OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae655
https://doi.org/10.1093/bib/bbae655
Problem Solving Protocol

scPAS: single-cell phenotype-associated subpopulation identifier

Aimin Xie[1,‡], Hao Wang[1,‡], Jiaxu Zhao[1,‡], Zhaoyang Wang[2], Jinyuan Xu[1,*], Yan Xu[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 157 Baojian Road, Heilongjiang 150081, China
[2]Genetron Health (Beijing) Co. Ltd, 1-2/F, Building 11, Zone 1, 8 Life Science Parkway, Changping District, Beijing 102208, China

[*]Corresponding authors. Jinyuan Xu, College of Bioinformatics Science and Technology, Harbin Medical University, 157 Baojian Road, Harbin, Heilongjiang 150081, China. E-mail: xujinyuan@hrbmu.edu.cn. X: @jinyuanxu; Yan Xu, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 157 Baojian Road, Heilongjiang 150081, China. E-mail: xuyan@ems.hrbmu.edu.cn. X: @yanxu
[‡]Aimin Xie, Hao Wang and Jiaxu Zhao contributed equally to this work.

Abstract

Despite significant advancements in single-cell sequencing analysis for characterizing tissue sample heterogeneity, identifying the associations between cell subpopulations and disease phenotypes remains a challenging task. Here, we introduce scPAS, a new bioinformatics tool designed to integrate bulk data to identify phenotype-associated cell subpopulations within single-cell data. scPAS employs a network-regularized sparse regression model to quantify the association between each cell in single-cell data and a phenotype. Additionally, it estimates the significance of these associations through a permutation test, thereby identifying phenotype-associated cell subpopulations. Utilizing simulated data and various single-cell datasets from breast carcinoma, ovarian cancer, and atherosclerosis, as well as spatial transcriptomics data from multiple cancers, we demonstrated the accuracy, flexibility, and broad applicability of scPAS. Evaluations on large datasets revealed that scPAS exhibits superior operational efficiency compared to other methods. The open-source scPAS R package is available at GitHub website: https://github.com/aiminXie/scPAS.

Keywords: single-cell; phenotype; cancer; spatial transcriptomics; data integration

Introduction

Single-cell RNA sequencing (scRNA-seq), with its ability to comprehensively characterize cells from complex tissues, has become an essential method in biological research [1]. Unlike bulk RNA-seq, scRNA-seq allows for the detailed cataloging of cell types, states, and lineages within heterogeneous tissue ecosystems [2, 3]. Identifying key cell subpopulations associated with specific phenotypes has become indispensable in disease research, particularly with advancements enabled by single-cell technology [4–6]. Despite significant progress in developing computational methods for different stages of scRNA-seq analysis, a bottleneck remains in the accurate identification of molecular relationships between phenotypes and cell populations. Large-scale cohort data with sufficient phenotypic information are scarce because of the high cost and labor associated with single-cell sequencing [7, 8]. Consequently, small-scale datasets have limited statistical power, making it challenging to establish statistically significant associations between cell subpopulations and the phenotypes of interest.

Fortunately, large-scale cohorts of bulk sequencing data, enriched with clinical information, are available from publicly accessible databases such as The Cancer Genome Atlas (TCGA) [9] and the International Cancer Genome Consortium [10]. As an alternative, many researchers leverage the available bulk RNA-seq data from patients to link scRNA-seq-derived cell subpopulations with clinical phenotypes by employing deconvolution or feature gene scoring methods [11, 12]. However, these approaches assess the association of disease phenotypes with predetermined cell clusters rather than with individual cells. This dependency on clustering results neglects transcriptional changes within cell clusters.

Recently, computational methods such as Scissor [13], scAB [14], and DEGAS [15] have systematically identified cell subpopulations that are highly associated with specific phenotypes by integrating bulk RNA-seq data. Unlike conventional methods, these approaches couple bulk and scRNA-seq data through specific associations and construct computational models to directly identify phenotype-associated cell subpopulations. Scissor and scAB employ Pearson correlations at the whole-transcriptome level to quantify similarities between cells and bulk samples. Using the correlation coefficients of cells to samples as training features, they established associations between phenotypes and cells using a sparse regression model and a knowledge- and graphically guided matrix decomposition model, respectively. DEGAS, on the other hand, leverages a combination of deep learning and transfer learning to transfer phenotypic information from patients to cells. Although these methods are pioneering to some extent, they exhibit several limitations: (i) similarity matrices

Received: August 22, 2024. **Revised:** October 13, 2024. **Accepted:** December 3, 2024

at the whole-transcriptome level may overlook the expression changes of a few key genes. (ii) High spatiotemporal complexity limits the use of these methods on large-scale datasets. (iii) They cannot provide both quantitative and qualitative estimations of the strength of association between each cell in the scRNA-seq data and phenotype.

We introduce a novel tool named scPAS (Single-Cell Phenotype-Associated Subpopulation identifier), designed to quantitatively estimate the association strength between each cell in scRNA-seq data and a given phenotype. This is achieved through the construction of a network-regularized sparse regression model that integrates bulk RNA-seq data, phenotype information, and the gene–gene similarity network derived from single-cell data. Additionally, scPAS can assess the statistical significance of these associations, providing a qualitative classification based on this evaluation. By applying scPAS to simulated data and various single-cell datasets from breast carcinoma, ovarian cancer (OV), and atherosclerosis, as well as spatial transcriptomics (ST) data from multiple cancers, we demonstrated its accuracy, flexibility, and broad applicability. Its application to large single-cell datasets revealed the excellent operational efficiency of scPAS. An application example incorporating ST effectively demonstrates the transferability of the scPAS model, illustrating that scPAS can be seamlessly applied to integrate bulk data, scRNA-seq, and ST data in disease research. In summary, our results suggest that scPAS is an efficient tool for exploring and analyzing single-cell data and elucidating disease mechanisms.

Materials and methods

scPAS workflow

Input data and data preprocessing

As illustrated in Fig. 1, scPAS requires the input of a single-cell expression matrix, a bulk expression matrix, phenotype labels for bulk samples, and an optional set of genes of interest. Phenotype data should correspond to the samples in the bulk expression profile and can be continuous variables, binary variables, or clinical survival data. scPAS processes the single-cell data using the Seurat [16] pipeline to obtain results suitable for visualization and extracts log-transformed expression values for subsequent analysis. For bulk data, it is recommended to use a log-transformed Transcripts Per Kilobase Million (TPM) or Fragments per Kilobase Million (FPKM) matrix as the input data. To mitigate potential batch effects, scPAS preprocesses bulk data using quantile normalization and z-score normalization. Subsequently, the bulk data expression submatrix $B_{n\times t}$ and single-cell data submatrix $S_{m\times t}$ are extracted for model construction. This step can optionally utilize a gene set of interest. Highly variable genes from the single-cell data are used by default. If the single-cell data have been well characterized, we recommend using marker genes that are specifically expressed in different cell subpopulations or clusters as input.

Construction and optimization of the regression model

Next, scPAS trains and optimizes a network-regularized sparse regression model using the bulk expression profile $B_{n\times t}$. The optimization model is formulated as follows:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} -\frac{1}{n} l(\beta) + \lambda\left[\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\beta^T L\beta\right] \tag{1}$$

where β is the regression coefficient vector of genes, and $l(\beta)$ corresponds to the log-likelihood function. The specific form of $l(\beta)$ depends on the input phenotype Y: linear regression for continuous variables, logistic regression for binary variables, and Cox regression for clinical survival data (more details at next section).

The scPAS incorporates two penalty terms into the regression model. The first is a sparse penalty (L1-norm) applied to select high-confidence genes relevant to the given phenotype. The second is a network-based penalty that encourages similar coefficients for tightly connected nodes (genes) in the network, thereby enhancing the consistency and interpretability of phenotype–gene associations. To achieve this, a gene–gene similarity network G is constructed in single-cell data using the shared nearest neighbor method (see next section). λ and α are two tunable optimization parameters, where λ controls the overall strength of the penalty terms and α balances the effects of the L1-norm and the network-based penalties.

Within the network-based penalty term ($\beta^T L\beta$), L is a symmetric normalized Laplacian matrix defined by the following formula:

$$L = D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{2}$$

where $A = (a_{ij})_{t\times t}$ is a binary adjacency matrix derived from a gene–gene similarity network graph G. The value of a_{ij} is either 0 or 1, indicating whether the existence of a connecting edge between gene i and gene j in the network. The degree matrix $D = (d_{ij})_{t\times t}$ is a diagonal matrix, where $d_{ii} = \sum_{j=1}^{t} a_{ij}$ and $d_{ij} = 0$ when $i \neq j$.

Log-likelihood functions

The likelihood function used in the network-regularized sparse regression model varies depending on the type of phenotypic data. There are three options: (i) linear regression for continuous phenotype; (ii) logistic regression for binary phenotype; and (iii) Cox regression for survival. Specifically, for expression data from n patients, if the given phenotype $Y = (y_1, y_2, y_3, \ldots, y_n)^T$ is continuous, the following linear regression likelihood function will be selected:

$$l(\beta) = -\sum_{i=1}^{n}\left(y_i - \beta^T B_i\right)^2 \tag{3}$$

where $B_i = \left(b_{i1}, b_{i2}, b_{i3}, \cdots, b_{it}\right)^T$ represents the expression profile of the i-th patient in the bulk dataset.

If Y is a binary group indicator vector, e.g. $y_i \in \{0, 1\}$, the logistic regression log-likelihood function will be used:

$$l(\beta) = \sum_{i=1}^{n}\left[y_i\beta^T B_i - \log\left(1 + \exp\left(\beta^T B_i\right)\right)\right] \tag{4}$$

When the phenotype is given as a two-column survival data containing survival time t and survival state s, Cox regression will be considered:

$$l(\beta) = \sum_{i=1}^{n}\left[\beta^T B_i - \log\left(\sum_{k\in R_i}\exp\left(\beta^T B_k\right)\right)\right] \tag{5}$$

where $R_i = \{k : t_k \geq t_i\}$ represents the set of samples that are still alive at time t_i.

Construction of gene–gene similarity network

To introduce network-based penalties in the regression model, we need to construct a gene–gene similarity network. This process involves two steps: first, we calculate the Pearson correlation coefficients for all pairs of genes to construct a correlation matrix.

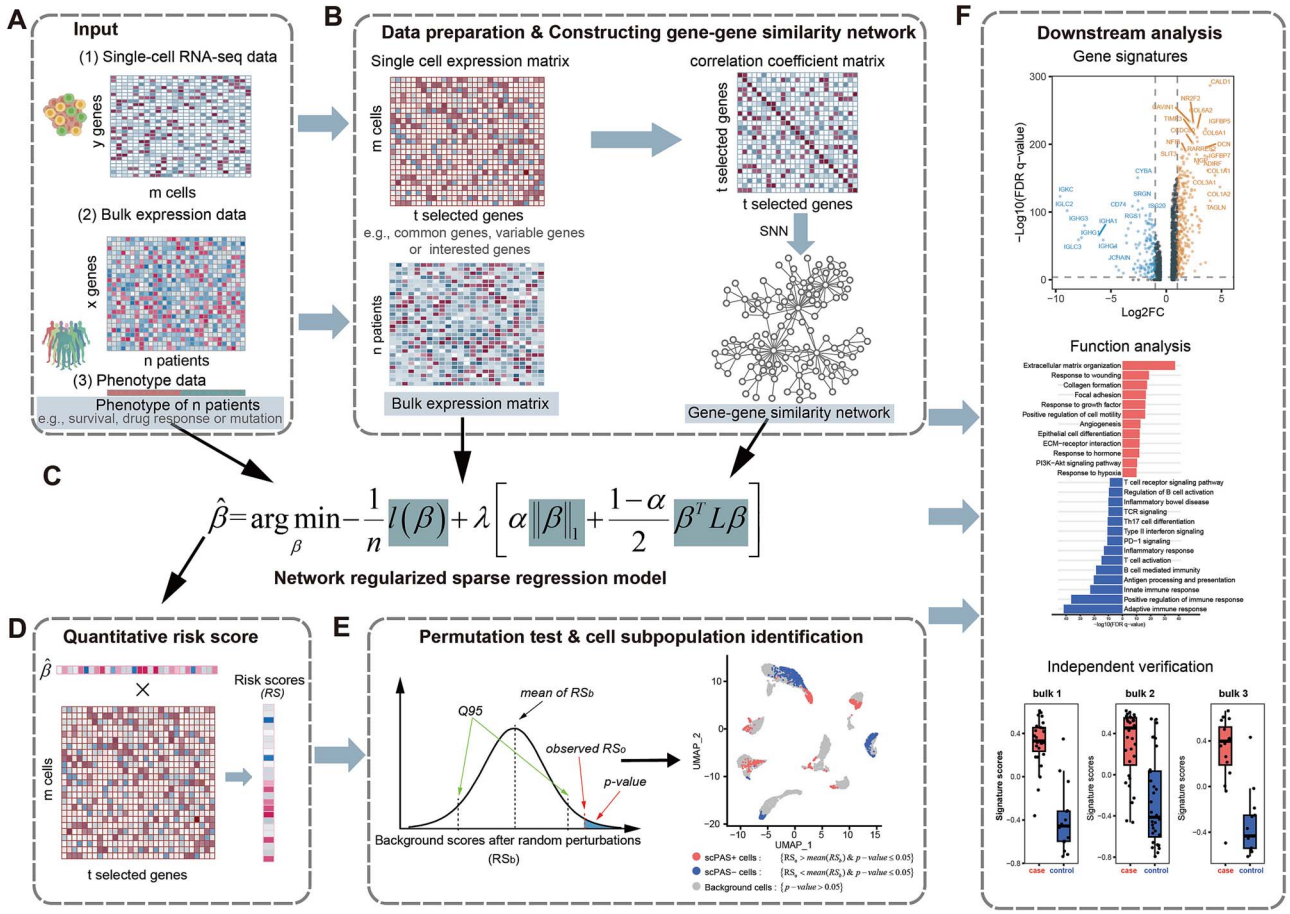

Figure 1. Overview of scPAS. (A) The inputs for scPAS were the scRNA-seq data, the bulk expression data, and phenotype labels corresponding to the bulk expression data. (B) Preparation of bulk and single-cell data and construction of a gene–gene similarity network based on single-cell data. (C) scPAS optimizes a graph-regularized sparse linear regression model by integrating bulk expression profiles and the gene–gene similarity network from single-cell data. This model comprises three components: the likelihood function, the regularization sparsity penalty, and the network penalty. (D) Based on the optimized model, scPAS calculated the risk score for each cell to quantify its potential association with the phenotypes. (E) scPAS employs a permutation test program to classify qualitative cells. (F) Cell subpopulations identified by scPAS were used for downstream analysis.

The distance between each pair of genes is defined as 1 minus the Pearson correlation coefficient. Using this distance, the K nearest neighbors for each gene are identified. The Jaccard coefficient is then computed based on the number of shared neighbors between gene pairs. This coefficient is ultimately used to measure the similarity between gene pairs. This process is implemented using the FindNeighbors function from the Seurat package, with K set to 20 by default.

Parameter tuning and implementation

Using the algorithm proposed by Li *et al.* [17], we solved the network-regularized sparse regression described above and implemented it using the APML0 R package. In scPAS, there are only two parameters to determine: λ and α. The parameter λ controls the overall strength of the penalty terms. Its value is determined through model optimization. For a fixed α, we considered 100 possible λ values and applied 10-fold cross-validation to select the optimal λ that minimizes the average error (for more details, please refer to the supplemental method). The hyperparameter $\alpha \in [0,1]$ is used to balance the amount of regularization for sparsity and smoothness. A larger α emphasizes the L1-norm to encourage sparsity, whereas a smaller α gives more weight to the network-based penalty term, encouraging the selection of similar features. In practical applications, the optimal λ tends to decrease gradually as α increases. A fixed α value is not suitable for all models because different datasets can have different sensitivities to changes in α. To select as many similar features as possible, we recommend setting α to a gradient of values from small to large and choosing the value where the rate of decrease in λ starts to slow down for the final model (refer to the supplemental materials; Fig. S2).

Quantitative risk score

The application of the model-trained coefficient β to single-cell data allows the calculation of a risk score (RS) for each cell with respect to the phenotype. The formula used is as follows:

$$\mathrm{RS}_i = \sum_{j=1}^{t} \beta_j S_{ij} \tag{6}$$

The RS value was used to quantify the association between each cell and the phenotype of interest. A higher score indicates a positive association between the cell and phenotype, whereas a lower score indicates a negative association.

In addition to calculating the risk scores at the single-cell level, this approach can also be applied to bulk data to derive sample-level risk scores. This facilitated the validation of the robustness of the trained scPAS model on independent bulk datasets.

Permutation test and qualitative identification of cells

To qualitatively assess phenotype-related cells, we perform a permutation test program based on random perturbations. First, for the trained coefficient vector β, we randomly perturbed its gene labels f times to obtain a randomized coefficient matrix $(\beta_b)_{f\times t}$. Next, we calculated the random risk score $(RS_b)_{f\times m} = \beta_b S^T$ for cells as the empirical background. When f is sufficiently large, it is assumed that the scores for each cell follow a normal distribution. Therefore, we constructed the following distribution:

$$(\mathrm{RS}_b)_i \sim N\left(\mu_i, {\sigma_i}^2\right)$$

where $\mu_i = \frac{\sum_{j=1}^{f}(\mathrm{RS}_b)_{ji}}{f}$ is the average background score for cell i after f random perturbations, and $\sigma_i = \sqrt{\frac{\sum_{j=1}^{f}\left[(\mathrm{RS}_b)_{ji}-\mu_i\right]^2}{f-1}}$ is the standard deviation of the background scores for cell i. Based on this distribution, we calculated the empirical P-value for the risk score of each cell:

$$p_i = P\left(|\mathrm{RS}_i| > \mu_i\right) = 1 - \frac{1}{\sqrt{2\pi}\sigma_i}\int_{-\infty}^{|\mathrm{RS}_i|} e^{-\frac{(t-\mu_i)^2}{2{\sigma_i}^2}}dt \tag{7}$$

The original RS was normalized based on this distribution to obtain a normalized risk score (NRS):

$$\mathrm{NRS}_i = \frac{\mathrm{RS}_i - \mu_i}{\sigma_i} \tag{8}$$

The absolute value of NRS reflects the degree to which the cell's association with the phenotype deviates from a random event. A positive and larger value indicates a higher likelihood of positive association with the phenotype, whereas a negative value suggests a higher likelihood of negative association.

Finally, based on the P-value and NRS, we discretized the risk scores for each cell to achieve a qualitative assessment of phenotype associations. The specific criteria for this judgment are as follows:

$$\mathrm{scPAS}_i = \begin{cases} 1, \mathrm{NRS}_i > 0, p_i \le 0.05 \\ 0, p_i > 0.05 \\ -1, \mathrm{NRS}_i < 0, p_i \le 0.05 \end{cases},$$

where $\mathbf{scPAS}_i = 1$ indicates that cell i is positively associated with the phenotype. Conversely, $\mathbf{scPAS}_i = -1$indicated that cell i is negatively associated with the phenotype. Cells that were not significantly associated with the phenotype were assigned a value of 0.

Simulated datasets setup

To assess the performance of scPAS, simulated single-cell data with 10 000 cells and 3000 genes were generated using Splatter [18]. These cells were from five groups with group probabilities of 0.2. The probabilities that a gene is differentially expressed (de.prob) in each of the five groups were set at 0.05, 0.05, 0.8, 0.8, and 0.8. The factor location of a differentially expressed gene (de.facLoc) in each of the five groups were set at 0.5, 0.5, 1, 1, and 1. Thus, these five groups of cells comprise four cell types, where groups 1 and 2 are grouped into the same cell type in different cell states [cell type 1(state 1) and cell type (1 state 2)], as they exhibit similar gene expression patterns (Fig. S1A). Subsequently, we randomly split these cells into two equal portions: one for constructing bulk data and the other for single-cell data usage. By assigning different proportions of cell-type composition, we were able to artificially simulate bulk samples with a specific phenotype. We conducted four simulation experiments, denoted Simulation 1–4, where Simulations 1 and 2 constructed binary phenotypes, and Simulations 3 and 4 constructed continuous phenotypes. Each patient's bulk tissue data were generated by randomly combining 500–900 single cells using the proportions shown in Fig. 2B. The specifics of each experiment were as follows: Simulation 1: 250 samples with high proportions of cell type 4 cells were selected as case phenotypes (Fig. S1B); Simulation 2: the two groups of samples, predominantly composed of the two cell states of cell type 1, were designated as the case and control phenotypes, respectively (Fig. S1C); Simulation 3: the proportions of cells in cell type 3 were utilized as continuous phenotypes (Fig. S1D); and Simulation 4: the relative proportions of cells in the two cell states of cell type 1 were utilized as continuous phenotypes (Fig. S1E).

Imputation

scPAS defaults to use the k-nearest neighbor (KNN)-smoothing algorithm for imputation [19], which aggregates information from similar cells based on the KNN idea. In simple terms, for each cell, the gene expression values were replaced with the average expression of its KNNs. The formula is as follows:

$$S' = \frac{\left(S^T N\right)^T}{k} \tag{9}$$

where $N = (n_{ij})_{m\times m}$ is the k-nearest neighbor matrix calculated in Seurat. The value of n_{ij} is either 0 or 1, indicating whether cell j is one of the top k neighbors of cell i.

The KNN-smoothing imputation method is akin to the concept of metacells [20]. Metacells are a common approach used to reduce the size and complexity of the data while preserving biologically relevant information. Additionally, other imputation methods, such as MAGIC [21], SAVER [22], and scImpute [23], can also be integrated during the preprocessing of single-cell data and included as input in the scPAS workflow. However, imputation inevitably introduces some false-positive signals. Therefore, if the data quality is reliable, it may be prudent to consider disabling imputation.

Datasets and preprocessing

We analyzed scRNA-seq data from three diseases, breast cancer, OV, and atherosclerosis, as well as 14 Visium spatial transcriptomic datasets from five distinct tumor types to evaluate the performance of scPAS. Detailed information on the datasets is provided in Table S1. The scRNA-seq data were preprocessed using the R package Seurat [16] (v4.0.2). Genes expressed in at least 10 cells were retained. Low-quality cells that have detected <200 feature genes and had a mitochondrial gene percentage >20% were excluded. The count matrix was normalized using the NormalizeData function with default parameters. Then, the highly variable genes were identified by using the FindVariableGenes function with the "vst" method. After transformation on these data using the "ScaleData" function, the scaled data were used for PCA dimension reduction. The top 30 principal components were selected for graph-based clustering. Finally, cells were projected into 2D spaces using the "RunUMAP" function. The same process was used for spatial transcriptomic data, but low-quality spots were not filtered out.

Benchmarking metrics

Precision (+), precision (−), and accuracy were used to evaluate the performance of the scPAS. Precision (+) is defined as the proportion of the scPAS-identified true-positive cells among all

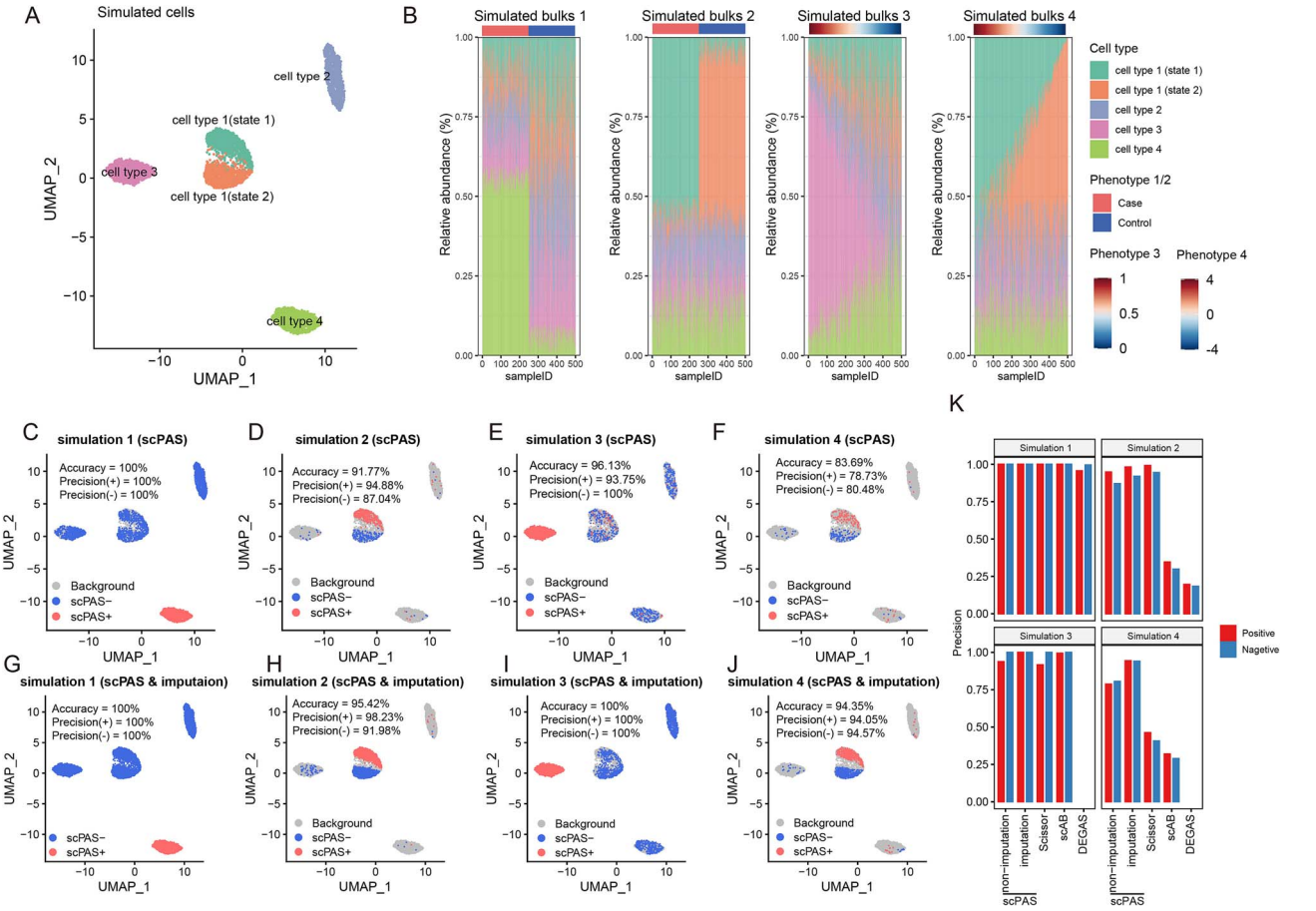

Figure 2. Simulation study and baseline comparisons of scPAS. (A) 10 000 simulated cells from Splatter with four cell types, where one of the cell types has two cell states. In total, 5000 of these cells were used to generate the 500 simulated patients in (B), and 5000 were used as the single-cell input for scPAS. (B) Four groups of simulated patients were generated by aggregating sets of 500–900 simulated cells with various proportions of the cell types. (C–F) The visualization of the result of scPAS results for each simulation without imputation. (G–J) The visualization of the result of scPAS for each simulation with imputation. (K) Comparison of the prediction performance of scPAS against different methods.

positive predictions. Precision (−) is defined as the proportion of the scPAS-identified true-negative cells among all negative predictions. Accuracy is defined as the proportion of the scPAS-identified true cells among all predictions.

Identification of signature genes and computation of signature scores

The FindAllMarkers function with the default Wilcoxon rank-sum test method was employed to identify differentially expressed genes in scPAS+ cell subpopulations versus scPAS− cell subpopulations. Genes significantly upregulated in scPAS+ cells were defined as positive signatures (log2 fold change >1 and FDR <0.001), whereas those significantly upregulated in scPAS− cells were defined as negative signatures (log2 fold change <-1 and FDR <0.001). Subsequently, the Gene Set Variation Analysis (GSVA) [24] was used to calculate the signature score for each sample. When both sets of signatures were considered simultaneously, the composite signature score was calculated by subtracting the negative signature score from the positive signature score.

Running time and memory requirements

We compared the running time and memory usage of scPAS and Scissor using the large-scale breast carcinoma scRNA-seq dataset. The experiments were performed on a Linux server equipped with a 2.75 GHz AMD EPYC 7B13 64-core processor and 2003 GB RAM, without employing parallel computing techniques. Specifically, we randomly sampled 5000, 10 000, 20 000, 40 000, and 80 000 cells for both scPAS and Scissor. Additionally, for scPAS, we varied the number of features from 1000 to 10 000. Each run was repeated five times, and the Rprof function was used to measure running time and memory usage. Notably, we excluded scAB and DEGAS from the comparison because of their significantly longer runtimes in the simulated data compared to scPAS and Scissor (Table S2).

Results

Overview of scPAS

scPAS requires three types of input data: single-cell RNA-seq data, bulk expression data, and phenotype data matched to the bulk samples (Fig. 1A). Phenotype annotations of bulk samples can be continuous variables, binary variables, or clinical survival data. Upon receiving the input data, the scPAS initially performs data preprocessing to obtain two normalized expression profiles with shared genes. Additionally, a gene–gene similarity network was constructed from the single-cell data using the shared nearest neighbor method with Pearson correlation coefficients between pairs of genes serving as the similarity or distance measures (Fig. 1B).

The key step in scPAS is the optimization of a network-regularized sparse linear regression model by integrating bulk expression profiles and the gene–gene similarity network derived from single-cell data. This step incorporates sparse and network-based penalties into the conventional regression model based on bulk data to select and weigh phenotype-associated genes (Fig. 1C). Subsequently, we directly applied the trained model to single-cell data to calculate the risk score for each cell and quantify its potential association with the phenotype (Fig. 1D). To qualitatively identify phenotype-associated subpopulations, a permutation test program was designed to calculate statistically significant *P*-values. Cells with significantly higher or lower risk scores than the background score are indicated as scPAS+ and scPAS− cells, which are positively and negatively associated with the phenotype of interest, respectively (Fig. 1E). Finally, the utility of scPAS-selected cells was demonstrated in downstream analyses, such as verification with independent data, signature gene identification, and functional analysis (Fig. 1F).

Validating scPAS using simulated single-cell data

To evaluate the performance of scPAS, we conducted experiments on a series of simulated datasets with known phenotypes. Our approach involved generating a single-cell dataset comprising 10 000 cells using Splatter [18] (Fig. 2A). Half of these cells were reserved to create simulated bulk samples. By aggregating sets of 500–900 simulated cells from this reserved group, we generated 500 simulated bulk samples. The phenotype of each sample was determined by the proportion of cell types within the aggregated cells. Four simulation experiments were designed as shown in Fig. 2B. Simulations 1 and 2 were designed for binary phenotypes, whereas Simulations 3 and 4 were designed for continuous phenotypes.

The results demonstrated that scPAS accurately identified the ground-truth phenotype-specific cell subpopulations (Fig. 2C–F). The accuracy ranged from 83.69% to 100%, while the precision of positive cells ranged from 78.73% to 100%, and the precision of negative cells ranged from 80.48% to 100%. Despite the relatively challenging task of distinguishing different cell states within the same cell type (as seen in Simulations 2 and 4), the scPAS still achieved satisfactory performance.

Furthermore, we evaluated scPAS's performance before and after imputation. The results indicated that imputation significantly enhanced scPAS's performance (Fig. 2G–J). For instance, in Simulation 2, more phenotype-associated cells were identified after imputation (293 versus 903 for scPAS+ cells and 193 versus 736 for scPAS− cells), resulting in an improvement in accuracy from 91.77% to 95.42% (Fig. 2K and Table S1). We also explored the impact of optimizing parameter α on performance by varying its gradient (refer to the supplementary materials; Fig. S2). Notably, λ decreases with increasing alpha and quickly converges to a relatively stable range. Once λ stabilized, the accuracy of phenotype-associated cells identified by scPAS also reached a stable state, depending on the proportion of true phenotype-associated features favored by the model. This provides a basis for selecting reasonable parameters for scPAS in practical applications.

To further assess scPAS's performance, we compared it with other previously reported methods, including Scissor, scAB, and DEGAS. We tested these methods using their default parameters as provided by tutorials (refer to the supplementary materials; Fig. S3). The results indicated that both scPAS and Scissor exhibited good prediction accuracy for binary phenotypes (Fig. 2K and Fig. S3A). However, scAB and DEGAS could not distinguish phenotype-associated cells from the two different cell states within the same cell type. For continuous phenotypes, the performance of Scissor was noticeably worse than that of scPAS (Fig. 2K and Fig. S3F). Specifically, in Simulation 4, Scissor identified 42.98% of the true background cells (unrelated to the phenotype) as Scissor+ cells, resulting in an accuracy of only 43.42% and a false-positive rate >30% (Fig. S3D–F).

Capturing subpopulations related to breast carcinoma

We applied scPAS to a large-scale breast carcinoma scRNA-seq dataset that incorporated tissue phenotypes [25] (Fig. 3A and B). Our approach leveraged tumor and normal phenotypes from bulk breast carcinoma samples in TCGA for model training to predict cells associated with either tumor (scPAS+ cells) or normal (scPAS− cells) phenotypes within heterogeneous cell populations (Fig. 3C and D). As a result, scPAS successfully identified 37 453 scPAS+ cells and 25 226 scPAS− cells from 106 469 cells spanning diverse cell types. Notably, >99% of the scPAS+ cells originated from tumor samples, predominantly consisting of epithelial cells, myeloid cells, and fibroblasts (Fig. 3E, left panel). In concordance with the cell types, scPAS exhibited nonrandom identification of cells originating from tumor tissue (Table S3). Conversely, scPAS− cells, 79.87% of which originated from normal tissue, predominantly represented nonmalignant cell types, including fibroblasts, basal cells, and endothelial cells (Fig. 3E, right panel). It is plausible that tumor tissue–derived cells also include nonmalignant cells associated with normal phenotypes, accounting for the presence of 37.54% scPAS− epithelial cells from tumor tissue. Using inferCNV to infer copy number variation (CNV) scores, we observed that scPAS− epithelial cells from tumor samples exhibited lower CNV scores than those of other epithelial cells from tumor samples (Fig. S5). The unsupervised clustering of all epithelial-like cells shows that scPAS− epithelial cells from tumor samples cluster together with those derived from normal tissues (Fig. S6E). Additionally, their specific expression signature is significantly upregulated in normal breast samples (Fig. S6F). This suggests that the scPAS− epithelial cells from tumor samples are more likely to be nonmalignant.

Furthermore, we systematically adjusted the scale of single-cell data by downsampling and varied the number of highly variable genes. This investigation aimed to assess their impact on the runtime and memory requirements of scPAS. Notably, while scPAS's runtime and memory requirements increased with cell count and number of features for training, the growth trend remained significantly lower than that observed for Scissor (Fig. 3F and G). In summary, scPAS can precisely identify the most phenotype-associated cells from single-cell data with guidance from phenotype information obtained from bulk data, and its performance surpasses that of the existing algorithms.

Identification of cell subpopulations associated with atherosclerosis and normal arterial tissue

In this case study, we meticulously annotated a single-cell expression dataset of atherosclerosis comprising 49 393 cells [26] (Fig. 4A). These cells were derived from calcified atherosclerotic core (AC) plaques and patient-matched proximal adjacent (PA) portions of the carotid artery, collected from three patients undergoing carotid endarterectomy. Guided by 104 bulk samples (comprising 69 atherogenesis patients and 35 healthy controls), scPAS identified 6824 scPAS+ cells associated with atherogenesis patients and 7689 scPAS− cells associated with healthy controls

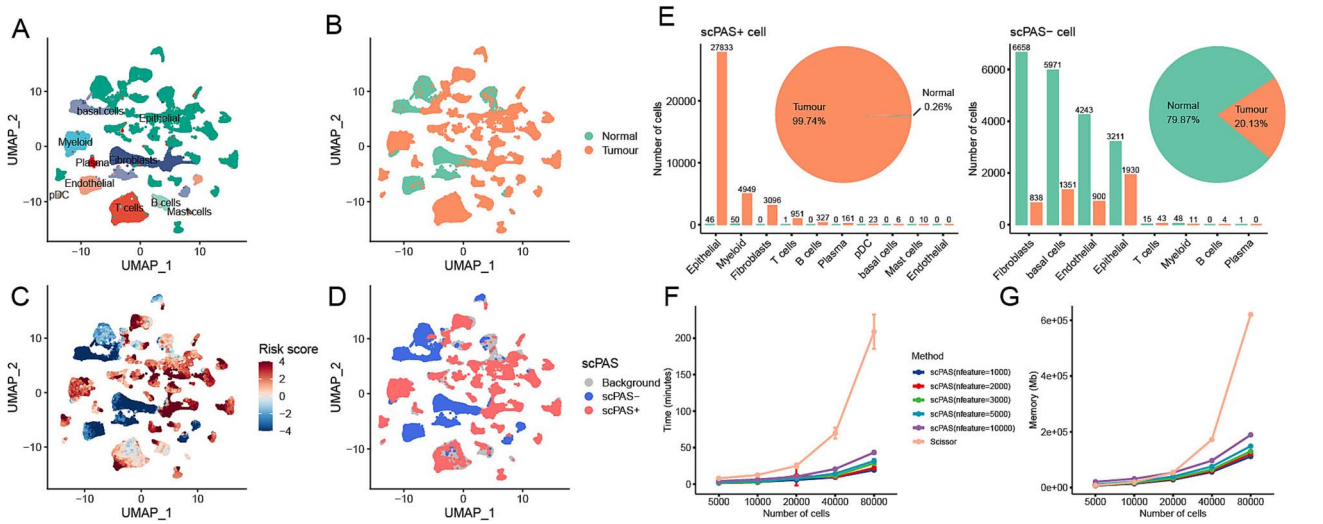

Figure 3. scPAS validation in the breast carcinoma scRNA-seq dataset. For peer review. (A, B) The Uniform Manifold Approximation and Projection (UMAP) visualization of 106 469 cells from the breast carcinoma samples and normal breast samples for their cell type annotation and tissue phenotypes. (C, D) The UMAP visualization of the scPAS-calculated risk scores and the scPAS selected cells. (E) The pie chart of the sample tissue phenotypes for the scPAS+ cells (left) or scPAS+ cells (right) with the corresponding bar plots showing the detailed constitutions in each cell type. (F, G) Curve of runtime and memory requirements of scPAS and Scissor with increasing cell number.

(Fig. 4B–D). Notably, scPAS+ cells predominantly originated from clusters annotated as immune cells, including T cells, macrophages, and B cells. In contrast, the scPAS− cells were primarily derived from vascular smooth muscle cells. This observation aligns with existing knowledge that atherosclerosis is a chronic inflammatory disease of the arterial wall, where a significant number of immune cells are recruited to atherosclerotic plaques by chemokines [27–29]. Consequently, compared with normal arterial walls, plaque tissues are enriched with immune cells.

The trained scPAS model can be directly applied to independent data to validate its robustness. Validation results revealed that the risk scores calculated for three independent bulk datasets could significantly distinguish advanced atherosclerosis samples from early atherosclerotic plaque samples [30–32] (Fig. 4E). To further elucidate the phenotypic associations of the cell subpopulations identified by scPAS, we conducted differential expression analysis between scPAS+ and scPAS− cells, constructing two molecular signatures: a positive signature for 213 upregulated genes and a negative signature for 272 downregulated genes (Fig. 4F). Subsequently, we calculated the GSVA score [24] for each sample to evaluate the predictive effectiveness of these signatures. As shown in Fig. 4G, the enrichment scores of the positive signature were significantly higher in advanced atherosclerosis and lesions with intraplaque hemorrhages, whereas the negative signature exhibited significantly lower scores.

We also applied Scissor to the atherosclerosis dataset (refer to the supplementary materials; Fig. S8). A notable difference was that Scissor identified a large number of T cells and macrophages as normal arterial-related cells (Scissor− cells). Validation with independent data revealed that the negative signature constructed from Scissor results did not show significant predictive effectiveness (Fig. S8E). When focusing on macrophages, we found that the signature up-regulated in Scissor− macrophages was significantly upregulated in atherogenesis samples (Fig. S8F and H). This indicates that Scissor incorrectly identified macrophages, which are positively associated with atherogenesis, as negatively associated cells.

Identifying prognosis-related cell subpopulations in ovarian cancer

To explore the ability of scPAS to identify cell subpopulations associated with prognosis, we applied it to a single-cell dataset containing 59 604 cells from 12 OV tumors [33] (Fig. 5A). By leveraging TCGA-OV bulk gene expression data and corresponding survival information, scPAS predicted the prognostic risk scores for each cell (Fig. 5B). Our results revealed that B/plasma cells confer a favorable prognosis in OV, whereas stromal cells, including fibroblasts and endothelial cells, were associated with poorer outcomes (Fig. 5C). Furthermore, qualitative analysis enabled the identification of phenotype-associated cells (Fig. 4D). Among the 1760 scPAS-captured cells, 18.52% (326/1760) were associated with poor survival, whereas 81.48% (1434/1760) were associated with good survival (Fig. 5E). Specifically, the 326 scPAS+ cells were predominantly localized to fibroblasts and epithelial cells, whereas the 1434 scPAS− cells were primarily immune cells, including plasma cells, myeloid cells, and T cells (Fig. 5E).

We evaluated the trained scPAS model using three independent bulk datasets [34–36], and the results consistently demonstrated that the bulk-level scPAS score significantly predicts OV survival (Fig. S9A). This validation underscores the robustness of this model in assessing OV prognosis risk. To delve into the underlying transcriptional patterns of scPAS-captured cells, we conducted differential gene expression analysis between scPAS+ cells and scPAS− cells to identify 142 down-regulated and 262 up-regulated signature genes (Fig. 5F and Table S5). Among these signature genes, those upregulated in scPAS+ cells are primarily associated with mesenchymal functions (Fig. 5G and H and Table S6), such as extracellular matrix organization, focal adhesion (involving genes such as COL1A1, COL1A2, THBS1/2, VWF, and MYL9), response to growth factors (including EGR1, FBN1, FOS, ZFP36, IER2, and POSTN), and angiogenesis (such as CYR61, CTGF, CALD1, FLNA, FN1, and NR4A1). Epithelial–mesenchymal transition (EMT) is a critical process in the malignant progression of tumors and is strongly associated with tumor metastasis and treatment resistance [37]. Conversely, the upregulated genes in scPAS− cells were predominantly related to immune cell activation functions

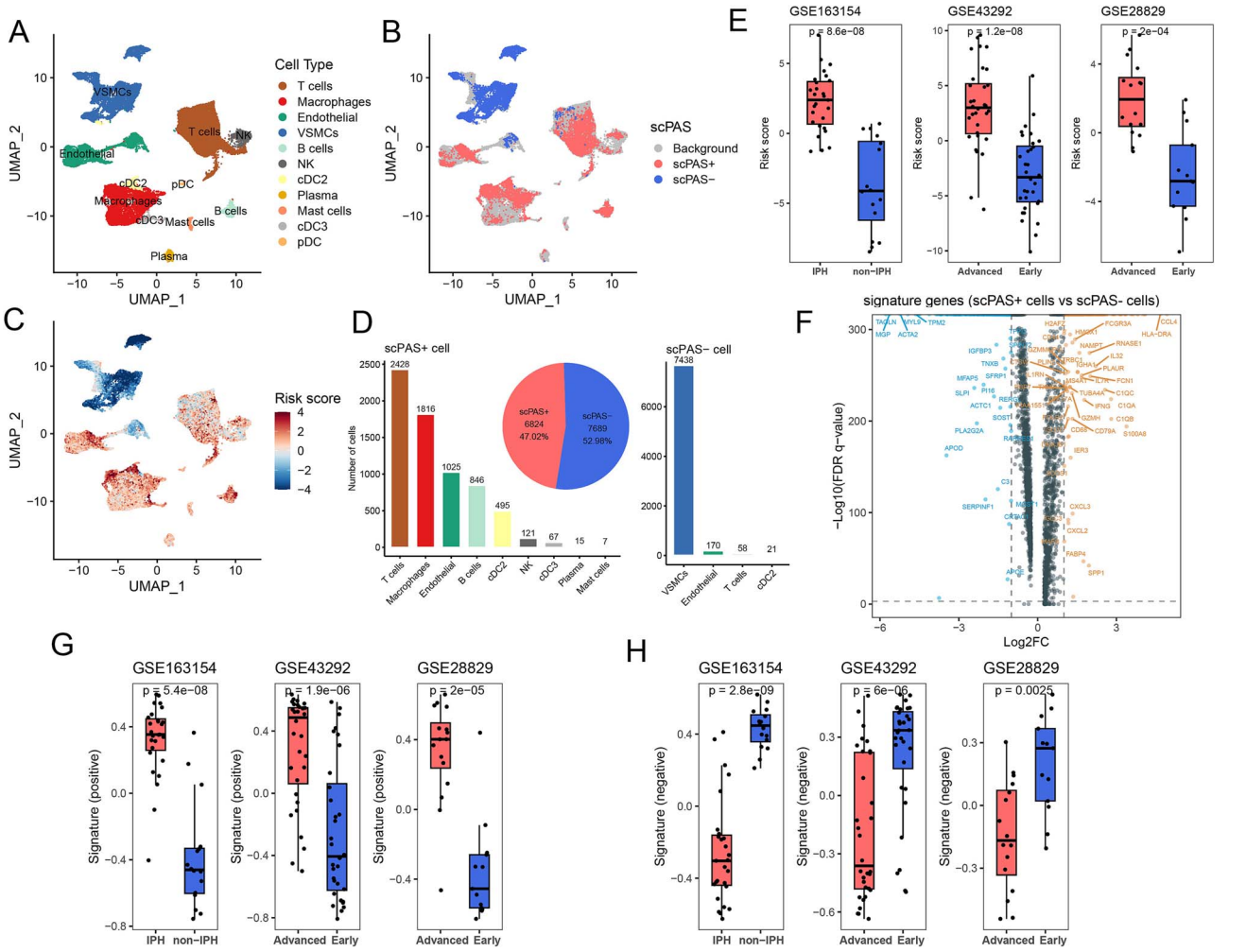

Figure 4. scPAS identification results on atherosclerosis. (A) The UMAP visualization of 49 393 cells derived from AC plaques and patient-matched PA portions of the carotid artery. (B) The UMAP visualization of the scPAS-selected cells. (C) The UMAP visualization of the scPAS-calculated risk scores. (D) The pie chart of the scPAS-selected cells with the corresponding bar plots shows the detailed constitutions in each cell type. (E) Boxplots display the comparison of the risk scores calculated by the trained scPAS model between advanced atherosclerosis samples and early atherosclerotic plaque samples from two independent validation datasets, as well as between lesions with and without intraplaque hemorrhage from one independent validation dataset. The Wilcoxon rank-sum test was used to assess the differences. (F) The volcano plot of differentially expressed genes in scPAS+ cells versus scPAS− cells. (G, H) Boxplots display the comparison of the GSVA scores of positive (G) and negative (H) signature genes, as defined in (F), between advanced atherosclerosis samples and early atherosclerotic plaque samples on three independent validation datasets.

(Fig. 5G and H and Table S6), including B-cell activation (involving genes such as CD27, CD38, CD74, and MZB1), T-cell activation (including SLAMF7, CD2, CD3D, LAG3, NKG7, and PTPRC), and inflammatory responses (such as AIF1, CD68, IFNG, CXCL10, CXCL9, CXCL13, CCL3/4/5, CXCR4, STAT1, and ISG15). This suggests that the favorable survival of patients with OV is related to an antitumor immune microenvironment.

To further assess the prognostic impact of these signature genes, we computed GSVA scores for the prognostic signature across three independent bulk datasets. Subsequent Cox survival analysis demonstrated a notable impact of our signature on survival across all three datasets (Fig. 5I–K). We also applied Scissor to these three validation cohorts (refer to the supplementary materials; Fig. S9B–E); however, we observed good performance in only one of the validation cohorts.

Application of scPAS to spatial transcriptomic data

To evaluate the applicability of scPAS in spatial transcriptomic data, we utilized scPAS to infer the locations of malignant cells across 13 Visium spatial slides from five distinct tumor types (Fig. 6A and Table S1), using disease phenotypes and bulk transcriptomic data from TCGA as references. To establish a gold standard for evaluation, the initial classification of malignant and nonmalignant spots was based on annotations from original research, which were subsequently refined by pathologists. In cases where slide annotations were unavailable, pathologists annotated the spots directly. As a result, scPAS demonstrated excellent performance in predicting malignant spots within ST data, achieving an area under the receiver operating characteristic curve (ROC-AUC) ranging from 0.815 to 0.968. (Fig. 6B and Fig. S12A). For instance, in a 10x Visium dataset of invasive ductal carcinoma, which encompassed 22 953 genes across 4727 spots, scPAS effectively identified tumor regions with elevated malignant risk (Fig. 6D, ROC-AUC = 0.815). These regions corresponded to predominantly invasive carcinoma, carcinoma *in situ*, and benign hyperplasia, as originally annotated by pathologists [38] (Fig. 6C). However, it is important to acknowledge that our reliance on TCGA bulk data as a reference may have blurred the distinction between benign hyperplasia and carcinoma regions.

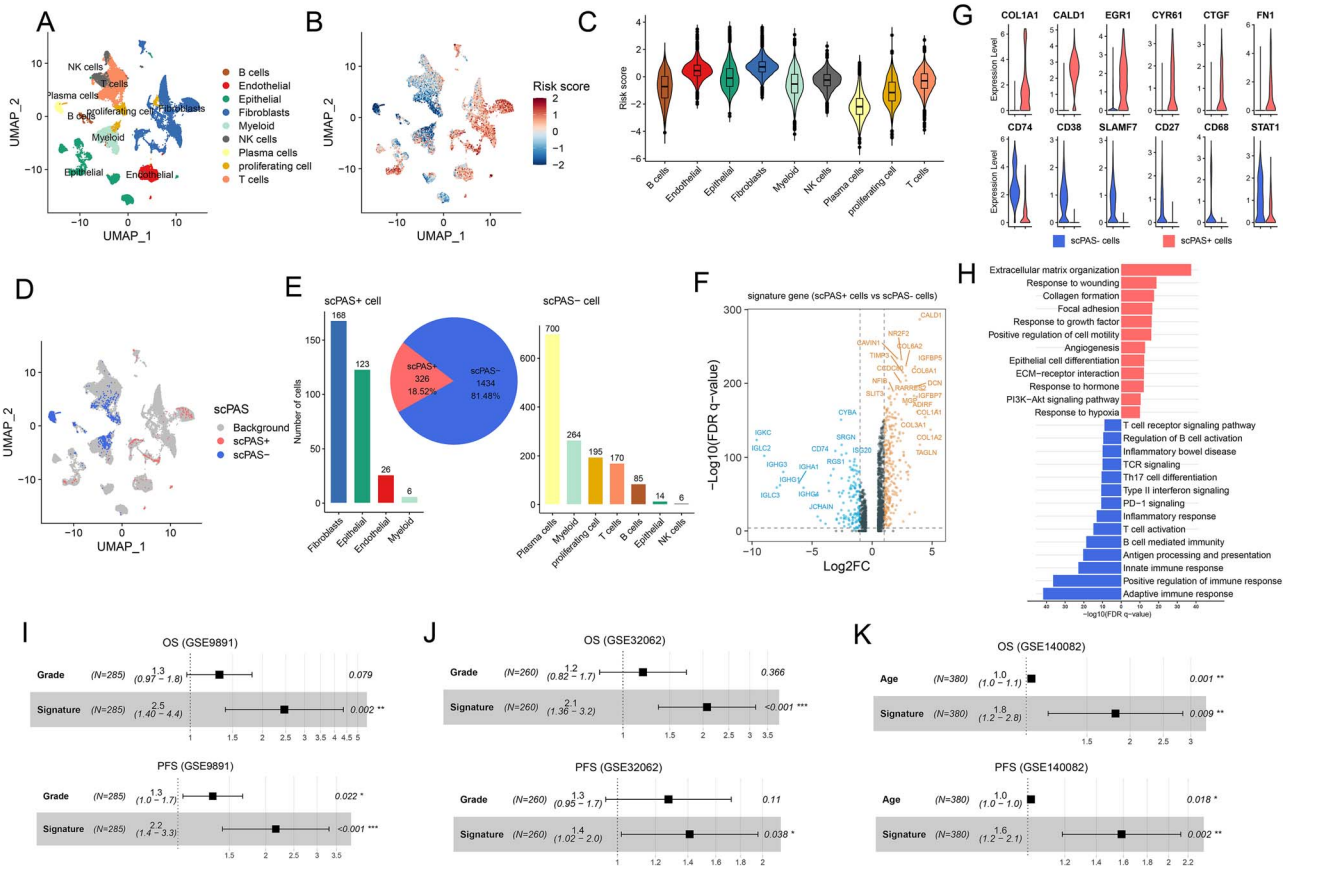

Figure 5. scPAS identification results of survival-relevant cells in OV. (A) The UMAP visualization of 59 604 cells from OV samples for their cell type annotation. (B) UMAP visualization of the scPAS-calculated risk scores. (C) The violin plot shows a comparison of the scPAS-calculated risk scores for cells from different cell types. (D) UMAP visualization of scPAS-selected cells. (E) Pie chart of the scPAS-selected cells with corresponding bar plots showing the detailed constitutions in each cell type. (F) Volcano plot of differentially expressed genes in scPAS+ cells versus scPAS− cells. (G) Violin plots of the expression levels of selected dysregulated genes in scPAS+ versus scPAS− cells. (H) Enrichment bar plot of selected pathways in the Reactome and Gene Ontology (GO) biological process domains. (I–K) Forest plots show the hazard ratios, 95% confidence intervals, and *P*-value for multivariate Cox survival analysis of our signature score on three independent datasets.

Achieving a more sensitive classification could necessitate the incorporation of finer-grained phenotypic labels from bulk data.

Advancements in technology have propelled high-definition spatial transcriptomics (HD-ST) into the forefront [39]. The analysis of extensive expression data from a single tissue section has become an inevitable task. Consequently, it is imperative to rigorously evaluate the performance of the computational tools designed to handle such substantial data volumes. In this context, we applied scPAS to HD-ST data from a colorectal cancer sample (Fig. 6E–G), comprising 137 051 bins. We used scPAS to perform two tasks on this dataset, aiming to identify bins enriched with malignant cells and those associated with survival outcomes, respectively. The results demonstrated that when leveraging tumor and normal phenotype information from TCGA bulk samples as a reference, scPAS effectively discriminated between regions enriched with malignant cells and those containing normal mucosal cells within the HD-ST data (Fig. 6H and Fig. S12B). Furthermore, when survival data served as the reference, scPAS successfully pinpointed stromal cell–enriched areas adjacent to malignant tumor cells (such as fibroblasts and endothelial cells) as poor survival-associated regions. Conversely, regions enriched with normal mucosal cells were identified as good survival-associated regions (Fig. 6I and Fig. S12C). Notably, we observed that both the runtime and memory usage of the scPAS were lower than those observed when running the Seurat pipeline on the same data (Fig. 6J). In fact, since the training of the regression model is independent of the size of the single-cell data, scPAS can theoretically process any expression profile that completes the Seurat pipeline with ease.

Additionally, we attempted to apply the pretrained models on single-cell data to HD-ST data to observe the portability of scPAS (refer to the supplementary materials; Fig. S13). The results showed that the pretrained model had highly consistent predictive results with the model trained from scratch on the HD-ST data. This characteristic highlights the advantage of scPAS in integrating single-cell data and ST data to explore disease mechanisms.

Discussion

The core of scPAS is a network-regularized sparse regression model. This model utilizes the expression profiles and phenotype information from bulk data and integrates the gene–gene similarity network from single-cell data to optimize a weighted regression model for predicting phenotype-associated risks. Based on the scoring of the model, scPAS provides each cell with a continuous metric that quantifies its phenotype-associated risk. Subsequently, a permutation test program was used to assess the significance of these risk scores, providing a qualitative distinction for identifying phenotype-associated cells. Throughout the

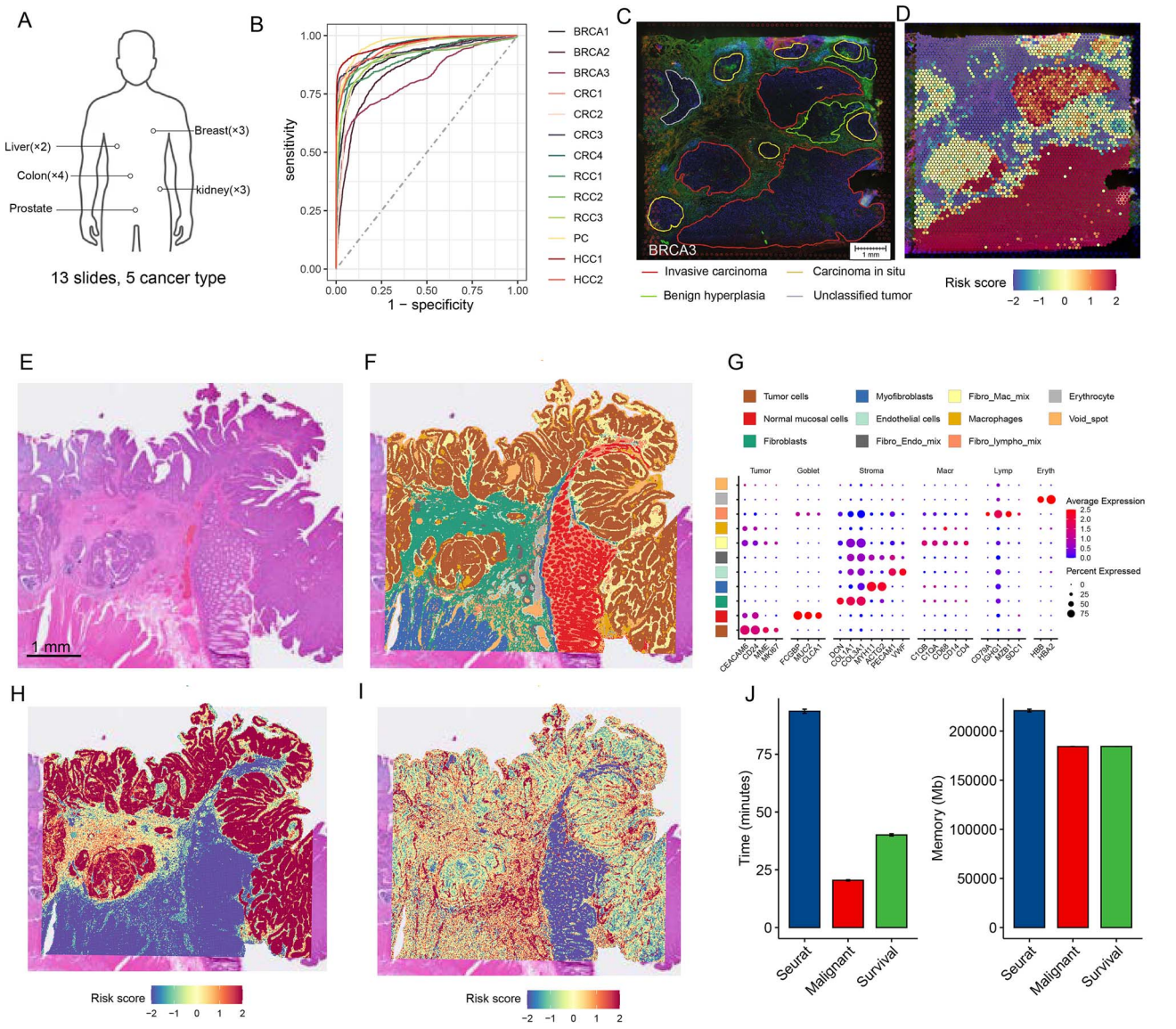

Figure 6. Expansion of scPAS to spatial transcriptomic annotation. (A) Multiple tumor ST datasets were used for the performance evaluation. Annotations from the pathologists were used to determine which spots were malignant. (B) Receiver operating characteristic curves of malignant spot prediction risk scores. (C) An example of immunofluorescent imaging of a tissue section and histopathological annotations of invasive ductal carcinoma (BRCA3). (D) Risk scores of malignant cells for each spot predicted using scPAS. (E) Hematoxylin and eosin staining of a tissue section from a colorectal cancer sample. (F) Unbiased clustering of HD-ST bins and defining cell types in each cluster of the HD-ST colorectal cancer data. (G) The dot plot shows the average expression of known markers in each cell type defined in (F). (H) Risk scores of malignant cells for each bin predicted by scPAS on the colorectal cancer HD-ST data. (I) Survival-related risk scores for each bin predicted by scPAS in the colorectal cancer HD-ST data. (J) The bar plots of the runtime and memory usage for running scPAS to identify bins enriched with malignant cells or associated with survival outcomes and running the Seurat pipeline on the same data.

process, single-cell data and bulk data remained independent, with connections established through the optimized feature genes. This approach avoids the subjective determination of cluster numbers in scRNA-seq datasets [40] and enhances the scalability of preprocessing single-cell data.

Unlike Scissor and scAB, which train models on a sample-cell similarity matrix and identify phenotype-associated cells through feature selection, scPAS directly uses genes as the features for model training. This approach avoids the potential impact of different similarity measures on the model and highlights the key phenotype-associated genes. In addition, gene-based models offer enhanced portability, enabling the direct transfer of trained models to other independent datasets. For instance, pretrained models can be applied to other bulk data to validate model robustness or to independent spatial transcriptomic data to identify phenotype-associated regions without necessitating model retraining from scratch. This characteristic enables scPAS to be seamlessly applied in studies integrating bulk data, single-cell data, and ST data to explore disease mechanisms.

Furthermore, the spatiotemporal complexity of the model optimization process is not influenced by single-cell data, enabling its easy application to large single-cell datasets. For instance, when applied to the large-scale breast carcinoma scRNA-seq dataset, as the number of cells increases, the running time and memory usage of scPAS are significantly lower than those of Scissor (Fig. 3F and G). Similarly, in the application to HD colorectal cancer ST data, the running time and memory usage of scPAS are also lower than those of the conventional Seurat pipeline (Fig. 6J).

Certainly, the scPAS methodology presents several limitations. First, as a linear regression model, its capacity to accommodate nonlinear relationships remains to be rigorously evaluated. Second, scPAS is currently limited to the analysis of three types of phenotypes and lacks the capability to address multiclass classification challenges. Developing effective strategies for analyzing cells associated with multiclass phenotypes represents a significant challenge that we intend to tackle in our subsequent research. Finally, although scPAS demonstrates strong performance on scRNA-seq data, its applicability and efficacy across other modalities of single-cell data, such as scATAC-seq and single-cell proteomics, have yet to be thoroughly validated. This will serve as a focal point for investigation in our future work.

Key Points

- We present scPAS, a bioinformatics tool designed to identify phenotype-associated cell subpopulations in single-cell data by integrating bulk data and bulk phenotypic information.
- scPAS provides both quantitative and qualitative estimates of the association between cells and phenotypes.
- scPAS was rigorously validated using simulated and real datasets, demonstrating its superior performance in phenotype-associated cell subpopulation identification.
- scPAS exhibits excellent computational efficiency, significantly outperforming Scissor in terms of processing speed on large single-cell datasets.
- As a gene-based training model, scPAS boasts high transferability. The trained model can be easily applied to other independent datasets, including bulk data, scRNA-seq data, and spatial transcriptomics data.

Supplementary data

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

Funding

This research was supported by the National Natural Science Foundation of China (grant numbers 81673036, 32170675 and 32000459).

Data availability

No new data were generated for this study. All the data used in this study are publicly available, as shown in Table S1.

Code availability

The open-source scPAS R package is available at the GitHub online repository: https://github.com/aiminXie/scPAS.

Author contributions

Aimin Xie: Conceptualization, Methodology, Investigation, Software, Writing-Original draft preparation. Hao Wang: Data curation, Investigation, Visualization, Writing-Original draft preparation. Jiaxu Zhao: Data curation, Investigation, Visualization. Zhaoyang Wang: Writing-Original draft preparation, Writing-Reviewing and Editing. Jinyuan Xu: Software, Writing- Reviewing and Editing, Validation, Funding acquisition. Yan Xu: Writing-Reviewing and Editing, Funding acquisition.

References

1. Potter SS, Single-cell RNA. Sequencing for the study of development, physiology and disease. *Nat Rev Nephrol* 2018;**14**:479–92. https://doi.org/10.1038/s41581-018-0021-7.
2. Zhang Q, He Y, Luo N. *et al.* Landscape and dynamics of Single immune cells in hepatocellular carcinoma. *Cell* 2019;**179**:829–845.e20. https://doi.org/10.1016/j.cell.2019.10.003.
3. Wagner J, Rapsomaniki MA, Chevrier S. *et al.* A Single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* 2019;**177**:1330–1345.e18. https://doi.org/10.1016/j.cell.2019.03.005.
4. Jagadeesh KA, Dey KK, Montoro DT. *et al.* Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat Genet* 2022;**54**:1479–92. https://doi.org/10.1038/s41588-022-01187-9.
5. van Galen P, Hovestadt V, Wadsworth MH. *et al.* Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* 2019;**176**:1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.
6. Qi J, Sun H, Zhang Y. *et al.* Single-cell and spatial analysis reveal interaction of FAP + fibroblasts and SPP1 + macrophages in colorectal cancer. *Nat Commun* 2022;**13**:1742. https://doi.org/10.1038/s41467-022-29366-6.
7. Lähnemann D, Köster J, Szczurek E. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:31. https://doi.org/10.1186/s13059-020-1926-6.
8. Suvà ML, Tirosh I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Mol Cell* 2019;**75**:7–12. https://doi.org/10.1016/j.molcel.2019.05.003.
9. Weinstein JN, Collisson EA, Mills GB. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764.
10. Zhang J, Baran J, Cros A. *et al.* International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* 2011;**2011**:bar026. https://doi.org/10.1093/database/bar026.
11. Chan JM, Quintanal-Villalonga Á, Gao VR. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* 2021;**39**:1479–1496.e18. https://doi.org/10.1016/j.ccell.2021.09.008.
12. Kim N, Eum HH, Lee HO. Clinical perspectives of single-cell RNA sequencing. *Biomolecules* 2021;**11**:1161. https://doi.org/10.3390/biom11081161.
13. Sun D, Guan X, Moran AE. *et al.* Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nat Biotechnol* 2022;**40**:527–38. https://doi.org/10.1038/s41587-021-01091-3.
14. Zhang Q, Jin S, Zou X. scAB detects multiresolution cell states with clinical significance by integrating single-cell genomics and bulk sequencing data. *Nucleic Acids Res* 2022;**50**:12112–30. https://doi.org/10.1093/nar/gkac1109.
15. Johnson TS, Yu CY, Huang Z. *et al.* Diagnostic evidence GAuge of Single cells (DEGAS): a flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome Med* 2022;**14**:11. https://doi.org/10.1186/s13073-022-01012-2.
16. Butler A, Hoffman P, Smibert P. *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and

species. *Nat Biotechnol* 2018;**36**:411–20. https://doi.org/10.1038/nbt.4096.
17. Li X, Xie S, Zeng D. *et al.* Efficient $\ell 0$ -norm feature selection based on augmented and penalized minimization. *Stat Med* 2018;**37**:473–86. https://doi.org/10.1002/sim.7526.
18. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**:174. https://doi.org/10.1186/s13059-017-1305-0.
19. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* 2018. https://doi.org/10.1101/217737.
20. Baran Y, Bercovich A, Sebe-Pedros A. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol* 2019;**20**:206. https://doi.org/10.1186/s13059-019-1812-2.
21. van Dijk D, Sharma R, Nainys J. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.
22. Huang M, Wang J, Torre E. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42. https://doi.org/10.1038/s41592-018-0033-z.
23. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**:9. https://doi.org/10.1038/s41467-018-03405-7.
24. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;**14**:997. https://doi.org/10.1186/1471-2105-14-7.
25. Pal B, Chen Y, Vaillant F. *et al.* A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J* 2021;**40**:e107333. https://doi.org/10.15252/embj.2020107333.
26. Alsaigh T, Evans D, Frankel D. *et al.* Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. *Commun Biol* 2022;**5**:1084. https://doi.org/10.1038/s42003-022-04056-7.
27. Moore KJ, Tabas I. Macrophages in the pathogenesis of atherosclerosis. *Cell* 2011;**145**:341–55. https://doi.org/10.1016/j.cell.2011.04.005.
28. Saigusa R, Winkels H, Ley K. T cell subsets and functions in atherosclerosis. *Nat Rev Cardiol* 2020;**17**:387–401. https://doi.org/10.1038/s41569-020-0352-5.
29. Tabas I, Lichtman AH. Monocyte-macrophages and T cells in atherosclerosis. *Immunity* 2017;**47**:621–34. https://doi.org/10.1016/j.immuni.2017.09.008.
30. Döring Y, Manthey HD, Drechsler M. *et al.* Auto-antigenic protein-DNA complexes stimulate plasmacytoid dendritic cells to promote atherosclerosis. *Circulation* 2012;**125**:1673–83. https://doi.org/10.1161/CIRCULATIONAHA.111.046755.
31. Jin H, Goossens P, Juhasz P. *et al.* Integrative multiomics analysis of human atherosclerosis reveals a serum response factor-driven network associated with intraplaque hemorrhage. *Clin Transl Med* 2021;**11**:e458. https://doi.org/10.1002/ctm2.458.
32. Ayari H, Bricca G. Identification of two genes potentially associated in iron-heme homeostasis in human carotid plaque using microarray analysis. *J Biosci* 2013;**38**:311–5. https://doi.org/10.1007/s12038-013-9310-2.
33. Xu J, Fang Y, Chen K. *et al.* Single-cell RNA sequencing reveals the tissue architecture in human high-grade serous ovarian cancer. *Clin Cancer Res* 2022;**28**:3590–602. https://doi.org/10.1158/1078-0432.CCR-22-0296.
34. Yoshihara K, Tsunoda T, Shigemizu D. *et al.* High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. *Clin Cancer Res* 2012;**18**:1374–85. https://doi.org/10.1158/1078-0432.CCR-11-2725.
35. Kommoss S, Winterhoff B, Oberg AL. *et al.* Bevacizumab may differentially improve ovarian cancer outcome in patients with proliferative and mesenchymal molecular subtypes. *Clin Cancer Res* 2017;**23**:3794–801. https://doi.org/10.1158/1078-0432.CCR-16-2196.
36. Tothill RW, Tinker AV, George J. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 2008;**14**:5198–208. https://doi.org/10.1158/1078-0432.CCR-08-0196.
37. Mittal V. Epithelial mesenchymal transition in tumor metastasis. *Annu Rev Pathol* 2018;**13**:395–412. https://doi.org/10.1146/annurev-pathol-020117-043854.
38. Zhao E, Stone MR, Ren X. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;**39**:1375–84. https://doi.org/10.1038/s41587-021-00935-2.
39. Vickovic S, Eraslan G, Salmén F. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90. https://doi.org/10.1038/s41592-019-0548-y.
40. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**:273–82. https://doi.org/10.1038/s41576-018-0088-9.

Briefings in Bioinformatics, 2025, **26(1)**, bbae655
https://doi.org/10.1093/bib/bbae655
Problem Solving Protocol