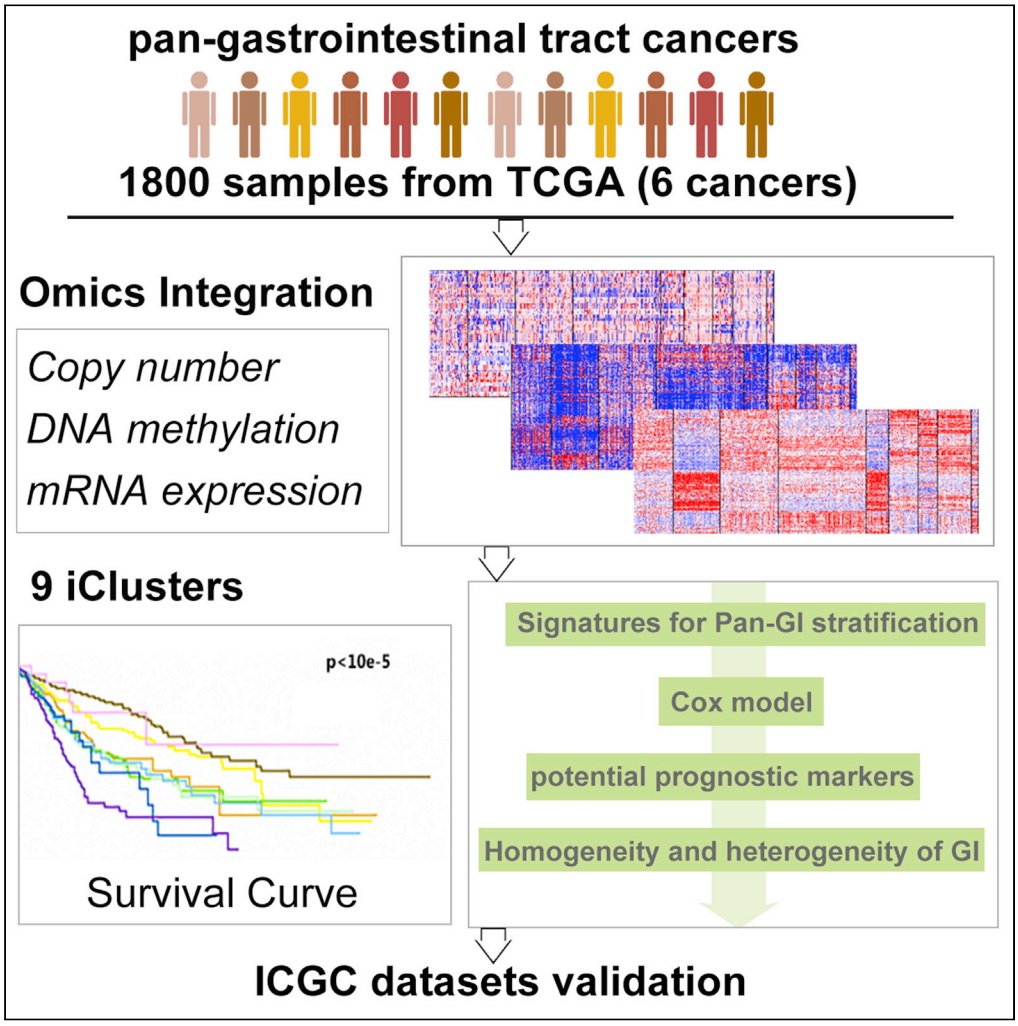
iScience

CellPress
OPEN ACCESS

Article

Integrative omics analysis reveals effective stratification and potential prognosis markers of pan-gastrointestinal cancers



Huiting Jiangzhou, Hang Zhang, Renliang Sun, ..., Zhiqiang Li, Yongyong Shi, Zhuo Wang

shiyongyong@gmail.com (Y.S.)
zhuowang@sjtu.edu.cn (Z.W.)

Highlights

Pan-cancer multi-omics strategy reveals homogeneity and heterogeneity of pan-GI cancers

Identify 9 iclusters with significantly different survival and molecular features

Potential prognostic markers have prominent power supported by concordance index

Jiangzhou et al., iScience *24*, 102824
August 20, 2021
https://doi.org/10.1016/j.isci.2021.102824

iScience

CellPress
OPEN ACCESS

Article

Integrative omics analysis reveals effective stratification and potential prognosis markers of pan-gastrointestinal cancers

Huiting Jiangzhou,[1] Hang Zhang,[1] Renliang Sun,[1] Aamir Fahira,[1] Ke Wang,[1] Zhiqiang Li,[1,2] Yongyong Shi,[1,2,*] and Zhuo Wang[1,3,*]

SUMMARY

Gastrointestinal (GI) tract cancers are the most common malignant cancers with high mortality rate. Pan-cancer multi-omics data fusion provides a powerful strategy to examine commonalities and differences among various cancer types and benefits for the identification of pan-cancer drug targets. Herein, we conducted an integrative omics analysis on The Cancer Genome Atlas pan-GI samples including six carcinomas and stratified into 9 clusters, i.e. 5 single-type-dominant clusters and 4 mixed clusters, the clustering reveals the molecular features of different subtypes, other than the organ and cell-of-origin classifications. Especially the mixed clusters revealed the homogeneity of pan-GI cancers. We demonstrated that the prognosis differences among pan-GI subtypes based on multi-omics integration are more significant than clustering by single-omics. The potential prognostic markers for pan-GI stratification were identified by proportional hazards model, such as *PSCA* (for colorectal and stomach cancer) and *PPP1CB* (for liver and pancreatic cancer), which have prominent prognostic power supported by high concordance index.

INTRODUCTION

Gastrointestinal (GI) cancers, including colorectal, gastric, esophageal, pancreatic, and hepatocellular cancer, are the most common human malignancies. Especially colorectal, gastric, and esophageal cancers ranked among the top 10 with a high global mortality rate (Bray et al., 2018). The GI tumors are homogeneous to some extent; almost all GI cancers tend to have subtypes of mesenchymal gene expression, immune infiltration, or metabolism (Bijlsma et al., 2017). On the other hand, cancer is a highly heterogeneous disease, different morphological and phenotypic features of different tumor regions resulting from the heterogeneity often cause drug resistance, fosters lethal outcomes, and the therapeutic failure (McGranahan and Swanton, 2017; Sun et al., 2020). Patients with the same type of cancer undergoing the same treatment have been shown to experience different prognosis. Pan-cancer research is an important strategy to study the homogeneity and heterogeneity among different tumors. Pan-cancer analysis of the molecular subtyping of different tumors through multi-omics data integration explores the underlying tumor mechanism and deepens our understanding of the process of tumor growth and lays a strong foundation for precision medicine.

With the development of high-throughput technologies, large amounts of molecular data have been generated and gathered. The Cancer Genome Atlas (TCGA) offers an incentive for pan-cancer research with abundant genome data sets (Hoadley et al., 2014). Some accomplishments were made by the individual omics platform in the pan-cancer study. For example, a previous study classified breast cancer into three subtypes through the gene mutation spectrum, and the results corresponded to clinical information (Vural et al., 2016). Moreover, Reis Filho et al. (Reis-Filho and Pusztai, 2011) used an expression profile for breast cancer via sample classification and obtained 7 subtypes of the tumor having a certain corresponding relationship with clinical information. At the methylation level, Yagi et al. (Yagi et al., 2010) obtained three types of colorectal cancer in which the high-methylation group was consistent with MSI high type and BRAF mutation type, the intermediate-methylation type was highly correlated with KRAS mutation.

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Centre for Brain Science, Shanghai Jiao Tong University, Shanghai 200030, China

[2]Affiliated Hospital of Qingdao University & Biomedical Sciences Institute of Qingdao University, Qingdao University, Qingdao 266003, China

[3]Lead contact

*Correspondence: shiyongyong@gmail.com (Y.S.), zhuowang@sjtu.edu.cn (Z.W.)

https://doi.org/10.1016/j.isci.2021.102824

Using multi-omics data acquiring from the same set of samples has the potential capacity to explore more accurate molecular subtypes with different survival than examining by one single-omics data (Rappoport and Shamir, 2018). To date, several studies have been conducted to stratify TCGA cancer samples using a multi-omics data integration analysis. For example, a previous study found 1954 pan-urologic cancer (kidney, testes, bladder, and prostate) that were typically classified into 9 major molecular subtypes based on DNA copy alteration, DNA methylation, mRNA expression, miRNA expression, and protein expression, whereas these various subtypes have different molecular signatures, such as prostate cases contain *ERG* fusion or *PTEN* loss were primarily associated with one specific subtype, and cases involving *FOXA1* or *SPOP* mutation are predominantly linked to another subtype (Chen et al., 2017). Another representative study is that the TCGA team conducted integrated cluster analysis on analyzing 10,000 samples from 33 cancer types by multi-omics integration including somatic copy number alterations, mRNA expression, and DNA methylation. Two-thirds of the samples were classified into 28 subtypes and some homogeneity and heterogeneity of tumors were found in the clustering results, for example, the group of tumors with four main cell or organ systems (pan-GI tract, pan-gynecology, pan-squamous, and pan-kidney) may indicate different types of tumors. The patients in each molecular subtype of the pan-cancer clustering have common molecular features and pathway characteristics. Moreover, the samples with similar tumor types, histopathology, and organ systems have obvious heterogeneity among different subtypes, which indicates that the histologic and pathologic criteria cannot completely determine the tumors subtypes based on molecular biology (Hoadley et al., 2018).

The study of TCGA team revealed global characteristics of 33 types of tumors, but further detail study on GI cancers is still required for precision stratification and treatment of such particular cancer types. Therefore, this research aimed to study the homogeneity and heterogeneity of the six forms of pan-GI cancer by using samples from the TCGA database, including esophageal carcinoma [ESCA], stomach adenocarcinoma [STAD], colon adenocarcinoma [COAD], rectum adenocarcinoma [READ], liver hepatocellular carcinoma [LIHC], and pancreatic adenocarcinoma [PAAD]. TCGA database contains 1801 samples of pan-GI cancers that have the complete data across three platforms including aneuploidy, DNA methylation, and mRNA. Using the unsupervised clustering process, we incorporated molecular data and categorized primary pan-GI tumor samples into 9 subtypes which were correlated with overall survival for each cancer. We found that the classifications were closely related to the molecular signaling pathway by enrichment analysis. Next, we identified the essential genes associated with stratification and prognosis that could become potential biomarkers. Finally, we compared the results of classification of the pan-GI samples from both TCGA and International Cancer Genome Consortium (ICGC) databases, and the similar results further confirmed our conclusion.

RESULTS

Clustering of pan-GI cancers by individual omics platform

In order to examine the CNV profile of the pan-GI cancers, we used Genomic Identification of Significant Targets in Cancer (GISTIC) 2.0 (Mermel et al., 2011) and performed hierarchical clustering on alterations score to identify 8 groups, as shown in Figure S1A. LIHC samples were separated into two classes, i.e. class3 and class7, where the class7 mainly contains copy number gains in 8q. Similarly, the ESCA samples were mainly separated into class2 and class6. The samples of ESCA in class2 contain copy number gains in 12p and class6 contains copy number gains in 11q. The COAD-READ (COAD and READ) samples were mainly divided into class1 and class4, both of which have deletions at 8p, 18q, and 17p, and significant copy number repeat variation at 20q and 13q. While the class 8 is a confounding category that contains samples of ESCA, STAD, and COAD-READ, which has significant copy number repeat variation in 17q.

Similarly, we performed unsupervised clustering on 1725 samples by the mRNA expression profile of 1037 selected genes and identified 5 groups. The Figure S1B demonstrated that the five groups are matched with the corresponding cancer types in high consistency, and LIHC exhibits different mRNA profiles from other GI cancers.

For the DNA methylation level, the unsupervised clustering of 1745 pan-GI samples by 875 selected CpG sites identified 8 subgroups. As shown in Figure S1C, PAAD samples were divided into class3 and class5. COAD-READ samples are mainly divided into class2, class3, class4. Almost all LIHC samples were in class1 and ESCA samples were in class5. STAD is heterogeneous and divided into class2, class3, class4, and class5. We found that the GI cancers exhibit higher homogeneity at the methylation level compared with the mRNA level.

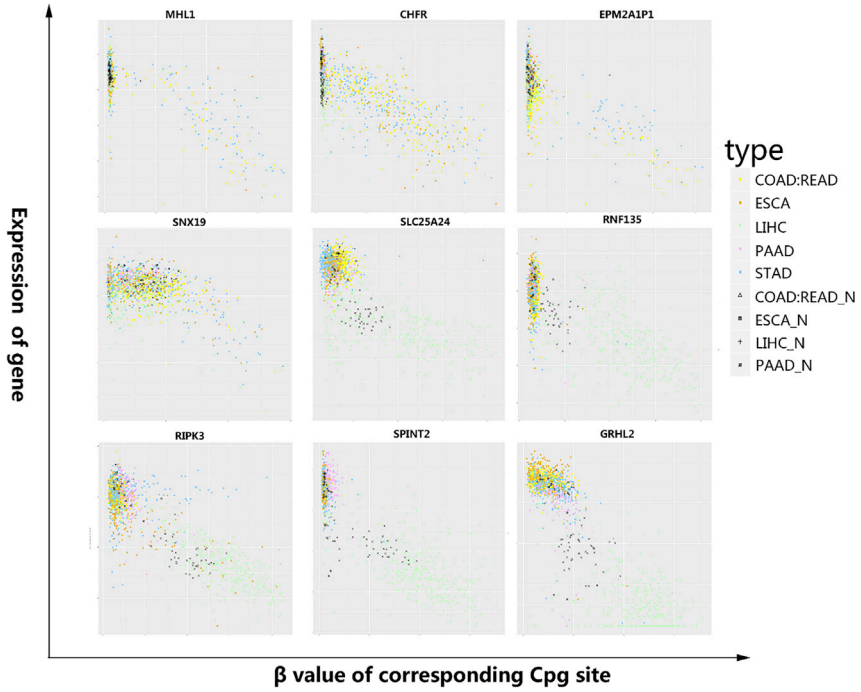


Figure 1. Scatterplots of representative cpg sites in gene promoters

The figure shows the relationship between the methylation level and the expression level of 9 genes in all gastrointestinal cancer samples. The abscissa represents the β value of methylation sites, and the ordinate represents the gene expression. Each dot in the figure represents a sample, including 1449 samples of primary pan-GI tumor represented by dots of different colors and 73 samples of precancerous tissue represented by black dots.

Generally speaking, each individual omics platform exhibited different patterns of pan-GI samples, the samples of the same tumor that divided into different clusters showed certain heterogeneity. We found both mRNA expression level and DNA methylation level revealed LIHC was most divergent from other GI tumors. But the clustering results by any individual omics platform could not well correspond to patient survival; therefore, we need a better clustering for personalized prognosis and treatment on GI cancer samples by multi-omics integration.

Identification of silenced genes

The silencing of some genes has a great influence on the occurrence and progression of tumor. We integrated gene expression and methylation data to identify the epigenetically silenced genes associated with pan-GI. Overall 9 genes were found including *MLH1, EPM2AIP1, SNX19, SLC25A24, RIPK3, RNF135, CHFR, SPINT2, and GRHL2*. Figure 1 shows the silent condition of 9 genes in different types of cancers.

The two silenced genes, *EPM2AIP1* and *MLH1,* are both regulated by the same promoter region and have previously been reported to be silent in colorectal cancer and gastric cancer (Neumeyer et al., 2019; Oue et al., 2019; Suter et al., 2004), concisely, the *MLH1* is a tumor suppressor gene that encodes a mismatch repair protein which is part of the DNA mismatch repair system (Esteller et al., 1998) and the function of the protein encoded by *EPM2AIP1* is currently unclear. Moreover, consistent with previous results, we observed epigenetic silencing of *MLH1* and *EPM2AIP1* in the STAD (19.7%; 14%) and COAD-READ (14%; 9.9%). Similarly, *CHFR* encodes a protein E3 ubiquitin ligase that maintains the mitotic checkpoint and directs cells into mitosis, which may play a key role in cell cycle progression and tumorigenesis (Scolnick and Halazonetis, 2000). Moreover, *CHFR* was observed to be silenced in gastric cancer and colorectal cancer (Cha et al., 2019; Dai et al., 2019). Our result showed that *CHFR* is an epigenetically silenced gene in STAD (29.3%), COAD-READ (32.0%), and ESCA (25.3%), whereas *SPINT2* is a candidate tumor suppressor gene that encodes a serine protease inhibitor called stem cell growth factor activation inhibitor type 2 (HAI-2) which acts to inhibit cytokine activators (Kawaguchi et al., 1997). Previous studies have shown that this gene has been epigenetically silenced in numerous cancers including hepatocellular carcinoma,

gastric cancer, and esophageal squamous cell carcinoma (Dong et al., 2010; Tung et al., 2009; Yue et al., 2014). Our results found that *SPINT2* obtained an epigenetically silenced gene in LIHC (81.4%). Similarly, *RNF135* is widely present in various organs of the human body, such as the liver and kidney, and is an immunologically related gene (Lai et al., 2019). Previous studies have reported that RNF135 is hyper-methylated in LIHC (Song et al., 2013). We found that *RNF135* obtained epigenetically silenced in LIHC (83.8%) and STAD (8.8%). *GRHL2* is a tumor suppressor gene that encodes a transcription factor induces the target gene to produce a variety of epithelial cell adhesion factors, thereby inhibiting cell migration and invasion. Low-expression *GRHL2* often occurs in breast, ovarian, and gastric cancers, with worse survival and higher metastatic risk (Chung et al., 2016; Cieply et al., 2013; Frisch et al., 2017; Werner et al., 2013; Xiang et al., 2013, 2017). Our results found that GRHL2 was epigenetically silenced in LIHC (92.6%) and STAD (9.6%). Despite that the other three genes, *SNX19, SLC25A24*, and *RIPK3*, were not previously identified as epigenetically silent genes in GI tumors. *SNX19* and *RIPK3* play important roles in apoptosis, whereas *SLC25A24* is involved in ATP energy conversion (Traba et al., 2012).

Generally speaking, *SPINT2, RNF135,* and *SLC25A24* were found to be epigenetically silenced genes in LIHC. *RIPK3* was found silenced in both LIHC and ESCA. *GRHL2* was found silenced in both LIHC and STAD. Moreover, *CHFR* was found silenced in COAD-READ, STAD, and ESCA. *SNX19* was found silenced in STAD, but this phenomenon was also present in some COAD, READ, ESCA, and LIHC samples. *MLH1* and *EPM2AIP1* were silenced in COAD-READ and STAD. However, no results were found for silence genes in PAAD.

Integrative molecular subtyping of pan-GI cancer by multi-omics data

Integrative molecular subtyping with iClusterPlus was performed using three data types (CNV, DNA methylation, mRNA) across 1673 tumor samples and identified 9 iClusters as shown in the Figure 2. Five iClusters were dominated by a single tumor type (iC2: LIHC, iC4: COAD-READ, iC5: LIHC, iC7: ESCA, iC9: COAD-READ). We retrieved 13 subgroups of pan-GI cancer each having at least 30 samples, including COAD_iC3 (samples of COAD-READ in iC3), COAD_iC4, COAD_iC9, LIHC_iC2, LIHC_iC5, ESCA_iC1, ESCA_iC7, PAAD_iC6, PAAD_iC8, STAD_iC1, STAD_iC3, STAD_iC6.

CHFR and MLH1 were the most significant genes we found in the Silenced Genes part. Silencing of MLH1 and CHFR can influent the colorectal tumor and the gastric tumor a lot. In the result of multi-omics analysis, COAD-READ_iC4, STAD_iC3, STAD iC1 displayed high silencing frequency of CHFR and MLH1 genes, showing the reliability of clustering. In addition, the molecular profile of the STAD iC6 subgroup showed a low silencing frequency of CHFR and MLH1 genes relative to STAD_iC1 and STAD_iC3. RAS genes are collectively mutated human cancers and we hope to find possible cause of tumor heterogeneity through analyzing the RAS gene. In the multi-omics based stratification, PAAD iC8 exhibited a high RAS mutation, while PAAD_iC6 exhibited nearly no RAS mutation.

To characterize the composition and relative homogeneity of each iCluster, we plotted the dominant-cancer-type proportion versus the mean iCluster silhouette width, a metric within-group homogeneity. As shown in Figure 3, the single-type-dominant iClusters (iC2, iC4, iC7, iC9) obtained the highest silhouette width. The multiple-type iClusters (iC1, iC3, iC6, iC8) had similar ranges of silhouette width as compared to single-type-dominant iClusters, which suggested classification is robust.

Compared with the results of individual omics analysis, the results of multi-omics analysis can better reflect the homogeneity and heterogeneity of GI tumors. For individual omics analysis, the CNV profile and the DNA methylation level reveal high similarity between different types of tumors whereas the clustering result of the mRNA expression profile corresponds to the tumor types. And the result of multi-omics clustering contains not only single-type-dominant iClusters but also multiple-type-dominant iClusters, while both the individual omics results and multi-omics results show greater differences between LIHC and other GI tumors.

The integrative clustering result was then evaluated by a survival analysis of each subgroup, which revealed significant differences in patient survival of each cancer stratification. As shown in Figure 4, within LIHC cases, LIHC_iC5 was associated with worse survival while the LIHC_iC2 was associated with better survival ($p = 0.00754$), whereas within PAAD cases, PAAD6 showed better survival than PAAD8 ($p = 0.00018$). As for STAD cases, the overall survival of STAD_iC1 was substantially decreased ($p = 0.02476$) compared to the

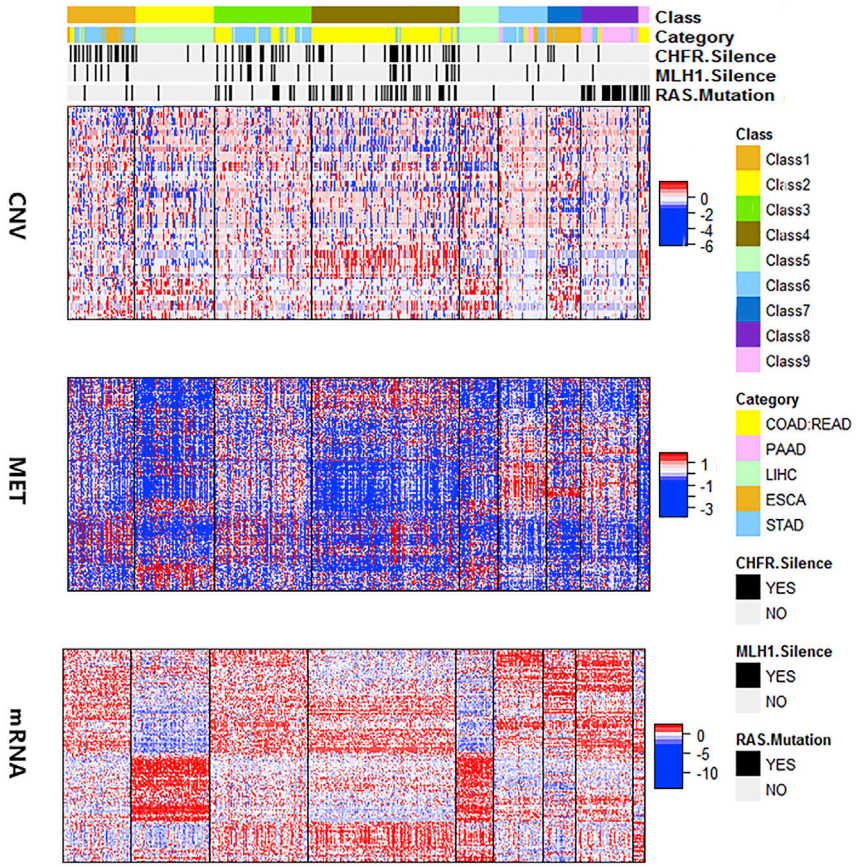


Figure 2. Multi-omics data clustering analysis of pan-gastrointestinal cancer by iClusterPlus

"Class" shows all the samples are divided into 9 classes represented by different colors. "Category" represents the cancer type of the samples, "CHFR. Silence" indicates the silence of the CHFR gene in the samples; "MLH1.Silence" means the silence of MLH1 gene in the samples. And "RAS Mutation" indicates the mutation of the RAS gene in the samples while black color indicates the mutation and gray indicates normal. CNV shows the copy number variation profile in each iCluster with each row represents a chromosome segment. MET shows the DNA methylation profile in each iCluster with each row represents a methylation site. mRNA shows the gene expression profile in each iCluster with each row represents a gene.

others. We observed that there were no substantial differences in overall survival between the two subgroups of ESCA, which may be due to the small sample size of ESCA patients. In general, the prognosis differences among pan-GI subtypes based on multi-omics integration are more significant than clustering by single-omics data.

Enrichment analysis of discriminative omics-signatures for pan-GI stratification

We obtained the most discriminative signatures for pan-GI stratification. Concisely, 22 segments from the CNV platform, 187 CpG sites from the methylation platform, and 250 genes from the expression platform were selected according to the fit score exported by iClusterPlus after setting the threshold of three quantiles.

As shown in Figure 5, the 250 genes from the expression platform significantly enriched in KEGG pathways including ribosome, ECM-receptor interaction, focal adhesion, protein digestion and absorption, estrogen signaling pathway, PI3K-Akt signaling pathway, and antigen processing and presentation. After mapping against the USGC data set, 245 genes from the methylation platform significantly enriched in KEGG pathways including ascorbate and aldarate metabolism, retinol metabolism, pentose and glucuronate interconversions, metabolism of xenobiotics by cytochrome P450, steroid hormone biosynthesis, porphyrin, and chlorophyll metabolism, and complement and coagulation cascades. The enrichment pathways at the

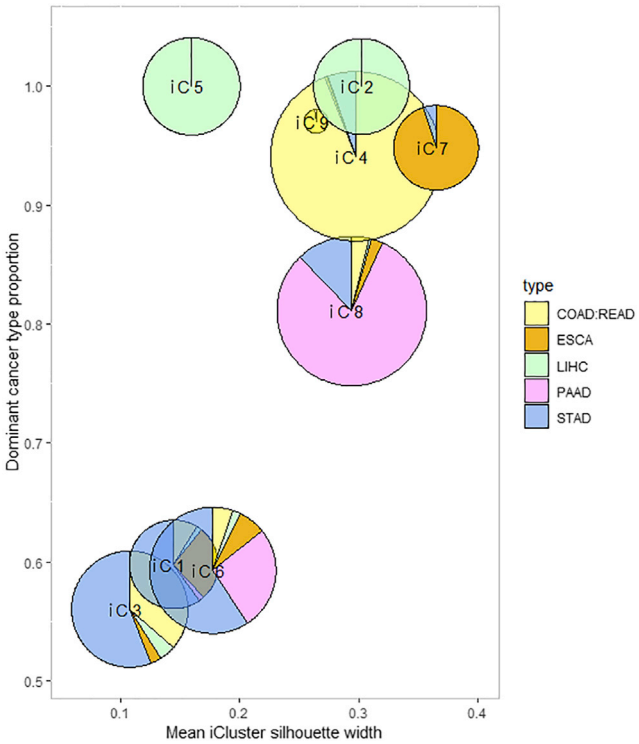


Figure 3. The proportion of pan-gastrointestinal cancer in each iCluster class

The abscissa means the mean silhouette width of each iCluster and the ordinate shows the proportion of the dominant tumor samples in each iCluster. The size of the circle is proportional to the number of samples; the proportion of different colors in the circle means the proportion of different tumors in each iCluster.

methylation level are mainly corresponding with metabolic-related pathways. Concernedly, pentose phosphate pathway (PPP) is one of the key pathways for the rapid proliferation of cancer cells, and the pentose phosphate synthesized by this pathway is the basis of nucleic acid synthesis. Moreover, PPP pathway also provides a large amount of NADPH for the synthesis of fatty acids in cancer cells (Patra and Hay, 2014). The results of enrichment analysis showed that metabolism may play an important role in cancer subtype classification, and there may have a potential relationship between metabolism and methylation.

Similarly, we analyzed the characteristics of the discriminative omics-signature genes in each iCluster. Hierarchical clustering was performed on the mean value of signature genes of samples in each of the 9 classes, as shown in Figure 6. Genes at the expression level are classified into five categories by hierarchical clustering, in which the genes in category 1 are significantly enriched in the estrogen signaling pathway, and the genes in category 2 are significantly enriched in the ECM-receptor interaction/PIK3-Akt signaling pathway. A considerable part of the genes in category 3 is related to oxidative phosphorylation. The genes in category 4 were significantly enriched in the ribosomes, and the genes in category 5 cover quite a few potential survival markers (1/3) found by the Cox model. The ESCA samples in iC7 have significant copy number repeat a variation on chromosome 3, while at the mRNA level, iC7 is significantly highly expressed in the estrogen signaling pathway compared with the others. There were almost no differences between iC4 and iC9 which were mainly composed of COAD-READ samples. However, at the mRNA level, the expression of ribosome-related genes in iC9 is significantly higher than that of iC4. For PAAD samples, the expression of ECM-receptor interactions in iC8 is significantly higher than iC6, and correspondingly the survival of iC8 is worse than iC6 as shown in Figure 4.

Significant prognostic markers obtained by the proportional hazards model

The proportional hazards model (Cox model) was used to identify potential prognostic markers among the characters generated from iClusterPlus for different cancers. Table 1 represents all the retrieved prognostic markers. Our results showed the *PSCA* gene is a significant risk factor for both COAD-READ and STAD,

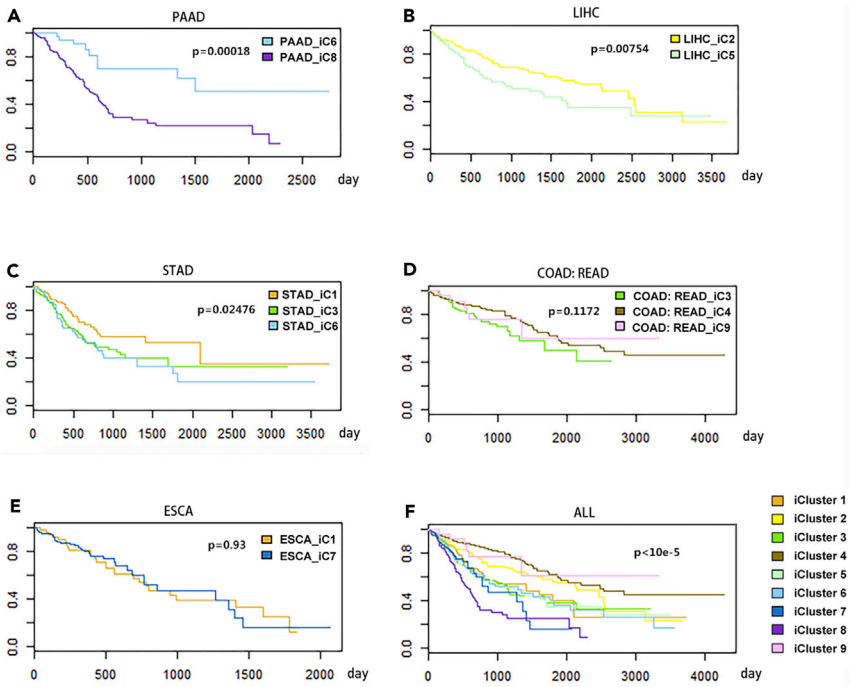


Figure 4. Survival curve of each gastrointestinal cancer by multi-omics stratification of pan-cancer

(A–E) The survival curve of different subgroups in each cancer divided by iClusterPlus method, 185 PAAD samples, 377 LIHC samples, 439 STAD samples, 626 COAD-READ samples, 183 ESCA samples, respectively.

(F) the survival curve of the 9 iClusters.

whereas *PSCA* encodes a glycosylphosphatidylinositol-anchored cell membrane glycoprotein, which is abundantly expressed in pancreatic cancer, and clinical trials on this target in pancreatic cancer are already underway (González et al., 2013; Liu and Saif, 2017). We found that the high expression of *PPP1CB* is associated with worth survival in both LIHC and *PAAD*. *PPP1CB* encodes one of the three catalytic subunits of protein phosphatase 1 which participate in a series of follow-up events such as cell division and glycogen metabolism through regulation of the oncogene RAF. It has been reported that this gene is a potential marker for malignant melanoma (Huang et al., 2019; Young et al., 2018).

Despite that, some of the predicted biomarkers have been under investigation by other studies. *TUBA1C* is an apoptosis-related target, being the target of many known drugs (Wang et al., 2019), including two important drugs, epothilone D, and patupilone. Epothilone D was investigated for use in colorectal cancer (Lee et al., 2007), and patupilone was also investigated for treatment in gastric cancer (Kim et al., 2012). *SERPINB5* is a target of the p53 signal pathway, which has been associated with colorectal tumor (Chang et al., 2018; Li et al., 2017), hepatocellular carcinoma (Yang et al., 2016), and gastric tumor (Lei et al., 2011). Our results showed that it is a significant marker of hepatocellular carcinoma which is consistent with the previous study. Similarly, the *CLCA2* is a tumor suppressor in colorectal tumors (Bustin et al., 2001), and our results showed that it is related to the survival of ESCA. At the methylation level, we observed *PSMB2* as an important prognostic marker. There is a publication suggesting knockdown of *PSMB2* suppresses hepatocellular carcinoma cell invasion and proliferation (Tan et al., 2018). Because of the certain homogeneity among pan-GI tumors, the identified markers that had been supported by the previous studies may have impact on other GI tumors.

The results of the proportional hazards model indicate that our analysis was compatible with previous studies, suggesting that our classifier is credible. In addition, our predicted biomarker may guide the future pan-cancer study.

We further calculated the C-index for each cancer type to quantify the discriminative power of the Cox model, as shown in Figure 7. All of the GI cancers have a C-index higher than 0.5, especially PAAD reveals

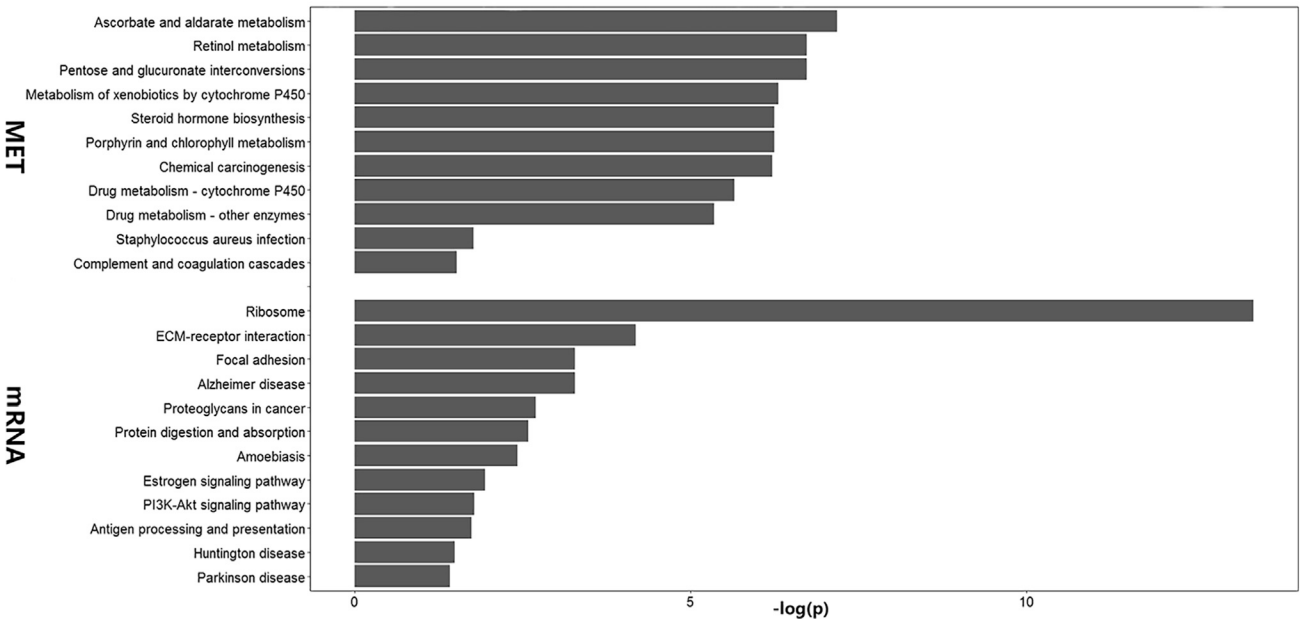


Figure 5. The enriched KEGG pathway of discriminative omics-signatures for pan-GI

The enriched pathways of methylation and the enriched pathways of mRNA expression

the highest C-index. We also evaluated the C-index by multi-omics data integration and by each single-omics data of PAAD. The results showed that the markers identified by multi-omics integration could better indicate the prognosis status of pan-GI cancers.

Mutational assessment of subgroups of pan-GI cancers by iClusterPlus

We calculated the tumor mutation burden (TMB) for each GI tumor sample (Figure 8). Concisely, iC8 has the lowest average TMB, which is mainly composed of PAAD samples, whereas within the categories containing a high proportion of STAD samples, the TMB of iC6 is significantly lower than iC1 and iC3. Moreover, the iC6 composed of a higher sample size of PAAD (n samples) and moderate sample size of STAD (n samples). The TMB for STAD (3.9) was found significantly lower as compared with iC1 (TMB = 13.9) and iC3 (TMB = 17.6). In addition, iC3 has the highest average TMB, which is mainly composed of STAD and COAD-READ, and the average TMB of COAD-READ in this category is 31.5, which is significantly higher than the average TMB of COAD-READ in iC4 (10.7).

In addition to TMB, the study of mutation frequency also brought us new discoveries. We clustered the weights of 30 mutation features summarized by Catalogue of Somatic Mutations in Cancer (COSMIC) in 9 iClusters to show the differences of iClusters at the mutation signatures level. As shown in Figure 8, the result was consistent with the result of TMB, where iC1, iC3, and iC4 were enriched in the hypermutation signature. The mutational signatures of iC6 and iC9 are apparent in the enrichment of Signature1, which may be age-related. The signatures of iC7 are similar to iC2 and iC5 compared to other iClusters.

Multi-omics clustering of pan-GI cancer samples from ICGC database

In order to further validate our results, we conducted a similar analysis of pan-GI cancer samples, including mRNA and CNV data, from the ICGC database released in 2019. The results by iClusterPlus show that the pan-GI samples can be divided into 7 clusters. Concisely, the LIHC and PAAD were divided into 2 clusters, and the COAD-READ was divided into3 clusters. These results were consistent with those of TCGA data.

To further demonstrate the classification of ICGC data, we conducted survival analysis and differential expression analysis of ICGC data, the survival analysis shows that the tumors were stratified successfully (the result of LIHC were showed in Figure 9A as an example), especially the LIHC ($p = 0.0039$) and STAD ($p = 0.0085$). For example, LIHC was significantly divided into 2 clusters in the results of TCGA. In the ICGC result, LIHC was also divided into LIHC3 and LIHC4, in which LIHC3 showed better survival than

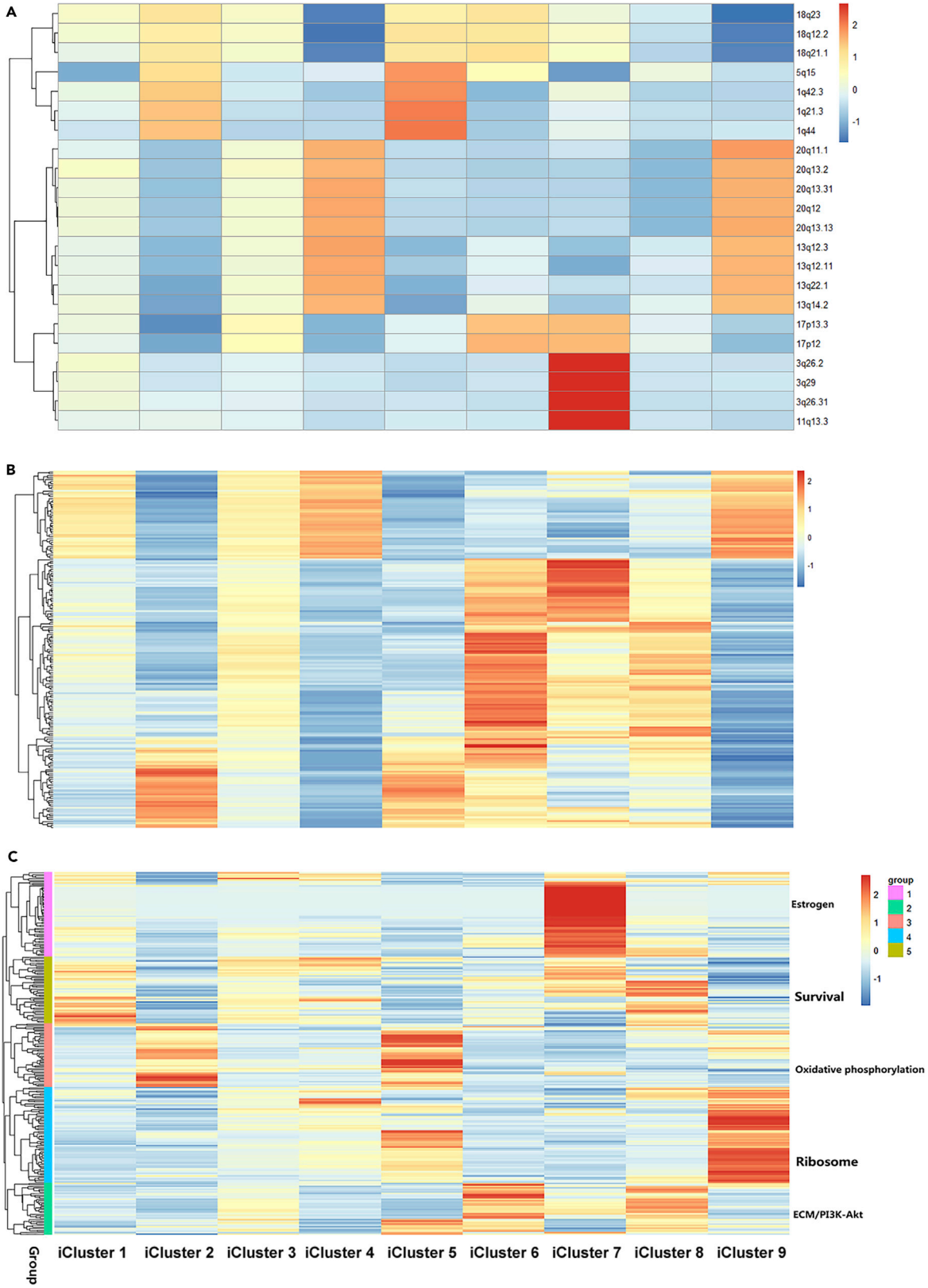

A
18q23
18q12.2
18q21.1
5q15
1q42.3
1q21.3
1q44
20q11.1
20q13.2
20q13.31
20q12
20q13.13
13q12.3
13q12.11
13q22.1
13q14.2
17p13.3
17p12
3q26.2
3q29
3q26.31
11q13.3
2
1
0
-1
B
2
1
0
-1
C
group
1
2
3
4
5
Estrogen
Survival
Oxidative phosphorylation
Ribosome
ECM/PI3K-Akt
group
iCluster 1
iCluster 2
iCluster 3
iCluster 4
iCluster 5
iCluster 6
iCluster 7
iCluster 8
iCluster 9

Figure 6. Heatmap of the discriminative omics-signature genes in 9 iCluster classes at three omics levels

(A) represents the copy number variation level (CNV), each row represents a chromosome segment.

(B) represents the methylation level (MET), each row represents a methylation site.

(C) represents the gene expression profile level (mRNA), and each row represents a gene. Different colors represent different groups.

LIHC4 with a p value of 0.0039. The differential expression analysis also showed that there were a lot of differential genes between the two subtypes.

In order to further verify our findings with ICGC data, we compared the ICGC genes and CNV fragments results with the TCGA result. We found 41 genes and 3 CNV fragments have discriminative effects on the classification of both the two data sets. In addition, we conducted the KEGG analysis for the significant genes predicted by using the ICGC data set and found that there were also some similarities between ICGC and TCGA results, which means our results are validated at the pathway level.

DISCUSSION

In this study, we leveraged the accurate stratification and prognosis markers identification of pan-GI cancer by multi-omics data integration. GI tract cancers have complex relationships in clinical course, and we systematically uncovered the homogeneity and heterogeneity among different subtypes, which can improve the personalized treatment. The identified prognostic biomarkers could predict the prognostic status of each patient, which could help to achieve precision medicine for GI tract cancers.

We performed the unsupervised consensus clustering of 3 platforms of molecular profiles of pan-GI samples in the TCGA database, focusing on the copy number variation, the methylation, and the transcriptome levels. The single-platform of copy number variation level showed high similarity among different types of tumors. For the methylation level, almost all of the LIHC were divided into one category, while other GI tumors showed some homogeneity. At the transcriptome level, the clustering results were correlated with tumor types, and the LIHC showed more differences from other GI tumors.

Moreover, by combining the expression profile and methylation data, we found nine genes, *MLH1, EPM2AIP1, SNX19, SLC25A24, RIPK3, RNF135, CHFR, SPINT2, GRHL2,* that were frequently silenced by promoter methylation in GI cancer. Among them, *EPM2AIP1*, *MLH1*, and *CHFR* have previously been reported to be silent in colorectal cancer and gastric cancer (Neumeyer et al., 2019; Oue et al., 2019; Suter et al., 2004), and RNF135 obtained an epigenetically silenced gene in LIHC (Song et al., 2013). In addition, *SNX19, SLC25A24*, and *RIPK3* were silenced due to hypermethylation.

By multi-omics clustering using the iClusterPlus algorithm, we identified nine major molecular subtypes of GI cancers, including 5 single-type-dominant clusters and 4 mixed clusters, and the clustering reveals not only the organ and cell-of-origin but also the molecular mechanisms. Especially the mixed clusters can show the homogeneity of pan-GI cancers. For some clusters, the specific significant pathways may be most relevant than the histology signatures, the cluster iC1, iC3, iC4 identified strong *MLH1, CHFR,* and TMB silencing features. PI3K-Akt pathway was highly associated with iC6. The single-type-dominant clusters also showed special molecular features, iC7 was mainly composed of ESCA, which was highly expressed in the estrogen signaling pathway. iC8 was composed of PAAD with obvious KRAS mutations. iC9 consisted of COAD-READ, which was highly expressed in the ribosomal pathway.

The pan-cancer study showed the homogeneity and heterogeneity of GI cancer. We found that the LIHC was significantly different from other GI tumors at the molecular level, while STAD samples had certain homogeneity with others. Moreover, there were significant differences in the survival of some tumors from the same organ but were belonging to different subtypes, especially the STAD, LIHC, and PAAD. The classification of tumors from the same organ or cell-of-origin may be related to some molecular mechanisms, for example, the worse survival subgroup PAAD_iC8 has significant RAS mutation as compared with PAAD_iC6. Meanwhile, a considerable part of the features for PAAD that had a significant effect on survival screened by the Cox model was associated with KRAS mutation. The result suggests that KRAS mutation may be associated with the survival of PAAD.

Molecular differences among GI cancers related to several processes and pathways. Based on the enrichment analysis of the discriminative signatures which played significant roles in the classification of pan-GI

Table 1. Significant prognostic marker selected by Cox model

	Significant factors	β-value	p value
COAD-READ	cg01401376/EYA4	0.44	0.00051
	TUBA1C	−0.56	0.0093
	PSCA	2.31	0.027
	KRT15	9.69	0.038
STAD	CCT6A	0.25	9.65e-05
	PSCA	0.15	0.0021
	RPL13	0.54	0.0038
LIHC	SERPINB5	26.15	0.0043
	ADAM9	1.1	0.034
	PPP1CB	0.77	0.039
ESCA	TMSL3	0.46	0.00013
	CLCA2	0.17	0.0047
	SAMD9	0.10	0.021
	HSPD1	0.24	0.039
	GFPT1	0.31	0.047
PAAD	cg25391023/BTNL2	6.41	7.49e-06
	cg24109894/PSMB2	-7.11	2.00e-05
	cg05316065/GSDMC	−6.70	0.00013
	PPP1CB	5.63	0.00030
	KRT6C	−306.79	0.00052
	MYOF	4.51	0.00054
	cg00412772/C19orf33	6.08	0.00059
	KRT6A	134.52	0.00080
	MXRA5	−3.82	0.0012
	cg15484375/SAA1	4.73	0.0013
	CAPN2	−3.03	0.0036
	CAP1	−2.30	0.011
	YWHAZ	5.49	0.013
	SH3BGRL3	−2.38	0.016
	SYPL1	−5.65	0.018
	13q12.3	7.94	0.020
	CAPN1	3.47	0.020
	CLIC4	−4.02	0.021
	KRT16	−25.91	0.024
	RHOC	−3.65	0.024
	ATP2A3	1.47	0.024
	PLAU	−2.22	0.026
	FKBP8	−4.58	0.028
	cg05659947/C4BPB	−3.24	0.036
	cg25577842/WT1-AS	−2.44	0.039
	17p12	3.05	0.044

Table 1 shows the significant prognostic markers that we selected by the Cox model in the first column, including 29 genes, 8 methylation sites, and 2 chromosome segments. The β value was a regression coefficient measure the effect of each factor on the Cox model, if $\beta > 0$ then the marker considered as risk factor, if the $\beta < 0$ then the marker considered as protective factor. We used p value ≤ 0.05 as the significance level to screen out the characteristics with a significant effect on survival.

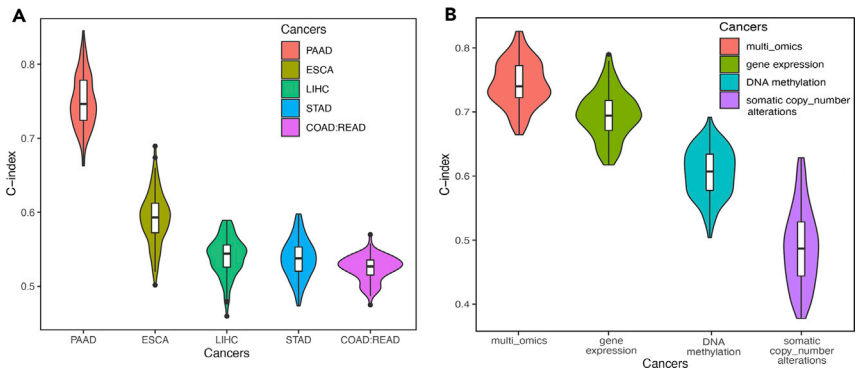


Figure 7. The C-index comparison of the biomarkers by Cox model

(A) The C-index comparison of the biomarkers by Cox model in pan-GI cancers.

(B) The C-index by multi-omics data integration and each single-omics data in PAAD.

cancers by the iClusterPlus model, we observed heterogeneity between genotypes that were concentrated with the metabolic pathways at methylation level, indicate that the metabolism could distinguish the subtypes. At the transcriptome level, two classic cancer pathways, ECM-receptor interaction and PI3K-Akt signaling pathway may play an important role in distinguishing different types of pan-GI cancer, indicating that some GI cancers may be carcinogenic by activating the PI3K-Akt signaling pathway and had certain homogeneity.

The significant molecular differences of the Pan-GI subtypes may have consequences for targeted therapy. We found two genes that were simultaneously present in cancers from different organs, *PSCA* was found in

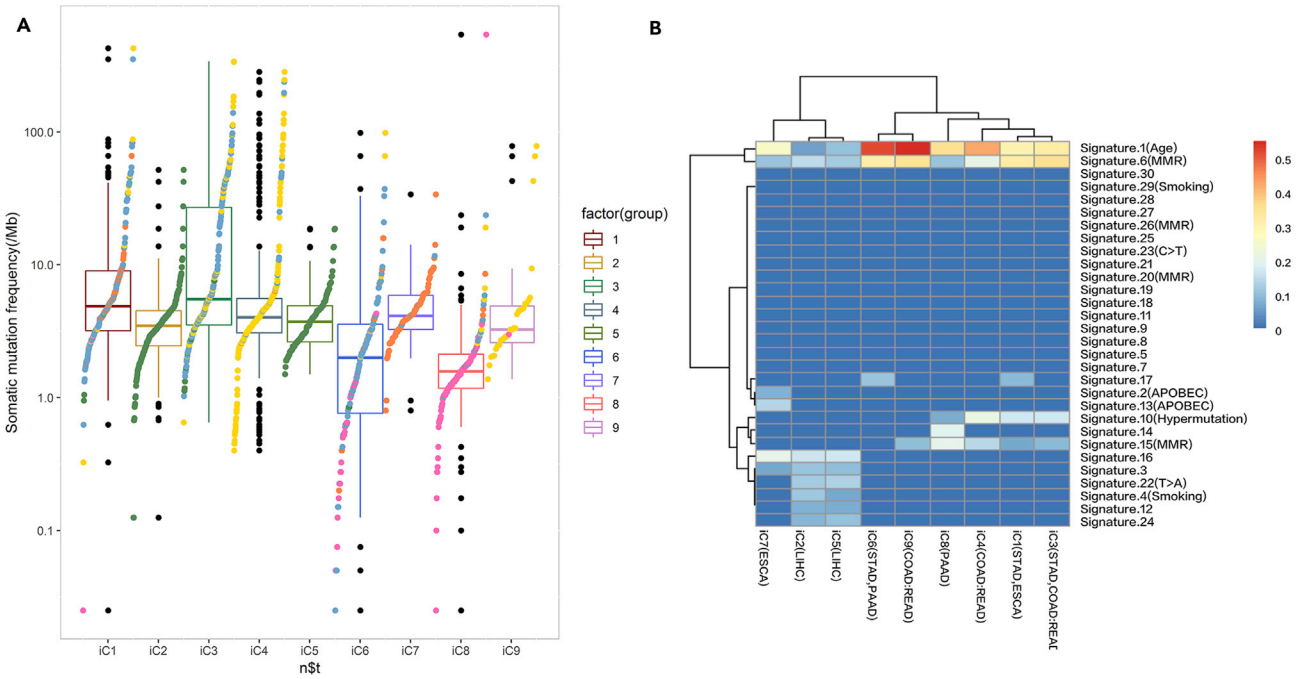


Figure 8. The mutational analysis of the iClustersPlus

(A) The boxplot for TMB of 9 iCluster classes. The ordinate represents the TMB value and the abscissa represents nine iClusters. The black dots on the boxplot represent outliers. All the colored dots represent different types of cancer samples.

(B) Heatmap of iCluster classes on mutational signatures. This figure shows the differences of iClusters at the 30 mutation signatures level, and the color reveals the enrichment.

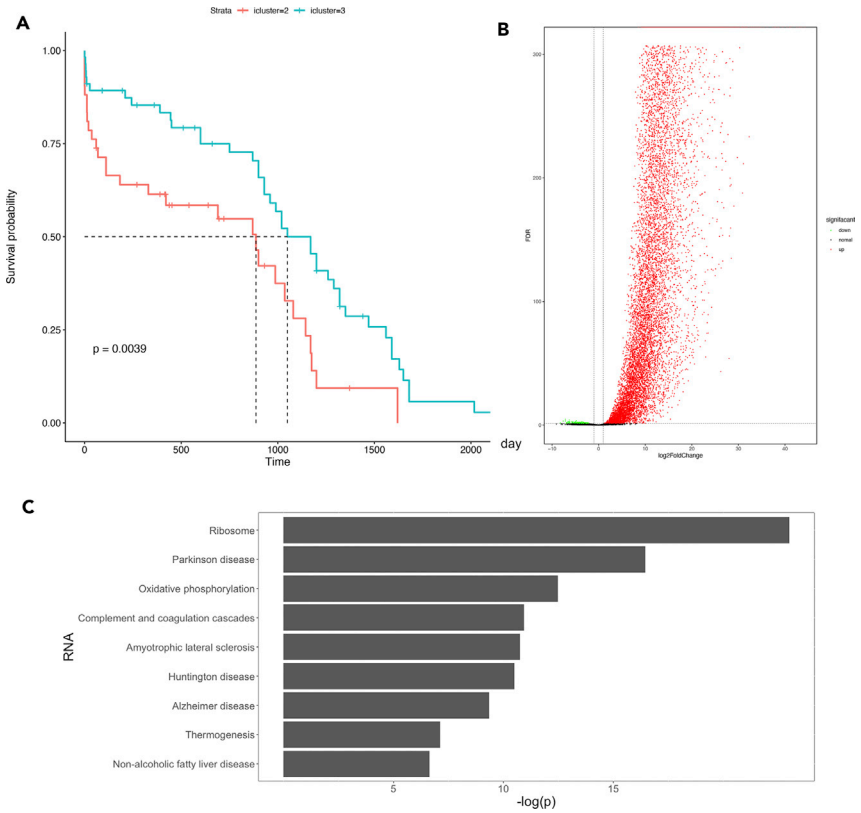


Figure 9. Stratification and expression analysis of LIHC based on integrative classification of pan-GI samples from ICGC

(A) The survival curve of two LIHC subtypes by iClusterPlus on pan-GI cancer.

(B) The differential gene expression analysis between two subtypes of LIHC samples.

(C) The enriched KEGG pathways of significantly differential expressed genes.

COAD-READ and STAD, and *PPP1CB* was both found in PAAD and LIHC; other targets such as *Epothilone D*, *TUBA1C, SERPINB5, CLCA2, PSMB2* also got attention in our study. There are strong overlaps between the previous research and our study. For example, *PSCA* is abundantly expressed in pancreatic cancer, and clinical trials on this target in pancreatic cancer are already underway (González et al., 2013; Liu and Saif, 2017); *SERPINB5* has been associated with the colorectal tumor, hepatocellular carcinoma, and gastric tumor (Chang et al., 2018; Lei et al., 2011; Li et al., 2017; Yang et al., 2016). This means that our method is effective in finding potential targets.

To further validate our results, we performed the same unsupervised consensus clustering for the data of the ICGC database. Our analysis reveals that the classification result of ICGC and TCGA did not vary widely including classification results and molecule features, verifying the reliability of our whole analysis process.

Previous pan-cancer studies by TCGA team focused on a variety of tumors from different tissues and organs, and most of pan-GI cancers was classified into one cluster (Hoadley et al., 2018), but there is no detail analysis on particular kinds of cancers. Our research focused on pan-GI cancer, which not only revealed multi-omics subtyping but also identified the significant prognosis markers by using Cox model. Furthermore, we demonstrated that liver cancer has quite different molecular characteristic from other pan-GI cancers, although all the GI cancer samples were grouped into one big cluster in previous TCGA study. In conclusion, by leveraging the multi-omics data from thousands of tumor samples of the TCGA and ICGC database, we uncovered the homogeneity and heterogeneity of GI cancer and provide a framework of pan-GI cancer classification which may help researchers in selecting potential prognostic markers for pan-GI cancer. This study will help the researchers to better understand the pathophysiology of the pan-GI cancer and converse more findings into clinical therapies.

Limitations of the study

The pan-cancer analyses based on 3 molecular platforms have limitations imposed by original omics data. The sample size of different cancers from TCGA is different, as an example, and there are smaller samples of ESCA than other GI tract cancers, which may cause deviation of statistical results. Another major limitation of the original omics data is the differences between different databases, and the ICGC database lacks methylation data so the verification just including mRNA and CNV data. In addition, the significant prognostic markers we found need to be further investigated and supported by the *in vitro* and *in vivo* experiments to further verify the impact on pan-GI cancers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Information of pan-gastrointestinal cancer datasets
  - SNV and somatic copy number alterations analysis
  - DNA methylation data analysis
  - mRNA expression data analysis
  - Identification of epigenetically silent genes
  - Integrative clustering using iClusterPlus algorithm
  - Data processing
  - Model selection
  - Feature selection
  - Silhouette width
  - Survival analysis of the Pan-GI subtypes from iClusters
  - Proportional hazards model (Cox model)
  - Mutation signature analysis
  - Evaluation of the prognostic power of the biomarkers by the concordance index
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102824.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program (2019YFA0905400, 2021YFC2100600, 2017YFC0908105), National Science Foundation of China (32070679), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Science Foundation of China (U1804284, 81421061, 81701321, 31571012, 81501154, 81871055), the Shanghai Natural Science Funding (16ZR1449700), Shanghai Hospital Development Center (SHDC12016115), Shanghai Science and Technology Committee (17JC1402900 and 17490712200), Shanghai Mental Health Commission (ZK2015B01 and 201540114), National Science Foundation of Shandong Province (ZR2019YQ14), Taishan Scholar Program of Shandong Province (tsqn201812153). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, Z.W., Y.S., and H.J.; methodology, Z.W., H.J., H.Z., and R.S.; formal analysis, Z.W., H.J., and H.Z.; investigation, Z.W., H.J., H.Z., and K.W.; resources, Z.W., Y.S.; writing—original draft, H.J., H.Z.; writing—review and editing, Z.W., Y.S., H.J., A.F., R.S., and Z.L.; visualization, H.J.; supervision, Z.W., Y.S.; project administration, Z.W., Y.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 16, 2020
Revised: May 1, 2021
Accepted: July 5, 2021
Published: August 20, 2021

REFERENCES

Alexandrov, L.B., and Stratton, M.R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr. Opin. Genet. Dev. *24*, 52–60. https://doi.org/10.1016/j.gde.2013.11.014.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415–421. https://doi.org/10.1038/nature12477.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259. https://doi.org/10.1016/j.celrep.2012.12.008.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407. https://doi.org/10.1038/ng.3441.

Bijlsma, M.F., Sadanandam, A., Tan, P., and Vermeulen, L. (2017). Molecular subtypes in cancers of the gastrointestinal tract. Nat. Rev. Gastroenterol. Hepatol. *14*, 333–342. https://doi.org/10.1038/nrgastro.2017.33.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424. https://doi.org/10.3322/caac.21492.

Bustin, S.A., Li, S.R., and Dorudi, S. (2001). Expression of the Ca2+-activated chloride channel genes CLCA1 and CLCA2 is downregulated in human colorectal cancer. DNA Cell Biol. *20*, 331–338. https://doi.org/10.1089/10445490152122442.

Cancer Genome Atlas Research Network (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell *169*, 1327–1341.e1323. https://doi.org/10.1016/j.cell.2017.05.046.

Cha, Y., Kim, S.Y., Yeo, H.Y., Baek, J.Y., Choi, M.K., Jung, K.H., Dong, S.M., and Chang, H.J. (2019). Association of CHFR promoter methylation with treatment outcomes of irinotecan-based chemotherapy in metastatic colorectal cancer. Neoplasia *21*, 146–155. https://doi.org/10.1016/j.neo.2018.11.010.

Chang, I.W., Liu, K.W., Ragunanan, M., He, H.L., Shiue, Y.L., and Yu, S.C. (2018). SERPINB5 expression: association with CCRT response and prognostic value in rectal cancer. Int. J. Med. Sci. *15*, 376–384. https://doi.org/10.7150/ijms.22823.

Chen, F., Zhang, Y., Bossé, D., Lalani, A.A., Hakimi, A.A., Hsieh, J.J., Choueiri, T.K., Gibbons, D.L., Ittmann, M., and Creighton, C.J. (2017). Pan-urologic cancer genomic subtypes that transcend tissue of origin. Nat. Commun. *8*, 199. https://doi.org/10.1038/s41467-017-00289-x.

Chung, V.Y., Tan, T.Z., Tan, M., Wong, M.K., Kuay, K.T., Yang, Z., Ye, J., Muller, J., Koh, C.M., Guccione, E., et al. (2016). GRHL2-miR-200-ZEB1 maintains the epithelial status of ovarian cancer through transcriptional regulation and histone modification. Sci. Rep. *6*, 19943. https://doi.org/10.1038/srep19943.

Cieply, B., Farris, J., Denvir, J., Ford, H.L., and Frisch, S.M. (2013). Epithelial-mesenchymal transition and tumor suppression are controlled by a reciprocal feedback loop between ZEB1 and Grainyhead-like-2. Cancer Res. *73*, 6299–6309. https://doi.org/10.1158/0008-5472.can-12-4082.

Dai, D., Zhou, B., Xu, W., Jin, H., and Wang, X. (2019). CHFR promoter hypermethylation is associated with gastric cancer and plays a protective role in gastric cancer process. J. Cancer *10*, 949–956. https://doi.org/10.7150/jca.27224.

Dong, W., Chen, X., Xie, J., Sun, P., and Wu, Y. (2010). Epigenetic inactivation and tumor suppressor activity of HAI-2/SPINT2 in gastric cancer. Int. J. Cancer *127*, 1526–1534. https://doi.org/10.1002/ijc.25161.

Esteller, M., Levine, R., Baylin, S.B., Ellenson, L.H., and Herman, J.G. (1998). MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. Oncogene *17*, 2413–2417. https://doi.org/10.1038/sj.onc.1202178.

Fang, F., Wang, X., and Song, T. (2018). Five-CpG-based prognostic signature for predicting survival in hepatocellular carcinoma patients. Cancer Biol. Med. *15*, 425–433. https://doi.org/10.20892/j.issn.2095-3941.2018.0027.

Frisch, S.M., Farris, J.C., and Pifer, P.M. (2017). Roles of Grainyhead-like transcription factors in cancer. Oncogene *36*, 6067–6073. https://doi.org/10.1038/onc.2017.178.

González, C.A., Sala, N., and Rokkas, T. (2013). Gastric cancer: epidemiologic aspects. Helicobacter *18*, 34–38. https://doi.org/10.1111/hel.12082.

Harrell, F.E., Jr., Lee, K.L., and Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. *15*, 361–387. https://doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. *15*, 585–598. https://doi.org/10.1038/nrg3729.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944. https://doi.org/10.1016/j.cell.2014.06.049.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell *173*, 291–304.e296. https://doi.org/10.1016/j.cell.2018.03.022.

Huang, Z., Lin, B., Pan, H., Du, J., He, R., Zhang, S., and Ouyang, P. (2019). Gene expression profile analysis of ENO1 knockdown in gastric cancer cell line MGC-803. Oncol. Lett. *17*, 3881–3889. https://doi.org/10.3892/ol.2019.10053.

Kawaguchi, T., Qin, L., Shimomura, T., Kondo, J., Matsumoto, K., Denda, K., and Kitamura, N. (1997). Purification and cloning of hepatocyte growth factor activator inhibitor type 2, a Kunitz-type serine protease inhibitor. J. Biol. Chem. *272*, 27558–27564. https://doi.org/10.1074/jbc.272.44.27558.

Kim, Y.H., Muro, K., Yasui, H., Chen, J.S., Ryu, M.H., Park, S.H., Chu, K.M., Choo, S.P., Sanchez, T., DelaCruz, C., et al. (2012). A phase II trial of ixabepilone in Asian patients with advanced gastric cancer previously treated with fluoropyrimidine-based chemotherapy. Cancer Chemother. Pharmacol. *70*, 583–590. https://doi.org/10.1007/s00280-012-1943-6.

Lai, Y., Liang, M., Hu, L., Zeng, Z., Lin, H., Yi, G., Li, M., and Liu, Z. (2019). RNF135 is a positive regulator of IFN expression and involved in RIG-I signaling pathway by targeting RIG-I. Fish Shellfish Immunol. *86*, 474–479. https://doi.org/10.1016/j.fsi.2018.11.070.

Lee, S.H., Son, S.M., Son, D.J., Kim, S.M., Kim, T.J., Song, S., Moon, D.C., Lee, H.W., Ryu, J.C., Yoon, D.Y., and Hong, J.T. (2007). Epothilones induce human colon cancer SW620 cell apoptosis via the tubulin polymerization independent activation of the nuclear factor-kappaB/IkappaB kinase signal pathway. Mol. Cancer Ther. *6*, 2786–2797. https://doi.org/10.1158/1535-7163.mct-07-0002.

Lei, K.F., Liu, B.Y., Wang, Y.F., Chen, X.H., Yu, B.Q., Guo, Y., and Zhu, Z.G. (2011). SerpinB5 interacts with KHDRBS3 and FBXO32 in gastric cancer cells. Oncol. Rep. *26*, 1115–1120. https://doi.org/10.3892/or.2011.1369.

Lengyel, A., and Botta-Dukát, Z. (2019). Silhouette width using generalized mean-A flexible method for assessing clustering efficiency. Ecol. Evol. *9*, 13231–13243. https://doi.org/10.1002/ece3.5774.

Li, H., Zhong, A., Li, S., Meng, X., Wang, X., Xu, F., and Lai, M. (2017). The integrated pathway of TGFβ/Snail with TNFα/NFκB may facilitate the tumor-stroma interaction in the EMT process and

colorectal cancer prognosis. Sci. Rep. *7*, 4915. https://doi.org/10.1038/s41598-017-05280-6.

Liu, F., and Saif, M.W. (2017). T cell optimization for the treatment of pancreatic cancer. Expert Opin. Biol. Ther. *17*, 1493–1501. https://doi.org/10.1080/14712598.2017.1369948.

Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics *9*, 271. https://doi.org/10.1186/1471-2105-9-271.

McGranahan, N., and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. Cell *168*, 613–628. https://doi.org/10.1016/j.cell.2017.01.018.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

Neumeyer, S., Popanda, O., Edelmann, D., Butterbach, K., Toth, C., Roth, W., Bläker, H., Jiang, R., Herpel, E., Jäkel, C., et al. (2019). Genome-wide DNA methylation differences according to oestrogen receptor beta status in colorectal cancer. Epigenetics *14*, 477–493. https://doi.org/10.1080/15592294.2019.1595998.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993. https://doi.org/10.1016/j.cell.2012.04.024.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54. https://doi.org/10.1038/nature17676.

Oue, N., Sentani, K., Sakamoto, N., Uraoka, N., and Yasui, W. (2019). Molecular carcinogenesis of gastric cancer: Lauren classification, mucin phenotype expression, and cancer stem cells. Int. J. Clin. Oncol. *24*, 771–778. https://doi.org/10.1007/s10147-019-01443-9.

Patra, K.C., and Hay, N. (2014). The pentose phosphate pathway and cancer. Trends Biochem. Sci. *39*, 347–354. https://doi.org/10.1016/j.tibs.2014.06.005.

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. *46*, 10546–10562. https://doi.org/10.1093/nar/gky889.

Reis-Filho, J.S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet *378*, 1812–1823. https://doi.org/10.1016/s0140-6736(11)61539-0.

Scolnick, D.M., and Halazonetis, T.D. (2000). Chfr defines a mitotic stress checkpoint that delays entry into metaphase. Nature *406*, 430–435. https://doi.org/10.1038/35019108.

Song, M.A., Tiirikainen, M., Kwee, S., Okimoto, G., Yu, H., and Wong, L.L. (2013). Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. PLoS One *8*, e55761. https://doi.org/10.1371/journal.pone.0055761.

Sun, R., Xu, Y., Zhang, H., Yang, Q., Wang, K., Shi, Y., and Wang, Z. (2020). Mechanistic modeling of gene regulation and metabolism identifies potential targets for hepatocellular carcinoma. Front. Genet. *11*, 595242. https://doi.org/10.3389/fgene.2020.595242.

Suter, C.M., Martin, D.I., and Ward, R.L. (2004). Germline epimutation of MLH1 in individuals with multiple cancers. Nat. Genet. *36*, 497–501. https://doi.org/10.1038/ng1342.

Tan, S., Li, H., Zhang, W., Shao, Y., Liu, Y., Guan, H., Wu, J., Kang, Y., Zhao, J., Yu, Q., et al. (2018). NUDT21 negatively regulates PSMB2 and CXXC5 by alternative polyadenylation and contributes to hepatocellular carcinoma suppression. Oncogene *37*, 4887–4900. https://doi.org/10.1038/s41388-018-0280-6.

Traba, J., Del Arco, A., Duchen, M.R., Szabadkai, G., and Satrústegui, J. (2012). SCaMC-1 promotes cancer cell survival by desensitizing mitochondrial permeability transition via ATP/ADP-mediated matrix Ca(2+) buffering. Cell Death Differ. *19*, 650–660. https://doi.org/10.1038/cdd.2011.139.

Tung, E.K., Wong, C.M., Yau, T.O., Lee, J.M., Ching, Y.P., and Ng, I.O. (2009). HAI-2 is epigenetically downregulated in human hepatocellular carcinoma, and its Kunitz domain type 1 is critical for anti-invasive functions. Int. J. Cancer *124*, 1811–1819. https://doi.org/10.1002/ijc.24115.

Vural, S., Wang, X., and Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC Syst. Biol. *10*, 62. https://doi.org/10.1186/s12918-016-0306-z.

Wang, C.C.N., Li, C.Y., Cai, J.H., Sheu, P.C., Tsai, J.J.P., Wu, M.Y., Li, C.J., and Hou, M.F. (2019). Identification of prognostic candidate genes in breast cancer by integrated bioinformatic analysis. J. Clin. Med. *8*. https://doi.org/10.3390/jcm8081160.

Werner, S., Frey, S., Riethdorf, S., Schulze, C., Alawi, M., Kling, L., Vafaizadeh, V., Sauter, G., Terracciano, L., Schumacher, U., et al. (2013). Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer. J. Biol. Chem. *288*, 22993–23008. https://doi.org/10.1074/jbc.M113.456293.

Xiang, J., Fu, X., Ran, W., Chen, X., Hang, Z., Mao, H., and Wang, Z. (2013). Expression and role of grainyhead-like 2 in gastric cancer. Med. Oncol. *30*, 714. https://doi.org/10.1007/s12032-013-0714-5.

Xiang, J., Fu, X., Ran, W., and Wang, Z. (2017). Grhl2 reduces invasion and migration through inhibition of TGFβ-induced EMT in gastric cancer. Oncogenesis *6*, e284. https://doi.org/10.1038/oncsis.2016.83.

Yagi, K., Akagi, K., Hayashi, H., Nagae, G., Tsuji, S., Isagawa, T., Midorikawa, Y., Nishimura, Y., Sakamoto, H., Seto, Y., et al. (2010). Three DNA methylation epigenotypes in human colorectal cancer. Clin. Cancer Res. *16*, 21–33. https://doi.org/10.1158/1078-0432.ccr-09-2006.

Yang, S.F., Yeh, C.B., Chou, Y.E., Lee, H.L., and Liu, Y.F. (2016). Serpin peptidase inhibitor (SERPINB5) haplotypes are associated with susceptibility to hepatocellular carcinoma. Sci. Rep. *6*, 26605. https://doi.org/10.1038/srep26605.

Young, L.C., Hartig, N., Boned Del Río, I., Sari, S., Ringham-Terry, B., Wainwright, J.R., Jones, G.G., McCormick, F., and Rodriguez-Viciana, P. (2018). SHOC2-MRAS-PP1 complex positively regulates RAF activity and contributes to Noonan syndrome pathogenesis. Proc. Natl. Acad. Sci. U S A *115*, e10576–e10585. https://doi.org/10.1073/pnas.1720352115.

Yue, D., Fan, Q., Chen, X., Li, F., Wang, L., Huang, L., Dong, W., Zhang, Z., Liu, J., Wang, F., et al. (2014). Epigenetic inactivation of SPINT2 is associated with tumor suppressive function in esophageal squamous cell carcinoma. Exp. Cell Res. *322*, 149–158. https://doi.org/10.1016/j.yexcr.2013.11.009.

Zhao, N., Guo, M., Wang, K., Zhang, C., and Liu, X. (2020). Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. Front. Bioeng. Biotechnol. *8*, 268. https://doi.org/10.3389/fbioe.2020.00268.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA cancer program	National Cancer Institute GDC legacy Archive	https://portal.gdc.cancer.gov/legacy-archive/
PCAWG dataset	ICGC Data Portal	https://dcc.icgc.org/releases/PCAWG
Software and algorithms		
R Version 3.5	The Comprehensive R Archive Network	https://cran.r-project.org/
iClusterPlus version 1.28.0	A.D., Thorsson, V., et al. (2018)	http://www.bioconductor.org/packages/release/bioc/html/iClusterPlus.html
MutSigCV	GenePattern	https:/www.genepattern.org/
Proportional hazards model (Cox model)	Cox.D.R (1972)	https://www.jstor.org/stable/2985181?seq=1#metadata_info_tab_contents
Concordance index (C- index)	Harrell et al. (1996)	https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zhuo Wang (zhuowang@sjtu.edu.cn), and Yongyong Shi (shiyongyong@gmail.com)

Materials availability

Our study did not generate new unique materials.

Data and code availability

Our study did not generate new unique data and the source code have been deposited in github at https://github.com/JZHT-jiangzhou/pan-cancer/tree/source.

METHOD DETAILS

Information of pan-gastrointestinal cancer datasets

We retrieved the omics data of gastrointestinal cancers from the Pan-Cancer project in TCGA Research Network (https://portal.gdc.cancer.gov/legacy-archive/). Overall, 1801 gastrointestinal samples (COAD, READ, ESCA, PAAD, LIHC, STAD) were examined on at least one molecular profiling platform, including 1741 Copy Number Alterations cases, 1745 methylation cases, and 1725 mRNA expression cases. Meanwhile, our study also included 151 normal samples. The omics data from the ICGC:PCAWG database (https://dcc.icgc.org/releases/PCAWG) was used to validate our results, including 806 Copy Number Alterations cases and 281 mRNA expression cases.

SNV and somatic copy number alterations analysis

We used the MutSigCV algorithm on the GenePattern website (https:/www.genepattern.org/) to identify nan-silence gene across the 1741 tumor sample, finding 128 important segments. Similarly, unsupervised hierarchical clustering was performed using Euclidean distance metric and Ward.D2 methods for 1741 samples on gene-level copy numbers generated by GISTIC 2.0 analysis (Mermel et al., 2011).

DNA methylation data analysis

CpG sites were selected based on the β value, concisely, we selected the CpG sites that were not methylated in normal tissues ($\beta < 0.2$) in the selected tumor type and then defined $\beta \geq 0.3$ (Fang et al., 2018) as a

positive group and $\beta < 0.3$ as a negative group. Overall 298 sites were found in at least one tumor, which explaining that more than 5% of the tumor samples belonged to the positive group. Meanwhile, we screened out 577 methylation sites that are methylated in normal samples of any cancer type ($\beta > 0.2$) and found more than 5% of tumor samples belonged to the negative group in at least one tumor. By this method, we selected 875 methylation sites, and then the hierarchical clustering was performed on 1745 samples using row-scaling, Euclidean distance, and ward.D2 clustering.

mRNA expression data analysis

We firstly filtered the genes expressed only in one organ according to the database of Tissue-specific Gene Expression and Regulation (TIGER) (Liu et al., 2008), similarly, we removed the genes which were not expressed in more than 75% of the TCGA pan-GI tumor samples or were absent in any tumor sample, then overall 16,630 genes were retrieved. Moreover, we selected 1037 genes by applying the standard deviation ≥ 0.5 after log10 transformation. Then the Hierarchical clustering was performed on the samples with the 'pheatmap' package in R, using row-scaling, Euclidean distance, and ward.D2 clustering.

Identification of epigenetically silent genes

All methylated probes overlapping with SNPs, repeats, or designed for allosome were removed. The remaining probes were mapped against UCSC genes using the GenomicFeatures R package. Then, we screened all the probes located in the promoter region (the region spanning from 1500 bp upstream to 1500 bp downstream of the transcription start sites) based on the absolute position of the chromosomes. Level 3 mRNA expression data were log10 transformed. The approach used to identify genes epigenetically silenced in gastrointestinal cancer was described as below.

First, all methylated sites in normal tissues with $\beta > 0.2$ were removed. Then, we used $\beta = 0.3$ as a threshold for positive DNA methylation and deleted CpG sites which were methylated in less than 10% of the samples in either type of cancer. Finally, we follow the steps described below to identify genes epigenetically silenced. 1) Divide all cancer samples into the methylated group ($\beta \geq 0.3$) and unmethylated group ($\beta < 0.3$). 2) Then calculate the mean expression in the methylated group and the unmethylated group. 3) Calculate the standard deviation of the expression level of the unmethylated group. 4) Select genes whose mean expression level in the methylated group was 1.64 times standard deviations greater than the unmethylated group. For each sample/gene pair, if a) the sample belongs to the methylated group. b) The expression level of the gene in the sample is lower than the mean expression level of the gene in the unmethylated group, the sample will be labeled as epigenetically silenced for a specific gene. In special cases, if there are multiple methylation sites associated with a gene, the sample will be labeled as epigenetically silenced when at least one methylation site is epigenetically silent.

Integrative clustering using iClusterPlus algorithm

The iClusterPlus is the enhancement of iCluster which uses generalized linear regression for the formulation of the joint model with a set of latent variables that represents driving factors (Cancer Genome Atlas Research Network, 2017). The implementation of iClusterPlus includes the following three steps.

Data processing

Three molecular platforms, SNV, DNA methylation and mRNA expression were provided as input to iClusterPlus. Data were processed using the following procedures. SNV data was filtered to remove genes that demonstrated a mutation rate below 0.05. For the DNA methylation, we select 1000 most variable CpG sites according to standard deviation. For mRNA sequence data, 1000 most variable genes were selected as an input to iClusterPlus after log10 transformed and scaled.

Model selection

iClusterPlus uses the Bayesian Information Criterion (BIC) to determine the number of clusters.

$$BIC = -2\log L + K\log N$$

Where L means the likelihood, N is the number of data that was used to train the model and K is the number of model parameters.

Feature selection

Lasso penalty terms added in iClusterPlus allow us to find the most important factors for dividing classes. For every platform, we select the factors whose sum of fit scores (output by iClusterPlus) across all tumor samples are greater than three quantiles of all features.

Silhouette width

The silhouette width is a widely used measure of within-group homogeneity, which can show the quality of clusters (Lengyel and Botta-Dukát, 2019), the form of silhouette width is:

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i}, & a_i < b_i \\ 0, & a_i = b_i \\ \frac{b_i}{a_i} - 1, & a_i > b_i \end{cases}$$

Where a_i is the average distance between sample i and other samples in the same cluster. b_{ij} is the average distance of sample i to all samples of other clusters C_j, it can also be defined as the inter-cluster dissimilarity of sample i. b_i is the minimum value of the dissimilarity between clusters of sample i. If s_i is close to 1, the clustering result of sample i is reasonable and if it is close to - 1, then the sample i is quite different from other samples.

Survival analysis of the Pan-GI subtypes from iClusters

Kaplan-Meier survival plots were performed with the survival R package. We used the Mantel-Haenszel method to identify the significant difference in the survival curve. For each of the 6 cancer types, we calculate the p value in the Cox model with subtypes of different clusters. If the p value is lower than 0.05, we can statistically believe the subtypes have significant differences in survival.

Proportional hazards model (Cox model)

The proportional Hazards Model (Cox model) is a semiparametric regression model proposed by D.R.Cox in 1972. By modeling the hazard rate function, the Cox model can indicate the association between predictor variables and the survival time of patients. The form of risk rate function is:

$$h(t, X) = lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t, X)}{\Delta t}$$

Cox model assumes that the hazard function has the following form:

$$h(t, X) = \lambda_0(t) \cdot \exp(\beta \cdot X)$$

where λ_0 denotes time-related basic hazard rate, X is the survival-related feature and β is the parameter of the model to be predicted corresponding to features.

For N events, the likelihood function of the i^{th} event when the risk characteristic is X_I *and*occurrence time is t_i as shown above:

$$L(\beta) = \prod_{i=1}^{N} \frac{exp(\beta \cdot X_i)}{\sum_{j:t_j \geq t_i} (\beta \cdot X_j)}$$

we used maximum likelihood estimation (MLE) to estimate the β. When$\beta > 0$, *h (t, X)* increases with *X*, which means *X* corresponds to risk factors for worse survival. On the contrary, when $\beta < 0$, *h (t, X)* decreases with the increase of *X*, so that *X* can be the protective factor for better survival.

To obtain biomarkers that can play a significant role in both stratification and survival, we choose 0.05 as the statistical significance and use Cox model to find the characteristics that play a significant role in survival from different histological levels and tumor types.

Furtherly, after standardizing and centralizing the features selected from the single factor Cox model, we constructed the multi-factor Cox model according to different cancers and screened out the characteristics that have significant effects on survival (choose 0.05 as the statistical significance).

Mutation signature analysis

All single nucleotide mutations can be classified into the following six types (C>A, C>G, C>T, T>A, T>C, T>G). Considering the upstream and downstream base information, for triplet bases, all mutations can be divided into 96 mutation types. Previous studies have found a series of mutation signatures according to the frequency of base mutation, while the Catalog of natural mutations in cancer (COSMIC) summarizes the 30 published mutation signatures of cancers (Alexandrov et al., 2013a, 2013b, 2015; Alexandrov and Stratton, 2014; Helleday et al., 2014; Nik-Zainal et al., 2012, 2016). we performed COSMIC on the mutation frequency of our samples with 'deconstructSigs' package in R.

We analyzed the mutation characteristics of 9 iClusters according to COSMIC with the 'deconstructSigs' package in R. Then we clustered the weights of 30 mutation features obtained from the COSMIC analysis using hierarchical clustering with Euclidean distance and ward.D2 clustering.

Evaluation of the prognostic power of the biomarkers by the concordance index

We applied the concordance index (C- index) to evaluate the prognostic power of the biomarker for each cancer type (Harrell et al., 1996). The C-index is a non-parametric measure to quantify the discriminatory power of a predictive model. If the prediction result is absolutely accurate, then the C-index is 1, while the C-index less than 0.5 indicates the prediction is not effective.

We first calculate the risk score (RS) for each sample of our identified biomarkers to predict the patients' risk (Zhao et al., 2020). The RS for the gene in each sample can be calculated:

$$RS = B \cdot V$$

Where $B = [\beta_1, \beta_2, \beta_3]$, each β is a risk coefficient so that the B can show the risk coefficient of each omics including gene expression, DNA methylation, and somatic copy number alterations. While $V = [v_1, v_2, v_3]$shows the value of three types of omics data in each sample. Finally, we use the boosting to generate 100 C-indexes of the RS. If the median value of the C-index was higher than 0.5, then our model can be credible.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed in R version 3.5.

For each single-omics platform, the details of data filtering, model selection and silent genes identification have been indicated in the respective method details.

For multi-omics analysis, iClusterPlus was used to classify the pan-GI cancer samples, in which we choose $\lambda = 35$ (lambda determines the number of molecular features that have nonzero weights on the fused eigen-feature) and $P \leq 0.05$ were considered significant.

The details of proportional hazards model and concordance index can be found in the respective method details.

The independent data from ICGC dataset is used for repetitive experiments and the selected features and the methodology is repeatable, the details have been included in the previous sections.