Xu *et al. BioData Mining* (2025) 18:9
https://doi.org/10.1186/s13040-025-00426-z

BioData Mining

METHODOLOGY Open Access

A generative deep neural network for pan-digestive tract cancer survival analysis

Lekai Xu[1], Tianjun Lan[2,3], Yiqian Huang[1], Liansheng Wang[2,3], Junqi Lin[1], Xinpeng Song[1], Hui Tang[1], Haotian Cao[2,3] and Hua Chai[1*]

Lekai Xu and Tianjun Lan are the co-first authors of this article.

*Correspondence:
Hua Chai
chaih1989@fosu.edu.cn
[1]School of Mathematics, Foshan University, Foshan 528000, China
[2]Department of Oral and Maxillofacial Surgery, Sun Yat-sen Memorial Hospital of Sun Yat-Sen University, Guangzhou 510010, China
[3]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, Medical Research Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China

Abstract

Background The accurate identification of molecular subtypes in digestive tract cancer (DTC) is crucial for making informed treatment decisions and selecting potential biomarkers. With the rapid advancement of artificial intelligence, various machine learning algorithms have been successfully applied in this field. However, the complexity and high dimensionality of the data features may lead to overlapping and ambiguous subtypes during clustering.

Results In this study, we propose GDEC, a multi-task generative deep neural network designed for precise digestive tract cancer subtyping. The network optimization process involves employing an integrated loss function consisting of two modules: the generative-adversarial module facilitates spatial data distribution understanding for extracting high-quality information, while the clustering module aids in identifying disease subtypes. The experiments conducted on digestive tract cancer datasets demonstrate that GDEC exhibits exceptional performance compared to other advanced methodologies and can separate different cancer molecular subtypes that possess both statistical and biological significance. Subsequently, 21 hub genes related to pan-DTC heterogeneity and prognosis were identified based on the subtypes clustered by GDEC. The following drug analysis suggested Dasatinib and YM155 as potential therapeutic agents for improving the prognosis of patients in pan-DTC immunotherapy, thereby contributing to the enhancement of cancer patient survival.

Conclusions The experiment indicate that GDEC outperforms better than other deep-learning-based methods, and the interpretable algorithm can select biologically significant genes and potential drugs for DTC treatment.

Key Points

- The proposed competitive Graph-based Deep Generative Neural Network (GDEC) presents a novel approach for cancer subtyping, enabling the identification of both statistically and biologically significant cancer subtypes.
- Based on the subtypes clustered by GDEC, we have identified 21 hub genes that are significantly associated with the heterogeneity and prognosis of pan-digestive tract cancer.
- The drug analysis suggested Dasatinib and YM155 as potential therapeutic agents for improving the prognosis of patients in pan-digestive tract cancer immunotherapy.

Keywords Cancer subtype, Tumor heterogeneity, Survival analysis, Pan-digestive tract cancer analysis

Introduction

The digestive tract cancer (DTC) heterogeneity driven by multiple genomic alterations, is a major obstacle that hampers the effectiveness of therapies for patients [1, 2]. Identifying cancer subtypes by unsupervised clustering methods can offer valuable insights to guide clinical decision-making and facilitate the identification of potential biomarkers [3]. According to the global cancer report, DTCs account for 25.8% of new cases and contribute to 35.4% of deaths worldwide in 2022 [4]. Accurate subtyping of cancer can provide valuable insights for designing treatment strategies and developing therapeutics [5, 6]. In this study, we investigate the pan-DTC subtype analysis by integrating six types of cancers (colon, esophageal, liver, pancreatic, rectum, and stomach) to identify molecular subtypes, recognize latent cancer-related biomarkers, and discover the potential drugs for DTC treatment.

In the original study, traditional machine learning techniques such as k-means and hierarchical clustering were employed to discern similarities and dissimilarities among cancer patients with different molecular subtypes. With advancements in sequencing technologies, there is an increasing demand to cluster high-throughput patient features for disease subtyping. Early studies primarily relied on dimensionality reduction algorithms to reconstruct the low-dimensional representation, which was subsequently followed by conventional clustering algorithms. For instance, Alex employed the principal component analysis (PCA) to extract gene information for breast cancer subtype clustering [7]; CIDR proposed by Lin quantified dissimilarity between cells based on Euclidean distance between the high-dimensional gene expression, and then performed hierarchical clustering to detect disease subtypes [8]; Becht designed a nonlinear dimensionality reduction method based on uniforming manifold approximation to analyze single-cell data of the cancers [9].

One limitation of these methods is that the majority of these algorithms employ linear transformation for data dimensionality reduction, which may not adequately address complex gene features for extracting the effective information to distinguish disease subtypes residing in non-linear manifolds within a high-dimensional space [3]. In recent studies, various deep learning (DL) -based technologies have been employed to address this tissue [10, 11]. Chaudhary initially utilized the Autoencoder for extracting disease multi-omics features in hepatocellular carcinoma survival analysis [12]. In Guo's work on cancer subtyping, the denoising Autoencoder was adopted as a replacement for the Autoencoder to enhance model robustness against the noise [13]. Yang designed Subtype-GAN to enhance the learning of diverse distribution knowledge about patients by leveraging the Gaussian mixture model [14]. Han proposed a data augmentation-based contrastive learning network for integrating single cell RNA-seq data in disease clustering [15]. To address the increasing dimensionality of gene features, Cheng proposed scGAC, which integrates gene connection information using a graph-based autoencoder to extract crucial patient information for cancer survival analysis [16].

However, in these DL-based methods, the representation learning module and clustering module are employed as separate components, which may result in suboptimal

clustering outcomes because the deep neural network cannot benefit from mutual task optimization. To address this issue, Guo proposed IDEC, a multi-task deep neural network that combines a local structure preservation module with a cluster module [17]. Chen developed a semi-supervised learning deep neural network for cancer subtyping [3]. However, the reliance on labeled samples in this framework restricts its applicability. In a recent study, Gan employed a self-supervised learning framework to integrate the data augmentation strategy with the deep clustering multi-task deep neural network [18]. Despite advancements in meticulously designed unsupervised clustering methods for capturing tumor heterogeneity, cancer clustering remains an exceedingly challenging task due to computational issues arising from high-dimensional omics data, which can lead to ambiguous and overlapping patient subtypes.

To address the limitation caused by the high-dimensional and complex cancer data, we propose GDEC, an end-to-end graph-based generative deep neural network for clustering DTC subtypes. GDEC is optimized using an integrated loss function comprising two modules: a generative-adversarial module to capture spatial distribution knowledge of the data, facilitating extraction of high-quality patient information; and a clustering module to identify the disease subtypes. The integration of the modules enables the deep neural network to benefit from the optimization of these two tasks. Moreover, a graph convolution layer in GDEC is employed to incorporate additional gene connection information associated with cancers.

The experiment demonstrates that GDEC exhibits a competitive performance in cancer subtyping, surpassing other methods in terms of clustering efficacy and enabling the identification of biologically significant cancer subtypes. Considering the commonly used TCGA pan-cancer atlas has classified cancers into a variety of new pan-cancer subtypes according to similar molecular expression [19]. Despite the apparent specificity of each tumor class, molecular variants are often integrated into established biological pathways that are shared by different tumor types [20]. Studies of relatively rare cancers would benefit from the results of pan-cancer analysis [21, 22]. Based on the pan-DTC subtypes clustered by GDEC, we identified 21 hub genes associated with intra-tumor heterogeneity and prognosis. Further drug analysis identified Dasatinib [23] and YM155 [24] as potential therapeutic agents for improving the prognosis of patients in DTCs. We hope our findings can contribute to prolonging the survival of these patients.

Methods

Datasets

Six datasets (COAD (colon), ESCA (esophageal), LIHC (liver), PAAD (pancreatic), READ (rectum), and STAD (stomach)) obtained from TCGA [25] and ten additional cancer datasets (GSE91061 [26], GSE10186 [27], GSE14333 [28], GSE17538 [26], GSE54236 [29], GSE57495 [30], GSE84437 [31], GSE78220 [28], GSE135222 [31], and IMvigor 210 [32]) collected from public databases were used in this study. The expression data underwent log transformation normalization, with features being excluded if they contained more than 20% missing values. For the remaining samples, missing values were imputed using median values. After preprocessing, the R package "*limma*" [33] was adjusted to remove batch effects. The detailed data information are given in Supplementary Table S1.

Deep learning framework for cancer subtyping

As depicted in Fig. 1(A), the proposed deep learning neural network incorporates a GCN layer to extract the effectiveness information of cancer features by leveraging the biological prior knowledge regarding gene connection pathways. Assuming the cancer dataset comprises N patients with gene features $X = (x_1, x_2, \ldots, x_p)$, and the input connection graph G is constructed using the KEGG pathway information. The output of the GCN layer is expressed as:

$$X_g = \sigma\left(\widetilde{D}^{-\frac{1}{2}} A' \widetilde{D}^{-\frac{1}{2}} X W^{(l)}\right) \tag{1}$$

where X' denotes the output of the GCN layer, is the adjacency matrix, and D is the degree matrix in the constructed graph G, $W^{(l)}$ represents the coefficient matrix in the deep neural network. The activation function $\sigma(\cdot)$ used in this study is *RELU*. Subsequently, the GCN output features are fed into a fully connected layer, which performs nonlinear transformations and enhances the graph structure features extracted by GCN. By leveraging the fully connected layer, we can effectively map the features generated by GCN to a potentially more representative space, thereby optimizing them further in the following generative-adversarial module.

To enhance the extraction of crucial information regarding cancer patients, a generative-adversarial module is incorporated into the deep neural network to capture knowledge about the spatial distribution of data.

For cancer genomics data with a complex distributed structure, the generative adversarial network can help extract representative and more discriminative potential features. This is particularly critical for clustering cancer subtypes, as the data structures of different cancer subtypes may have highly nonlinear patterns that may be difficult to fully capture using autoencoders alone. Additionally, the generator of a generative adversarial network can effectively handle data noise and provide robust feature representations.

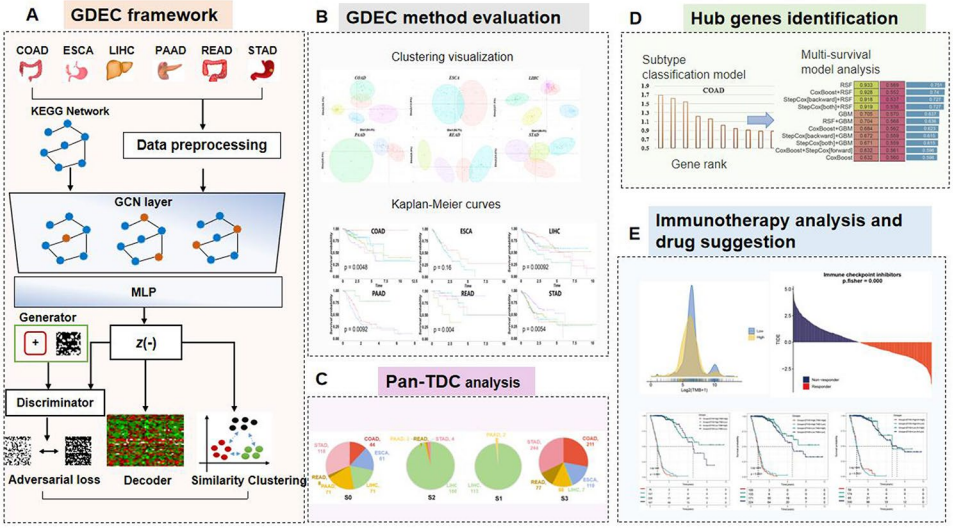

Fig. 1 The architecture of proposed GDEC framework for digestive tract tumor subtyping. (**A**). The deep neural network in GDEC for digestive tract tumor clustering. (**B**). The clustering performance obtained by GDEC evaluation in different datasets. (**C**). The Pan-DTC analysis by subtyping with GDEC. (**D**). Hub gene identification by using different interpretable machine learning methods. (**E**). The pan-DTC immunotherapy analysis and drug suggestion with identified hub genes

The generative-adversarial module comprises three components: an encoder, a decoder, and a discriminator. The encoder generates a non-linear low-dimensional representation of the cancer patient features Z, while the decoder reconstructs the compressed features back to the original input. Furthermore, the discriminator is used to force the compressed features to adhere to the prior distribution. The loss l_e that is used to calculate the Euclidean distance between input X_g and output X'_g which can be expressed as:

$$l_e = ||X_g - X'_g||_2^2 \tag{2}$$

The loss function in the discriminator can be decomposed into two components: the generator loss and the discriminator loss. The generator loss quantifies the discrepancy between the compressed data generated by the encoder and real samples, which the loss function is written as:

$$l_g = -\frac{1}{n}\sum\nolimits_{i=1}^{n}(log(d_{fake_i})) \tag{3}$$

The objective of the discriminator loss is to accurately discern between the real and generated sample while minimizing the misclassification by the discriminator:

$$\begin{aligned}\min_{Q}\max_{D} & E_{Z'\sim P(Z)}\left(\log\left(D\left(Z'\right)\right)\right)+\\ & E_{Z\sim Q(Z)}\left(\log\left(1-D\left(Z\right)\right)\right)\end{aligned} \tag{4}$$

The discriminator loss l_d can be expressed as:

$$\begin{aligned}l_d = & -E_{Z'\sim P(Z)}\left(\log\left(D\left(Z'\right)\right)\right)\\ & -E_{Z\sim Q(Z)}\left(\log\left(1-D\left(Z\right)\right)\right)\\ & -E_{Z\sim Q(Z)}\left(\log\left(D\left(Z\right)\right)\right)\end{aligned} \tag{5}$$

Hence the total loss of the generative-adversarial module is given as:

$$l_a = l_e + \alpha\, l_d \tag{6}$$

To facilitate the integration of diverse tasks in algorithmic optimization and enhance method usability, we constructed the end-to-end deep neural network by merging the generative-adversarial module with the clustering task. The clustering model was completed to minimize the following objective for cancer subtyping:

$$l_c = \sum\nolimits_i \sum\nolimits_j p_{ij} log\frac{p_{ij}}{q_{ij}} \tag{7}$$

where q_{ij} can describe the similarity between the cluster center μ_j and embedded point z_j by Student's t-distribution:

$$q_{ij} = \frac{\left(1+||z_i-\mu_j||^2\right)^{-1}}{\sum_j\left(1+||z_i-\mu_j||^2\right)^{-1}} \tag{8}$$

The p_{ij} is the target distribution written as:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j \left(q_{ij}^2 / \sum_i q_{ij} \right)} \tag{9}$$

The total loss function in GDEC can be calculated as: $l_{GDEC} = l_c + \gamma L_a$, where the value of γ was set to 0.1 to balance the importance of generative-adversarial loss l_a and clustering loss l_c.

Feature importance evaluation

Enhanced interpretability of deep neural networks is crucial for our tasks, as a comprehensive understanding of the decision-making process employed by models can foster trust in the clustering outcomes and offer valuable biological insights. The interpretative algorithm multiple classification random forest was used to fit the clustering results of black box deep neural networks, for calculating the individual contribution of each feature gene. This approach provides a more intuitive interpretation of the clustering outcomes and facilitates the identification of crucial gene features associated with tumor subtypes. Random Forest (RF) is an ensemble learning-based approach that combines with K trees $(T_1(x_1, y_1) \ldots T_k(x_n, y_n)$, where x represents the input features and y denotes the cancer subtype labels distinguished by the deep neural network. The RF-based classification model can be formulated as follows:

$$\widehat{y_i} = \varnothing(x_i) = \sum_{k=1}^{K} f_k(x_i), \; f_k \in \mathcal{F} \tag{10}$$

where $\mathcal{F} = \left\{ f(x) = w_{q(x)} \right\} \left(q : \mathbb{R}^m \to T, \; w \in \mathbb{R}^T \right)$ can be regarded as the classification trees integration, q represents the classification trees structure, T denotes the leave numbers in each tree, and f_k corresponds to the tree with weight w and structure q. Based on the results calculated by RF, the genes Gini scores>0.2 are considered significant in relation to pan-DTC heterogeneity.

Pan-DTC biomarker identification

To further narrow down the range of potential targets related to pan-DTC, we further explored the relationship between the prognosis and the important genes screened by RF, The impact of these genes on cancer prognosis was evaluated by constructing a comprehensive set of six machine learning algorithms (CoxBoost, stepwise Cox regression, generalized boosted regression models (GBMs), supervised principal components (SuperPC), partial least squares Cox regression (plsRcox), and Random Survival Forests (RSF)). The TCGA-DTC cohort was utilized as the training set, while DTC datasets collected from the GEO database were defined as the validation set. The development pipeline for biomarker identification is outlined as follows:

1. Univariate Cox regression analysis was performed within both TCGA-DTC and GEO-DTC cohorts. Genes that exhibited a p-value of less than 0.05 and maintained consistent hazard ratio (HR) directionality across both two cohorts were identified as Stable Prognosis-Related Genes (SPRGs).
2. A comprehensive approach was employed by utilizing six machine learning algorithms. In total, 32 unique combinations of these algorithms were explored to formulate the most predictive and interpretable prognosis evaluation model, with the primary objective of achieving optimal concordance index (C-index) performance.

3. Upon establishing the model using the training set, we proceeded to evaluate its accuracy across all validation cohorts. This involved calculating the average C-index for each model configuration. The interpretable model with the highest average C-index was considered as the optimal solution, identifying the prognosis-related genes that could serve as potential biomarkers.

Method performance evaluation

The -$\log_{10}(p)$ value is employed as a main metric to assess the disparity in survival among patients within distinct cancer subtypes. A higher -$\log_{10}(p)$ value indicates superior clustering performance, while a lower value suggests no significant variation in the survival of cancer patients with different subtypes. Additionally, the average silhouette coefficient (SC), Davies-Bouldin Index (DBI), Dunn Index (DI), and Squared Error with cosine distance (SSE) were employed to assess the clustering performance under various hyperparameters in GDEC.

The hyperparameters (batch size, learning rate and the node number in the middle-hidden layer) in GDEC were determined based on the average silhouette coefficients of the clustering results. The batch size (BS) for the deep neural network was set [8, 16, 3], the learning rate (LR) was set [1e-4, 1e-5, 1e-6], and the node number was set [10, 20, 50]. The hyperparameter sensitivity analysis results are given in the Supplementary Table S2.

The selection of cluster values k, ranging from 2 to 10 for different types of DTCs, is determined by calculating the within sum of squares (wss) using the k-means algorithm in the initial stage. The determined subtypes of different cancers are given in Table 1. Additionally, to verify whether our results are biologically meaningful, this study applied a variety of biological analysis means, as detailed in the *Biological Analysis* section in the supplementary file.

Results

GDEC clustering performance evaluation

In Table 1, we present a concise summary of the DTC subtype identification performance achieved by GDEC. COAD and LIHC are clustered into five subtypes, while READ and STAD are clustered into four subtypes. ESCA has two subtypes, whereas PAAD has three. The log-rank p-values among different cancer subtypes obtained by our method are below 0.01(-$\log_{10}(p) = 2$), except for ESCA, which exhibits significant differences in survival between patients with distinct subtypes. The values of SC, DBI, DI and SSE are listed in the following columns of Table 1.

Table 1 The results of the clustered DTC subtypes

Cancer	Subtypes	-$\log_{10}(p)$	SC	DBI	DI	SSE
COAD	5	2.317	0.531	0.704	0.094	0.003
ESCA	2	0.796	0.443	0.798	0.021	1.482
LIHC	5	3.035	0.628	0.512	0.184	0.003
PAAD	3	2.036	0.579	0.506	0.029	1.039
READ	4	2.401	0.562	0.466	0.054	6.001
STAD	4	2.267	0.396	0.861	0.028	0.0001
Pan-DTC	4	5.585	0.807	0.203	0.090	0.0004
Average	-	2.636	0.564	0.579	0.071	1.218

As depicted in Fig. 2A, we present the visualization of GDEC clustering results across six DTCs. The distinct subtypes of cancer patients are color-coded and projected onto a two-dimensional plane using principal component analysis. The visualization of clustering results reveals a discernible disparity in the distribution of patients across distinct subtypes within the same cancer. Our findings demonstrate that GDEC exhibits favorable visualization performance in most cancers; however, ESCA presents mixed clustered cancer subtypes (p=0.16). Furthermore, in Fig. 2B we present the Kaplan-Meier survival curves of the DTCs drawn based on the distinguished subtypes identified by GDEC. These results demonstrate that for most cancers the p-values among different cancer subtypes were consistently below 0.01. Thus, our method successfully identifies statistically and biologically meaningful cancer subtypes with significant differences in patient survival observed across various subgroups.

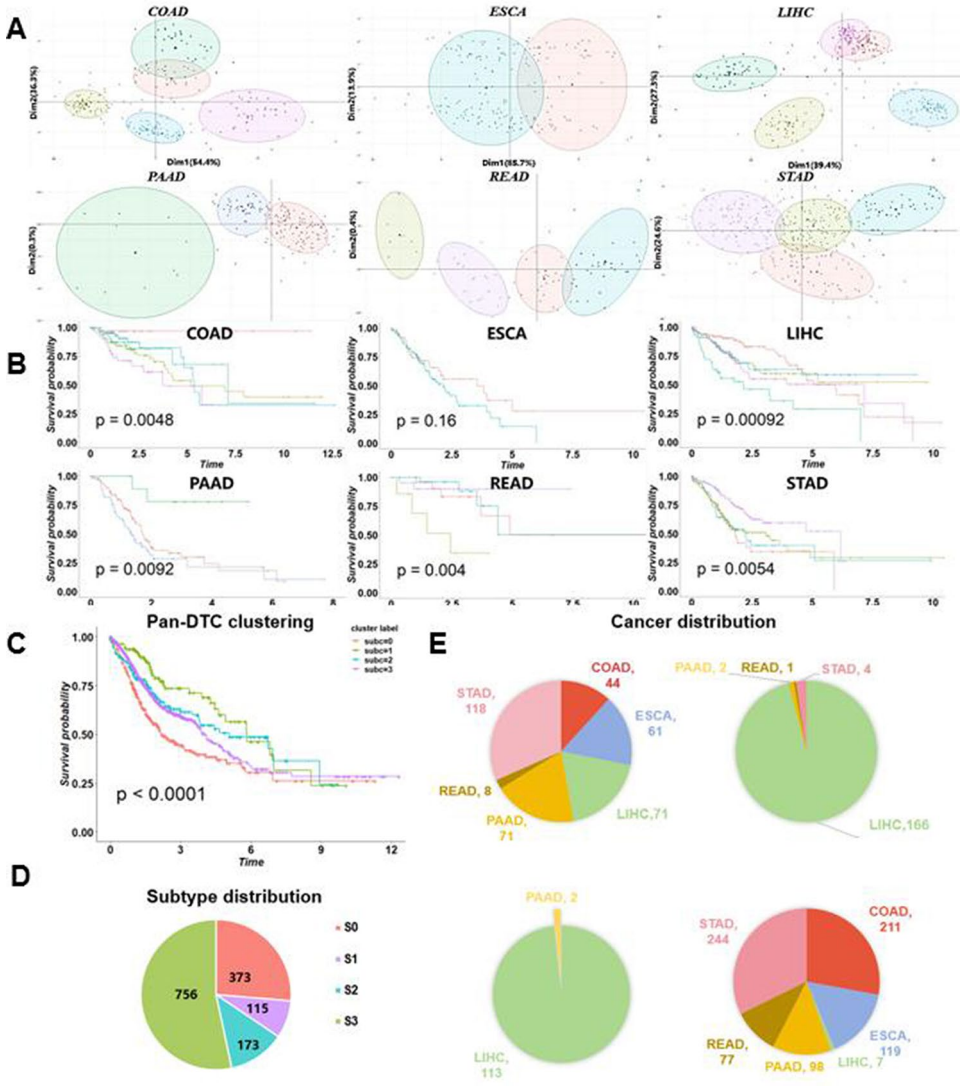

Fig. 2 Clustering results evaluation obtained by GDEC. (**A**). The visualization of GDEC clustering results in the DTC datasets. (**B**). The Kaplan-Meier survival curves of six DTCs (COAD, ESCA, LIHC, PAAD, READ, STAD) drawn based on the distinguished subtypes by GDEC. (**C**). The survival curves drawn based the clustered subtypes by GDEC. (**D**). The number of pan-DTC patients distributed in different subtypes. (**E**). The individual proportion of different cancers in each pan-DTC subtype

To assess the effectiveness of our approach in cancer subtyping, GDEC was compared with six methods: the k-means, the sparse k-means (sparseK), denoising Autoencoder with k-means (DAE-KM), scGAC [16], DECC [34] and ProgCAE [35]. While the k-means and sparseK belong to traditional clustering methods, the others are deep learning-based clustering frameworks proposed in recent years. The performance obtained by GDEC was compared with the mentioned methods on the DTC datasets collected from the TCGA database. Table 2 gives the $-\log_{10}(p)$ values obtained by different computational methods. As shown in Table 2, GDEC gets the highest scores ranging from 0.796 (ESCA) to 3.025 (LIHC), with an average of 2.142. Compared to other methods, GDEC demonstrated a significant improvement of 0.728 (average $-\log_{10}(p) = 1.414$). The two traditional methods (k-means and sparseK) obtained lower average $-\log_{10}(p)$ values compared to DL-based methods (average $-\log_{10}(p) = 1.58$). ProgCAE outperformed DAE-KM and scGA but performed worse than DECC. Among all comparison methods, DECC achieved the highest average $-\log_{10}(p)$ value of 1.813; however, it still falls short compared to GDEC performance (2.142).

The survival analysis of pan-DTC subtypes

After verifying the reliability of our method, we performed GDEC in pan-DTC subtyping. Figure 2C shows four subtypes were clustered and the significant differences in pan-DTC patient survival are observed across various subgroups ($p < 0.001$). In these four pan-DTC subtypes, there are 373 individuals in subtype S0, 115 individuals in subtype S1, 173 individuals in subtype S2, and 756 individuals in subtype S3 (Fig. 2D). In Fig. 2E, we also present the individual proportion of different cancers in each pan-DTC subtype.

Notably, subtype S1 exhibited the most favorable prognosis among all evaluated subtypes. Presently, the molecular subtyping of DTC predominantly hinges on molecular expression profiles, often associated with discrete biological functionalities. Thus, our investigation sought to delve into the unique molecular attributes characterizing these subtypes. Recognizing the pivotal influence of tumor immune microenvironment on tumorigenesis and disease progression, we meticulously quantified the infiltration levels of immune cells. Compared to other subtypes, S1 demonstrated a higher abundance of immune cells within its TME, particularly anti-tumor immune cells as determined by various algorithms including TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, EPIC and ssGSEA (Fig. 3A, C). These findings suggest that subtype S1 represents an immune-activated microenvironment.

To delve deeper into transcriptomic disparities, our study undertook an exhaustive analysis targeting putative regulators intricately linked with cancer chromatin remodeling, encompassing examination of 23 transcription factors (TFs) pertinent to DTC. (Fig. 3B). The strong correlation between regulator activity and subtypes confirms

Table 2 log10(p) values obtained by various methods across nine cancer datasets

	k-means	sparseK	DAE-KM	scGAC	DECC	ProgCAE	GDEC
COAD	0.661	0.844	1.200	1.465	1.278	1.522	2.317
ESCA	0.151	0.770	0.277	0.316	0.569	0.630	0.796
LIHC	0.982	1.725	1.668	2.015	1.301	2.122	3.035
PAAD	1.528	0.839	2.091	1.378	3.293	1.873	2.036
READ	1.054	1.224	1.460	2.524	2.255	1.954	2.401
STAD	1.832	1.381	1.507	1.076	2.184	1.970	2.267
Average	1.035	1.131	1.367	1.462	1.813	1.678	2.142

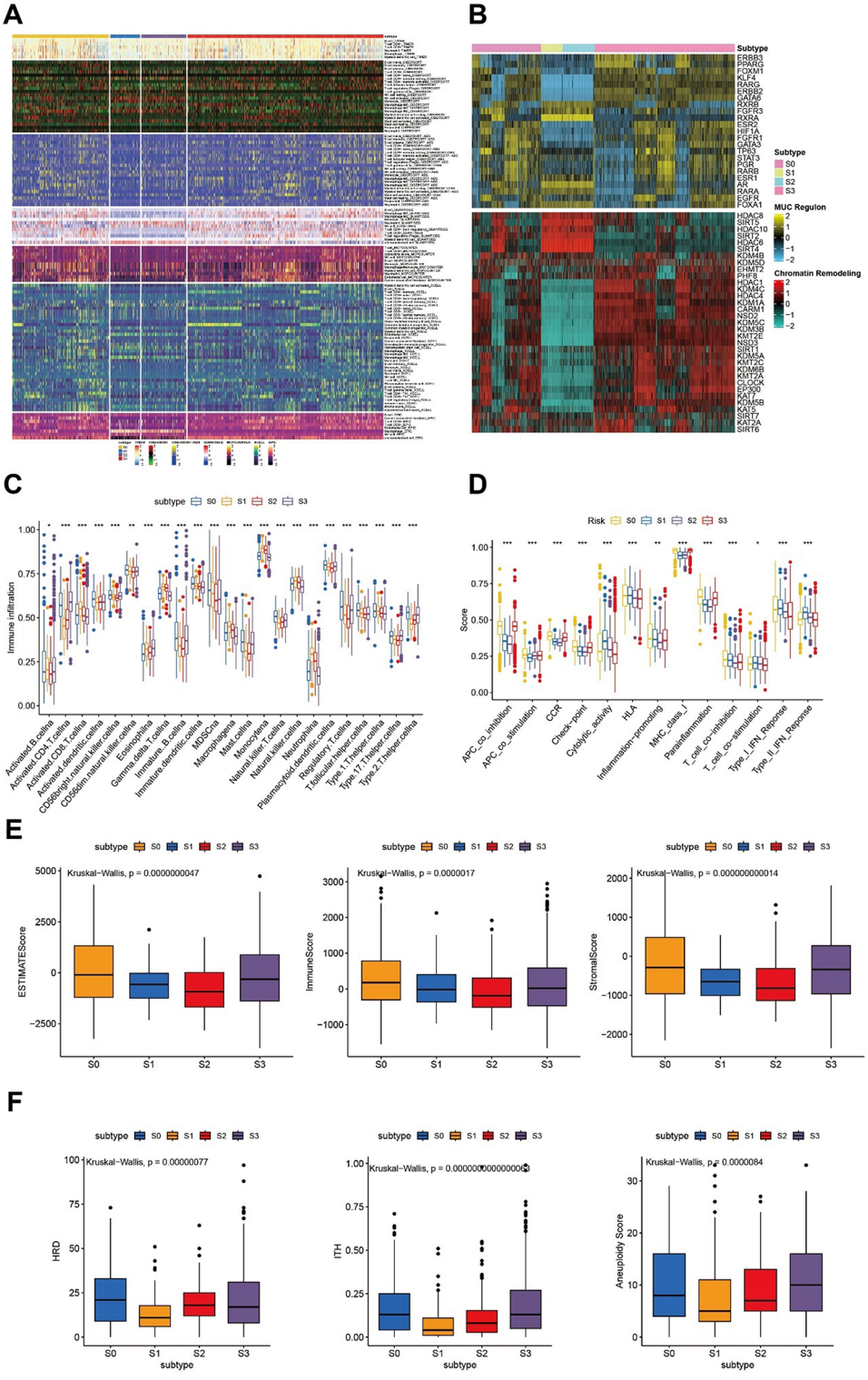

Fig. 3 Molecular landscapes of DTC subtypes. (**A**). Heatmap displaying the infiltration of immune cells across four subtypes, analyzed using a suite of algorithms including TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC. (**B**). Regulon activity profiles of 23 transcription factors (TFs) (top) and potential regulators involved in chromatin remodeling (bottom) across four subtypes. (**C**). Analysis of immune cell infiltration in four subtypes using the ssGSEA method. (**D**). Immune function analysis across four subtypes employing the ssGSEA method. (**E**). Comparison of ImmuneScore, StromalScore, and EstimateScore for the four subtypes using ESTIMATE algorithms. (**F**). Comparison of homologous recombination deficiency (HRD), intratumor heterogeneity (ITH), and aneuploidy scores across four subtypes

the biological relevance of these subtypes. Notably, Androgen Receptor (AR), ERBB2, RXRA, FGFR3, and FOXA1 regulators were significantly activated in S1 and S2, while EGFR, GATA3, TP63, HIF1A, and STAT3 showed specific enrichment in S0 and S3 (Fig. 3B). The regulon activity profiles associated with cancer-related chromatin remodeling underscore potential patterns of divergent regulation across molecular subtypes. This observation suggests that epigenetically mediated transcriptional networks play a pivotal role as discerning factors among these distinct molecular subtypes (Fig. 3B). Subsequently, we explored the relationship between subtypes and immune patterns. As depicted in Fig. 3C and D using the ssGSEA method, patients classified under S1 exhibited more prominent immune-related functions. Consistently supporting this finding is our demonstration that patients belonging to S1 displayed elevated immune and stromal scores (Fig. 3E).

Interestingly, we observed higher expression levels of immune checkpoint-related genes, as well as MHC I and II related genes in patients belonging to S1 (Fig. 4A and C). Therefore, we utilized immune phenotype score (IPS) data to assess the response of four

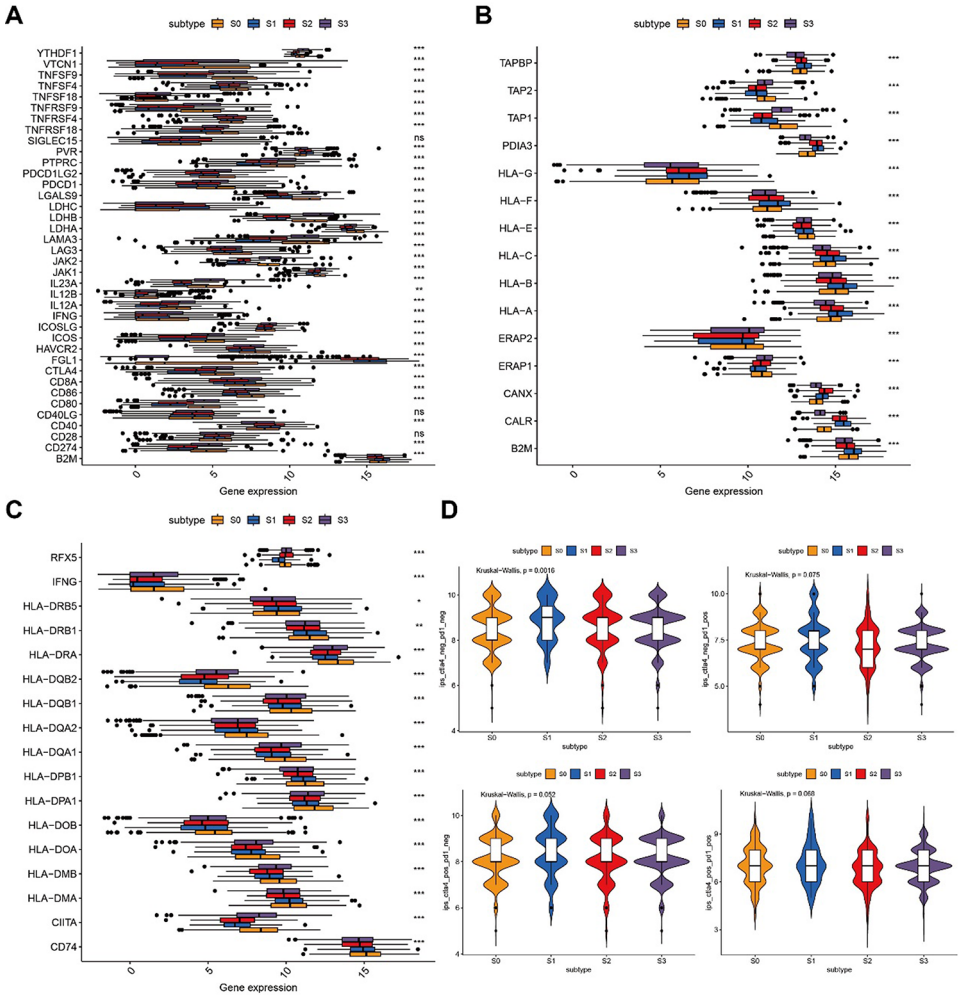

Fig. 4 The association between subtypes and immune-related functions. (**A**). Comparison of immune checkpoint genes among four subtypes. (**B**). Comparison of MHC I genes among four subtypes. (**C**). Comparison of MHC II genes among four subtypes. (**D**). Comparison of IPS scores of four subtypes among CTLA4_Positive PD1_Positive, CTLA4_Negative PD1_Negative, CTLA4_Negative PD1_Positive and CTLA4_Positive and PD1_Negative subgroups (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$)

patient subgroups treated with different immune checkpoint inhibitors (ICIs), including anti-PD-1 and anti-CTLA-4. As depicted in Fig. 4D, patients within the S1 group exhibited significantly elevated IPS scores in the CTLA4_Negative PD1_Negative subgroup ($p<0.05$). Additionally, we evaluated other immunogenic biomarkers such as HRD, ITH, and aneuploidy across all four clusters. Notably, compared to the other clusters, cluster S0 demonstrated heightened tumor immunogenicity (Fig. 3F).

Hub gene validation

Based on the pan-DTC biomarker identification results (Fig. S1), 21 hub genes were identified as associated with intra-tumor heterogeneity and prognosis. The GSCALite public server (http://bioinfo.life.hust.edu.cn/web/GSCALite/) was used to analyze their expression patterns across diverse tumor types in TCGA. Our analysis revealed that several genes, including LAMC, TNFRSF12A, and C2, exhibited significantly elevated expression levels in multiple cancer tissues (Fig. 5A). Furthermore, we observed a positive correlation between mRNA expression levels and copy number variations (CNVs) of the selected genes across most cancer types, particularly highlighting DVL3 as a prominent example (Fig. 5B). Analysis of CNV frequency alterations demonstrated substantial disparities in the CNVs of the identified genes among different cancer types, with PEA15 displaying the highest frequencies primarily characterized by heterozygous amplification events (Fig. 5C and D).

Furthermore, we observed substantial differences in the methylation levels of these genes between tumor and normal samples in most cancer specimens (Fig. 5E). Notably, there was a negative correlation between the methylation levels and mRNA expression levels of these genes across various cancers (Fig. 5F), suggesting that epigenetic modifications mediated by these genes may impact patient prognosis. Furthermore, our results elucidated that the identified genes play a pivotal role in activating the epithelial-mesenchymal transition (EMT) pathway, concurrently exerting a substantial inhibitory influence on the cell cycle pathway. (Figure 6A and B). To further validate the prognostic significance of these hub genes, we performed Kaplan-Meier analysis using data from the Biomarker Exploration for Solid Tumors (BEST) database (https://rookieutopia.com/app_direct/BEST/). The results obtained were consistent with those derived from Cox regression analysis (Fig. 6C). Additionally, these genes exhibited significant associations with both disease-specific survival (DSS) and progression-free survival (PFS) in DTC patients, underscoring their profound prognostic relevance for individuals affected by this condition (Fig. 6C).

Potential therapeutic drug screening

Based on the identified hub genes in this study, we try to screen the potential therapeutic drugs for pan-DTC treatment. Significant differences in prognosis were observed between high-risk and low-risk populations (Fig. S2-S3), as evidenced by Gene Set Enrichment Analysis (GSEA) which revealed significant activation of pathways such as epithelial-mesenchymal transition (EMT), angiogenesis, and TGF-beta in high-risk patients (Fig. 7A). Considering the poor response to immunotherapy observed in high-risk patients (Fig. S4), we employed the Cancer Therapeutics Response Portal (CTRP) and Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) databases to identify potential therapeutic drugs for this patient group. To validate the reliability of our

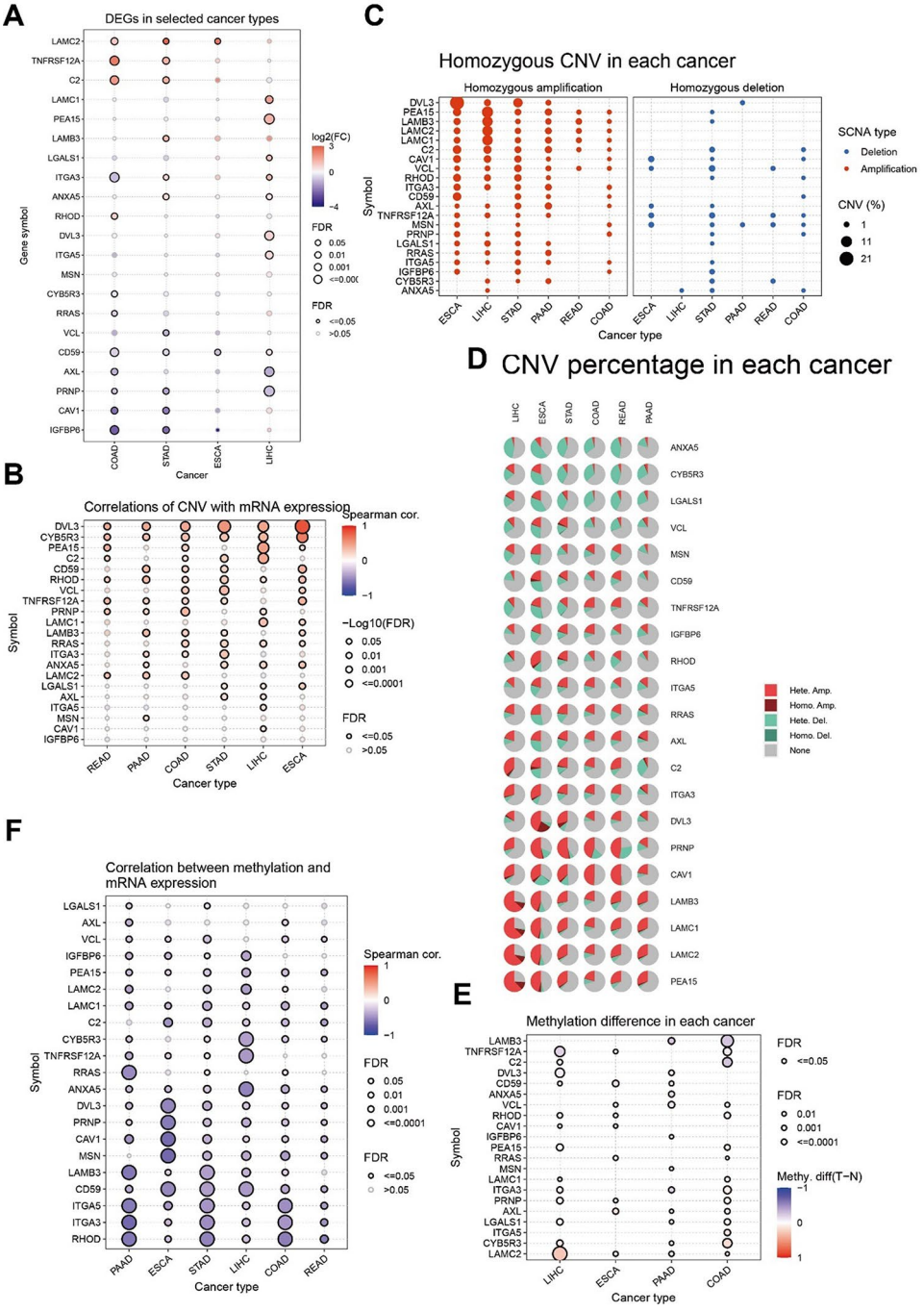

Fig. 5 Differential expression analysis and CNV-related analysis of selected hub genes among six DTCs. (**A**). Differential expression analysis of hub genes. (**B**). Correlation analysis between CNV and mRNA expression level of hub genes. (**C**). The CNV pie distribution indicates the constitution of Heterozygous/Homozygous CNV of each hub genes in each cancer. (**D**). The distribution of heterozygous and homozygous CNV of hub genes in each cancer. (**E**). Differential methylation expression analysis of hub genes. (**F**). Correlation analysis between methylation and mRNA expression level of hub genes

methodology, cisplatin was used as a benchmark since it is commonly employed for DTC treatment. Our algorithm predicted sensitivities consistent with established clinical outcomes, corroborating a previous study that identified ERCC1 as a prognostic biomarker in advanced DTC patients receiving cisplatin-based chemotherapy. Notably, our

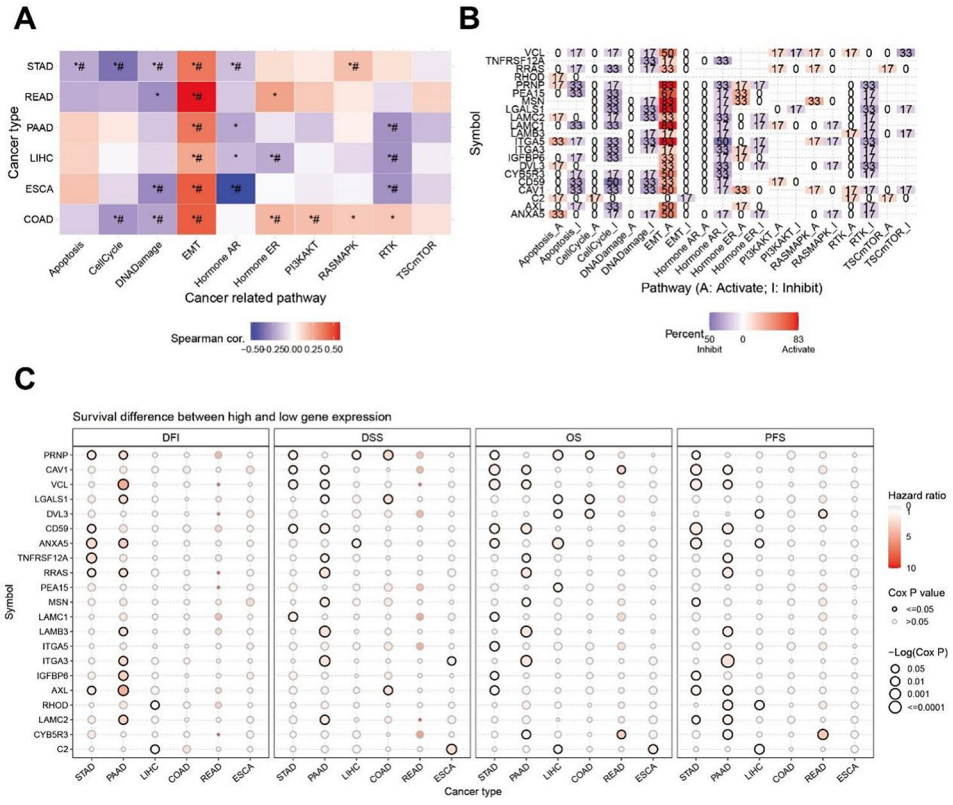

Fig. 6 Pathway analysis and prognosis analysis of hub genes among six DTCs. (**A**). The heatmap shows the activation of pathways among six DTCs. (**B**). The heatmap shows the activation and repression status of hub genes on related pathways. (**C**). Correlation analysis between prognosis and mRNA expression level of hub genes

analysis demonstrated that low expression levels of ERCC1 were associated with a favorable response to cisplatin therapy, suggesting a potential therapeutic advantage (Fig. 7B).

Subsequently, we conducted a systematic investigation to identify potential drugs for high-risk patients based on previous research findings. From the CTRP database, we identified five promising candidates (Tamatinib, Dasatinib, YM-155, Birinapant, Canertinib; Fig. 7C), and from the PRISM database, six candidates showed promise (tedizolidphosphate, Dasatinib, YM-155, PHA-793887, Litronesib and LY2606368; Fig. 7D). To further validate the efficacy of these compounds in treating high-risk patients, comprehensive literature searches were performed using PubMed database. Ultimately, our analysis revealed that Dasatinib and YM-155 exhibited significant potential as therapeutic agents for pan-DTC patients.

Discussion

Although our method has been demonstrated to provide reliable cancer subtype labels, there are still several pertinent questions that warrant discussion. Firstly, the integration of different omics information has been proven to yield valuable insights into cancer patients [36, 37, 38]. Hence, incorporating multimodal data (such as images, epigenome data, clinical data, etc.) may obtain more comprehensive information and improve the clustering performance of our proposed method. Additionally, patient data is characterized by high value and low volume. Currently, single cell sequencing technology is rapidly advancing and generating vast amounts of data. The combination of single cell

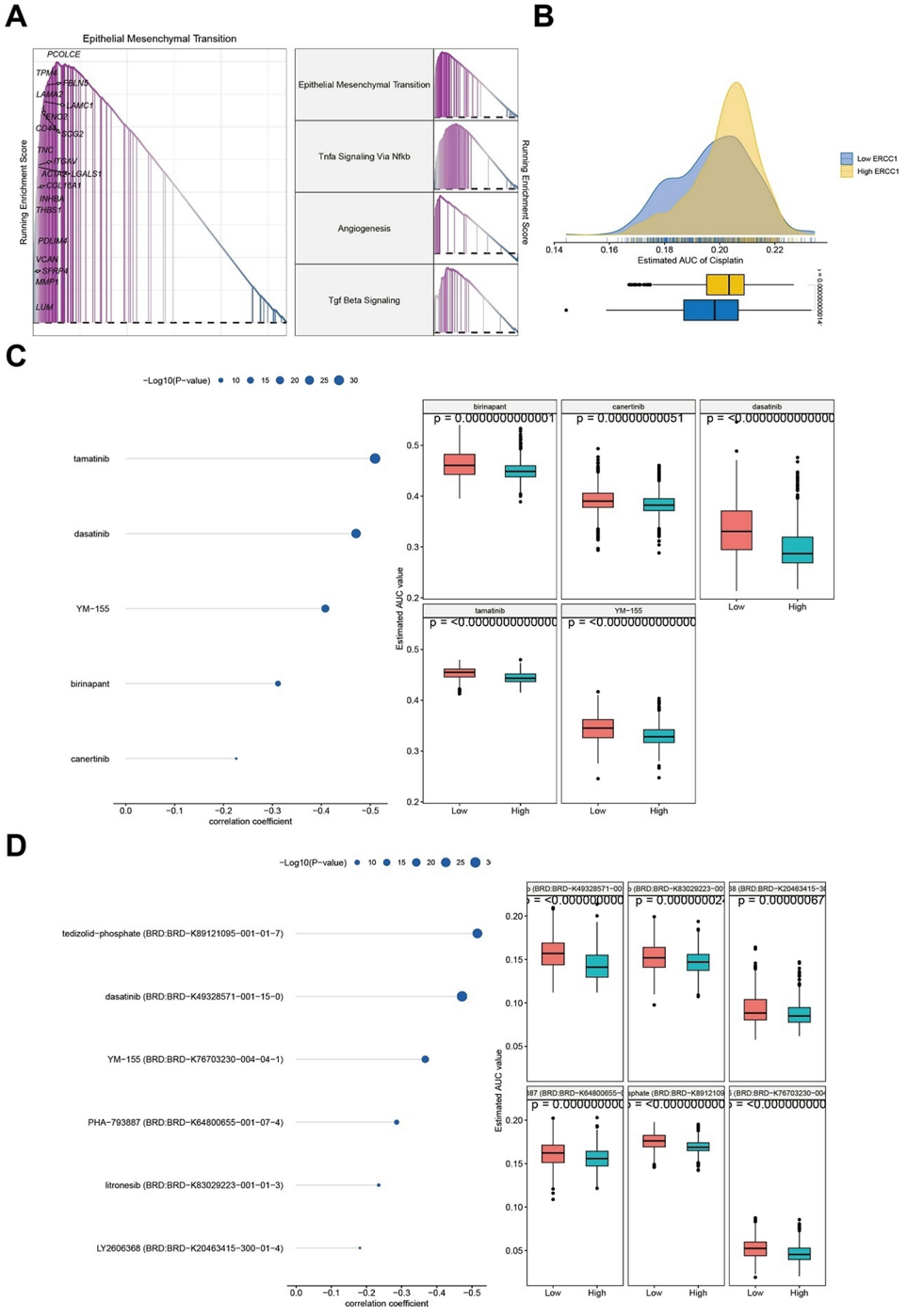

Fig. 7 Identification of potential therapeutic agents for patients in pan-DTC. (**A**). Utilization of GSEA algorithm to identify pathways significantly activated in the high-risk cohort. (**B**). Preliminary assessment of cisplatin sensitivity to validate the computational algorithm's feasibility. (**C**-**D**). Correlation and differential analysis of drug sensitivity for potential therapeutic agents identified from the CTRP and PRISM datasets

data with patient data may facilitate the construction of high-quality clustering models. Thirdly, our finding leads us to propose that neutrophils may contribute to the nuanced heterogeneity observed within gastrointestinal malignancies. However, further comprehensive investigation is needed to delve deeper into the precise role of neutrophils in shaping this heterogeneity.

Considering the potential ways discussed, in the future we aim to integrate multimodal information on cancer and incorporate single cell data as prior knowledge for model training, to continuously enhance the performance of our method. Moreover, deepening our understanding of the intricate interplay between neutrophils and the complex landscape of gastrointestinal tumor heterogeneity holds great potential in refining therapeutic strategies and optimizing patient outcomes. Therefore, future research endeavors aimed at unraveling the intricate mechanistic underpinnings of neutrophil involvement in tumor heterogeneity are imperative for advancing our therapeutic arsenal.

Conclusion

Various computational approaches have been proposed for cancer subtyping through the patients' omics data analysis. However, due to the limited sample size and nonlinearity of high-dimensional data, clustering may result in ambiguous and overlapping cancer subtypes. To address this challenge, we proposed GDEC, an end-to-end generative deep neural network for tumor subtyping. By applying it to six DTC datasets, our results indicate that GDEC outperforms other methods compared and can cluster biologically meaningful tumor subtypes for selecting potential cancer-related genes. Based on the clustered subtypes, we have identified 21 hub genes that are associated with pan-DTC heterogeneity and prognosis, exhibiting a strong correlation with immunotherapy response. Our findings suggest that pan-DTC patients may benefit from targeted therapies such as Dasatinib and YM-155. By integrating advanced computational algorithms and biological analysis, this study lays the foundation for pan-DTC patient precision treatment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00426-z.

Supplementary Material 1

Acknowledgements
We thank Mr. Xuan He in Foshan University for data collection in our research.

Author contributions
CH and LT conceived the study. XL, HY and SX performed the data analysis. LT, WL, CH, and TH interpreted the biological results. CH, and LT wrote the manuscript.

Funding
This work was funded by the Jihua laboratory scienctific project (X210101UZ210),
the Natural Science Foundation of Guangdong Province of China (No. 2022A1515110759),
and the National Natural Science Foundation of China (62201150,62301006).

Data availability
All the data analyzed during the current study are available in the TCGA dataset (https://tcga-data.nci.nih.gov/tcga/). The method codes are available at https://github.com/starlightyouth/GDEC.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All the authors listed have approved the manuscript.

Competing interests
The authors declare no competing interests.

Received: 17 May 2024 / Accepted: 20 January 2025
Published online: 27 January 2025

References

1. Xu J, Wu P, Chen Y, Meng Q, Dawood H, Dawood H. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinformatics. 2019;20(1):1–11.
2. Hemminki K, Liu X, Ji J, Sundquist J, Sundquist K. Autoimmune disease and subsequent digestive tract cancer by histology. Ann Oncol. 2012;23(4):927–33.
3. Chen R, Yang L, Goodison S, Sun Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics. 2020;36(5):1476–83.
4. Chhikara BS, Parang K. Global Cancer statistics 2022: the trends projection analysis. Chem Biology Lett. 2023;10(1):451–451.
5. Malone ER, Oliva M, Sabatini PJ, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med. 2020;12:1–19.
6. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021;19:949–60.
7. Qarmiche N, El Kinany K, Otmani N, El Rhazi K, Chaoui NEH. Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case–control study. BMJ Health Care Inf 2023, 30(1).
8. Lin P, Troup M, Ho JW. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18(1):1–11.
9. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
10. Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, Vermeulen L, Wang X. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis. 2019;8(9):44.
11. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron J, Perou CM, Troester MA, Niethammer M. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast cancer. 2018;4(1):30.
12. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based Multi-omics Integration robustly predicts survival in Liver Cancer. Clin Cancer Res. 2018;24(6):1248–59.
13. Guo L-Y, Wu A-H, Wang Y-x, Zhang L-p, Chai H, Liang X-F. Deep learning-based ovarian cancer subtypes identification using multi-omics data. BioData Min. 2020;13(1):1–12.
14. Yang H, Chen R, Li D, Wang Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics. 2021;37(16):2231–7.
15. Han W, Cheng Y, Chen J, Zhong H, Hu Z, Chen S, Zong L, Hong L, Chan T-F, King I. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. Brief Bioinform. 2022;23(5):bbac377.
16. Cheng Y, Ma X. scGAC: a graph attentional architecture for clustering single-cell RNA-seq data. Bioinformatics. 2022;38(8):2187–93.
17. Guo X, Gao L, Liu X, Yin J. Improved deep embedded clustering with local structure preservation. In: *Ijcai: 2017*; 2017: 1753–1759.
18. Gan Y, Chen Y, Xu G, Guo W, Zou G. Deep enhanced constraint clustering based on contrastive learning for scRNA-seq data. Brief Bioinform 2023:bbad222.
19. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of Cancer. Cell. 2018;173(2):291–304. e296.
20. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic signaling pathways in the Cancer Genome Atlas. Cell. 2018;173(2):321–e337310.
21. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, et al. Perspective on oncogenic processes at the end of the beginning of Cancer Genomics. Cell. 2018;173(2):305–20. e310.
22. Aran D, Sirota M, Butte A. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6(1):8971.
23. Kantarjian H, Jabbour E, Grimley J, Kirkpatrick P. Dasatinib. Nat Rev Drug Discovery. 2006;5(9):717–9.
24. Winter GE, Radic B, Mayor-Ruiz C, Blomen VA, Trefzer C, Kandasamy RK, Huber KV, Gridling M, Chen D, Klampfl T. The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. Nat Chem Biol. 2014;10(9):768–73.
25. Tomczak K, Czerwińska P, Wiznerowicz M. Review the Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncology/Współczesna Onkologia. 2015;2015(1):68–77.
26. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, Hodi FS, Martín-Algarra S, Mandal R, Sharfman WH. Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell. 2017;171(4):934–49. e916.
27. Hoffman-Censits JH, Grivas P, Van Der Heijden MS, Dreicer R, Loriot Y, Retz M, Vogelzang NJ, Perez-Gracia JL, Rezazadeh A, Bracarda S. IMvigor 210, a phase II trial of atezolizumab (MPDL3280A) in platinum-treated locally advanced or metastatic urothelial carcinoma (mUC). In: American Society of Clinical Oncology; 2016.
28. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell. 2016;165(1):35–44.
29. Villa E, Critelli R, Lei B, Marzocchi G, Cammà C, Giannelli G, Pontisso P, Cabibbo G, Enea M, Colopi S. Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut. 2016;65(5):861–9.
30. Chen D-T, Davis-Yadley AH, Huang P-Y, Husain K, Centeno BA, Permuth-Wey J, Pimiento JM, Malafa M. Prognostic fifteen-gene signature for early stage pancreatic ductal adenocarcinoma. PLoS ONE. 2015;10(8):e0133562.
31. Jung H, Kim HS, Kim JY, Sun J-M, Ahn JS, Ahn M-J, Park K, Esteller M, Lee S-H, Choi JK. DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. Nat Commun. 2019;10(1):4278.
32. Balar AV, Galsky MD, Rosenberg JE, Powles T, Petrylak DP, Bellmunt J, Loriot Y, Necchi A, Hoffman-Censits J, Perez-Gracia JL. Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. Lancet. 2017;389(10064):67–76.
33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
34. Yu L, Liu C, Yang JYH, Yang P. Ensemble deep learning of embeddings for clustering multimodal single-cell omics data. Bioinformatics 2023:btad382.
35. Liu Q, Song K. ProgCAE: a deep learning-based method that integrates multi-omics data to predict cancer subtypes. Brief Bioinform 2023:bbad196.

36. Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med. 2021;134:104481.
37. Dhillon A, Singh A, Bhalla VK. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. Arch Comput Methods Eng. 2023;30(2):917–49.
38. Tsai P-C, Lee T-H, Kuo K-C, Su F-Y, Lee T-LM, Marostica E, Ugai T, Zhao M, Lau MC, Väyrynen JP. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nat Commun. 2023;14(1):2102.

Publisher's note