Research Paper

Integrative analysis of DNA methylation and gene expression reveals distinct hepatocellular carcinoma subtypes with therapeutic implications

Xiaowen Huang[1,*], Chen Yang[2,*], Jilin Wang[1], Tiantian Sun[1], Hua Xiong[1]

[1]State Key Laboratory of Oncogenes and Related Genes, Key Laboratory of Gastroenterology and Hepatology, Ministry of Health, Division of Gastroenterology and Hepatology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University, Shanghai Cancer Institute, Shanghai Institute of Digestive Disease, Shanghai, China
[2]State Key Laboratory of Oncogenes and Related Genes, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
*Equal contribution

Correspondence to: Hua Xiong, Tiantian Sun; **email:** huaxong88@126.com, suntt2005@126.com
Keywords: hepatocellular carcinoma, DNA methylation-driven genes, classification, integrative analysis, gene expression
Received: December 9, 2019 **Accepted:** March 2, 2020 **Published:** March 22, 2020

ABSTRACT

We aimed to develop an HCC classification model based on the integrated gene expression and methylation data of methylation-driven genes. Genome, methylome, transcriptome, proteomics and clinical data of 369 HCC patients from The Cancer Genome Atlas Network were retrieved and analyzed. Consensus clustering of the integrated gene expression and methylation data from methylation-driven genes identified 4 HCC subclasses with significant prognosis difference. HS1 was well differentiated with a favorable prognosis. HS2 had high serum α-fetoprotein level that was correlated with its poor outcome. High percentage of *CTNNB1* mutations corresponded with its activation in WNT signaling pathway. HS3 was well differentiated with low serum α-fetoprotein level and enriched in metabolism signatures, but was barely involved in immune signatures. HS3 also had high percentage of *CTNNB1* mutations and therefore enriched in WNT activation signature. HS4 was poorly differentiated with the worst prognosis and enriched in immune-related signatures, but was barely involved in metabolism signatures. Subsequently, a prediction model was developed. The prediction model had high sensitivity and specificity in distributing potential HCC samples into groups identical with the training cohort. In conclusion, this work sheds light on HCC patient prognostication and prediction of response to targeted therapy.

INTRODUCTION

Hepatocellular carcinoma (HCC) is the sixth most common cancer and the third leading cause of cancer-related death worldwide [1]. It is estimated that by 2020 the number of HCC cases will reach 78,000 in Europe and 27,000 in the United States [1]. A better understanding of the underlying mechanisms of HCC diversity will increase the chances for effective treatment and improvement in survival rate.

Genome-wide analyses of mRNA expression profiles have contributed to developing HCC targeted therapies over the past two decades. Boyault et al. performed transcriptome analyses on 57 HCCs and 3 hepatocellular adenomas. Six robust subgroups of HCC (G1-G6) associated with clinical and genetic characteristics were identified [2]. Hoshida et al. classified a total of 603 patients into 3 robust HCC subclasses (S1, S2, and S3) based on gene expression profiles. Each subclass was correlated with clinical

parameters such as tumor size and extent of cellular differentiation [3]. Chiang et al. divided 91 HCC samples into 5 subclasses based on gene expression profiles [4]. Lee et al. analyzed global gene expression patterns of 91 HCCs. The samples were classified into two distinctive subclasses that were highly associated with patient survival [5]. The existing classifications are mainly based on gene expression profiles, and few of them are based on DNA methylation profiles. However, HCC is a complex disease arising from accumulation of both genetic and epigenetic alterations [6]. Transcriptome data alone is insufficient for revealing the heterogeneity of HCC. It has been demonstrated that classification of HCC with DNA methylation data is clinically significant [7].

As one of the core elements in epigenetic modifications, DNA methylation participates in a diverse range of cellular and biological processes such as cell differentiation, aging, tissue-specific gene expression, genome stability and genomic imprinting [8]. In addition to the implication during normal development, DNA methylation involves in pathologies such as carcinogenesis [9]. Hypermethylation of CpG islands in promoter sequences can cause epigenetic inactivation of tumor suppressor genes followed by mRNA transcript repression [9]. Unlike DNA aberrations, epigenetic changes are reversible, which makes them potential therapeutic targets [9].

Aberrant methylation of several tumor suppressor genes and tumor-related genes such as *RASSF1A*, *hMLH1* and *SOCS1* is constantly identified in HCC [10]. *TMS1* is a proapoptotic gene with promoter methylation observed in 80% HCC patients [11]. Aberrant methylation of *SEMA3B* is reported in 80% HCCs [11]. *SEMA3B* induces apoptosis and is detected in lung cancers and gliomas [11]. A number of studies on these DNA methylation-driven genes have already been published [12, 13].

To obtain a better understanding of HCC heterogeneity, we established an HCC classification based on integrated gene expression and methylation data of methylation-driven genes (MDGs). Consensus clustering identified 4 HCC subclasses significantly associated with prognosis value. The 4 subclasses showed distinct clinical features and enrichment in different signatures. Somatic mutations and copy number mutations data were analyzed and visualized. Besides, HCC patients were clustered into distinct CpG island methylator phenotype (CIMP) based on the methylation level of 674 most variable CpGs. The accuracy of the transcriptome-based prediction model constructed by machine learning algorithms was favorable.

RESULTS

Identification of 4 HCC subclasses

Messenger RNA expression data and methylation data were integrated under the same sample with the *MethylMix* R package [14] to identify MDGs. 401 MDGs with $|logFC| > 0$, $P < 0.05$ and $|Cor| > 0.3$ were reserved for subsequent analyses (Supplementary Table 1). Then, 369 HCC patients were clustered based on the integrated mRNA expression and methylation data of 401 MDGs by "ExecuteCNMF" function in *CancerSubtypes* package [15]. Optimal number of clusters was determined according to comprehensive consideration of Silhouette width value and clinical significance (Figure 1A, 1B and Supplementary Figure 1). When the samples were classified into 2, 3 and 4 subtypes, average silhouette widths were 0.93, 0.97 and 0.94, respectively. If Silhouette width is close to 1, it means the samples are well classified. Silhouette widths for 2, 3 and 4 clusters were all close to 1. Besides, when the samples were classified into 3 groups, no significance in survival was identified (p=0.0692). We considered it more appropriate to divide the samples into 4 subclasses to provide more information for diagnosis based on their different molecular features. The 4 HCC subclasses identified were named HCC Subclass 1 (HS1), HCC Subclass 2 (HS2), HCC Subclass 3 (HS3) and HCC Subclass 4 (HS4). To validate subclasses' assignments, we performed t-distributed stochastic neighbor embedding (t-SNE) to decrease the dimension of features and found that subtype designations were largely concordant with two-dimensional t-SNE distribution patterns (Figure 1C).

Survival analysis was conducted, and significant prognostic difference was observed when using overall survival (OS) as an endpoint (log-rank test $P = 0.0057$, Figure 1D). A longer median survival time (MST) was detected for HS1 (MST=2839 days, 95% CI: 1749-3929 days) compared with HS2 (MST= 1622 days, 95% CI: 929-2315 days, $P = 0.0609$), HS3 (MST=1818 days, 95% CI: 1213-2423 days, $P = 0.5308$) and HS4 (MST= 1135 days, 95% CI: 450-1820 days, $P = 0.0034$). However, when using recurrence free survival (RFS) as an endpoint, there was no significant prognostic difference among HCC classifications (Figure 1D and Supplementary Table 2).

The characteristics of 401 MDGs were then investigated. Metabolism and immune relevant gene lists were obtained from previous studies [16, 17]. Through intersecting these gene lists with 401 MDGs, we identified metabolism and immune associated MDGs (100 MDGs for metabolism and 51 for immunity). Besides, considering that DNA methylation alterations in tumor suppressor genes (TSGs)

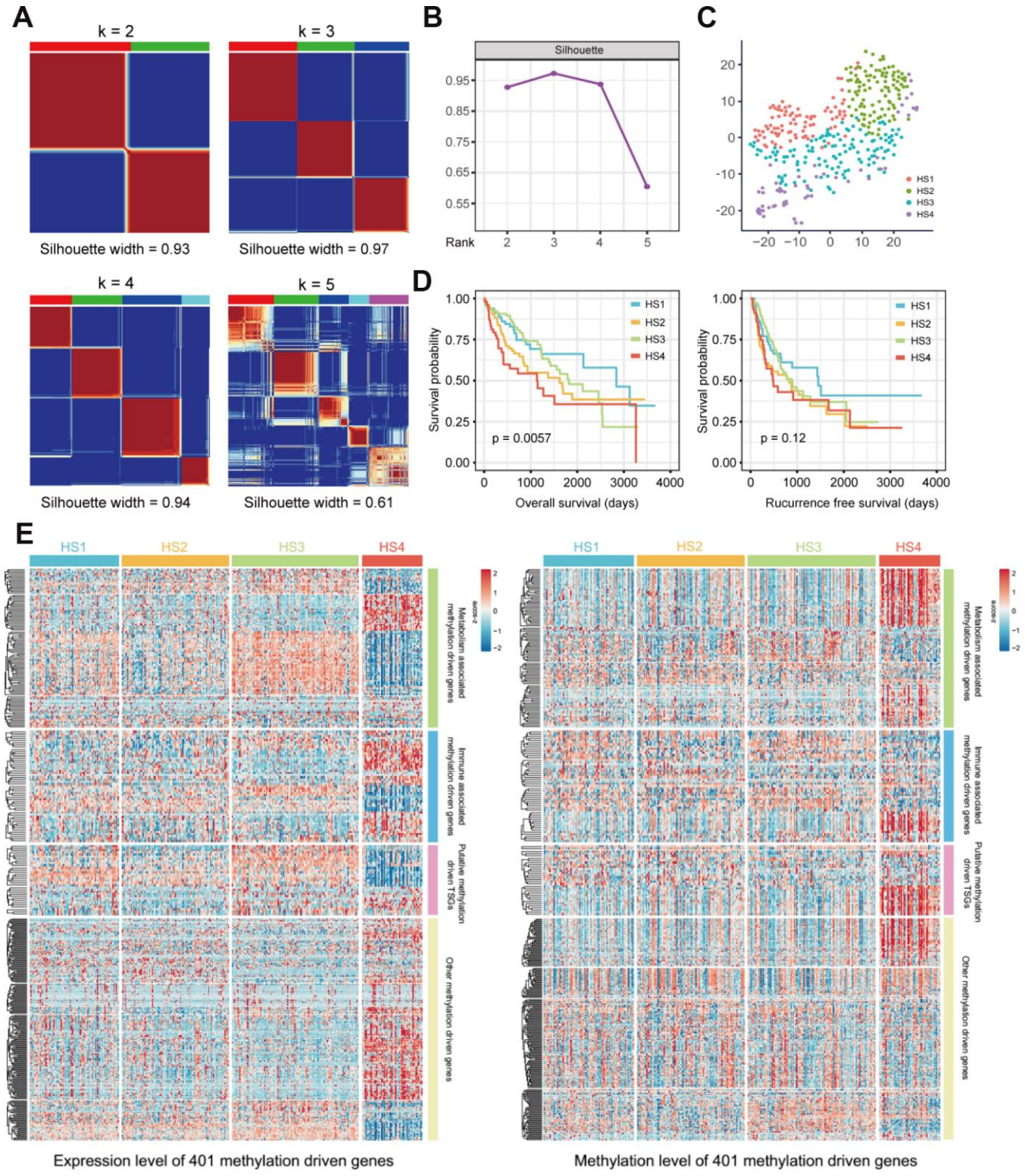


Figure 1. Identification of HCC subclasses based on integrated transcriptome and methylation data of MDGs. (**A**) Consensus matrix for k = 2 to k = 5. (**B**) Silhouette values under corresponding k values. (**C**) T-SNE analysis of mRNA expression data from tumor samples included in the cluster analysis (**D**) OS and RFS of 4 HCC subclasses. Statistical significance of differences was determined by Log-rank test. (**E**) Heatmaps show the expression and methylation level of 401 MDGs in HCC subclasses. 401 MDGs were divided into 4 groups, including metabolism associated MDGs, immune associated MDGs, putative methylation driven TSGs and other MDGs. HCC: hepatocellular carcinoma; MDG: methylation driven gene; t-SNE: t-distributed stochastic neighbor embedding; OS: overall survival; RFS: recurrence free survival; TSG: tumor suppressor genes.

were involved in carcinogenesis, we intersected 401 MDGs with putative TSGs to obtain putative methylation driven TSGs. The expression and methylation levels of these MDGs were both visualized in Figure 1E and detailed information was listed in Supplementary Table 3.

Correlation of the HCC subclasses with clinical characteristics and classical classification

The relationships between HCC classifications and clinical characteristics were then investigated (Figure 2 and Supplementary Table 4). Results revealed that HS2 was associated with histologic grade G3/G4 (46/99 vs 82/259, P = 0.0089) and high serum α-fetoprotein (AFP) level (37/75 vs 58/201, P = 0.0014). HS3 was associated with lower proportion of virus infection (44/87 vs 58/166, P = 0.0160), histologic grade G1/G2 (93/120 vs 137/238, P = 0.0002), and low serum AFP level (77/90 vs 105/186 in the rest, P < 0.0001).

Then, our classification was also compared with previously reported HCC molecular subclasses, including Boyault's classification [2] (G1 to G6), Chiang's classification [4] (5 classes), Hoshida's classification [3] (S1, S2, and S3), and The Cancer Genome Atlas (TCGA) classification [18] (iCluster1, iCluster2, and iCluster3). Results suggested that HS1 was significantly associated with Chiang's Proliferation class (31/85 vs 52/278 in the rest, P = 0.0006). HS2 was significantly associated with Hoshida's S2 (47/100 vs 53/263 in the rest, P < 0.0001). HS3 was significantly associated with Boyault's G5/G6 (66/122 vs 65/241 in the rest, P < 0.0001), Chiang's *CTNNB1* class (45/122 vs 44/241 in the rest, P = 0.0001), and Hoshida's S3 (111/122 vs 114/241 in the rest, P < 0.0001). HS4 was significantly associated with Boyault's G3 (45/56 vs 97/307 in the rest, P < 0.0001), Hoshida's S1 (28/56 vs 10/307 in the rest, P < 0.0001), and TCGA iCluster1 (20/33 vs 41/145 in the rest, P = 0.0004).

Correlation between HCC subclasses and CIMP

Considering that MDGs based classification may result in different methylation status among subclasses, we then explored the methylation characteristics of 4 HCC subclasses. First, according to previously mentioned approach to find CIMP in HCC [7], we clustered samples into distinct groups using K-means method based on the methylation level of 674 most variable CpGs. Among these groups, C2 was defined as non-CIMP group with the lowest methylation level of 674 CpGs. C7 was defined as CIMP-H group with the

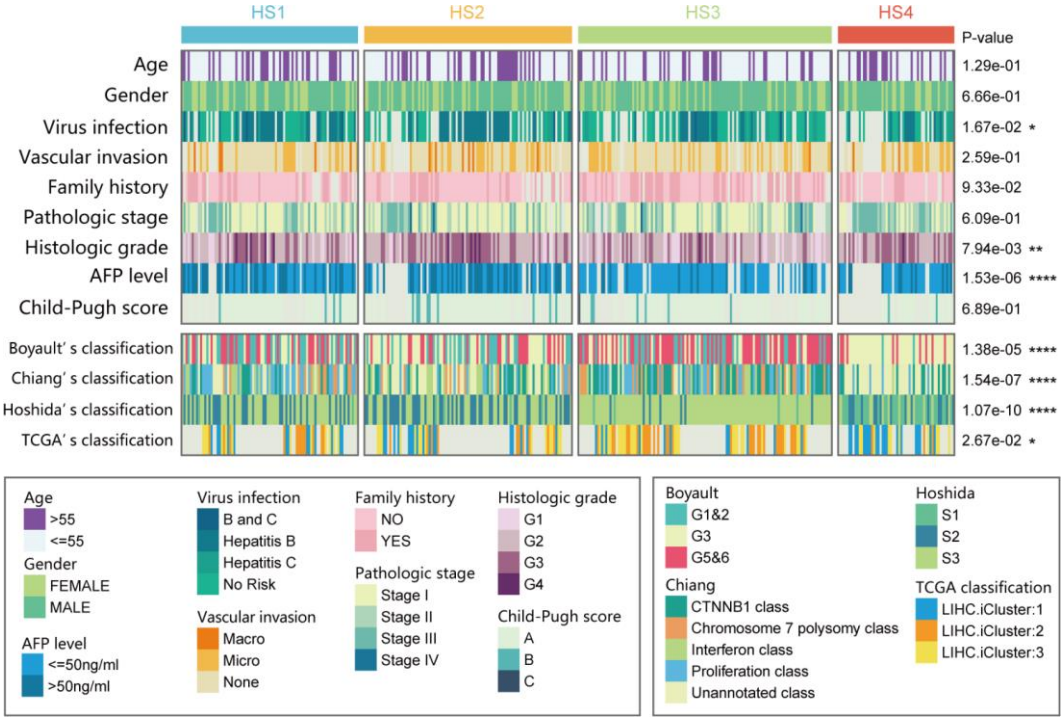


Figure 2. Correlation of our classification (HS1, HS2, HS3 and HS4) with distinct clinical characteristics and previously published HCC subclasses. Prediction of previously published HCC classifications was performed with NTP analyses. Statistical significance of differences was determined by Chi-square test (ns represents no significance, * represents P < 0.05, ** represent P < 0.01, *** represent P < 0.001, **** represent P < 0.0001). HCC: hepatocellular carcinoma; NTP: Nearest Template Prediction.

highest methylation level of 674 CpGs. The remaining groups with moderate methylation level of 674 CpGs were defined as CIMP-L group (Supplementary Figure 2A). Although no significant prognostic difference was observed among groups, CIMP-H (C7) group still showed a trend towards poorer prognosis (Supplementary Figure 2B and 2C). The relationship between our classification and CIMP was visualized in Supplementary Figure 2D, and results of statistical analysis revealed that samples in non-CIMP were more enriched in HS3 and HS4 than HS1 and HS2 (63/180 vs 44/189, $P = 0.0131$).

Correlation of HCC subclasses with metabolism and immune associated signatures

The outcome that 100 of the 401 MDGs were involved in metabolism and 51 were involved in immunity drove us to investigate the characteristics of metabolism and immunity in HCC subclasses (Figure 3). First,

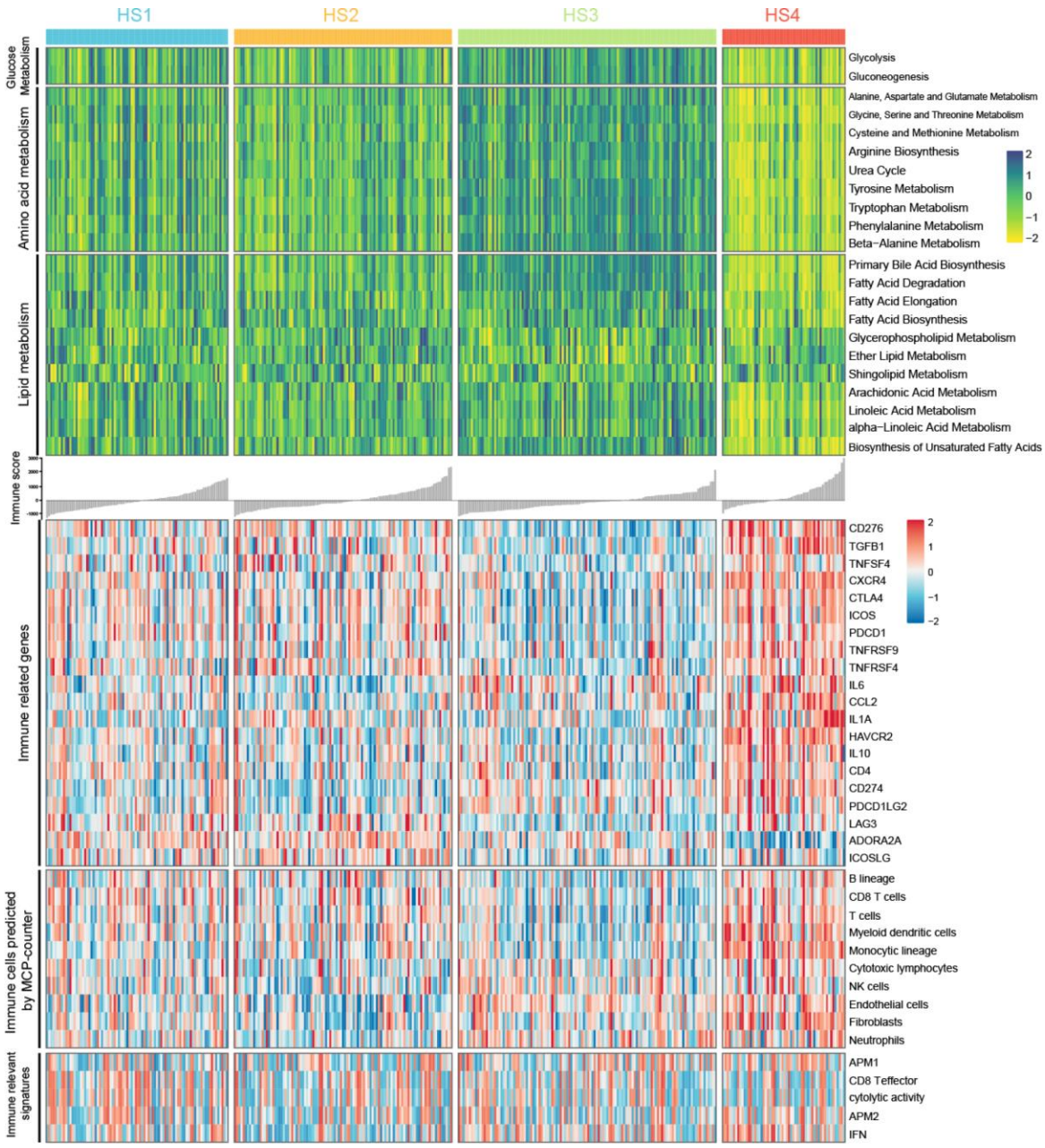


Figure 3. Heatmaps show difference in metabolism signatures (glucose metabolism, amino acid metabolism, and lipid metabolism), immune related genes expression, immune-associated signatures and other signatures, immune and stromal cell populations predicted by MCP-counter among 4 HCC subclasses (see detailed information in Supplementary Figure 2).

metabolism and immune associated processes were quantified using Gene Set Variation Analysis (GSVA) and microenvironment cell populations-counter (MCP-counter) methods. Then, statistical analyses were conducted, and results suggested that metabolic and immune processes in distinct classifications differed greatly (detailed statistical analyses were shown in Supplementary Figure 3A). Particularly, HS3 had higher signature scores for metabolism than other subclasses, except several lipid metabolic processes, including glycerophospholipid metabolism, ether lipid metabolism, shingolipid metabolism, arachidonic acid metabolism, and alpha−linoleic acid metabolism. HS4 exhibited lower enrichment in these metabolic processes than other subclasses. HS1 and HS2 had moderate signature scores, and there was also no significant difference between HS1 and HS2.

For immune associated processes, we first investigated the association between subclasses and the expression of 20 potentially targetable immune related genes, and results indicated that HS4 exhibited higher expression for multiple immune related genes (*CD276*, *TGFB1*, *CXCR4*, *CTLA4*, *ICOS*, *TNFRSF9*, *CCL2*, *IL1A*, *HAVCR2*, *IL10*, *CD274*, and *PDCD1LG2*) and lower expression for *ADORA2A* than other subclasses (Supplementary Figure 3B). HS3 exhibited lower expression for *CD276*, *TGFB1*, *CTLA4*, *ICOS*, *PDCD1*, *TNFRSF4*, *CD274*, and *LAG3* than other subclasses. No significant difference for immune related gene expression was detected between HS1 and HS2. We then explored immune infiltration of 4 subclasses. The abundance of 10 immune and stromal related cell types was calculated using MCP-counter algorithm. Significant difference was observed between HS4 and other 3 subclasses, with higher abundance of 4 cell populations (T cells, myeloid dendritic cells, monocytic lineage, and Fibroblasts) for HS4 compared with other 3 subclasses. In addition, HS3 exhibited lower enrichment of B lineage, CD8 T cells, T cells, and myeloid dendritic cells. There was no significant difference of cell abundance in most cell populations between HS1 and HS2 (Supplementary Figure 3C). For immune associated signatures, HS4 exhibited higher enrichment for interferon (IFN) signature than HS1 and HS2 (Supplementary Figure 3D).

The difference of other critical signatures among HCC subclasses

The associations between our HCC classification and several critical signatures involved in oncogenesis and progression of HCC were also investigated, including extracellular matrix (ECM) signature, epithelial mesenchymal transition (EMT) signature, TGF-β signature, mismatch repair signature, DNA damage repair signature, angiogenesis signature, cell cycle signature, differentiation signature, mTOR pathway signature, stem signature, and WNT activation signature (Figure 4A and 4B). Results showed that HS4 demonstrated a higher enrichment of stromal relevant signature (ECM signature and TGF-β signature), DNA repair relevant signature, cell cycle signature, mTOR signature and lower enrichment of differentiation signature compared with other 3 subclasses. HS3 exhibited lower enrichment of stem signature than other subclasses, and higher enrichment of differentiation signature than HS2. In addition, no significant difference of WNT activation signature was observed between HS2 and HS3, and both of them showed a higher enrichment of WNT activation signature than HS1 and HS4. HS1 and HS2 showed no significant difference in enrichment of ECM signature, TGF-β signature, DNA repair relevant signature, cell cycle signature, and mTOR signature. HS2 showed higher enrichment of stem signature compared with HS4. HS1 showed higher level of angiogenesis signature and differentiation signature compared with HS3.

Considering the limited evidence provided by transcriptome data, we further analyzed proteomic data to validate the conclusion. Reverse Phase Protein Array (RPPA) based proteomic data was download from The Cancer Proteome Atlas (TCPA) database. All proteins were annotated according to their corresponding genes. Because of the limited proteins detected by protein array, we only chose to investigate the difference of protein levels in PI3K/mTOR pathway, p53/Cell cycle pathway and TGF-β/Smad pathway among 4 HCC subclasses. In PI3K/mTOR pathway, HS4 exhibited higher expression of S6_pS240/S244, X4EBP1 and X4EBP1_pT70 than other 3 groups. HS3 had higher expression of AKT_pS473, Tuberin_pT1462 and P70S6K_pT389 than other groups (Figure 5A and Supplementary Figure 4A). In P53/Cell cycle pathway, HS4 had higher expression of ATM and CHK1_pS296, while HS3 had higher expression of CHK1, CHK1_pS345, P53, CDK1 and CDK1_pY15 (Figure 5B and Supplementary Figure 4B). In TGF-β/Smad pathway, HS3 had lower expression of Smad3 and higher expression of Snail than other 3 groups (Figure 5C and Supplementary Figure 4C). Significance was detected between HS1 and HS2 for the expression of AKT and AKT_pT308.

Mutations and copy number alterations associated with HCC subclasses

To investigate differences in mutations and copy number alterations among HCC subclasses, we analyzed the somatic mutation and copy number data. The mutation status of genes in p53/Cell cycle pathway,

Wnt/beta-catenin pathway, hepatic differentiation, and DNA methylation was visualized in Supplementary Figure 5A. Results of statistical analysis revealed that HS1 was associated with a low percentage of alterations in *CTNNB1* (13/84 vs 70/265 in the rest, P = 0.0402) and a high percentage of alterations in *AXIN1* (11/84 vs 17/265 in the rest, P = 0.0495). HS2 was associated with a high percentage of alterations in *AXIN1* (13/99 vs 15/250 in the rest, P = 0.0271). HS3 was associated with a low percentage of alterations in *TP53* (19/112 vs 81/237 in the rest, P = 0.0009), *MUC16* (10/112 vs 44/237 in the rest, P = 0.0201) and *AXIN1* (2/112 vs 26/237 in the rest, P = 0.0032), and a high percentage of alterations in *CTNNB1* (35/112 vs 48/237 in the rest, P = 0.0243). HS4 was associated with a low percentage of alterations in *CTNNB1* (5/54 vs 77/295 in the rest, P = 0.0174). Detailed results of the above statistical analyses were shown in Supplementary Table 5. Subsequently, mutation signatures in subclasses were investigated. First, we explored the proportion of 6 single-nucleotide substitutions (C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G, and T>G/A>C) in each HCC subclass (Supplementary Figure 5B). Then we computed sample-wise signature profiles, and filtered out mutation signatures with no prognostic significance (P > 0.15 in Cox regression). 4 mutation

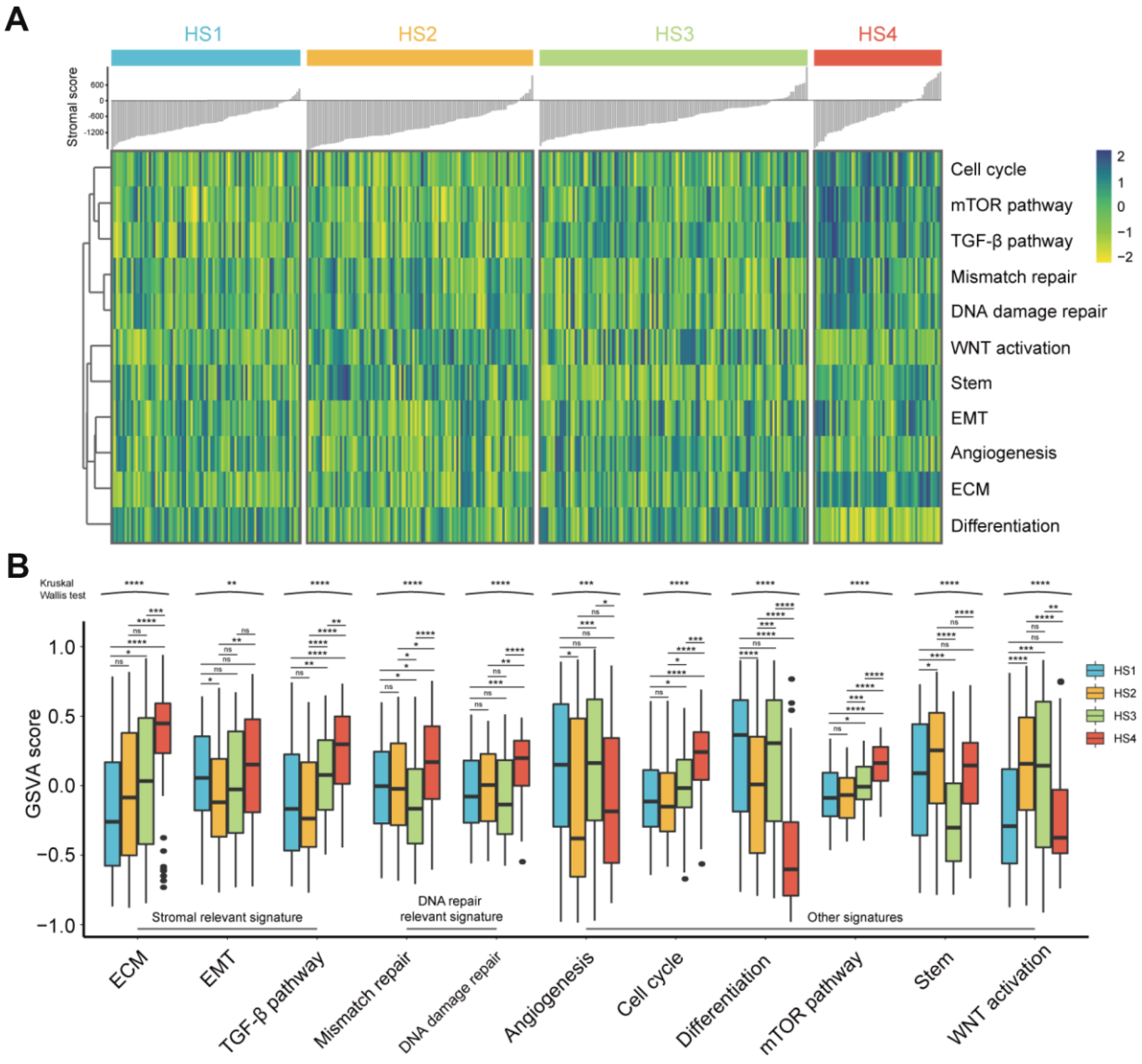


Figure 4. Difference of progression-relevant signatures among HCC subclasses. (**A**) Heatmap of progression-relevant signatures in 4 HCC subclasses. (**B**) Box plots (from 25th percentile to the 75th percentile with a line at the median) show the abundance of progression-associated signatures. Statistical significance of overall differences was determined by Kruskal Wallis test (ns represents no significance, * represents P < 0.05, ** represent P < 0.01, *** represent P < 0.001, **** represent P < 0.0001).

signatures (Signature 4, 18, 22, and 24) were remained after filtration (Supplementary Table 6), and signature weight was transformed into mutation number for comparison among groups. Significant difference of Signature 24 among 4 subclasses was observed, with more mutations of Signature 24 in HS4 than in HS3. (Supplementary Figure 5C and 5D).

Aside from point mutations and short insertions/deletions, we also analyzed DNA copy number alterations across distinct classifications based on segmentation data obtained from TCGA by using GISTIC2. Genome-wide focal amplification (red) and deletion (blue) peaks identified in different subclasses were presented in Supplementary Figure 6A. The number of specific amplification regions for HS1, 2, 3 and 4 were 18, 6, 37 and 8, respectively. The number of specific deletion regions for HS1, 2, 3 and 4 were 15, 7, 13 and 7, respectively. The common amplification regions of 4 subclasses were 5p15.33, 6q12, 11q13.3

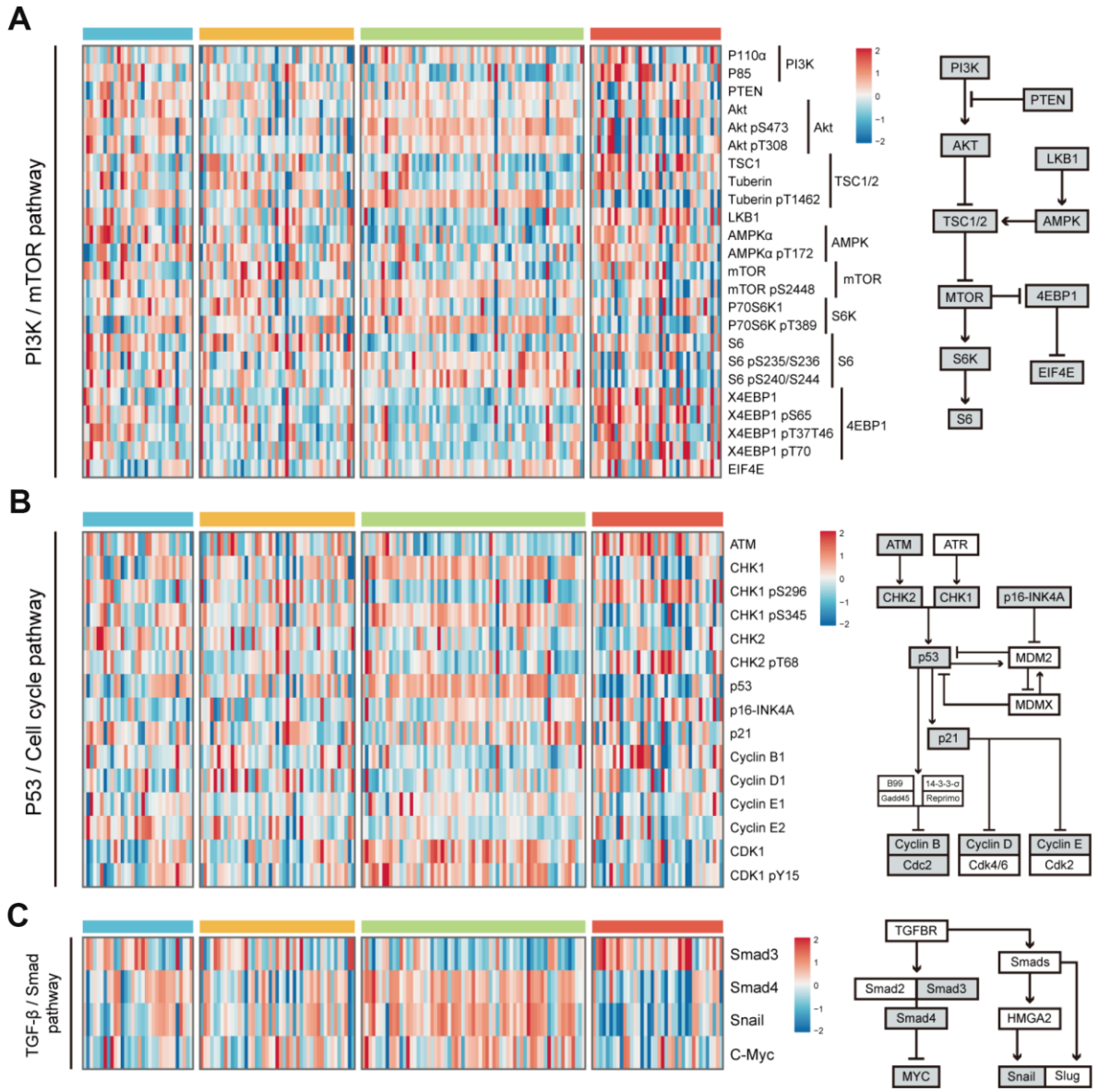


Figure 5. Difference of protein expression levels in PI3K/mTOR pathway, P53/Cell cycle pathway, and TGF-β/Smad pathway among 4 HCC subclasses. (**A**) Heatmap shows the expression level of 24 proteins in PI3K/mTOR pathway. The right half of the figure shows the basic components of PI3K/mTOR pathway. (**B**) Heatmap shows the expression level of 15 proteins in P53/Cell cycle pathway. The right half of the figure shows the basic components of P53/Cell cycle pathway. (**C**) Heatmap shows the expression level of proteins in TGF-β/Smad pathway. The right half of the figure shows the basic components of TGF-β/Smad pathway (see detailed information in Supplementary Figure 3). Data is available for proteins inside grey boxes. HCC: hepatocellular carcinoma.

and 19p13.12, while common deletion region was 9p21.3 (Supplementary Figure 6B and Supplementary Table 7).

Class prediction of HCC patients based on transcriptome data

We labeled each sample with its assigned cluster according to the HCC classification we established. A classification model was developed to investigate whether potential HCC samples can be distributed into groups identical with the training cohort based on transcriptome data of 2835 differentially expressed genes (DEGs) (Supplementary Table 8). The workflow was shown in Figure 6A. A transcriptome-based prediction model was constructed by random forest (RF) and Least Absolute Shrinkage and Selector Operation (LASSO) algorithm. The accuracy of the model in training cohort and testing cohort were 97.3% and 79.7%, respectively (Figure 6B). Then, we performed receiver operating characteristic (ROC) curves that can illustrate the relationship between TPR (sensitivity) and FPR (1-specificity) for each class. Area under the curve (AUC) close to 1 indicates that the classifier is predicting with maximum TP and minimum FP. Results of AUC for HS1, 2, 3 and 4 in training cohort were 1.000, 0.999, 0.999 and 1.000, respectively. In the testing cohort, AUC for HS1, 2, 3 and 4 were 0.950, 0.939, 0.960 and 0.980, respectively (Figure 6C).

DISCUSSION

This integrative analysis based on DNA methylation and gene expression profiles of MDGs in HCC revealed 4 subclasses with distinct features (Figure 6D). HS1 was well differentiated with the best prognosis and high percentage of *AXIN1* mutations. HS2 had high serum AFP level that was correlated with its poor outcome. High percentage of *CTNNB1* mutations corresponded with HS2's activation in WNT signaling pathway. HS3 was well differentiated with low serum AFP level and enriched in metabolism signatures, but was barely involved in immune signatures. HS3 also had high percentage of *CTNNB1* mutations and enriched in WNT activation signature. HS4 was poorly differentiated with the worst prognosis and enriched in immune-related signatures, but was barely involved in metabolism signatures. HS3 and HS4 both enriched in non-CIMP. Machine learning algorithms were applied to building a prediction model, and results showed that the model had high sensitivity and specificity in distributing potential HCC samples into distinct classifications.

The best prognosis value of HS1 was associated with high GSVA score in differentiation and lower score in WNT activation signature. The poor prognosis of HS2 was associated with higher enrichment in stemness and WNT signaling pathway activation. WNT signaling activation is mainly due to mutations in *CTNNB1*, a β-catenin gene [19]. The frequent mutations of *CTNNB1* in HS2 corresponded with its activation in WNT signaling. HS3 patients also had high percentage of alterations in *CTNNB1*. HS3 presented lower score in stemness and higher score in differentiation. On the contrary, HS4 had higher score in ECM, TGF- β pathway, mismatch repair, DNA damage repair, cell cycle, mTOR pathway and lower score in differentiation. The worst prognosis of HS4 was correlated with its involvement in the above mentioning carcinogenesis signatures.

HS2 and HS3 patients may be beneficial from therapeutic approaches that aim to target the Wnt–β-catenin pathway. For example, a small peptide called CGX1321 can inhibit Wnt lipid modifications. It is tested by a phase I trial in patients with advanced solid tumors, including HCC (https://clinicaltrials.gov/ct2/show/NCT02675946). Another phase I trial tests the efficacy of OMP-54F28, a fusion protein targeting Wnt ligands (https://clinicaltrials.gov/ct2/show/NCT02069145). DKN-01 inhibits non-canonical β-catenin pathway and is currently being investigated in a phase I trial in combination with gemcitabine and cisplatin in various cancers including HCC (https://clinicaltrials.gov/ct2/show/NCT02375880). The efficacy of these therapies for HS2 and HS3 patients requires further investigation. HS4 patients may be beneficial from therapies targeting the mTOR pathway. It has been reported that Everolimus is an mTOR inhibitor that can prevent tumor progression and improve survival in preclinical HCC models. A phase III study tested the efficacy of Everolimus in patients with advanced HCC after failure of sorafenib [20].

HS3 exhibited higher expression level of Akt involving in PI3K/mTOR pathway, indicating that HS3 patients may be beneficial from Akt inhibitors. As the key component of PI3K signaling pathway, Akt is considered to be an attractive target for cancer therapy [21]. Multiple Akt inhibitors such as ATP-competitive inhibitors (GSK690693, GDC0068, and AZD5363) and allosteric inhibitors (MK-2206) have been investigated in clinical trials against tumors. The results are promising [22]. HS4 had higher expression of ATM, a core component of the DNA repair system [23]. Targeting ATM may be a promising strategy for cancer treatment [23]. Currently, ATM inhibitors such as AZD0156 and AZD1390 are under investigation in phase I clinical trials [23]. HS4 patients may be beneficial from ATM inhibitors. HS3 had higher expression of CDK1, indicating HS3 patients may be beneficial from inhibitors targeting CDK1. BEY1107, an anti-cancer agent that selectively acts on CDK1, is in phase I/II clinical trial [24].

It appears that human cancer mutations and cancer genes constantly affect metabolism processes including aerobic glycolysis, glutaminolysis and one-carbon metabolism that produce amino acids, nucleotides, fatty acids and other substances for cell growth and proliferation [25]. Metabolic therapies targeting certain metabolism process provide alternatives for chemoresistant patients. For example, it has been

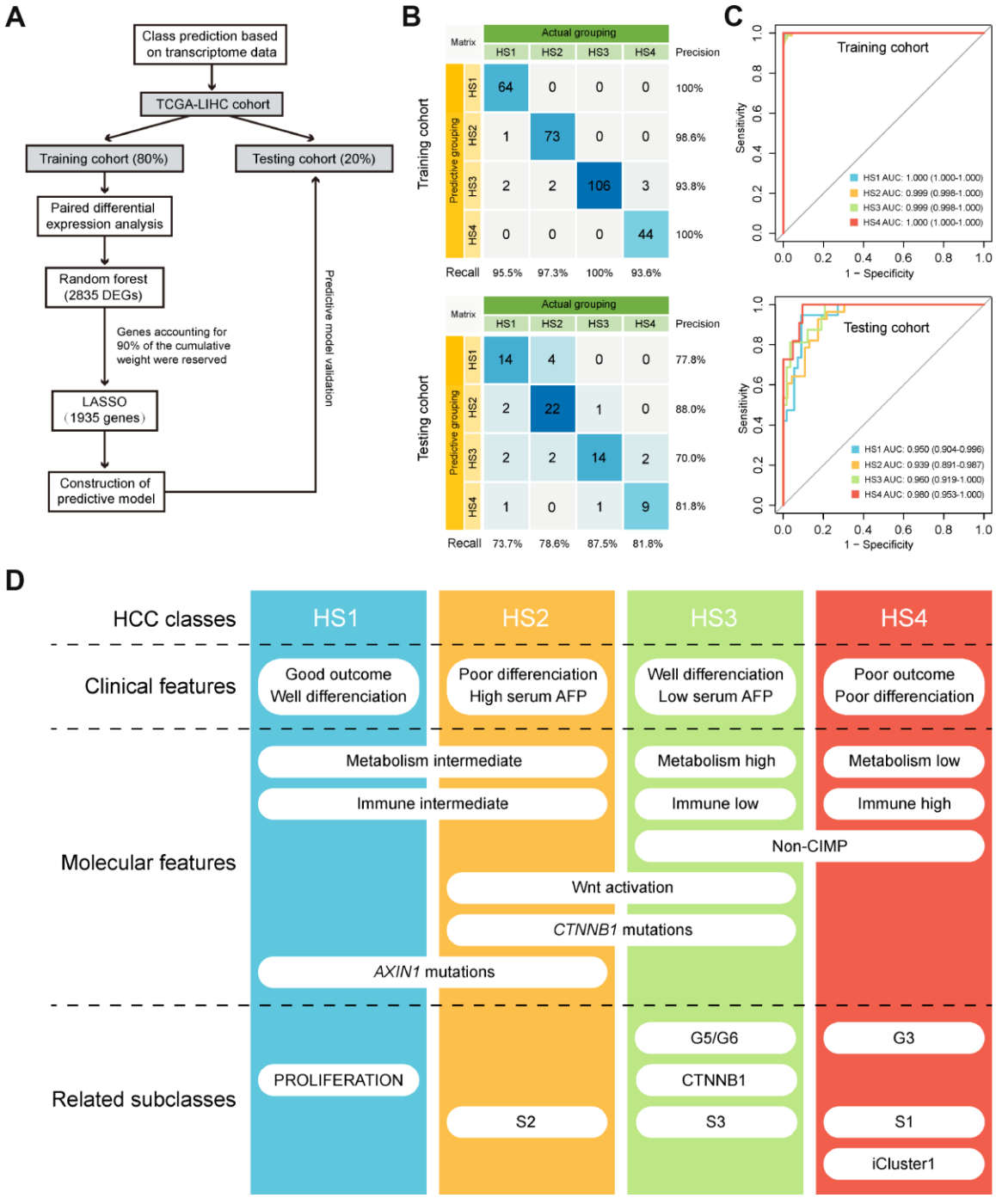


Figure 6. Class prediction of HCC patients. (**A**) Flow chart shows the process of prediction model construction. (**B**) Confusion matrix evaluations of prediction model within the training cohort and testing cohort. A perfect prediction model (100% accuracy) have 0 counts for all non-diagonal entries (that is, no misclassified samples). (**C**) ROC curves in training and testing cohort depict trade-offs between true and false positive rates as classification stringency varies. AUC values close to 1 indicate that a high true positive rate was achieved with low false positive rate, while AUC values close to 0.5 indicate random performance. (**D**) Overview of the characteristics of 4 HCC subclasses. HCC: hepatocellular carcinoma; ROC: Receiver operating characteristic; AUC: Area under the curve.

reported that metformin can prevent liver carcinogenesis [26] and treatment with metformin is associated with favorable prognosis in patients with HCC [27]. Determining the responders of metabolic therapies has proven to be challenging [28]. This study provided insights into predicting potential responders towards metabolic therapies. HS3 enriched in metabolism signatures including glucose metabolism and amino acid metabolism, indicating that HS3 patients may be beneficial from metabolic therapies like metformin. On the other hand, HS4 patients presented low enrichment in metabolic processes, suggesting that they may be non-responders towards metabolic therapies. These assumptions require further experimental validation.

In the last decades, immunotherapy has been investigated and applied in multiple tumors including HCC. Immune checkpoints play an essential role in maintaining tolerance and preventing T cell over-activation [29]. Immune checkpoint expression can lead to T cell exhaustion and immune tolerance [29]. PD-1 is expressed by activated T cells, B cells, NK cells and myeloid cells [29]. In physical conditions, when PD-L1 is expressed on antigen presenting cells (APC), the interaction between PD-L1 and PD-1 will maintain self-tolerance and prevent the activation of T cells [30]. However, tumor cells can also express PD-L1 thus inducing immune tolerance [31]. The multiplicity of infiltrating $PD1^+$ $CD8^+$ cells and the expression of PD-L1 in HCC cells have been proven to be associated with worse prognosis [32]. Other checkpoint molecules including CTLA4, TIM3 and LAG3 are also implicated in the suppression of immune response against HCC [29]. Immune checkpoint inhibitors (ICIs) can unleash cytotoxic T cells against tumors to strengthen immune response thus showing anti-tumor efficacy [33]. ICIs including CTLA-4 and PD-1/PD-L1 inhibitors have been investigated in clinical trials of HCC. Nivolumab is a PD-1 immune checkpoint inhibitor. Promising results regarding its efficiency have been achieved in a clinical trial on advanced HCC patients [34]. Pembrolizumab is another PD-1 immune checkpoint inhibitor which has been proven to be effective for advanced HCC patients who was previously treated with sorafenib [35]. In this study, HS4 showed higher expression for most of the immune checkpoint genes, while HS3 exhibited lower expression for *CD276*, *TGFB1*, *CTLA4*, *ICOS*, *PDCD1*, *TNFRSF4*, *CD274*, and *LAG3* than other subclasses. Results indicated that HS4 patients may be responders towards ICIs while HS3 patients were less likely to respond to ICIs. In addition, HS4 also exhibited higher enrichment for IFN signature than HS1 and HS2. IFNγ is one of the cytokines that can induce PD-1 expression in T cells [29], which may be associated with the highest expression of immune checkpoint genes in HS4. The worst prognosis of HS4 was associated with its high expression of immune checkpoint molecules.

CIMP is a phenomenon of simultaneous methylation in multiple genes [7]. Although the fraction of CIMP is smaller in HCC compared with other cancer types, the CIMP group still requires special attention because of its poor prognosis [7]. Based on the methylation level of 674 most variable CpGs, HCC patients were clustered into 7 groups, which was consistent with a previous study [7]. Although no significant outcome was identified, consistent with previous results [7], the overall survival time of CIMP patients was statistically shorter than that of non-CIMP patients. Several drugs that modify DNA methylation by targeting DNA methyltransferases have been investigated. For example, it has been reported that Zebularine (1-(β-(D)-ribofuranosyl)-1,2-dihydropyrimidin-2-one) inhibits DNA methylation and induces apoptosis in HepG2 cell line [36]. Another study reported that Zebularine inhibits tumor growth in xenograft models. Genes involved in apoptosis, cell cycle, and tumor suppression were demethylated in liver cancer cell lines [37].

In conclusion, this classification based on integration of DNA methylation and transcriptome profiles revealed distinct characteristics of HCC subtypes, which provided novel clinical insights into predicting both the prognosis of HCC and prospective therapies. Future research will accelerate the clinical validation of HCC classification and will promote precision diagnostics as well as therapeutics for HCC patients.

MATERIALS AND METHODS

Data preparation

Multiplatform genomics data, including mRNA expression data (raw counts), gene somatic mutation data (MAF files), DNA copy data (segment file) (March 27, 2019), DNA methylation array data (July 27, 2019), RPPA data and corresponding clinical information (August 19, 2019) of TCGA-LIHC cohort were retrieved from TCGA database (http://cancergenome.nih.gov/).

Transcripts per kilobase million (TPM) values were calculated based on raw counts. DNA methylation array data was generated from the Illumina Infinium HumanMethylation450 BeadChip array. Methylation level of each probe was represented by β value (ranging from 0 to 1). Probes containing 'NA' marked data points or located on sex chromosomes were removed. Then, probes residing in gene promoter regions including the upstream 2.5 kb from TSS, 5'UTR and first-exon regions were mapped to their corresponding

genes. Methylation level of a certain gene was determined as the average methylation level of corresponding probes residing in promoter regions.

Identification of methylation driven genes-associated classification

First, MDGs were identified based on mRNA expression data from tumor samples and methylation data from tumor and normal samples by using *MethylMix* package in R [14]. 369 HCC and 50 normal non-paired samples were used to explore differentially expressed MDGs. The MethylMix algorithm can explore different methylation level and calculate the correlation between gene expression and gene methylation level. We defined MDGs as genes with |logFC| > 0, P < 0.05 and |Cor| > 0.3. Subsequently, consensus nonnegative matrix factorization (CNMF) was applied to conduct consensus clustering based on the integrated gene expression and methylation data of MDGs by the function "ExecuteCNMF" from the R package *CancerSubtypes* [15]. T-SNE based approach was then applied to validate subtype assignments based on mRNA expression data of MDGs. Prediction of previously published HCC molecular classifications [2–4, 18] was performed by conducting nearest template prediction (NTP) analyses (Gene Pattern modules). DEGs among HCC subclasses were identified using *edgeR* package based on raw counts. Genes with an absolute log2 fold change (FC) > 1 (adjusted $P < 0.01$) were defined as DEGs [38].

Identification of CpG island methylator phenotype

To investigate the relationship between methylation driven genes associated classification and CpG island methylator phenotype, we used previously described approach [7] to identify distinct CpG island methylator phenotype of HCC. In specific, CpGs in the promoter region that have a high standard deviation (SD > 0.2) of methylation level in 369 tumor tissues and low methylation level (mean β value < 0.05) in 50 normal tissues were selected. K-means consensus clustering was performed on these CpGs using the *ConsensusClusterPlus* package in R [39].

Estimation of metabolism and immune-associated signatures

GSVA is a gene set enrichment method that can estimate the score of certain signatures based on transcriptomic data [40]. Metabolism-relevant (glucose metabolism, amino acid metabolism, lipid metabolism), immune-relevant (antigen presentation MHC class I/II, CD8 T effector, cytolytic activity, IFN), and other HCC progression (ECM, EMT, TGF-β pathway, mismatch repair, DNA damage repair, angiogenesis, cell cycle, differentiation, mTOR pathway, stem, and WNT activation) signatures were achieved from previously published studies [28, 41]. We can quantitatively measure these biological processes by *GSVA* R package. Besides, the absolute abundance of 8 immune cell populations (T cells, $CD8^+$ T cells, natural killer cells, cytotoxic lymphocytes, B cell lineage, monocytic lineage cells, myeloid dendritic cells, neutrophils) and 2 nonimmune stromal cell populations (endothelial cells and fibroblasts) was also quantified using MCP-counter algorithm [42].

Mutation signature and copy number analysis

A predefined set of 30 mutational signatures from the Wellcome Trust Sanger Institute was obtained [43]. Each signature represented a characteristic pattern of 96 possible nucleotide substitution motifs. Relative contribution of each mutational signature for tumor samples was quantified using *deconstructSigs* R package [44], and the parameters were set as following: 'exome2genome' trinucleotide-count normalization and signature cutoff at 6%. Prognosis associated mutational signatures ($P < 0.15$) were identified using Cox regression in *survival* package. Copy number variation (CNV) data was downloaded from GDAC Firehose. Then, GISTIC 2.0 (Gene Pattern modules) was used to investigate the significant amplification or deletion events in the regions of the genome [45].

Development of classification model

The full TCGA dataset (n=369) was randomly split into training and testing cohorts according to the ratio of 4:1, corresponding to 295 and 74 samples. Then, two machine learning (ML) algorithms were used to develop the classification model based on all DEGs. First, RF based variable selection method using OOB error was applied to preliminarily screen for DEGs, and genes accounting for 90% of the cumulative weight were reserved. After primary filtration, a LASSO algorithm, with penalty parameter tuning conducted by k-folds cross-validation (k=20), was used to build the final classification model. Subsequently, the LASSO based classification model was applied to the testing cohort. The predictability of the model was evaluated by confusion matrix and receiver operating characteristic (ROC) curves.

Statistical analysis

All the computational and statistical analyses were performed using R version 3.6.0 software. The difference between 2 groups was compared using unpaired Student t test (for normally distributed

variables) or Mann-Whitney U test (for non-normally distributed variables). For comparisons of 3 or more groups, one-way analysis and Kruskal-Wallis tests were used as parametric and non-parametric methods, respectively. Contingency table variables were analyzed by Chi-square test or Fisher's exact tests. Survival analysis was carried out using Kaplan Meier methods and was compared by the Log-rank test. A two-tailed p value less than 0.05 was statistically significant.

Abbreviations

AFP: α-fetoprotein; APC: antigen presenting cells; CIMP: CpG island methylator phenotype; CNMF: consensus nonnegative matrix factorization; CNV: copy number variation; DEG: differentially expressed genes; ECM: extracellular matrix; EMT: epithelial mesenchymal transition; GSVA: Gene Set Variation Analysis; HCC: hepatocellular carcinoma; HS: HCC Subclass; ICI: immune checkpoint inhibitors; IFN: interferon; LASSO: Least Absolute Shrinkage and Selector Operation; MCP-counter : microenvironment cell populations-counter; MDG: methylation-driven gene; ML: machine learning; MST: median survival time; NTP: nearest template prediction; OS: overall survival; RF: random forest; RFS: recurrence free survival; ROC: receiver operating characteristic; RPPA: Reverse Phase Protein Array; TCGA: The Cancer Genome Atlas; TPM: Transcripts per kilobase million; TSG: tumor suppressor genes.

CONFLICTS OF INTEREST

All authors declare no conflicts of interest.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (31701250).

REFERENCES

1. European Association For The Study Of The Liver, European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol. 2012; 56:908–43. https://doi.org/10.1016/j.jhep.2011.12.001 PMID:22424438

2. Boyault S, Rickman DS, de Reyniès A, Balabaud C, Rebouissou S, Jeannot E, Hérault A, Saric J, Belghiti J, Franco D, Bioulac-Sage P, Laurent-Puig P, Zucman-Rossi J. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology. 2007; 45:42–52. https://doi.org/10.1002/hep.21467 PMID:17187432

3. Hoshida Y, Nijman SM, Kobayashi M, Chan JA, Brunet JP, Chiang DY, Villanueva A, Newell P, Ikeda K, Hashimoto M, Watanabe G, Gabriel S, Friedman SL, et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. Cancer Res. 2009; 69:7385–92. https://doi.org/10.1158/0008-5472.CAN-09-1089 PMID:19723656

4. Chiang DY, Villanueva A, Hoshida Y, Peix J, Newell P, Minguez B, LeBlanc AC, Donovan DJ, Thung SN, Solé M, Tovar V, Alsinet C, Ramos AH, et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res. 2008; 68:6779–88. https://doi.org/10.1158/0008-5472.CAN-08-0742 PMID:18701503

5. Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, Durnez A, Demetris AJ, Thorgeirsson SS. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. Hepatology. 2004; 40:667–76. https://doi.org/10.1002/hep.20375 PMID:15349906

6. Shen J, Wang S, Zhang YJ, Kappil M, Wu HC, Kibriya MG, Wang Q, Jasmine F, Ahsan H, Lee PH, Yu MW, Chen CJ, Santella RM. Genome-wide DNA methylation profiles in hepatocellular carcinoma. Hepatology. 2012; 55:1799–808. https://doi.org/10.1002/hep.25569 PMID:22234943

7. Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, Wu L, Hou T, Xie L, Ding G, Li H, Li Y. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Med. 2018; 10:42. https://doi.org/10.1186/s13073-018-0548-z PMID:29848370

8. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002; 16:6–21. https://doi.org/10.1101/gad.947102 PMID:11782440

9. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002; 3:415–28. https://doi.org/10.1038/nrg816 PMID:12042769

10. Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. Cancer Res. 2001; 61:3225–29. PMID:11309270

11. Tischoff I, Tannapfe A. DNA methylation in hepatocellular carcinoma. World J Gastroenterol. 2008; 14:1741–48. https://doi.org/10.3748/wjg.14.1741 PMID:18350605

12. Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biol. 2015; 16:17. https://doi.org/10.1186/s13059-014-0579-8 PMID:25631659

13. Gao C, Zhuang J, Li H, Liu C, Zhou C, Liu L, Sun C. Exploration of methylation-driven genes for monitoring and prognosis of patients with lung adenocarcinoma. Cancer Cell Int. 2018; 18:194. https://doi.org/10.1186/s12935-018-0691-z PMID:30498398

14. Gevaert O. MethylMix: an R package for identifying DNA methylation-driven genes. Bioinformatics. 2015; 31:1839–41. https://doi.org/10.1093/bioinformatics/btv020 PMID:25609794

15. Xu T, Le TD, Liu L, Su N, Wang R, Sun B, Colaprico A, Bontempi G, Li J. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. Bioinformatics. 2017; 33:3131–33. https://doi.org/10.1093/bioinformatics/btx378 PMID:28605519

16. Li B, Cui Y, Nambiar DK, Sunwoo JB, Li R. The Immune Subtypes and Landscape of Squamous Cell Carcinoma. Clin Cancer Res. 2019; 25:3528–37. https://doi.org/10.1158/1078-0432.CCR-18-4085 PMID:30833271

17. Possemato R, Marks KM, Shaul YD, Pacold ME, Kim D, Birsoy K, Sethumadhavan S, Woo HK, Jang HG, Jha AK, Chen WW, Barrett FG, Stransky N, et al. Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature. 2011; 476:346–50. https://doi.org/10.1038/nature10350 PMID:21760589

18. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017; 169:1327–41.e23. https://doi.org/10.1016/j.cell.2017.05.046 PMID:28622513

19. Delgado E, Okabe H, Preziosi M, Russell JO, Alvarado TF, Oertel M, Nejak-Bowen KN, Zhang Y, Monga SP. Complete response of Ctnnb1-mutated tumours to β-catenin suppression by locked nucleic acid antisense in a mouse hepatocarcinogenesis model. J Hepatol. 2015; 62:380–87. https://doi.org/10.1016/j.jhep.2014.10.021 PMID:25457204

20. Zhu AX, Kudo M, Assenat E, Cattan S, Kang YK, Lim HY, Poon RT, Blanc JF, Vogel A, Chen CL, Dorval E, Peck-Radosavljevic M, Santoro A, et al. Effect of everolimus on survival in advanced hepatocellular carcinoma after failure of sorafenib: the EVOLVE-1 randomized clinical trial. JAMA. 2014; 312:57–67. https://doi.org/10.1001/jama.2014.7189 PMID:25058218

21. Fruman DA, Chiu H, Hopkins BD, Bagrodia S, Cantley LC, Abraham RT. The PI3K Pathway in Human Disease. Cell. 2017; 170:605–35. https://doi.org/10.1016/j.cell.2017.07.029 PMID:28802037

22. Revathidevi S, Munirajan AK. Akt in cancer: mediator and more. Semin Cancer Biol. 2019; 59:80–91. https://doi.org/10.1016/j.semcancer.2019.06.002 PMID:31173856

23. Jin MH, Oh DY. ATM in DNA repair in cancer. Pharmacol Ther. 2019; 203:107391. https://doi.org/10.1016/j.pharmthera.2019.07.002 PMID:31299316

24. An anti-cancer agent that selectively acts on CDK1 among the cyclin-dependent kinases (CDKs) that regulate the cell cycle. http://www.beyondbio.co.kr/eng/sub/sub3_02.php

25. Fiehn O, Showalter MR, Schaner-Tooley CE, and Reproducibility Project: Cancer Biology, and Reproducibility Project Cancer Biology. Registered report: the common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. eLife. 2016; 5:5. https://doi.org/10.7554/eLife.12626 PMID:26943899

26. Shankaraiah RC, Callegari E, Guerriero P, Rimessi A, Pinton P, Gramantieri L, Silini EM, Sabbioni S, Negrini M. Metformin prevents liver tumourigenesis by attenuating fibrosis in a transgenic mouse model of hepatocellular carcinoma. Oncogene. 2019; 38:7035–45. https://doi.org/10.1038/s41388-019-0942-z PMID:31409896

27. Schulte L, Scheiner B, Voigtländer T, Koch S, Schweitzer N, Marhenke S, Ivanyi P, Manns MP, Rodt T, Hinrichs JB, Weinmann A, Pinter M, Vogel A, Kirstein MM. Treatment with metformin is associated with a prolonged survival in patients with hepatocellular carcinoma. Liver Int. 2019; 39:714–26. https://doi.org/10.1111/liv.14048 PMID:30663219

28. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. Nat Commun. 2018; 9:5330. https://doi.org/10.1038/s41467-018-07232-8 PMID:30552315

29. Cariani E, Missale G. Immune landscape of

hepatocellular carcinoma microenvironment: Implications for prognosis and therapeutic applications. Liver Int. 2019; 39:1608–21. https://doi.org/10.1111/liv.14192 PMID:31314948

30. Hato T, Goyal L, Greten TF, Duda DG, Zhu AX. Immune checkpoint blockade in hepatocellular carcinoma: current progress and future directions. Hepatology. 2014; 60:1776–82. https://doi.org/10.1002/hep.27246 PMID:24912948

31. Shi F, Shi M, Zeng Z, Qi RZ, Liu ZW, Zhang JY, Yang YP, Tien P, Wang FS. PD-1 and PD-L1 upregulation promotes CD8(+) T-cell apoptosis and postoperative recurrence in hepatocellular carcinoma patients. Int J Cancer. 2011; 128:887–96. https://doi.org/10.1002/ijc.25397 PMID:20473887

32. Gao Q, Wang XY, Qiu SJ, Yamato I, Sho M, Nakajima Y, Zhou J, Li BZ, Shi YH, Xiao YS, Xu Y, Fan J. Overexpression of PD-L1 significantly associates with tumor aggressiveness and postoperative recurrence in human hepatocellular carcinoma. Clin Cancer Res. 2009; 15:971–9. https://doi.org/10.1158/1078-0432.CCR-08-1608 PMID:19188168

33. Buonaguro L, Mauriello A, Cavalluzzo B, Petrizzo A, Tagliamonte M. Immunotherapy in hepatocellular carcinoma. Ann Hepatol. 2019; 18:291–97. https://doi.org/10.1016/j.aohep.2019.04.003 PMID:31047849

34. El-Khoueiry AB, Sangro B, Yau T, Crocenzi TS, Kudo M, Hsu C, Kim TY, Choo SP, Trojan J, Welling TH 3rd, Meyer T, Kang YK, Yeo W, et al. Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial. Lancet. 2017; 389:2492–502. https://doi.org/10.1016/S0140-6736(17)31046-2 PMID:28434648

35. Zhu AX, Finn RS, Edeline J, Cattan S, Ogasawara S, Palmer D, Verslype C, Zagonel V, Fartoux L, Vogel A, Sarker D, Verset G, Chan SL, et al, and KEYNOTE-224 investigators. Pembrolizumab in patients with advanced hepatocellular carcinoma previously treated with sorafenib (KEYNOTE-224): a non-randomised, open-label phase 2 trial. Lancet Oncol. 2018; 19:940–52. https://doi.org/10.1016/S1470-2045(18)30351-6 PMID:29875066

36. Nakamura K, Aizawa K, Nakabayashi K, Kato N, Yamauchi J, Hata K, Tanoue A. DNA methyltransferase inhibitor zebularine inhibits human hepatic carcinoma cells proliferation and induces apoptosis. PLoS One. 2013; 8:e54036. https://doi.org/10.1371/journal.pone.0054036 PMID:23320119

37. Andersen JB, Factor VM, Marquardt JU, Raggi C, Lee YH, Seo D, Conner EA, Thorgeirsson SS. An integrated genomic and epigenomic approach predicts therapeutic response to zebularine in human liver cancer. Sci Transl Med. 2010; 2:54ra77. https://doi.org/10.1126/scitranslmed.3001338 PMID:20962331

38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID:19910308

39. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010; 26:1572–73. https://doi.org/10.1093/bioinformatics/btq170 PMID:20427518

40. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013; 14:7. https://doi.org/10.1186/1471-2105-14-7 PMID:23323831

41. Désert R, Rohart F, Canal F, Sicard M, Desille M, Renaud S, Turlin B, Bellaud P, Perret C, Clément B, Lê Cao KA, Musso O. Human hepatocellular carcinomas with a periportal phenotype have the lowest potential for early recurrence after curative resection. Hepatology. 2017; 66:1502–18. https://doi.org/10.1002/hep.29254 PMID:28498607

42. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, de Reyniès A. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016; 17:218. https://doi.org/10.1186/s13059-016-1070-5 PMID:27765066

43. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, et al, and Australian Pancreatic Cancer Genome Initiative, and ICGC Breast Cancer Consortium, and ICGC MMML-Seq Consortium, and ICGC PedBrain. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. https://doi.org/10.1038/nature12477 PMID:23945592

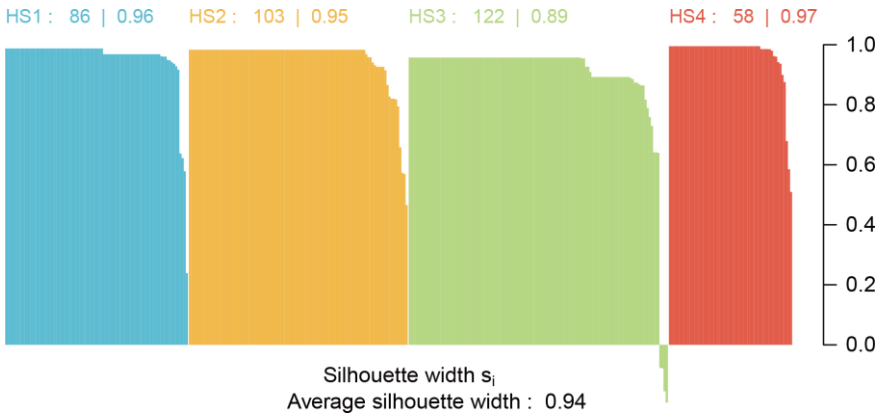
44. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair

deficiencies and patterns of carcinoma evolution. Genome Biol. 2016; 17:31.
https://doi.org/10.1186/s13059-016-0893-4
PMID:26899170

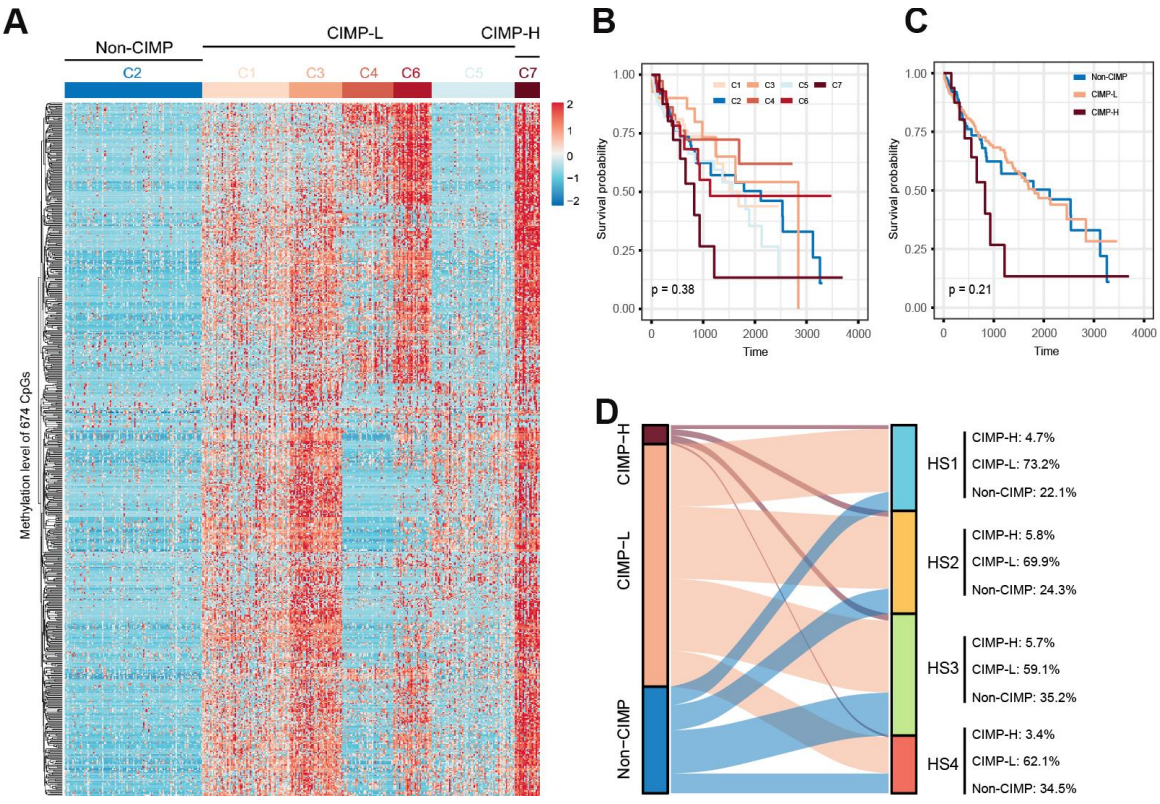
45. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011; 12:R41.
https://doi.org/10.1186/gb-2011-12-4-r41
PMID:21527027

SUPPLEMENTARY MATERIALS

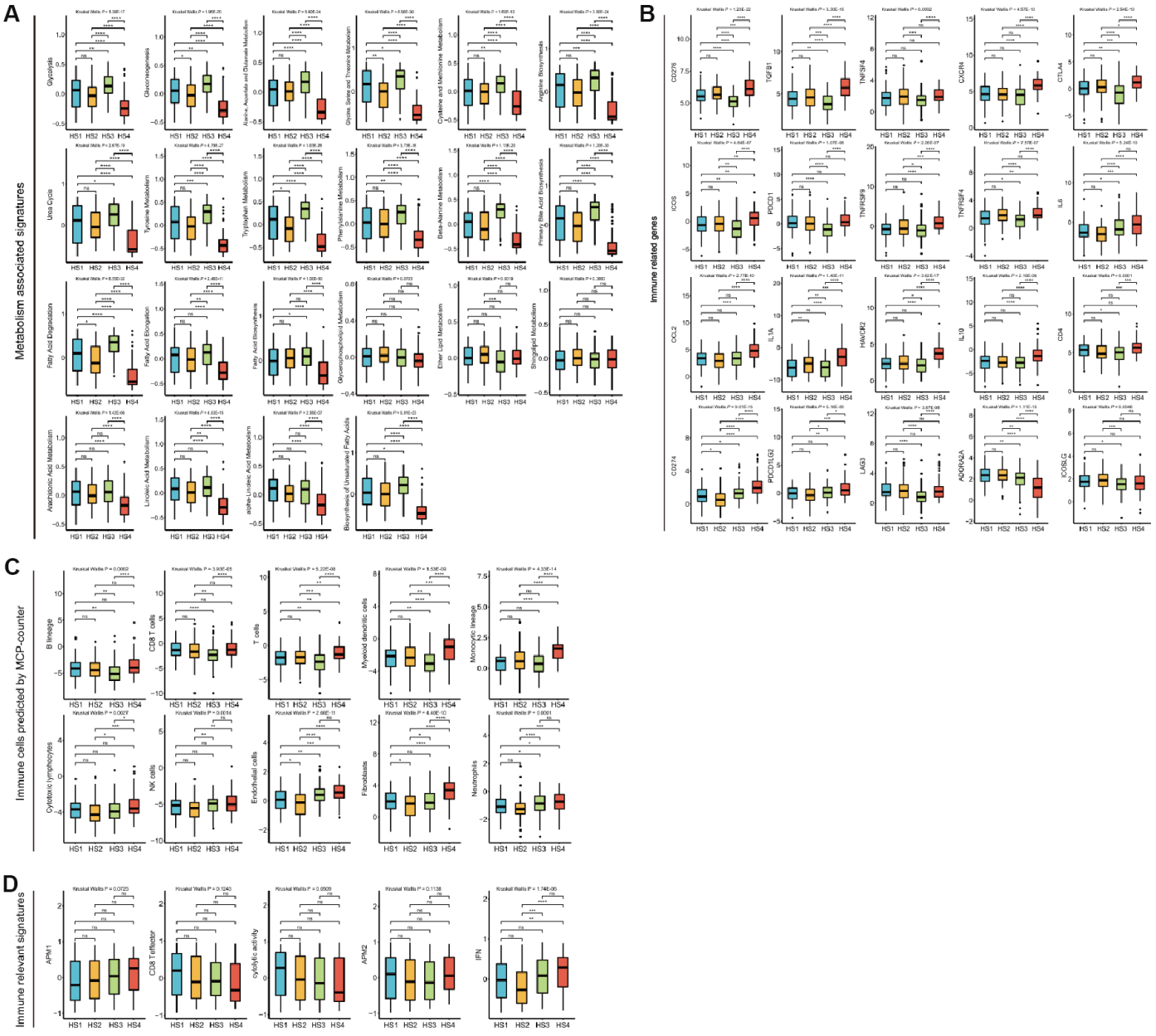
Supplementary Figures



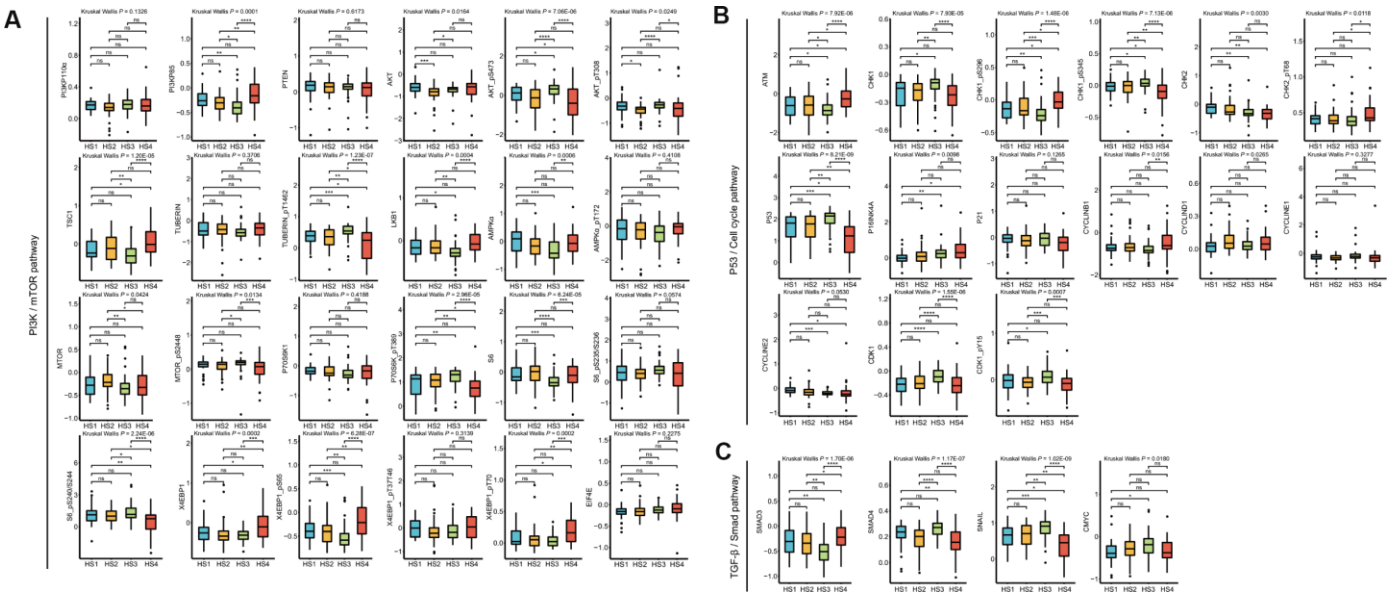
Supplementary Figure 1. Silhouette plot for k = 4 classes.



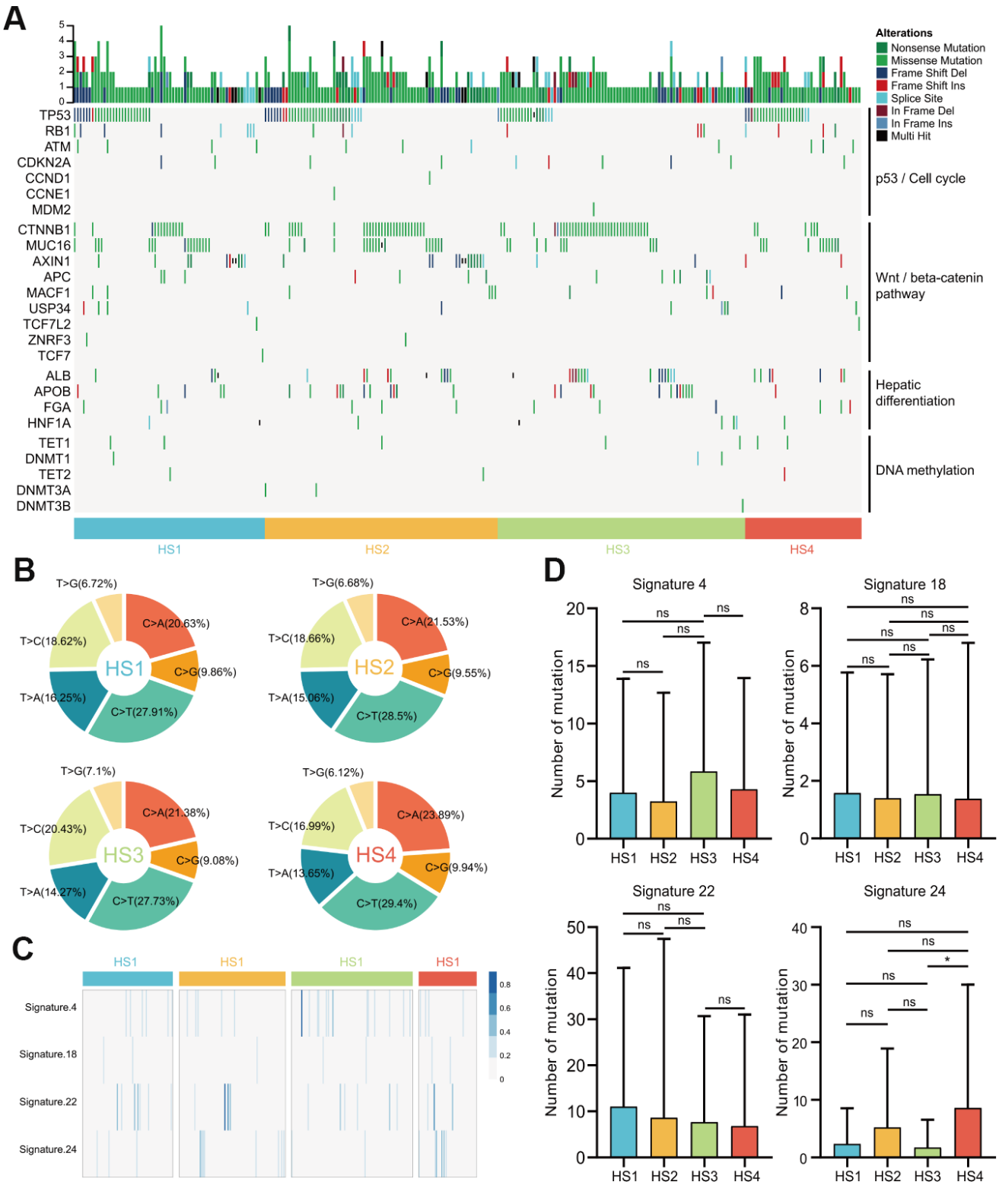
Supplementary Figure 2. Correlation of the HCC subclasses with methylation clusters. (**A**) 7 methylation clusters were obtained using k-means consensus clustering. These clusters were then divided into 3 groups, namely non-CIMP, CIMP-H and CIMP-L. (**B**) Kaplan-Meier survival curves of 7 methylation clusters. Statistical significance of differences was determined by Log-rank test. (**C**) Kaplan-Meier survival curves of 3 methylation groups. (**D**) Sankey plot shows that HS3 and HS4 are associated with non-CIMP group. Statistical significance of differences was determined by Chi-square test. HCC: hepatocellular carcinoma; CIMP: CpG island methylator phenotype.



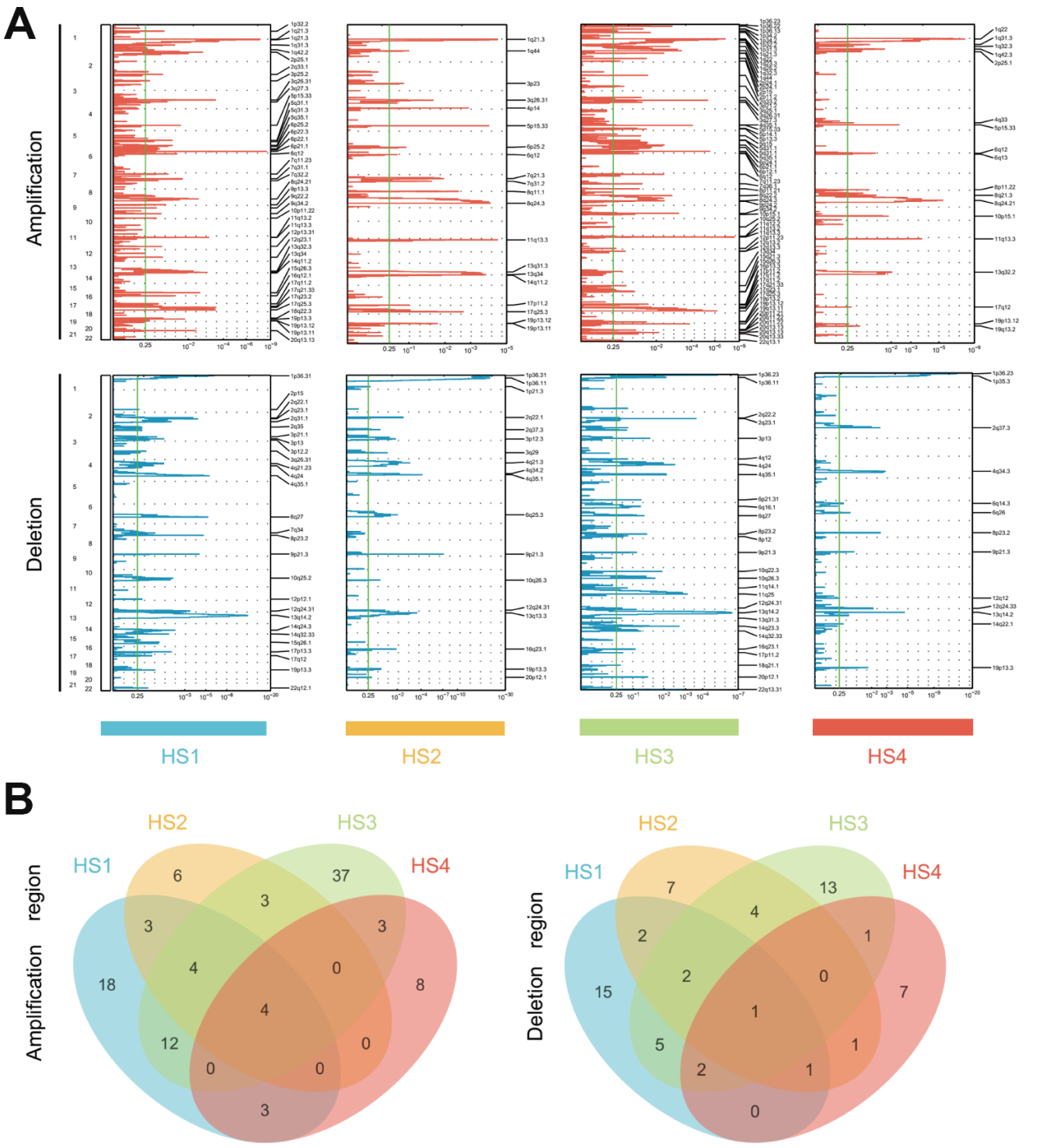
Supplementary Figure 3. Difference of abundance in metabolism and immune-associated signatures among HCC subclasses. (**A**) Box plots (from 25th percentile to the 75th percentile with a line at the median) show the range of abundance in metabolism associated signatures. (**B**) Box plots show the expression level of immune related genes. (**C**) Box plots show the abundance of immune and stromal cell populations. (**D**) Box plots show the abundance of immune-relevant signatures.



Supplementary Figure 4. Association between HCC classifications and expression level of proteins in 3 pathways. (**A**) Box plots (from 25th percentile to the 75th percentile with a line at the median) show the expression level of proteins in PI3K/mTOR pathway. (**B**) Box plots show the expression level of proteins in P53/Cell cycle pathway. (**C**) Box plots show the expression level of proteins in TGF-β/Smad pathway. HCC: hepatocellular carcinoma.



Supplementary Figure 5. Association between HCC classifications and somatic mutation alterations. (**A**) Oncoprint showing mutation status of genes in P53/Cell cycle pathway, Wnt/beta-catenin pathway, hepatic differentiation and DNA methylation (see detailed statistical analysis in Supplementary Table 3). (**B**) Proportion of 6 different single-nucleotide substitutions in HCC classifications are shown. (**C**) Heatmap shows signature weight of 4 prognosis-associated signatures among 4 subclasses. (**D**) Histograms show the difference of mutation number in 4 signatures among HCC subclasses. HCC: hepatocellular carcinoma.



Supplementary Figure 6. Association between HCC classifications and DNA copy number alterations. (**A**) Genome-wide focal amplification (red) and deletion (blue) peaks in 4 HCC subclasses identified by GISTIC2.0. (**B**) Venn diagrams identify the specific/common significant amplification and deletion regions in different HCC subclasses. HCC: hepatocellular carcinoma.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 1.

Supplementary Table 1. Results of MethylMix analysis.

Supplementary Table 2. Detailed results from statistical analysis of overall survival and recurrence free survival.

Results of survival analysis (OS) of distinct HCC classification					
	Median survival time	**95%CI**	**p value (vs HS2)**	**p value (vs HS3)**	**p value (vs HS4)**
HS1	2839	1749-3929	0.0609	0.5308	0.0034
HS2	1622	929-2315	/	0.0786	0.1821
HS3	1818	1213-2423	/	/	0.0048
HS4	1135	450-1820	/	/	/
Results of survival analysis (RFS) of distinct HCC classification					
	Median survival time	**95%CI**	**p value (vs HS2)**	**p value (vs HS3)**	**p value (vs HS4)**
HS1	1453	806-2100	0.0723	0.6013	0.0911
HS2	828	396-1260	/	0.1228	0.907
HS3	893	587-1199	/	/	0.1196
HS4	489	198-780	/	/	/

Please browse Full Text version to see the data of Supplementary Table 3.

Supplementary Table 3. List of 401 methylation driven genes.

Supplementary Table 4. Clinical Characteristics of TCGA-LIHC cohort.

Clinical Characteristics of TCGA cohort	
Variable	**TCGA set (n=369)**
Age	
>55	241
<=55	122
Gender	
female	118
male	245
Viral infection	
HBV	95
HCV	49
HBV and HCV	7
No infection	102
Child-Pugh score	
A	216
B/C	22
Histologic grade	
G1	55
G2	175
G3	116
G4	12
TNM stage	
I/II	254
III/IV	85
AFP level	
Low	181
High	95
Vascular invasion	
None	205
Micro	90
Macro	14
Family history	
No	204
Yes	110

Supplementary Table 5. Detailed results from statistical analysis of mutation characteristics in HCC classifications. HCC: hepatocellular carcinoma.

Mutation characteristics in distinct HCC classification												
	HS1	**percentage (%)**	**p-value**	**HS2**	**percentage (%)**	**p-value**	**HS3**	**percentage (%)**	**p-value**	**HS4**	**percentage (%)**	**p-value**
Number of patients	84	100.00		99	100.00		112	100.00		54	100.00	
TP53	26	30.95	ns	33	33.33	ns	19	16.96	0.0009	22	40.74	ns
CTNNB1	13	15.48	0.0402	29	29.29	ns	35	31.25	0.0243	6	11.11	0.0174
MUC16	16	19.05	ns	17	17.17	ns	10	8.93	0.0201	11	20.37	ns
ALB	4	4.76	ns	10	10.10	ns	15	13.39	ns	7	12.96	ns
APOB	5	5.95	ns	11	11.11	ns	12	10.71	ns	1	1.85	ns
AXIN1	11	13.10	0.0495	13	13.13	0.0271	2	1.79	0.0032	2	3.70	ns
RB1	8	9.52	ns	3	3.03	ns	4	3.57	ns	4	7.41	ns
APC	3	3.57	ns	4	4.04	ns	6	5.36	ns	0	0.00	ns
MACF1	2	2.38	ns	4	4.04	ns	3	2.68	ns	2	3.70	ns
FGA	3	3.57	ns	3	3.03	ns	3	2.68	ns	3	5.56	ns
ATM	4	4.76	ns	3	3.03	ns	1	0.89	ns	3	5.56	ns
USP34	4	4.76	ns	1	1.01	ns	4	3.57	ns	1	1.85	ns
CDKN2A	2	2.38	ns	2	2.02	ns	5	4.46	ns	1	1.85	ns
HNF1A	2	2.38	ns	1	1.01	ns	4	3.57	ns	1	1.85	ns
TET1	2	2.38	ns	1	1.01	ns	2	1.79	ns	2	3.70	ns
DNMT1	1	1.19	ns	0	0.00	ns	2	1.79	ns	0	0.00	ns
DNMT3A	0	0.00	ns	2	2.02	ns	0	0.00	ns	0	0.00	ns
TET2	1	1.19	ns	1	1.01	ns	0	0.00	ns	1	1.85	ns
TCF7	1	1.19	ns	0	0.00	ns	0	0.00	ns	0	0.00	ns
TCF7L2	1	1.19	ns	0	0.00	ns	0	0.00	ns	1	1.85	ns
ZNRF3	1	1.19	ns	1	1.01	ns	0	0.00	ns	0	0.00	ns
CCND1	0	0.00	ns	1	1.01	ns	0	0.00	ns	0	0.00	ns
CCNE1	0	0.00	ns	1	1.01	ns	0	0.00	ns	0	0.00	ns
DNMT3B	0	0.00	ns	0	0.00	ns	1	0.89	ns	0	0.00	ns
MDM2	0	0.00	ns	0	0.00	ns	1	0.89	ns	0	0.00	ns

Supplementary Table 6. Results from cox regression of mutation signature.

The results of cox regression of mutation signature		
	HR	**p value**
Signature.1	1.011480944	0.638463658
Signature.2	1.002796051	0.947780338
Signature.3	1.000279303	0.959690395
Signature.4	1.012134008	0.13054495
Signature.5	0.993570767	0.225181567
Signature.6	0.959881321	0.373625593
Signature.7	0.973549374	0.514468948
Signature.8	1.002630783	0.734655486
Signature.9	0.986273393	0.25153733
Signature.10	0.943404048	0.409932352
Signature.11	0.975163806	0.524308952
Signature.12	0.986674705	0.277786476
Signature.13	1.016710568	0.649605185
Signature.14	1.034850808	0.471043598
Signature.15	1.040357259	0.190478444
Signature.16	1.000884331	0.616969517
Signature.17	1.052131131	0.228968673
Signature.18	1.030359459	0.096865864
Signature.19	0.99322056	0.764671586
Signature.20	1.03315935	0.279846514
Signature.21	1.010874489	0.555240576
Signature.22	1.005884371	0.020551777
Signature.23	0.986966228	0.754838115
Signature.24	1.019837628	0.0000821
Signature.25	0.992860584	0.616339237
Signature.26	1.002719008	0.810417508
Signature.27	1.01248891	0.669391861
Signature.28	0.985549785	0.443335344
Signature.29	1.004175822	0.825991515
Signature.30	1.019595427	0.177302858

Please browse Full Text version to see the data of Supplementary Tables 7, 8

Supplementary Table 7. Detailed information of specific amplification and deletion regions in HCC classifications. HCC: hepatocellular carcinoma.

Supplementary Table 8. Information of 2835 genes for machine learning.