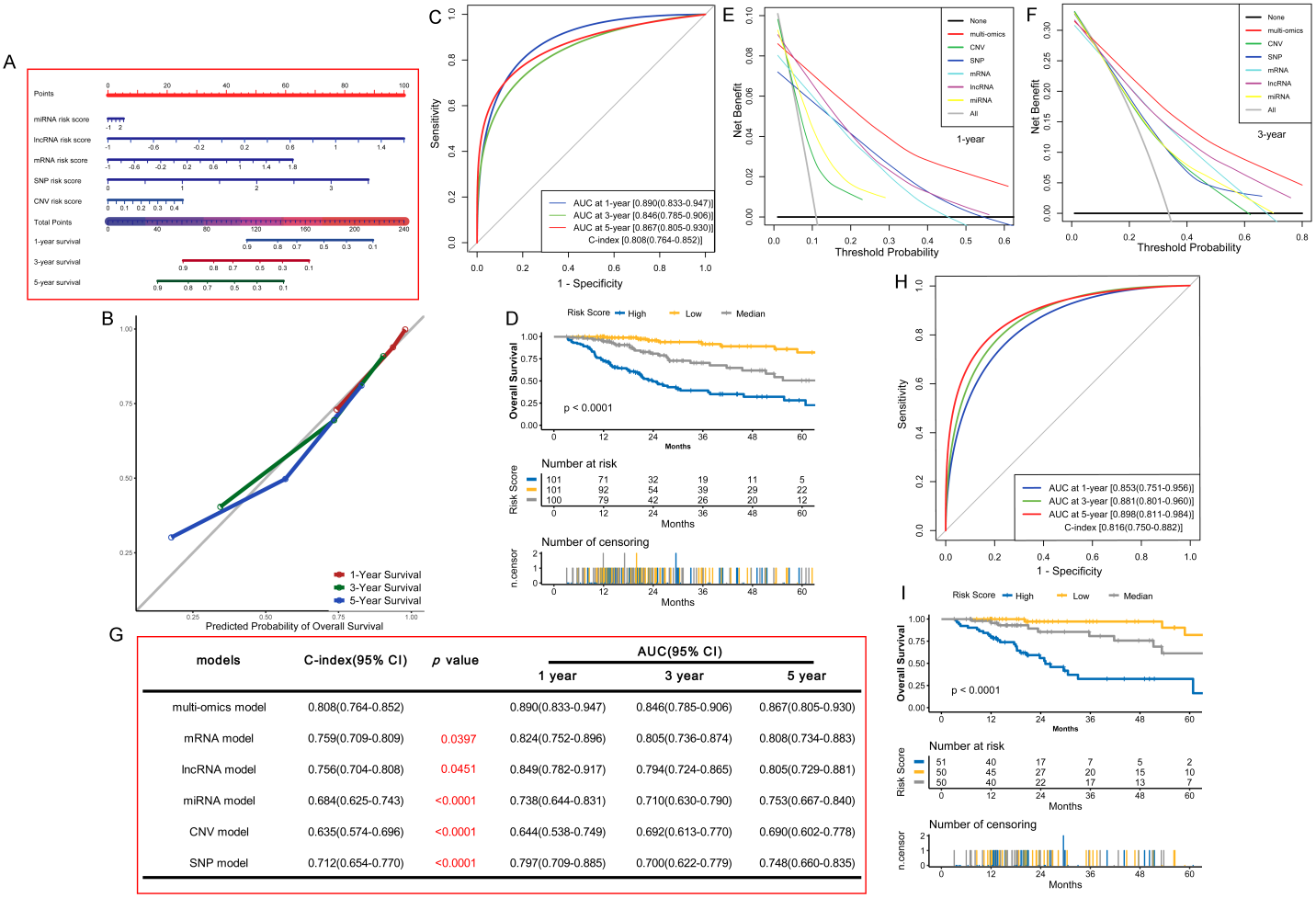


Figure 6

Construction and validation of the SNP model. (A) Distributions of various mutation types of the 16/85 high-frequency SNPs. The histogram in the top indicates the sum of non-synonymous and synonymous mutations in every case, and the histogram on the right stands for the sample number suffering from gene mutation. In the heat map, the various colors stand for various mutation types, whereas the white color represents no mutation. (B) The evaluation of the SNP model via the ROC curve and C-index in the TCGA training set. (C) Kaplan-Meier survival analysis of the different risk groups stratified with the SNP risk score in the TCGA training set. Patients with no mutation of the 7 key SNPs were attributed to the low risk group, and the others were attributed to the high risk group. (D) The verification of the SNP model via the ROC curve and C-index in the TCGA test set. (E) The validation of the SNP model with Kaplan-Meier survival analysis in the TCGA test set. (F) The external validation of the SNP model with Kaplan-Meier survival analysis in the LICA-FR dataset.



G

models	C-index(95% CI)	*p* value	AUC(95% CI)		
			1 year	3 year	5 year
multi-omics model	0.808(0.764-0.852)		0.890(0.833-0.947)	0.846(0.785-0.906)	0.867(0.805-0.930)
mRNA model	0.759(0.709-0.809)	0.0397	0.824(0.752-0.896)	0.805(0.736-0.874)	0.808(0.734-0.883)
lncRNA model	0.756(0.704-0.808)	0.0451	0.849(0.782-0.917)	0.794(0.724-0.865)	0.805(0.729-0.881)
miRNA model	0.684(0.625-0.743)	<0.0001	0.738(0.644-0.831)	0.710(0.630-0.790)	0.753(0.667-0.840)
CNV model	0.635(0.574-0.696)	<0.0001	0.644(0.538-0.749)	0.692(0.613-0.770)	0.690(0.602-0.778)
SNP model	0.712(0.654-0.770)	<0.0001	0.797(0.709-0.885)	0.700(0.622-0.779)	0.748(0.660-0.835)

Figure 7

Construction and validation of the multi-omics model. (A) Nomogram of the multi-omics model for predicting 1-, 3-, and 5-year OS in the TCGA training set. (B) Calibration plot for 1-, 3-, and 5-year OS of the multi-omics model in the TCGA training set. (C) The evaluation of the multi-omics model via the ROC curve and C-index in the TCGA training set. (D) Kaplan-Meier survival analysis of the different risk groups stratified with the trisection of the total point of the proposed nomogram in the TCGA training set. (E, F) Decision curve analysis for the multi-omics model and the 5 single-omic models at 1- and 3-year point in the TCGA training set. (G) Comparison of the predictive power of the multi-omics model and the 5 single-omic models with C-index, and ROC analysis in the TCGA training set. (H) The verification of the multi-omics model via the ROC curve and C-index in the TCGA test set. (I)The validation of the multi-omics model with Kaplan-Meier survival analysis in the TCGA test set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- Supplementaryfigures.pdf