scientific reports

OPEN

Pan-cancer gene set discovery via scRNA-seq for optimal deep learning based downstream tasks

Jong Hyun Kim, Soonyoung Lee & Jongseong Jang

The application of machine learning to transcriptomics data has led to significant advances in cancer research. However, the high dimensionality and complexity of RNA sequencing (RNA-seq) data pose significant challenges in pan-cancer studies. This study hypothesizes that gene sets derived from single-cell RNA sequencing (scRNA-seq) data will outperform those selected using bulk RNA-seq in pan-cancer downstream tasks. We analyzed scRNA-seq data from 181 tumor biopsies across 13 cancer types. High-dimensional weighted gene co-expression network analysis (hdWGCNA) was performed to identify relevant gene sets, which were further refined using XGBoost for feature selection. These gene sets were applied to downstream tasks using TCGA pan-cancer RNA-seq data and compared to six reference gene sets and oncogenes from OncoKB evaluated with deep learning models, including multilayer perceptrons (MLPs) and graph neural networks (GNNs). The XGBoost-refined hdWGCNA gene set demonstrated higher performance in most tasks, including tumor mutation burden assessment, microsatellite instability classification, mutation prediction, cancer subtyping, and grading. In particular, genes such as *DPM1*, *BAD*, and *FKBP4* emerged as important pan-cancer biomarkers, with *DPM1* consistently significant across tasks. This study presents a robust approach for feature selection in cancer genomics by integrating scRNA-seq data and advanced analysis techniques, offering a promising avenue for improving predictive accuracy in cancer research.

Recent advancements in oncology have leveraged machine learning techniques to enhance the analysis of transcriptomic data across various cancer types[1,2]. Integrating transcriptomic data with advanced algorithms has enhanced accuracy of downstream tasks such as mutation prediction, cancer subtyping and survival prediction[3–5]. In particular, models like gradient boosting and neural network architectures, including multilayer perceptrons (MLPs) and graph neural networks (GNNs), have demonstrated their capability in extracting meaningful patterns from complex genomic datasets[6,7]. These approaches enable more precise predictions and provide new clinical insights, supporting the development of targeted therapies and personalized cancer therapies.

RNA sequencing (RNA-seq) is a powerful technique in cancer research, providing profound insights into the genetic landscape of tumors. However, the high dimensionality and complexity of RNA-seq data present significant analytical challenges, requiring robust feature selection methods to identify biologically relevant genes[8]. Feature selection is critical for both biological interpretation and predictive model performance, particularly in deep learning applications[9]. Although various feature selection methods have been developed for RNA-seq data in cancer studies, no generalized method performs optimally across all studies. For instance, Chen et al.[10], utilized pathway-based gene sets relevant to cancer from the MsigDB hallmark database. While Elbashir et al.[11], focused on differentially expressed genes, and Silva et al.[12], selected genes with the highest variance across samples. This diversity underscores the lack of standardization in feature selection, making it challenging to identify consistent patterns across studies. This challenge is particularly notable in pan-cancer studies, where identifying commonalities across cancer types is essential.

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of the tumor microenvironment by revealing the cellular diversity and dynamics often obscured in bulk RNA sequencing data[13]. It provides detailed views of immune cell heterogeneity and tumor-immune interactions that strongly influence cancer progression and treatment response[14]. Recent advances in scRNA-seq have illuminated specific cellular mechanisms, enhancing our understanding of the immune landscape across multiple cancers[15–17]. These insights enable high-resolution biomarker discovery and predictive modeling. By capturing gene expression at the cellular level, scRNA-seq allows feature selection enriched with detailed biological context, in contrast to bulk RNA-based approach. Despite these advances in scRNA-seq technology, most current pan-cancer studies still rely on bulk RNA-seq data or predefined pathway gene sets, overlooking the opportunity to leverage single-cell heterogeneity for improved predictive modeling.

LG AI Research, Seoul, South Korea. email: j.jang@lgresearch.ai

In this study, we propose a novel framework that integrated high dimensional weighted gene co-expression network analysis (hdWGCNA) from pan-cancer scRNA-seq data with XGBoost-based feature refinement. To our knowledge, this represents the first application of this combined approach in pan-cancer feature selection. We hypothesize that gene sets derived from scRNA-seq data outperform those from bulk RNA-seq in predictive modeling and downstream cancer tasks. To demonstrate this hypothesis, we analyzed scRNA-seq data from 181 tumor biopsies across 13 cancer types, identifying co-expression modules through hdWGCNA and subsequently refining them using XGBoost for task-specific feature selection. The resulting gene sets were evaluated on TCGA pan-cancer RNA-seq data across multiple downstream tasks. We compared our approach against six established reference gene sets and OncoKB curated oncogenes using two different deep learning models. Our framework demonstrated improved predictive accuracy and also identified biologically meaningful pan-cancer biomarkers, offering a robust strategy for gene selection in cancer genomics.

Results

Overview of study

Our study workflow, as illustrated in Fig.1, we conducted a comprehensive analysis of scRNA-seq data from 181 tumor biopsy samples, encompassing 13 distinct cancer types (Fig.1a). This dataset represents 13 tumor types, with their proportional distribution shown in Fig. 1b. From the scRNA-seq data, we applied hdWGCNA to construct a gene co-expression network (Fig. 1c, d). This analysis allowed us to identify clusters of co-expressed genes with shared biological functions. These co-expressed genes served as gene sets for downstream tasks. We selected genes participating in the co-expression network and excluded those without network connectivity**.** These selected genes were then applied to feature selection processes of RNA-seq data for downstream modeling tasks, as shown in Figure 1e (Supplementary Table 1).

We further refined these features using XGBoost to select the most relevant genes for each downstream task. Downstream tasks performed using TCGA pan-cancer RNA-seq data included tumor mutation burden (TMB) assessment, microsatellite instability (MSI) classification, mutation prediction, cancer subtyping, and grading. Our approach integrates single-cell data with advanced machine learning techniques to improve predictive modeling and facilitate the development of targeted therapies.

Combining high-resolution scRNA-seq data with advanced analytical methods refines biomarker selection and enhances model accuracy. These results highlight the importance of advanced feature selection methods combined with powerful analytical tools in advancing cancer research.

High-dimensional co-expression network through scRNA-seq

We generated a scRNA-seq dataset from 181 tumor biopsy samples, including 87,659 genes and 317,111 cells[18]. This dataset captures distinct immune cell subsets and provides a comprehensive view of the tumor microenvironment. UMAP visualization (Fig. 2a) illustrates the cellular diversity and distribution of 25 immune cell types across 181 samples.

For hdWGCNA preprocessing, genes expressed in fewer than 5% of cells were excluded, yielding 6,617 common genes across all cancer types. This filtering step reduced noise and ensured that downstream analyses focused on consistently expressed genes. This processed resulted in 11 modules containing 1,857 genes (Fig. 2b). Each module is represented by the top 25 hub genes with the highest interconnectivity, ranked by kME (Fig.2c), and all module-associated genes are listed in Supplementary Table 2.

To assess the biological significance of the identified modules, we conducted a gene ontology (GO) enrichment analysis. The heatmap in Fig. 2d summarizes the enrichment results, highlighting major biological processes associated with each module. Module 1 is involved in immune and apoptotic processes, which regulate tumor immunity and cell death across cancer types. Module 2 focuses on mitochondrial energy metabolism, including cellular respiration and oxidative phosphorylation, which support the high metabolic demands of proliferating tumor cells. Module 4 is involved in ribosome biogenesis and protein synthesis, including translation and gene expression, which are fundamental to cancer cell growth and proliferation. Module 5 focuses on protein folding and stability, such as the cellular response to unfolded proteins and protein stabilization. Module 6 involves vascular and immune cell migration, such as endothelial cell movement and actin filament formation, which are crucial for tumor angiogenesis and metastasis. Module 7 focuses on cell adhesion and phagocytosis, involving leukocyte adhesion and plasma membrane invagination, highlighting the tumor microenvironment's role in cancer immunity and inflammation. Module 8 is linked to T cell activation and cytokine responses, including T cell activation pathways and interleukin signaling, emphasizing the importance of adaptive immunity in tumor recognition and elimination. Module 9 covers vesicle transport, such as golgi vesicle transport and endoplasmic reticulum to golgi transport, vital for protein trafficking and secretion in cancer cells. Module 10 involves stress response and growth factor signaling, including responses to fibroblast growth factors and interferon-beta, reflecting adaptive mechanisms for cancer cell survival. Finally, module 11 is associated with transcriptional regulation and RNA processing, including mRNA splicing, which is essential for gene expression regulation and tumor adaptability under stress.

Comparative performance analysis of gene sets for downstream tasks

After identifying gene modules through hdWGCNA, we applied XGBoost-based feature selection to refine these modules for specific downstream tasks (Supplementary Table 3). We evaluated feature importance and retained genes with scores higher than 0.001 for each task. These tasks included TMB assessment, MSI classification, mutation prediction for *TP53*, *EGFR*, and *KRAS* genes, cancer subtyping, and grading. For validation, we compared our refined gene sets against six reference gene sets and the OncoKB curated oncogene set (Supplementary Table 4 and 5)[9,19–24]. We performed stratified 5-fold cross-validation to ensure robust performance evaluation. Results from each fold were averaged and bootstrapped to calculate 95% confidence intervals (CI) for the mean

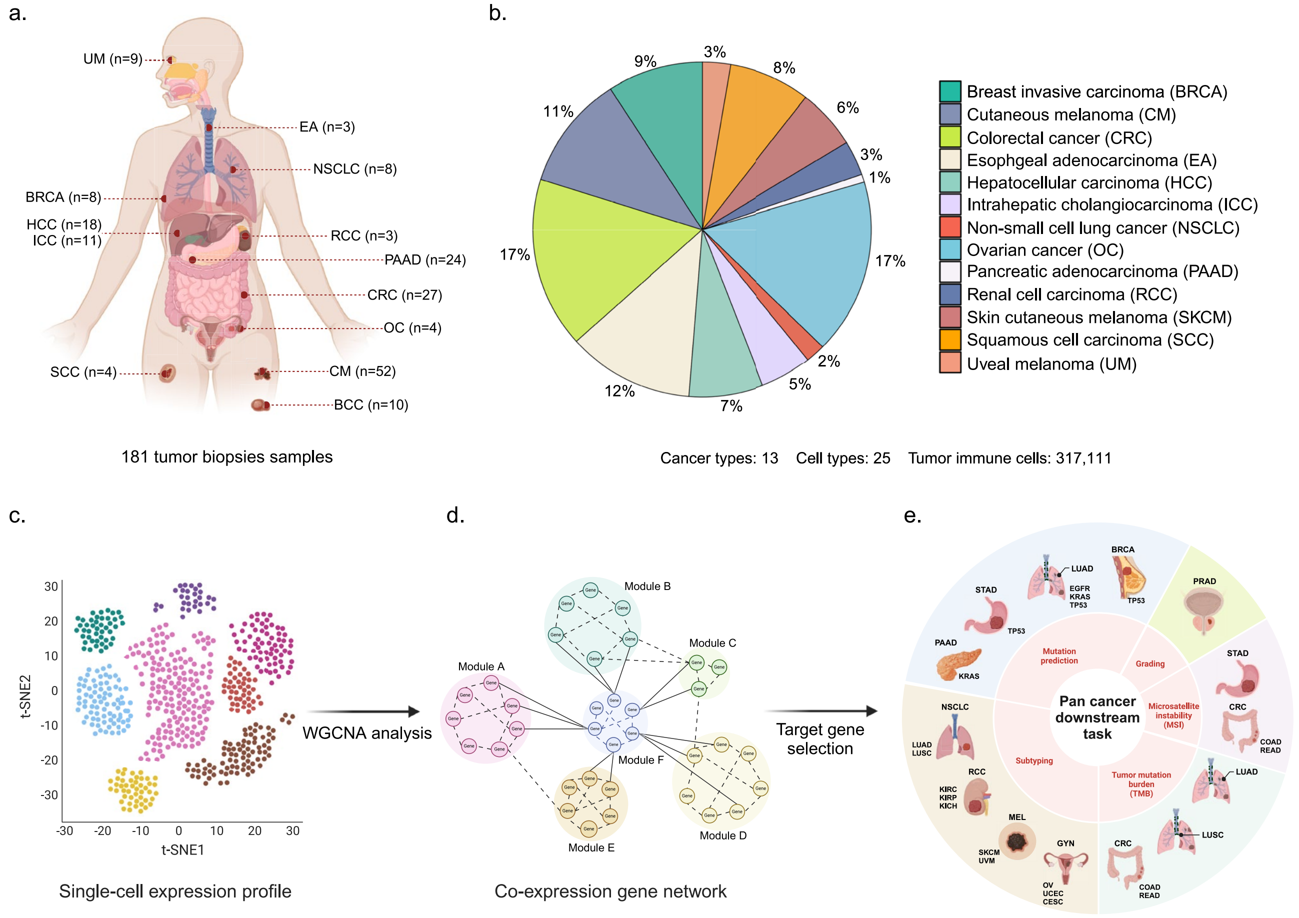

Fig. 1. Schematic overview of the study workflow. **a**. Each cancer type is indicated along with the number of samples analyzed: Uveal melanoma (UM, n=9), esophageal adenocarcinoma (EA, n=3), non-small cell lung cancer (NSCLC, n=8), breast invasive carcinoma (BRCA, n=8), hepatocellular carcinoma (HCC, n=18), intrahepatic cholangiocarcinoma (ICC, n=11), renal cell carcinoma (RCC, n=3), pancreatic adenocarcinoma (PAAD, n=24), colorectal cancer (CRC, n=27), ovarian cancer (OC, n=4), squamous cell carcinoma (SCC, n=4), cutaneous melanoma (CM, n=52), basal cell carcinoma (BCC, n=10). **b**. Proportional distribution of the 13 cancer types within the scRNA-seq dataset, highlighting the diversity and representation of different tumor types used in the study. The dataset includes 317,111 tumor immune cells classified into 25 distinct cell types. **c**. Schematic UMAP visualization of single-cell expression profiles, which were prepared for hdWGCNA. **d**. Schematic of the co-expression gene network derived from hdWGCNA. The genes comprising this biological module were used for various downstream analyses through the target gene process. **e**. (e) Application of selected co-expressed gene modules to various pan-cancer downstream tasks. These tasks include mutation prediction, cancer subtyping, tumor mutation burden (TMB) assessment, microsatellite instability (MSI) classification, cancer subtyping, and grading.

area under the receiver operating characteristic curve (AUROC). Model performance was evaluated using two different models: MLP and GNN (see more details in the method).

Overall, the XGBoost-refined hdWGCNA gene sets outperformed the reference gene sets across downstream tasks, except for the BRCA-*TP53* mutation prediction task (Fig. 3, Supplementary Table 6). To access statistical significance, we performed pairwise z-tests with Bonferroni correction, showing that our method achieved statistically significant gains ($p < 0.005$) in 88% of all comparisons across downstream tasks (Supplementary Table 7)

The results are as follows:

Tumor mutation burden (TMB): As shown in Fig.3a, our refined gene sets yielded consistently high performance. Specifically, for lung adenocarcinoma (LUAD), the average AUROC was 0.791 (95% CI: 0.782–0.800.782.800) with the MLP model and 0.764 (95% CI: 0.754–0.773.754.773) with the GNN model. In the case of lung squamous cell carcinoma (LUSC), the MLP model achieved an average AUROC of 0.647 (95% CI: 0.636–0.659.636.659), while the GNN model achieved 0.604 (95% CI: 0.592–0.616.592.616). For skin cutaneous melanoma (SKCM), the average AUROC was 0.772 (95% CI: 0.762–0.782.762.782) with the MLP model and 0.747 (95% CI: 0.736–0.758.736.758) with the GNN model. Lastly, for colorectal cancer (CRC), the average AUROC was 0.826 (95% CI: 0.818–0.835.818.835) with the MLP model and 0.808 (95% CI: 0.799–0.817.799.817) with the GNN model.

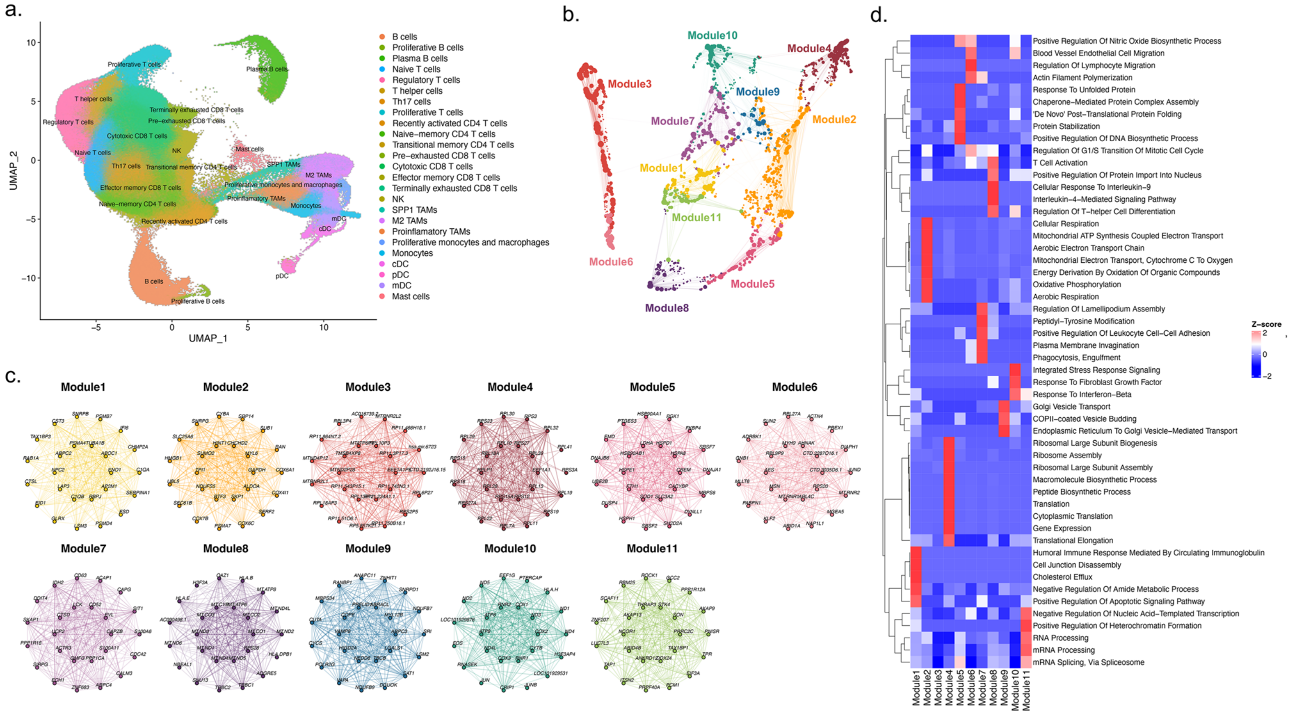

Fig. 2. Comprehensive gene co-expression networks reveal key functional modules in the tumor immune microenvironment. **a**. UMAP embeddings display the distribution of 25 immune cell types across 181 tumor biopsy samples, visualizing all 317,111 cells. Each color represents a distinct immune cell type. **b**. UMAP representation of the gene co-expression network. A total of 11 modules were identified. Nodes indicate individual genes, and edges signify co-expression relationships between genes and hub genes within modules. Node sizes are proportional to their kME (eigengene-based connectivity) values. Colors denote different co-expression modules. **c**. Visualization of hub gene networks for each spatial co-expression module. The 25 highest-ranked hub genes based on kME are presented. In the network, nodes represent genes, while edges indicate co-expression links. **d**. Heatmap summarizing GO pathway enrichment analysis for each module. Each row represents a biological process, and columns correspond to modules, with the color scale indicating Z-score values.

Microsatellite instability (MSI): For stomach adenocarcinoma (STAD), our refined gene sets achieved the highest AUROC with the MLP model, 0.990 (95% CI: 0.988–0.991.988.991). However, with the GNN model, the OncoKB gene set slightly outperformed ours, achieving an AUROC of 0.982 (95% CI: 0.979–0.985.979.985). For CRC, the average AUROC was 0.931 (95% CI: 0.925–0.937.925.937) with the MLP model and 0.936 (95% CI: 0.930–0.942.930.942) with the GNN model, both outperforming the reference gene sets.

Mutation prediction (MUT): In mutation prediction tasks, the OncoKB gene set achieved the highest performance for BRCA-*TP53*, with an average AUROC of 0.951 (95% CI: 0.949–0.954.949.954) using the MLP model, and 0.935 (95% CI: 0.932–0.938.932.938) with the GNN model. For LUAD-*TP53*, the average AUROC was 0.869 (95% CI: 0.862–0.876.862.876) with the MLP model and 0.872 (95% CI: 0.866–0.880.866.880) with the GNN model. Similarly, for LUAD-*EGFR*, the average AUROC was 0.868 (95% CI: 0.858–0.878.858.878) with the MLP model and 0.865 (95% CI: 0.855–0.875.855.875) with the GNN model. For LUAD-*KRAS*, the MLP model achieved an average AUROC of 0.845 (95% CI: 0.835–0.854.835.854), outperforming the GNN model, which had an AUROC of 0.792 (95% CI: 0.782–0.801.782.801). In the case of pancreatic adenocarcinoma (PAAD)-*KRAS*, the average AUROC was 0.904 (95% CI: 0.894–0.915.894.915) with the MLP model and 0.892 (95% CI: 0.879–0.904.879.904) with the GNN model. For STAD-*TP53*, both the MLP and GNN models exhibited similar performance, achieving average AUROC of approximately 0.790 (MLP: 95% CI: 0.779–0.800.779.800; GNN: 95% CI: 0.780–0.799.780.799).

Grading (GRAD): For prostate adenocarcinoma (PRAD) grading, our refined gene sets showed the highest performance among all gene sets with the MLP mode, achieving an average AUROC of 0.779 (95% CI: 0.771–0.787.771.787). The GNN model exhibited slightly lower performance, with an AUROC of 0.695 (95% CI: 0.686–0.704.686.704).

Despite these consistent results, the XGBoost-refined hdWGCNA gene sets did not outperform reference sets in all tasks. In MSI classification for STAD, the GNN model using OncoKB performed slightly better, with an AUROC only 0.003 higher than ours. Similarly, for BRCA-*TP53* mutation prediction, the OncoKB gene set performed best, but the differences were minimal. The MLP model had an AUROC difference of only 0.006 and the GNN model had an AUROC difference of only 0.001. For cancer subtyping, all evaluated gene sets already

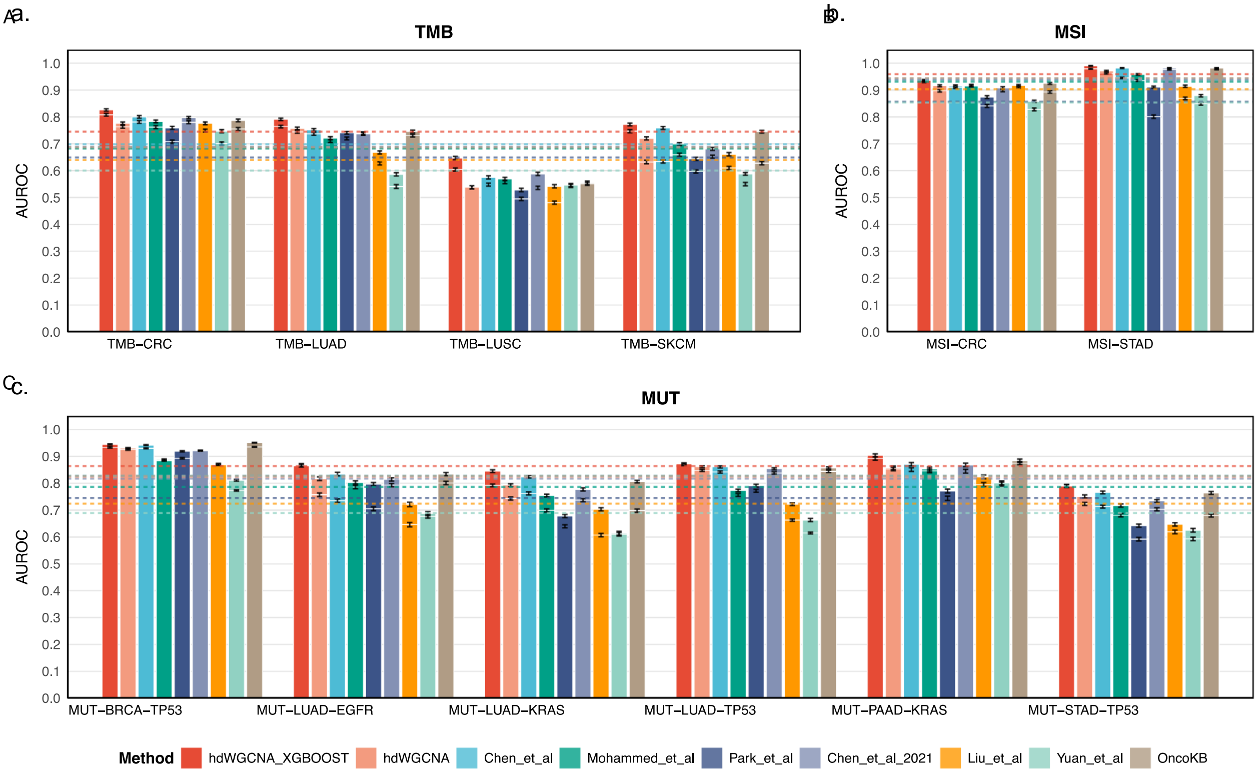

Fig. 3. Comparative performance analysis of gene sets across downstream tasks. Bar plots showing the performance of nine different gene selection methods evaluated across downstream tasks using the TCGA pan-cancer RNA-seq dataset. Each bar represents the mean AUROC across 5-fold cross-validation, with error bars indicating 95% confidence intervals obtained from bootstrapping. Dashed horizontal lines mark the average AUROC across all methods within each task category for visual reference. **a**. Tumor mutation burden (TMB) assessment performances across four cancer types: CRC, LUAD, LUSC, and SKCM. **b**. Microsatellite instability (MSI) classification for CRC and STAD. c. Mutation prediction (MUT) performance for six cancer gene combinations: BRCA-TP53, LUAD-EGFR, LUAD-KRAS, LUAD-TP53, PAAD-KRAS, and STAD-TP53. All models were trained using a multi-layer perceptron (MLP) architecture with identical hyperparameters to ensure fair comparison. Higher AUROC values indicate stronger discriminative performance of the corresponding gene set.

exhibited high AUROCs, resulting in minimal performance gaps. Thus, detailed results are provided separately in Supplementary Table 8.

In summary, the XGBoost-refined hdWGCNA gene sets achieved robust and statistically significant performance across most downstream tasks, highlighting their effectiveness in optimizing gene selection for pan-cancer analyses.

Biological significance of XGBoost-refined hdWGCNA gene sets

To evaluate the biological relevance of the XGBoost-refined hdWGCNA gene sets, we analyzed feature importance of selected genes within each downstream task. For each task, the top 10 genes with the highest importance values were identified (Fig. 4a-d). Many of these genes are known to be associated with cancer biology, supporting the biological validity of our findings[25–37].

For *TP53* mutation prediction in BRCA, *CIRBP*, a regulator of DNA repair and cell proliferation has been implicated in triple-negative breast cancer[25]. Similarly, *BTF3* has been reported to be up-regulated in individuals with *TP53* mutations across various cancer types[26]. In LUAD, *KPNA2*, a nuclear export protein essential for tumor formation, has been recognized for its significant role in human tumors[27]. Additionally, mutations in *CDKN2A* have been associated with *TP53* mutations, underscoring its relevance in this context[28]. For *TP53* mutation prediction in STAD, *DDB2*, closely linked to the p53 pathway, acts as a key regulator of p21 Waf1/Cip1 following DNA damage or p53-mediated induction[29]. In the LUAD-*EGFR* mutation prediction task, *GGA2* was identified as a critical gene due to its interaction with *EGFR*, which increases *EGFR* protein levels and modulates its degradation[30]. For *KRAS* mutation prediction in LUAD, *RHOB*, known to be down-regulated in non-small-cell lung cancer (NSCLC), is directly related to lung cancer progression[31]. In PAAD, *S100A11*, a well-established pancreatic tissue marker, was identified as crucial due to its association with tumor stage, drug resistance, and shorter overall survival[32]. *YWHAZ*, also identified in PAAD, is recognized as a promising therapeutic target due to its role in promoting cell proliferation, migration, and invasion[37]. For MSI classification tasks, *RPL22L1* was

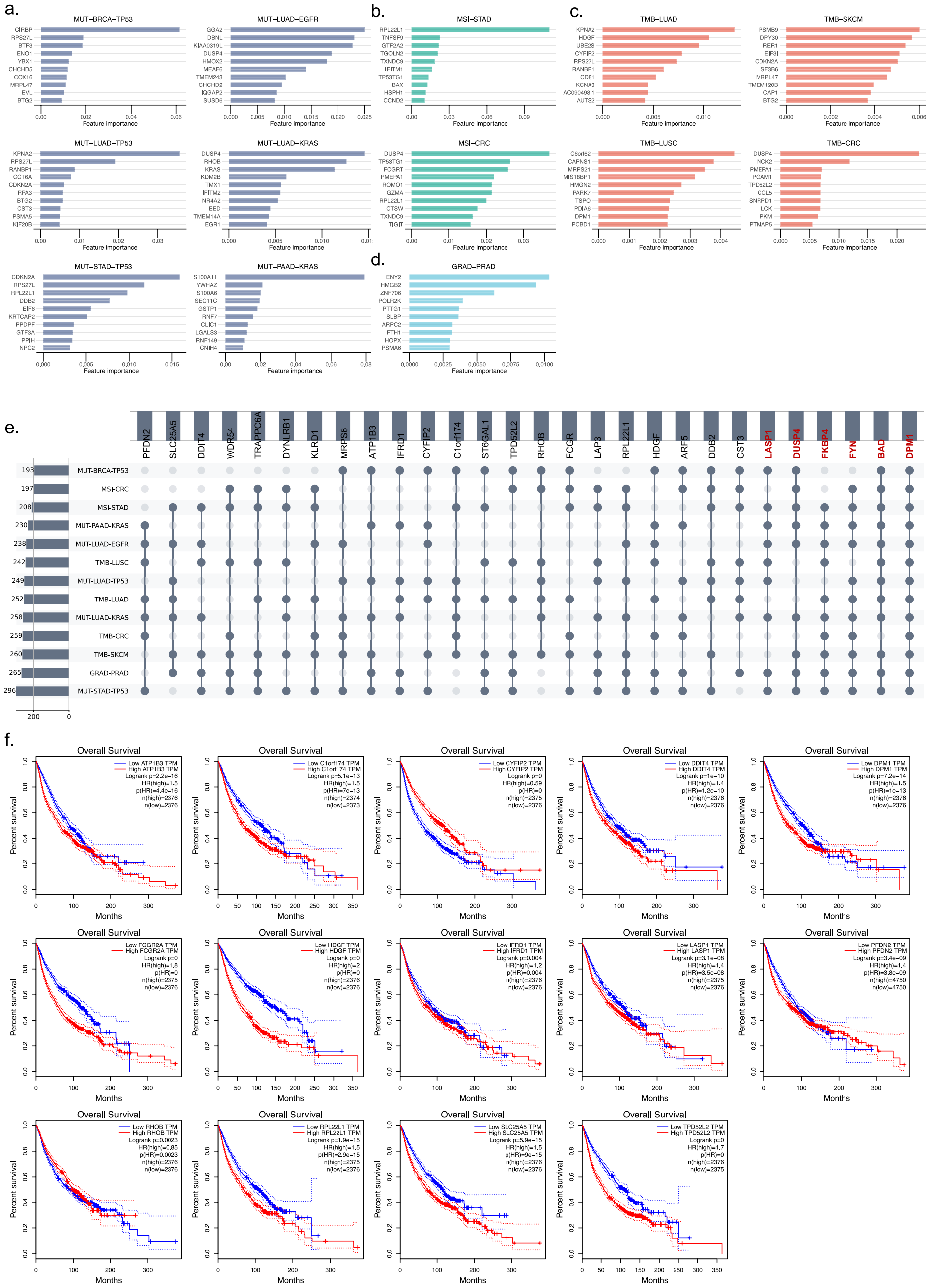

Fig. 4. Feature importance analysis and overlap of key genes across downstream tasks. **a**-**d**. Feature importance scores for the top 10 genes in each downstream task. Bar plots display XGBoost-derived feature importance scores (x-axis) for genes (y-axis). Panels represent: a. Mutation prediction (MUT) tasks across six gene-cancer combinations, **b**. Microsatellite instability (MSI) classification in CRC and STAD. **c**. Tumor mutation burden (TMB) assessment in four cancer types, **d**. Cancer grading (GRAD) in PRAD. **e**. Upset plot illustrating gene overlap across all 13 downstream tasks. The left vertical bar plot shows the total number of features selected for each task. The matrix indicates which tasks share common genes (connected dots), showing the size of each intersection. **f**. Kaplan-Meier survival curves for 14 genes showing significant pan-cancer prognostic value ($p < 0.005$). Each panel shows overall stratified by gene expression level: high expression (red line, top quartile) vs low expression (blu line, bottom quartile). Each plot displays the log-rank p-value, hazard ratio (HR) with 95% confidence interval, and HR p-value. All survival analyses were performed across 33 TCGA cancer types using the GEPIA platform.

significant in MSI-STAD due to its up-regulation in high-MSI groups across various cancer types[33]. In MSI-CRC, *PMEPA1* was notable for its association with tumor immunity in pan-cancer studies[34]. In TMB tasks, *PSMB9* was identified as a strong predictor of immune response in melanoma patients, significantly influencing their response to checkpoint therapy[35]. In TMB-CRC, *CCL5* was identified as a prognostic biomarker associated with the efficacy of immunotherapeutic interventions in various cancers[36].

To identify the overlap of selected genes across downstream tasks, we visualized them using upset plots (Figure 3e). These plots highlight genes that consistently ranked high in feature importance across multiple tasks. In particular, *DPM1* was significant in all 13 tasks, highlighting its critical role in pan-cancer studies. *DPM1* has previously been identified as a prognostic indicator in hepatocellular carcinoma (HCC), further supporting its importance in cancer biology[38]. Moreover, *BAD* ranked high in all tasks except TMB-CRC, indicating its broad involvement across multiple cancer pathways. *FKBP4*, which appeared in 10 tasks, has been associated with prognostic and immunological roles in several cancers[39]. *DUSP4*, associated with treatment resistance in breast cancer was identified[40]. Similarly, *LASP1* which is involved in metastasis in pancreatic cancer, was present in 9 tasks[41]. These recurrent genes highlight shared oncogenic mechanisms across cancer types and represent promising targets for future pan-cancer research.

To further validate the clinical relevance, we performed pan-cancer survival analysis using GEPIA with TCGA clinical data. We focused on genes with high feature importance across multiple tasks (Fig. 4e), stratifying patients into high and low expression groups using quartile cutoffs. Remarkably, 14 genes showed significant associations with overall survival ($p < 0.005$) at the pan-cancer level (Fig. 4f). These results provide independent clinical validation of our computation findings, showing that the hdWGCNA-XGBoost approach identifies features that are both statistically robust and biologically meaningful markers with direct clinical implications.

Beyond statistical associations, thirteen of the fourteen survival-associated genes identified in our analysis have prior evidence supporting their clinical and biological relevance. High expression of *DPM1* has been associated with poor prognosis in hepatocellular carcinoma, reflecting its involvement in aberrant protein glycosylation[42]. *LASP1* has been shown to drive metastasis and unfavorable clinical outcomes in several malignancies, including pancreatic adenocarcinoma[43]. *HDGF* has been widely recognized as a prognostic marker and therapeutic candidate across solid tumors[44]. In contrast, reduced expression of *RHOB* and *CYFIP2* correlates with aggressive phenotypes and poorer survival, consistent with their tumor-suppressive functions[45,46]. *DDIT4* mediates therapy resistance through mTORC1 inhibition and metabolic adaptation[47]. Upregulation of *PFDN2*, *TPD52L2*, and *IFRD1* has been reported in association with tumor aggressiveness, recurrence, or resistance phenotypes[48–50]. Expression of *ATP1B3* correlates with survival outcomes in hepatocellular and gastric cancers[51], while *C1orf174* shows diagnostic and prognostic potential in colorectal cancer[52]. *RPL22L1* promotes invasion and metastasis, contributing to adverse clinical outcomes in multiple cancer types. Although *SLC25A5* has been implicated in tumor metabolism and treatment response in preclinical studies[53], there is currently no direct evidence that genetic variation in this gene modifies the clinical benefit of antibody-based therapies. Accordingly, *SLC25A5* is conservatively presented as a potential biomarker warranting further validation.

Discussion

This study highlights the potential of integrating scRNA-seq data with advanced feature selection to improve predictive modeling in cancer genomics. By combining hdWGCNA with XGBoost-based refinement, we generated scRNA-seq-derived gene sets that outperformed conventional bulk RNA-seq gene sets across multiple pan-cancer tasks. The high-resolution single-cell data captured cellular heterogeneity and tumor microenvironment interactions, thereby enhancing biological interpretability. Additionally, XGBoost refinement enabled the selection of the most significant genes tailored to specific tasks, improving model accuracy and robustness. Many of the identified genes are well-established in cancer research, confirming the validity of our framework. In addition, several novel genes emerged as a potential pan-cancer biomarkers, offering new avenues for therapeutic exploration. Overall, our findings underscore the value of integrating scRNA-seq data with machine learning based feature selection, providing a scalable and biologically grounded strategy for discovering key genetic drivers in cancer.

Traditional feature selection approaches for RNA-seq data, such as those using MsigDB or OncoKB, typically rely on predefined pathways or cancer gene-based sets[24,54]. These databases have shown high performance due to their curated biological foundation, effectively capturing relevant pathways and processes. However, methods that depend solely on bulk RNA-seq or static gene lists often fail to reflect the complexity of the tumor microenvironment. For example, Mohammed et al. used LASSO regression to screen 173 genes with 10-fold cross-validation[11]. Similarly, Chen, Joe W., et al. selected the top 500 differentially expressed genes by p-value[21], and Silva., et al chose genes with the highest variance across samples[12]. While these methods are statistically robust, those approaches overlook cellular heterogeneity and gene interactions within tumors, potentially missing genes active in rare but critical cell populations. Furthermore, the resulting feature sets are typically tailored to single-cell analyses, limiting their generalizability for identifying shared biomarkers across multiple cancer types.

Our framework introduces several key advances over existing approaches. First, by leveraging scRNA-seq data with hdWGCNA, we capture cell-type specific co-expression patterns that remain hidden in bulk analyses. Second, the integration with XGBoost enables task-specific gene prioritization, moving beyond static pathway- or variance-based selection. Unlike previous studies that relied on statistical filters, differential expression, or variance-based selection, our method combines the biological insights from hdWGCNA with the predictive power of XGBoost. To our knowledge, this represents the first application of this integrated approach in pan-cancer features selection. Furthermore, our refined gene sets consistently outperformed reference gene sets across most tasks, demonstrating their robustness and broad applicability. This approach improves predictive accuracy while enhancing biological interpretability and generalizability across cancer types. The success of our gene sets in these tasks underscores their potential utility in multi-modal research, where combining diverse omics data could further enhance predictive modeling.

Our study highlights several genes, including *DPM1*, *BAD*, and *FKBP4*, as promising pan-cancer biomarkers. *FKBP4* has been previously reported in a pan-cancer context, supporting its relevance across multiple cancer types[39]. Similarly, *DPM1* has been identified as a prognostic marker in hepatocellular carcinoma and breast cancer[38], and also showed consistent importance across all 13 tasks in our analysis, suggesting its potential as a pan-cancer biomarker. Furthermore, our pan-cancer survival analysis revealed that 14 high-importance genes such as *DPM1*, *LASP1*, and *RHOB*, demonstrated significant associations with patient survival. This independent clinical validation confirms that our computational approach identifies not only statistically significant features but also biomarkers with direct prognostic relevance. Notably, many of the survival-associated genes identified in our analysis correspond to known prognostic or therapeutic biomarkers in clinical oncology. Their consistent recurrence across multiple cancer types underscores the translational value of our framework, bridging computational discovery with clinically actionable insights that may guide patient stratification and targeted therapy development.

Like other studies, our study has certain limitations. Frist, although the scRNA-seq dataset included 13 cancer types, it did not cover all possible tumor types, which may limit the generalizability of our findings. However, our framework maintained strong performance even in unseen cancers, such as the grading task in PRAD, highlighting its broad applicability. Second, while we propose that our gene sets could enhance multi-modal model performance, this remains to be experimentally validated. Recent pan-cancer studies have increasingly adopted multi-modal approaches, integrating transcriptomic data with other multi-omics or whole-slide pathology images[55,56]. In particular, foundation models trained on whole-slide images (WSI) have shown that incorporating RNA features through self-supervised fine-tuning can improve downstream task performance. Furthermore, as our downstream tasks align with precision medicine goals, recent advances in contrastive learning show promise in biomedical applications. Contrastive learning has been successfully applied in both drug-target affinity prediction, and broader computational drug repositioning under label sparsity highlight the potential of representation learning in biomedical domains[57–59]. While these methods have primarily focused on drug-molecular interactions, incorporating transcriptomics features from targeted gene sets could enable these frameworks to account for patient-specific molecular contexts, potentially improving their predictive accuracy in clinical settings.

Beyond these advances, several recent studies in omics-driven drug discovery, such as OmicsTweezer and GraphCL-DTA[60,61], have demonstrated that effective feature selection and representation learning from omics data can substantially enhance predictive modeling in drug-target interaction, drug repositioning, and multi-omics integration. These frameworks share conceptual similarities with our hdWGCNA-XGBoost approach, as they all rely on extracting biologically meaningful molecular features that drive downstream predictive performance. Integrating scRNA-seq-derived gene sets such as those proposed in our study could further complement these models by introducing patient-specific transcriptomic signals, thereby enriching molecular context in drug response prediction and other translational applications.

In conclusion, our study provides a robust framework to optimize feature selection in cancer genomics by integrating high-resolution scRNA-seq data with advanced analytical techniques. This approach improves predictive performance across diverse pan-cancer tasks and identifies gene sets with strong biological and clinical relevance. By revealing biomarkers with pan-cancer significance, our framework supports the development of more accurate and reliable diagnostics and prognostic models. Moreover, our results offer valuable insights that can be applied to multi-modal frameworks, thereby advancing personalized cancer treatment strategies and contributing to the broader field of cancer genomics.

Methods

Datasets

To characterize the gene expression landscape across a wide range of cancers, we obtained scRNA-seq data for 181 tumor biopsy samples from 13 different cancer types, which were collected by Nieto., et al.[18]. We obtained RNA-seq data for 7,178 tumor samples across 16 cancer types from The Cancer Genome Atlas (TCGA), accessed through the NIH Genomic Data Commons Portal[62]. All datasets used in this study were downloaded using the GenomicDataCommons package in R.

To facilitate various downstream tasks, cancer types were grouped as follows. For the MSI classification, CRC, specifically colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ), were collectively analyzed as a combined TCGA project. Similarly, for subtype classification, lung cancers were grouped as non-small cell lung cancer (NSCLC), combining both LUAD and LUSC. Kidney cancers were classified as renal cell carcinoma (RCC), comprising kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), and kidney chromophobe (KICH). Melanomas (MEL), which include skin cutaneous melanoma (SKCM) and uveal melanoma (UVM), were classified separately due to their distinct anatomical origins, despite both being melanomas. Finally, gynecologic cancers (GYN) were comprised into ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma, and endocervical adenocarcinoma (CESC).

RNA-seq data pre-processing

RNA-seq data were pre-processed to ensure consistency across samples and to facilitate accurate comparative analysis. Read counts were first normalized by adjusting for gene length to accounts for differences in transcript length. The normalized counts were then scaled to counts per million to standardize for sequencing depth, providing a standardized comparison of gene expression levels. Then, the read counts were log-transformed to stabilize variance[63].

Data preprocess for pan-cancer downstream tasks

Based on this processed data, TMB scores were calculated for each sample by further normalizing read counts against gene lengths and adjusting for sequencing depth. First, the RNA-seq data were mapped to barcodes for each sample. Then, read counts were divided by gene lengths and were scaled by a factor of 10^6 bases to derive

TMB scores for each sample. Samples with a TMB score higher than 10 were categorized as TMB-high, while those with a TMB score below 10 were classified as TMB-low[64].

To classify MSI status, we accessed missense mutations in key DNA mismatch repair (*MMR*) genes, including *MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1*, and *PMS2*[65]. Samples were classified as microsatellite unstable (MSI-H) or microsatellite stable (MSS).

For prostate cancer grading, we used gleason scores to categorize samples into three groups: low, moderate, and high. Clinical data were downloaded from cBioPortal, and samples were grouped based on their gleason scores[66]. Samples with a score of 6 were assigned to the low-grade group, a score of 7 in the moderate group, and scores of 8, 9, and 10 in the high group.

For the mutation prediction tasks, we used TCGA somatic mutation data to identify mutations. We determined the mutation status of target genes in each sample. This mutation information was then cross-referenced with RNA-seq data to ensure consistency in the samples. Only patients with both mutation data and corresponding RNA-seq data were included in the final dataset for further analysis.

Feature selection through high dimensional weighted gene co-expression network analysis

To identify co-expressed modules in the pan-cancer scRNA-seq data, we performed weighted gene co-expression network analysis (WGCNA)[67]. First, metacells were constructed for each immune cell subset to optimize the data sparsity and reduce technical noise. We applied principal component analysis (PCA) for dimensionality reduction, followed by the k-nearest neighbors (k=25) aggregation to generate metacells within each cell type-patient group. The metacell gene expression matrix was then normalized and used as input for WGCNA. The soft-thresholding power (β=9) was selected using the TestSoftPower function to achieve approximate scale-free topology, as recommended in the hdWGCNA pipeline. Co-expression modules were then identified using the dynamic tree-cut method with a minimum module size of 30. Harmonized module eigengenes (hMEs) were calculated across patients to summarize module-level activity. For network visualization, we generated a UMAP embedding based on the topological overlap matrix, which illustrating the modular structure and relationships among genes.

Pathway enrichment analysis

Pathway enrichment analysis was performed using the GO Biological Process database[68]. For each hdWGCNA module, we selected the top 100 genes based on kME scores. Enrichment significance was accessed using log-transformed P-values. The top 5 enriched terms were selected for further analysis. Analyses were performed using the Enrichr package in R[69].

Model development

We developed two separate model architectures optimized for RNA-seq data: a multi-layer perceptron (MLP) model and a graph neural network (GNN) model.

MLP

The MLP model comprises of an embedding module and a classifier module. The embedding module reduces input dimensional input features into a lower-dimensional space through a single linear layer. This dimensionality reduction is followed by batch normalization and Leaky ReLU activation, which stabilize training and enhance pattern recognition[70]. Dropout is applied to prevent overfitting. The classifier module processes these features through three sequential linear layers, with SELU activation and dropout after each layer[71]. This design preserves feature variability and enhances classification robustness. Additionally, the number of neurons in each layer is adjusted in each layer according to the specific needs of each downstream task.

GNN

The GNN model was adapted from previous study[72]. The gene embedding module contains an input embedding layer that projects input features to a hidden dimension using a linear layer, followed by dropout to prevent overfitting. This is followed by multiple graph transformer layers that perform message passing and update node features, incorporating batch normalization and layer normalization for training stability. Each graph transformer layer uses multi-head attention to capture diverse relational patterns. Graph layer outputs are reshaped and processed by a fully connected linear layer, followed by batch normalization and dropout. This embedding is then flattened to form the input to the classifier module. The classifier module consists of three sequential linear layers with SELU activation functions and dropout. Batch normalization is applied after each linear layer to maintain feature stability.

Model training

Both models were trained using binary cross-entropy loss for binary classification and cross-entropy loss functions for multi-class classification. Models were optimized using the Adam optimizer with a mini-batch size of 32 and a learning rate of 0.01[73]. Early stopping with a patience of 30 epochs was implemented. Model performance was evaluated using stratified cross-validation[74]. For each fold, we conducted 100 bootstrap iterations on the test set to obtain 95% confidence interval. This approach ensures robust performance estimates across different patient samples. All experiments were conducted using PyTorch (version 2.0.1) on an NVIDIA A100 GPU (40 Gb) equipped with CUDA version 11.7.

Feature importance through XGBoost

We used XGBoost to access feature importance from genes[75]. For each downstream task, we performed 100 bootstrap iterations to assess the importance of gene features. In each iteration, we resampled the dataset, trained the XGBoost classifier, and evaluated its performance using metrics such as AUROC and AUPRC. This bootstrap approach identified genes with consistently high feature importance, highlighting their importance in biological processes relevant to our study.

Pan-cancer survival analysis

To validate the clinical relevance of our identified biomarkers, we performed pan-cancer survival analysis using the GEPIA2 platform (http://gepia.cancer-pku.cn/). For each gene, we evaluated overall survival across all 33 available TCGA cancer types. Patients were stratified into high and low expression groups using quartile cutoffs (top 25% vs bottom 25%). Kaplan-Meier survival curves were generated, and statistical significance was accessed using log-rank test. Hazard ratios (HR) with 95% confidence intervals were calculated using Cox proportional hazards regression. Genes with p(HR) < 0.005 were considered to have significant prognostic value.

Data availability

All scRNA-seq data utilized in this study are publicly accessible. The datasets were obtained from links provided by Nieto et al.[18] https://doi.org/10.5281/zenodo.4263972. The remaining gene sets used for downstream tasks were also downloaded from links provided in their respective publications: Chen et al.[10], Mohammed et al.[19,20], Park et al., Chen et al.[21], Liu et al.[22], Yuan et al[23]. Lastly, cancer-specific gene lists were obtained directly from the OncoKB website https://www.oncokb.org/cancer-genes.

Code availability

The source code and software pipeline to reproduce our study can be accessed at https://github.com/kimjh0107/2024_pancancer_scRNA.git.

Received: 23 July 2024; Accepted: 3 November 2025
Published online: 02 December 2025

References

1. Tran, K. A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine* 13 Preprint at https://doi.org/10.1186/s13073-021-00968-x (2021).
2. Zhang, B., Shi, H. & Wang, H. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. *Journal of Multidisciplinary Healthcare* 16 1779–1791 Preprint at https://doi.org/10.2147/JMDH.S410301 (2023).
3. Yang, W., Qiang, Y., Wu, W. & Xin, J. Graph-ETMB: A graph neural network-based model for tumour mutation burden estimation. *Comput Biol Chem* **105**, 107900 (2023).
4. Alharbi, F. & Vakanski, A. Machine learning methods for cancer classification using gene expression data: A Review. *Bioengineering* vol. 10 Preprint at https://doi.org/10.3390/bioengineering10020173 (2023).
5. Vale-Silva, L. A. & Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **11**, 13505 (2021).
6. Zelli, V. et al. Classification of tumor types using XGBoost machine learning model: a vector space transformation of genomic alterations. *J. Transl. Med.* **21**, 836 (2023).
7. Taghizadeh, E., Heydarheydari, S., Saberi, A., JafarpoorNesheli, S. & Rezaeijo, S. M. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinform.* **23**, 410 (2022).
8. Arowolo, M. O., Adebiyi, M. O., Aremu, C. & Adebiyi, A. A. A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector. *J. Big Data* **8**, 50 (2021).
9. Chen, Z. et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* **36**, 1542–1552 (2020).
10. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865-878.e6 (2022).
11. Elbashir, M. K., Ezz, M., Mohammed, M. & Saloum, S. S. Lightweight convolutional neural network for breast cancer classification using RNA-Seq gene expression data. *IEEE Access* **7**, 185338–185348 (2019).
12. *IEEE ISBI 2020 International Symposium on Biomedical Imaging : 2020 Symposium Proceedings : April 3-7, 2020, Iowa City, Iowa*. (IEEE, 2020).
13. Lee, H. W. et al. Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. *Genome. Med.* **12**, 47 (2020).
14. Lei, Y. *et al.* Applications of single-cell sequencing in cancer research: progress and perspectives. *Journal of Hematology and Oncology* 14 Preprint at https://doi.org/10.1186/s13045-021-01105-2 (2021).
15. Ren, X., Kang, B. & Zhang, Z. Understanding tumor ecosystems by single-cell sequencing: Promises and limitations. *Genome Biology* 19 Preprint at https://doi.org/10.1186/s13059-018-1593-z (2018).
16. Huang, D. *et al.* Advances in single-cell RNA sequencing and its applications in cancer research. *Journal of Hematology and Oncology* 16 Preprint at https://doi.org/10.1186/s13045-023-01494-6 (2023).
17. Zhang, Y. *et al.* Single-cell RNA sequencing in cancer research. *Journal of Experimental and Clinical Cancer Research* 40 Preprint at https://doi.org/10.1186/s13046-021-01874-1 (2021).
18. Nieto, P. et al. A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**, 1913–1926 (2021).
19. Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K. & Omolo, B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Sci. Rep.* **11**, 15626 (2021).
20. Park, S. et al. Wx: a neural network-based feature selection algorithm for transcriptomic data. *Sci. Rep.* **9**, 10500 (2019).
21. Chen, J. W. & Dhahbi, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci. Rep.* **11**, 13323 (2021).
22. Liu, S. & Yao, W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. *BMC Bioinform.* **23**, 175 (2022).
23. Yuan, F., Lu, L. & Zou, Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochim. Biophys. Acta Mol. Basis Dis.* **1866**, 165822 (2020).
24. *OncoKB: A precision oncology knowledge base*. http://oncokb.org (2017).
25. Adinew, G. M., Messeha, S., Taka, E. & Soliman, K. F. A. The prognostic and therapeutic implications of the chemoresistance gene BIRC5 in triple-negative breast cancer. *Cancers (Basel)* **14**, 5180 (2022).
26. Parikh, N. et al. Effects of TP53 mutational status on gene expression patterns across 10 human cancer types. *J. Pathol.* **232**, 522–533 (2014).

27. Sun, Y. et al. Oncogenic role of karyopherin α2 (KPNA2) in human tumors: A pan-cancer analysis. *Comput. Biol. Med.* **139**, 104955 (2021).
28. Deneka, A. Y. et al. Association of TP53 and CDKN2A mutation profile with tumor mutation burden in head and neck cancer. *Clin. Cancer Res.* **28**, 1925–1937 (2022).
29. Stoyanova, T., Roy, N., Kopanja, D., Bagchi, S. & Raychaudhuri, P. *DDB2 Decides Cell Fate Following DNA Damage*. www.pnas.org/cgi/content/full/.
30. O'Farrell, H. et al. Integrative genomic analyses identifies GGA2 as a cooperative driver of egfr-mediated lung tumorigenesis. *J. Thoracic Oncol.* **14**, 656–671 (2019).
31. Sato, N. et al. RhoB is frequently downregulated in non-small-cell lung cancer and resides in the 2p24 homozygous deletion region of a lung cancer cell line. *Int. J. Cancer* **120**, 543–551 (2007).
32. Li, X., Qiu, N. & Li, Q. Prognostic Values and Clinical Significance of S100 Family Member's Individualized mRNA Expression in Pancreatic Adenocarcinoma. *Front Genet* **12**, (2021).
33. Zheng, Y. et al. Multi-omics characterization and validation of MSI-related molecular features across multiple malignancies. *Life Sci.* **270**, 119081 (2021).
34. Wang, B. et al. Diagnostic and therapeutic values of PMEPA1 and its correlation with tumor immunity in pan-cancer. *Life Sci.* **277**, 119452 (2021).
35. Kalaora, S. et al. Immunoproteasome expression is associated with better prognosis and response to checkpoint therapies in melanoma. *Nat. Commun.* **11**, 896 (2020).
36. Huang, Y. *et al.* CCL5 might be a prognostic biomarker and associated with immuno-therapeutic efficacy in cancers: A pan-cancer analysis. *Heliyon* **9** (2023).
37. Liu, S. et al. Endoplasmic reticulum stress regulators exhibit different prognostic, therapeutic and immune landscapes in pancreatic adenocarcinoma. *J. Cell Mol. Med.* **28**, e18092 (2024).
38. Li, M., Xia, S. & Shi, P. DPM1 expression as a potential prognostic tumor marker in hepatocellular carcinoma. *PeerJ* **8**, e10307 (2020).
39. Xiong, H. *et al.* Pan-cancer analysis of the prognostic and immunological role of FKBP4. *Heliyon* **10** (2024).
40. Menyhart, O. *et al.* DUSP4 is associated with increased resistance against anti-HER2 therapy in breast cancer. *Oncotarget* **8**(44), 77207 (2017).
41. Zhao, T. et al. LASP1 is a HIF1α target gene critical for metastasis of pancreatic cancer. *Cancer Res.* **75**, 111–119 (2015).
42. Li, M., Xia, S. & Shi, P. DPM1 expression as a potential prognostic tumor marker in hepatocellular carcinoma. *PeerJ* **8**, e10307 (2020).
43. Zhao, T. et al. LASP1 is a HIF1α target gene critical for metastasis of pancreatic cancer. *Cancer Res.* **75**, 111–119 (2015).
44. Bao, C. H. et al. Prognostic role of hepatoma-derived growth factor in solid tumors of eastern Asia: a systematic review and meta-analysis. *Asian Pac. J. Cancer Prev.* **16**, 1803–1811 (2015).
45. Ju, J. A. & Gilkes, D. M. Rhob: Team oncogene or team tumor suppressor? *Genes* 9 Preprint at https://doi.org/10.3390/genes9020067 (2018).
46. Tong, J. et al. The downregulation of prognosis-and immune infiltration-related gene CYFIP2 serves as a novel target in CcRCC. *Int. J. Gen. Med.* **14**, 6587–6599 (2021).
47. Foltyn, M. et al. The physiological mTOR complex 1 inhibitor DDIT4 mediates therapy resistance in glioblastoma. *Br. J. Cancer* **120**, 481–487 (2019).
48. Herranz-Montoya, I., Park, S. & Djouder, N. A comprehensive analysis of prefoldins and their implication in cancer. *iScience* **24**, 103273 (2021).
49. Wang, H. et al. TPD52L2 as a potential prognostic and immunotherapy biomarker in clear cell renal cell carcinoma. *Front. Oncol.* **13**, 1210910 (2023).
50. Lewis, M. A. et al. Increased IFRD1 expression in human colon cancers predicts reduced patient survival. *Dig. Dis. Sci.* **62**, 3460–3467 (2017).
51. Lu, S., Cai, S., Peng, X., Cheng, R. & Zhang, Y. Integrative transcriptomic, proteomic and functional analysis reveals ATP1B3 as a diagnostic and potential therapeutic target in hepatocellular carcinoma. *Front. Immunol.* **12**, 636614 (2021).
52. Wu, N. et al. Ribosomal L22-like1 (RPL22L1) promotes ovarian cancer metastasis by inducing epithelial-to-mesenchymal transition. *PLoS One* **10**, e0143659 (2015).
53. Seneviratne, J. A. et al. Inhibition of mitochondrial translocase SLC25A5 and histone deacetylation is an effective combination therapy in neuroblastoma. *Int. J. Cancer* **152**, 1399–1413 (2023).
54. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
55. Ding, T. *et al.* Multimodal whole slide foundation model for pathology. http://arxiv.org/abs/2411.19666 (2024).
56. Chen, W. et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nat. Methods* https://doi.org/10.1038/s41592-025-02707-1 (2025).
57. Yang, X., Yang, G. & Chu, J. GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. http://arxiv.org/abs/2307.08989 (2023).
58. Yang, X., Yang, G. & Chu, J. The computational drug repositioning without negative sampling. http://arxiv.org/abs/2111.14696 (2022).
59. He, J., Yang, X., Gong, Z. & Zamit, L. Hybrid attentional memory network for computational drug repositioning. *BMC Bioinform.* **21**, 566 (2020).
60. Yang, X. et al. OmicsTweezer: A distribution-independent cell deconvolution model for multi-omics data. *Cell Genom.* https://doi.org/10.1016/j.xgen.2025.100950 (2025).
61. Yang, X., Yang, G. & Chu, J. GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. http://arxiv.org/abs/2307.08989 (2023).
62. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
63. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* **19**, 776–792 (2018).
64. Marabelle, A. et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol.* **21**, 1353–1365 (2020).
65. Zaravinos, A. et al. Cytolytic activity correlates with the mutational burden and deregulated expression of immune checkpoints in colorectal cancer. *J. Exp. Clin. Cancer Res.* **38**, 364 (2019).
66. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
67. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E. & Swarup, V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods* **3** (2023).
68. Harris, M. A. *et al.* The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Res* **32**, (2004).
69. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
70. Xu, B., Wang, N., Chen, T. & Li, M. Empirical evaluation of rectified activations in convolutional network. http://arxiv.org/abs/1505.00853 (2015).

71. Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S. Self-normalizing neural networks. http://arxiv.org/abs/1706.02515 (2017).
72. Jeong, D., Koo, B., Oh, M., Kim, T. B. & Kim, S. GOAT: Gene-level biomarker discovery from multi-Omics data using graph ATtention neural network for eosinophilic asthma subtype. *Bioinformatics* **39**, btad582 (2023).
73. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 (2017).
74. Prusty, S., Patnaik, S. & Dash, S. K. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front. Nanotechnol.* **4**, 972421 (2022).
75. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016 785–794 (Association for Computing Machinery 2016).

Author contributions

J.H.K., S.L, and J.J. conceived of and designed the study. J.H.K analyzed the data and performed data processing and visualization. S.L and J.J. supervised the project. J.H.K. and J.J. wrote the manuscript.

Declarations

Competing interest

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-27296-z.

Correspondence and requests for materials should be addressed to J.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.