OXFORD

Briefings in Bioinformatics, 2025, **26(6)**, bbaf600
https://doi.org/10.1093/bib/bbaf600
Problem Solving Protocol

DSCC: disease subtyping using spectral clustering and community detection from consensus networks

Dao Tran[1], Van-Dung Pham[1], Ha Nguyen[1], Phi Bya[1], Aiham Qdaisat[2], Liem Minh Phan[3], Sai-Ching Jim Yeung[2], Tin Nguyen[1,*]

[1]Department of Computer Science and Software Engineering, Auburn University, Auburn, 36849 Alabama, United States
[2]Department of Emergency Medicine, The University of Texas MD Anderson Cancer Center, Houston, 77030 Texas, United States
[3]David Grant USAF Medical Center – Clinical Investigation Facility, 60th Medical Group, Defense Health Agency, Travis Air Force Base, 94535 California, United States
*Corresponding author. Department of Computer Science and Software Engineering, Auburn University, Auburn, 36849 Alabama, United States. E-mail: tinn@auburn.edu

Abstract

Molecular subtyping is fundamental in cancer research and clinical management of cancer, guiding treatment planning, monitoring therapeutic response, and informing prognosis. Early methods were designed specifically for gene expression data due to the lack of other molecular data types. Thanks to breakthroughs in high-throughput technologies, recent subtyping tools have shifted their focus to integrating multi-omics profiles to uncover novel subtypes that better reflect genetic variation, molecular pathogenesis, tumor heterogeneity, and host response biological mechanisms. However, these integrative approaches have not been able to fully exploit the complementary potentials of diverse molecular data types. They often rely on specific omics types with large common sample size and fail to incorporate important biological knowledge in their models. Here, we introduce Disease subtyping using Spectral clustering and Community detection from Consensus networks (DSCC), a method designed to identify meaningful disease subtypes from a wide range of molecular data, including gene expression, miRNA expression, DNA methylation, copy number variation, somatic mutations, protein abundance, and metabolite levels. We demonstrate the superiority of DSCC over state-of-the-art cancer subtyping methods using 43 cancer datasets with more than 11,000 patients. Furthermore, the incorporation of DSCC-derived subtype information as a covariate in prognostic models improves survival prediction accuracy and robustness. The DSCC source code, data, and scripts for reproducing all results in this study are available at https://github.com/tinnlab/DSCC.

Keywords: Cancer subtyping; multi-omics; data integration; consensus network

Introduction

Cancer is an umbrella term that includes a spectrum of disease severity, from those that are malignant, metastatic, and aggressive to benign lesions with low potential for progression or death. Various genome-wide profiling techniques have been developed to capture the dynamics of cancer development at multiple levels, including genomics, transcriptomics, epigenomics, proteomics and metabolomics. By analyzing multi-omics data, researchers can obtain a comprehensive view of cancer evolution, molecular subtypes, and potential risks. In turn, these novel insights can lead to effective personalized treatment and prognosis [1–5]. As such, Granja et al. [6] combined transcriptomic, epigenomic, and proteomic data of leukemic blood cells to identify cancer-specific processes involved in blood differentiation and critical markers of leukemia subtypes. Other multi-omics studies resulted in the discovery of novel subgroups and new therapeutic targets of breast cancer [7, 8], liver cancer [9, 10], lung cancer [11, 12], brain cancer [13, 14] and other cancer types [15–18].

Many integrative methods have been developed for disease subtyping. They can be grouped into three main categories: consensus-based models, shared representation methods, and similarity-based approaches. Early methods in the first category, such as BCC [19] and MDI [20], identify clusters for each data type and combine them into an optimal grouping while recent approaches, including MOVICS [21], ClustOmics [22] and Subtype-WESLR [23], integrate cluster assignments from multiple clustering algorithms instead. Methods in the second category, including intNMF [24], LRACluster [25], iClusterBayes [26], iClusterPlus [27], iCluster [28, 29], MRGCN [30], DLSF [31], and DSIR [32], generate a shared representation across data types and apply clustering algorithms for this representation to discover subtypes. Methods in the third category construct similarity matrices for each data type, and combine them into an overall similarity matrix for clustering. Methods in this category include SNF [33], NEMO [34], PINS [35–37], CIMLR [38], ANF [39], hMKL [40], and MDICC [41].

Current disease subtyping approaches exhibit important challenges. First, most methods rely on magnitude-based distance metrics (e.g. Euclidean distance) while neglecting directional differences between feature vectors. Recent deep learning studies have concluded that using direction-based metrics, such as angular distance, for constructing loss functions can improve the learning of more discriminative features, thereby enhancing classification performance [42–44]. Therefore, the addition of direction-awareness to subtyping tools may enhance the ability to differentiate molecular patterns between subtypes. Second,

Received: July 19, 2025. **Revised:** October 12, 2025. **Accepted:** October 19, 2025

available subtyping methods usually do not investigate the complementary strengths of different clustering algorithms. Since each of these algorithms has its own strength and weakness, combining them could yield more robust and biologically meaningful subtypes. For example, spectral clustering [45] is effective in identifying global structures while community detection methods such as Louvain [46] are better at capturing local patterns. Third, existing approaches often overlook the role of pathway information, which is crucial for understanding how apparently distinct carcinogenetic changes might be functionally similar in terms of biological processes. This limitation stems partly from the fact that current pathway databases such as KEGG [47] or Reactome [48] focus on genes/proteins interactions, lacking knowledge for other types of molecular data such as miRNAs and methylation. Finally, current methods often require matched samples across all omics types, which leads to substantial data loss when some modalities are missing. Only a few methods, including NEMO [34] and MRGCN [30], are capable of handling missing data across data types. As a result, available approaches usually prioritize data types with large common sample sizes (gene expression, miRNA expression, and DNA methylation) over less commonly obtained data modalities (e.g. protein quantification and metabolite levels).

In this article, we introduce a new subtyping approach, named Disease subtyping using Spectral clustering and Community detection from Consensus networks (DSCC), that addresses the above-mentioned challenges. DSCC leverages different distance metrics in constructing multi-omics consensus networks, and combines multiple clustering approaches to partition consensus networks into meaningful cancer subtypes. In addition, DSCC aggregates multi-omics features into KEGG-compatible representations to facilitate the incorporation of pathway information. The method also handles the problem of missing data across all omics types, allowing inclusion of all the samples and flexibility to add new molecular data to the analysis. To demonstrate the advantage of the proposed method, we compare DSCC against 13 current state-of-the-art approaches: CC [49], CIMLR [38], SNF [33], LRACluster [25], intNMF [24], ANF [39], NEMO [34], MOVICS [21], MRGCN [30], hMKL [40], MDICC [41], DLSF [31], and DSIR [32]. Our benchmarking involves the extensive analysis of 43 cancer datasets with over 11,000 patients obtained from Genomic Data Common (GDC/TCGA) [50] and other public databases.

Materials and methods

Figure 1 shows the high-level workflow of DSCC. The method first aggregates multi-omics data into gene-level features and then constructs patient similarity networks using multiple distance metrics, followed by an ensemble clustering that integrates analysis results from spectral and community detection methods.

Data processing

The datasets analyzed in this study contain up to seven omics types: mRNA, miRNA expression, DNA methylation, copy number variations (CNVs), somatic mutations, protein, and metabolite levels. Among the seven omics types, only CNVs contains gene-level features, while other data types consist of features at different levels. Each data type may include multiple data formats. For example, mRNA encompasses TPM, FPKM, FPKMuq, and raw counts, which are treated as separate data matrices in our analysis because they capture complementary characteristics of the expression data. Raw counts preserve the original statistical properties of the data but suffer from gene length and sequencing depth biases [51, 52]. FPKM corrects for these biases but is suboptimal for cross-sample comparison [53, 54]. FPKMuq (upper quartile normalized FPKM) offers a more robust, though imperfect, basis for inter-sample comparison at the cost of potential information loss [55, 56]. Finally, TPM provides a distinct advantage for cross-sample comparison but remains susceptible to highly expressed genes biases [57] and inaccuracies for very short transcripts [58]. Together, these four quantification units present a more comprehensive interpretation of gene expression, which provides more discriminatory power than each unit alone. (see Supplementary Section 1 and Figure S1 for more details).

The data pre-processing procedure performs gene-level aggregation for mRNA, miRNA, DNA methylation, and protein quantification. For mRNA, we average expression values of all transcripts for each gene. For miRNA, we map miRNA IDs to target genes using miRTarBase [59] and then calculate the average expression for individual target genes. For protein data, we average the measurements of all proteins encoded by the same gene. For DNA methylation, we map CpG sites to genes using the manufacturer-provided annotation and then calculate the median methylation level for each gene. Next, we remove genes that are not associated with KEGG pathways. For metabolomics data, we apply log2 transformation and replace missing values with zero. For somatic mutations, we quantify the six single-base substitution types (among C>A, C>G, C>T, T>A, T>C, T>G) for each sample using allelic information.

Our choice to use gene-level aggregation is primarily a measure to create a consistent gene-level framework across diverse data types and platforms, thereby enhancing stability and avoiding overfitting (see Supplementary Section 2 and Figure S2). However, we acknowledge users' need to customize their own data processing. Therefore, DSCC allows users to input their pre-processed data directly instead of mandating our gene-level aggregation step.

Patient network construction

After data processing, we obtain a set of M molecular data matrices, where each matrix X_k ($k \in [1, M]$) consists of N_k patients (rows) and P_k features (columns). For each data matrix X_k, we compute two types of affinity matrices using the following formulation:

$$A_k^{(t)}(i,j) = \exp\left(-\frac{d^{(t)}(X_{ki}, X_{kj})^2}{2\sigma_t^2}\right), t \in \{E, A\} \quad (1)$$

Here, $A_k^{(E)}$ denotes the Euclidean affinity matrix and $A_k^{(A)}$ denotes the Angular affinity matrix. X_{ki} and X_{kj} are the feature vectors of the patients i-th and j-th in the k-th omic, $d^{(t)}(X_{ki}, X_{kj})$ represents the distance between two feature vectors X_{ki} and X_{kj}, and σ is a scaling parameter that determines how quickly the similarity decays with distance. The two corresponding distances are defined as:

$$d^{(E)}(X_{ki}, X_{kj}) = \|X_{ki} - X_{kj}\|_2 \quad (2)$$

$$d^{(A)}(X_{ki}, X_{kj}) = \frac{\theta_{kij}}{\pi} = \frac{1}{\pi}\arccos\left(\frac{X_{ki} \cdot X_{kj}}{\|X_{ki}\| \cdot \|X_{kj}\|}\right) \quad (3)$$

where θ_{kij} is the angle between X_{ki} and X_{kj}, $\|\cdot\|$ indicates the L_2-norm of a vector and $X_{ki} \cdot X_{kj}$ represents the dot product of the two vectors. We choose the scaling parameter σ_E as the median of all pairwise Euclidean distances and set $\sigma_A = 0.5$.

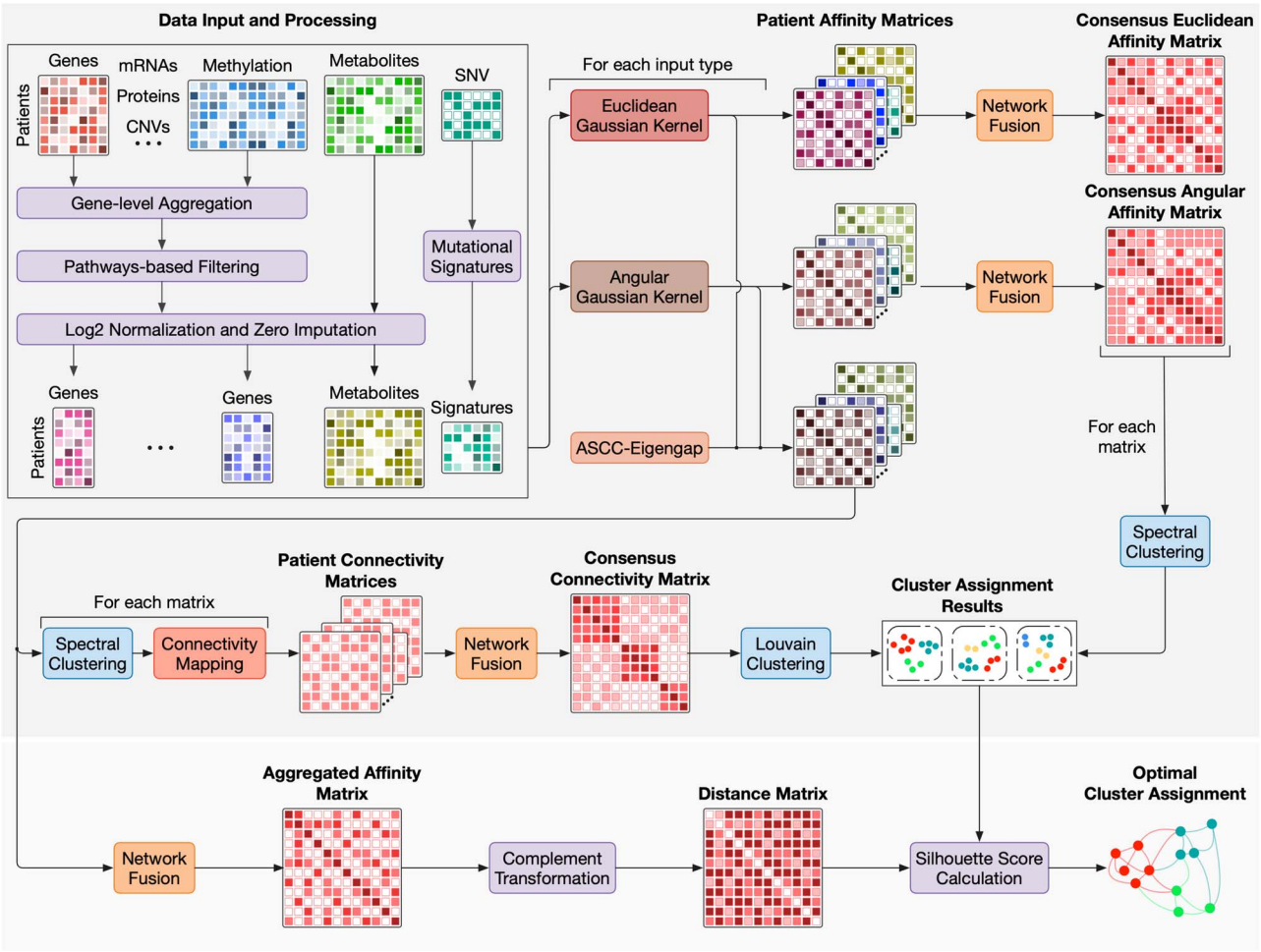

Figure 1. The overall analysis pipeline of DSCC. The method takes as input the multi-omics data matrices of multiple data types such as gene expression, microRNA, protein level, copy number variations (CNVs), methylation, metabolites, and single nucleotide variant (SNV). In each matrix, rows represent patients and columns represent molecular features. Each data type undergoes specific pre-processing steps: 1) gene-level aggregation, pathway-based filtering, log2 normalization, and zero imputation for mRNA, miRNA, methylation, and protein levels; 2) pathway-based filtering, log2 normalization, and zero imputation for CNVs; 3) log2 normalization and zero imputation for metabolomics data; and 4) mutational signature computation for SNV data. We use three methods to capture patient similarity. The first method generates patient affinity matrices – one per data type – using a Euclidean Gaussian Kernel, which are then combined to create the Consensus Euclidean Affinity Matrix. The second method employs an Angular Gaussian Kernel to create the affinity matrices before fusing them into a Consensus Angular Affinity Matrix. The third method generates the affinity matrices using the ASCC-Eigengap algorithm and applies Spectral Clustering with Connectivity Mapping to produce the Consensus Connectivity Matrix. The three consensus matrices are subsequently partitioned using spectral and Louvain clustering. To choose the optimal partitioning, we calculate the Silhouette score for each of the three clustering assignments and output the partitioning with the highest Silhouette score.

Using both Angular and Euclidean affinity matrices helps DSCC to capture complementary views of each omics type, which in turn improves the method's performance (see Supplementary Section 3 and Figure S3). After computing the Euclidean affinity $A_k^{(E)}$ and the Angular affinity matrices $A_k^{(A)}$ for all data types, DSCC combines these matrices into three different consensus matrices: 1) consensus Euclidean affinity, 2) consensus Angular affinity, and 3) consensus connectivity.

Constructing consensus affinity matrices

It is known that affinity-based clustering is often sensitive to the size of the data. This is because local structures can be strong in large datasets while they might be weak and unstable in small datasets [60, 61]. To overcome this, we adopt two different approaches for constructing consensus affinity matrices.

For small datasets (contain less samples than the sample size threshold of 200), we utilize the full affinity matrices and average them to compute the Full Consensus Affinity (FCA) matrices as follows:

$$FCA^{(t)}(i,j) = \frac{1}{|OT(i,j)|} \sum_{k \in OT(i,j)} A_k^{(t)}(i,j), t \in \{E, A\} \quad (4)$$

where $OT(i,j)$ is the set of omics types available for both i-th and j-th samples, and $|OT(i,j)|$ represents the cardinality.

For large datasets (at least 200 samples), we apply a K-nearest neighbor (KNN) approach [34] to emphasize local structure and mitigate noisy global connections. For each patient i, we retain the scores of its top 10 nearest neighbors in the affinity matrix $A_k^{(t)}$, while setting all other scores to zero. We then normalize the scores so that the scores are summed up to one for each patient. To ensure symmetry and incorporate mutual neighborhood information, we add the matrix to its transpose. The resulting local

affinity matrices, $LA_k^{(E)}$ and $LA_k^{(A)}$, are defined as follows:

$$LA_k^{(t)}(i,j) = \frac{A_k^{(t)}(i,j)\cdot I(j\in\eta_{ki})}{\sum_{r\in\eta_{ki}} A_k^{(t)}(i,r)} + \frac{A_k^{(t)}(i,j)\cdot I(i\in\eta_{kj})}{\sum_{r\in\eta_{kj}} A_k^{(t)}(r,j)},\quad t\in\{E,A\} \quad (5)$$

where η_{ki} denotes the set of 10 nearest neighbors of the i-th sample, and $I(j \in \eta_{ki})$ is an indicator function equal to 1 if $j \in \eta_{ki}$, and 0 otherwise. We then use Equation (4) and replace full affinity matrices with local affinity matrices to derive the Local Consensus Affinity (LCA) matrices.

Analytically, the LCA matrices alone prove highly effective for downstream tasks. However, the sparsity introduced by the LCA's KNN step artificially inflates the resulting number of clusters in small datasets. Therefore, we use the full, dense FCA matrices for small datasets and reserve the LCA matrices for large datasets where the cluster inflation issue is mitigated (see Supplementary Section 4 and Figure S4.

Constructing consensus connectivity matrix

Given a Euclidean affinity matrix $A_k^{(E)}$ and an Angular affinity matrix $A_k^{(A)}$ for each data type k, we aggregate them using the AASC-Eigengap algorithm, a modified version of the AASC algorithm [62]. The objective of AASC-Eigengap is to find the best weights $\upsilon_k^{(t)}$ with $t \in \{A, E\}$ for the omics type k so that its aggregated affinity matrix $\bar{A}_k = \sum_{t\in\{A,E\}} \upsilon_k^{(t)} \times A_k^{(t)}$ gives the largest eigengap. The pseudocode for the algorithm is illustrated in Algorithm 1. For each data type, we perform spectral clustering on the corresponding aggregated affinity matrix to obtain a temporary partitioning of patients. This partitioning is then used to construct a binary connectivity matrix, where 1 indicates that the two samples belong to the same cluster, and 0 indicates that they belong to different clusters. We subsequently leverage Equation (4) and replace the affinity matrices with the connectivity matrices of all omics types to combine them into a consensus connectivity matrix.

Algorithm 1 AASC-Eigengap (for each omics type k)

Input: Affinity matrices $A^{(t)} \in \mathbb{R}^{n\times n}$ for $t = 1, \ldots, \tau$
▷ All variables are defined for a fixed omics type k
Output: *best_weights* for combining affinity matrices

Initialize $\upsilon^{(t)} \leftarrow 1/\tau$, $best\_eigengap \leftarrow -\infty$
$best\_weights \leftarrow \upsilon^{(t)}$
for *iteration* = 1 to 20 **do**
  for $t = 1$ to τ **do**
    $D^{(t)}(i,i) \leftarrow \sum_j A^{(t)}(i,j)$
    $L^{(t)} \leftarrow D^{(t)^{-1/2}}(D^{(t)} - A^{(t)})D^{(t)^{-1/2}}$
  end for
  $\bar{A} \leftarrow \sum_{t=1}^{\tau} \upsilon^{(t)} A^{(t)}$
  $D(i,i) \leftarrow \sum_j \bar{A}(i,j)$, $L \leftarrow D^{-1/2}(D - \bar{A})D^{-1/2}$
  Compute eigenvalues E_i and eigenvectors G_i of L
  Compute $eigengap_i = i \cdot |E_{i+1} - E_i|$
  Find $max\_eigengap = \max(eigengap_i)$
  if $max\_eigengap < best\_eigengap$ **then**
    $best\_eigengap \leftarrow max\_eigengap$
    $best\_weights \leftarrow \upsilon^{(t)}$
  end if
  for $t = 1$ to τ **do**
    $\alpha^{(t)} \leftarrow \text{mean}\left(\sum_{i=2}^{10} G_i^T D^{(t)} G_i\right)$
    $\beta^{(t)} \leftarrow \text{mean}\left(\sum_{i=2}^{10} G_i^T L^{(t)} G_i\right)$
    $\gamma^{(t)} \leftarrow \beta^{(t)}/\alpha^{(t)}$
  end for
  Solve for λ_1 such that:

$$\left(\sum_t \frac{1}{(\gamma^{(t)} - \lambda_1)\alpha^{(t)}}\right)^2 = \sum_t \frac{1}{(\gamma^{(t)} - \lambda_1)^2\alpha^{(t)}}$$

  Then compute:

$$\lambda_2 \leftarrow \left(\sum_t \frac{1}{(\gamma^{(t)} - \lambda_1)\alpha^{(t)}}\right)^{-1}$$

  for $t = 1$ to τ **do**
    $\upsilon^{(t)} \leftarrow \frac{\lambda_2}{(\gamma^{(t)} - \lambda_1)\alpha^{(t)}}$
  end for
  Normalize: $\upsilon^{(t)} \leftarrow \upsilon^{(t)} / \sum_t \upsilon^{(t)}$
  Recompute $\bar{A}$, D, L and eigengaps with updated $\upsilon^{(t)}$
  Select $\upsilon^{(t)}$ with largest eigengap
end for
return *best_weights*

Ensemble clustering

After constructing consensus matrices, we obtain three matrices: 1) consensus Euclidean affinity, 2) consensus Angular affinity, and 3) consensus connectivity. We subsequently apply spectral clustering to the first two matrices (Euclidean and Angular affinity matrices). The optimal number of clusters is determined using the eigengap method. We also use the Louvain community detection algorithm to partition the consensus Connectivity matrix. Since Louvain is stochastic, we perform multiple runs and choose the result with the highest modularity score [63].

After clustering, we obtain three different partitionings for the same set of patients. We evaluate each partitioning using the Silhouette score [64] based on a distance matrix derived by fusing the aggregated affinity matrices $\bar{A}_k$ of all omics types using Equation (4), and subtracting the resulting matrix from 1. At the end, we output the partitioning with the highest Silhouette score.

Results

We perform an extensive analysis of 43 cancer datasets (33 TCGA datasets and 10 metabolomics datasets from published articles) covering 28 tissues and over 11,000 samples (see Supplementary Table S1 for more details). We first compare the performance of DSCC against Consensus Clustering (CC) [49], CIMLR [38], SNF [33], LRACluster [25], intNMF [24], ANF [39], NEMO [34], MOVICS [21], MRGCN [30], hMKL [40], MDICC [41], DLSF [31] and DSIR [32] using Cox P-values. We then analyze the usefulness of using DSCC-derived subtypes as an additional covariate for risk prediction. Finally, we provide an in-depth pathway analysis of the discovered subtypes for the adrenocortical carcinoma (TCGA-ACC) dataset.

For TCGA data, we download all 12 matrices measuring mRNA expression, miRNA expression, DNA methylation, CNV, and somatic mutations from Genomic Data Commons (https://gdc.cancer.gov/), together with clinical variables and survival information. We also download their protein data from Proteomic Data Commons (https://pdc.cancer.gov/pdc/) and metabolite data from Benedetti et al. [65]. We obtain seven of the additional datasets from published articles: P23918603 [66], P24316975 [67], P30244973 [68], P33577785 [69], P38007532 [70], P38395893 [71], and P40016594 [72]. We download the remaining three datasets (ST001235, ST001236, and ST001237) from the Metabolomics

Workbench repository (https://www.metabolomicsworkbench.org/). For P23918603, P24316975, P30244973, and P33577785, we use pre-processed metabolite levels and gene expression data from Benedetti et al. [65]. As P38007532 lacks overall survival information, we focus on the subset of samples in this dataset that developed cancer during the follow-up period. We define survival time for such samples as the duration from the start of follow-up until cancer diagnosis.

Subtyping on 43 cancer datasets

We compare DSCC against 13 approaches, including Consensus Clustering (CC) [49], CIMLR [38], SNF [33], LRACluster [25], intNMF [24], ANF [39], NEMO [34], MOVICS [21], MRGCN [30], hMKL [40], MDICC [41], DLSF [31] and DSIR [32]. For these methods, we select the top 8,000 highly variable features from each data matrix and apply the processing procedure suggested by each method. The feature size of 8,000 is chosen to align with DSCC's post-processing data dimensionality, ensuring a fair assessment across all evaluated methods. These methods can automatically determine the number of subtypes (i.e. the number of clusters). We execute each method using their default parameters.

We use three metrics to measure the performance of each method for each dataset. First, we use the Cox proportional hazards model [73] to assess the statistical significance in survival differences of identified subtypes (the smaller the *P*-values, the better). Second, we assess the statistical significance in survival differences of identified subtypes using an empirical *P*-value that was introduced by Rappoport and Shamir [74]. Third, we use the Concordance Index (C-Index) [75] to evaluate whether incorporating clustering information improves discrimination power of survival prediction (the higher C-Indices, the better).

Table 1 shows the Cox *P*-values computed using the *survival* package [76]. NA entries indicate analysis failures. Only DSCC, NEMO, and MRGCN are able to perform subtyping for all datasets. All other methods crash when analyzing TCGA-BRCA, TCGA-COAD, and TCGA-OV datasets. A potential cause might be insufficient numbers of matched samples across all data matrices. DSCC, NEMO, and MRGCN are capable of handling missing data across data types, which allows us to use all samples for each dataset when running the three methods. In contrast, the remaining methods require completely matched samples among data types, restricting them to sample intersection. DSCC outperforms all competitors in identifying subtypes with significantly different survival profiles in most datasets (27 out of 43). DSCC also maintains this robust performance across different settings of established hyperparameters including σ_E, σ_A, and the sample size threshold. (see Supplementary Section 5 and Figure S5). The next best methods are NEMO, hMKL, intNMF and ANF, with significant results in 19, 17, 15, and 15 datasets, respectively. All remaining methods achieve significant results in less than 15 datasets.

Supplementary Table S2 shows the empirical *P*-values obtained for each method. Instead of calculating the *P*-values directly from a log-rank test [73], we calculate the χ^2 statistic using the *survdiff* function from the *survival* package [73, 76]. To construct the null distribution for each dataset, we then permute the cluster labels across all samples 10,000 times and compute the χ^2 statistics under the null. The empirical *P*-value is derived by calculating the proportion of the null distribution that was more extreme than the observed χ^2 statistic. The performance assessment using the empirical *P*-values is consistent with that from the Cox *P*-values. DSCC outperforms all competitors in identifying subtypes with significantly different survival profiles in most datasets (26 out of 43). The next best methods are NEMO, intNMF and ANF, with significant results in 16, 15, and 14 datasets, respectively. All remaining methods achieve significant results in less than 14 datasets.

Figure 2 shows the minus log10 *P*-values of both Cox and empirical *P*-values of all methods across 43 cancer datasets. In this figure, the distribution of the minus log10 Cox *P*-values of DSCC has a median value of 1.5, which is substantially higher than those of all other methods. This median, obtained using all available omics types, is also the highest for DSCC across varying number of omics types tested (see Supplementary Section 6 and Figure S6). The second best method, NEMO, has a median of 0.94, which falls well below the commonly used 5% significance threshold (minus log10 of 0.05 is approximately 1.3). Similarly, the minus log10 empirical *P*-value of DSCC has a median of 1.47, considerably exceeding the second best method, intNMF, with a median of 0.88. DSCC also has the highest number of significant datasets: 27 by Cox *P*-value and 26 by empirical *P*-value.

Next, we evaluate the usefulness of DSCC-derived subtypes for risk prediction. Specifically, we examine the performance of a risk prediction method under two different scenarios: (1) using only clinical variables (age, height, gender, etc.) as predictors, and (2) incorporating the subtype assignment of each subtyping method as an additional covariate. We adopt blockForest [77] for this validation due to its simplicity and proven effectiveness [78, 79]. For each dataset, we perform 5-fold cross-validation and measure the C-Index on the validation sets. As shown in Fig. 3, DSCC exhibits the highest mean C-Index value of 0.71 (more details in Supplementary Table S3). Notably, DSCC improves survival prediction performance when their subtype assignments are added as a covariate in survival prediction. Specifically, incorporating DSCC-identified groupings increases the average C-Index from 0.69 (clinical variables only) to 0.71.

Pathway analysis of adrenocortical carcinoma

We investigate the mechanisms underlying the different survival outcomes of the discovered subtypes for TCGA-ACC. Figure 4 shows the Kaplan-Meier survival analysis of the three ACC subtypes, in which subtype 1 (red) exhibits a significantly lower survival probability than the other two. To characterize the molecular differences between subtypes, we conduct a consensus pathway analysis using the Intelligent Platform for Systems-level Analysis (IPSA), an upgraded version of our previously published Consensus Pathway Analysis (CPA) platforms [80–82]. Due to the small sample size of subtype 2 (blue) in the TCGA-ACC dataset (only two patients), we merge subtypes 2 and 3 to form a single high-survival group for the differential analysis. Briefly, IPSA performs differential analysis to identify the differentially expressed (DE) genes and then performs consensus analysis of four pathway analysis methods (Over-Representation Analysis (ORA) [83, 84], Gene Set Analysis (GSA) [85], Fast Gene Set Enrichment Analysis (FGSEA) [86], and Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) [87], using KEGG pathways. IPSA outputs the list of DE genes, and the list of pathways that are identified as significantly impacted across all four pathway methods.

IPSA identifies 925 DE genes, with 896 genes (97%) upregulated in the poor-survival group (Supplementary Table S4). We also plot the significant pathways and their statistics (minus log10 adjusted *P*-values vs. normalized enrichment scores) in Fig. 5, which highlights the top twenty significantly enriched pathways, including *Cell Cycle*, *DNA Replication*, and *Base Excision Repair*, with adjusted *P*-values of 4.3×10^{-4}, 0.020, and 0.024, respectively (Supplementary Table S5). This suggests elevated mitotic activity and increased DNA maintenance processes in the poor-survival

Table 1. Cox *P*-values and number of subtypes (formatted as **P-value (#Subtypes)**) identified by DSCC, CC, CIMLR, SNF, LRACluster, intNMF, ANF, NEMO, MOVICS, MRGCN, hMKL, MDICC, DLSF and DSIR for 43 cancer datasets. Cells highlighted in grey have significant *P*-values (less than 0.05). Cells with bold text have the most significant *P*-value in each row. The last two rows report the number of datasets with significant *P*-values (less than 0.05) and the average number of subtypes identified by each method. DSCC outperforms other methods by having the highest number of significant *P*-values (27 out of 43 datasets)

Dataset	DSCC	CC	CIMLR	SNF	LRA- Cluster	intNMF	ANF	NEMO	MOVICS	MRGCN	hMKL	MDICC	DLSF	DSIR
TCGA-ACC	5E-7(3)	2E-2(3)	6E-2(6)	2E-4(3)	7E-1(2)	8E-4(4)	1E-3(3)	**3E-8(5)**	3E-3(4)	6E-1(2)	5E-3(8)	1E-1(3)	7E-3(4)	5E-4(5)
TCGA-BLCA	**1E-4(10)**	1E-1(3)	3E-1(8)	1E-1(2)	3E-2(2)	5E-2(2)	8E-1(2)	7E-4(7)	3E-1(3)	6E-1(2)	3E-1(9)	2E-1(2)	5E-1(7)	2E-1(10)
TCGA-BRCA	6E-1(7)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**5E-1(2)**	NA(NA)	5E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-CESC	2E-2(10)	6E-1(3)	4E-2(7)	1E-1(2)	3E-1(2)	5E-1(4)	9E-2(2)	1E-1(10)	2E-1(3)	9E-1(2)	5E-2(10)	6E-1(2)	**1E-2(5)**	7E-2(8)
TCGA-CHOL	6E-1(3)	7E-1(3)	1E+0(2)	3E-1(2)	3E-1(2)	4E-1(3)	**2E-1(2)**	5E-1(7)	3E-1(2)	2E-1(2)	NA(NA)	1E+0(2)	NA(NA)	2E-1(2)
TCGA-COAD	**2E-2(10)**	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	8E-1(3)	NA(NA)	1E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-DLBC	3E-1(4)	6E-1(3)	4E-1(2)	9E-1(2)	6E-1(4)	**2E-1(2)**	4E-1(2)	6E-1(3)	1E+0(2)	3E-1(2)	NA(NA)	8E-1(2)	NA(NA)	NA(NA)
TCGA-ESCA	2E-1(2)	2E-3(5)	3E-1(2)	4E-3(2)	**4E-4(3)**	2E-2(2)	4E-3(2)	4E-1(2)	4E-2(2)	1E-1(2)	1E-2(3)	5E-3(2)	3E-1(3)	9E-1(9)
TCGA-GBM	**7E-3(9)**	2E-2(3)	6E-1(2)	6E-1(2)	3E-1(2)	1E-1(2)	4E-2(3)	2E-1(8)	3E-1(3)	6E-2(2)	9E-1(3)	2E-2(4)	NA(NA)	3E-2(3)
TCGA-HNSC	6E-3(6)	5E-1(3)	1E-1(8)	3E-1(2)	1E+0(2)	1E-1(3)	2E-1(2)	**2E-3(6)**	7E-2(2)	1E-1(2)	4E-1(4)	4E-1(2)	2E-1(4)	7E-1(9)
TCGA-KICH	8E-5(2)	**4E-6(3)**	6E-1(5)	1E-1(2)	3E-1(4)	2E-5(4)	4E-3(2)	2E-5(8)	6E-1(2)	2E-1(2)	4E-2(5)	2E-2(2)	9E-3(6)	4E-1(3)
TCGA-KIRC	**7E-13(10)**	3E-1(3)	2E-1(6)	6E-1(2)	5E-1(2)	5E-1(3)	7E-1(2)	2E-10(5)	3E-1(2)	8E-10(2)	1E-1(7)	4E-1(3)	2E-1(3)	8E-2(10)
TCGA-KIRP	**2E-16(6)**	4E-1(3)	8E-3(10)	2E-2(2)	2E-2(2)	4E-2(4)	3E-2(2)	1E-4(2)	1E-1(2)	3E-1(2)	1E-4(9)	6E-1(5)	8E-2(6)	2E-2(9)
TCGA-LAML	6E-3(3)	1E-1(3)	4E-1(8)	3E-2(3)	2E-2(2)	2E-1(5)	3E-2(3)	**2E-4(6)**	1E-1(4)	3E-1(2)	5E-1(10)	8E-1(2)	1E-1(6)	3E-2(4)
TCGA-LGG	**2E-30(10)**	6E-8(3)	4E-7(4)	1E-15(2)	4E-3(2)	6E-3(2)	1E-10(2)	7E-30(4)	9E-3(2)	7E-8(2)	3E-2(2)	7E-1(3)	2E-7(7)	9E-1(8)
TCGA-LIHC	2E-3(9)	4E-1(3)	1E-1(2)	8E-1(2)	4E-2(2)	1E-1(2)	7E-1(2)	5E-1(2)	4E-1(2)	**3E-4(2)**	8E-1(2)	6E-1(3)	1E-3(5)	3E-2(8)
TCGA-LUAD	**2E-2(9)**	5E-1(3)	4E-1(3)	4E-1(2)	2E-1(2)	7E-2(5)	9E-2(2)	2E-1(2)	5E-2(6)	5E-2(2)	5E-1(3)	3E-1(2)	9E-1(4)	8E-1(6)
TCGA-LUSC	4E-2(2)	5E-2(3)	7E-1(3)	9E-1(2)	2E-1(3)	**3E-2(2)**	2E-1(2)	3E-2(2)	8E-1(2)	1E-1(2)	3E-1(2)	3E-2(2)	2E-1(5)	1E-1(8)
TCGA-MESO	4E-3(3)	2E-1(3)	3E-1(6)	3E-2(2)	5E-1(2)	1E-2(3)	3E-2(3)	1E-1(3)	2E-1(4)	**2E-3(2)**	2E-3(8)	2E-1(2)	9E-3(3)	3E-1(2)
TCGA-OV	4E-2(8)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	**1E-3(4)**	NA(NA)	4E-1(2)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
TCGA-PAAD	1E-1(2)	1E-2(3)	2E-1(3)	1E-2(3)	2E-2(3)	4E-1(3)	1E-2(3)	5E-3(2)	2E-2(3)	3E-1(2)	9E-2(6)	5E-1(2)	2E-1(5)	**8E-5(7)**
TCGA-PCPG	8E-1(2)	6E-1(3)	2E-1(2)	1E-1(3)	4E-1(2)	6E-1(2)	**1E-1(2)**	7E-1(2)	4E-1(3)	2E-1(2)	2E-1(9)	NA(NA)	2E-1(4)	3E-1(5)
TCGA-PRAD	5E-1(10)	3E-1(3)	9E-1(6)	7E-1(2)	9E-1(2)	7E-1(2)	9E-1(2)	4E-1(2)	3E-1(2)	2E-1(2)	3E-1(8)	8E-1(2)	7E-1(5)	**6E-2(9)**
TCGA-READ	3E-1(2)	5E-1(3)	6E-1(7)	3E-1(2)	3E-1(2)	7E-1(3)	7E-1(2)	7E-1(3)	1E-1(2)	7E-1(2)	**3E-6(9)**	5E-1(3)	2E-1(6)	4E-1(3)
TCGA-SARC	3E-4(9)	1E-3(4)	4E-1(6)	8E-1(2)	8E-1(2)	2E-1(2)	2E-2(3)	3E-1(3)	1E-4(5)	6E-1(2)	**4E-9(8)**	3E-1(2)	6E-2(6)	4E-2(7)
TCGA-SKCM	9E-2(2)	1E+0(3)	**8E-2(8)**	5E-1(5)	5E-1(3)	3E-1(2)	4E-1(5)	2E-1(10)	3E-1(2)	3E-1(2)	5E-1(10)	5E-1(3)	7E-1(6)	5E-1(3)
TCGA-STAD	2E-1(10)	2E-1(3)	**3E-2(3)**	3E-1(3)	4E-1(3)	1E-1(5)	4E-1(3)	4E-2(3)	4E-1(2)	8E-1(2)	3E-1(6)	9E-2(2)	4E-1(7)	8E-2(6)
TCGA-TGCT	9E-1(2)	3E-1(3)	6E-1(7)	1E-1(2)	1E-1(2)	**1E-1(4)**	1E-1(2)	7E-1(3)	3E-1(3)	5E-1(2)	2E-1(5)	4E-1(3)	6E-1(5)	7E-1(7)
TCGA-THCA	**6E-3(10)**	1E+0(3)	4E-2(10)	6E-3(3)	4E-1(2)	7E-1(3)	2E-2(3)	4E-2(4)	5E-1(3)	3E-1(2)	2E-2(5)	NA(NA)	4E-1(4)	5E-1(10)
TCGA-THYM	9E-3(2)	7E-1(3)	8E-1(6)	4E-1(3)	4E-1(3)	6E-1(3)	3E-1(3)	**1E-3(5)**	4E-1(3)	1E-2(2)	6E-3(9)	4E-1(2)	5E-1(4)	8E-1(6)
TCGA-UCEC	3E-3(10)	3E-1(3)	2E-2(7)	1E-3(2)	3E-2(2)	6E-3(2)	5E-3(2)	**1E-5(4)**	3E-2(3)	1E-3(2)	3E-2(10)	NA(NA)	5E-4(4)	1E+0(8)
TCGA-UCS	4E-1(4)	4E-1(3)	6E-1(4)	5E-1(3)	2E-1(2)	5E-1(3)	4E-1(3)	3E-2(10)	5E-1(2)	8E-1(2)	3E-1(6)	**3E-2(2)**	6E-1(4)	4E-1(3)
TCGA-UVM	**5E-5(2)**	5E-1(3)	1E-1(5)	7E-3(2)	1E-1(2)	2E-2(2)	7E-3(2)	2E-2(9)	5E-2(3)	9E-1(2)	3E-1(8)	1E-2(3)	3E-1(7)	1E-1(4)
P23918603	4E-2(2)	1E-1(3)	4E-2(4)	2E-1(2)	6E-2(2)	3E-2(2)	2E-1(2)	4E-1(6)	3E-2(2)	5E-1(2)	5E-2(5)	1E-1(4)	**2E-2(6)**	1E-1(5)
P24316975	**3E-2(2)**	5E-2(3)	1E-1(2)	3E-1(2)	9E-2(2)	5E-2(2)	3E-1(2)	2E-1(2)	9E-2(3)	1E-1(2)	2E-1(2)	8E-1(2)	5E-1(3)	4E-1(10)
P30244973	5E-2(2)	4E-1(3)	**2E-4(10)**	6E-1(3)	4E-1(3)	2E-2(4)	1E-1(3)	9E-1(6)	3E-1(5)	6E-1(2)	7E-1(3)	6E-1(3)	1E-1(5)	5E-1(7)
P33577785	9E-2(2)	4E-1(3)	**7E-3(4)**	8E-1(2)	4E-1(2)	7E-1(2)	8E-1(2)	8E-1(4)	1E+0(2)	6E-1(2)	2E-2(7)	1E+0(2)	6E-1(3)	8E-1(9)
P38007532	6E-1(2)	8E-1(3)	8E-1(3)	**2E-1(2)**	4E-1(2)	3E-1(2)	**2E-1(2)**	6E-1(4)	NA(NA)	1E+0(2)	8E-1(10)	9E-1(3)	9E-1(3)	NA(NA)
P38395893	9E-1(2)	5E-1(3)	6E-1(3)	7E-1(2)	1E+0(3)	7E-1(2)	4E-1(2)	4E-1(7)	NA(NA)	5E-1(2)	2E-1(3)	2E-1(3)	**1E-1(4)**	NA(NA)
P40016594	3E-2(2)	1E-3(3)	6E-5(10)	1E-1(2)	2E-2(2)	2E-3(2)	3E-2(2)	1E-1(2)	1E-2(2)	4E-2(2)	8E-6(8)	3E-2(2)	**5E-11(3)**	2E-1(9)
ST001235	5E-2(2)	**3E-3(4)**	2E-1(9)	7E-2(2)	8E-2(2)	3E-2(2)	9E-2(2)	1E-2(7)	NA(NA)	4E-1(2)	5E-3(8)	1E+0(2)	3E-2(4)	NA(NA)
ST001236	1E-1(2)	**1E-2(4)**	3E-1(9)	4E-1(3)	2E-1(2)	3E-1(2)	4E-1(3)	8E-2(9)	NA(NA)	9E-1(2)	3E-2(9)	8E-1(2)	1E-1(3)	NA(NA)
ST001237	2E-4(7)	2E-5(3)	5E-6(6)	1E-6(2)	5E-3(2)	3E-3(2)	5E-6(2)	**2E-7(6)**	NA(NA)	9E-2(2)	2E-6(8)	1E-2(2)	5E-3(5)	NA(NA)
Total sig. datasets	**27**	**12**	**11**	**11**	**10**	**15**	**15**	**19**	**9**	**8**	**17**	**8**	**11**	**7**
Average clus. no.	**5.07**	**3.13**	**5.45**	**2.33**	**2.3**	**2.75**	**2.38**	**4.74**	**2.77**	**2**	**6.5**	**2.49**	**4.7**	**6.53**

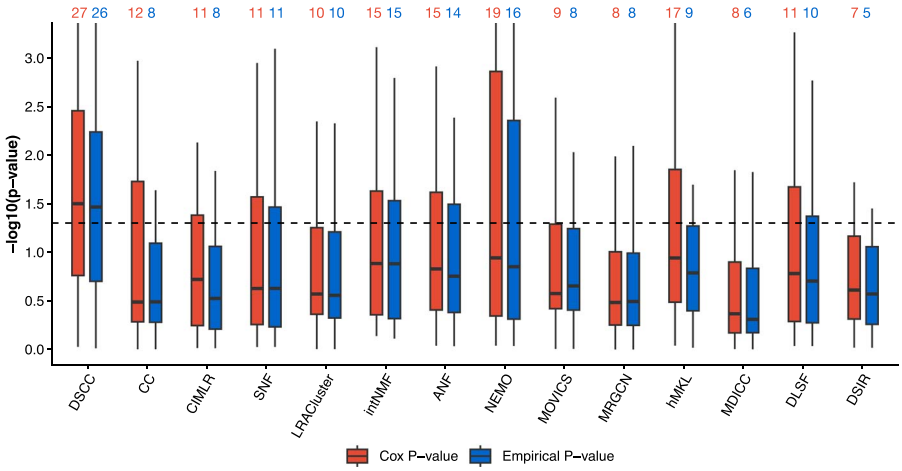

Figure 2. Distribution of Cox *P*-values and empirical *P*-values (in minus log10) for subtypes identified by each method across 43 cancer datasets. Higher values indicate stronger associations between discovered subtypes and patient survival. The black horizontal line inside each box represents the median value of the minus log10 *P*-values. The number on top of each box represents the total number of datasets with significant *P*-values. The dashed horizontal line denotes the statistical significance threshold corresponding to 5%. DSCC has median values of 1.5 and 1.47 (for Cox and empirical *P*-values, respectively), which are substantially higher than those obtained from all other methods. This demonstrates the advantage of DSCC in identifying cancer subtypes with significantly different survival profiles.

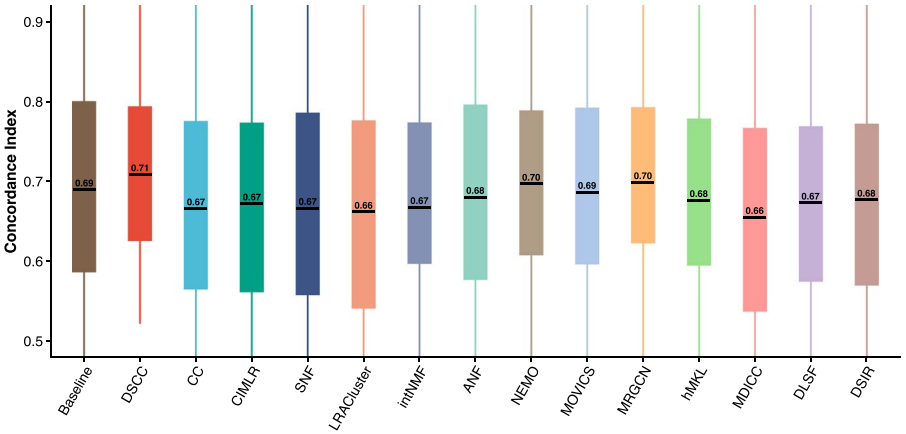

Figure 3. Evaluation of survival prediction models by incorporating subtype labels identified by DSCC, CC, CIMLR, SNF, LRACluster, intNMF, ANF, NEMO, MOVICS, MRGCN, hMKL, MDICC, DLSF and DSIR. For each dataset, we calculate the concordance index (C-Index) for a blockForest model that was trained using either clinical variables alone (baseline) or clinical variables combined with subtype labels from each method. Higher C-Index values indicate better alignment between predicted and real survival. The black horizontal line and the number inside each box both represent the mean value of the C-Index. The subtype information returned by DSCC achieves the highest mean C-Index, demonstrating the usefulness of its subtype labels in survival prediction.

group – a pattern consistent with previous studies reporting that high expression of cell cycle and DNA repair genes correlates with unfavorable prognosis in ACC [88, 89]. Moreover, the association between elevated base excision repair activity and tumor aggressiveness has also been observed in other cancer types [90].

Additional enriched pathways (*Proteasome*, *Spliceosome*, and *Ribosome*) reflect increased demand for transcription, translation, and protein degradation-hallmarks of rapidly proliferating tumors. It has been reported that upregulation of proteasome components is widely implicated in cancer progression and poor outcomes [91], while dysregulation of spliceosomal machinery and ribosome biogenesis contributes to oncogenic proliferation and survival [92, 93]. The pathway network (Fig. 6) also shows that these translational and protein-turnover pathways share numerous DE genes with the core cell-cycle pathway, underscoring their concerted activation. Significant enrichment of *ATP-dependent Chromatin Remodeling* and *Polycomb Repressive Complex* pathways further suggests widespread epigenetic alterations, consistent with dedifferentiation and aggressive tumor phenotypes. This aligns with the over-expression of EZH2, a key Polycomb complex member previously reported in adrenocortical carcinoma and associated with disease progression [94]. Together, these findings indicate that the poor-survival ACC subtype exhibits transcriptional signatures of hyperproliferation, replication stress, and epigenomic instability, aligning with its poor clinical prognosis.

In addition to proliferative signatures, several pathways enriched in the poor-survival ACC subtype are associated with cellular stress responses and tumor suppression. Specifically, the *Cellular Senescence* and *p53 Signaling* pathways are significantly enriched (Supplementary Table S5). These pathways are typically activated in response to genomic instability, oncogenic signaling, or DNA damage, and regulate checkpoint arrest, senescence, and apoptosis. Their enrichment in the poor-survival subtype may reflect cellular attempts to restrain unregulated proliferation or compensatory pathway activation following dysfunction in upstream regulators such as TP53, which is frequently altered in adrenocortical carcinoma [88, 94]. Although these stress-related programs are canonically tumor-suppressive,

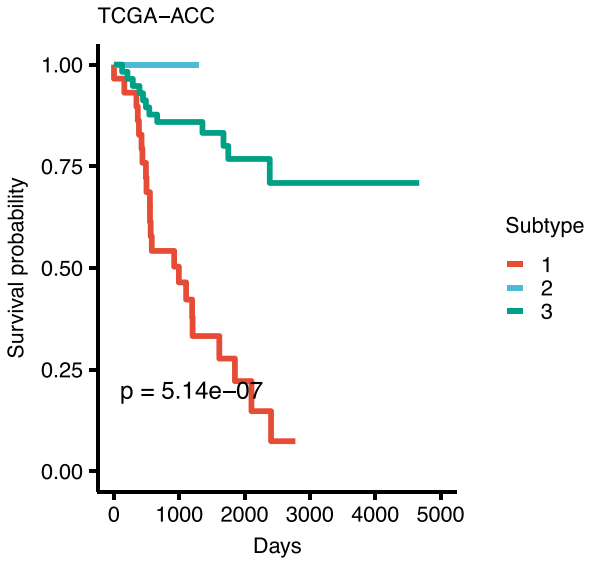

Figure 4. Kaplain-Meier survival analysis of the adrenocortical carcinoma (TCGA-ACC) dataset. The horizontal axis represents the time (days) passed after entry into the study while the vertical axis represents the estimated survival probability. The subtypes are identified by DSCC.

their dysregulation in aggressive tumors may promote therapy resistance, clonal selection, or intratumoral heterogeneity. Recent studies suggest that cellular senescence may even facilitate tumor progression through secretory phenotypes and immune modulation [95]. These findings suggest that even in the context of widespread proliferation, the poor-survival ACC subtype retains molecular signatures of intrinsic stress regulation and checkpoint signaling.

Several significantly enriched pathways in the poor-survival ACC subtype are annotated as viral or immune-related, including *Human T-cell Leukemia Virus 1 Infection*, *Epstein-Barr Virus Infection*, and *Kaposi Sarcoma-Associated Herpesvirus Infection* (Supplementary Table S5). These annotations do not imply actual viral infection, but rather reflect convergence on shared signaling mechanisms such as inflammatory responses, immune evasion, and anti-apoptotic mechanisms, which are commonly hijacked by tumors to sustain growth and to evade immune surveillance. This viral mimicry phenomenon, driven by derepressed endogenous retroviruses and innate immune signaling, has been reported in a number of other cancers [96]. In the pathway network (Fig. 6), the immune-annotated pathways share DE genes with core stress-response pathways, highlighting transcriptional overlap between immune modulation and checkpoint signaling programs. These results suggest that the poor-survival ACC subtype engages transcriptional programs associated with replication stress, immune modulation, and surveillance evasion-features commonly linked to tumor aggressiveness.

Conclusion

We present DSCC, a robust integrative framework for cancer subtyping and multi-omics integration. DSCC is among the first methods that integrate all types of available omics data (mRNA, miRNA, DNA methylation, CNV, somatic mutation, protein, and metabolite levels). While many methods can be extended to include a variety of omics types, DSCC is the first to validate its

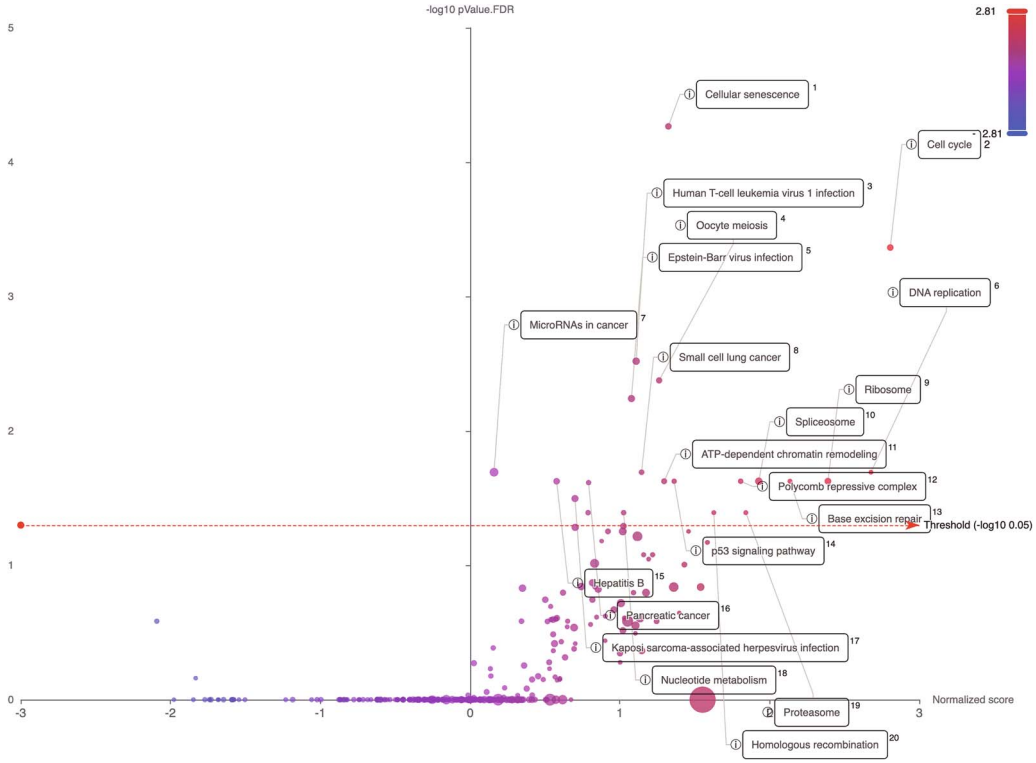

Figure 5. Pathway volcano plot for the consensus pathway analysis of the TCGA-ACC gene expression dataset with subgroups identified by DSCC. The x-axis shows the normalized enrichment score while the y-axis shows the minus log10 of FDR-adjusted *P*-value (pFDR). Each point on the figure represents a pathway or gene set. The size of a point is proportional to the number of genes in the corresponding gene set. The color of each point is determined by the normalized enrichment score. The top 20 pathways with the smallest pFDR are labeled with the pathway names.

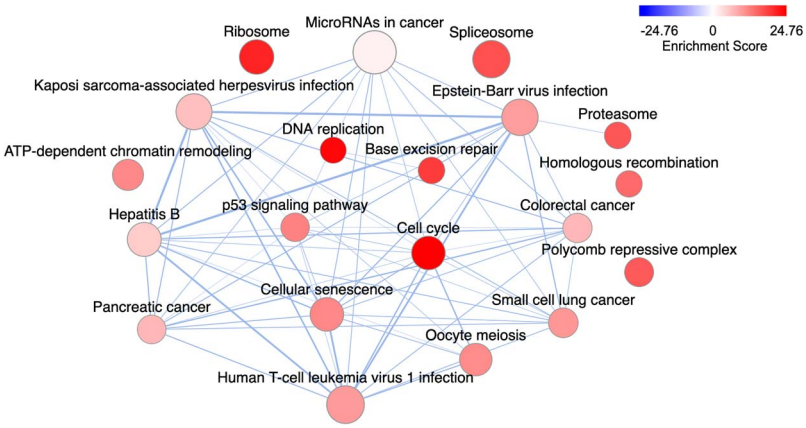

Figure 6. Pathway network of the top 20 significantly enriched pathways in the TCGA-ACC dataset based on the consensus analysis. Each node represents a biological pathway, with node size proportional to the number of genes it contains. Edges indicate gene overlap between pathways, and edge width reflects the number of shared genes. The color intensity within each node represents the enrichment score from the consensus result (as shown in the color bar). The network reveals a high degree of interconnectivity among pathways.

performance using 43 cancer datasets. DSCC makes three distinct technical contributions. First, DSCC achieves strong results with an interpretable pipeline that avoids complex architectures like deep learning, suggesting that a carefully designed pipeline can be effective in this research area. Second, DSCC introduces a dual-affinity matrix framework (Angular and Euclidean affinity matrices) that captures complementary views of each omics type. Finally, DSCC establishes a rigorous integration pipeline (flexible integration of available quantification units and gene-level aggregation), and allows for maximal sample retention.

We benchmark DSCC against 13 state-of-the-art methods (CC, CIMLR, SNF, LRACluster, intNMF, ANF, NEMO, MOVICS, MRGCN, hMKL, MDICC, DLSF, and DSIR) using 43 cancer datasets and three evaluation metrics (Cox *P*-values, empirical *P*-values, and C-Index). DSCC consistently outperforms other subtyping approaches by having more significant Cox *P*-values, empirical *P*-values, and higher C-Indices. Additionally, a detailed pathway analysis of the TCGA-ACC dataset demonstrates DSCC's ability to recover known oncogenic processes and uncover potential therapeutic targets in aggressive disease subtypes.

Several directions for future research could further enhance capabilities of DSCC. The current method uses pathway knowledge for data processing, but it has not fully modeled inter-layer biological dependencies. One potential direction is to integrate omics-specific regulatory relationships within pathway structures to extract mechanistic insights [97, 98]. Another direction is to integrate advanced techniques developed for other fields into the pipeline of DSCC, including low-rank symmetric affinity graphs [99], contrastive clustering [100], and Large Language Models (LLMs) [101]. Specifically, LLMs can process valuable, unstructured information available from electronic health records, pathology reports, and clinical information [102–104]. These data can serve as vital input for DSCC, supplementing molecular data for a more complete analysis.

Key Points

- Molecular subtyping is pivotal in cancer research, prognosis and treatment.
- Integrative analysis of diverse molecular data has the potentials to discover meaningful cancer subtypes.
- This article introduces DSCC, a novel method in cancer subtyping, which can work on a wide range of multi-omics data.
- The proposed method outperforms state-of-the-art approaches in identifying cancer subtypes with distinct survival profiles.
- The paper presents performance results comparing DSCC with other methods on 43 cancer datasets and a case study of Adrenocortical Carcinoma.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work is partially supported by the National Science Foundation (NSF: # 2343019 and # 2203236), the National Cancer Institute (NCI: # 1U01CA274573-01A1), and the National Institute of General Medical Sciences (NIGMS: # 1R44GM152152-01). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

Author contributions

D.T. developed the method, V.D.P., H.N., and P.B. helped with the case studies. T.N., A.Q., L.M.P. and S.C.J.Y. provided guidance for analysis and results' interpretation. D.T. and V.D.P. wrote the manuscript. All the authors reviewed the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: No competing interest is declared.

References

1. de Anda-Jáuregui G, Hernández-Lemus E. Computational oncology in the multi-omics era: State of the art. *Front Oncol* 2020;**10**:423. https://doi.org/10.3389/fonc.2020.00423

2. Menyhárt O, Gyoőrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput Struct Biotechnol J* 2021;**19**:949–60. https://doi.org/10.1016/j.csbj.2021.01.009
3. Karaman ED, Işik Z. Multi-omics data analysis identifies prognostic biomarkers across cancers. *Medical Sciences* 2023;**11**:1–24. https://doi.org/10.3390/medsci11030044
4. Kumar-Sinha C, Chinnaiyan AM. Precision oncology in the age of integrative genomics. *Nat Biotechnol* 2018;**36**:46–60. https://doi.org/10.1038/nbt.4017
5. Senft D, Leiserson MDM, Ruppin E. *et al*. Precision oncology: The road ahead. *Trends Mol Med* 2017;**23**:874–98. https://doi.org/10.1016/j.molmed.2017.08.003
6. Granja JM, Klemm S, McGinnis LM. *et al*. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* 2019;**37**:1458–65. https://doi.org/10.1038/s41587-019-0332-7
7. Curtis C, Shah SP, Chin S-F. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012;**486**:346–52. https://doi.org/10.1038/nature10983
8. Burstein MD, Tsimelzon A, Poage GM. *et al*. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 2015;**21**:1688–98. https://doi.org/10.1158/1078-0432.CCR-14-0432
9. Chaudhary K, Poirion OB, Liangqun L. *et al*. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853
10. Yang H, Chen R, Li D. *et al*. Subtype-GAN: A deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 2021;**37**:2231–7. https://doi.org/10.1093/bioinformatics/btab109
11. Wang C, Li J, Chen J. *et al*. Multi-omics analyses reveal biological and clinical insights in recurrent stage I non-small cell lung cancer. *Nat Commun* 2025;**16**:1–19. https://doi.org/10.1038/s41467-024-55068-2
12. Zhao Y, Gao Y, Xiaodong X. *et al*. Multi-omics analysis of genomics, epigenomics and transcriptomics for molecular subtypes and core genes for lung adenocarcinoma. *BMC Cancer* 2021;**21**:257. https://doi.org/10.1186/s12885-021-07888-4
13. Migliozzi S, Young Taek O, Hasanain M. *et al*. Integrative multi-omics networks identify PKCδ and DNA-PK as master kinases of glioblastoma subtypes and guide targeted cancer therapy. *Nature Cancer* 2023;**4**:181–202. https://doi.org/10.1038/s43018-022-00510-x
14. Herrera-Oropeza GE, Angulo-Rojo C, Gástelum-López SA. *et al*. Glioblastoma multiforme: A multi-omics analysis of driver genes and tumour heterogeneity. *Interface Focus* 2021;**11**:1–22.
15. Lindskrog SV, Prip F, Lamy P. *et al*. Aurélien de Reyniès, roman Nawroth, and Lars Dyrskjøt. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nature. Communications* 2021;**12**:1–18. https://doi.org/10.1038/s41467-021-22465-w
16. Mulong D, Dongying G, Xin J. *et al*. Integrated multi-omics approach to distinct molecular characterization and classification of early-onset colorectal cancer. *Cell Reports Medicine* 2023;**4**:1–13.
17. Zhao Z, Ding Y, Tran LJ. *et al*. Innovative breakthroughs facilitated by single-cell multi-omics: Manipulating natural killer cell functionality correlates with a novel subcategory of melanoma cells. *Front Immunol* 2023;**14**:1–24. https://doi.org/10.3389/fimmu.2023.1196892
18. Charoentong P, Finotello F, Angelova M. *et al*. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;**18**:248–62. https://doi.org/10.1016/j.celrep.2016.12.019
19. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics* 2013;**29**:2610–6. https://doi.org/10.1093/bioinformatics/btt425
20. Kirk P, Griffin JE, Savage RS. *et al*. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 2012;**28**:3290–7. https://doi.org/10.1093/bioinformatics/bts595
21. Xiaofan L, Meng J, Zhou Y. *et al*. MOVICS: An R package for multi-omics integration and visualization in cancer subtyping. *Bioinformatics* 2020;**36**:5539–41.
22. Brière G, Darbo É, Thébault P. *et al*. Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics* 2021;**22**:361. https://doi.org/10.1186/s12859-021-04279-1
23. Song W, Wang W, Dai D-Q. Subtype-WESLR: Identifying cancer subtype with weighted ensemble sparse latent representation of multi-view data. *Brief Bioinform* 2022;**23**:1–12.
24. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS One* 2017;**12**:1–18. https://doi.org/10.1371/journal.pone.0176278
25. Dingming W, Wang D, Zhang MQ. *et al*. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC Genomics* 2015;**16**:1022.
26. Mo Q, Shen R, Guo C. *et al*. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* 2018;**19**:71–86. https://doi.org/10.1093/biostatistics/kxx017
27. Mo Q, Wang S, Seshan VE. *et al*. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci* 2013;**110**:4245–50.
28. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**:2906–12. https://doi.org/10.1093/bioinformatics/btp543
29. Shen R, Mo Q, Schultz N. *et al*. Integrative subtype discovery in glioblastoma using iCluster. *PloS One* 2012;**7**:e35236. https://doi.org/10.1371/journal.pone.0035236
30. Yang B, Yang Y, Wang M. *et al*. MRGCN: Cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset. *Bioinformatics* 2023;**39**:1–8.
31. Zhang C, Chen Y, Zeng T. *et al*. Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Brief Bioinform* 2022;**23**:1–15. https://doi.org/10.1093/bib/bbab600
32. Yang B, Yang Y, Xueping S. Deep structure integrative representation of multi-omics data for cancer subtyping. *Bioinformatics* 2022;**38**:3337–42. https://doi.org/10.1093/bioinformatics/btac345
33. Wang B, Mezlini AM, Demir F. *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**:333–7. https://doi.org/10.1038/nmeth.2810
34. Rappoport N, Shamir R. NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 2019;**35**:3348–56. https://doi.org/10.1093/bioinformatics/btz058
35. Nguyen H, Tran B, Tran D. *et al*. PINSPlus: Clustering algorithm for data integration and disease subtyping. *R package version* 2020. 2.0.4.

36. Nguyen T, Tagett R, Diaz D. *et al*. A novel approach for data integration and disease subtyping. *Genome Res* 2017;**27**:2025–39. https://doi.org/10.1101/gr.215129.116
37. Arslanturk S, Draghici S, Nguyen T. Integrated cancer subtyping using heterogeneous genome-scale molecular datasets. In: Altman R (ed.), *Pacific Symposium on Biocomputing*. Singapore: World Scientific, 2020, 551–62.
38. Ramazzotti D, Lal A, Wang B. *et al*. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun* 2018;**9**:4453. https://doi.org/10.1038/s41467-018-06921-8
39. Ma T, Zhang A. Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering. In: Hu X (ed.), *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. New York: IEEE, 2017, 398–403.
40. Wei Y, Li L, Zhao X. *et al*. Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning. *Brief Bioinform* 2023;**24**:1–13. https://doi.org/10.1093/bib/bbac488
41. Yang Y, Tian S, Yushan Qiu P. *et al*. MDICC: Novel method for multi-omics data integration and cancer subtype identification. *Brief Bioinform* 2022;**23**:1–13.
42. Liu W, Wen Y, Yu Z. *et al*. Sphereface: Deep hypersphere embedding for face recognition. In: Liu Y (ed.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2017, 212–20.
43. Wang H, Wang Y, Zheng Z. *et al*. Cosface: Large margin cosine loss for deep face recognition. In: Forsyth D (ed.), *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*. New York: IEEE, 2018, 5265–74.
44. Deng J, Guo J, Xue N. *et al*. Arcface: Additive angular margin loss for deep face recognition. In: Gupta A (ed.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2019, 4690–9.
45. Andrew Y N, Michael JI, Weiss Y. *et al*. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 2002;**2**:849–56.
46. Blondel VD, Guillaume J-L, Lambiotte R. *et al*. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008;**2008**:1–12. https://doi.org/10.1088/1742-5468/2008/10/P10008
47. Kanehisa M, Furumichi M, Tanabe M. *et al*. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61. https://doi.org/10.1093/nar/gkw1092
48. Matthews L, Gopinath G, Gillespie M. *et al*. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**:D619–22. https://doi.org/10.1093/nar/gkn863
49. Monti S, Tamayo P, Mesirov J. *et al*. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003;**52**:91–118. https://doi.org/10.1023/A:1023949509487
50. Genomic data commons. *GDC Data Portal https://gdccancergov/* 2021.
51. Tarazona S, García-Alcalde F, Dopazo J. *et al*. Differential expression in rna-seq: A matter of depth. *Genome Res* 2011;**21**:2213–23. https://doi.org/10.1101/gr.124321.111
52. Risso D, Schwartz K, Sherlock G. *et al*. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 2011;**12**:480. https://doi.org/10.1186/1471-2105-12-480
53. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Brief Bioinform* 2018;**19**:776–92. https://doi.org/10.1093/bib/bbx008
54. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 2020;**26**:903–9. https://doi.org/10.1261/rna.074922.120
55. Bullard J, Purdom E, Hansen K. *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC bioinformatics* 2010;**11**:94. https://doi.org/10.1186/1471-2105-11-94
56. Frederick MJ, Perez-Bello D, Yadollahi P. *et al*. Reliable rna-seq analysis from ffpe specimens as a means to accelerate cancer-related health disparities research. *PloS One* 2025;**20**:e0321631. https://doi.org/10.1371/journal.pone.0321631
57. Zhao Y, Li M-C, Konaté MM. *et al*. A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *J Transl Med* 2021;**19**:1–15. https://doi.org/10.1186/s12967-021-02936-w
58. Hong LZ, Li J, Schmidt-Küntzel A. *et al*. Digital gene expression for non-model organisms. *Genome Res* 2011;**21**:1905–15. https://doi.org/10.1101/gr.122135.111
59. Cui S, Sicong Y, Huang H-Y. *et al*. miRTarBase 2025: Updates to the collection of experimentally validated microRNA–target interactions. *Nucleic Acids Res* 2025;**53**:D147–56. https://doi.org/10.1093/nar/gkae1072
60. Xia D-Y, Fei W, Zhang X-Q. *et al*. Local and global approaches of affinity propagation clustering for large scale data. *Journal of Zhejiang University-Science A* 2008;**9**:1373–81. https://doi.org/10.1631/jzus.A0720058
61. Zhu X, Loy CC, Gong S. Constructing robust affinity graphs for spectral clustering. In: Kacprzyk J (ed.), *Proceedings of the IEEE conference on computer vision and pattern recognition*. New York: IEEE, 2014, 1450–7.
62. Huang H-C, Chuang Y-Y, Chen C-S. Affinity aggregation for spectral clustering. In: Stelluto G (ed.), *2012 IEEE Conference on computer vision and pattern recognition*. New York: IEEE, 2012, 773–80.
63. Mark EJN. *Modularity and community structure in networks Proceedings of the National Academy of Sciences* 2006;**103**:8577–82. https://doi.org/10.1073/pnas.0601602103
64. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987;**20**:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
65. Benedetti E, Liu EM, Tang C. *et al*. A multimodal atlas of tumour metabolism reveals the architecture of gene–metabolite covariation. *Nat Metab* 2023;**5**:1029–44.
66. Zhang G, He P, Tan H. *et al*. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clin Cancer Res* 2013;**19**:4983–93. https://doi.org/10.1158/1078-0432.CCR-13-0209
67. Terunuma A, Putluri N, Mishra P. *et al*. Myc-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J Clin Invest* 2014;**124**:398–412. https://doi.org/10.1172/JCI71180
68. Gentric G, Kieffer Y, Mieulet V. *et al*. PML-regulated mitochondrial metabolism enhances chemosensitivity in human ovarian cancers. *Cell Metab* 2019;**29**:156–173.e10. https://doi.org/10.1016/j.cmet.2018.09.002
69. Wang L-B, Karpova A, Gritsenko MA. *et al*. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* 2021;**39**:509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006
70. Benedetti E, Chetnik K, Flynn T. *et al*. Plasma metabolomics profiling of 580 patients from an early detection research network

prostate cancer cohort. *Scientific Data* 2023;**10**:1–8. https://doi.org/10.1038/s41597-023-02750-7
71. Chen Y, Wang B, Zhao Y. *et al*. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. *Nature. Communications* 2024;**15**:1–13. https://doi.org/10.1038/s41467-024-46043-y
72. Coyle S, Chapman E, Hughes DM. *et al*. Urinary metabolite model to predict the dying process in lung cancer patients. *Commun Med* 2025;**5**:1–11. https://doi.org/10.1038/s43856-025-00764-3
73. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. *Crit Care* 2004;**8**:389–94. https://doi.org/10.1186/cc2955
74. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res* 2018;**46**:10546–62. https://doi.org/10.1093/nar/gky889
75. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat Med* 2005;**24**:3927–44. https://doi.org/10.1002/sim.2427
76. Therneau TM, Lumley T. Package 'survival'. *R Top Doc* 2015;**128**:28–33.
77. Hornung R, Wright MN. Block forests: Random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics* 2019;**20**:1–17.
78. Herrmann M, Probst P, Hornung R. *et al*. Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief Bioinform* 2021;**22**:1–15. https://doi.org/10.1093/bib/bbaa167
79. Tran D, Nguyen H, Pham V-D. *et al*. A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. *Brief Bioinform* 2025;**26**:1–17. https://doi.org/10.1093/bib/bbaf150
80. Nguyen H, Tran D, Galazka JM. *et al*. CPA: A web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Res* 2021;**49**:W114–24. https://doi.org/10.1093/nar/gkab421
81. Nguyen H, Nguyen H, Maghsoudi Z. *et al*. RCPA: An open-source r package for data processing, differential analysis, consensus pathway analysis, and visualization. *Current Protocols* 2024;**4**:e1036. https://doi.org/10.1002/cpz1.1036
82. Nguyen H, Pham V-D, Nguyen H. *et al*. CCPA: Cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and Reactome. *Brief Bioinform* 2024;**25**:bbae222.
83. Tavazoie S, Hughes JD, Campbell MJ. *et al*. Systematic determination of genetic network architecture. *Nat Genet* 1999;**22**:281–5. https://doi.org/10.1038/10343
84. Perroud B, Lee J, Valkova N. *et al*. Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol Cancer* 2006;**5**:64. https://doi.org/10.1186/1476-4598-5-64
85. Efron B, Tibshirani R. On testing the significance of sets of genes. *The Annals of Applied Statistics* 2007;**1**:107–29. https://doi.org/10.1214/07-AOAS101
86. Korotkevich G, Sukhov V, Budin N. *et al*. Fast gene set enrichment analysis. 2016;060012. http://biorxiv.org/content/early/2016/06/20/060012
87. Tarca AL, Drăghici S, Bhatti G. *et al*. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 2012;**13**:136. https://doi.org/10.1186/1471-2105-13-136
88. Subramanian C, Cohen MS. Over expression of DNA damage and cell cycle dependent proteins are associated with poor survival in patients with adrenocortical carcinoma. *Surgery* 2019;**165**:202–10. https://doi.org/10.1016/j.surg.2018.04.080
89. Qiqi L, Nie R, Luo J. *et al*. Identifying immune-specific subtypes of adrenocortical carcinoma based on immunogenomic profiling. *Biomolecules* 2023;**13**:104. https://doi.org/10.3390/biom13010104
90. Somuncu B, Keskin S, Antmen FM. *et al*. Non-muscle invasive bladder cancer tissues have increased base excision repair capacity. *Sci Rep* 2020;**10**:1–17. https://doi.org/10.1038/s41598-020-73370-z
91. Zhou X, Ruqing X, Yue W. *et al*. The role of proteasomes in tumorigenesis. *Genes & Diseases* 2024;**11**:101070. https://doi.org/10.1016/j.gendis.2023.06.037
92. Ivanova OM, Anufrieva KS, Kazakova AN. *et al*. Non-canonical functions of spliceosome components in cancer progression. *Cell Death Dis* 2023;**14**:1–17. https://doi.org/10.1038/s41419-022-05470-9
93. Bastide A, David A. The ribosome,(slow) beating heart of cancer (stem) cell. *Oncogenesis* 2018;**7**:34. https://doi.org/10.1038/s41389-018-0044-8
94. Drelon C, Berthon A, Mathieu M. *et al*. EZH2 is overexpressed in adrenocortical carcinoma and is associated with disease progression. *Hum Mol Genet* 2016;**25**:2789–800. https://doi.org/10.1093/hmg/ddw136
95. Liu B, Peng Z, Zhang H. *et al*. Regulation of cellular senescence in tumor progression and therapeutic targeting: Mechanisms and pathways. *Mol Cancer* 2025;**24**:106. https://doi.org/10.1186/s12943-025-02284-z
96. Chiappinelli KB, Strissel PL, Desrichard A. *et al*. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* 2015;**162**:974–86. https://doi.org/10.1016/j.cell.2015.07.011
97. Lan W, Liao H, Chen Q. *et al*. DeepKEGG: A multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Brief Bioinform* 2024;**25**:1–16.
98. Lan W, Tang Z, Liao H. *et al*. MULGONET: An interpretable neural network framework to integrate multi-omics data for cancer recurrence prediction and biomarker discovery. *Fundamental Research* 2025;**20**:1–13. https://doi.org/10.1016/j.fmre.2025.01.004
99. Lan W, Yang T, Chen Q. *et al*. Multiview subspace clustering via low-rank symmetric affinity graph. *IEEE Transactions on Neural Networks and Learning Systems* 2023;**35**:11382–95. https://doi.org/10.1109/TNNLS.2023.3260258
100. Lan W, Zhou G, Chen Q. *et al*. Contrastive clustering learning for multi-behavior recommendation. *ACM Transactions on Information Systems* 2024;**43**:1–23. https://doi.org/10.1145/3698192
101. Lan W, Tang Z, Liu M. *et al*. The large language models on biomedical data analysis: A survey. *IEEE J Biomed Health Inform* 2025;**29**:4486–97. https://doi.org/10.1109/JBHI.2025.3530794
102. Lee J, Yoon W, Kim S. *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40. https://doi.org/10.1093/bioinformatics/btz682
103. Luo R, Sun L, Xia Y. *et al*. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;**23**:1–11.
104. Li Y, Li Z, Zhang K. *et al*. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* 2023;**15**:1–12.

Briefings in Bioinformatics, 2025, **26(6)**, bbaf600
https://doi.org/10.1093/bib/bbaf600
Problem Solving Protocol