ORIGINAL ARTICLE

GMJ.2026;15:e4304
www.salviapub.com

Received 2026-04-28
Revised 2026-06-13
Accepted 2025-07-29

Pan-Cancer Transcriptomic Clustering Reveals Conserved Molecular Programs and Cluster-Associated Compound Signatures

Mahnaz Mohammadi [1], Maryam Arabi [2] , Ahmad Bereimipour [2]✉

[1] University of North Texas, Department of Information Sciences, Denton, Texas, 76203, USA
[2] Department of Stem Cells and Developmental Biology, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran

Abstract

Background: Pan-cancer transcriptomic analyses can identify molecular programs shared across tumor types. This study performed integrative unsupervised clustering of RNA-seq data to characterize expression-defined tumor groups and associated biological programs.

Materials and Methods: RNA-seq data from 11,296 TCGA primary tumor samples across 33 cancer types were analyzed using preprocessing, principal component analysis, exploratory k-means clustering, differential-expression analysis, pathway enrichment, DSigDB compound-signature enrichment, independent CCLE analysis, and exploratory PRISM assessment.

Results: Six exploratory transcriptomic clusters showed variable relationships with tissue of origin and were associated with immune activation, cell-cycle regulation, metabolic processes, extracellular matrix remodeling, developmental signaling, and stress-response programs. DSigDB identified non-directional compound-signature overlaps involving retinoic acid derivatives, HDAC inhibitors, glucocorticoids, metabolic modulators, and toxicogenomic agents. CCLE analysis identified a stable six-group partition with partial concordance to TCGA clusters, and PRISM provided limited compound-specific pharmacogenomic support.

Conclusion: This study characterizes transcriptomic heterogeneity across cancers and identifies biological and compound-signature associations that require further transcriptomic, clinical, pharmacogenomic, and experimental validation.

Keywords: Pan-Cancer Transcriptomics; Unsupervised Transcriptomic Clustering; Pathway Enrichment; Drug-Gene Perturbation Signatures; Cross-Tumor Molecular Stratification

[GMJ.2026;15:e4304] DOI:10.31661/gmj.v15i0.4304

Introduction

Cancer represents a heterogeneous group of diseases driven by diverse genetic, epigenetic, and microenvironmental alterations [1,2]. Although individual

Copyright© 2026, Galen Medical Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)
Email:gmj@salviapub.com

✉ **Correspondence to:**
Ahmad Bereimipour
Department of Stem Cells and Developmental Biology, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran,
TelephoneNumber:+1 (940)629-2360
Email Address: ahmadbereimipour@my.unt.edu

tumor types have traditionally been classified according to tissue of origin, transcriptomic and multi-omics evidence shows that cancers may converge on shared molecular programs across anatomical boundaries [3,4]. Pan-cancer analyses aim to identify these conserved biological programs, improve understanding of tumor biology, and support stratification frameworks that extend beyond organ-specific classifications [5,6].

Gene expression profiling is a powerful approach for delineating cross-cancer patterns because it captures tumor functional states and integrates upstream genomic events into coherent transcriptional outputs [7]. Unsupervised clustering of large transcriptomic cohorts can identify latent expression subtypes associated with pathways, immune states, metabolic reprogramming, or stromal interactions across cancers [8]. Such transcriptomic subtypes may refine tumor taxonomy and identify biological programs not evident in conventional classifications.

However, many pan-cancer studies remain limited in integrating pathway information with compound-associated perturbation signatures [9]. Systematic linkage of expression-defined clusters to functional pathways and pharmacologic perturbation profiles can characterize cluster-associated biological programs and identify compound-associated transcriptional signatures for further investigation.

Here, we conducted an integrative pan-cancer transcriptomic analysis to identify expression-defined clusters and characterize their associated biological programs. Cluster-associated genes were subsequently compared with compound- and chemical-related gene sets in DSigDB. The resulting overlaps are presented as hypothesis-generating compound associations rather than evidence of drug response, therapeutic efficacy, or repurposing potential.

Materials and Methods

The Cancer Genome Atlas (TCGA) Pan-Cancer Transcriptomic and Clinical Data Acquisition

RNA-seq expression profiles and matched clinical metadata were obtained from the TCGA Pan-Cancer cohort through the Genomic Data Commons (GDC) Pan-Cancer (PANCAN) resource. The initial dataset contained 17,180 samples representing 33 cancer types and approximately 20,000 gene features. Sample-type information was determined using TCGA barcode annotations [10]. Only primary tumor samples were retained for the main analysis. Solid-tissue normal, metastatic, recurrent, additional primary tumor, and other non-primary samples were excluded. When multiple aliquots or samples were available for the same patient, one representative primary tumor sample was retained. Samples lacking valid cancer-type annotations, clinical metadata, or complete expression profiles were also removed. After these sample-selection steps, 11,296 primary tumor samples remained and constituted the final cohort used for transcriptomic preprocessing and unsupervised clustering.

Preprocessing, Normalization, and Log Transformation

Gene identifiers were harmonized by removing Ensembl version suffixes and mapping Ensembl identifiers to official HGNC gene symbols. Genes that could not be mapped to an official gene symbol were excluded. When multiple Ensembl identifiers mapped to the same gene symbol, they were collapsed by retaining the feature with the highest median expression across all samples. Low-abundance genes were removed by

requiring an FPKM value of at least 1 in at least 20% of the samples. The retained expression values were transformed using log2(FPKM + 1).

Because FPKM values had already undergone library-size normalization within the harmonized GDC pipeline, no additional count-based normalization was applied. Sample-wise median scaling was used to reduce global expression-intensity differences among samples. Genes falling within the lowest 25% of the variance distribution were excluded as low-variance features. The remaining genes were ranked according to their variance across samples, and the 5,000 most variable genes were selected for principal component analysis (PCA) and unsupervised clustering.

Available TCGA technical metadata, including sequencing center, tissue source site, and plate information, were reviewed as potential sources of technical variation. However, formal batch correction was not applied because these variables were incompletely available and were strongly confounded with cancer type and tissue lineage. Sample-wise median scaling was therefore used only as a normalization procedure and was not considered a formal batch-correction method. Cancer type and tissue lineage were retained as biologically relevant sources of variation and were not regressed out. Tumor purity was considered during interpretation where corresponding purity estimates were available.

Unsupervised Clustering of Pan-Cancer Transcriptomes

To investigate the latent transcriptomic structure across tumor types, the processed expression matrix containing the 5,000 most variable genes was subjected to principal component analysis. The first 50 principal components were used as input for k-means clustering. Candidate cluster solutions ranging from k = 2 to k = 10 were systematically evaluated using within-cluster sum of squares, mean silhouette coefficients, and repeated clustering stability analysis.

The first 50 principal components were used as input for k-means clustering with k-means++ initialization, 100 random initializations, a maximum of 300 iterations, and a fixed random seed of 42. The six-cluster solution used in the present study was generated during the original exploratory clustering workflow and was retained for subsequent biological characterization. However, the run-level diagnostic outputs required reporting within-cluster sum of squares, silhouette coefficients, repeated-clustering stability, and consensus matrices were not retained. Therefore, the six-cluster solution is presented as an exploratory partition rather than as a uniquely optimal or quantitatively validated clustering solution.

Cluster Annotation and Cancer-Type Composition Analysis

Each TCGA sample was assigned to one of the six transcriptomic clusters. Cancer-type labels that were derived from TCGA phenotype files were mapped onto cluster assignments in order to quantify tumor type distribution across clusters. Cluster-cancer associations were visualized in the form of stacked bar plots, proportional composition diagrams, and cross-tabulations. Analysis along these lines allowed the biological heterogeneity within clusters to be interpreted and cancer types disproportionately enriched within transcriptional states to be identified.

Calculation of the DriverScore

To rank genes according to their contribution to separation among the exploratory transcriptomic clusters, we calculated a study-specific Fisher-

type variance-ratio metric, referred to as the DriverScore. For each gene g, the between-cluster mean-square component was calculated as:

The within-cluster mean-square component was calculated as:
The DriverScore for gene g was then defined as:
where K is the number of transcriptomic clusters; C_k is the set of samples assigned to cluster k; n_k is the number of samples in cluster k; N is the total number of samples included in the clustering analysis; x_ig is the normalized expression value of gene g in sample i; μgk is the mean normalized expression of gene g among samples assigned to cluster k; and μg is the mean normalized expression of gene g across the complete clustering cohort. The constant $\varepsilon = 10^{-8}$ was added to the denominator to prevent division by zero. Higher DriverScore values indicate greater between-cluster expression variability relative to within-cluster variability. DriverScore was used only as a study-specific gene-ranking metric based on the conventional Fisher/ANOVA variance-ratio framework and is not presented as an independently validated biomarker score.

Differential Expression and Pathway Enrichment Analysis

Cluster-specific gene-expression analysis was performed using the limma package in R/Bioconductor. The analysis was conducted on the preprocessed FPKM expression matrix after low-abundance filtering, log2(FPKM + 1) transformation, and sample-wise median normalization across the complete pan-cancer cohort. For each gene, a linear model was fitted with transcriptomic cluster as the primary variable of interest and TCGA cancer type as a categorical covariate to reduce potential confounding by tissue of origin. Cluster-specific contrasts were defined using a one-versus-rest framework, in which each transcriptomic cluster was compared with the combined samples from the remaining five clusters. Statistical significance was evaluated using empirical Bayes-moderated t-statistics implemented in limma. P-values were adjusted for multiple testing using the Benjamini-Hochberg false-discovery-rate procedure. Genes with an adjusted P-value <0.05 and a log2 fold change ≥1 were classified as cluster-associated upregulated genes.

The upregulated gene set identified for each cluster was submitted to Enrichr for functional enrichment analysis using GO Biological Process, GO Molecular Function, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome libraries. Enrichment terms with a Benjamini-Hochberg-adjusted P-value <0.05 were considered statistically significant.

Driver-Gene Correlation and Hierarchical-Clustering Analysis

Exploratory correlation analysis was performed using the 20 genes with the highest DriverScore values. Pairwise Pearson correlation coefficients were calculated across the complete pan-cancer cohort using the normalized expression matrix. The resulting correlation matrix was hierarchically clustered using correlation distance, defined as 1−r, and average-linkage clustering. The clustered heatmap was used to visualize descriptive blocks of positively or negatively correlated genes. WGCNA, dynamic tree cutting, community detection, or another formal module-detection method was not applied. Therefore, the observed gene groups were interpreted only as descriptive correlation blocks.

Quantification of Tissue-of-Origin Dependence

The dependence between transcriptomic cluster assignment and TCGA cancer type

was quantified using the adjusted Rand index, normalized mutual information, and Cramér's V. A Pearson chi-square test was applied to the cancer type-by-cluster contingency table. Cancer-type diversity within each cluster was assessed using Shannon entropy normalized by the maximum possible entropy across the 33 cancer-type categories. Higher normalized entropy indicated greater cancer-type heterogeneity within a cluster. Cancer-type overrepresentation within individual clusters was evaluated using one-sided hypergeometric tests, followed by Benjamini-Hochberg false-discovery-rate correction.

External Replication Analysis Using CCLE

External replication was assessed using RNA-seq TPM profiles from the Cancer Cell Line Encyclopedia (CCLE) [11]. The expression matrix contained 1,019 cell lines and 57,820 gene features. Expression values were transformed as log2(TPM + 1). Genes were retained when TPM was at least 1 in at least 20% of cell lines, and the 5,000 genes with the highest variance were selected for principal component analysis. The first 50 principal components were used for k-means clustering with k-means++ initialization, 100 random initializations, a maximum of 300 iterations, and a fixed random seed of 42. Candidate solutions from k = 2 to k = 10 were evaluated using within-cluster sum of squares and mean silhouette coefficients. Stability was assessed across 30 repeated clustering runs using pairwise adjusted Rand indices. For the six-cluster solution, consensus clustering was performed using 100 iterations in which 80% of the cell lines were sampled without replacement. CCLE clusters were compared with the TCGA clusters using cancer-type composition across 23 shared cancer categories. Because CCLE consists of cancer cell lines rather than primary tumor samples, this analysis was interpreted as independent replication rather than direct validation of identical cluster assignments.

Drug Signature Enrichment Using Drug Signatures Database (DSigDB)

Cluster-specific upregulated genes were queried against the DSigDB drug-gene signature database to identify statistically significant overlaps with compound-associated transcriptional signatures. Enrichment was assessed using Fisher's exact test, and compounds were ranked according to adjusted P-values and combined enrichment scores. The resulting associations were interpreted as similarities between cluster-specific expression programs and previously reported drug- or chemical-induced gene signatures. They were not considered evidence of drug sensitivity, therapeutic efficacy, or clinical response. Toxicants and environmental chemicals were categorized as toxicogenomic markers, whereas pharmacologically relevant compounds were separately identified as candidates for future experimental evaluation. Although the input lists consisted of cluster-associated upregulated genes, DSigDB [12] enrichment was based on non-directional gene-set overlap. The analysis did not determine whether a compound-associated transcriptional signature was concordant with or opposed to the cluster-associated expression program. No connectivity score, reversal score, or direction-resolved perturbation analysis was calculated. Enriched terms were therefore interpreted only as compound-signature overlaps.

Construction of Cluster-Cancer Relationship Visualizations

Several visualization frameworks were implemented as a way to represent

heterogeneous connections between cancer types and transcriptomic clusters. Plotly was used to generate Sankey diagrams that display the flow of samples from each cancer type into its designated cluster, highlighting cross-cancer convergence and divergence. Lastly, chord diagrams or circular cluster-cancer interfaces were generated in Python using libraries based on Holoviews. These visualizations provided a clear overview of how cancers distribute across expression-defined states.

t-Distributed Stochastic Neighbor Embedding (t-SNE) Visualization of Sample-Level Transcriptomic Structure

Dimensionality-reduction analysis was performed using t-distributed stochastic neighbor embedding to visualize global sample relationships. High-variance genes were used as input, and t-SNE embeddings were computed with optimized perplexity values tailored to sample size. Samples were colored by transcriptomic cluster and cancer type, allowing local and global similarities across tumors to be explored. This afforded a complementary representation of the clustering structure beyond linear PCA projections.

Integration of Drug Signatures with Transcriptomic States

Drug- and chemical-signature enrichment results were summarized according to adjusted P values and combined enrichment scores. Compounds were annotated according to their reported biological context and classified as pharmacologic compounds, environmental chemicals, or toxicants. The results were interpreted solely as non-directional overlaps between cluster-associated genes and compound-annotated gene sets. Toxicants were retained only as toxicogenomic markers and were not considered therapeutic candidates. No inference was made regarding drug sensitivity, treatment response, therapeutic efficacy, or repurposing potential.

Statistical Analysis and Reproducibility

All statistical analyses and visualizations were performed in Python using pandas, NumPy, SciPy, scikit-learn, Matplotlib, Seaborn, and Plotly. PCA and k-means clustering were implemented using scikit-learn. K-means clustering used k-means++ initialization, 100 random initializations, a maximum of 300 iterations, and a fixed random seed of 42. The number of clusters was evaluated using silhouette coefficients, within-cluster sum of squares, and stability across repeated initializations.

For enrichment analyses involving multiple comparisons, P-values were adjusted using the Benjamini-Hochberg false-discovery-rate procedure. Enriched terms with an adjusted P-value below 0.05 were considered statistically significant. t-SNE was performed using a fixed random seed of 42, an initial PCA reduction, a learning rate set to "auto," and a perplexity of 30. All sample-exclusion criteria, gene-filtering thresholds, cluster assignments, enrichment outputs, and random seeds were maintained within the same computational workflow to support reproducibility.

Cluster-specific expression differences were assessed using linear models and empirical Bayes-moderated t-statistics implemented in the limma R/Bioconductor package. One-versus-rest contrasts were constructed for each cluster, with cancer type included as a categorical covariate. Multiple-testing correction was performed using the Benjamini-Hochberg procedure. Cluster-associated upregulated genes were defined by an adjusted P-value <0.05 and a log2 fold change ≥ 1.

All computational analyses were performed using Python version 3.12.2 with pandas version 2.2.2, NumPy version

1.26.4, SciPy version 1.13.1, scikit-learn version 1.5.1, Matplotlib version 3.9.2, Seaborn version 0.13.2, and Plotly version 5.24.1. Differential-expression analyses were performed using R version 4.4.1 with Bioconductor version 3.19 and the limma package version 3.60.6. Enrichr [13] and DSigDB were accessed on July 30, 2026. Random seeds, filtering thresholds, clustering parameters, and multiple-testing procedures are reported in the corresponding Methods subsections.

Clinical outcome analyses were not performed because sample-level cluster assignments were not linked to harmonized survival, stage, grade, treatment-response, immune-infiltration, mutation, or copy-number data in the present workflow. Accordingly, the study was not designed to assess prognostic, predictive, or treatment-response associations.

Results

Exploratory Clustering Identifies Six Transcriptomic Groups Across TCGA Pan-Cancer Samples

Following sample-level filtering, gene-identifier harmonization, and expression preprocessing, the final TCGA cohort comprised 11,296 primary tumor samples. The transcriptomic data were represented using the first 50 principal components, and the exploratory k-means solution divided the cohort into six groups, designated C0-C5. The clusters contained 433, 2,611, 1,722, 903, 1,830, and 3,797 samples, respectively, totaling 11,296 samples. These cluster sizes and their cancer-type compositions are reported in Supplementary Table-S1. PCA visualization showed separation among several assigned groups, although this visualization was not considered quantitative evidence that six was the uniquely optimal number of clusters. Because the original elbow, silhouette, stability, and consensus-clustering outputs were not available, the six-cluster structure should be interpreted as exploratory.

To rank genes according to their contribution to cluster separation, we calculated the DriverScore, a Fisher-type variance-ratio statistic comparing between-cluster expression variability with within-cluster variability. The top 20 genes exhibited the greatest discriminatory power (Figure-1B), with many falling into cytoskeletal, junctional, or epithelial integrity pathways (e.g., *KRT18*, *STMN4*, *HEPACAM*). A number of these candidates represent known drivers in cancer; others emerge here as novel, cluster-specific biomarkers not previously implicated in pancancer differentiation. This ranking provided a systematic basis for the identification of transcripts defining major axis variations across tumor transcriptomes. Heatmap profiling of the top driver genes showed pronounced heterogeneity between clusters (Figure-1C). For instance, genes such as *KRT18* and *ITIH1* showed strong enrichment in Cluster 0, indicating active epithelial differentiation programs. Other driver genes, such as *CEACAM5*, *NECTIN4*, and *SH2D3A*, marked specific clusters but were minimally expressed in others, denoting lineage-restricted or microenvironment-associated pathways. A number of neuronal-signature genes (e.g., *ELAVL3*, *CACNG7*, *ATCAY*) were preferentially activated in Clusters 3 and 5, pointing toward possible neural-like or dedifferentiated states. In general, these patterns suggest that each transcriptomic cluster is driven by a unique combination of pathway-level regulators.

Correlation Analysis Reveals Coordinated Expression Patterns Among Top Driver Genes

To examine coordinated expression among the most discriminative genes, pairwise Pearson correlation coefficients were calculated for the 20 genes with the

highest DriverScore values, followed by hierarchical clustering of the correlation matrix. The resulting heatmap revealed several blocks of positively and negatively correlated genes (Figure-2A). These patterns indicate coordinated expression among subsets of driver genes and were interpreted only as descriptive correlation blocks because no formal gene-group detection algorithm was applied. For each gene, DriverScore was calculated as the ratio of between-cluster mean-square variability to within-cluster mean-square variability. Genes with large expression differences among clusters and relatively compact expression distributions within clusters received higher scores. The DriverScore values were highly right-skewed, as shown in Figure-2B, meaning that only a small subset of genes exhibited prominent influence while the majority showed minimal discriminative power. Such a pattern describes the classical "long-tail" architecture of cancer transcriptomes in which a limited number of transcripts

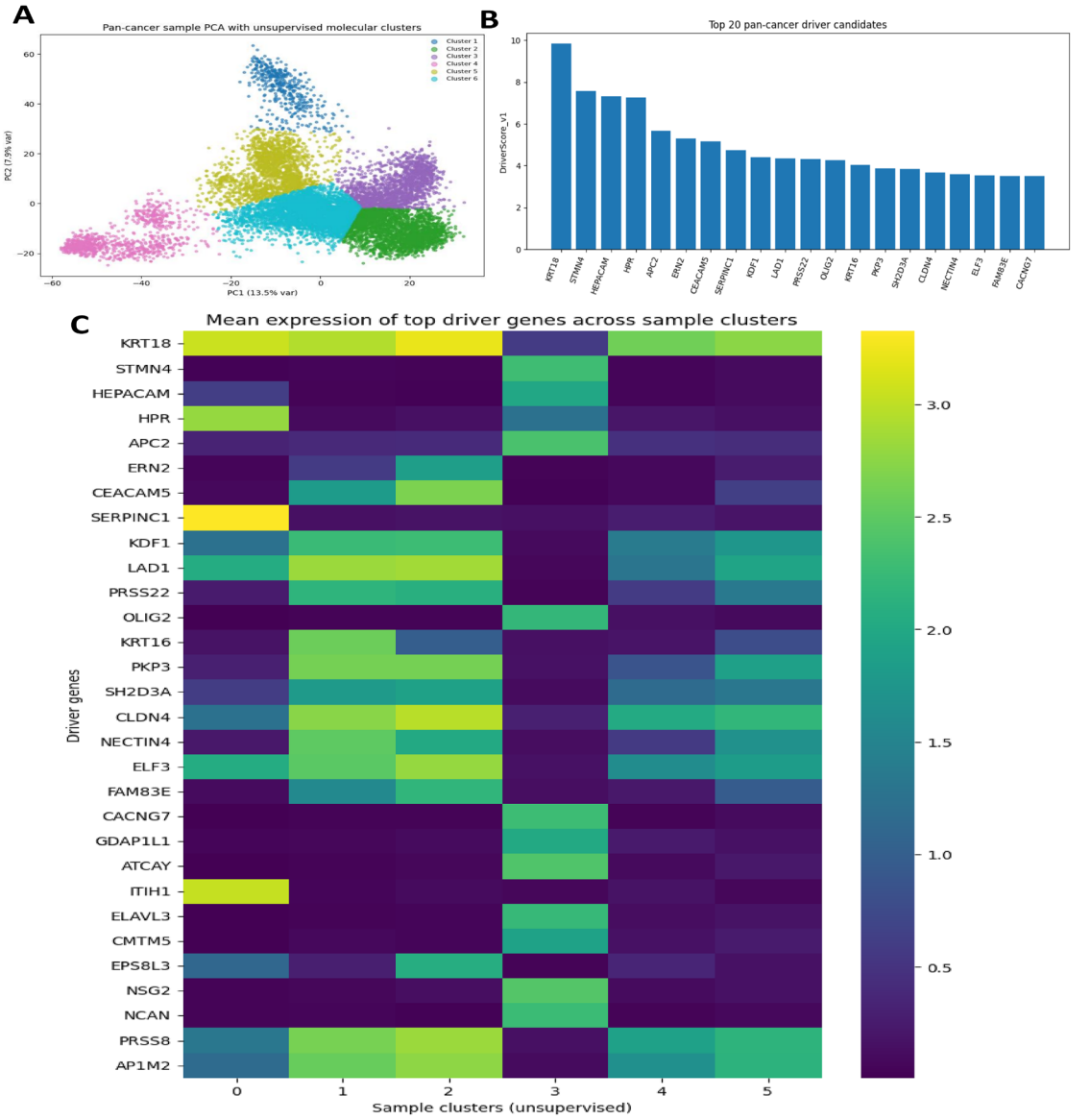


Figure 1. Transcriptomic clustering and driver-gene characterization across 11,296 TCGA primary tumor samples. (A) PCA projection of tumor samples colored according to the six exploratory transcriptomic clusters. (B) Top 20 genes ranked by DriverScore, showing the strongest contribution to cluster separation. (C) Heatmap of mean expression values for top driver genes across clusters, illustrating cluster-specific molecular programs and distinct pathway activation profiles.

serve as major transcriptional anchors that define global tumor stratification across the pan-cancer cohort. High-scoring genes thus represented core biomarkers and potential mechanistic regulators of tumor heterogeneity.

Degree and betweenness centrality were calculated to describe the relative connectivity of individual genes within the thresholded correlation network. Genes with higher degree generally also showed higher betweenness, indicating that highly connected genes tended to occupy central positions within the constructed network. "These centrality measures provide a descriptive summary of network topology but do not establish discrete gene-group membership, higher-order network organization, regulatory causality, or master-regulator function."

Cluster-Specific Expression Patterns Reveal Distinct Molecular Identities

To dissect the transcriptional heterogeneity among the six unsupervised pan-cancer transcriptomic clusters, we assessed cluster-specific z-scored expression profiles of the top driver genes. The resulting heatmap demonstrated sharp, nonoverlapping blocks of high expression. Each cluster expressed a unique molecular signature that was enriched for distinct sets of driver transcripts. For instance, Clusters 0 and 1 demonstrated strong and mutually exclusive activation of epithelial-associated transcripts, whereas Cluster 3 was characterized by broad repression of most drivers except a subset involved in stromal or immune-related functions. These expression patterns show differences among the six exploratory transcriptomic clusters. The observed block-like expression patterns suggest coordinated variation among subsets of driver genes, but they should be interpreted descriptively because the clustering solution was not independently validated.

To examine coordinated expression beyond individual genes, we inspected the hierarchically ordered correlation matrix together with the cluster-level z-scored expression patterns. This analysis revealed descriptive blocks of genes with similar expression profiles across the six transcriptomic clusters. Cluster 5 showed relatively higher expression of blocks containing proliferation- and metabolism-related genes, whereas Cluster 2 showed higher expression of blocks containing neural-like or developmental genes. Cluster 3 displayed lower expression across several descriptive gene blocks. Because no formal gene-group detection algorithm was applied, these patterns were interpreted only as descriptive correlation and expression blocks. No quantitative activity score was calculated for these descriptive blocks.

We then used t-SNE to project the full transcriptome onto a two-dimensional manifold and observed compact sample groupings corresponding to the six transcriptomic clusters (Figure-3C). The limited overlap among these groups provides a visual representation of the exploratory clustering structure but does not establish the robustness or optimality of the six-cluster solution. The t-SNE landscape suggested that Cluster 5 was the largest and most compositionally diverse region, whereas Clusters 0 and 3 formed more tightly condensed regions, suggestive of lower intra-cluster heterogeneity. Samples from different tumor types co-localized within several clusters, indicating that some cluster boundaries reflect shared transcriptional programs rather than tissue of origin alone.

Unsupervised Transcriptomic Clusters Are Broadly but Non-Uniformly

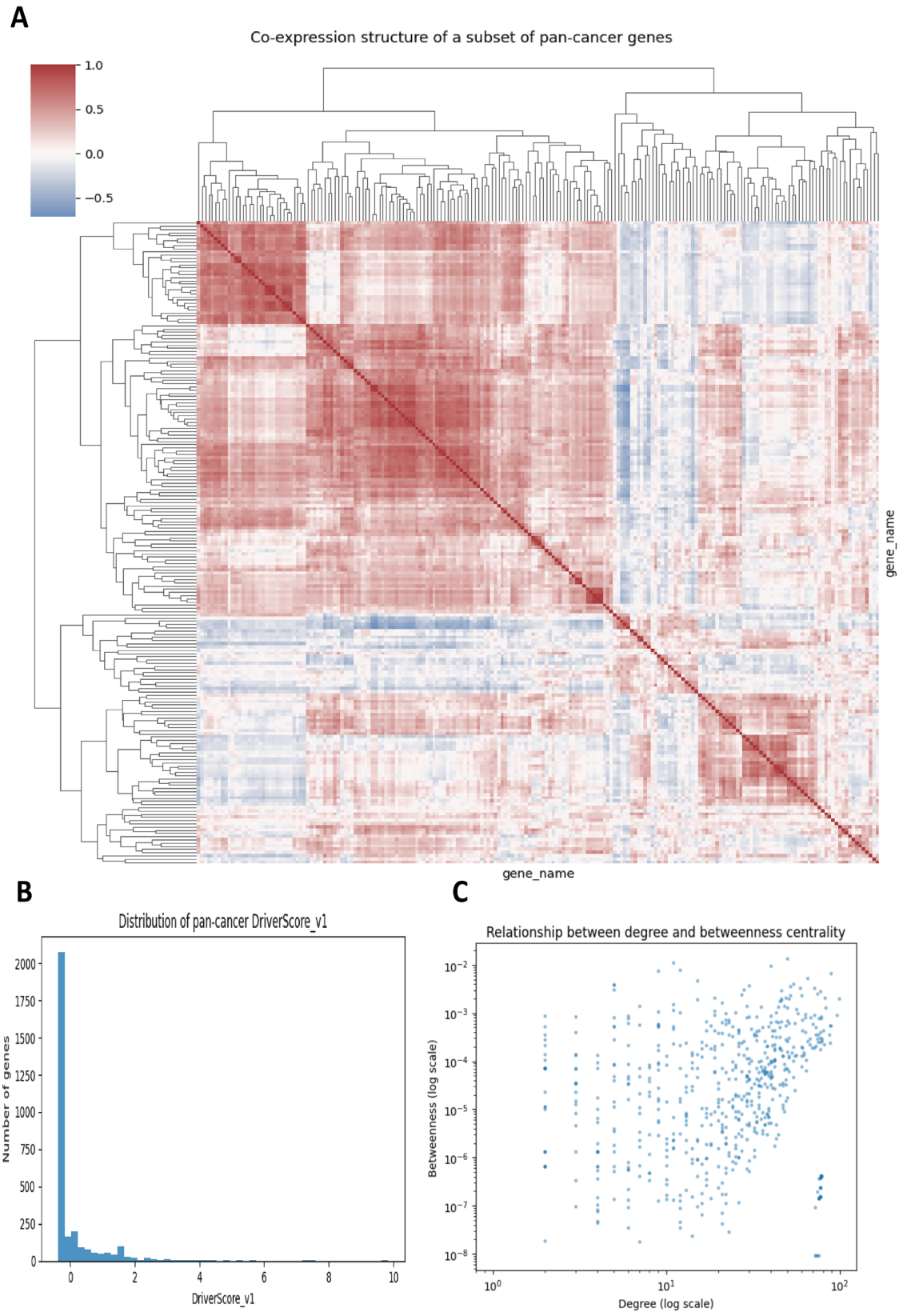


Figure 2. Correlation structure and centrality analysis of pan-cancer driver genes. (A) Hierarchical clustering of the gene-gene correlation matrix showing descriptive blocks of positively and negatively correlated genes. These blocks were not identified using a formal module-detection method. (B) Distribution of the DriverScore metric. (C) Relationship between degree and betweenness centrality in the thresholded correlation network.

Distributed Across Cancer Types

We next examined the distribution of the six exploratory transcriptomic clusters across 33 TCGA cancer types. Cancer-type composition was markedly non-uniform across clusters (Figure-4A-C). Some clusters were strongly enriched for particular tissue lineages, whereas others contained samples from a broader range of cancer types. For example, Cluster C0 was dominated by liver hepatocellular carcinoma, which accounted for 397 of 433 samples (91.7%), whereas Cluster C3 was dominated by brain lower-grade glioma, which accounted for 536 of 903 samples (59.4%). In contrast, Clusters C1, C2, C4, and C5 showed greater cancer-type heterogeneity.

Quantitative analysis confirmed substantial dependence between cancer type and transcriptomic cluster assignment. The cancer type-by-cluster contingency table showed a significant association (Pearson χ^2 = 40,037.16, df = 160, $P<1 \times 10^{-300}$), with a Cramér's V of 0.842. The adjusted Rand index was 0.165, and normalized mutual information was 0.449. These values indicate a strong association between cancer type and cluster membership, although the cancer-type and transcriptomic-cluster partitions were not equivalent.

Cancer-type diversity also varied across individual clusters. Normalized Shannon entropy values were 0.108, 0.638, 0.597, 0.299, 0.624, and 0.736 for C0-C5, respectively. Thus, C0 and C3 showed relatively lineage-restricted compositions, whereas C1, C2, C4, and particularly C5 contained samples from a wider range of cancer types. Consistent with these quantitative results, the cancer type-by-cluster distribution and low-dimensional visualization showed that tissue lineage is an important determinant of the observed clustering structure. At the same time, the presence of multiple cancer types within several clusters indicates partial sharing of transcriptomic patterns across anatomical sites. The six groups are therefore interpreted as exploratory transcriptomic clusters influenced substantially by tissue lineage, with some evidence of shared expression patterns across cancer types, rather than as tissue-independent pan-cancer molecular states.

Independent analysis of the CCLE cohort identified a reproducible six-group transcriptomic partition. The six clusters contained 113, 172, 127, 257, 108, and 242 cell lines, respectively. The mean silhouette coefficient for k = 6 was 0.185. Although this value did not identify six clusters as the uniquely optimal solution, repeated clustering showed high stability, with a mean pairwise adjusted Rand index of 0.939 across 30 runs. Resampling-based consensus analysis produced a mean within-cluster consensus of 0.899 and a mean between-cluster consensus of 0.023. The proportion of ambiguous consensus values between 0.1 and 0.9 was 0.117.

Cluster assignment was strongly associated with cancer lineage among annotated CCLE cell lines, with a normalized mutual information value of 0.419 and a Cramér's V value of 0.741. Comparison of cancer-type composition across 23 categories shared by TCGA and CCLE showed partial concordance after one-to-one cluster matching, with a mean cosine similarity of 0.503. Two matched cluster pairs showed stronger similarities of 0.838 and 0.731, whereas the remaining matches were moderate or weak. These findings support the reproducibility of a stable six-group partition in an independent expression dataset but do not establish that the CCLE groups are identical to the TCGA-derived clusters.

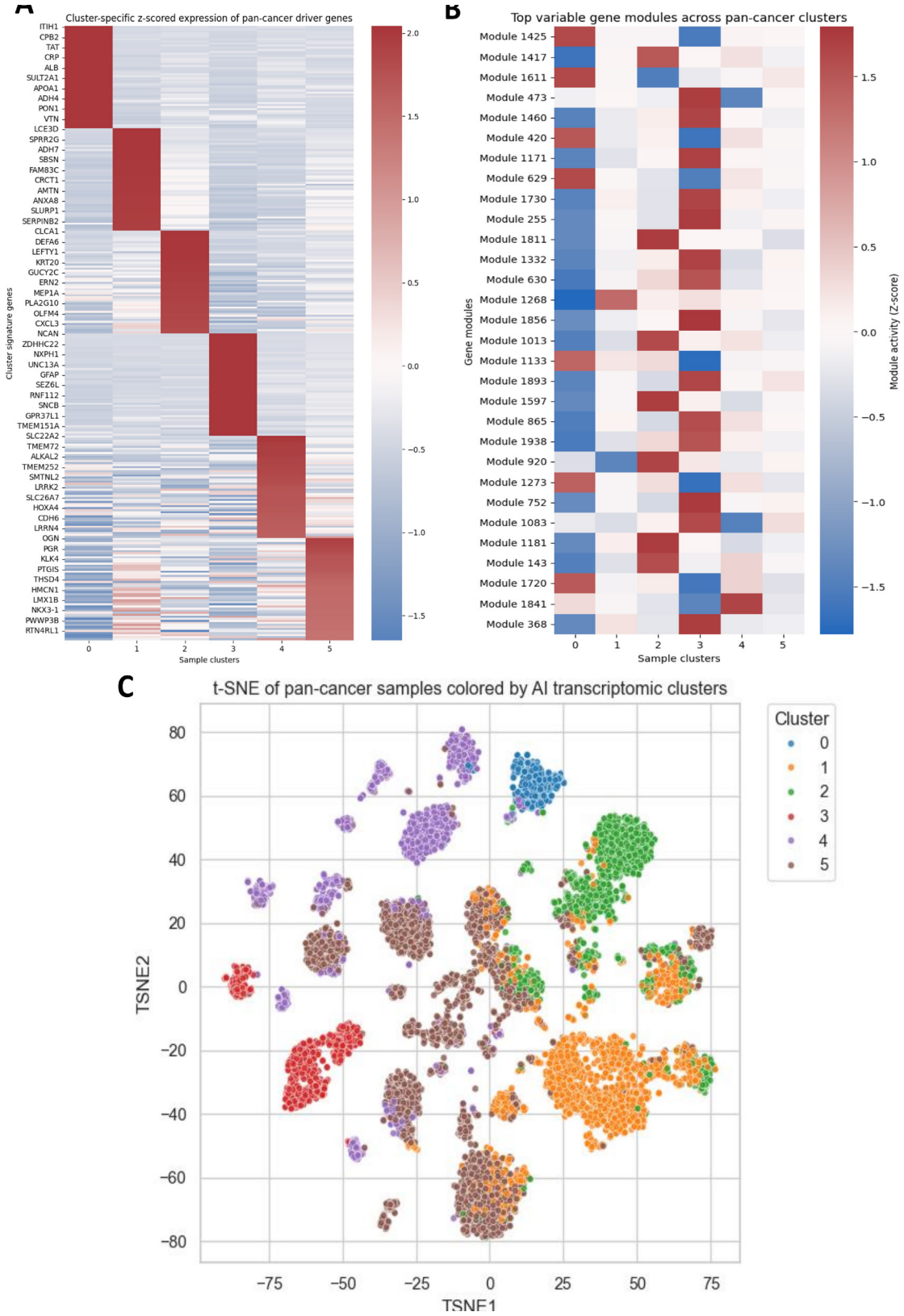


Figure 3. Driver-gene expression patterns and low-dimensional structure of pan-cancer transcriptomic clusters. (A) Heatmap of z-scored expression patterns for driver genes across the six transcriptomic clusters. (B) Heatmap showing descriptive expression blocks among selected variable genes across clusters. These patterns were used for descriptive visualization only and were not treated as formally defined gene groups. (C) t-SNE representation of pan-cancer samples colored according to transcriptomic cluster.

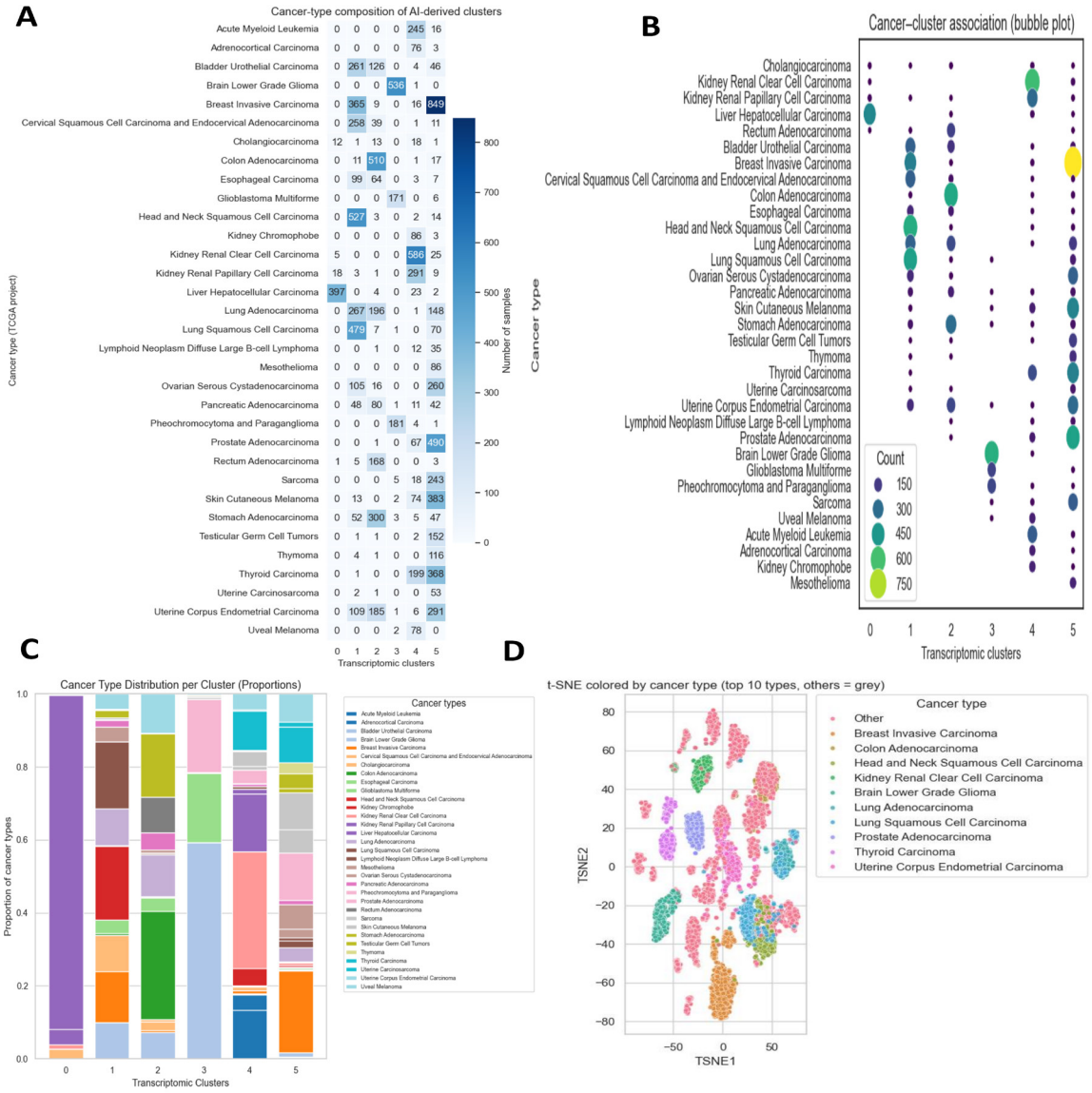


Figure 4. Cancer-type distribution across unsupervised pan-cancer transcriptomic clusters. (A) Heatmap showing the absolute distribution of 33 TCGA cancer types across six transcriptomic clusters. Several clusters exhibit broad multi-cancer representation, whereas others show tissue-skewed enrichment patterns. (B) Bubble plot illustrating cancer-cluster associations, with bubble size reflecting sample count. Distinct enrichment hotspots highlight preferential mapping of certain tumor types to specific clusters. (C) Proportional representation of cancer types within each cluster, illustrating differences in cancer-type heterogeneity and the presence of both lineage-enriched and more compositionally diverse clusters. (D) t-SNE projection of samples colored by cancer type, with the top 10 cancer types highlighted, illustrating the contribution of tissue lineage to the clustering structure together with partial overlap of cancer types across several exploratory transcriptomic clusters.

Each transcriptomic cluster is characterized by distinct biological programs.

To further resolve the biological identity of the six transcriptomic clusters, we performed pathway and Gene Ontology (GO) enrichment analysis using cluster-specific upregulated genes. The exploratory clusters showed different pathway-enrichment profiles, indicating variation in the biological programs associated with each cluster. Across KEGG pathways (Table 1), Cluster 0 was strongly enriched for immune-related signaling, including TLR4, NF-κB, and antigen-presentation pathways, suggesting an inflammation-associated transcriptional state. Cluster 1 showed dominant enrichment for epithelial polarity and adhesion programs, while Cluster

Table 1. Summary of pathway and Gene Ontology enrichment across the six transcriptomic clusters

Transcriptomic cluster	Top KEGG pathways	Top GO Biological Processes	Top GO Molecular Functions
C0	PPAR signaling pathway; NF-κB signaling pathway; Toll-like receptor signaling; Sphingolipid signaling pathway; Phospholipase D signaling pathway; cGMP-PKG signaling pathway; PI3K-Akt signaling pathway	Chylomicron remodeling; Negative regulation of fibrinolysis; Regulation of very-low-density lipoprotein-related processes; Chylomicron assembly; Fibrinolysis; Triglyceride-rich lipoprotein-related processes	Apolipoprotein receptor binding; Phosphatidylcholine-sterol O-acyltransferase-related activity; Lipoprotein particle receptor binding; Oxidoreductase activity; Peptidase inhibitor activity; Alcohol dehydrogenase activity; Endopeptidase regulator activity
C1	IL-17 signaling pathway; GnRH signaling pathway; Phosphatidylinositol signaling system; Glucagon signaling pathway; C-type lectin receptor signaling; Sphingolipid signaling pathway; Neurotrophin signaling pathway	Epidermis development; Keratinocyte differentiation; Regulation of steroid-related processes; Regulation of serine-type processes; Epidermal cell differentiation; Regulation of phosphatase activity	Peptidoglycan immune receptor activity; Peptidoglycan binding; Serine-type endopeptidase activity; DNA binding; Alcohol dehydrogenase-related activity; Sphingosine N-acyltransferase-related activity
C2	TGF-beta signaling pathway; IL-17 signaling pathway; NOD-like receptor signaling pathway; AGE-RAGE signaling pathway; TNF signaling pathway; Thyroid hormone signaling pathway; AMPK signaling pathway	Intracellular pH elevation; Type B pancreatic cell differentiation; Epithelial structure maintenance; Maintenance of gastrointestinal epithelium; Organic heteropentacyclic compound-related processes; Xenobiotic metabolic process; Oxidation-related metabolic processes	Bicarbonate transmembrane transporter activity; C-C chemokine receptor activity; All-trans retinol binding; Guanylate cyclase activator activity; Chloride transmembrane transporter activity; C-X-C chemokine receptor binding; Glycosyltransferase-related activity
C3	cAMP signaling pathway; Retrograde endocannabinoid signaling; Hippo signaling pathway; Estrogen signaling pathway; Adrenergic signaling; Oxytocin signaling pathway	Neuron fate specification; Dermatan sulfate biosynthetic process; Dermatan sulfate metabolic process; Chondroitin sulfate catabolic process; Neuron fate commitment; Dermatan sulfate proteoglycan-related process; Cerebral cortex neuron differentiation	Lysophosphatidic-acid-related activity; Ionotropic glutamate receptor activity; ATP-activated inward rectifier-related activity; Dehydropyrimidinase activity; Glutamate receptor activity; Cadherin light-chain binding; Cuprous ion binding
C4	AGE-RAGE signaling pathway; PI3K-Akt signaling pathway; Relaxin signaling pathway; Hedgehog signaling pathway; Adipocytokine signaling pathway; cAMP signaling pathway; NF-κB signaling pathway	Metanephric distal tubule development; Metanephric nephron tubule development; Metanephric epithelium development; Metanephric renal vesicle morphogenesis; Nephron tubule formation; Pronephros development; Mesenchymal-to-epithelial transition	Monoamine transmembrane transporter activity; Choline transmembrane transporter activity; L-arginine transmembrane transporter activity; Phosphodiesterase activity; GPCR-linked ephrin receptor activity; Histone serine kinase activity; Sulfur amino-acid transmembrane transporter activity
C5	PPAR signaling pathway; Estrogen signaling pathway; Wnt signaling pathway; Adipocytokine signaling pathway; Prolactin signaling pathway; NF-κB signaling pathway; Thyroid hormone signaling pathway	Negative regulation of insulin-related processes; Keratan sulfate catabolic process; Intracellular steroid hormone receptor signaling; Negative regulation of steroid-related processes; Steroid hormone-mediated signaling; Androgen receptor signaling; Cellular response to heparin	

2 was enriched for metabolic signaling and xenobiotic processing pathways. Cluster 3 exhibited signatures associated with extracellular matrix remodeling and cell–cell communication, consistent with mesenchymal or stromal-like molecular programs. Cluster 4 demonstrated enrichment for developmental and morphogen signaling pathways, including Wnt and Hedgehog pathways. Finally, Cluster 5 displayed strong enrichment for cell cycle regulation and mitotic progression, marking it as a proliferation-dominant molecular phenotype.

Gene Ontology biological processes further supported these distinctions: Cluster 0 was characterized by immune activation processes; RNA processing and translation were prominent in Cluster 1; lipid metabolism and redox activity defined Cluster 2; extracellular matrix organization marked Cluster 3; morphogenesis-related processes distinguished Cluster 4; and chromosome segregation and DNA replication enriched Cluster 5.

At the molecular function level, Clusters again showed highly specific biochemical identities: receptor binding and cytokine interaction in Cluster 0; nucleic acid–binding functions in Cluster 1; oxidoreductase activity in Cluster 2; integrin and cadherin binding in Cluster 3; transcription factor activity in Cluster 4; and enzymatic regulators of mitosis in Cluster 5.

Drug and Chemical Signature Enrichment Identifies Cluster-Associated Gene-Set Overlaps

To determine whether cluster-associated genes overlapped with compound- or chemical-annotated gene sets, we performed enrichment analysis using DSigDB. The resulting associations indicate gene-set overlap but do not establish drug sensitivity, resistance, therapeutic efficacy, or clinical response. Figure-6 shows selective overlaps between cluster-specific genes and compound-derived transcriptional signatures. These associations indicate shared transcriptional responses but do not establish pharmacologic sensitivity or therapeutic efficacy.

Cluster 0 showed significant enrichment for chemical signatures associated with oxidative and xenobiotic stress, including aflatoxin B1 and cyclosporin A. Aflatoxin B1 represents a toxicogenomic marker rather than a therapeutic candidate, and its enrichment is consistent with activation of detoxification and stress-response pathways in this cluster. Cluster 3 showed overlap with signatures associated with epigenetic modulators, including trichostatin A and MS-275. Cluster 4 was associated with hormonal and differentiation-related compounds, including estradiol and retinoic acid, whereas Cluster 5 showed enrichment for signatures linked to valproic acid and topiramate. These non-directional compound-signature overlaps provide hypotheses for future direction-resolved perturbation analysis. The enrichment patterns were non-uniform, with several compounds showing cluster-specific transcriptional overlaps. However, these computational associations should be interpreted as hypothesis-generating drug or chemical signatures rather than evidence of therapeutic profiles, clinical drug response, or treatment efficacy.

The direction of the relationship between each compound-associated signature and the corresponding cluster program could not be determined. A significant overlap may reflect transcriptional similarity, an opposing perturbation response, shared compound targets, or a common cellular stress program. Therefore, the enrichment results do not indicate whether a compound

would reproduce or reverse the cluster-associated transcriptional state.

Exploratory Pharmacogenomic Evaluation Using PRISM

PRISM measurements were available for 475 independently clustered CCLE cell lines. Cluster C1 was not represented in the secondary screen; the remaining coverage included 51 C0, 50 C2, 180 C3, 31 C4, and 163 C5 cell lines. A total of 1,446 compounds met the prespecified coverage requirement, of which 357 showed an AUC difference among clusters at FDR< 0.05. This result indicates substantial pharmacologic heterogeneity across the CCLE expression-defined groups, although the differences may partly reflect their tissue-lineage composition. Eight compounds discussed in the manuscript were identified in PRISM [14] . Cyclosporin A could not be evaluated against its matched cluster because CCLE Cluster C1 was absent from the PRISM-tested subset. Rifampin, trichostatin A, and entinostat showed significant overall AUC differences among the available clusters, with global FDR values of 0.013, 0.042, and 0.047, respectively; however, their lowest-AUC clusters did not correspond to the CCLE clusters matched to the original TCGA associations. Estradiol showed the clearest concordant result. The matched CCLE Cluster C3 had the lowest median estradiol AUC (1.177) compared with the combined remaining clusters (1.225; one-sided P=0.0036, targeted FDR = 0.025). The global five-cluster comparison for estradiol did not remain significant after correction across all PRISM compounds (FDR = 0.105). Simvastatin, dexamethasone acetate, and cyproterone acetate did not show significant cluster-wide AUC differences. These findings provide limited, compound-specific pharmacogenomic support but do not constitute broad validation of the DSigDB associations.

Cluster-Specific Chemical Perturbation Signatures Identified by DSigDB Enrichment

DSigDB enrichment analysis revealed that each transcriptomic cluster displayed a unique pattern of small-molecule and chemical perturbation signatures, reflecting the underlying biological programs responsible for cluster behavior. Cluster C0 was enriched for strong xenobiotic-response signatures, including aflatoxin B1 and benzo[a]pyrene, suggesting activation of DNA-damage and detoxification pathways. Clusters C1 and C2 showed enrichment for steroid-related and anti-inflammatory signatures such as estradiol, dexamethasone, and retinoids, consistent with transcriptional programs involving hormonal regulation and immune modulation. Cluster C3 displayed enrichment for neuromodulatory and metabolic agents, while clusters C4 and C5 were associated with diverse epigenetic modifiers and stress-response compounds. Together, these results show that each cluster is associated with a distinct set of chemical-induced transcriptional signatures. Toxicant-associated signatures primarily serve as markers of activated stress or detoxification pathways, whereas pharmacologic signatures provide hypotheses that require direct experimental testing.

Cancer-Type Composition Across the Six Transcriptomic Clusters

Bar plots depict the distribution of major cancer types within each transcriptomic cluster. Each subplot represents one cluster (C0-C5) and depicts the most abundant tumor types contributing to that cluster. C0 is dominated by

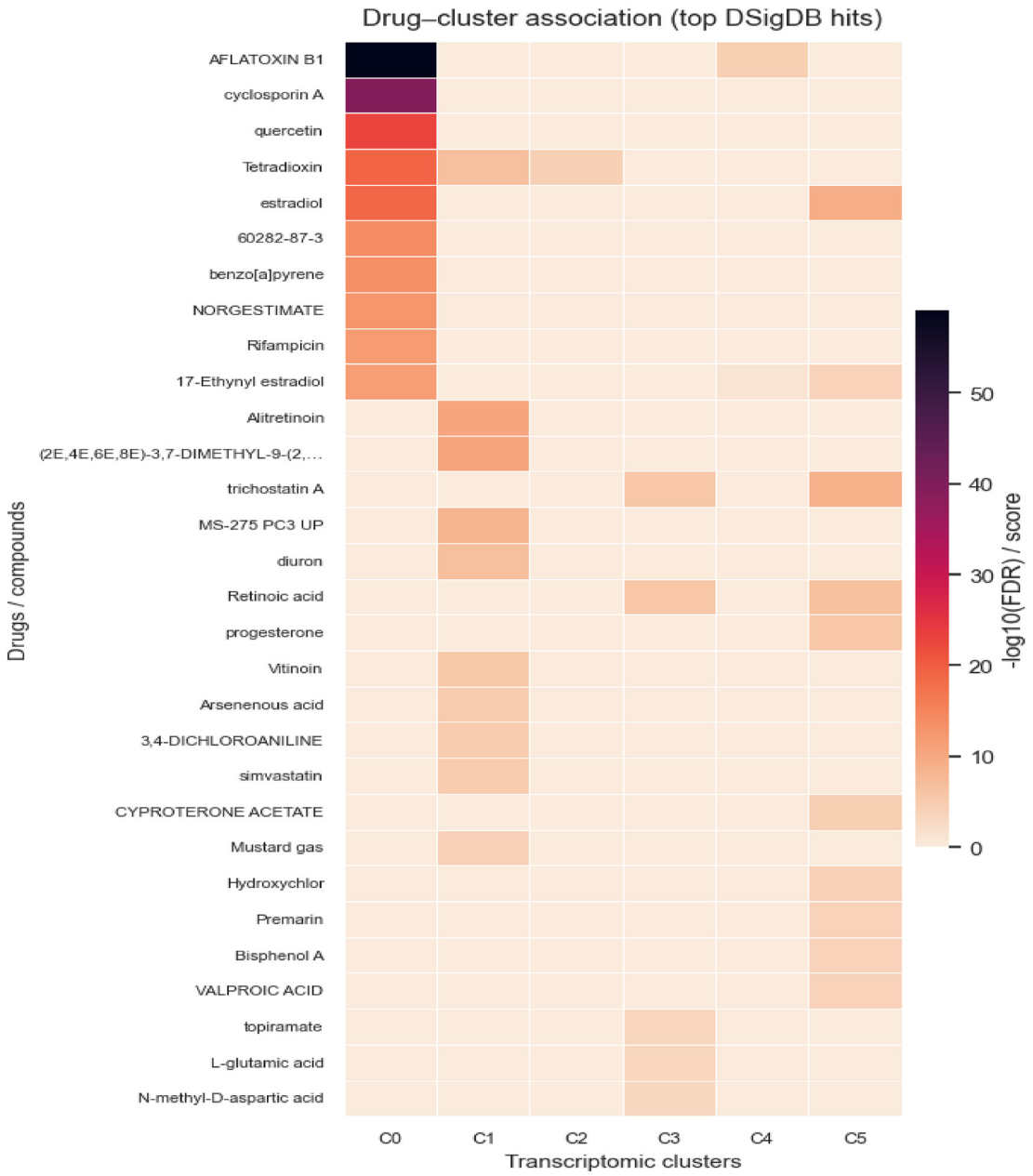


Figure 5. Cluster-associated compound-signature overlaps identified by DSigDB enrichment. Heatmap shows significant drug-gene signature enrichment scores (-log10 false-discovery rate) across the six transcriptomic clusters. Each row corresponds to a drug or chemical compound; each column represents a transcriptomic cluster (C0-C5). Several compounds exhibit strong cluster-specific associations, including aflatoxin B1 and cyclosporin A (Cluster 0), quercetin (Cluster 1), HDAC inhibitors such as trichostatin A and MS-275 (Cluster 3), hormonal agents such as estradiol and retinoic acid (Cluster 4), and signatures associated with compounds such as valproic acid (Cluster 5). Enrichment values represent non-directional overlaps between cluster-associated genes and drug- or chemical-associated gene sets. They do not indicate drug sensitivity, therapeutic efficacy, or clinical response. Aflatoxin B1, benzo[a]pyrene, mustard gas, and other toxicants are shown only as toxicogenomic markers. Enrichment scores represent non-directional gene-set overlaps and do not indicate whether a compound-associated signature mimics or reverses the cluster-associated transcriptional program.

hepatocellular carcinoma, whereas C1 and C2 are more heterogeneous, with strong contributions from head and neck, lung, breast, and gastrointestinal cancers. C3 is enriched for central nervous system tumors, including lower-grade glioma and glioblastoma. C4 consists mainly of kidney malignancies, while C5 contains a mixture of breast, prostate, melanoma, thymoma, and gynecologic cancers.

These profiles illustrate differences in cancer-type composition across the exploratory clusters and indicate that some tumors from different cancer types share similar transcriptional patterns (Supplementary Figure-S2).

Discussion

The integrative pan-cancer analysis identified an exploratory six-cluster transcriptomic partition characterized by differences in expression patterns, enriched pathways, and hypothesis-generating compound-signature overlaps. In the context of prior pan-cancer studies, these clusters appear to recapitulate and, in some cases, refine known lineage-dependent and lineage-agnostic oncogenic programs. Large-scale studies, including TCGA Pan-Cancer Atlas analyses, have sought to stratify tumor transcriptomes into broad categories such as immune-active, proliferative, or metabolic programs across histologies. The present clustering results align with this paradigm while highlighting narrower, biologically meaningful subdivisions that require further validation.

Cluster C0, which was strongly dominated by hepatocellular carcinoma samples, reveals a highly cohesive expression program centered on xenobiotic metabolism, lipid signaling, and chemical detoxification pathways. This finding has its roots in prior work detailing that HCC tumors display distinct transcriptomic signatures that are driven by liver-specific metabolic circuitry and nuclear receptors such as *HNF4A* [15]. Previous pan-cancer efforts have often noted that hepatocellular carcinoma forms a relatively isolated node in transcriptomic space; our results reaffirm that independence and, through enrichment of cytochrome P450-linked signaling, extend prior findings suggesting the metabolic plasticity of HCC may shape its response to environmental toxicants [16,17]. Notably, DSigDB drug associations returned in this cluster included compounds like aflatoxin B1-not as therapeutic candidates but rather as known inducers of transcriptional stress programs. This is consistent with toxicogenomic studies indicating that aflatoxin-induced gene signatures are enriched in HCC tissue as a result of chronic exposure-related reprogramming [18].

Cluster C1 was more heterogeneous in its cancer-type composition, containing both breast, lung, head-and-neck, and bladder tumors. The enrichment of immune-related pathways and drug signatures, including rifampicin and simvastatin, implies that this group may reflect tumors with heightened inflammatory or stress-response programs [19]. This notion echoes from the "immune-hot" clusters within multi-omic analyses where numerous epithelial tumors show a convergence upon a shared interferon-γ axis regardless of tissue origin. Previous reports have suggested that simvastatin-mediated transcriptional programs intersect with immune-modulatory signatures in both cancer cells and macrophages; the overlap observed here is consistent with these previously reported transcriptional associations and provides a hypothesis for further perturbation-based investigation.

Cluster C2 was enriched for hormone-associated signaling, particularly that involving retinoid and glucocorticoid pathways. Indeed, various reports have described subsets of gastrointestinal and gynecologic tumors assuming a "differentiation-linked" transcriptional state in response to retinoic acid or corticosteroids. This cluster's association with dexamethasone-like signatures is consistent with reports of glucocorticoid-

associated transcriptional programs across diverse tumor types. This pattern of hormone-responsive clusters has been described in endometrial and colorectal cancers, reinforcing the notion of a pan-tissue endocrine-tuned gene expression program [20,21].

Cluster C3 was dominated by gliomas and exhibited a well-defined CNS-specific regulatory state. The enrichment of glutamatergic signaling components, kainate receptor-related compounds, and excitatory synapse pathways closely parallels known glioma biology. The presence of such neuronal-like transcriptional programs as a basis for clustering in this analysis is concordant with integrative studies of glioma subtyping [22,23]. The emergence of small-molecule signatures corresponding to kainic acid or N-methyl-D-aspartate is also consistent with toxicogenomic datasets where signs of neuronal dysfunction emerge in both glial tumors and experimental models of neurotoxicity [24,25].

Cluster C4 was primarily made up of renal malignancies. Past TCGA and CPTAC proteogenomic studies have found that renal clear-cell and papillary carcinomas cluster tightly due to shared metabolic hallmarks, including oxidative phosphorylation and lipid metabolism [26,27]. Our results capture these same features, with pathway enrichment pointing toward metabolic reprogramming and transport-related signaling. The appearance of estrogenic signatures in this cluster aligns with recent reports that renal tumors can activate steroid metabolic pathways despite not being classically hormone driven.

Cluster C5 included a diverse but biologically coherent mixture of breast, prostate, thymoma, melanoma, and gynecologic cancers. Pathway enrichment emphasized cell-cycle regulation, DNA repair, and steroid receptor signaling-patterns that overlap with proliferation-high tumors in prior pan-cancer analyses. The DSigDB associations highlight compounds such as valproic acid, hydroxychloroquine, and cyproterone acetate, which have been linked in independent studies to modulation of epigenetic and hormone signaling programs [28,29]. The presence of breast-specific estrogenic signatures and prostate-specific androgenic patterns suggests that this cluster represents tumors with broadly endocrine-responsive transcriptional states, independent of their tissue of origin.

Integration of DSigDB signatures identified overlaps between cluster-specific genes and transcriptional responses induced by drugs, environmental chemicals, and toxicants. These results should not be interpreted as direct evidence of drug sensitivity or therapeutic efficacy. Compounds such as aflatoxin B1, benzo[a]pyrene, and mustard gas are toxicogenomic markers that reflect activation of DNA-damage, detoxification, or cellular stress pathways and are not therapeutic candidates. Pharmacologically relevant compounds, including retinoic acid, quercetin, simvastatin, and valproic acid, are presented separately as hypothesis-generating associations. Their potential relevance requires validation using drug-response datasets and experimental models before any therapeutic conclusions can be drawn.

The use of FPKM expression values imposes methodological limitations. Although FPKM data are suitable for exploratory clustering, dimensionality reduction, and visualization, they are less optimal than raw counts for variance modeling and formal differential-expression analysis and may retain

technical or compositional variation. In addition, DSigDB enrichment was based on non-directional gene-set overlap and did not determine whether compound-associated signatures mimic or reverse cluster-associated transcriptional programs. These associations therefore do not demonstrate drug sensitivity, therapeutic efficacy, or repurposing potential.

Finally, PRISM was used only as an exploratory, indirect, compound-specific pharmacogenomic assessment linked to the independently derived CCLE clusters and should not be considered definitive validation of the TCGA-derived DSigDB associations. PRISM coverage was incomplete, one CCLE cluster was not represented, several DSigDB compounds were unavailable or could not be matched exactly, and concordant evidence was observed for only a limited subset of compounds. Therefore, the PRISM findings provide limited compound-specific pharmacogenomic support rather than validation of drug sensitivity, therapeutic efficacy, or clinical response. Experimental validation in appropriate cellular, organoid, or in vivo models remains necessary.

This study provides a discovery-stage characterization of transcriptomic heterogeneity across TCGA cancer types and identifies six transcriptomic clusters associated with distinct biological and compound-related gene signatures. The clustering structure reflects both shared expression patterns and tissue-lineage effects. DSigDB enrichment identifies non-directional overlaps with drug- and chemical-associated signatures and should be interpreted as hypothesis-generating rather than as evidence of drug sensitivity or therapeutic efficacy. Independent transcriptomic validation, clinical-outcome integration, direction-resolved perturbation analysis, and experimental or pharmacogenomic testing will be required to establish the reproducibility, biological significance, and potential clinical relevance of these findings. The DSigDB analysis did not distinguish compounds whose transcriptional effects might reproduce the cluster-associated program from those that might oppose it. Consequently, no compound can be interpreted as a candidate reversal agent on the basis of the present enrichment analysis.

Limitations

Several limitations should be considered when interpreting the present findings. First, tissue of origin contributed substantially to the clustering structure, and several clusters were enriched for specific cancer lineages. Although cluster-cancer-type associations were quantified and considered during interpretation, residual lineage-related confounding remains. The identified clusters should therefore not be regarded as fully tissue-independent molecular states.

Second, independent transcriptomic replication was performed using CCLE expression data; however, this analysis does not constitute direct validation of the TCGA clusters. CCLE consists of cancer cell lines rather than primary tumors, and the original TCGA PCA loadings and cluster centroids were not available for direct projection of CCLE samples. The correspondence between TCGA and CCLE clusters was therefore assessed indirectly using cancer-type composition and should be interpreted cautiously.

Third, clinical outcomes, including overall survival, progression-free survival, tumor stage, grade, treatment history, and patient-level therapeutic response, were not incorporated into the analysis. Consequently, the identified transcriptomic clusters cannot currently be interpreted as prognostic, predictive, or

clinically validated molecular subtypes.
Fourth, the TCGA analysis was based on FPKM expression values. Although FPKM data are suitable for exploratory clustering, dimensionality reduction, and visualization, they are less appropriate than raw counts for formal variance modeling and differential-expression inference and may retain technical or compositional variation.
Fifth, DSigDB enrichment was based on non-directional gene-set overlap. The analysis did not determine whether compound-associated signatures were concordant with or opposed to the cluster-associated expression programs, and no connectivity or reversal score was calculated. The resulting associations therefore do not establish whether a compound would mimic or reverse a transcriptional state.
Sixth, exploratory pharmacogenomic assessment was performed using PRISM dose-response AUC data linked to the independent CCLE clusters. However, PRISM coverage was incomplete, one CCLE cluster was not represented, several DSigDB compounds were unavailable or could not be matched exactly, and concordant evidence was observed only for a limited subset of compounds. These findings provide partial, compound-specific support rather than definitive validation of drug sensitivity, therapeutic efficacy, or clinical response. Experimental validation in appropriate cellular, organoid, or in vivo models remains necessary.
Finally, public reproducibility remains limited. The primary TCGA, CCLE, DepMap, and PRISM datasets are publicly accessible, and the principal processed summaries are provided in the Supplementary Materials. However, the complete processed expression matrices and fully documented analysis scripts have not yet been deposited in a public version-controlled repository. The remaining outputs and available scripts are obtainable from the corresponding author upon reasonable request.

Conclusion

This study provides a discovery-stage characterization of transcriptomic heterogeneity across multiple TCGA cancer types. The analysis identified six expression-defined clusters associated with distinct gene-expression patterns, biological pathways, and compound-signature overlaps. Quantitative cancer-type composition analysis indicated that the clustering structure reflects both shared transcriptional patterns and substantial tissue-lineage effects.
Independent analysis of CCLE expression profiles provided indirect evidence that broadly comparable transcriptomic groupings can be observed in a separate cancer model system. In addition, exploratory PRISM analysis identified cluster-associated differences in compound-response AUC and limited compound-specific concordance with selected DSigDB associations. However, these findings do not establish clinically validated tumor subtypes, patient-level treatment response, therapeutic efficacy, or immediate translational utility.
Further validation in independent primary-tumor transcriptomic cohorts, integration of clinical outcomes and treatment information, direction-resolved perturbation analysis, and additional pharmacogenomic and experimental testing will be required before the biological, clinical, or therapeutic relevance of these findings can be established.

Conflicts of Interest

The authors declare that they have no competing interests.

Declaration of Generative AI Use

Generative AI was used solely for language editing and assistance in drafting responses to reviewers. It was not used for data generation, data analysis, statistical testing, interpretation of numerical results, or figure preparation. All revisions were critically reviewed and approved by the authors, who take full responsibility for the final content.

Acknowledgements

The authors gratefully acknowledge the developers of open-source scientific tools that made this work possible, including Python, NumPy, pandas, PyTorch, scikit-learn, Matplotlib, Seaborn, and Plotly. We also thank the TCGA consortium for generating and maintaining high-quality multi-cancer molecular datasets that serve as an invaluable resource for systems oncology research.

References

1. Yuan S, Almagro J, Fuchs E. Beyond genetics: driving cancer with the tumour microenvironment behind the wheel. Nat Rev Cancer. 2024;24(4):274-86.
2. Jacquemin V, Antoine M, Dom G, Detours V, Maenhaut C, Dumont JE. Dynamic cancer cell heterogeneity: diagnostic and therapeutic implications. Cancers (Basel). 2022;14(2):280.
3. Bansal A, Goyal S, Goyal A, Jana M. WHO classification of soft tissue tumours 2020: an update and simplified approach for radiologists. Eur J Radiol. 2021;143:109937.
4. Kallen ME, Hornick JL. The 2020 WHO classification: what's new in soft tissue tumor pathology? Am J Surg Pathol. 2021;45(1):e1-e23.
5. Gondal MN, Chaudhary SU. Navigating multi-scale cancer systems biology towards model-driven clinical oncology and its applications in personalized therapeutics. Front Oncol. 2021;11:712505.
6. Li J, Wang D, Tang F, Ling X, Zhang W, Zhang Z. Pan-cancer integrative analyses dissect the remodeling of endothelial cells in human cancers. Natl Sci Rev. 2024;11(9):nwae231.
7. Creighton CJ. Gene expression profiles in cancers and their therapeutic implications. Cancer J. 2023;29(1):9-14.
8. Cao J, Gong J, Li X, Hu Z, Xu Y, Shi H, et al. Unsupervised hierarchical clustering identifies immune gene subtypes in gastric cancer. Front Pharmacol. 2021;12:692454.
9. Zhu L, Miao Y, Xi F, Jiang P, Xiao L, Jin X, et al. Identification of potential biomarkers for pan-cancer diagnosis and prognosis through the integration of large-scale transcriptomic data. Front Pharmacol. 2022;13:870660.
10. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Onkol. 2015;2015(1):68-77.
11. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V, Lo CC, McDonald III ER, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature. 2019;569(7757):503-8.
12. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015;31(18):3069-71.
13. Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90-W97.
14. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nat Cancer. 2020;1(2):235-48.
15. Teeli AS, Łuczyńska K, Haque E, Gayas MA, Winiarczyk D, Taniguchi H. Disruption of tumor suppressors HNF4α/HNF1α causes tumorigenesis in liver. Cancers (Basel). 2021;13(21):5357.
16. Song F, Wang CG, Mao JZ, Wang TL,

Liang XL, Hu CW, et al. PANoptosis-based molecular subtyping and HPAN-index predicts therapeutic response and survival in hepatocellular carcinoma. Front Immunol. 2023;14:1197152.

17. Cao L, Li F, Cai S, Zhang J, Guo C, Ali S, et al. Pan-cancer analysis and the oncogenic role of Glypican 1 in hepatocellular carcinoma. Sci Rep. 2024;14(1):15870.
18. Zhang BA, Miao B, Fu S, Coarfa C, Kumar R, Kuntala PK, et al. Toxicogenomic insights into environmental toxicant exposures: the TaRGET II resource. Research Square. 2025 Aug 20:rs-3.
19. Sernoskie SC. Characterization of the early immune response to clozapine: relevance for idiosyncratic drug-induced agranulocytosis [dissertation]. Toronto, Canada: University of Toronto; 2023.
20. Dasgupta S, Das A, Bandyopadhyay U. Clinicopathological correlation of hormone receptors, angiogenesis, and tumor budding in endometrial carcinoma: a tertiary care center study. J Obstet Gynecol India. 2025 Nov 3:1-7.
21. Kadry MO, Serag A, Ammar NM. Metabolomics reveals the therapeutic efficacy of liposomal resvida in endometrial cancer through regulating autophagy-related gene expression. Cancer Cell Int. 2025;25(1):400.
22. Tian RY. The landscape and role of structural variants in gliomas: integrative analysis with linked-read whole genome sequencing and single-cell transcriptomics [master's thesis]. Toronto, Canada: University of Toronto; 2024.
23. Cihan M, Schmauck G, Sprang M, Andrade-Navarro MA. Unveiling cell-type-specific microRNA networks through alternative polyadenylation in glioblastoma. BMC Biol. 2025;23(1):15.
24. Klima S. Developmental neurotoxicity in a human model system [dissertation]. Konstanz, Germany: Universität Konstanz; 2021.
25. Emon MAEK. Mechanism-based stratification of Alzheimer's and Parkinson's disease using artificial intelligence [dissertation]. Bonn, Germany: Rheinische Friedrich-Wilhelms-Universität; 2021.
26. Czyzyk-Krzeska MF, Landero Figueroa JA, Gulati S, Cunningham JT, Meller J, ShamsaeI B, et al. Molecular and metabolic subtypes in sporadic and inherited clear cell renal cell carcinoma. Genes (Basel). 2021;12(3):388.
27. Qu Y, Feng J, Wu X, Bai L, Xu W, Zhu L, et al. A proteogenomic analysis of clear cell renal cell carcinoma in a Chinese population. Nat Commun. 2022;13(1):2052.
28. Kringel D, Malkusch S, Lötsch J. Drugs and epigenetic molecular functions. A pharmacological data scientometric analysis. Int J Mol Sci. 2021;22(14):7250.
29. Sarwar S, Alamro A, Huq F, Alghamdi A. Insights into the role of epigenetic factors determining the estrogen response in estrogen-positive ovarian cancer and prospects of combining epi-drugs with endocrine therapy. Front Genet. 2022;13:812077.