Computers in Biology and Medicine 171 (2024) 108113

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

SIMarker: Cellular similarity detection and its application to diagnosis and prognosis of liver cancer

Mengsha Tong [a,b,1,**], Shijie Luo [a,b,1], Lin Gu [a,1], Xinkang Wang [a,b], Zheyang Zhang [a,b], Chenyu Liang [a,b], Huaqiang Huang [a], Yuxiang Lin [b], Jialiang Huang [a,b,*]

[a] State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian 361102, China
[b] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, 316005, China

ARTICLE INFO

Keywords:
Hepatocellular carcinoma
Cirrhosis-like signatures
Relative expression
Orderings
Early diagnosis and prognosis
Single-cell transcriptome

ABSTRACT

Background: The emergence of single-cell technology offers a unique opportunity to explore cellular similarity and heterogeneity between precancerous diseases and solid tumors. However, there is lacking a systematic study for identifying and characterizing similarities at single-cell resolution.
Methods: We developed SIMarker, a computational framework to detect cellular similarities between precancerous diseases and solid tumors based on gene expression at single-cell resolution. Taking hepatocellular carcinoma (HCC) as a case study, we quantified the cellular and molecular connections between HCC and cirrhosis. Core analysis modules of SIMarker is publicly available at https://github.com/xmuhuanglab/SIMarker ("SIM" means "similarity" and "Marker" means "biomarkers).
Results: We found *PGA5*$^{+}$ hepatocytes in HCC showed cirrhosis-like characteristics, including similar transcriptional programs and gene regulatory networks. Consequently, the genes constituting the gene expression program of these cirrhosis-like subpopulations were designated as cirrhosis-like signatures (CLS). Strikingly, our utilization of CLS enabled the development of diagnosis and prognosis biomarkers based on within-sample relative expression orderings of gene pairs. These biomarkers achieved high precision and concordance compared with previous studies.
Conclusions: Our work provides a systematic method to investigate the clinical translational significance of cellular similarities between HCC and cirrhosis, which opens avenues for identifying similar paradigms in other categories of cancers and diseases.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful for characterizing cellular diversity of tumors at single-cell resolution. In recent years, an increasing number of single-cell transcriptome studies have been focused on the investigation of links between pre-cancerous lesions and cancer samples. For example, Li et al. reported shared cellular networks of gastric cancer and pre-cancerous lesions in epithelial cells [1]. Li et al. found that AT2 cells is the most likely origin of cancer cells in the development of lung adenocarcinoma [2]. Becker et al. used multi-omics single-cell datasets from healthy, adenomas, and colorectal cancer to discover the progression from pre-CAF to CAF [3]. Sun et al. unveiled the progression of CAFs, myeloid cells and T cells in the immune microenvironment between pre-cancerous lesions and oral cancer [4]. These studies extend our understanding about the molecular mechanisms of cancer initiation and progression. However, there is still lacking a useful and systematic toolkit for characterizing cellular similarities of tumor microenvironment between precancerous diseases and solid tumors. Application of cellular similarities in translational research remains a challenge.

HCC is one of the most leading causes of cancer deaths globally [5]. As shown in Table S1, lots of clinical reports revealed that cirrhosis is

* Corresponding author. State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian 361102, China.
** Corresponding author. State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian 361102, China.
E-mail addresses: mstong@xmu.edu.cn (M. Tong), jhuang@xmu.edu.cn (J. Huang).
[1] These authors contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.compbiomed.2024.108113
Received 11 October 2023; Received in revised form 9 January 2024; Accepted 4 February 2024
Available online 14 February 2024

one of the most important risk factors for HCC [6–16]. Many studies have reported similar signaling pathways between cirrhosis and liver cancer (Table S2). To improve the prognosis of HCC, it is important to develop biomarkers to distinguish HCC samples from patients with cirrhosis. It has been reported that tissue biopsy is important for early diagnosis of HCC. However, due to the inaccurate biopsy location for adjacent non-tumor tissues such as cirrhosis and normal of HCC, the false negative rate of diagnosis could be about 30%–50% [17]. Previous studies have explored several diagnostic biomarkers based on transcriptomes. Hoshida et al. identified a 186-gene signature to predict outcomes of patients with HCC and patients with early-stage cirrhosis [18]. Ning et al. developed a risk signature for early prediction of HCC development in early-stage cirrhosis patients [19]. These reported diagnostic biomarkers were based on risk scores summarized from expression levels of signature genes, which depends on data normalization due to experimental batch effects. These kinds of risk classification methods could not be diagnosed at the individualized level. Recent studies have reported the within-sample relative expression orderings (REOs) of genes are robust against to experimental batch effects, which could be directly applied to samples at the individualized level [20,21]. Moreover, the REOs-based risk model has been applied in the scRNA-seq. Wang et al. have proposed single-cell pair-wise gene expression (scPAGE) to aid in diagnosing acute myeloid leukemia [22]. We previously developed single-cell gene pair signatures to predict recurrence risk for colorectal patients [23]. Few studies focus on how to develop individualized biomarkers based on scRNA-seq for HCC. Most of studies relied on bulk transcriptomes or statistical studies of clinical information [24–33] (Table S2), which might partially obscure the characteristics of different cellular subpopulations. scRNA-seq technology could assist us in finding the cellular and molecular characteristics of microenvironment within cirrhotic and liver cancer samples. In this study, taking hepatocellular carcinoma (HCC) as a case study, we developed a novel computational workflow to identify diagnosis and prognosis biomarkers based on cellular similarities between HCC and cirrhosis.

2. Materials and methods

2.1. Datasets collection

In Table S3, the scRNA-seq data of healthy, cirrhosis, and liver cancer samples from Gene Expression Omnibus (GEO) were summarized. GSE136103 and GSE151530 were used as discovery datasets. GSE156337, GSE166635, GSE149614 and Mendeley Datasets were used as the validation sets [34–37]. In addition, we collected a dataset (GSE134520) of gastric cancer with different stages including intestinal metaplasia (IM), non-atrophic gastritis (NAG), chronic atrophic gastritis (CAG), and early gastric cancer (EGC), totaling 32,332 cells [38]. The colorectal cancer (CRC) dataset (GSE201349) consists of samples from 70 specimens at normal, tubular adenomas, and carcinoma sites, with a total of 201,884 cells [3].

Seventeen independent bulk transcriptome datasets were collected from the GEO and TCGA databases, as described in Table S4. For the prediction of cirrhosis progression to HCC, gene expression profiles of 216 patients with hepatitis C-associated early cirrhosis were available in GSE15654. In this dataset, 65 patients with cirrhosis eventually developed HCC were collected. To develop biomarkers for diagnosing HCC patients at an early stage, we collected four kinds of tissue samples: HCC tissues, adjacent cirrhosis tissues of HCC patients (CHCC), adjacent normal tissues of HCC patients (Adj) and cirrhosis tissues of non-HCC patients (CoHCC). The corresponding sample numbers for each kind of tissue were shown in Table S4. Among these datasets, samples from GSE54236, GSE64041 and GSE15654 were biopsy specimens. The remaining datasets were surgical resection samples. For the prognosis of early-stage HCC patients, we collected 572 samples with survival information from TCGA, GSE116174, GSE14520, and GSE76427 (Table S5). The details of HBV, HCV or non-viral for HCC patients were summarized in Tables S3–4.

2.2. Preprocessing of single-cell RNA datasets

Cells with fewer than 300 detected genes were excluded. Gene expression level was calculated as the fraction of its unique molecular identifier (UMI) count with respect to the total UMI count, which was then multiplied by scale and log2-transformed. Normalization, selection of highly-variable genes, dimension reduction and clustering of single-cell transcriptome data was performed using Scanpy, a Python-based package [39]. *FindIntegrationAnchors* function in Seurat was used to correct the batch effect among samples from different resources. Then, major immune cell types were annotated according to these markers [40–44]: Hepatocyte (ALB, TF, TTR), Cholangiocyte (EPCAM, KRT19, CD24), T cell (CD3D, CD3E, CD3G), Mesenchyme (PDGFRB, ACTA2, COLIA1), Endothelial (PECAM1, CDH5, ICAM2), B cell (CD79A, CD79B, CD19), MPs (CD68, CD14, ITGAX), etc. Minor cell subpopulations were named according to the marker genes found by re-clustering (Table S6). Finally, the accuracy of the annotated results was checked by using *sc. pl. correlation_matrix* function. For datasets of gastric cancer and colorectal cancer, the labels of cell annotations were obtained from the original studies [3,38].

2.3. Preprocessing of bulk transcriptome datasets

For the Affymetrix microarray datasets, the Robust Multi-array Average algorithm [45] was used to normalized the matrix. For the RNA-seq datasets, we normalized the matrix by *vst* method from DEseq2 package [46]. Clinical information including gender, age, disease stages, history of cirrhosis and survival time were summarized in Table S5.

2.4. Overview of SIMarker construction

SIMarker consists of three main parts (Fig. 1).

2.4.1. Part 1: identification of cirrhosis-like subpopulations

In order to calculate the similarity of two target subpopulations from cirrhotic and liver cancer samples, we designed a modified statistical method based on the hypothesis that cells resemble transcriptome characteristics in different samples might have similar microenvironments or niches [47]. We firstly normalized the raw count expression matrix, then screened the genes with coefficient of variation ranked at top 500 by calculating $hvg = sd/mean$. The top 500 genes with high specificity for all subpopulations were further screened using Entropy Weighting method (1, 2). After that, a query-reference framework based on k-nearest neighbors (KNN) was employed: healthy/cirrhosis subpopulations were used as reference objects and tumor subpopulations were used as query. Each object ($O = \{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$) have N cells (x) and labels (y). The K cells closest to x in the reference object according to the Euclidean distance were searched (3), and the set represented by k cells was denoted as N_K(x), Similarly, 10-fold cross-validation was employed to select the optimal K value during each cell type query. Then, cell (x) was determined to belong to which particular cell subpopulation (y) based on the majority voting principle (4). A phenotypic marker index for each query type was defined by calculating the proportion of query cells labeled with the corresponding sample type. In addition to KNN, we offered other commonly used clustering methods including Partition clustering, Hierarchical clustering, Mixture models, Density-based and Neural networks to calculate cellular similarity (Table S7). The performance of these methods was compared in the datasets of gastric cancer, colorectal cancer, and hepatocellular carcinoma, respectively.

$$H(x) = -\sum_{i=1}^{n} p_i \log_n p_i \quad (1)$$

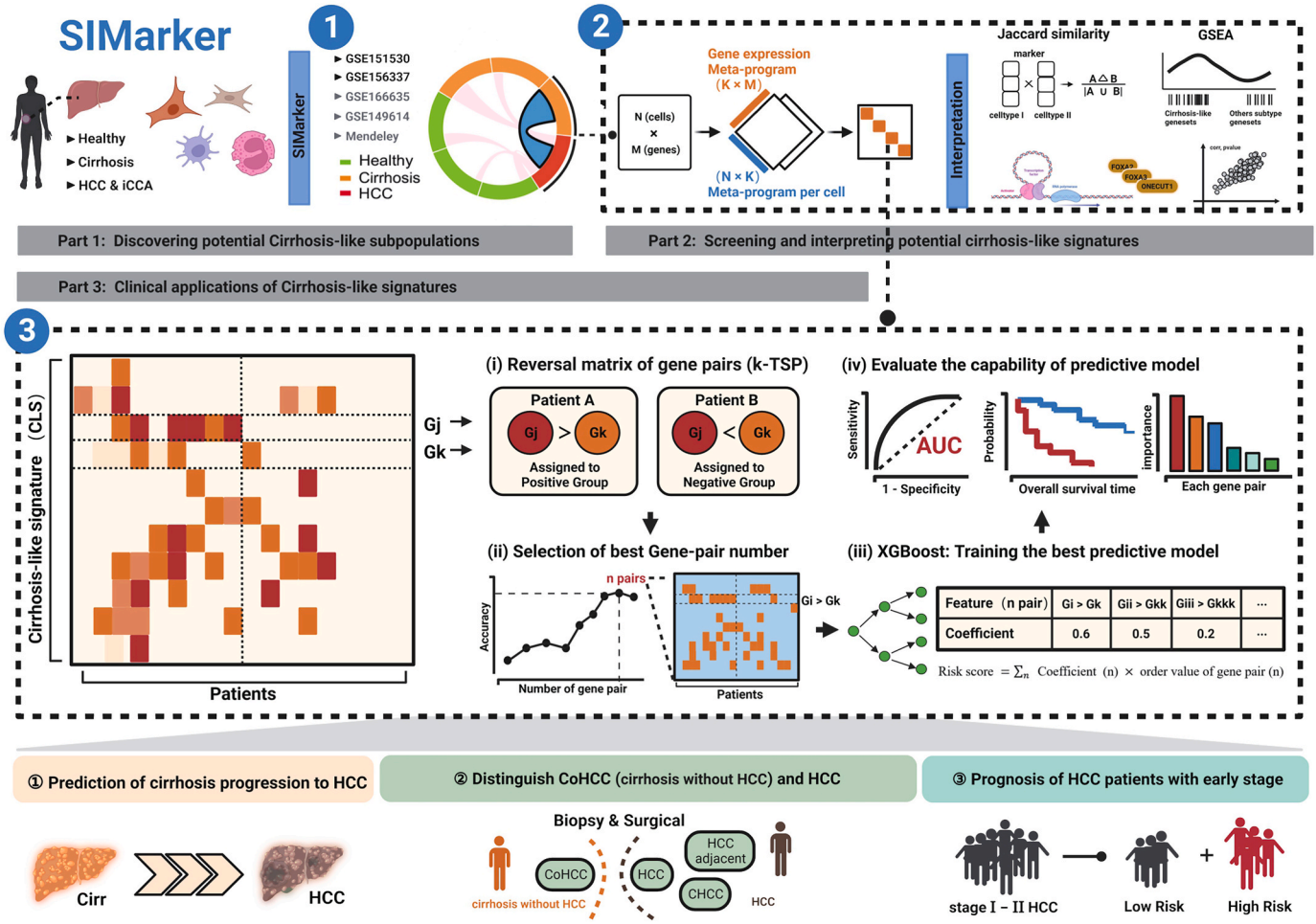

Fig. 1. The framework of this study.
The workflow of SIMarker. The workflow was drawn at https://biorender.com. "SIM" means "similarity" and "Marker" means "biomarkers.

$$W_i = \frac{1 - E_i}{k - \sum E_i} \tag{2}$$

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{3}$$

$$y = \operatorname{argmax} \sum_{x_i \in N_k(x)} I(y_i, c_j), i = 1,2,\ldots,N; j = 1,2,\ldots,K \tag{4}$$

2.4.2. Part 2: screening and interpreting cirrhosis-like signatures

2.4.2.1. Definition of cirrhosis-like signatures. Genes within the gene expression program from cirrhosis-like subpopulations, were defined as cirrhosis-like signatures (CLS). We dissected gene expression programs based on non-negative matrix factorization (NMF), which were repeated 100 times for each cell type and a set of consensus programs were computed by aggregating results from all 100 runs. The cNMF v1.4 python package was used to perform the analysis of consensus NMF [48]. The optimal number of programs were determined for each cell type by maximizing stability and minimizing error of the cNMF solution as well as ensuring the programs were biologically coherent.

2.4.2.2. Jaccard similarity for cirrhosis-like subpopulations. Transcriptional similarity between two subpopulations were also evaluated by the Jaccard similarity coefficient (5). We further selected marker genes for cirrhosis-like subpopulations using *scanpy. tl.rank_genes_groups* function. We selected top 60/100/200 marker genes according to the characteristics of subpopulations of different scales. A and B represented marker gene lists of subpopulations.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

2.4.2.3. Gene regulatory network inference analysis. Single-cell regulatory network inference and clustering (SCENIC) [49] from the python package 'pySCENIC' was used to explore similar characteristics at the gene network regulatory level. The potential transcription factors were found by GENIE3 using co-expression between genes. The high-frequency motifs appearing near the gene transcription initiation site were used. The motifs (NES>3) found in the database were compared to calculate the regulators enriched by the target gene. Finally, based on the expression value of the gene, the specific regulator activity specific to each cell was calculated with AUCell function. To obtain the correlation between each subpopulation, Spearman correlation was used.

2.4.2.4. Deconvolution of cell abundance. Normalized gene expression matrix from 33,686 single cells, belonging to 21 HCC subpopulations were used for GSVA [50] function. Deconvolution was performed on bulk transcriptome, including 572 samples from 4 cohorts (TCGA, GSE14520, GSE116174, GSE76427). The abundance of the 21 HCC subpopulations were normalized to a sum of 1. The score represents the estimated proportion for each cell type. The normalized HCC cell type

compositions of these patients were used for the PCA analysis. Euclidean distance was used to detect neighbors. We project the query dataset onto the PCs of the reference dataset to assign the corresponding cluster labels.

2.4.3. Part 3: clinical applications of cirrhosis-like signatures

In this study, we used the CLS obtained from cNMF as input, and we developed three kinds of individualized biomarkers independently (Fig. 1): ① Risk prediction of cirrhosis progressing to HCC; ② Accurately distinguish cirrhosis without HCC (CoHCC), HCC, HCC adjacent and HCC accompanied with cirrhosis (CHCC) in biopsy and surgical samples; ③ Prognosis of HCC patients with early stage. Here, we used k-TSP to screen out candidate gene pairs in different bulk transcriptome datasets, and feature importance scores of gene pairs were generated by XGBoost to indicate the usefulness of each feature in constructing the model.

2.4.3.1. Conversion of gene pair matrix. We converted the gene expression matrices into gene pair matrices. G represents the genes from CLS. Specifically, the transcriptome profile $E \in \mathbb{R}^{K\times N}$ composed of the expression levels of K ($G = \{g_1, \cdots, g_K\}$) in N samples ($P = \{p_1, \cdots, p_N\}$), L denotes the largest number of gene pairs, and $i, j \in \{1, \cdots, K\}$, the gene pair matrix M (6) was built based on the order value as follows:

$$M_{\mathrm{LN}} = \begin{cases} 1, E_{ik} - E_{jk} > 0 \\ 0, E_{ik} - E_{jk} \leq 0 \end{cases} \tag{6}$$

2.4.3.2. Feature selection based on k–Top Scoring Pairs. k-Top Scoring Pairs (k-TSP) [51] is a classification method for making predictions from transcriptome data based on a set of gene pairs. Each of relative expression orderings (REOs) of gene pairs is associated with one of the two categories (for example: cancer and normal). The k-TSP prediction rule, is a voting summary of this separate two-feature decision rule based on conversion of REOs. K-TSP, as well as its precursor, Top Scoring Pair (TSP), relies on the ranking of only a fraction of features, making it robust across datasets. Compared to TSP, k-TSP has comparable accuracy to standard classification methods. In this study, we used the CLS as input for k-TSP-based prediction, using the R package 'switchBox' [52].

2.4.3.3. Predicting model based on XGBoost. Firstly, we set the following labels for each sample: Part1: cirrhosis: "0", HCC: "1"; Part2: CoHCC: "0", HCC: "1" and tissues from HCC patients: "1"; Part3: survival: "0", death: "1". True positive (TP) means that the predicted label "1" is correct and false positive (FP) means that the predicted label "1"is wrong. True negative (TN) means that the predicted label "0" is correct and false negative (FN) means that the predicted label "0" is wrong. Each point on the ROC curve represents a random pair of sensitivity and specificity values (7,8), and the Accuracy was calculated as follows:

$$\text{Sensitivity} = \frac{\mathrm{TP}}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{\mathrm{TN}}{FP + TN} \tag{8}$$

We randomly divided the bulk data into training cohort, test cohort and validation cohort as described in Table S4. In order to select the optimal number of gene pairs, we repeated XGBoost based on the highest number of gene pairs that could be found by k-TSP, and made selection based on the best predicted value (Accuracy = Predicted label/Actual label) (9). Extreme gradient boosting (XGBoost, https://xgboost.read thedocs. io/en/stable/index.html), is an ensemble algorithm of

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FP + TN + FN} \tag{9}$$

decision trees. The final prediction of a specific instance is the sum of the predicted values in each tree. Next, we measured the accuracy on the training, test, and validation cohorts by calculating the area under the receiver operating characteristic curve (AUC) based on ROC or time-dependent ROC.

After that, the prognostic and predictive values of gene pairs were built by using the regression coefficients derived from XGBoost analysis for each sample based on the training cohort. The risk score formula was established as follows:

$$\text{Risk score} = \sum_j \text{Coefficient (j)} \times \text{order value of gene pair (j)}. \tag{10}$$

The coefficient of gene pair (j) is the regression coefficient of the gene pair (j), and the order of the gene pair (j) is from the transformed gene pair matrix M(6). The risk score of each sample was calculated by using the 'predict' function from XGBoost R package. We selected 50% of risk score as the thresholds for high and low risk classification in Part1 and Part2 by ranking the risk scores in descending order. In Part3, we selected the risk score with the best generalization ability in the testing set as the threshold. Survival curves were estimated using the Kaplan–Meier method and were compared using the log-rank test. Multivariate Cox regression analyses were used to examine whether the risk model was an independent prognostic factor.

3. Results

3.1. Detecting cellular similarities between precancerous diseases and solid tumors

By utilizing single-cell transcriptome data from diverse samples, the core analysis of similarity is clustering cells. Here, we employed commonly used clustering techniques including Partition clustering, Hierarchical clustering, Mixture models, Density-based and Neural networks to tackle this challenging task (Table S7).

To demonstrate the capability of SIMarker to detect cellular similarities, we used cancer cells, mesenchymal stem cells (MSCs) and proliferative cell as positive controls in a comparative analysis of gastric cancer and the corresponding precancerous lesions reported by Zhang et al. [38]. The similarity between MSCs subpopulations at IM stage and EGC subpopulations at EGC stage was observed through connectivity analysis (Fig. S1A), which consisted with the previous finding. In addition to KNN, we also used other four clustering methods including CellTree, TSCAN, Monocle and scDeepcluster (Table S7). Interestingly, we found that KNN-based method was more consistent with the original findings compared to other methods (Fig. S1A). Then, we used SIMarker to identify cellular similarities in the CRC dataset. Becker et al. identified an enrichment of stem-like epithelial cells and a depletion of mature enterocytes in cells originating from polyps and CRC samples [3]. In the different stages of the CRC samples, we found that KNN could better align with the results of previous studies compared with other methods, reproducing a higher similarity in MSCs with cancer cell types at IM stages (Fig. S1B). These results demonstrate the feasibility of SIMarker in identifying cellular similarities between precancerous lesions and cancer samples.

3.2. SIMarker uncovers $PGA5^+$ hepatocytes in HCC exhibit cirrhosis-like features

Single-cell transcriptome datasets from 5 healthy, 4 cirrhosis and 22 HCC samples, were served as the discovery datasets. These samples contained 22,853, 16,522, and 34,287 cells, respectively (Fig. 2A, Table S3). Next, we employed lineage specific markers to annotate these cells. We identified 8 major cell types, each of which exhibited enrichment preferences across different samples (Fig. 2B, Figs. S2A–C, Table S6, $p < 0.001$, Wilcoxon test). Furthermore, we subdivided these major cell types into 31 subpopulations in HCC samples and 37 subpopulations in healthy and cirrhosis samples (Figs. S2D–E).

Then, we applied SIMarker in HCC. Similarly, we also compared the performance of five clustering methods between HCC and cirrhosis. Both

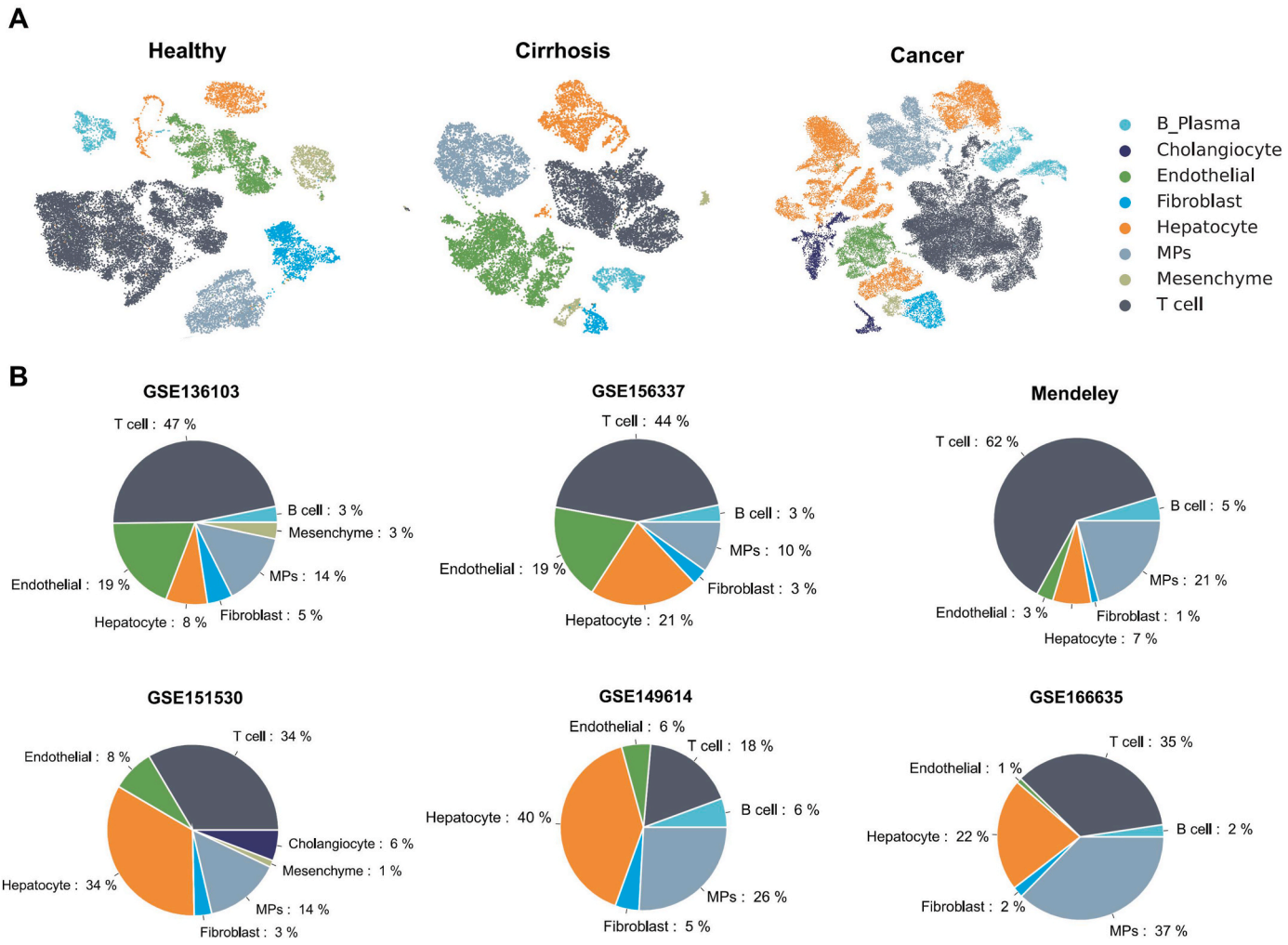

Fig. 2. Annotation and composition of cell subpopulations.
(A) UMAP visualization of cell compositions of healthy liver, cirrhosis and cancer in GSE151530. **(B)** Cell compositions of cell types in different datasets.

the KNN and Neural network revealed a high degree of similarity between *KNG1*$^+$ hepatocytes and *PGA5*$^+$ hepatocytes (Fig. 1C). Consequently, in the subsequent analysis of other cell types, we employ the KNN method to calculate similarity across different datasets.

We observed that in the discovery dataset, *PGA5*$^+$ hepatocytes within HCC samples exhibited higher degree of similarity compared to other subpopulations (Fig. 3A and B, Fig. S3). Additionally, they displayed more pronounced similarities with *KNG1*$^+$ hepatocytes in cirrhosis samples (Fig. 3C). The similar results were obtained in the four validation datasets (Fig. 3D). This finding was consistent with the results from rat model of drug-induced cirrhosis progression to HCC reported by Takuma T et al. (Table S2). They found that the gene dysregulation modules were comparable between these two samples, as well as growth signals and stress responses were more strongly shared in the hepatocytes [53]. In addition, we also discovered that *IGFBP4*$^+$ endothelials, *C1QC* + tumor-associated macrophages and CD4$^+$ central memory T cells, exhibited cirrhosis-like features (Fig. 4A–C).

3.3. Identifying and characterizing cirrhosis-like signatures in *PGA5*$^+$ hepatocytes

Results above suggested that hepatocyte subpopulations had the highest proportion of similarity between cirrhosis and HCC samples (Fig. 5A). We further explored the cirrhosis-like signatures (CLS) in *PGA5*$^+$ hepatocytes using cNMF method. We determined the most appropriate number of programs and focused on programs shared between *PGA5*$^+$ hepatocytes and *KNG1*$^+$ hepatocytes. 9 programs were obtained (Fig. 5B and C) and meta-programs 2 had the most similar molecular characteristics (Fig. 5D). Pathway enrichment analysis revealed that genes within Meta2 program were mainly involved in complement and coagulation cascades, tyrosine metabolism, and fat digestion and absorption (Fig. 5E). We calculated the Jaccard similarity based on specific markers from each gene program and hepatocyte subtypes, revealing a strong association between Meta2 and *KNG1*$^+$ hepatocytes as well as *PGA5*$^+$ hepatocytes (Fig. 5F). Besides, GSEA was performed to confirm that *KNG1*$^+$ signature gene enriched in the HCC-specific gene set compared with other hepatocytes of cirrhosis, likewise, *PGA5*$^+$ signature enriched in the cirrhosis-specific gene set compared with other hepatocytes of HCC (Fig. 5G). Furthermore, we collected two gene sets of cirrhosis related genes from DisGeNET dataset (1182 genes for C0023890, 919 genes for C1623038) [54]. We found 20 genes overlap between our cirrhosis-like genes and genes in C0023890 dataset (hypergeometric distribution test, $p = 4.09e^{-8}$, Fig. S4). For the C1623038 dataset, the number of overlapped genes is 8 (hypergeometric distribution test, $p = 0.00021$, Fig. S4). Several overlapped genes have been reported to be associated with cirrhosis. RBP4 and TTR involved in hormone and vitamin transportation to be altered in patients with cirrhosis [55]. The change in ApoA-I and ApoB mRNA level is associated with severe alcohol-induced cirrhosis [56]. ApoE genetic polymorphisms may also influence the progression of liver cirrhosis [57]. AZGP1 expression in HCC significantly associated with liver cirrhosis [58].

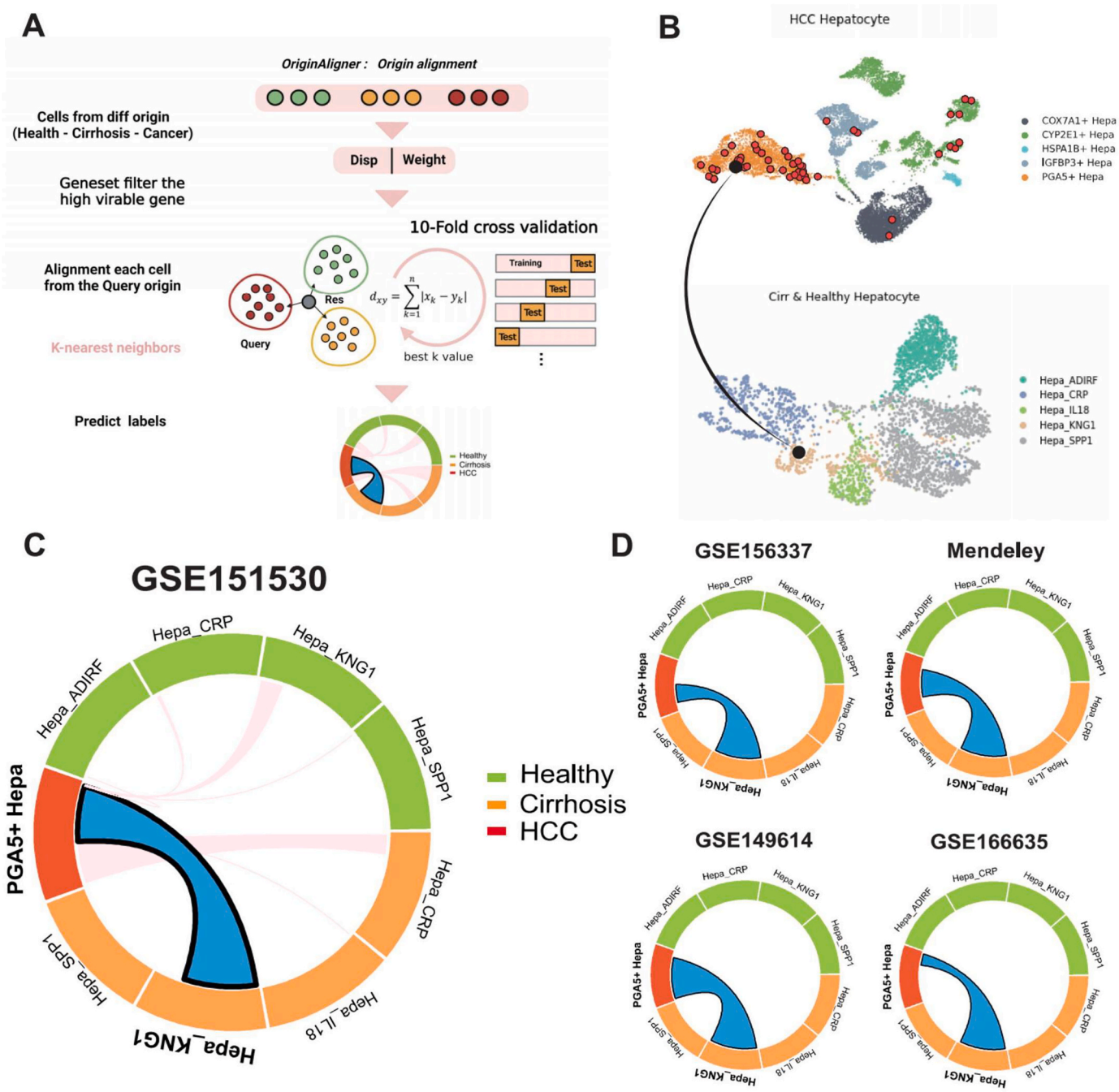

Fig. 3. Identification of cirrhosis-like cell subpopulations in hepatocytes.
(A) Cells from cirrhosis and healthy sample were aligned to cells from HCC samples (reference cells) according to their molecular similarities based on KNN. (**B**) UMAP visualization of the hepatocyte subpopulations within HCC and cirrhosis samples. The red points represent the proportion of cells from each subpopulation that can be mapped to cirrhosis samples. (**C**) Circle plot showing the molecular similarities of various subpopulations of hepatocytes in three tissues in GSE151530. **(D)** Circle plot showing the molecular similarities of various subpopulations of hepatocytes in three tissues in validation datasets (GSE156337, Mendeley, GSE149614 and GSE166635).

Then, we observed that certain regulators exhibited significant regulatory capabilities in both *KNG1*$^{+}$ and *PGA5*$^{+}$ hepatocytes (Fig. 5H). Previous findings revealed XBP1 promoted tumorigenesis through increased cholesterol biosynthesis in hepatocarcinoma [59]. Knocking out NR2F6 significantly inhibited HCC growth and metastasis [60]. MYC has been reported to control the hallmark expression programs responsible for cancer development [61]. HNF4A stimulates the EGFR-mediated proliferative response during liver cancer development [62]. E2F1 promoted HCC cell proliferation, migration and invasion [33]. Knockdown of CEBPD promotes cancer cell death [63].

We further inferred the abundances of the 37 subpopulations from the bulk transcriptomes to explore the prognostic value of cirrhosis-like subpopulations (Fig. 6A, Figs. S5–S8, see Method). The correlation between subpopulation abundances and survival status was calculated using Cox proportional hazards regression. After dimensionality reduction analysis of the abundance matrices from four datasets (GSE14520, GSE76427, GSE116174, and TCGA) (Fig. 6B), we observed that the potential cirrhosis-like subpopulations such as *PGA5*$^{+}$ hepatocytes, *IGFBP4*$^{+}$ endothelials, *C1QC* $^{+}$ tumor-associated macrophages, CD4^{+} central memory, and CD8^{+} effector T cells (Fig. 6C). Moreover, we

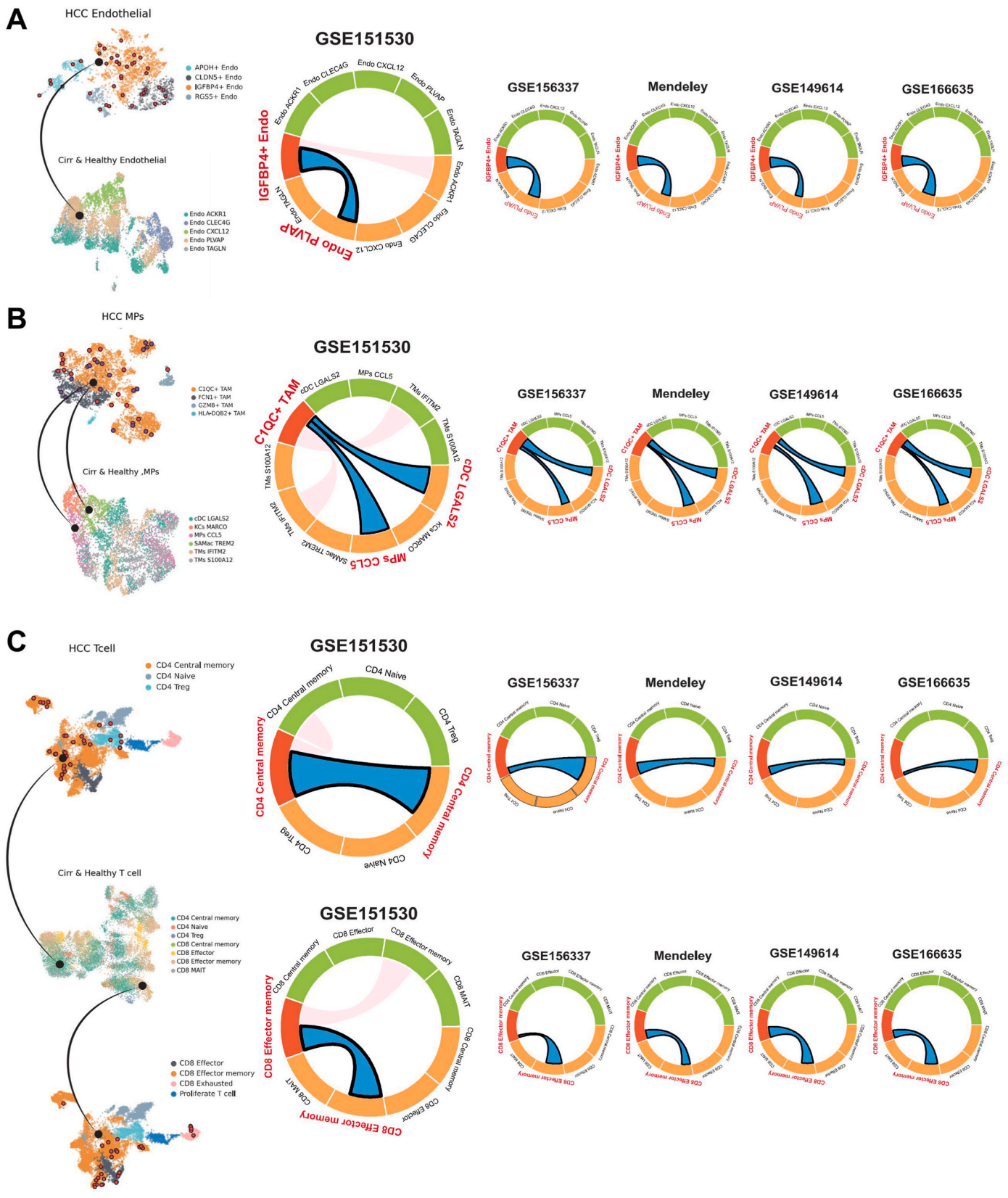

Fig. 4. Identification of cirrhosis-like cell subpopulations in other cell types.
(A–C) UMAP visualization of the Endothelial, MPs and T cell subpopulations within HCC and cirrhosis samples. The red points represent the proportion of cells from each subpopulation that can be mapped to cirrhosis samples. Circle plot showing the molecular similarities of various subpopulations of Endothelial, MPs and T cells in three tissues in GSE151530 and validation datasets.

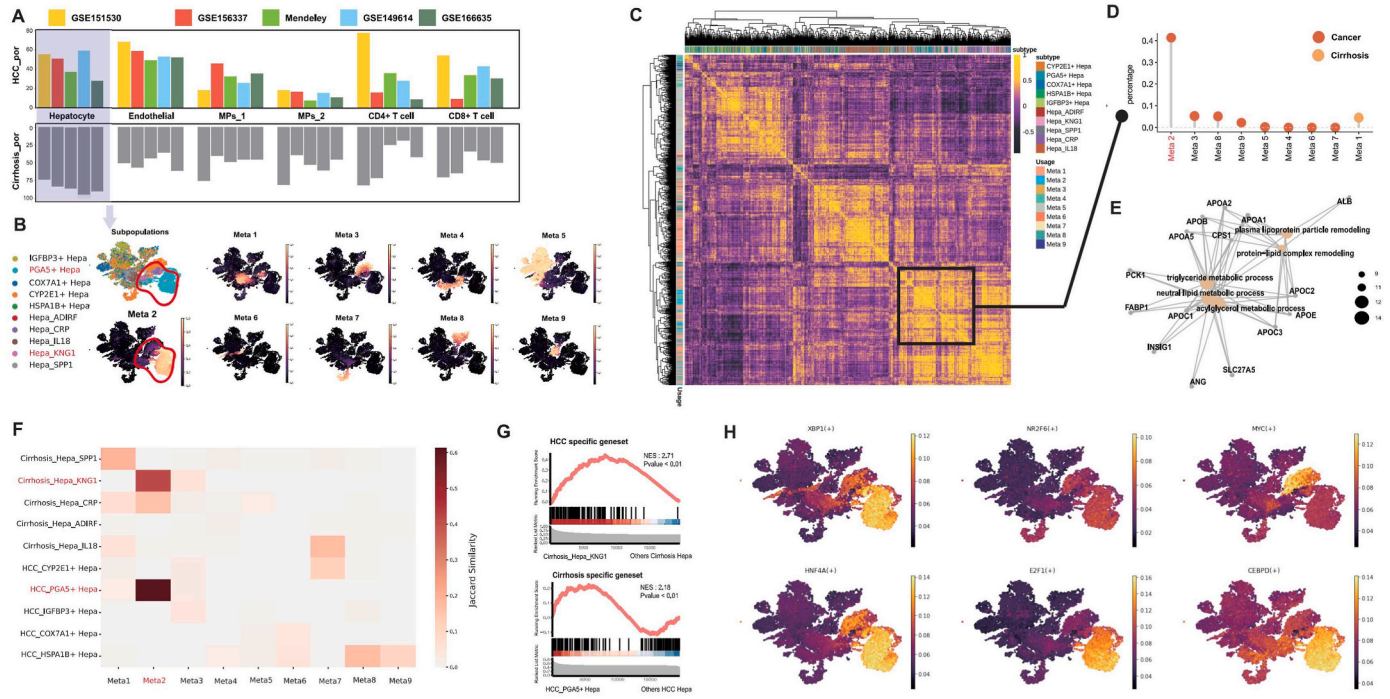

Fig. 5. Identification of cirrhosis-like signatures (CLS).
(**A**) Proportion of similarity between cirrhosis and HCC samples of different cell types in GSE151530 and validation datasets. **(B)** t-SNE of hepatocytes colored by subpopulations and meta-program 1–9. (**C**) Heatmap showing correlation of all meta-programs of cNMF and hepatocyte subpopulations of HCC and cirrhosis (k = 15). (**D**) Percentage of cells within different meta programs between cirrhosis and HCC. **(E)** GO pathway enrichment analysis of genes within Meta2 program. (**F**) Jaccard similarities of nine meta programs (x axis) with the signatures of ten subpopulations of hepatocyte (y axis). (**G**) GSEA analysis for HCC and cirrhosis-specific gene sets. Genes were ranked by logarithmic fold change in the mean expression values of *KNG1*$^{+}$ hepatocyte and *PGA5*$^{+}$ hepatocyte, respectively. **(H)** t-SNE visualization of the SCENIC-regulon activity of six regulons (XBP1, NR2F6, MYC, HNF4A, E2F1 and CEBPD) in hepatocyte subpopulations.

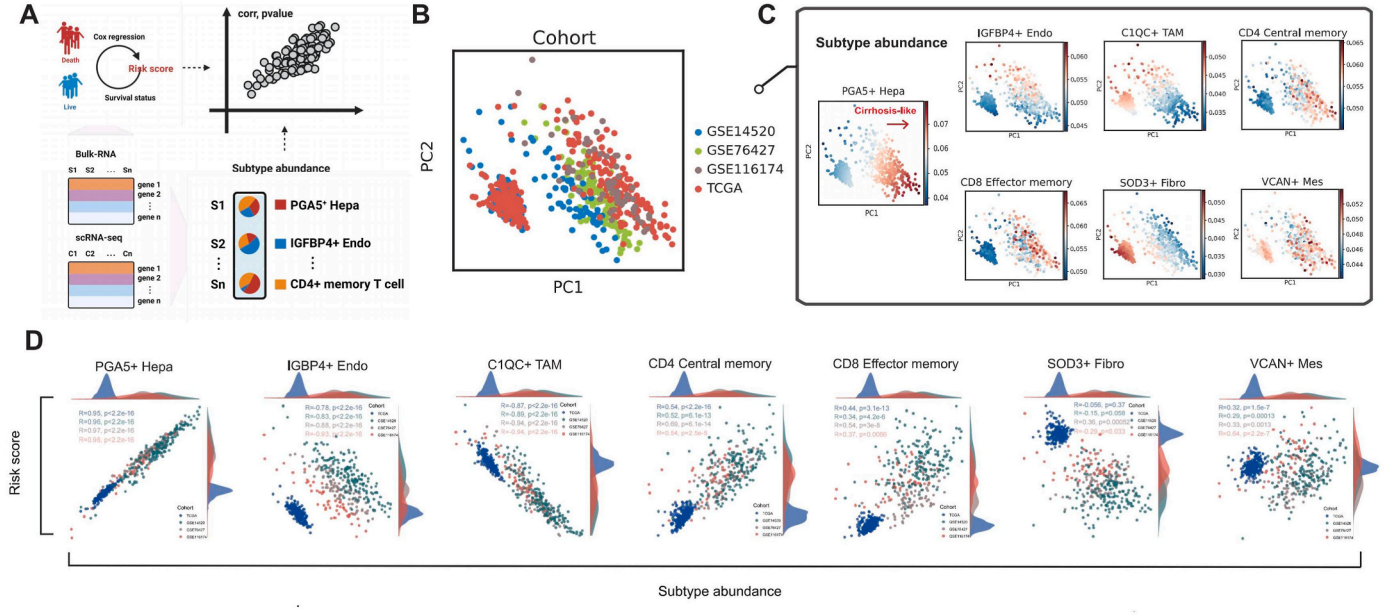

Fig. 6. The *PGA5*$^{+}$ Hepatocyte lead axis governs HCC poor prognosis.
(**A**) Deconvolution was used to construct the hierarchy of patients and demonstrate the prognostic value of cirrhosis-like subpopulations. (**B**) PCA of 572 patients with HCC from 4 cohorts. **(C)** PCA of 572 patients with HCC based on the compositions of their cellular hierarchy. (**D**) Correlation between a prognostic score trained by regularized cox regression using HCC subtypes abundances with the *PGA5*$^{+}$ hepatocyte lead proportion axis (PC1) within the 4 cohorts.

observed that higher abundances of these subpopulations, particularly *PGA5*$^{+}$ hepatocytes, were associated with a higher risk of prognosis (Fig. 6D).

3.4. Clinical applications of cirrhosis-like signatures

Taking the Cirrhosis-like signatures from *PGA5*$^{+}$ hepatocytes as input, we identified three kinds of individualized biomarkers based on

within-sample relative expression orderings of gene pairs (**see Method**). The parameters of each step in SIMarker were summarized in this Table S8.

3.4.1. *Construction of a risk signature for prediction of cirrhosis progression to HCC*

First, the 216 cirrhosis patients in GSE15654 cohort were divided into a training cohort (N = 152) and a validation cohort (N = 64). Next, 35 gene pairs were selected as diagnostic features according to the process described in Methods (Fig. 7A and B). Based on the set of gene pairs described above, we calculated the coefficient weights of each gene pair by XGBoost to construct a risk prediction model. In this model, we calculated the risk score of each patient in the training set, and patients with risk scores greater than the median were classified into the high-risk group (N = 57) and those less than the median were classified into the low-risk group (N = 95), while the validation set was also divided according to this threshold. Strikingly, patients in the high-risk group had a shorter time to HCC development than the low-risk group (P = 0.0032, Fig. 7C). Multivariate Cox regression analysis also showed that our constructed gene pair risk profile was an independent prognostic factor for HCC development (HR = 6.40, 95% CI: 1.82–22.55, p = 0.004, Fig. 7D). In addition, the area under the ROC curve (AUC) of risk features in the training set for predicting HCC progression at 3, 5, and 10 years was 0.83, 0.864, and 0.915, respectively. The AUC at 3, 5, and 10 years of validation dataset were 0.708, 0.802, and 0.8, respectively (Fig. 7E).

Finally, mapping the risk scores constructed by XGBoost back to the single-cell datasets of cirrhosis, we still observe that *KNG1*$^{+}$ hepatocytes (Fig. 7F) exhibit high risk of progression to hepatocellular carcinoma. Taken together, these suggests that the risk signatures obtained from cirrhosis-like gene pairs has good performance in predicting the progression of cirrhosis to HCC.

3.4.2. *Distinguish CoHCC and HCC in biopsy and surgical samples*

To improve the prognosis of HCC, it is important to develop biomarkers for the diagnosis of HCC at an early stage and distinguish HCC

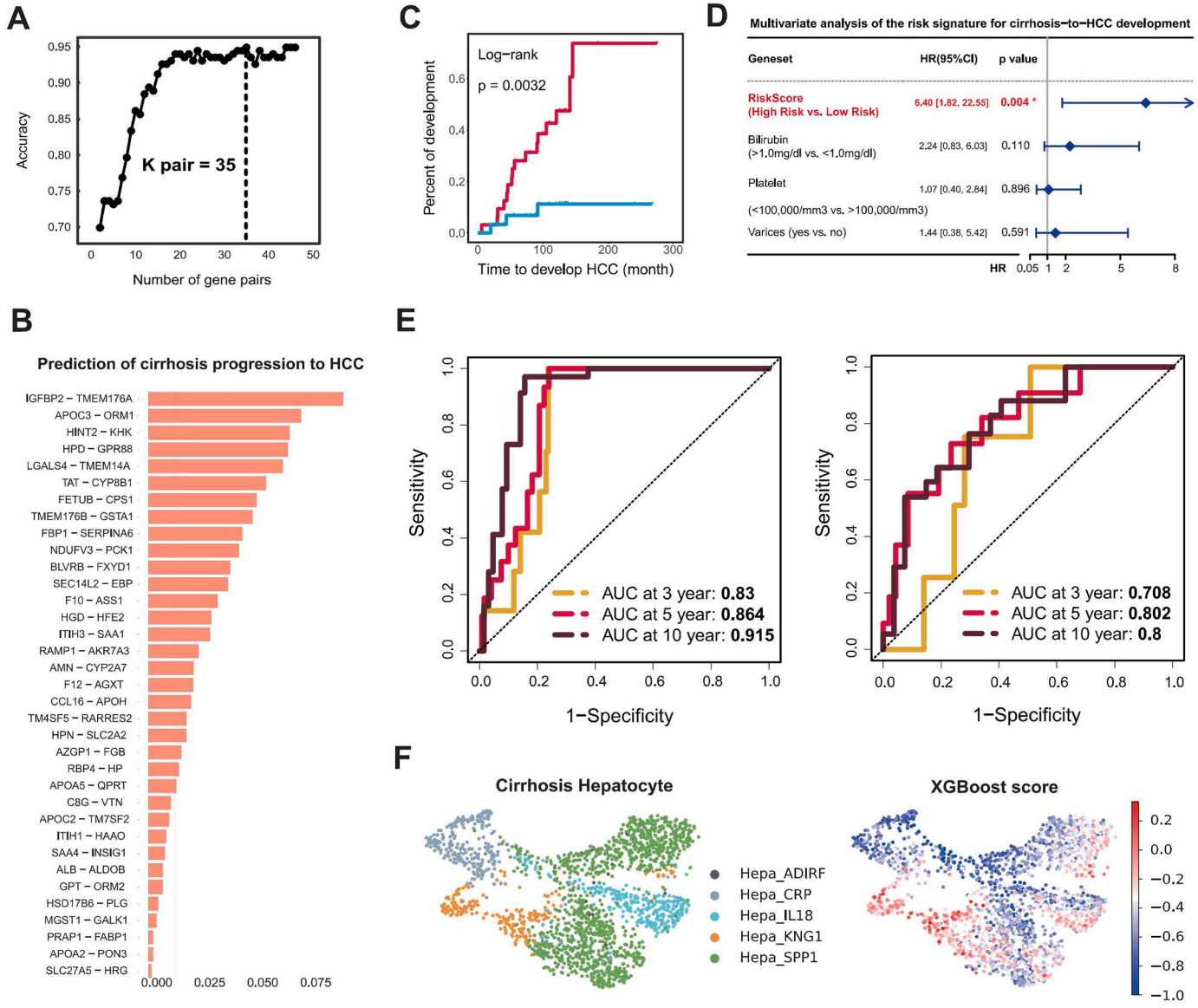

Fig. 7. Prediction of cirrhosis progression to HCC.
(**A**) Accuracy of the top-ranked gene pairs in the 35 gene pairs selected from the bulk transcriptome by k-TSP. (B) Risk score of gene pairs in XGBoost model. (**C**) Kaplan-Meier analysis of the HCC development time in patients with different risk scores. (**D**) Multivariate analysis of risk signatures for cirrhosis to HCC progression. (**E**) Time-dependent receiver characteristic operating curves. AUC of the risk scores for predicting the development of HCC in patients with cirrhosis at 3, 5 and 10 years. (F) UMAP of hepatocytes in cirrhosis based on subpopulations and XGBoost score.

samples from cirrhosis patients. The details of the datasets were described in Table S4. 37 gene pairs were selected as diagnostic features (Fig. 8A–B). Overall, 96.4% of samples in the biopsy-independent validation datasets were correctly classified (AUC = 0.95) (Fig. 8C). For the biopsy specimens in GSE64041, 100% of the 120 HCC and HCC adjacent samples were correctly classified to HCC. In GSE54236, 99.4% of the 161 HCC and CHCC samples were classified as HCC, while in GSE15654-V, 87.9% of 108 CoHCC samples were correctly classified (Fig. 8D).

Similarly, for the surgical resection samples, 93.3% of the seven data sets with 1572 samples in total were correctly classified (AUC = 0.93) (Fig. 8C). As shown in Fig. 8E, there were 268 HCC, 40 CHCC and 243 adjacent of HCC samples, of which 94.7% were correctly classified as HCC In GSE25097. In GSE36376, 99.8% of 240 HCC samples and 193 adjacent of HCC samples were classified to HCC. In GSE39791, 97.9% of 72 HCC samples and 72 adjacent of HCC samples were classified to HCC. These results suggest that these individualized biomarkers are robust for the diagnosis of samples obtained by samples from biopsy and surgical resection.

3.4.3. Prognosis of HCC patients with early stage

Next, we attempted to investigate the prognosis of HCC at early stage. Firstly, 178 HCC samples with early stage in the TCGA dataset as the training set , 75 TCGA HCC samples as the internal test set, and GSE14520, GSE76427, and GSE116174 serve as independent validation sets with 174, 53, and 90 samples, respectively. 34 gene pairs were selected as diagnostic features (Fig. 9A–B). The AUC for predicting survival status in the test dataset and validation datasets were 0.72, 0.71, 0.73 and 0.66, respectively (Fig. 9C). We could find that patients in the high-risk group had a significantly shorter survival time than the low-risk group in all four data sets (Fig. 9D). Multivariate Cox regression analysis also showed that the risk signatures, compared to those constructed from other hepatocyte subpopulation markers, was an independent prognostic factor (HR = 6.07, $p < 0.001$, Fig. 9E). Mapping the risk scores constructed by XGBoost back to the single-cell datasets of HCC, we still observe that *PGA5*$^{+}$ hepatocytes (Fig. 9F) exhibit high risk of progression to HCC.

Finally, we performed univariate Cox regression on all datasets for published models and observed that only our model was significantly associated with prognosis in all cohorts, indicating the stability of the

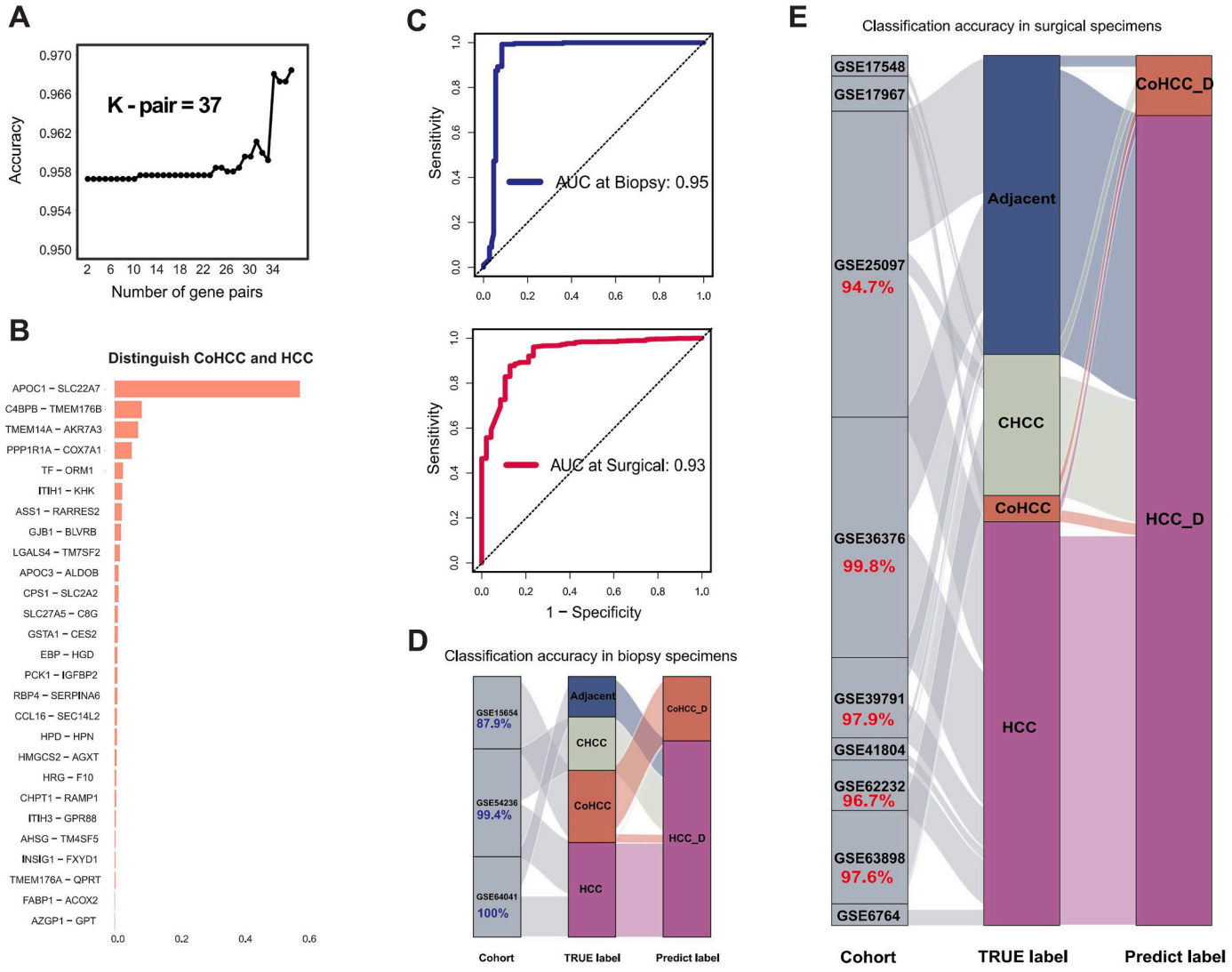

Fig. 8. Classification of CoHCC, CHCC, HCC, and HCC adjacent samples in biopsy and surgical specimens.
(**A**) Accuracy of the top-ranked gene pairs in the 37 gene pairs selected from the bulk transcriptome by k-TSP. (**B**) Risk score of gene pairs in XGBoost model. (**C**) The performance of the signature in the validation data sets from biopsy (top) and surgical resection (bottom). (**D**) The classification accuracy in biopsy specimens. The biopsy specimens included 141 HCC tissues and 108 cirrhosis tissues from non-HCC patients. (**E**) The classification accuracy in surgical specimens. The surgical resection specimens included 733 HCC tissues and 47 cirrhosis tissues from non-HCC patients.

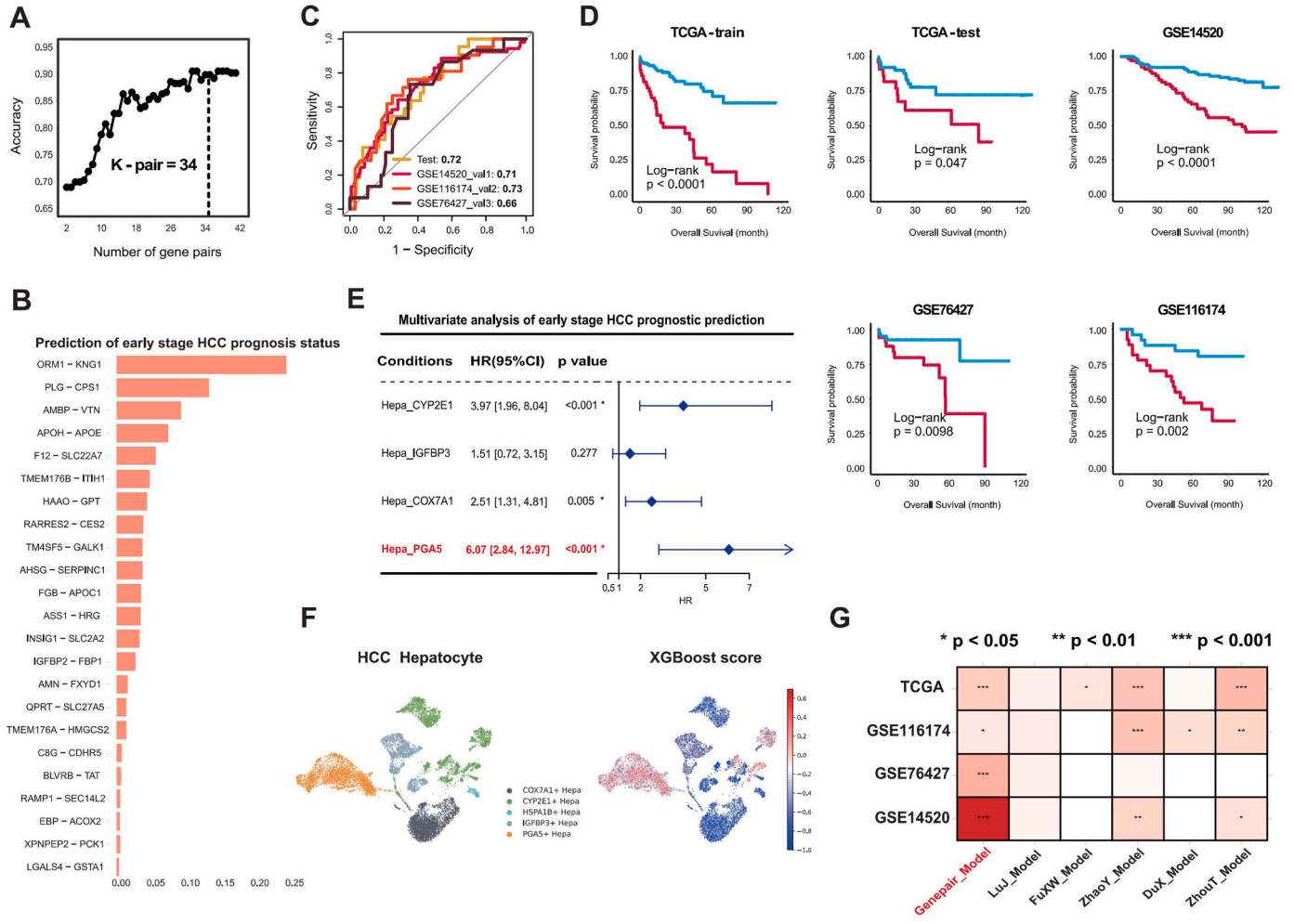

Fig. 9. Prognosis of HCC patients with early stage.
(**A**) Accuracy of the top-ranked gene pairs in the 34 gene pairs selected from the bulk transcriptome by k-TSP. **(B)** Risk score of gene pairs in XGBoost model. (**C**) The receiver characteristic operating curves. The information of test set and independent validation set were summarized in Table S4. (**D**) Kaplan-Meier curves for overall survival. Based on the Youden's index of the training set, each dataset is classified as high (red line) or low risk (blue line) group. (**E**) Multivariable analysis of cirrhosis-like gene pair model. (F) UMAP of hepatocytes in HCC based on subpopulations and XGBoost score. (**G**) Univariate Cox regression analysis of our cirrhosis-like gene pairs model and five published biomarkers in four cohorts (*p < 0.05; **p < 0.01; ***p < 0.001).

risk signature (Fig. 9G). These results highlighted the robustness of gene pair model constructed from CLS, demonstrating superior performance compared to nearly all other models across all datasets.

4. Discussion

scRNA-seq could help us understand the cell type heterogeneity within inter-tumor, intra-tumor, and precancerous tissue, as well as provide insights into the inflammatory and tumor microenvironment. A crucial challenge lies in characterizing similarity through the integration of single-cell transcriptomic data from different sources or periods. By aligning cellular subpopulations using KNN algorithm, we found remarkable similarity between *KNG1*$^{+}$ hepatocytes in cirrhosis and *PGA5*$^{+}$ hepatocytes in HCC. Furthermore, we proposed a hypothesis that similar characteristics of cirrhosis and HCC might provide new biomarkers for cancer diagnosis and prognosis. The gene pair models achieved higher accuracy compared with previous studies based on transcriptomes (Table 1) [18,20,64–68]. Compared to clinically used markers (AFP, AFP-L3, DCP, Table S9) [69], our model performed better than most of the other biomarkers with 88.9% of specificity and 68.4% of sensitivity. However, clinically used markers are usually proteins. Further analysis is needed to explore the performance of protein expression in the gene pair models.

5. Limitations

Several limitations should be mentioned. First, the cellular similarities deserve further validated in more scRNA-seq datasets of cirrhosis and HCC. The viral types (e.g. HBV and HCV), one of risk factors for HCC, should be considered in the cellular similarity identification. In addition, it has been reported that network-based biomarkers such as EdgeMarker [70], could be used to detect early signals for complex diseases. These methods could reveal essential mechanisms on disease initiation and progression at a network level [71]. More advanced methods and cohorts should be used to investigate the clinical value of cellular similarities in the future study. Second, the similar transcriptional programs should be further validated through spatial and imaging technologies. Third, we should notice that gain of cirrhosis-like signatures might not necessarily cause worse prognosis for the patients. The poor prognostic patients might undergo accumulation of different oncogenesis properties during malignant progression. For prediction of cirrhosis progression to HCC, only one public dataset was used in this study. More cohorts with short-term prediction (e.g. 3 and 5 years) are needed to collected to investigate the clinical value of cirrhosis-like signatures in the future study.

Table 1
Comparisons of individualized biomarkers with previous studies.

Model name	AUC or Accuracy	Type of signature (number)	Screening Method	Prediction model	Datasets segmentation
Application 1: Construction of a risk signature for prediction of cirrhosis progression to HCC					
Genepair_Model	training AUC: 0.83 (3 year)/0.86 (5 year)/0.92 (10 year) validation AUC: 0.71 (3 year)/0.80 (5 year)/ 0.80 (10 year)	gene pairs (35)	single-cell resolution similarity genes; k-TSP	XGBoost	152:64
Journal of oncology Model	training AUC: 0.77 (2 year)/0.91 (5 year)/0.86 (10 year) validation AUC: 0.83 (2 year)/0.75 (5 year)/ 0.66 (10 year)	genes (42)	DEG	Cox regression	108:108
Application 2: Distinguish CoHCC and HCC in biopsy and surgical samples					
Genepair_Model	biopsy accuracy: HCC (98.6%)/CHCC (100%)/HCC_adj (100%) surgical accuracy: HCC (99.4%)/CHCC (95.3%)/HCC_adj (100%)	gene pairs (35)	single-cell resolution similarity genes; k-TSP	XGBoost	634 : 389 : 1572
Liver interactional Model	biopsy accuracy: HCC (92.6%)/CHCC (77.5%)/HCC_adj (100%) surgical accuracy: HCC (99.7%)/CHCC (96%)/HCC_adj (95.9%)	gene pairs (19)	DEG	Majority voting rule	634 : 389 : 1572
Application 3: Prognosis of HCC patients with early stage					
Genepair_Model	TCGA: 0.83 GSE116174: 0.73 GSE76427: 0.66 GSE14520: 0.71	gene pairs (34)	single-cell resolution similarity genes; k-TSP	XGBoost	**255:53:174:90**
LuJ_Model	TCGA: 0.56 GSE116174: 0.61 GSE76427: 0.69 GSE14520: 0.55	genes (3)	DEG	LASSO regression	**255:53:174:90**
FuXW_Model	TCGA: 0.58 GSE116174: 0.55 GSE76427: 0.61 GSE14520: 0.53	genes (3)	DEG	LASSO regression	**255:53:174:90**
ZhaoY_Model	TCGA: 0.62 GSE116174: 0.65 GSE76427: 0.62 GSE14520: 0.61	genes (9)	DEG	LASSO regression	**255:53:174:90**
DuX_Model	TCGA: 0.52 GSE116174: 0.61 GSE76427: 0.58 GSE14520: 0.52	genes (7)	DEG	Cox regression	**255:53:174:90**
ZhouT_Model	TCGA: 0.63 GSE116174: 0.53 GSE76427: 0.54 GSE14520: 0.61	genes (10)	DEG	LASSO regression	**255:53:174:90**

6. Conclusions

In summary, we provide a systematic analysis workflow, named SIMarker, to quantify similarities between HCC and cirrhosis at single-cell resolution. Moreover, robust individualized signatures for early diagnosis and prognosis of HCC based on within-sample REOs were developed. Our work opens avenues for the exploration of similarity in other types of cancers and diseases based on single cell transcriptomes.

Funding

This work was supported by National Natural Science Foundation of China (Grant No. 82002529 to M.T., 32070635, 32370586 to J.H.), the National Key Research and Development Program of China (2023YFC2508100 to J.H.), the Fundamental Research Funds for the Central Universities (20720230068 to J.H.).

Data availability

All scRNA-seq and bulk transcriptome datasets were summarized in Tables S3–S4.

CRediT authorship contribution statement

Mengsha Tong: Supervision, Writing – original draft, Writing – review & editing. **Shijie Luo:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Lin Gu:** Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft. **Xinkang Wang:** Validation, Visualization, Writing – original draft. **Zheyang Zhang:** Investigation, Supervision. **Chenyu Liang:** Formal analysis, Supervision. **Huaqiang Huang:** Conceptualization, Data curation, Supervision. **Yuxiang Lin:** Supervision. **Jialiang Huang:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108113.

References

[1] P. Zhang, et al., Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer, Cell Rep. 30 (12) (2020) 4317.
[2] Z. Wang, et al., Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing, Nat. Commun. 12 (1) (2021) 6500.
[3] W.R. Becker, et al., Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer, Nat. Genet. 54 (7) (2022) 985–995.
[4] L. Sun, et al., Single-cell and spatial dissection of precancerous lesions underlying the initiation process of oral squamous cell carcinoma, Cell Discov 9 (1) (2023) 28.
[5] J.A. Davila, et al., Diabetes increases the risk of hepatocellular carcinoma in the United States: a population based case control study, Gut 54 (4) (2005) 533–539.
[6] G. Chang, et al., Hypoexpression and epigenetic regulation of candidate tumor suppressor gene CADM-2 in human prostate cancer, Clin. Cancer Res. 16 (22) (2010) 5390–5401.
[7] P. Gines, et al., Liver cirrhosis, Lancet 398 (10308) (2021) 1359–1376.
[8] Y. Hoshida, et al., Pathogenesis and prevention of hepatitis C virus-induced hepatocellular carcinoma, J. Hepatol. 61 (1 Suppl) (2014) S79–S90.
[9] Y. Hoshida, et al., Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis, Gastroenterology 144 (5) (2013) 1024–1030.
[10] D.Q. Huang, et al., Global epidemiology of alcohol-associated cirrhosis and HCC: trends, projections and risk factors, Nat. Rev. Gastroenterol. Hepatol. 20 (1) (2023) 37–49.
[11] N. Kawada, et al., Hepatocellular carcinoma arising from non-cirrhotic nonalcoholic steatohepatitis, J. Gastroenterol. 44 (12) (2009) 1190.
[12] M.H. Lee, et al., Development and validation of a clinical scoring system for predicting risk of HCC in asymptomatic individuals seropositive for anti-HCV antibodies, PLoS One 9 (5) (2014) e94760.
[13] A.G. Singal, et al., HCC surveillance improves early detection, curative treatment receipt, and survival in patients with cirrhosis: a meta-analysis, J. Hepatol. 77 (1) (2022) 128–139.
[14] V.W. Wong, et al., Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers, J. Clin. Oncol. 28 (10) (2010) 1660–1665.
[15] X.F. Xu, et al., Risk factors, patterns, and outcomes of late recurrence after liver resection for hepatocellular carcinoma: a multicenter study from China, JAMA Surg 154 (3) (2019) 209–217.
[16] M.F. Yuen, et al., Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B, J. Hepatol. 50 (1) (2009) 80–88.
[17] Hepatocellular carcinoma, Nat. Rev. Dis. Prim. 7 (1) (2021) 7.
[18] G. Ning, et al., Identification of new biomarker for prediction of hepatocellular carcinoma development in early-stage cirrhosis patients, JAMA Oncol. 2021 (2021) 9949492.
[19] C.H. Jiang, et al., Bioinformatics-based screening of key genes for transformation of liver cirrhosis to hepatocellular carcinoma, J. Transl. Med. 18 (1) (2020) 40.
[20] L. Ao, et al., A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings, Liver Int. 38 (10) (2018) 1812–1819.
[21] L. Qi, et al., Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer, Briefings Bioinf. 17 (2) (2016), 233-42.
[22] R. Wang, et al., Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia, Briefings Bioinf. 23 (2) (2022).
[23] M. Tong, et al., Prioritizing prognostic-associated subpopulations and individualized recurrence risk signatures from single-cell transcriptomes of colorectal cancer, Briefings Bioinf. 24 (3) (2023).
[24] E. Kim, et al., Promotion of growth factor signaling as a critical function of β-catenin during HCC progression, Nat. Commun. 10 (1) (2019) 1909.
[25] F. Li, et al., LncRNA MNX1-AS1 promotes progression of intrahepatic cholangiocarcinoma through the MNX1/Hippo axis, Cell Death Dis. 11 (10) (2020) 894.
[26] S. Lu, et al., NNMT promotes the progression of intrahepatic cholangiocarcinoma by regulating aerobic glycolysis via the EGFR-STAT3 axis, Oncogenesis 11 (1) (2022) 39.
[27] R. Moreno Traspas, et al., Loss of FOCAD, operating via the SKI messenger RNA surveillance pathway, causes a pediatric syndrome with liver cirrhosis, Nat. Genet. 54 (8) (2022) 1214–1226.
[28] S. Song, et al., EGFR/MET promotes hepatocellular carcinoma metastasis by stabilizing tumor cells and resisting to RTKs inhibitors in circulating tumor microemboli, Cell Death Dis. 13 (4) (2022) 351.
[29] Q. Su, et al., Sanguinarine inhibits epithelial-mesenchymal transition via targeting HIF-1α/TGF-β feed-forward loop in hepatocellular carcinoma, Cell Death Dis. 10 (12) (2019) 939.
[30] J. Sun, et al., Long noncoding RNA SNHG1 silencing accelerates hepatocyte-like cell differentiation of bone marrow-derived mesenchymal stem cells to alleviate cirrhosis via the microRNA-15a/SMURF1/UVRAG axis, Cell Death Dis. 8 (1) (2022) 77.
[31] T. Yamanaka, et al., Nintedanib inhibits intrahepatic cholangiocarcinoma aggressiveness via suppression of cytokines extracted from activated cancer-associated fibroblasts, Br. J. Cancer 122 (7) (2020) 986–994.
[32] Y.M. Yang, et al., Hyaluronan synthase 2-mediated hyaluronan production mediates Notch1 activation and liver fibrosis, Sci. Transl. Med. 11 (496) (2019).
[33] Y. Yu, et al., E2F1 mediated DDX11 transcriptional activation promotes hepatocellular carcinoma progression through PI3K/AKT/mTOR pathway, Cell Death Dis. 11 (4) (2020) 273.
[34] Y. Liu, et al., Identification of a tumour immune barrier in the HCC microenvironment that determines the efficacy of immunotherapy, J. Hepatol. 78 (4) (2023) 770–782.
[35] Y. Lu, et al., A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma, Nat. Commun. 13 (1) (2022) 4594.
[36] Y. Meng, et al., A TNFR2-hnRNPK Axis promotes primary liver cancer development via activation of YAP signaling in hepatic progenitor cells, Cancer Res. 81 (11) (2021) 3036–3050.
[37] A. Sharma, et al., Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma, Cell 183 (2) (2020) 377–394 e21.
[38] P. Zhang, et al., Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer, Cell Rep. 27 (6) (2019) 1934–1947. e5.
[39] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (1) (2018) 15.
[40] L. Ma, et al., Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma, J. Hepatol. 75 (6) (2021) 1397–1408.
[41] A. Sharma, et al., Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma, Cell 183 (2) (2020) 377–394.e21.
[42] Y. Sun, et al., Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma, Cell 184 (2) (2021) 404–421.e16.
[43] R. Xue, et al., Liver tumour immune microenvironment subtypes and neutrophil heterogeneity, Nature 612 (7938) (2022) 141–147.
[44] Q. Zhang, et al., Landscape and dynamics of single immune cells in hepatocellular carcinoma, Cell 179 (4) (2019) 829–845.e20.
[45] R.A. Irizarry, et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2) (2003) 249–264.
[46] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550.
[47] Y. Liu, et al., Immune phenotypic linkage between colorectal cancer and liver metastasis, Cancer Cell 40 (4) (2022) 424–437 e5.
[48] D. Kotliar, et al., Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Single-Cell RNA-Seq, vol. 8, Elife, 2019.
[49] S. Aibar, et al., SCENIC: single-cell regulatory network inference and clustering, Nat. Methods 14 (11) (2017) 1083–1086.
[50] S. Hänzelmann, R. Castelo, J. Guinney, GSVA: gene set variation analysis for microarray and RNA-seq data, BMC Bioinf. 14 (2013) 7.
[51] A.C. Tan, et al., Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (20) (2005) 3896–3904.
[52] B. Afsari, et al., switchBox: an R package for k-Top Scoring Pairs classifier development, Bioinformatics 31 (2) (2015) 273–274.
[53] S. Nakagawa, et al., Molecular liver cancer prevention in cirrhosis by organ transcriptome analysis and lysophosphatidic acid pathway inhibition, Cancer Cell 30 (6) (2016) 879–890.
[54] A. Bauer-Mehren, et al., DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks, Bioinformatics 26 (22) (2010) 2924–2926.
[55] L. Niu, et al., Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease, Mol. Syst. Biol. 15 (3) (2019) e8793.
[56] P. Mathurin, et al., Quantification of apolipoprotein A-I and B messenger RNA in heavy drinkers according to liver disease, Hepatology 23 (1) (1996) 44–51.
[57] J.C.R. Nascimento, et al., Impact of apolipoprotein E genetic polymorphisms on liver disease: an essential review, Ann. Hepatol. 19 (1) (2020) 24–30.
[58] Y. Huang, et al., Decreased expression of zinc-alpha2-glycoprotein in hepatocellular carcinoma associates with poor prognosis, J. Transl. Med. 10 (2012) 106.
[59] M. Wei, et al., Unspliced XBP1 contributes to cholesterol biosynthesis and tumorigenesis by stabilizing SREBP2 in hepatocellular carcinoma, Cell. Mol. Life Sci. 79 (9) (2022) 472.
[60] L. Wang, et al., Circular RNA circRHOT1 promotes hepatocellular carcinoma progression by initiation of NR2F6 expression, Mol. Cancer 18 (1) (2019) 119.
[61] H. Wang, et al., TAZ is indispensable for c-MYC-induced hepatocarcinogenesis, J. Hepatol. 76 (1) (2022) 123–134.
[62] X. Zhang, et al., Ferroptosis is governed by differential regulation of transcription in liver cancer, Redox Biol. 24 (2019) 101211.
[63] X. Sun, et al., Dominant-negative ATF5 compromises cancer cell survival by targeting CEBPB and CEBPD, Mol. Cancer Res. 18 (2) (2020) 216–228.
[64] X. Du, Y. Zhang, Integrated analysis of immunity- and ferroptosis-related biomarker signatures to improve the prognosis prediction of hepatocellular carcinoma, Front. Genet. 11 (2020) 614888.
[65] X.W. Fu, C.Q. Song, Identification and validation of pyroptosis-related gene signature to predict prognosis and reveal immune infiltration in hepatocellular carcinoma, Front. Cell Dev. Biol. 9 (2021) 748039.
[66] J. Lu, et al., A novel prognostic model based on single-cell RNA sequencing data for hepatocellular carcinoma, Cancer Cell Int. 22 (1) (2022) 38.

[67] Y. Zhao, et al., Identification and validation of a nine-gene amino acid metabolism-related risk signature in HCC, Front. Cell Dev. Biol. 9 (2021) 731790.
[68] T. Zhou, et al., A novel ten-gene signature predicting prognosis in hepatocellular carcinoma, Front. Cell Dev. Biol. 8 (2020) 629.
[69] H. Pinto Marques, et al., Emerging biomarkers in HCC patients: current status, Int. J. Surg. 82s (2020) 70–76.
[70] W. Zhang, T. Zeng, L. Chen, EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers, J. Theor. Biol. 362 (2014) 35–43.
[71] T. Zeng, et al., Network biomarkers reveal dysfunctional gene regulations during disease progression, FEBS J. 280 (22) (2013) 5682–5695.