*Briefings in Bioinformatics*, 2026, 27, bbag160
https://doi.org/10.1093/bib/bbag160
Published: 17 April 2026
**Problem Solving Protocol**

OXFORD

# scGPD: single-cell informed gene panel design for targeted spatial transcriptomics

Yunshan Guo[1], Jia Zhao[1], Rui B. Chang[2,3], Hongyu Zhao[1,*]

[1]Department of Biostatistics, School of Public Health, Yale University, 300 George Street, Ste 503, New Haven, CT 06511, United States
[2]Department of Neuroscience, School of Medicine, Yale University, 200 S Frontage Rd, New Haven, CT 06510, United States
[3]Department of Cellular and Molecular Physiology, School of Medicine, Yale University, 333 Cedar St, New Haven, CT 06510, United States

*Corresponding author. Department of Biostatistics, School of Public Health, Yale University, 300 George Street, Ste 503, New Haven, CT 06511, United States. E-mail: hongyu.zhao@yale.edu

## Abstract

**In targeted spatial transcriptomics technologies, a key challenge is to select an informative gene panel that captures the complexity of cellular and spatial heterogeneity within tissues. Many existing methods use prior knowledge or heuristic selection rules, such as selecting highly variable genes, which overlook gene–gene correlations and may consequently result in suboptimal coverage. To address the limitations of the existing methods, we introduce single-cell informed Gene Panel Design (scGPD), a deep learning-based framework for gene panel design that leverages single-cell RNA-seq data to identify compact, nonredundant sets of genes for spatial profiling. scGPD uses a gene–gene correlation-aware gating mechanism to extract informative features from data, encouraging diversity among selected genes and eliminating redundancy. Across diverse single-cell datasets, scGPD outperforms existing gene panel design methods in recovering transcriptome-wide expression using a limited number of genes. When applied to spatial transcriptomics data, it achieves competitive and robust cell-type classification performance, demonstrating strong generalization across modalities. The gene panels selected by scGPD further exhibit well-defined spatial expression patterns, highlighting their robustness and relevance for spatial analysis. The scGPD framework is flexible and can be adapted to multiple use cases, enabling the prioritization of genes relevant to specific diseases or phenotypes. Together, these results demonstrate that scGPD provides a robust and adaptable solution to design efficient gene panels for spatial transcriptomics, with broad applicability to tissue mapping and disease characterization.**

***Keywords*** spatial transcriptomics, deep learning, gene panel design

## Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the study of cellular heterogeneity by enabling transcriptomic profiling at single-cell resolution across diverse tissues and biological systems [1–4]. This technology has been instrumental in uncovering novel cell types, developmental trajectories, and disease-associated transcriptional programs. However, a fundamental limitation of scRNA-seq is the loss of spatial context, as tissues must be dissociated into single-cell suspensions. This dissociation disrupts native cell–cell interactions and obliterates spatial patterns of gene expression that are essential for understanding tissue organization and microenvironmental influences [5].

To overcome this limitation, spatial transcriptomics (ST) technologies have been developed to enable spatially resolved gene expression profiling within intact tissue sections. These methods preserve tissue architecture while allowing for the detection of messenger RNA (mRNA) abundance at various spatial resolutions [6–9]. ST technologies can be broadly categorized into two categories: spot-based ST (spot-ST) and single-cell ST (sc-ST) approaches.

Spot-ST methods, such as 10X Visium [10] and HDST [11], quantify gene expression within spatially resolved spots. While these approaches provide transcriptome-wide coverage, they lack single-cell resolution, making them less suitable for cell-type-specific analyses, such as studying spatial variation in ligand–receptor interactions.

In contrast, sc-ST technologies, such as MERFISH [9], seqFISH, seqFISH+ [12], and osmFISH [8], provide subcellular spatial precision and single-molecule sensitivity, which makes them particularly well suited for high-resolution tissue mapping [5]. Importantly, these sc-ST techniques are imaging-based single-molecule assays, where individual mRNA molecules are first detected *in situ* and subsequently assigned to individual cells through image-based cell segmentation.

**Received:** October 9, 2025. **Revised:** March 2, 2026. **Accepted:** March 11, 2026

As a result, downstream single-cell level analyses critically depend on the quality and accuracy of the segmentation step.

Despite these advantages, being fluorescence in situ hybridization (FISH)-based, each sc-ST dataset is limited by the number of genes that can be simultaneously measured, typically in the range of a few hundred to about a thousand genes, which restricts the full potential of sc-ST data in capturing comprehensive transcriptomic landscapes. Consequently, the design of targeted gene panels is a critical step in sc-ST workflows. The selection of a biologically informative and functionally relevant subset of genes is essential to accurately capture the cellular diversity and spatial organization within tissues. Gene selection for targeted ST is frequently guided by heuristic strategies, such as choosing well-characterized marker genes or those with high expression in specific cell subsets. Although these approaches are straightforward, they often fail to capture genes with more subtle, heterogeneous, or spatially complex expression patterns. A key limitation of such methods is that they do not account for correlations between genes, which can substantially reduce the informativeness of the selected panel. As a result, co-expressed or functionally related genes may be selected together, yielding overlapping signals that add little additional discriminatory power and ultimately limit the panel's ability to capture cellular and spatial diversity [13–15].

To address these challenges, we reformulate gene selection as a feature selection problem and propose single-cell informed Gene Panel Design (scGPD), a principled, data-driven framework grounded in deep learning. The framework leverages single-cell RNA-seq data to guide panel selection, systematically prioritizing informative and complementary genes to yield compact panels that preserve biological signal. In contrast to methods such as scGIST [16] and PERSIST [17], scGPD explicitly models gene–gene dependencies during panel selection rather than assuming independence among candidate genes.

A key innovation of scGPD lies in its correlation-aware gating mechanism [18], which explicitly accounts for gene–gene dependencies during training. Rather than allowing correlated genes to be selected together, our gating strategy promotes competition among related features, encouraging the model to retain only the most informative representative within each correlated group. This results in gene panels optimized for downstream tasks such as transcriptome reconstruction or cell type classification. Specifically, scGPD employs a correlated Gaussian copula-based gating mechanism to model dependency structure among genes, enabling principled control of feature overlap while promoting complementary markers that jointly capture cell-state heterogeneity. In contrast to existing methods, scGPD is also adaptable: it can scale to large datasets via minibatched training and accommodate diverse experimental designs and panel size requirements.

We demonstrate the efficacy and versatility of scGPD through extensive benchmarking on multiple publicly available single-cell RNA-seq datasets, covering a wide range of dataset sizes and cellular diversity. Across these datasets, scGPD consistently selects informative genes that preserve key biological variation. Beyond quantitative benchmarking, we show that the genes selected by scGPD from single-cell data also exhibit coherent spatial expression patterns in the corresponding ST datasets, highlighting the method's ability to identify biologically meaningful markers of spatial relevance. Furthermore, we demonstrate that scGPD can be adapted to prioritize disease-associated genes by modifying the objective of gene selection, enabling the targeted discovery of candidate biomarkers for spatial profiling in pathological contexts.

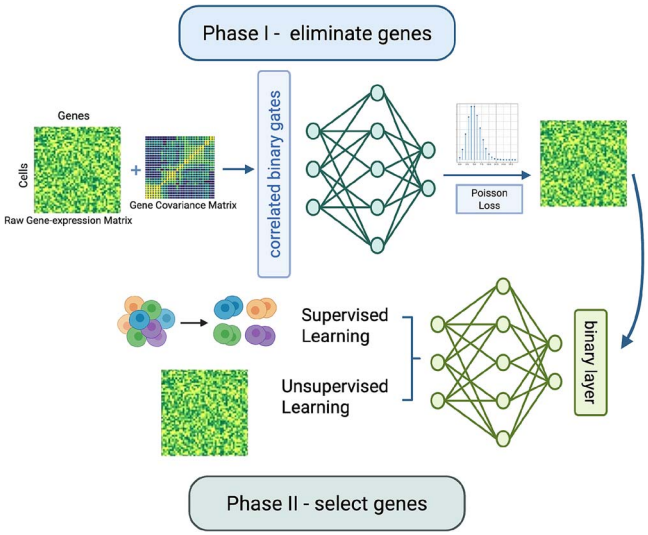



**Figure 1** Overview of scGPD, a dual-stage gene selection framework that first removes redundant genes using correlated binary gating and then selects a fixed number of informative genes via application-specific loss functions and a binary mask.

## Methods

### Method overview

scGPD is a deep learning framework that leverages reference single-cell RNA-seq data to construct gene panels optimized for specific experimental objectives. Through its specialized architecture with differentiable feature selection layers, the method identifies a compact set of highly informative genes that effectively capture transcriptome-wide expression patterns and preserve biologically meaningful cellular variation.

The framework consists of two sequential stages (Fig. 1). In the first stage, scGPD reconstructs transcriptome-wide expression profiles from a reduced set of candidate genes. Rather than evaluating genes independently, the model employs a correlation-aware binary gating mechanism that identifies interdependent gene groups using a Gaussian copula to model gene dependencies [19]. This gating mechanism utilizes the Binary Concrete distribution [20] to select the most informative representative from each correlated group while eliminating redundant genes, resulting in a narrowed gene pool containing the most salient candidate features. To appropriately match the distributional characteristics of scRNA-seq data, scGPD employs Poisson loss as its objective function [21–24], which handles the count-based nature of single-cell expression data. In the second stage, scGPD further refines the gene selection by choosing exactly $k$ genes from the reduced candidate set (of size $d$) obtained from the first stage. This refinement again employs a binary gating mechanism, but now uses masks derived from Concrete distributions [20] to enforce a strict selection budget of exactly $k$ genes. To ensure that the final gene panel is optimized for specific biological applications, the model employs task-specific loss functions tailored to downstream objectives, such as cell-type classification or spatial expression prediction.

### Poisson loss function

Poisson loss is a suitable choice for modeling scRNA-seq data due to the distributional properties of gene expression counts [21–24].

This characteristic makes Poisson loss particularly well suited for predicting transcriptome-wide expression from a selected subset of genes, as it captures the inherent statistical properties of scRNA-seq data while remaining computationally efficient.

Given observed gene expression counts $y_i$ and predicted values $\hat{y}_i$, the Poisson loss is defined as

$$\ell^{(i)} = \hat{y}_i - y_i \log \hat{y}_i + \log(y_i!)$$

In scGPD, Poisson loss is used as the default objective in the first training stage, where the goal is to reconstruct transcriptome-wide expression and reduce the influence of uninformative genes. This loss function is well suited for modeling the count-based structure of scRNA-seq data, encouraging the model to prioritize genes that provide informative signals for transcriptome reconstruction. Poisson loss can incorporate cell-specific size factors, allowing the model to account for sequencing depth differences, thereby improving robustness to varying data coverage. While Poisson loss is the default, other loss functions such as mean squared error can also be used, depending on the specific characteristics of the dataset or the downstream application. This flexibility allows scGPD to adapt to a variety of modeling needs while maintaining its focus on compact, informative gene panel design.

### Feature selection layers

Our methodology implements a two-stage feature selection framework. In the first stage, a correlated binary gating mechanism is used to filter out redundant and uninformative genetic markers. This design directs computational resources toward salient features by applying learned correlated binary gates (Supplementary Simulation Study, Figure S10). Rather than presuming selection independence, we characterize the gate vector distribution $B_i$ using a correlated binary concrete formulation, expressed as $B_i \sim \text{BinConcrete}(\beta_i, \tau)$, where $\beta_i$ denotes non-normalized logarithmic probabilities [20]. These probabilities are formulated via a Gaussian copula [19], introducing interdependencies between binary variables.

A Gaussian copula constitutes a multivariate cumulative distribution function for random variables $U_1, \ldots, U_p$, defined across the unit hypercube $[0,1]^p$ with uniform marginal distributions where $U_k \sim \text{Uniform}(0,1)$ for all $k \in [p]$. Given correlation matrix $R \in [-1,1]^p$ encoding the feature correlation structure of $X$, the Gaussian copula is formally defined as

$$C_R(U_1, \ldots, U_p) = \Phi_R\left(\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_p)\right), \tag{1}$$

where $\Phi_R$ represents the joint cumulative distribution function of a multivariate Gaussian with zero mean and correlation matrix $R$, while $\Phi^{-1}$ denotes the inverse cumulative distribution function of a standard univariate Gaussian.

The logarithmic probabilities $\beta_i$ are computed according to

$$\beta_i = \sigma\left(\frac{\log(U_i) + G_i}{\tau}\right), \tag{2}$$

where $\sigma(x)$ represents the sigmoid activation function, and $\tau$ constitutes the temperature hyperparameter. As $\tau$ approaches zero, the distribution converges to a Bernoulli distribution, whereas increasing $\tau$ toward infinity renders the distribution increasingly continuous.

The utilization of a Gaussian copula ensures that the gate vector distribution $B_i$ accurately captures the correlation structure among input features, enabling the model to account for gate interdependencies during genetic marker selection.

From a biological perspective, the correlated binary gating mechanism can be viewed as selecting representative genes from groups of co-expressed or functionally related markers. In many biological contexts, such correlated gene groups arise from shared regulatory programs, pathway membership, or cell-type-specific expression patterns. By explicitly modeling dependencies among genes, scGPD avoids selecting multiple redundant markers from the same correlated group and instead prioritizes a minimal set of representative genes that collectively capture the underlying biological signal.

The output from the correlated binary gates is determined by the Hadamard product $\mathbf{b} \odot \mathbf{x}$, where $\mathbf{x}$ represents the input vector and $\mathbf{b} = [b_1, \ldots, b_p] \in \mathbb{R}^p$ comprises samples $b_i$ drawn from the respective random variables $B_i \sim \text{BinConcrete}(\beta_i, \tau)$. This product subsequently traverses a neural architecture $f_\theta$ to predict gene expression counts $\mathbf{y}$. Genetic marker elimination occurs when the model converges toward low values of $\beta_i$, a process incentivized through the incorporation of a regularization term on BinConcrete samples within the objective function:

$$\min_{\theta,\beta} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{b}}\left[\ell\left(f_\theta(\mathbf{b} \odot \mathbf{x}), \mathbf{y}\right) + \lambda \mathbf{1}^\top \mathbf{b}\right]. \tag{3}$$

The regularization component $\lambda \mathbf{1}^\top \mathbf{b}$ penalizes excessive feature selection, where hyperparameter $\lambda > 0$ acts as a tuning parameter that balances predictive accuracy against feature sparsity, and $\ell$ denotes the Poisson loss function.

In the second stage, after constraining the candidate pool to size $d$, we again apply a binary mask to select exactly $k$ genetic markers from the $d$ candidates. The mask is generated via the element-wise maximum of $k$ Concrete random variables, denoted as $\mathbf{A}_i \sim \text{Concrete}(\alpha_i, \tau)$ for $i = 1, \ldots, k$. Each Concrete distribution is parameterized by non-normalized probabilities $\alpha_i \in \mathbb{R}_+^d$ and temperature parameter $\tau > 0$, with each distribution asymptotically approaching a multinomial distribution with probabilities defined by $\frac{\alpha_i}{\mathbf{1}^\top \alpha_i}$ as $\tau$ approaches zero [20].

The binary mask layer processes normalized gene expression profiles $\mathbf{x} \in \mathbb{R}^d$, yielding the Hadamard product $\mathbf{a} \odot \mathbf{x}$, where $\mathbf{a} = \max_i \mathbf{a}_i \in \mathbb{R}^d$ constitutes the element-wise maximum of samples $\mathbf{a}_i$ drawn from each concrete distribution random variable. This product subsequently traverses a neural architecture $f_\gamma$, which predicts the target variable $\mathbf{y}$ based on $\mathbf{a} \odot \mathbf{x}$. The model parameters are optimized according to

$$\min_{\theta,\alpha} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{a}}\left[l\left(f_\gamma(\mathbf{a} \odot \mathbf{x}), \mathbf{y}\right)\right], \tag{4}$$

where $l$ represents an application-specific loss function.

## Results

### scGPD enables accurate cell-type annotation and scRNA-seq expression profiles reconstruction

We evaluated scGPD on three scRNA-seq datasets derived from distinct biological sources: pancreas [25], heart [26], and peripheral blood mononuclear cells [4]. Starting from an initial set of 5000 highly

variable genes and the gene covariance estimated from CS-CORE [27], we applied scGPD and three baseline gene selection methods: PERSIST [17], scGIST [16], and geneBasis [28] to identify gene panels ranging in size from 32 to 256 genes, a range that encompasses the majority of panel sizes used in FISH experiments.

As a flexible analytical framework, scGPD supports both supervised and unsupervised learning paradigms, making it applicable to a wide range of downstream tasks. To evaluate its effectiveness, we first assessed how well-selected gene panels could reconstruct the original scRNA-seq expression profiles. Specifically, we computed Pearson correlation coefficients [29] between the original and reconstructed gene expression matrices across genes for each cell. To ensure a fair comparison, we trained the same multilayer perceptron (MLP) network architecture to map each method's selected gene panels to the full expression profiles. Across all datasets and panel sizes, scGPD consistently achieved higher reconstruction accuracy than competing methods (Fig. 2A). An ablation analysis further demonstrates that explicitly modeling gene–gene correlations contributes substantially to this improvement, with correlation-aware scGPD outperforming its noncorrelation variant across gene panel sizes (Figure S2). These results underscore the robustness and reliability of scGPD for guiding targeted sequencing.

We further evaluated the utility of selected gene panels in a cell-type classification task, a key downstream analysis in single-cell transcriptomics. For each method, we trained identical MLP classifiers using the selected gene panels and assessed classification accuracy across varying panel sizes. As expected, classification performance generally improved with increasing panel size across all methods. Overall, scGPD achieved classification performance that was comparable with or better than existing methods and consistently maintained competitive accuracy across different panel sizes (Fig. 2B, Supplementary Figure S1). As with reconstruction accuracy, absolute performance differences across methods are relatively small, particularly at larger panel sizes where accuracy approaches saturation, indicating broadly comparable performance when accuracy is the primary metric. Nevertheless, scGPD shows stable, consistently strong performance across a wide range of panel sizes.

We further visualized UMAP embeddings derived from the raw data and from the 256-gene panel selected by scGPD. Cells cluster according to annotated cell types, demonstrating concordance between the selected gene panel and the underlying transcriptional structure (Fig. 2C). While similar patterns can be observed for several methods with sufficiently large panels, scGPD preserves clear cell-type separation using a compact gene set, highlighting its practical utility under panel size constraints. Collectively, these results indicate that scGPD is a robust and versatile approach for gene panel selection, with modest but consistent advantages across tasks and panel sizes.

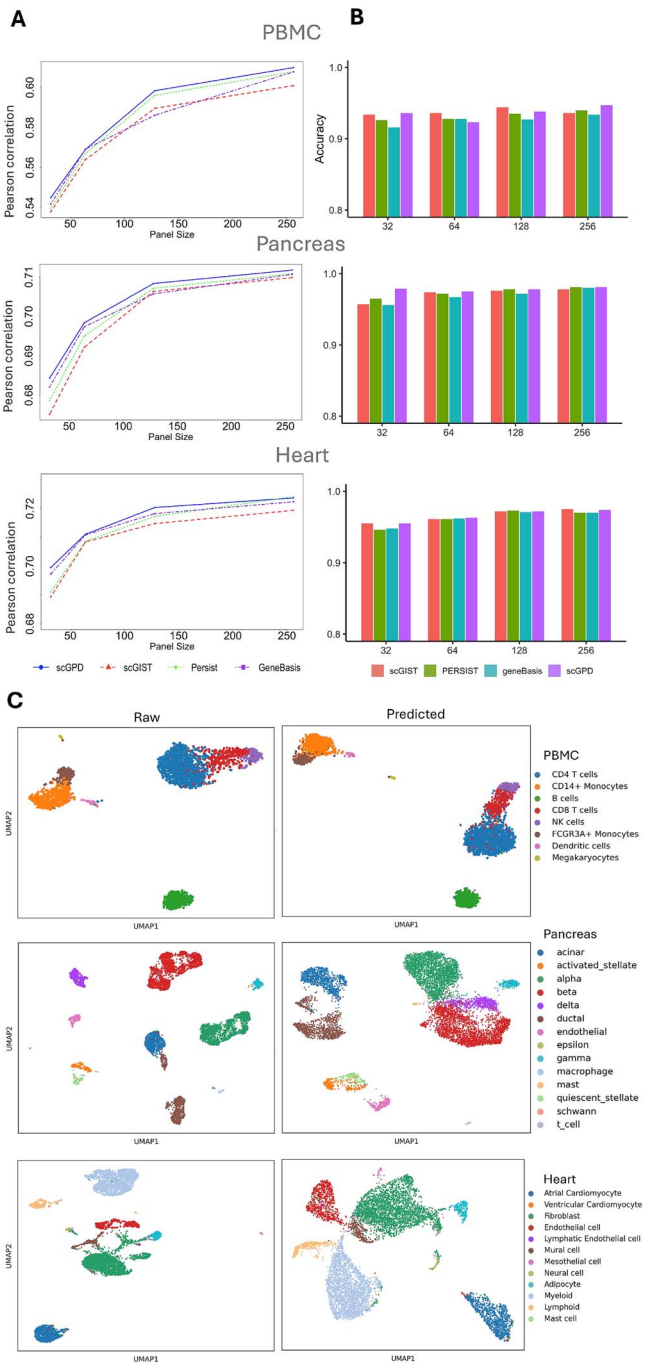



**Figure 2** Performance of various methods on three different datasets. A. Pearson correlation coefficients for variable panel sizes. B. Accuracies for variable panel sizes. C. UMAP plots using the gene panel of size 256 selected by scGPD for the three datasets.

## Performance evaluation on a spatial transcriptomics dataset

scGPD can identify informative marker genes for a variety of experimental objectives. Here, we applied the genes selected using scRNA-seq to data collected from ST studies. We analyzed a seqFISH+ dataset [12] from mouse olfactory, where a corresponding scRNA-seq dataset [30] is also available. To ensure consistency, we restricted our analysis to the 9913 common genes present in both datasets. We then used scGPD to generate gene panels from the scRNA-seq data and predicted cell types in the seqFISH+ dataset based on these panels. The seqFISH+ dataset's preexisting cell-type annotations served as ground truth labels, enabling quantitative assessment of how effectively scGPD-selected gene panels preserve cell-type discrimination when transferred from single-cell to ST contexts.

We benchmark the performance of scGPD with the other three methods: scGIST, geneBasis, and PERSIST. Although scGPD shows slightly lower performance for smaller panels, when panel size increases, scGPD outperforms the other methods, achieving the

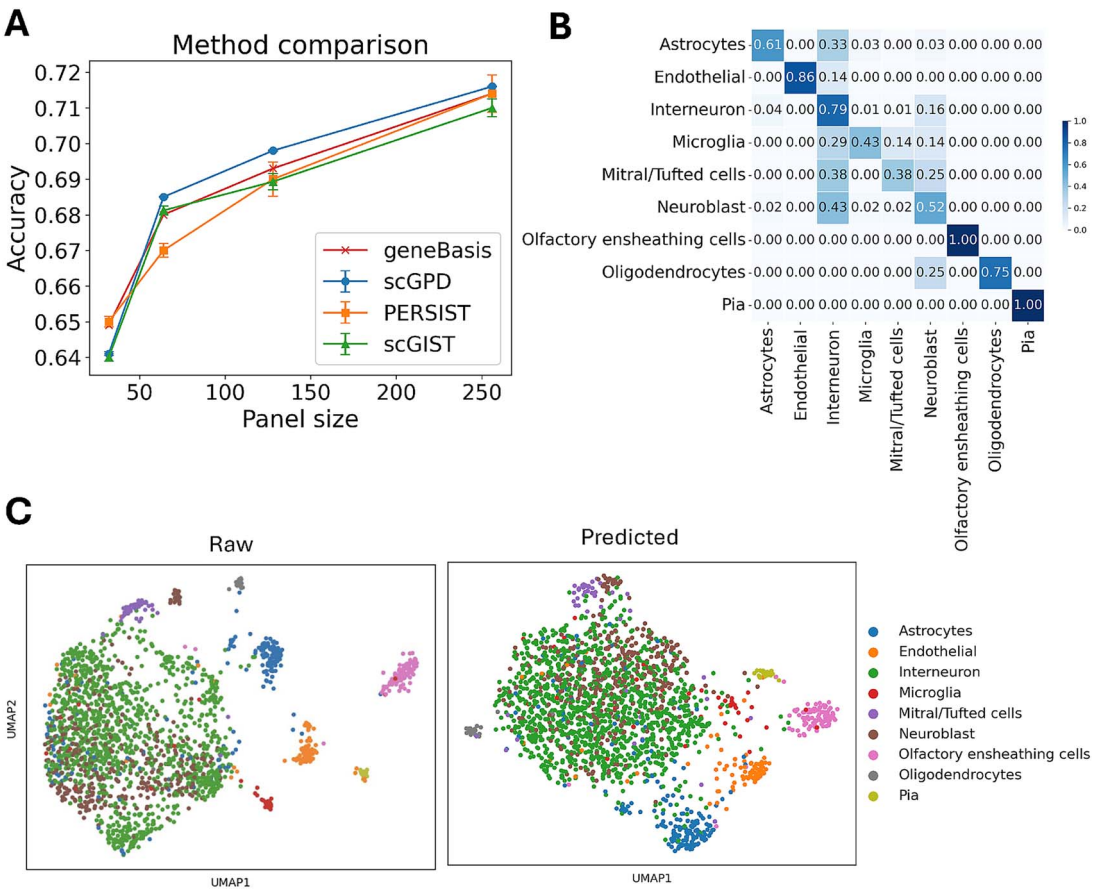



**Figure 3** Performance evaluation of gene panel selection on an ST dataset. A. Cell-type classification accuracy using gene panels designed from scRNA-seq data and evaluated on the seqFISH+ test set. B. Confusion matrix of cell-type predictions by scGPD using a panel of 256 genes, evaluated on the seqFISH+ test dataset. C. UMAP visualization of the seqFISH+ data derived based on raw data and the 256-gene panel selected by scGPD, demonstrating alignment between predicted cell types and the original clustering structure.

highest accuracy at larger panel sizes (Fig. 3A). PERSIST and scGIST exhibit similar trends and GeneBasis is the most competitive baseline method. Figure 3B presents the confusion matrix of predicted cell by scGPD using a 256-gene panel, evaluated on the seqFISH+ test dataset.

Supplementary Figure S3 presents the PAGA similarity [31] scores for varying-size gene panels (32 to 256 genes), evaluating the ability of each method to preserve the structure of cell–cell similarity. Across all panel sizes, scGPD demonstrates consistently strong performance, maintaining high PAGA similarity scores and outperforming scGIST and PERSIST at all panel sizes. This pattern underscores the ability of scGPD to select gene panels that faithfully retain the structure of the cellular manifold.

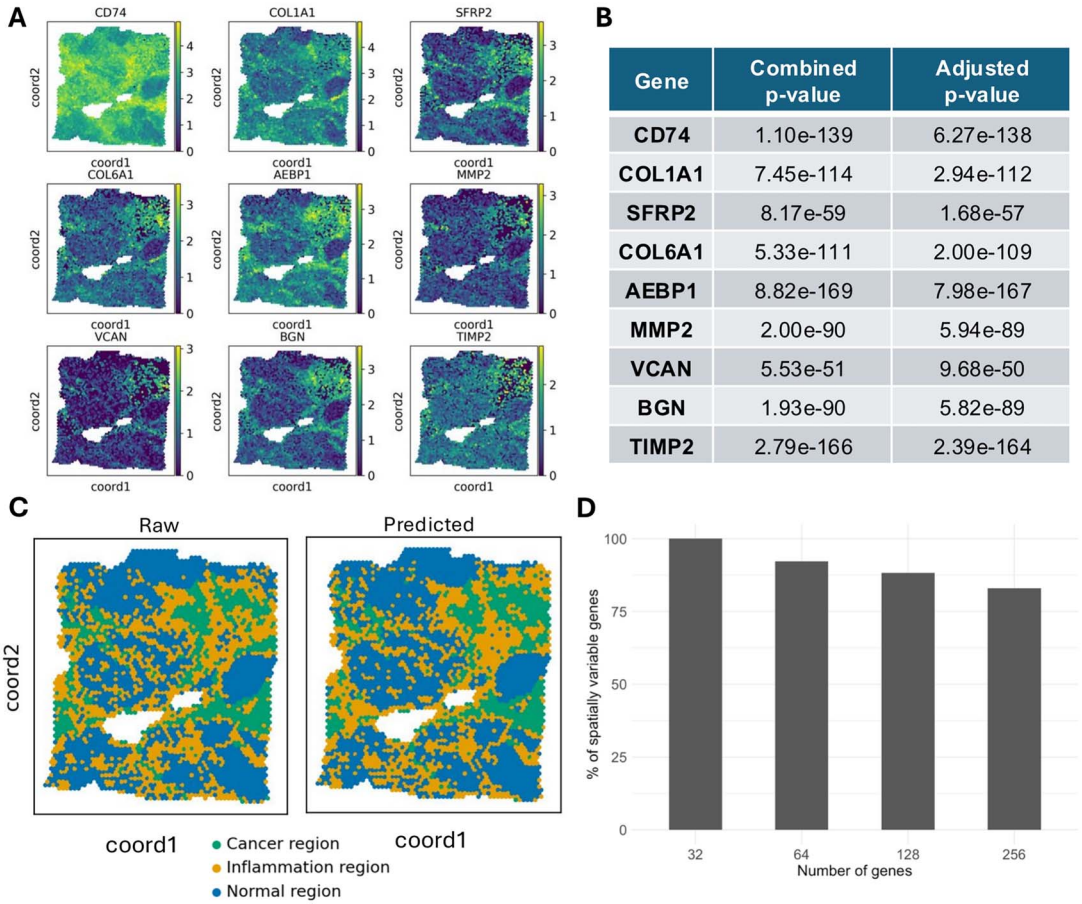
Using the 256 genes identified from scGPD, we show a UMAP projection derived based on raw data and selected gene set to compare the cell-type distributions (Fig. 3C). The predicted UMAP maintains the overall structure of the raw dataset, demonstrating that our model successfully captures the differences among cell types. Together, these results illustrate the effectiveness of scGPD in selecting informative gene panels that generalize well to ST data.

## Identifying informative genes reveals spatially distinct expression patterns in spatial transcriptomics data

Our proposed method, scGPD, effectively identifies genes with spatially distinct expression patterns in ST data. Using a human breast cancer 10x Genomics scRNA-seq dataset [32] as a reference, we applied scGPD-identified genes to data from breast cancer tissue [32] profiled by 10x Visium to explore their role in dissecting the tumor microenvironment. The selected genes exhibit clear spatial variation, with specific regions showing elevated expression, indicating their significance in defining distinct tissue microenvironments and potentially influencing tumor heterogeneity (Fig. 4A).

To statistically evaluate the spatial variability of these genes, we employed SPARK [21], a robust statistical framework for detecting spatially expressed genes. The analysis (Fig. 4B, Supplementary Tables S1–S4) revealed highly significant combined and adjusted *P*-values, confirming the meaningful spatial expression of the identified genes and reinforcing the robustness of scGPD in capturing genes relevant to spatial tissue organization.

Importantly, the genes identified, such as CD74, COL1A1, SFRP2, and COL6A1, have well-established biological relevance in breast cancer. CD74 promotes triple-negative breast cancer progression by expanding immunosuppressive cells [33]. COL6A1 facilitates breast cancer growth and spread by supporting tumor cell proliferation and creating a favorable tumor environment and COL1A1 is a structural protein in the extracellular matrix and is involved in cancer spread and metastasis [34, 35]. The spatial expression patterns of these genes likely reflect critical biological processes in the tumor microenvironment, such as immune cell localization, stromal activation, and epithelial-mesenchymal transitions. Together, these findings




| Gene | Combined p-value | Adjusted p-value |
|---|---|---|
| CD74 | 1.10e-139 | 6.27e-138 |
| COL1A1 | 7.45e-114 | 2.94e-112 |
| SFRP2 | 8.17e-59 | 1.68e-57 |
| COL6A1 | 5.33e-111 | 2.00e-109 |
| AEBP1 | 8.82e-169 | 7.98e-167 |
| MMP2 | 2.00e-90 | 5.94e-89 |
| VCAN | 5.53e-51 | 9.68e-50 |
| BGN | 1.93e-90 | 5.82e-89 |
| TIMP2 | 2.79e-166 | 2.39e-164 |

**Figure 4** Spatial validation and functional relevance of scGPD-selected genes in breast cancer tissue. A. Spatial expression patterns of representative scGPD-identified genes in 10x Visium breast cancer ST data, showing clear spatially distinct expression domains. B. Statistical validation of spatial variability using SPARK, with significant adjusted *P*-values confirming spatial expression of selected genes. C. Tissue segmentation using scGPD-selected genes compared with histological annotations from transcriptome-wide profiles, demonstrating strong concordance and preservation of tumor architecture. D. Bar plot showing the proportion of spatially variable genes retained across increasing gene panel sizes.

underscore the ability of scGPD to identify genes that are not only statistically significant in their spatial distribution but also biologically meaningful in the context of cancer progression and tissue heterogeneity. We also evaluated the impact of gene panel size on spatial variability detection. By systematically varying panel sizes from 32 to 256, we assessed how the proportion of spatially variable genes changes (Fig. 4D, Figure S9). The results indicate that even with an expanded gene set, scGPD retains at least 80% of spatially variable genes (adjusted *P*-value $<.05$), demonstrating its ability to efficiently prioritize informative genes while maintaining biological interpretability.

To further assess the functional utility of the identified genes, we evaluated their ability to preserve large-scale spatial organization in an unsupervised tissue segmentation task. We first constructed a reference tissue segmentation by identifying cell-type–associated gene signatures from the matched single-cell RNA-seq data via differential expression analysis, projecting these signatures onto the ST data to obtain spot-level cell-type scores, and applying K-means clustering to partition the tissue into three major spatial regions. This reference segmentation serves as a proxy for large-scale tissue domains rather than a supervised ground truth. Using the same unsupervised clustering procedure, we then performed spatial segmentation based only on the 256 genes selected by scGPD (trained in a supervised manner to select genes that are informative for cell-type discrimination) and compared the resulting spatial partitions with the reference annotations. We additionally included a simple and interpretable baseline based on spatially variable genes with minimal bivariate spatial association by selecting 256 genes using Moran's I statistic and repeating the same segmentation and evaluation procedure; in contrast to this baseline, which yields less spatially coherent segmentation (Figure S5), the strong concordance between the scGPD-based segmentation and the reference domains (Fig. 4C), quantified by an adjusted Rand index of 0.63, demonstrates that scGPD effectively preserves biologically meaningful spatial organization under substantial gene panel reduction.

These findings underscore the power of scGPD in optimizing gene selection for ST, enhancing tissue characterization, and enabling deeper insights into spatial gene expression. By identifying genes with strong spatial signals, scGPD provides a scalable and efficient solution for uncovering the spatial organization of complex tissues.

## Discovery of disease-related genes for lung adenocarcinoma

scGPD is a powerful and versatile tool for uncovering disease-related genes by systematically analyzing scRNA-seq data from samples under different conditions. By leveraging its ability to distinguish

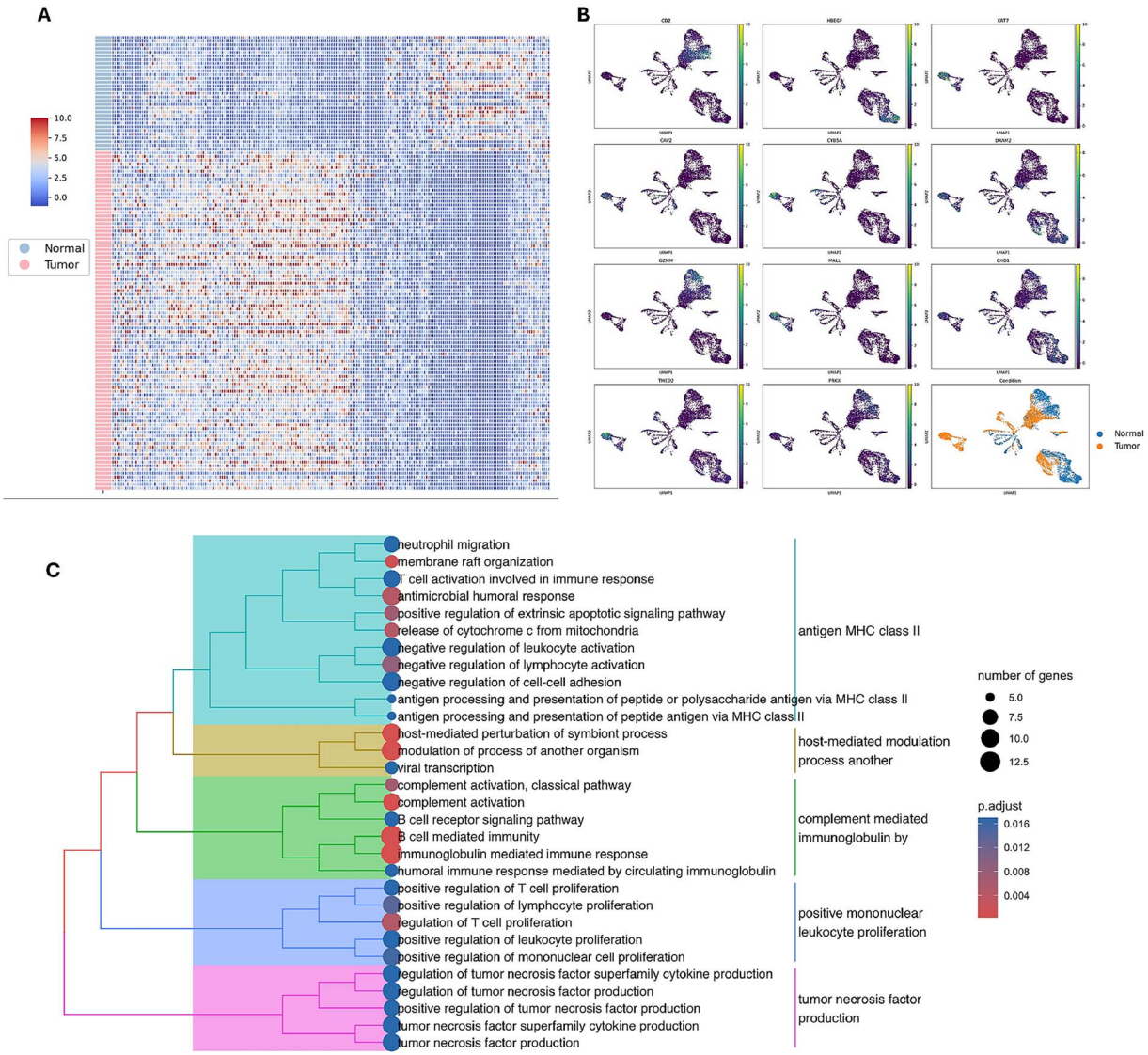



**Figure 5** Identification of LUAD-related genes and their expression dynamics across cell types. A. Heatmap showing DEGs between HC and LUAD samples identified by scGPD (rows represent individual cells and columns represent genes, grouped by sample condition). B. UMAP projections of single-cell transcriptomes colored by gene expression and sample condition. C. GO enrichment analysis of scGPD identified genes.

gene expression variations at the single-cell level, scGPD enables the identification of key molecular signatures associated with disease progression. This capability is particularly valuable for studying complex diseases where cellular heterogeneity plays a crucial role in pathogenesis.

In this study, we apply scGPD to lung adenocarcinoma (LUAD), a highly prevalent and aggressive subtype of non-small cell lung cancer (NSCLC). LUAD is characterized by tumor heterogeneity and clinical symptoms such as persistent coughing, chest pain, and respiratory decline. Understanding its underlying genetic risk factors is essential for unraveling its molecular mechanisms and developing targeted therapeutic strategies.

To comprehensively investigate LUAD-specific gene expression patterns, we analyzed a single-cell RNA-seq dataset [36] comprising LUAD and healthy control (HC) samples. Figure 5A presents a heatmap of gene expression across normal and tumor samples, revealing key differentially expressed genes (DEGs) that may serve as LUAD biomarkers [37–39]. Because this single-cell dataset contains explicit condition labels indicating whether each cell originates from a LUAD or HC sample, scGPD was applied in a supervised setting in this analysis. Using scGPD in a supervised setting, we identified disease-related genes by training the model to distinguish between HC and LUAD samples based on single-cell transcriptomic profiles. Specifically, scGPD was optimized using a standard cross-entropy loss on the LUAD versus HC labels, such that gene selection was driven by their contribution to accurate disease-status classification while accounting for gene–gene correlation. The genes selected by scGPD were those most informative for accurately classifying cells by disease status. For example, among the highlighted genes, FABP4 modulates lipid metabolism and immune responses by influencing NK and macrophage function [40]; MMP7 facilitates tumor proliferation and metastasis through matrix degradation and angiogenesis [41]; and C1QB contributes to immune microenvironment regulation via the complement cascade [42]. These findings provide insights into tumor microenvironment heterogeneity and highlight potential driver genes contributing to LUAD pathogenesis.

The UMAP plot in the bottom right of Fig. 5B shows the distribution of diseased conditions. The remaining UMAP projections in Fig. 5B and Supplementary Figure S7 illustrate the spatial distribution of these genes across different cell clusters, demonstrating their selective enrichment in specific LUAD subpopulations. These complementary visualizations collectively demonstrate the ability of scGPD to robustly classify disease-associated genes.

Our results demonstrate that scGPD could successfully identify genes with significantly higher expression in LUAD-associated cell types compared with their healthy counterparts, as visualized in Fig. 5B. These expression patterns confirm the genes' specificity to the tumor microenvironment and suggest their potential roles in LUAD development and progression. To understand the biological significance of these identified genes, we performed GO enrichment analysis on the scGPD-selected gene panel. The analysis reveals that these genes are enriched in tumor-specific pathways and functionally associated with oncogenic signaling cascades, immune evasion mechanisms, and metastatic progression (Fig. 5C). Compared with the low-rank baseline (Figure S6), scGPD yields gene panels with more coherent and tumor-relevant functional profiles, reinforcing the link between improved robustness and biological interpretability.

Beyond gene identification, scGPD offers a robust framework for biomarker discovery and precision medicine applications. The genes identified through scGPD provide valuable candidate targets for early diagnosis, patient stratification, and therapeutic intervention. Moreover, additional genes identified by scGPD, which are not shown here, can be found in Supplementary Figure S6, providing an expanded repository of potential LUAD biomarkers and therapeutic targets.

These findings emphasize the critical role of scGPD can play in discovering disease-associated genes and advancing our understanding of LUAD pathogenesis. By efficiently identifying key molecular drivers, scGPD enables deeper insights into the complex transcriptional landscape of cancer biology.

## Discussion

We introduced scGPD, a deep learning framework for gene panel design in targeted ST using scRNA-seq as a reference. By explicitly modeling gene–gene correlations through a correlation-aware gating mechanism, scGPD selects compact, nonredundant gene panels that capture diverse biological programs within a limited panel size.

Across multiple benchmarks, scGPD demonstrates strong performance across panel sizes and task settings. Genes selected from single-cell data transfer robustly to spatial contexts, preserving coherent and biologically meaningful spatial expression patterns. Compared with existing approaches that treat genes as independent features, scGPD reduces redundancy among correlated genes and yields more informative panels, which is particularly important for FISH-based spatial technologies with constrained target capacity.

In addition, scGPD is flexible with respect to analysis objectives. Through supervised training, the framework can prioritize genes associated with specific phenotypes, such as disease states or functional cellular properties, making it suitable for translational and clinical applications including spatial pathology and targeted biomarker discovery.

We note that scGPD is primarily designed to optimize gene panel selection at the gene level, with objectives defined on aggregate gene expression patterns derived from single-cell RNA-seq data. As such, scGPD does not explicitly model subcellular RNA localization, transcript isoform diversity, or single-molecule spatial distributions within individual cells. For imaging-based ST experiments that aim to resolve RNA localization at subcellular or single-molecule resolution, additional factors, such as probe design constraints, transcript length, local RNA density, optical crowding, and subcellular compartmentalization, may play a critical role and are not directly accounted for in the current framework.

Accordingly, while scGPD is well suited for identifying gene panels informative for cell-type composition, tissue architecture, and spatial domain analysis, it may be less appropriate for studies focused on fine-scale intracellular RNA organization or molecular trafficking. Extending the framework to incorporate subcellular spatial priors, isoform-level information, or imaging-specific constraints represents an important direction for future methodological development.

Future work will include experimental validation of scGPD-designed panels on spatial profiling platforms, particularly multiplexed FISH-based methods [43], as well as integration of spatial priors and multiomic information to further enhance panel robustness and generalizability [44–46].

In summary, scGPD provides an interpretable and flexible solution to gene panel design that bridges single-cell and ST, enabling efficient and targeted exploration of tissue organization and disease-relevant biology.

**Key Points**

- scGPD is a deep learning-based framework for gene panel design that leverages single-cell RNA-seq data to identify compact and nonredundant gene sets for spatial profiling.
- Its correlation-aware gating mechanism extracts informative and diverse genes while effectively eliminating redundancy.
- scGPD achieves outstanding performance in transcriptome recovery and cell-type classification, and its flexible framework can be adapted to prioritize genes relevant to specific diseases or phenotypes.

## Author contributions

Yunshan Guo (Conceptualization, Investigation, Methodology, Writing—original draft), Jia Zhao (Conceptualization, Investigation, Methodology, Writing—original draft), Rui B. Chang (Writing—review & editing), and Hongyu Zhao (Conceptualization, Investigation, Methodology, Supervision, Writing—original draft)

## Supplementary material

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

## Conflicts of interest

None declared.

## Funding

This work was supported in part by National Institutes of Health grants U01 HG013840 and U24 HG012108.

## Data availability

We provide a summary of the sources and statistics for all datasets used in Supplementary Table S5. All datasets are accessible through the links included in this file. The codes of scGPD are available at https://github.com/TinaGuo/scGPD. We follow the MIT license for usage.

## References


1. Jovic D, Liang X, Zeng H *et al.* Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694.
2. Macosko EZ, Basu A, Satija R *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14.
3. Tang F, Barbacioru C, Wang Y *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**:377–82.
4. Zheng GX, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049.
5. Du J, Yang Y-C, An Z-J *et al.* Advances in spatial transcriptomics and related data analysis strategies. *J Transl Med* 2023;**21**:330.
6. Liang G, Yin H, Ding F. Technical advances and applications of spatial transcriptomics. *GEN Biotechnol* 2023;**2**:384–98.
7. Femino AM, Fay FS, Fogarty K *et al.* Visualization of single RNA transcripts in situ. *Science* 1998;**280**:585–90.
8. Codeluppi S, Borm LE, Zeisel A *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* 2018;**15**:932–5.
9. Chen KH, Boettiger AN, Moffitt JR *et al.* Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**:aaa6090.
10. 10x Genomics. *Visium Spatial Gene Expression* 2020. Available at: https://www.10xgenomics.com/spatial-transcriptomics (3 April 2026, date last accessed).
11. Vickovic S, Eraslan G, Salmén F *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90.
12. Eng C-HL, Lawson M, Zhu Q *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqfish+. *Nature* 2019;**568**:235–9.
13. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemom Intel Lab Syst* 2005;**78**:103–12.
14. Katrutsa A, Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst Appl* 2017;**76**:1–11.
15. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons; 2005.
16. Yafi MA, Hisham MHH, Grisanti F *et al.* scGIST: gene panel design for spatial transcriptomics with prioritized gene sets. *Genome Biol* 2024;**25**:57.
17. Covert I, Gala R, Wang T *et al.* Predictive and robust gene selection for spatial transcriptomics. *Nat Commun* 2023;**14**:2091.
18. Lee C, Imrie F, van der Schaar M. Self-supervision enhanced feature selection with correlated gates. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Conference: OpenReview; April 25–29, 2022.
19. Nelsen RB. *An Introduction to Copulas*. New York: Springer; 2006.
20. Maddison CJ, Mnih A, Teh YW. The concrete distribution: a continuous relaxation of discrete random variables. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France: OpenReview; 2017.
21. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* 2020;**17**:193–200.
22. Cable DM, Murray E, Zou LS *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022a;**40**:517–26.
23. Cable DM, Murray E, Shanmugam V *et al.* Cell type-specific inference of differential expression in spatial transcriptomics. *Nat Methods* 2022b;**19**:1076–87.
24. Wang G, Zhao J, Yan Y *et al.* Construction of a 3D whole organism spatial atlas by joint modelling of multiple slices with deep neural networks. *Nat Mach Intell* 2023;**5**:1200–13.
25. Baron M, Veres A, Wolock SL *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst* 2016;**3**:346–60.
26. Litviňuková M, Talavera-López C, Maatz H *et al.* Cells of the adult human heart. *Nature* 2020;**588**:466–72.
27. Su C, Xu Z, Shan X *et al.* Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nat Commun* 2023;**14**:4846.
28. Missarova A, Jain J, Butler A *et al.* geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq. *Genome Biol* 2021;**22**:1–22.
29. Sedgwick P. Pearson's correlation coefficient. *BMJ* 2012;**345**:e4483.
30. Brann DH, Tsukahara T, Weinreb C *et al.* Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. *Sci Adv* 2020;**6**:eabc5801.
31. Wolf FA, Hamey FK, Plass M *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**:1–9.
32. Wu SZ, Al-Eryani G, Roden DL *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* 2021;**53**:1334–47.
33. Pellegrino B, David K, Rabani S *et al.* CD74 promotes the formation of an immunosuppressive tumor microenvironment in triple-negative breast cancer in mice by inducing the expansion of tolerogenic dendritic cells and regulatory B cells. *PLoS Biol* 2024;**22**:e3002905.
34. Li X, Li Z, Gu S *et al.* A pan-cancer analysis of collagen vi family on prognosis, tumor microenvironment, and its potential therapeutic effect. *BMC Bioinformatics* 2022;**23**:390.
35. Yin P, Bai Y, Wang Z *et al.* Non-canonical Fzd7 signaling contributes to breast cancer mesenchymal-like stemness involving Col6a1. *Cell Commun Signal* 2020;**18**:1–13.
36. Kim N, Kim HK, Lee K *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;**11**:2285.
37. Myers DJ, Wallen JM. Lung adenocarcinoma. In: *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing; 2023.
38. Zhou H, Zhang Y, Liu J *et al.* Education and lung cancer: a Mendelian randomization study. *Int J Epidemiol* 2019;**48**:743–50.
39. Sun X, Yi J, Yang J *et al.* An integrated epigenomic-transcriptomic landscape of lung cancer reveals novel methylation driver genes of diagnostic and therapeutic relevance. *Theranostics* 2021;**11**:5346.
40. Tian H, Li K, Chen J *et al.* Impaired natural killer cell maturation in lung adenocarcinoma driven by FABP4 and SPON2 downregulation through disrupted lipid metabolism. *Transl Lung Cancer Res* 2025;**14**:1660.
41. Duffy M, McCarthy K. Matrix metalloproteinases in cancer: prognostic markers and targets for therapy. *Int J Oncol* 1998;**12**:1343–51.

42. Mamidi S, Höne S, Kirschfink M. The complement system in cancer: ambivalence between tumour destruction and promotion. *Immunobiology* 2017;**222**:45–54.
43. Merritt CR, Ong GT, Church SE *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol* 2020;**38**: 586–99.
44. Arora R, Cao C, Kumar M *et al.* Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nat Commun* 2023; **14**:5029.
45. Kleshchevnikov V, Shmatko A, Dann E *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:661–71.
46. Mimitou EP, Lareau CA, Chen KY *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* 2021;**39**:1246–58.


*Briefings in Bioinformatics*, 2026, **27**, bbag160
https://doi.org/10.1093/bib/bbag160
Problem Solving Protocol