Deep Learning based multi-omics integration robustly predicts survival in liver cancer

Kumardeep Chaudhary[1$], Olivier Poirion[1$], Liangqun Lu[1,2], Lana X. Garmire[1,2*]

[1] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA.

[2] Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA.

[$] These authors contributed equally to the work.

[*] To whom correspondence should be addressed. Email address: lgarmire@cc.hawaii.edu

Abstract

Identifying robust survival subgroups of hepatocellular carcinoma (HCC) will significantly improve patient care. Currently, endeavor of integrating multi-omics data to explicitly predict HCC survival from multiple patient cohorts is lacking. To fill in this gap, we present a deep learning (DL) based model on HCC that robustly differentiates survival subpopulations of patients in six cohorts. We train the DL based, survival-sensitive model on 360 HCC patient data using RNA-seq, miRNA-seq and methylation data from TCGA. This model provides two optimal subgroups of patients with significant survival differences (P=7.13e-6) and good model fitness (C-index=0.68). More aggressive subtype is associated with frequent *TP53* inactivation mutations, higher expression of stemness markers (*KRT19*, *EPCAM*) and tumor marker *BIRC5*, and activated Wnt and Akt signaling pathways. We validated this multi-omics model on five external datasets of various omics types: LIRI-JP cohort (n=230, c-index=0.75), NCI cohort (n=221, c-index=0.67), Chinese cohort (n=166, c-index=0.69), E-TABM-36 cohort (n=40, c-index=0.77), and Hawaiian cohort (n=27, c-index=0.82). This is the first study to employ deep learning to identify multi-omics features linked to the differential survival of HCC patients. Given its robustness over multiple cohorts, we expect this model to be clinically useful for HCC prognosis prediction.

Keywords:
Hepatocellular carcinoma, Multi-omics, Deep learning, Survival

Introduction

Hepatocellular carcinoma (HCC) is the most prevalent type (70-90%) of liver cancer, and 2nd leading cancer responsible for the mortality in men[1]. In USA, it has the 2nd highest incidence rate and highest mortality rate[2]. HCC is aggravated by various risk factors, including HBV/HCV infection, nonalcoholic steatohepatitis (NASH), alcoholism, and smoking. These confounding factors along with high level of heterogeneity have rendered HCC prognosis a much challenging task[3,4]. HCC is a detrimental disease with poor prognosis in general, where median survival is less than 2 years[5]. In particular, 5-year survival rate of HBV-associated HCC is less than 30% in multiple studies[5-8]. Treatment strategies in HCC are very limited, imposing additional urgent needs for developing tools to predict patient survival[9].

To understand the HCC heterogeneity among patients, a considerable amount of work has been done to identify the HCC molecular subtypes[10-16]. A variety of numbers of subtypes were identified, ranging from 2 to 6, based on various omics data types, driving hypotheses and computational methods. Besides most commonly used mRNA gene expression data, a recent study integrated copy number variation (CNV), DNA methylation, mRNA and miRNA expression to identify the 5 HCC molecular subtypes from 256 TCGA samples[17]. However, most of these studies explored the molecular subtypes without relying on survival during the process of defining subtypes. Rather, survival information was used *post hoc* to evaluate the clinical significance of these subtypes[17]. As a result, some molecular subtypes showed converging and similar survival profile, making them redundant subtypes in terms of survival differences[13]. New approaches to discover survival-sensitive and multi-omics data based molecular subtypes are much needed in HCC research.

To address these issues, for the first time, we have utilized deep learning (DL) computational framework on multi-omics HCC data sets. We chose autoencoder framework as the implementation of DL for multi-omics integration. Autoencoders have already been proved to be efficient approaches to produce features linked to clinical outcomes[18]. And it was successfully applied to analyze high-dimensional gene expression data[19,20], and to integrate heterogeneous data[21,22]. Notably, autoencoder transformation tends to aggregate genes sharing similar pathways[23], therefore making it appealing to interpret the biological functions. The contributions of this study to HCC field is not only manifested in its thorough and integrative computational rigor, but also unify the discordant molecular subtypes into robust subtypes that withstand the testing of various cohorts, even when they are in different omics forms.

We derived the model from 360 HCC samples in TCGA multi-omics cohort, which have mRNA expression, miRNA expression, CpG methylation and clinical information. We discovered two subtypes with significant differences in survival. These subtypes hold independent predictive values on patient survival, apart from clinical characteristics. Most importantly, the two subtypes obtained from our DL framework are successfully validated in five independent cohorts, which have miRNA or mRNA or DNA methylation results. Functional analysis of these two subtypes identified that gene expression signatures (*KIRT19*, *EPCAM* and *BIRC5*) and Wnt signaling pathways are highly associated with poor survival. In summary, the survival-sensitive subtypes model reported here is significant for both HCC prognosis prediction and therapeutic intervention.

Results

Two differential survival subtypes are identified in TCGA multi-omics HCC data

From the TCGA HCC project, we obtained 360 tumor samples that had coupled RNA-seq, miRNA-seq and DNA methylation data. For these 360 samples, we pre-processed the data as described in the ‘Materials and Methods’ section, and obtained 15,629 genes from RNA-seq, 365 miRNAs from miRNA-seq, and 19,883 genes from DNA methylation data as input features. These three types of omics features were stacked together using autoencoder, a deep learning framework[24]. The architecture of autoencoder is shown in Figure 1A. It has 5 layers of nodes: an input layer, a hidden layer for encoding, a bottleneck layer, a hidden layer for decoding, and an output layer. We used *tanh* function for activation and retrieved 100 new transformed features from the bottleneck layer. We then conducted univariate Cox-PH regression on each of the 100 features, and identified 37 features significantly (log-rank p-value <0.05) associated with survival. These 37 features were subjective to K-means clustering, with cluster number K ranging from 2 to 6 (Figure 1B). Using silhouette index and the Calinski-Harabasz criterion, we found that K=2 was the optimum with the best scores for both metrics (Supplementary Fig. 1A). Further, the survival analysis on the full TCGA HCC data shows that the survivals in the two sub-clusters are drastically different (log-rank p-value =7.13e-6, Figure 2A). Moreover, K=2 to 6 yielded

KM survival curves that essentially represent 2 significantly different survival groups (Supplementary Fig. 1B). Thus, we determined that K=2 was the classification labels for the subsequent supervised machine learning processes.

We next used the 2 classes determined above as the labels to build a classification model using the support vector machine (SVM) algorithm with cross-validation (CV) (Fig. 1B). We split the 360 TCGA samples into 10 folds using 60/40 ratio for training and testing data. We chose 60/40 split, rather than a conventional 90/10 split, in order to have sufficient testing samples for sensible log-rank p-values in the survival analysis (see 'Materials and Methods'). Additionally, we assessed the accuracy of the survival subtype predictions using C-index, which measures the fraction of all pairs of individuals whose predicted survival times are ordered correctly[25]. We also calculated the error of the model fitting on survival data using Brier score[26]. On average, the training data generated high C-index (0.70±0.04), low brier score (0.19±0.01), and significant average log-rank p-value (0.001) on survival difference (Table 1). Similar trend was observed for the 3-omics held-out testing data, with C-index=0.69±0.08, Brier score=0.20±0.02, and average survival p-value=0.005 (Table 1). When tested on each single omic layer of data, this multi-omics model also has decent performances, in terms of C-index, low Brier scores and log-rank p-values (Table 1). These results demonstrate that the classification model using cluster labels is robust to predict survival-specific clusters.

The performance of the model described in Fig. 1B is superior to the alternative model, where autoencoder is replaced by traditional dimension reduction approach namely Principal Component Analysis (PCA). The PCA approach failed to give significant log-rank p-value (α=0.05) in survival subgroups. It also yielded significantly lower C-indices for both the training (0.63, p-value<0.01) and testing (0.62, p-value < 0.05) data (Supplementary Table 1**)**, as compared to the model using autoencoder. Worth noticing, the 3-omics based DL model gives better prediction metrics in CV, when compared to single-omics based DL models (Supplementary Table 2), suggesting that indeed multi-omics data are better than single-omics data for model building.

The survival subtypes are robustly validated in five independent cohorts

To demonstrate the robustness of the classification model at predicting survival outcomes, we validated the model on a variety of five independent cohorts, each of which had only mRNA, or miRNA or methylation omics data (Table 2 and Fig. 2 B-F). LIRI-JP dataset is the RNA-seq data set with the most number of patients (n=230); we achieved a good C-index 0.75, a low Brier error rate of 0.16 and the log-rank p-values of 4.4e-4 between the two subtypes. For the second largest (n=221) NCI cohort (GSE14520), the two subgroups have decent C-index of 0.67 and low Brier error rate of 0.18 with log-rank p-value of 1.05e-3 (Table 2). For Chinese cohort (GSE31384), the miRNA array data with 166 samples, the two subgroups have C-index of 0.69, low Brier error rate of 0.21, and log-rank p-value of 8.49e-4 (Table 2). Impressively, the C-indices for the two smallest cohorts, E-TABM-36 (40 samples) and Hawaiian cohorts (27 samples) are very good, with values of 0.77 and 0.82, respectively. The p-values obtained for the small cohorts after resampling are also significant, with values of 3.06e-4 and 1.19e-8, respectively (Fig. 2B-F).

Adding clinical information does not improve DL-based multi-omics model

It remains to see if the DL based multi-omics model will improve the predictability, by adding clinical information. Therefore, we assessed the performance of alternative models with clinical variables as the features, either alone or in combination with previous DL-based multi-omics model (Table 3). When clinical features were used as the sole feature set for survival prediction, the models' performances were much poorer (Table 3), when compared to the DL-based genomic model (Table 2). Then we combined the clinical features with the 3 omics layers before the k-means clustering step in Fig. 1B**.** Surprisingly, the C-indices of the combined model were not better on the validation cohorts with larger sample sizes (LIRI-JP, NCI and E-TABM-36 cohorts), compared to those of DL-based multi-omics model. C-index and p-value were only slightly but not statistically significantly better for the Hawaiian cohort, which has only 27 samples. We thus conclude that the DL-based multi-omics model performs sufficiently well even without clinical features. We speculate the reason is due to the unique advantage of DL neural network, which can capture the redundant contributions of clinical features through their correlated genomic features.

Associations of survival-subgroups with clinical covariates

We performed the Fisher's exact test between the two survival subgroups and the clinical variables from TCGA cohort, and found that only grade (P=0.0004) and stage (P=0.002) were significantly associated with survival, as expected. Since HCC is aggravated by the multiple risk factors including HBC, HCV, and alcohol, we also tested our model within subpopulations stratified by individual risk factors (Table 4). Impressively, our model performed very well on all the risk factor categories with C-indices ranging from 0.69-0.79, and Brier scores between 0.19 and 0.20. Log-rank P-values were significant in HBV infected patients (P=0.04), alcohol consumers (P=0.005) and other category (P=0.0035). The only non-significant p-value (P=0.20) was obtained from the HCV infected patients, probably attributed to the small group size (n=31).

TP53 is one of the most frequently mutated genes in HCC, and its inactivation mutations have been reported to be associated with poor survival in the HCC[27]. Between the 2 survival subgroups S1 and S2 in TCGA samples, TP53 is more frequently mutated in the aggressive subtype S1 (Fisher's test p-value=0.042). Further, TP53 inactivation mutations are associated with the aggressive subtype S1 in LIRI-JP cohort, where whole genome sequencing data are available (p-value=0.024).

Functional analysis of the survival-subgroups in TCGA HCC samples

We used DESeq2 package[28] for differential gene expression between the two identified subtypes. After applying the filter of log2 fold change >1 and FDR <0.05, we obtained 820 up-regulated and 530 down-regulated genes in the aggressive sub-cluster S1. Fig. 3 shows the comparative expression profile of these 1350 genes after normalization. The up-regulated genes in the S1 cluster include the stemness marker gene, *EPCAM* (P=5.7e-6), *KRT19* (P=6.7e-15) and tumor marker *BIRC5* (P=1.2e-13) genes, which were also reported earlier to be associated with aggressive HCC subtype[29-31]. Additionally, 18 genes (*ADH1B*, *ALDOA*, *APOC3*, *CYP4F12*, *EPHX2*, *KHK*, *PFKFB3*, *PKLR*, *PLG*, *RGN*, *RGS2*, *RNASE4*, *SERPINC1*, *SLC22A7*, *SLC2A2*, *SPHK1*, *SULT2A1*, *TM4SF1*) differentially expressed in the two subtypes have similar trends of expression as in the previous study, where a panel of 65-gene signature was associated with the HCC survival[32].

Using the differentially expressed genes above, we conducted KEGG pathway analysis to pinpoint the pathways enriched in two subtypes. These subtypes have different and (almost) disjoint active pathways, confirming that they are distinct subgroups at the pathway level (Fig. 4**).** Aggressive subtype S1 is enriched with cancer related pathways, Wnt signaling pathway, PI3K-Akt signaling pathway etc. (Fig. 4A). Wnt signaling pathway was reported being associated with aggressive HCC previously[33]. In contrast, the moderate subtype S2 has activated metabolism related pathways including drug metabolism, amino acid and fatty acid metabolism etc. (Fig. 4B**)**. We performed similar differential analysis for miRNA expression and methylation data, and detected 23 miRNAs and 55 genes' methylation statistically different between the two subgroups (Supplementary Fig. 2 and File 1).

Discussion

Heterogeneity is one of the bottlenecks for understanding the HCC etiology. Though there are many studies for subtype identification of the HCC patients, embedding survival outcome of the patients as part of the procedure of identified subtypes has not been reported before. Moreover, most reported HCC subtype models have either no or very few external validation cohorts. This calls for better strategies, where the identified subtypes could reflect the phenotypic outcome of the patients i.e. the survival directly. Present work includes the integration of the multi-omics data from the same patients, giving an edge by exploiting the improved signal-to-noise ratio. To our knowledge, we are the first to use the deep learning framework to integrate multi-omics information in HCC. It propels deep learning to develop risk stratification model, not only for prognostication but also instrumental for improvising risk-adapted therapy in HCC.

We have identified two subtypes from the molecular level. This model is robust and perhaps more superior than other approaches, manifested in several levels. First, CV results gave the consistent performance in TCGA HCC testing samples, implying the reliability and robustness of the model. Secondly, deep-learning technique used in the model has captured sufficient variations due to potential clinical confounders, such that it performs as accurately or even better than, having additional clinical features in the model. Thirdly, autoencoder framework has much more efficiency to infer features linked to survival, compared to PCA. Lastly and most importantly, this model is repetitively validated in five additional cohorts, ranging from RNA-seq, mRNA microarray, miRNA array, and DNA methylation platforms.

In association with clinical characteristics, the more aggressive subtype (S1) has consistent trends of association with higher TP53 inactivation mutation frequencies in the TCGA and LIRI-JP cohorts, which is in concordance with the previous study[27]. Association of stemness markers (*KRT19*, *EPCAM*) with S1 subtype is also in congruence with the literature[29,30]. Moreover, S1 subtype is enriched with activated Wnt signaling pathway[33]. Despite our effort, the one to one comparison with the previous studies is not feasible due to the absence of cluster label information in original reports, and lack of survival data in some cases. Fortunately, we were able to identify five external validation cohorts encompassing different omic dataset, and succeeded in validating the subtypes among them. These results gave enough confidence that the 2 survival subtype model proposed in this report is of direct clinical importance, and maybe useful to improve HCC patients survival.

Methods

Datasets

TCGA Training dataset: We trained the model on multi-omics HCC data from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/). We used R package TCGA-assembler (v1.0.3)[34] and obtained 360 samples with RNA-seq data (UNC IlluminaHiSeq_RNASeqV2; Level 3), miRNA-seq data (BCGSC IlluminaHiSeq_miRNASeq; Level 3), DNA methylation data (JHU-USC HumanMethylation450; Level 3), and the clinical information. For the DNA methylation, we mapped CpG islands within 1500 bp of transcription start sites (TSS) of genes and averaged their methylation values. In dealing with the missing values (preprocessing of data), three steps were performed as elsewhere[35]. First, the biological features (e.g. genes) were removed if having zero value in more than 20% of patients. The samples were removed if missing across more than 20% features. Then we used *impute* function from R impute package[36], to fill out the missing values. Lastly, we removed genes with zero values across all samples.

Validation dataset 1 (LIRI-JP cohort, RNA-seq): 230 samples with RNA-seq data were obtained from ICGC portal (https://dcc.icgc.org/projects/LIRI-JP). These samples belong to Japanese population primarily infected with HBV/HCV[37]. We used the normalized read count values given in the gene expression file.

Validation dataset 2 (NCI cohort, GSE14520-microarray gene expression): 221 samples with survival information were chosen from GSE14520 Affymetrix high-throughput GeneChip HG-U133A microarray dataset, from an earlier study of HCC patients[38]. This is a Chinese population primarily associated with HBV infection. Log2 Robust Multi-array Average (RMA)-calculated signal intensity values provided by the authors were used for analysis.

Validation dataset 3 (Chinese cohort, GSE31384-miRNA expression): 166 pairs of HCC/matched noncancerous normal tissue samples were downloaded, with CapitalBio custom Human miRNA array data (GSE31384)[39]. Since the data were already log2 transformed, we used unit-scale normalization**.**

Validation dataset 4 (E-TABM-36-microarray gene expression): 40 HCC samples were used, with survival information and transcriptional profiling from Affymetrix HG-U133A GeneChips arrays platform[13]. We used the CHPSignal values for the further processing as a measure of gene expression.

Validation dataset 5 (DNA Methylation): 27 samples were used, with genome-wide methylation profiling from Illumina HumanMethylation450 BeadChip platform[40]. Probe to gene conversion was done the same way as for TCGA HCC methylation data.

All the available clinical information for the validation cohorts is listed in Supplementary Table 3.

Deep Learning framework

We used the 3 pre-processed TCGA HCC omics data sets as the input for the autoencoders framework. We stacked the 3 matrices that are unit-norm scaled by sample, in order to form a unique matrix as reported before[41]. An autoencoder is a feedforward, non-recurrent neural network[24]. Given an input layer x, the objective of an autoencoder is to reconstruct x by the output layer x' (x and x' have the same dimension), via transforming x through successive hidden layers. For a given layer *i*, we used *tanh* as activation function between input layer *x* and output layer y. That is:

$$y = f_i(x) = tanh(W_i x + \ b_i)$$

Where W_i is the coefficient matrix and b_i the intercept. For an autoencoder with k layers, x' is then given by:

$$x' = \ F_{1\to k}(x) = \ f_1 \circ \ldots \circ f_k(x)$$

We chose *logloss* as objective function:

$$logloss(x, x') = \sum_{k=1}^{d} (x_k \log(x'_k) + (1 - x_k) \log(1 - \ x'_k))$$

In order to control overfitting, we added a L1 regularization penalty on the coefficient weight W_i, and a L2 regularization penalty on the hidden nodes activities: $F_{1\to k}(x)$. Thus the objective function to minimize becomes:

$$L(x, x') = logloss(x, x') + \sum_{i=1}^{k} (\alpha_w \|W_i\|_1 + \alpha_a \|F_{1\to i}(x)\|_2^2)$$

New feature selection and K-means clustering

We used the bottleneck layer of the autoencoder to select features linked to survival. For each node of this layer, we computed the activity of the node for every sample from the training set and built a Cox-PH model using the survival data. We selected nodes from which a significant Cox-PH model is obtained (log-rank p-value < 0.05). We then used these new features to cluster the samples using the k-means clustering algorithm. We determined the optimal number of clusters with two metrics: Silhouette index[42] and Calinski-Harabasz criterion[43]. We used the *scikit-learn* package as the K-Means implementation[44].

Supervised classification

Using the labels obtained from K-means clustering, we built a supervised classification model using Support Vector machine (SVM) algorithm. We first selected common features between the training and the validation datasets. We then applied robust-scaling on the validation dataset, using the means and the standard deviations of the training dataset[45]. Finally, we selected the top N features which are most correlated with the cluster labels, using ANOVA F-values. We set default N values as 200 for mRNAs, 200 for methylation and 50 for miRNAs. We used grid search approach to find the best hyperparameters of the SVM classifier, using 5-fold CV for each set of parameters. We used the *scikit-learn* package to build the SVM models, perform the grid search and compute the ANOVA[44].

Robustness assessment

We performed robustness assessment using a CV like procedure. We used a 60/40% split (training/test sets) of the TCGA data, in order to have sufficient number of test samples that generate evaluation metrics. We first randomly split the 360 samples from TCGA into 5 folds and used each pair of folds as a new fold (40% of the data), thus obtaining 10 new folds. For each fold, we constructed a model using the 60% remaining samples and predicted the labels for the sample from the fold.

Evaluation metrics for models

The metrics used closely reflects the accuracy of survival prediction in the subgroups identified. Three sets of evaluation metrics were used.

Concordance index (C-index): The c-index can be seen as the fraction of all pairs of individuals whose predicted survival times are correctly ordered[25] and is based on Harrel's C statistics[46]. A C-index score around 0.70 indicates a good model, whereas a score around 0.50 means random background.

To compute the c-index, we first built a Cox-PH model using the training dataset (cluster labels and survival data) and predict survival using the labels of the test/validation dataset. We then calculated the concordance index (c-index) using function concordance.index in R *survcomp* package [47]. To compute the C-index using the multiple clinical features, we built a Cox-PH using the *glmnet* package [48] instead, which enables penalization through ridge regression. Before building the Cox-PH model, we performed a 10-fold cross-validation to find the best lambda.

Log-rank p-value of Cox-PH regression: We plotted the Kaplan-Meier survival curves of the two risk groups, and calculated the log-rank p-value of the survival difference between them. We used Cox proportional hazards (Cox-PH) model for survival analysis[49], similar to described before[50,51], using R *survival* package[52]. For the two datasets with a low number of samples, E-TABM-36 (40 samples) and the Hawaiian methylation dataset (27 samples), we amplified the number of samples by randomly selecting with 200 replacement samples, in order to obtain the reliable statistical power.

Brier score: It is another score function that measures the accuracy of probabilistic prediction[26]. In survival analysis, the brier score measures the mean of the difference between the observed and the estimated survival beyond a certain time[53]. The score ranges between 0 and 1 and a larger score indicates higher inaccuracy. We used the implementation of Brier score from R *survcomp* package.

Functional analysis

A number of functional analyses were performed to understand the characteristics of 2 survival risk subtypes of TCGA HCC samples.

TP53 mutation analysis: We analyzed the somatic mutation frequency distributions in the survival subtypes for the *TP53* gene, among TCGA and LIRI-JP cohorts. TCGA and LIRI-JP cohorts have exome sequencing and whole genome sequencing data for 186 and 230 samples with survival data, respectively. We performed Fisher's test on *TP53* mutation between two survival risk groups.

Clinical covariate analysis: We tested the associations of our identified subtypes with other clinical characters, including gender, race, grade, stage and risk factors, using Fisher's exact tests. To test if the two survival risk subtypes have prognostic values in addition to clinical characteristics, we built a combined Cox-PH model with survival risk classification and clinical data, and compared it to the one with only clinical data (stage, grade, race, gender, age and risk factor).

Differential Expression: In order to identify the differential expressed genes between the two survival risk subtypes, we performed the differential expression analysis for the mRNA, miRNA expression and methylation genes. We used DESeq2 package[28] to identify the differential gene and miRNA expression between the 2 subtypes (false discovery rate, or FDR <0.05). Additionally, we used log2 fold change greater than 1 as filtering for mRNA/miRNA. For methylation data, we transformed the beta values into M values as elsewhere[54,55] using the *lumi* package in R[56]. We fit the linear model for each gene using *lmFit* function followed by empirical Bayes method, using *limma* package in R[57]. It uses moderate t-tests to determine significant difference in methylation for each gene between S1 and S2 subtypes (Benjamin-Hochberg corrected P<0.05). Additionally, we used averaged M value differences greater than 1 as filtering. We used volcano plot to show the differentially methylated genes in two subtypes.

Enriched pathway analysis: We used upregulated and downregulated genes for the KEGG pathway analysis, using the functional annotation tool from the online DAVID interface[58,59]. We used a p-value threshold of 0.10 to consider a pathway significant. We plot the gene-pathway network using Gephi[60].

Acknowledgements

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (http://datascience.nih.gov/bd2k), P20 COBRE GM103457 awarded by NIH/NIGMS, NICHD R01 HD084633 and NLM R01LM012373 and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to LX Garmire.

Author contributions

LXG envisioned the project, LL and KC prepared the dataset. OP, KC and LL developed the pipeline. KC, OP, LL and LXG wrote the manuscript. All authors have read, revised, and approved the final manuscript.

Conflict of interest

The authors declared no conflict of interest.

References

1 Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* **65**, 87-108 (2015).

2 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2015. *CA Cancer J Clin* **65**, 5-29 (2015).

3 Marrero, J. A., Kudo, M. & Bronowicki, J. P. The challenge of prognosis and staging for hepatocellular carcinoma. *Oncologist* **15 Suppl 4**, 23-33 (2010).

4 Colagrande, S. *et al.* Challenges of advanced hepatocellular carcinoma. *World J Gastroenterol* **22**, 7645-7659 (2016).

5 Nguyen, V. T., Law, M. G. & Dore, G. J. Hepatitis B-related hepatocellular carcinoma: epidemiological characteristics and disease burden. *J Viral Hepat* **16**, 453-463 (2009).

6 Trevisani, F. *et al.* Impact of etiology of cirrhosis on the survival of patients diagnosed with hepatocellular carcinoma during surveillance. *Am J Gastroenterol* **102**, 1022-1031 (2007).

7 Chen, C. H. *et al.* Hepatitis B- and C-related hepatocellular carcinomas yield different clinical features and prognosis. *Eur J Cancer* **42**, 2524-2529 (2006).

8 Chen, C. H. *et al.* Long-term trends and geographic variations in the survival of patients with hepatocellular carcinoma: analysis of 11,312 patients in Taiwan. *J Gastroenterol Hepatol* **21**, 1561-1566 (2006).

9 Llovet, J. M. *et al.* Sorafenib in advanced hepatocellular carcinoma. *N Engl J Med* **359**, 378-390 (2008).

10 Chen, X. *et al.* Gene expression patterns in human liver cancers. *Mol Biol Cell* **13**, 1929-1939 (2002).

11 Lee, J. S. *et al.* Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* **40**, 667-676 (2004).

12 Breuhahn, K. *et al.* Molecular profiling of human hepatocellular carcinoma defines mutually exclusive interferon regulation and insulin-like growth factor II overexpression. *Cancer Res* **64**, 6058-6064 (2004).

13 Boyault, S. *et al.* Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology* **45**, 42-52 (2007).

14 Chiang, D. Y. *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* **68**, 6779-6788 (2008).

15 Hoshida, Y. *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res* **69**, 7385-7392 (2009).

16 Goossens, N., Sun, X. & Hoshida, Y. Molecular classification of hepatocellular carcinoma: potential therapeutic implications. *Hepat Oncol* **2**, 371-379 (2015).

17 Liu, G., Dong, C. & Liu, L. Integrated Multiple "-omics" Data Reveal Subtypes of Hepatocellular Carcinoma. *PLoS One* **11**, e0165457 (2016).

18 Tan, J., Ung, M., Cheng, C. & Greene, C. S. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp*

Biocomput, 132-143 (2015).

19 Chen, L., Cai, C., Chen, V. & Lu, X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17 Suppl 1**, 9 (2016).

20 Khalili, M., Alavi Majd, H., Khodakarim, S., Ahadi, B. & Hamidpour, M. Prediction of the Thromboembolic Syndrome: an Application of Artificial Neural Networks in Gene Expression Data Analysis. *2016 %9 Thromboembolic syndrome; gene expression data; principal component analysis (PCA);auto-encoder neural networks %! Prediction of the Thromboembolic Syndrome: an Application of Artificial Neural Networks in Gene Expression Data Analysis* **7**, 8 (2016).

21 Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* **6**, 26094 (2016).

22 Chen, Q., Song, X., Yamada, H. & Shibasaki, R. in *Thirtieth AAAI Conference on Artificial Intelligence.*

23 Tan, J., Hammond, J. H., Hogan, D. A. & Greene, C. S. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems* **1** (2016).

24 Bengio, Y. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* **2**, 1-127 (2009).

25 Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P. & Raykar, V. C. in *Advances in neural information processing systems.* 1209-1216.

26 Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review* **78**, 1-3 (1950).

27 Villanueva, A. & Hoshida, Y. Depicting the role of TP53 in hepatocellular carcinoma

progression. *J Hepatol* **55**, 724-725 (2011).

28 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

29 Yamashita, T. *et al*. EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. *Cancer Res* **68**, 1451-1461 (2008).

30 Andersen, J. B. *et al*. Progenitor-derived hepatocellular carcinoma model in the rat. *Hepatology* **51**, 1401-1409 (2010).

31 Cao, L. *et al*. OCT4 increases BIRC5 and CCND1 expression and promotes cancer progression in hepatocellular carcinoma. *BMC Cancer* **13**, 82 (2013).

32 Kim, S. M. *et al*. Sixty-five gene-based risk score classifier predicts overall survival in hepatocellular carcinoma. *Hepatology* **55**, 1443-1452 (2012).

33 White, B. D., Chien, A. J. & Dawson, D. W. Dysregulation of Wnt/beta-catenin signaling in gastrointestinal cancers. *Gastroenterology* **142**, 219-232 (2012).

34 Zhu, Y., Qiu, P. & Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods* **11**, 599-600 (2014).

35 Wang, B. *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-337 (2014).

36 Xiang, Q. *et al*. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinformatics* **9**, 252 (2008).

37 Fujimoto, A. *et al*. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* **48**, 500-509 (2016).

38 Roessler, S. *et al*. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res* **70**, 10202-10212 (2010).

39 Wei, R. *et al*. Clinical significance and prognostic value of microRNA expression signatures in

hepatocellular carcinoma. *Clin Cancer Res* **19**, 4780-4791 (2013).

40 Song, M. A. *et al.* Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PLoS One* **8**, e55761 (2013).

41 Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci Rep* **6**, 28517 (2016).

42 Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65 (1987).

43 Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1-27 (1974).

44 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

45 Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, 878 (2016).

46 Lee, K. & Mark, D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**, 361-387 (1996).

47 Schröder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**, 3206-3208 (2011).

48 Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1 (2010).

49 Cox, D. R. in *Breakthroughs in statistics* 527-541 (Springer, 1992).

50 Wei, R. *et al.* Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget* **7**, 55249-55263 (2016).

51 Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A novel model to combine clinical and

pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Comput Biol* **10**, e1003851 (2014).

52 Therneau, T. A package for survival analysis in S. R package version 2.38. (2015).

53 Zhang, X. *et al.* Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach. *bioRxiv*, 043661 (2016).

54 Ching, T. *et al.* Genome-scale hypomethylation in the cord blood DNAs associated with early onset preeclampsia. *Clin Epigenetics* **7**, 21 (2015).

55 Ching, T. *et al.* Genome-wide hypermethylation coupled with promoter hypomethylation in the chorioamniotic membranes of early onset pre-eclampsia. *Mol Hum Reprod* **20**, 885-904 (2014).

56 Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-1548 (2008).

57 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

58 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).

59 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

60 Bastian, M., Heymann, S. & Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. (2009).

Figure Legends

Figure 1: Overall workflow

(A) Autoencoder architecture used to integrate 3 omics of HCC data. (B) Workflow combining deep learning and machine learning techniques to infer HCC subtypes in an unsupervised manner.

Figure 2: Significant survival differences for TCGA and external validation cohorts

(A) TCGA cohort, (B) LIRI-JP cohort, (C) NCI cohort, (D) Chinese cohort, (E) E-TABM-36 cohort, and (F) Hawaiian cohort.

Figure 3: Differentially expressed genes and their enriched pathways in the two subtypes from TCGA cohort

S1: aggressive (higher-risk survival) subtype; S2: moderate (lower-risk survival) subtype.

Figure 4: Bipartite graph for significantly enriched KEGG pathways and upregulated genes in two subtype.

Enriched pathway-gene analysis for upregulated genes in the (A) S1 aggressive tumor sub-group and (B) less aggressive S2 sub-group.

Tables

Table 1: Cross-validation based performance robustness of SVM classifier on training and testing dataset in TCGA cohort.

Dataset	10-folds CV	C-index	Brier score	Log-rank p-value (geo. mean)
Train	3-omics train (60%)	0.70 (± 0.04)	0.19 (± 0.01)	0.001
Test	3-omics test (40%)	0.69 (± 0.08)	0.20 (± 0.02)	0.005
	RNA only	0.68 (± 0.07)	0.20 (± 0.02)	0.01
	MIR only	0.69 (± 0.07)	0.20 (± 0.02)	0.003
	METH only	0.66 (± 0.07)	0.20 (± 0.02)	0.031

Table 2: Performance of classifier for the five external validation dataset.

Validation cohort	Omics Data type	Reference	# samples	C-index	Brier score	Log-rank p-value
LIRI-JP	RNA-Seq	[37]	230	0.75	0.16	4.4e-4
NCI	mRNA microarray	[38]	221	0.67	0.18	1.05e-3
Chinese	miRNA array	[39]	166	0.69	0.21	8.49e-4
E-TABM-36	mRNA microarray	[13]	40	0.77	0.19	3.06e-4*
Hawaiian	DNA methylation	[40]	27	0.82	0.19	1.19e-8*

*p-value obtained after resampling, due to small sample sizes.

Table 3: Performance of the model using clinical features on validation datasets.

Validation cohort	C-index (clinic only)	C-index (Combined[#])	Brier score	Log-rank p-value
LIRI-JP	0.55	0.74	0.16	0
NCI	0.45	0.65	0.19	0.007
E-TABM-36	0.50	0.75	0.19	0.007*
Hawaiian	0.70	0.87	0.19	4.05e-11*

#Combined = clinical + DL-based class labels

*p-value obtained using resampling datasets

Table 4: Full model performance within each subpopulation stratified by the clinical confounders in TCGA cohort.

Confounder	**# samples**	**C-index**	**Brier score**	**Log-rank p-value**
HBV	74	0.74	0.20	0.04
HCV	31	0.69	0.19	0.20
Alcohol	67	0.79	0.20	0.005
Others	59	0.77	0.19	0.0035

Supplementary Materials

Supplementary Figure 1: (A) Selection of the best subcluster K according to Silhouette score and Calinski-Harabasz score. (B) Kaplan-Meier plots show the separation of subtypes in terms of survival profiles from K=2 to 6.

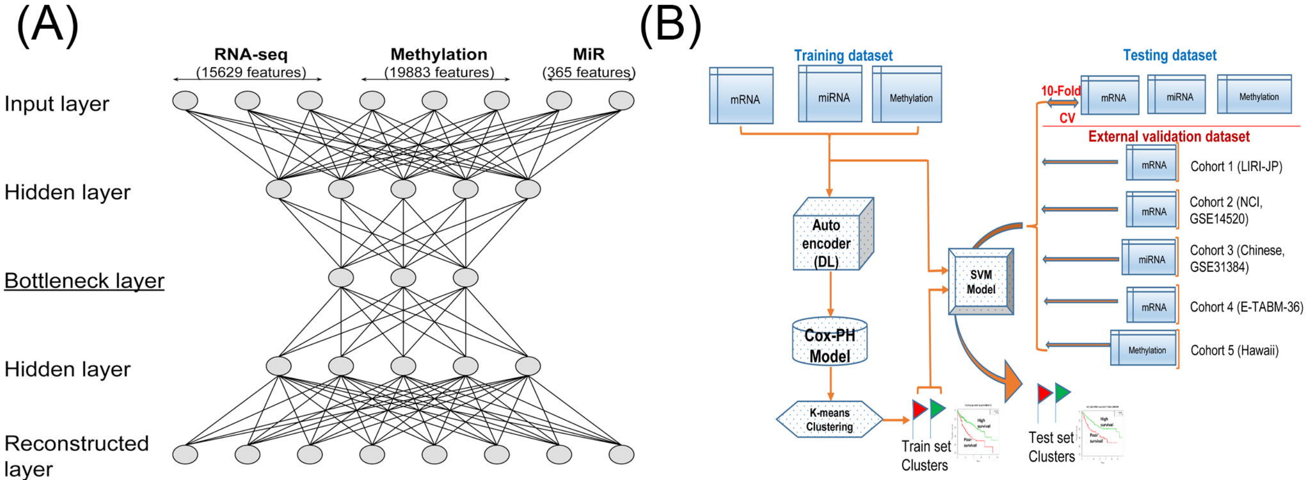
Supplementary Figure 2: Differential tests for miRNAs and Methylation (A) heatmap shows the differentially expressed miRNAs in two subtypes and (B) Volcano plot showing the differentially methylated genes in two subtypes. Red dotted line: BH adjusted p-value=0.05; blue dotted line: p-value=0.05 without adjustment. Red color: genes differentially methylated with BH adjusted p-value <0.05 and absolute mean difference >1 between the two subtypes.

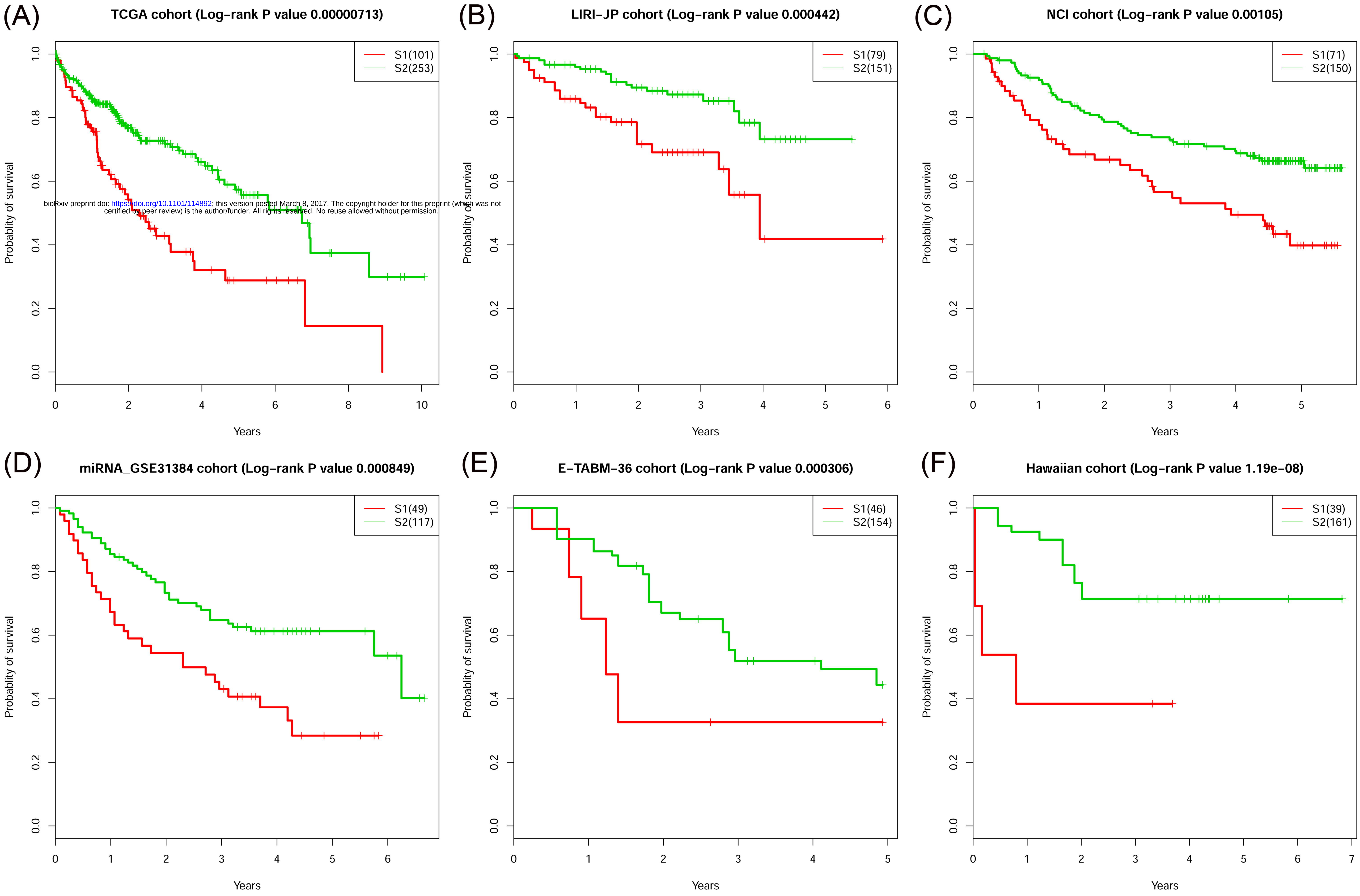
Supplementary Table 1: Performance of PCA on training and testing dataset on different metrics.

Supplementary Table 2: Performance of Models constructed with one omic only.

Supplementary Table 3: Clinical Characteristics of the HCC cohorts used in this study.

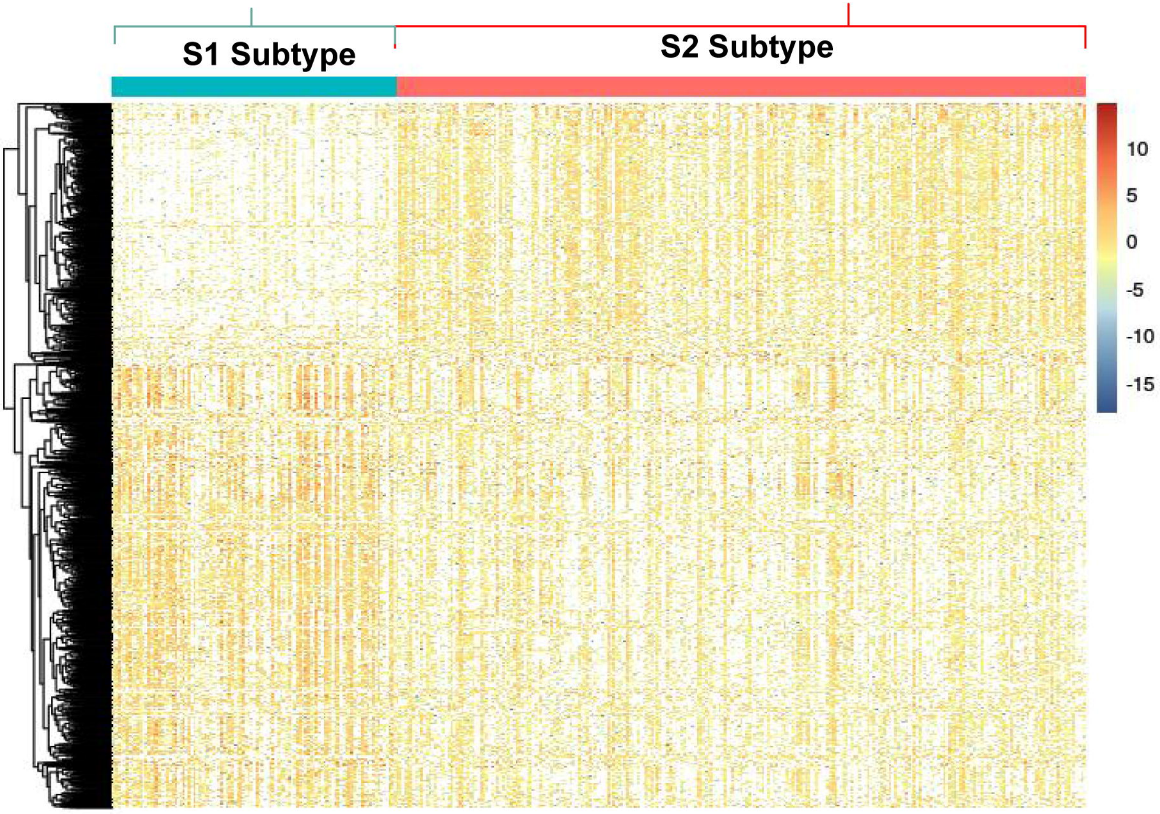
Supplementary File 1: Description of results on differential tests of miRNAs and methylation.

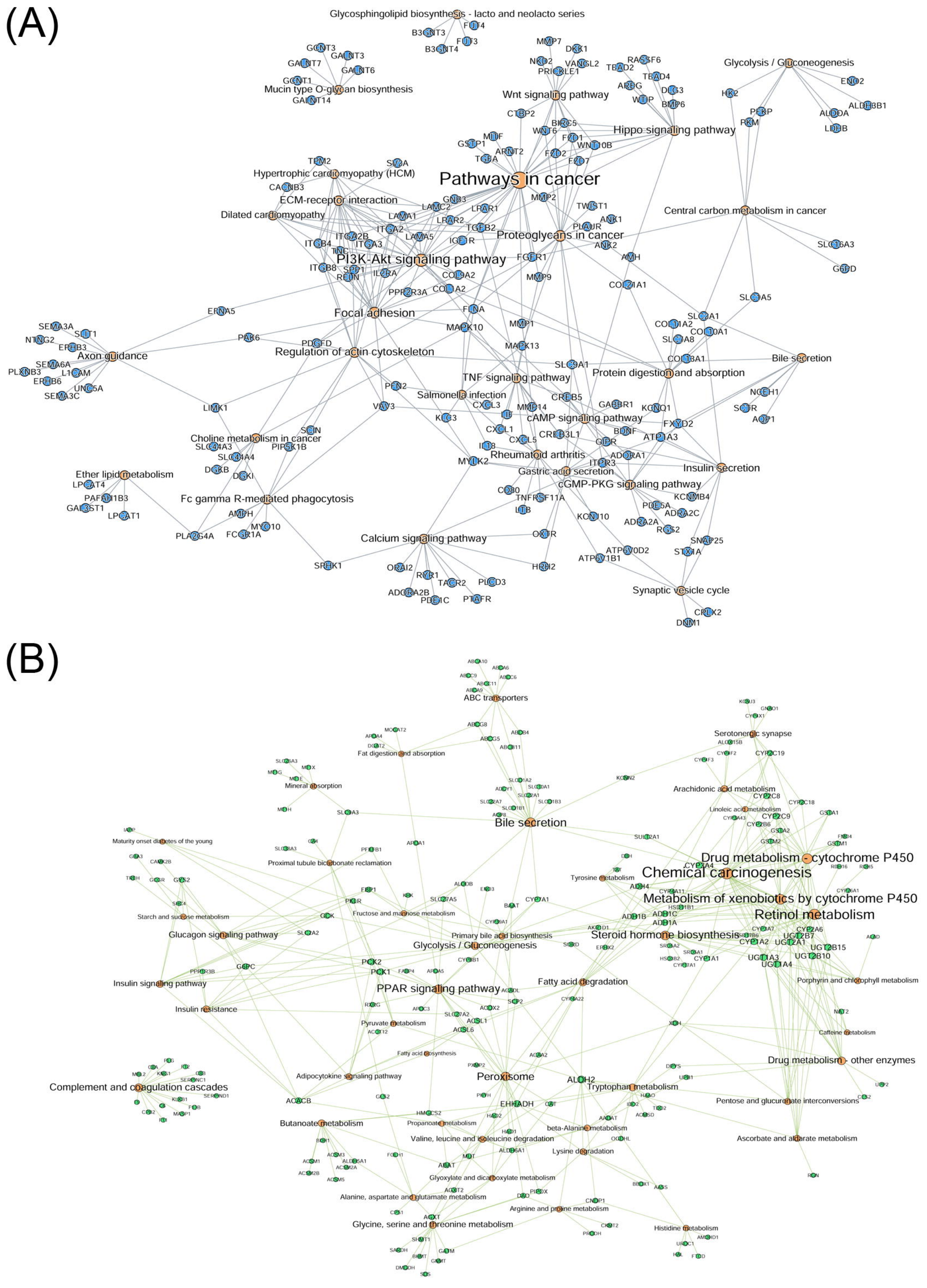

(A)
RNA-seq
(15629 features)
Methylation
(19883 features)
MiR
(365 features)
Input layer
Hidden layer
Bottleneck layer
Hidden layer
Reconstructed layer
(B)
Training dataset
mRNA
miRNA
Methylation
Auto encoder (DL)
Cox-PH Model
K-means Clustering
Train set Clusters
High survival
Poor survival
SVM Model
Testing dataset
10-Fold CV
mRNA
miRNA
Methylation
External validation dataset
mRNA
Cohort 1 (LIRI-JP)
mRNA
Cohort 2 (NCI, GSE14520)
miRNA
Cohort 3 (Chinese, GSE31384)
mRNA
Cohort 4 (E-TABM-36)
Methylation
Cohort 5 (Hawaii)
Test set Clusters
High survival
Poor survival


(A)
TCGA cohort (Log−rank P value 0.00000713)
S1(101)
S2(253)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Years
(B)
LIRI−JP cohort (Log−rank P value 0.000442)
S1(79)
S2(151)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Years
(C)
NCI cohort (Log−rank P value 0.00105)
S1(71)
S2(150)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
Years
(D)
miRNA_GSE31384 cohort (Log−rank P value 0.000849)
S1(49)
S2(117)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Years
(E)
E−TABM−36 cohort (Log−rank P value 0.000306)
S1(46)
S2(154)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
Years
(F)
Hawaiian cohort (Log−rank P value 1.19e−08)
S1(39)
S2(161)
Probablity of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
Years

Pathway name	EASE score	# genes
Pathways in cancer	0.024	27
PI3K-Akt signaling pathway	0.0214	24
Focal adhesion	0.0178	20
Proteoglycans in cancer	0.0169	19
Hippo signaling pathway	0.0134	15
Regulation of actin cytoskeleton	0.0134	15
ECM-receptor interaction	0.0125	14
Axon guidance	0.0116	13
cAMP signaling pathway	0.0116	13
Wnt signaling pathway	0.0107	12
cGMP-PKG signaling pathway	0.0107	12
Calcium signaling pathway	0.0107	12
Protein digestion and absorption	0.0098	11

Pathway name	EASE score	# genes
Metabolic pathways	0.19	123
Chemical carcinogenesis	0.0417	27
Biosynthesis of antibiotics	0.0417	27
Retinol metabolism	0.0371	24
Drug metabolism - cytochrome P450	0.034	22
Metabolism of xenobiotics by cytochrome P450	0.034	22
Steroid hormone biosynthesis	0.0278	18
Bile secretion	0.0278	18
PPAR signaling pathway	0.0263	17
Peroxisome	0.0263	17
Carbon metabolism	0.0263	17
Complement and coagulation cascades	0.0232	15
Drug metabolism - other enzymes	0.0216	14
Glycolysis / Gluconeogenesis	0.0201	13
Fatty acid degradation	0.0185	12
Glycine, serine and threonine metabolism	0.017	11
Tryptophan metabolism	0.017	11




(A)
Glycosphingolipid biosynthesis - lacto and neolacto series
FUT4
B3GNT3
FUT3
B3GNT4
GCNT3
GALNT3
GALNT7
GALNT6
GCNT1
Mucin type O-glycan biosynthesis
GALNT14
MMP7
DKK1
NKD2
VANGL2
PRICKLE1
RASSF6
TEAD2
TEAD4
AREG
DLG3
WTIP
BMP6
Wnt signaling pathway
Glycolysis / Gluconeogenesis
HK2
ENO2
ALDH3B1
PFKP
PKM
ALDOA
LDHB
CTBP2
BIRC5
WNT6
FZD1
Hippo signaling pathway
WNT10B
GSTP1
MIF
ARNT2
TGFA
FZD2
FZD7
TPM2
SV2A
Hypertrophic cardiomyopathy (HCM)
CACNB3
Pathways in cancer
ECM-receptor interaction
GNB3
LAMC2
LPAR1
MMP2
TWIST1
Dilated cardiomyopathy
LAMA1
LPAR2
ANK1
Central carbon metabolism in cancer
ITGA2
TGFB2
PLAUR
ITGB4
ITGA2B
LAMA5
IGF1R
Proteoglycans in cancer
ANK2
SLC16A3
TNC
ITGA3
PI3K-Akt signaling pathway
FGFR1
AMH
ITGB8
SPP1
G6PD
RELN
IL2RA
COL9A2
MMP9
COL2A1
PPP2R3A
COL1A2
SLC1A5
ERNA5
FLNA
Focal adhesion
SLC2A1
SEMA3A
COL11A2
SLIT1
MMP1
COL10A1
NTNG2
PAK6
MAPK10
SLC7A8
ERHB3
PDGFD
MAPK13
COL18A1
Bile secretion
SEMA6A
Axon guidance
Regulation of actin cytoskeleton
PLXNB3
L1CAM
SLC9A1
Protein digestion and absorption
ERHB6
UNC5A
TNF signaling pathway
SEMA3C
PFN2
Salmonella infection
NCEH1
CREB5
GABBR1
SCTR
LIMK1
CXCL3
KCNQ1
AQP1
VAV3
KLC3
LIF
MMP14
cAMP signaling pathway
FXYD2
SCIN
CXCL1
BDNF
CREB3L1
ATP1A3
Choline metabolism in cancer
CXCL5
GIPR
SLC44A3
PIP5K1B
IL18
SLC44A4
Rheumatoid arthritis
ADORA1
DGKB
MYLK2
ITPR3
Insulin secretion
DGKI
Gastric acid secretion
Ether lipid metabolism
cGMP-PKG signaling pathway
LPCAT4
CD80
TNFRSF11A
KCNMB4
PAFAH1B3
Fc gamma R-mediated phagocytosis
PDE5A
LTB
ADRA2C
GAL3ST1
LPCAT1
AMPH
KCNJ10
ADRA2A
RGS2
MYO10
PLA2G4A
FCGR1A
OXTR
SNAP25
Calcium signaling pathway
ATP6V0D2
STX1A
ATP6V1B1
SPHK1
ORAI2
HRH2
RYR1
TACR2
PLCD3
ADORA2B
Synaptic vesicle cycle
PDE1C
PTAFR
CPLX2
DNM1
(B)
ABCA10
ABCA6
ABCC9
ABCC6
ABCC11
ABCA9
ABC transporters
KCNJ3
GNAO1
CYP4X1
ABCG8
ABCB4
Serotonergic synapse
MOGAT2
APOA4
ABCG5
ALOX15B
DGAT2
ABCB11
CYP4F2
CYP2C19
Fat digestion and absorption
CYP4F3
SLC26A3
MT1X
KCNN2
MT1G
MT1E
SLCO1A2
Mineral absorption
SLC10A1
Arachidonic acid metabolism
ADCY1
CYP2C8
SLC22A1
SLC22A7
SLCO1B3
CYP2C18
Linoleic acid metabolism
MT1H
SLC9A3
SLCO1B1
AQP8
GSTA1
CYP3A43
CYP2C9
Bile secretion
CYP2B6
GSTA2
IAPP
SULT2A1
FMO4
GSTM2
GSTM1
Maturity onset diabetes of the young
CA4
SLC38A3
PFKFB1
APOA1
Drug metabolism - cytochrome P450
DBH
CYP3A4
RDH16
RDH5
GBA3
CAMK2B
Proximal tubule bicarbonate reclamation
TAT
Chemical carcinogenesis
Tyrosine metabolism
TREH
GCGR
GYS2
ALDOB
ADH4
CYP26A1
ENO3
FBP1
KHK
CYP4A11
Metabolism of xenobiotics by cytochrome P450
SLC27A5
PKLR
CYP7A1
SI
HSD11B1
BAAT
Retinol metabolism
ADH1C
Starch and sucrose metabolism
GCK
Fructose and mannose metabolism
ADH1B
ADH1A
CYP39A1
AKR1D1
CYP3A7
CYP2A6
Primary bile acid biosynthesis
SLC2A2
Steroid hormone biosynthesis
HSD17B6
UGT2B7
Glucagon signaling pathway
CYP1A2
UGT2A1
UGT2B15
AOAD
Glycolysis / Gluconeogenesis
SORD
EPHX2
SRD5A2
SRD5A1
HSD3B2
CYP1A1
UGT1A3
UGT2B10
PCK2
CYP8B1
CYP17A1
UGT1A4
G6PC
PPP1R3B
PCK1
FABP4
APOA5
Insulin signaling pathway
Porphyrin and chlorophyll metabolism
PPAR signaling pathway
Fatty acid degradation
ACADL
SCP2
CYP4A22
RXRG
APOC3
ACOX2
Insulin resistance
SLC27A2
NAT2
ACSL1
Pyruvate metabolism
XDH
ACOT12
ACSL6
Caffeine metabolism
Fatty acid biosynthesis
ACAA2
PXMP2
DPYS
Drug metabolism - other enzymes
PLG
F12
C4A
C8B
MBL2
KNG1
SERPINC1
Adipocytokine signaling pathway
UPB1
Peroxisome
ALDH2
UPP2
Complement and coagulation cascades
Tryptophan metabolism
HAAO
CES2
SERPIND1
F9
KLKB1
F13B
ACACB
GCS2
PHYH
EHHADH
CAT
IDO2
TDO2
Pentose and glucuronate interconversions
C8
CPB2
MASP1
F11
HMGCS2
ACMSD
HAO2
AADAT
Butanoate metabolism
Propanoate metabolism
beta-Alanine metabolism
HAO1
BDH1
Valine, leucine and isoleucine degradation
OGDHL
Ascorbate and aldarate metabolism
ACSM3
ALDH6A1
FOLH1
MUT
Lysine degradation
ACSM1
ALDH5A1
ACSM2A
ABAT
ACSM2B
ACSM5
Glyoxylate and dicarboxylate metabolism
BBOX1
AASS
RGN
AGXT2
PIPOX
DAO
CNDP1
Alanine, aspartate and glutamate metabolism
Arginine and proline metabolism
CPS1
AGXT
CKMT2
Glycine, serine and threonine metabolism
Histidine metabolism
PRODH
AMDHD1
SHMT1
UROC1
SARDH
GATM
BHMT
GNMT
HAL
FTCD
DMGDH
SDS