OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf160
https://doi.org/10.1093/bib/bbaf160
Problem Solving Protocol

Deep scSTAR: leveraging deep learning for the extraction and enhancement of phenotype-associated features from single-cell RNA sequencing and spatial transcriptomics data

Lianchong Gao[1,‡], Yujun Liu[2,‡], Jiawei Zou[3], Fulan Deng[4], Zheqi Liu[5], Zhen Zhang[6], Xinran Zhao[6], Lei Chen[7], Henry H.Y. Tong[4], Yuan Ji[8], Huangying Le[1,*], Xin Zou[9,*], Jie Hao[10,11,*]

[1]Shanghai Center for Systems Biomedicine, Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Jiao Tong University, 800# Dong Chuan Road, Minhang District, Shanghai 200240, China
[2]Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Fudan University, Shanghai 200433, China
[3]Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China
[4]Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China
[5]Department of Oral and Maxillofacial Surgery, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[6]Department of Oral and Maxillofacial-Head and Neck Oncology, Ninth People's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200011, China
[7]Renji Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China
[8]Molecular Pathology Center, Dept. Pathology, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[9]Digital Diagnosis and Treatment Innovation Center for Cancer, Institute of Translational Medicine, Shanghai Jiao Tong University, 800# Dong Chuan Road, Shanghai 200240, China
[10]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Chen Hua Road, Songjiang District, Shanghai 201602, China
[11]Institute of Clinical Science, Zhongshan Hospital, Fudan University, No.180 Fenglin Road, Xuhui District, Shanghai 200032, China

*Corresponding authors. Huangying Le, Shanghai Center for Systems Biomedicine, Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai JiaoTong University, Shanghai 200240, China. E-mail: hyle@sjtu.edu.cn; Xin Zou, Digital Diagnosis and Treatment Innovation Center for Cancer, Institute of Translational Medicine, Shanghai Jiao Tong University, Shanghai, China. E-mail: albumxin@qq.com; Jie Hao, Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, 201602 Shanghai, China, Institute of Clinical Science, Zhongshan Hospital, Fudan University, Shanghai 200032, China. E-mail: jhao@fudan.edu.cn
‡Equal contribution.

Abstract

Single-cell sequencing has advanced our understanding of cellular heterogeneity and disease pathology, offering insights into cellular behavior and immune mechanisms. However, extracting meaningful phenotype-related features is challenging due to noise, batch effects, and irrelevant biological signals. To address this, we introduce Deep scSTAR (DscSTAR), a deep learning-based tool designed to enhance phenotype-associated features. DscSTAR identified HSP+ FKBP4+ T cells in CD8+ T cells, which linked to immune dysfunction and resistance to immune checkpoint blockade in non-small cell lung cancer. It has also enhanced spatial transcriptomics analysis of renal cell carcinoma, revealing interactions between cancer cells, CD8+ T cells, and tumor-associated macrophages that may promote immune suppression and affect outcomes. In hepatocellular carcinoma, it highlighted the role of S100A12+ neutrophils and cancer-associated fibroblasts in forming tumor immune barriers and potentially contributing to immunotherapy resistance. These findings demonstrate DscSTAR's capacity to model and extract phenotype-specific information, advancing our understanding of disease mechanisms and therapy resistance.

Keywords: scRNA-seq; spatial transcriptomics; deep learning; phenotype-associated features; tumor microenvironment

Introduction

Recent advances in single-cell sequencing have revolutionized biomedical research and clinical diagnostics, unveiling cellular diversity at an unparalleled scale, providing insights into developmental biology [1], disease progression [2], immune function [3–6], and prognostication [7, 8], but also introduce challenges. One of the most significant hurdles is accurately identifying phenotype-associated cellular subtypes within the vast single-cell data landscape [2, 3]. This challenge is compounded by data processing complexities, such as batch effect correction, and by the presence of biological signals that may not correlate with the phenotypes being studied. Thus, innovative solutions are needed to address these complexities.

Current methods such as Harmony [9], scMerge [10], scMerge2 [11], MNN [12], Seurat [13], and LIGER [14] aim to enhance data observability and reduce batch effects, but they may distort or

Received: November 17, 2024. **Revised:** February 28, 2025. **Accepted:** March 19, 2025

eliminate crucial phenotypic features [15], obscuring valuable insights into cell heterogeneities and interactions.

Approaches like HiDDEN [16] refine cell-level labels in single-cell datasets by transforming sample-level labels into cell-specific continuous scores using dimensionality reduction and predictive modeling. This process enhances the identification of phenotype-associated cell types by accurately distinguishing affected from unaffected cells based on their transcriptional profiles. Our initial endeavor, scSTAR, utilized a partial least squares (PLS) model [17], reconstructing expression data using phenotype information to reveal phenotype-associated cell subtypes.

Although scSTAR has made progress [18, 19], its primary limitations lie in handling complex features and larger datasets that are increasingly common in contemporary research settings. Deep learning has been widely applied in various bioinformatics studies [20–24], demonstrating its potential to address challenges in large-scale and complex biological datasets. To address these issues, we introduce Deep scSTAR (DscSTAR), an advancement over the original scSTAR [17], which integrates deep learning to manage large-scale single-cell datasets. DscSTAR extracts phenotype-associated features by preserving biological signals through a balance of reconstruction and orthogonal loss, uncovering previously hidden cellular interactions.

DscSTAR has revealed that elevated expression of HSP genes and FKBP4 in CD8+ T cells correlates with immune dysfunction, offering potential markers for resistance to immune checkpoint blockade (ICB) therapy in non-small cell lung cancer (NSCLC). These genes are also enriched in CD8+ T cells unresponsive to ICB therapy in basal cell carcinoma (BCC) and adoptive cell transfer therapy in skin cutaneous melanoma (SKCM). In renal cell carcinoma (RCC), DscSTAR's spatial transcriptomics (ST) analysis identified interactions between mesenchymal-like cancer cells, CD8+ T cells, and tumor-associated macrophages (TAMs), which may contribute to immune suppression and worsen patient outcomes. Furthermore, DscSTAR highlighted the critical role of S100A12+ neutrophils and cancer-associated fibroblasts (CAFs) in forming tumor immune barriers in hepatocellular carcinoma (HCC), potentially contributing to immunotherapy resistance.

Results

Overview of deep scSTAR

DscSTAR is designed to compare two groups of specific cell types from different biological or clinical conditions, such as disease severity, tissue origin, spatial location, or custom phenotype labels, like binary classification based on the expression of specific genes. DscSTAR involves three major steps (see Materials and methods section): (i) Unchanged cell recognition between the two conditions using scCURE [25]. (ii) Constructing a PLS discriminant analysis (PLS-DA) model on the unchanged cells and reducing noises. (iii) Extracting the phenotype-associated cellular features by leveraging a supervised multitask learning (MTL) model.

The workflow of DscSTAR begins by identifying unchanged cells across different conditions using scCURE. scCURE employs a Gaussian mixture model (GMM) to cluster cells with consistent biological traits. It can automatically determine the optimal number of clusters through the Akaike Information Criterion, offering flexibility in clustering without the need to pre-specify cluster counts. Additionally, users have the option to manually set a cluster count if domain knowledge suggests a more appropriate value. This approach aligns each cell with the most suitable Gaussian model and accurately identifies "unchanged cells" by evaluating the Kullback–Leibler (KL) divergence. The unchanged cells are considered consistent and stable between pairs of conditions, with observed differences primarily attributed to noise rather than significant biological influences [25]. While subtle biological differences may exist, scCURE models these differences using phenotype-labeled single-cell expression data, ensuring that the variations among unchanged cells are minimally correlated with phenotype differences [25]. After distinguishing unchanged cells, the next step employs a PLS-DA [17] model trained on these cells to reduce random noise, irrelevant biological signals, and batch effects, focusing on minimizing variability unrelated to the phenotype labels being modeled. This denoising step is intentionally aggressive to maximize signal clarity. Recognizing the potential for over-filtering of subtle biological differences, we provide users with the option to skip this step and directly enhance phenotype-relevant features. This ensures flexibility and adaptability to diverse datasets.

The final step leverages a supervised MTL model to further process the noise-reduced data. The aim of this step is to isolate and extract the phenotype-associated components. The MTL model achieves this by employing a denoising autoencoder (DAE) to map the original data into a latent feature space. Within this space, multi-layer perceptron (MLP) is utilized to introduce classification loss for distinguishing cells from different phenotypes to extract phenotype-associated dynamic features while balancing reconstruction loss and orthogonal loss, ensuring that phenotype-related features are faithfully retained during data reconstruction. The workflow of DscSTAR is illustrated in Fig. 1.

Evaluation of DscSTAR's ability to enhance phenotype-related features on simulated datasets

To evaluate the effectiveness of DscSTAR in revealing phenotype-associated cell subtypes, we first compared the full version of DscSTAR, as well as a variant of DscSTAR that only underwent the step2 noise reduction (DscSTAR_nr) without proceeding to step3, with unprocessed data and four expression matrix correction algorithms: scSTAR [17], scMerge2 [11], Harmony [9], MNN [12], and SAVER [26].

Following the protocol derived from Zou et al. [17], we simulated a pair of phenotype data with a control group and a case group. The control group was generated by distributing cells via a bivariate normal distribution, which were then projected into a high-dimensional gene expression space with normally distributed random noise added to create the initial case group. Subsequently, phenotype differences were introduced to the initial case group to generate cell subtypes characterized by phenotype-specific differentially expressed genes (DEGs). These methods were rigorously tested across 15 different scenarios in the case group, with an adjusted number of subpopulations (2, 3, and 4) and degree of FC (fold change, 1.2, 1.3, 1.4, 1.5, and 2) of DEGs. Specifically, we generated 10 simulated datasets for each of the 15 different combinations of clusters and FC, resulting in a total of 150 synthetic datasets. After processing, we performed unsupervised dimensionality reduction and clustering on the case group to evaluate the algorithms' ability to explain phenotype-related subtypes.

First, the adjusted rand index (ARI) [27] was used to assess the alignment between clustering outcomes and ground truth labels. At low signal strength (FC = 1.2), DscSTAR excelled in identifying phenotype-associated cell subpopulations (Fig. 2a), achieving higher ARI scores than scSTAR across all conditions. While scMerge2 and MNN performed well at higher signal strengths (FC = 2, Fig. 2a), DscSTAR maintained high ARI scores

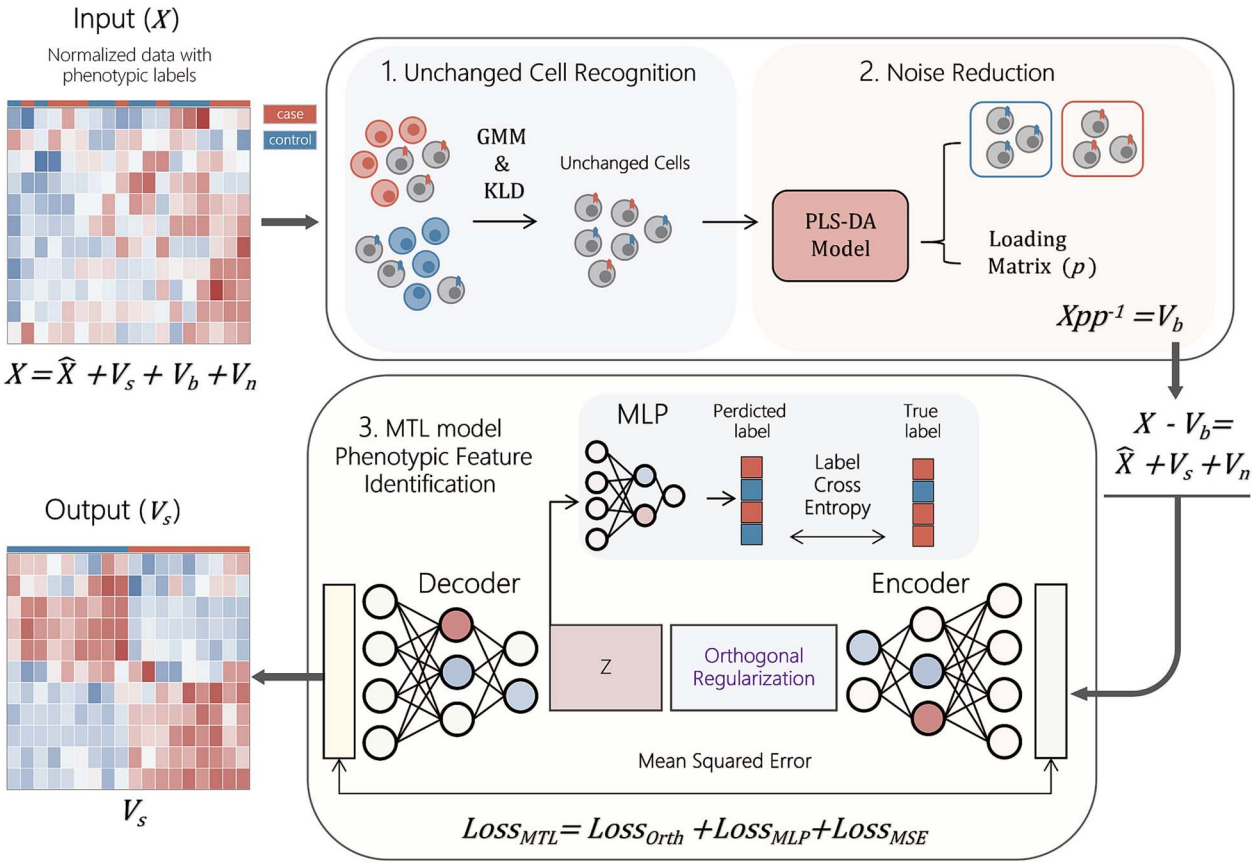

Figure 1. The workflow design of the DscSTAR. The input is the expression data with phenotype labels, three steps are then applied.

under various conditions, demonstrating robustness in preserving phenotype-associated cell heterogeneities. The average silhouette width (ASW) metric (Fig. 2b, Supplementary Fig. S1a) also showed DscSTAR's superiority in cluster separation. Unlike scSTAR, which overemphasizes features in larger datasets (high ASW but low F1 score), DscSTAR preserved data structure and ensured balanced clustering.

In addition, the F1 score was used to evaluate DEG accuracy with Seurat's "findallmarker" function, using the ground-truth labels to assume optimal clustering for all methods. This allows for an evaluation of DEG recovery, independent of clustering performance (Fig. 2c). The F1 score reflects the method's ability to recover DEGs, and an increase suggests better retention of biological signals, while a decrease indicates loss of important details for phenotypic analysis. Both SAVER and DscSTAR outperform the unprocessed data (Fig. 2c), demonstrating that both methods recover DEGs that were masked in the raw data. However, SAVER imputes missing values across all data, excelling in DEG identification but also amplifying unrelated signals, which may distort the topological structure of data. As a result, SAVER showed lower ARI and ASW (Fig. 2a and b). DscSTAR, designed to preserve phenotype-related heterogeneity, performed well in clustering while also recovering DEGs. Its overall F1 score is second only to SAVER (Fig. 2d, Supplementary Fig. S1b). Additionally, we assessed DscSTAR's performance at clusters as 6 and FC levels of 1.3, 1.5, and 2.0. The ARI and ASW results were consistent with previous evaluations, and it is worth noting that DscSTAR performed best in terms of F1 score at the low FC (1.3) condition (Supplementary Fig. S1c).

Optimal single-cell RNA sequencing analysis should be able to balance noise reduction with the preservation and enhancement of biological signal integrity. DscSTAR excels in this regard, underscoring DscSTAR's strengths in single-cell analysis (Fig. 2d). Additionally, uniform manifold approximation and projection (UMAP) mapping of cell topological structures illustrated the ability of DscSTAR to clearly delineate different cell subtypes, even in low signal scenarios (3 subclusters, FC = 1.2, Fig. 2e).

DscSTAR detects the key cell subtypes missed by the standard analysis in the simulated ground truth datasets of cell-type mixtures

To evaluate the performance of DscSTAR in identifying key cell subtypes in heterogeneous cell mixtures, we utilized simulated datasets designed to mimic biologically relevant perturbation scenarios. These datasets were constructed using single-cell RNA-seq profiles of Naive B and Memory B cells derived from peripheral blood mononuclear cells. Specifically, the control group consisted entirely of Naive B cells, while the case group was composed of 95% Naive B cells and 5% Memory B cells, simulating a highly imbalanced condition where Memory B cells were present as a rare subtype.

Under these conditions, standard dimensionality reduction methods generated a highly heterogeneous latent space that obscured the cluster of perturbed Memory B cells (Fig. 2f). Attempts to adjust parameters such as the number of principal components, clustering resolution, or integration methods (e.g., BBKNN or Harmony) did not resolve this issue (Supplementary Fig. S1d).

In contrast, DscSTAR, by levering case/control labels, successfully resolved five subclusters (Dsc_0 to Dsc_4). Among these, Dsc_1 and Dsc_4 effectively captured the Memory B cell signal, with Dsc_4 specifically identifying an activated transitional state

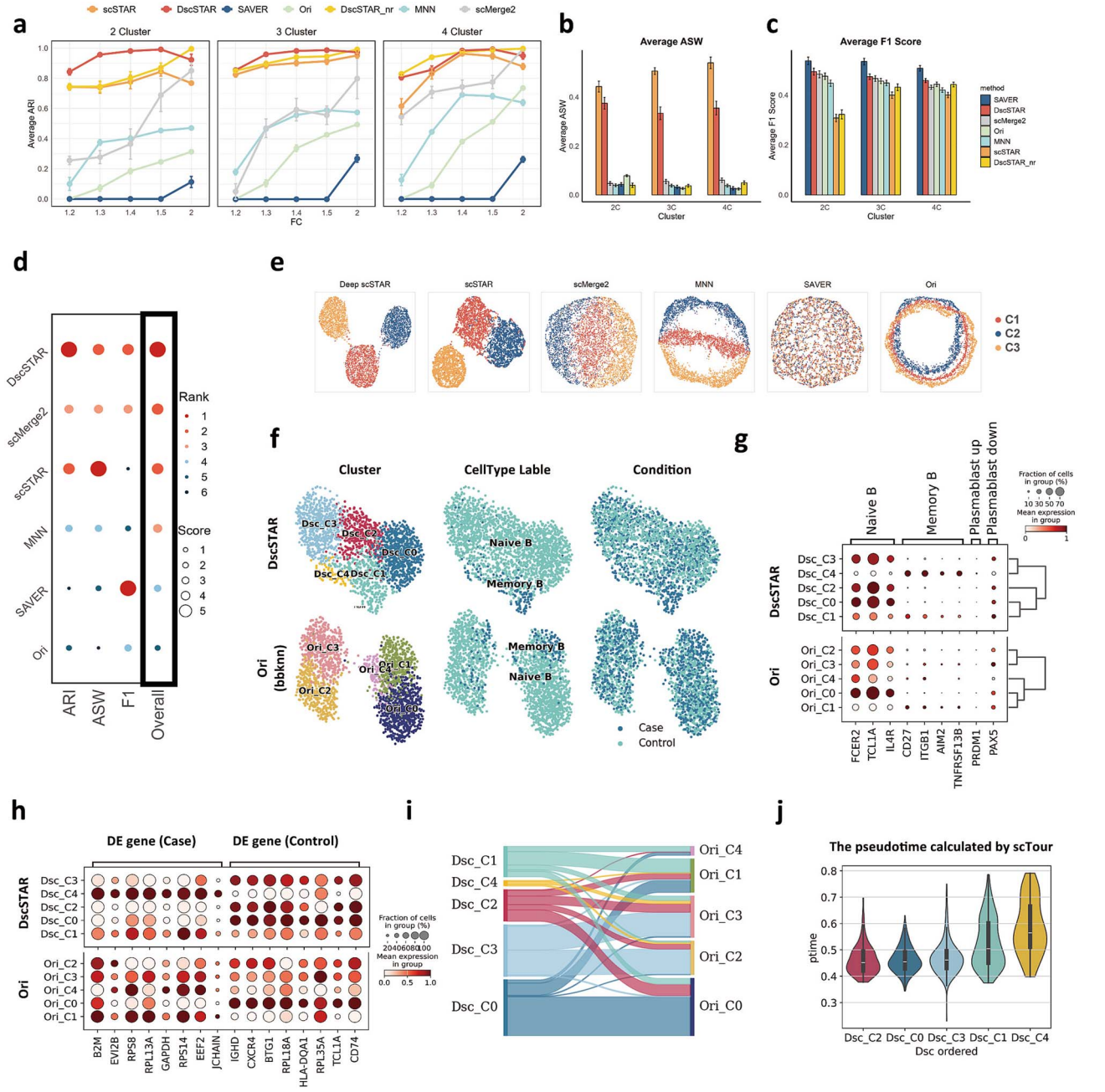

Figure 2. Evaluation on simulated datasets. (a) ARI-based clustering performance comparison across three cluster settings at different fold changes (FC = 1.2, 1.3, 1.4, 1.5, and 2.0). (b) ASW-based clustering performance, averaged over all tested cluster numbers and FC conditions. (c) F1-scores of DEGs identified by three cluster settings, averaged over all cluster numbers and FC conditions. (d) Dot plot ranking of tested methods across three evaluation metrics. The cumulative ranking is based on the average ranking across these metrics. (e) The UMAP plots depicting the distinctiveness of cell subclusters at 3C, FC = 1.3. (f) UMAP embeddings comparing DscSTAR and traditional analysis (BBKNN). The top row represents DscSTAR-processed data, while the bottom row represents data processed using the standard BBKNN pipeline. From left to right, UMAPs are labeled by Leiden cluster assignments, ground truth cell type labels, and condition labels. (g) Dot plot showing the mean expression levels of Naive B, Memory B, and Plasmablast marker genes. The size of each dot represents the percentage of cells expressing the marker, while color indicates mean expression. (h) Dot plot showing the mean expression levels of DE genes in the case group and control group. (i) Sankey diagram illustrating the relationship between DscSTAR-identified subclusters (Dsc labels) and the original clusters (Ori labels) from the standard analysis. (j) Violin plot displaying pseudotime (ptime) distributions across different clusters.

shared by Memory B cells and plasmablasts (Fig. 2g). Furthermore, Dsc_4 clearly revealed the DEGs between case and control samples (Fig. 2h). In comparison, the standard analysis merged Dsc_4 within larger clusters such as Ori_C3 and Ori_C2, masking the transitional state from Memory B cells to plasmablasts (Fig. 2i). We further compared our results with Hidden [16]. While hidden revealed an association between Memory B cells and the case condition, it failed to resolve finer subpopulation structures [16]. In contrast, DscSTAR not only detected the Memory B cell signal but also uncovered a distinct case-associated subcluster (Dsc_4) representing the transition from Memory B cells to plasmablasts, thereby providing a more detailed view of the perturbation response.

To validate the DscSTAR-discovered subclusters, we conducted pseudotime analysis using scTour, which confirmed a developmental trajectory transitioning from Naive B cells

to Memory B cells and subsequently to the intermediate Memory B–plasmablast subtype (Fig. 2j). Standard analysis failed to produce similar pseudotime results (Supplementary Fig. S1e).

Identifying an HSP-associated exhausted CD8+ T cell subpopulation related to immunotherapy

A recent pan-cancer T cell atlas revealed stress response state T cells (TSTR) with elevated HSPA1A expression, suggesting these cells may contribute to immunotherapy resistance [3]. However, the temporal dynamics and functions of TSTR cells remain unclear. Additionally, the ubiquitous expression of heat shock proteins (HSPs) [3, 28, 29] complicates the study of T cells associated with HSPs. Using DscSTAR, 32,528 CD8+ T cells from NSCLC tumor environments with PD-1 therapy [30] were analyzed to identify key groups related to HSPs. Cells expressing HSPA1A were designated as the case group, with the remaining cells serving as controls. Previously, CD8+ T cells were classified into terminally exhausted (Tex), non-exhausted (Non-ex), and proliferative (PROLIF) groups (Fig. 3a); here we focus on Tex cells due to their potential as tumor-reactive T cells [30, 31].

Subsequent analysis using hypergeometric testing linked the DscSTAR-clusters with the initial cell type, identifying five distinct subclusters. These included three Tex-related subclusters—Tex_C1, Tex_C2, and Tex_C3—along with a PROLIF cluster and a Non-ex cluster (Fig. 3a). Tex_C1 was characterized by the highest expression of exhaustion signatures, including PDCD1, CTLA4, and HAVCR2. Apart from GZMB, other cytotoxic genes were expressed at low levels (Fig. 3b). Interestingly, both Tex_C1 and Tex_C3 exhibited high expression of HSPs like HSP90AA1 and HSPA1A, but FKBP4 was uniquely highly expressed in Tex_C1 (Fig. 3b). A study of LUAD cancer cells revealed that FKBP4 integrates FKBP4/HSP90/IKK and FKBP4/HSP70/RelA complexes, promoting cancer progression and associating with worse overall survival [32]. However, co-expression of FKBP4 with HSP90 and HSP70 family members in T cells has not been previously reported. This leads us to the functional enrichment analysis, which indicated that Tex-associated clusters exhibited downregulation of several immune function-related terms, suggesting immune dysfunction (Fig. 3c, Supplementary Fig. S2). Similar results could not be obtained from subgroups derived from the raw data (Supplementary Fig. S3). Notably, unique functional downregulation was uncovered by DscSTAR clusters, including depletion in immune synapse formation and receptor-mediated endocytosis in Tex_C1. This impairment parallels findings in other studies, where disrupted immune synapse formation within tumor microenvironments (TMEs) leads to reduced tumor cell elimination [33, 34]. Furthermore, Gene sets downregulated in these clusters were found to be significantly higher in ICB responders compared to non-responders in an independent NSCLC ICB dataset [35] (Fig. 3d). Survival analysis incorporating marker genes from Tex-related clusters demonstrated that patients with higher expression levels of Tex_C1 and Tex_C2 marker genes had significantly worse survival outcomes in the TCGA-LUAD dataset, accessed via the GEPIA2 [10] (Fig. 3e, Supplementary Table S1). Results for other clusters were not significant. Given the association of Tex_C1 with immune resistance and poor survival, as well as high expression of HSPs, we termed Tex_C1 as HSP-related Tex.

To capture the temporal changes in CD8+ T cell differentiation within the DscSTAR clusters, we utilized the vector field-based deep learning algorithm scTour [36], revealing a continuum of cell differentiation. Naive (from the Non-ex cluster), positioned at the earliest pseudotime, with PROLIF cells at the final stage. Tex_C1 and Tex_C2 were positioned at the terminal pseudotime before proliferation, confirming they are in a terminal exhaustion state (Fig. 3f).

Since differentiation trajectories did not reveal differences between Tex_C1 and Tex_C2, we turned to TCR information, a reliable molecular marker for tracking T cell lineage [3, 29], thus enabling a better trace of the origins of the HSP-related Tex. TCR clonotype overlap analysis showed that Tex_C1 and Tex_C2 shared only a few clonotypes (Fig. 3g), proving intrinsic differences between them. To explore the origin and characteristics of HSP-related Tex using TCR information, CD8+ T cells sharing TCRs with Tex were selected, and unsupervised clustering was applied to these cells (Fig. 3i and j). Besides the terminal and PROLIF Tex cells, we also identified two Non-ex T cell states that share TCRs with the terminal Tex subset. They were defined as precursor exhausted T cells, namely Texp1 (GZMK+NR4A2- T cells) and Texp2 (GZMK+NR4A2+ T cells), each characterized by unique signature genes (Fig. 3h and i). Studies have shown that lung cancer with ICB immunotherapy leads to an increase in Texp populations, and the post-treatment high abundance of Texp correlates with ICB response [30]. These Texps originate from the expansion of pre-existing Texp rather than a reversal from Tex.

Interestingly, the number of Texps sharing TCRs with HSP-related Tex was lower than those not sharing TCRs with HSP-related Tex, implying that HSP-related Texps do not undergo notable proliferation or revival post-treatment (Fig. 3j and k). Specifically, Texp1 proliferates less, and Texp2 expands but may quickly transition to an exhausted state (Fig. 3j and k). We propose that HSP-related Tex represents a distinct exhaustion state that readily transitions to Tex and struggles to maintain their numbers, which may contribute to immunotherapy resistance.

Therefore, cells sharing TCRs with HSP-related Tex were labeled as HSP-related clones, while those not sharing TCRs with HSP-related Tex were labeled as HSP-unrelated clones. Subsequently, we compared the differences between HSP-related and unrelated clones within these four subgroups. Activity scoring was conducted for various gene signatures to identify the cellular programs underlying their transcriptional differences (Fig. 3l). Results showed that HSP-related clones scored higher for gene sets associated with alpha-beta T cell activation, endoplasmic reticulum stress responses, TNF signaling pathways, and MAPK signaling pathways. Additionally, only in Texp1, HSP-related Tex clones had higher apoptosis scores, suggesting that HSP-related clones might be in a more severe chronic inflammatory environment leading to Texp apoptosis and reduced numbers. Interestingly, HSP-unrelated clones scored higher at all stages for peptide biosynthesis processes, indicating enhanced protein synthesis and processing activities, suggesting robust immune functions in HSP-unrelated clones. These results highlight the differences between HSP-related and unrelated clones, warranting further exploration.

Finally, given that HSP-related Tex specifically overexpresses HSP90 and FKBP4, we used these genes to annotate different datasets for HSP-related Tex. First, correlation analysis in NSCLC CD8+ T single-cell data and two independent NSCLC ICB bulk datasets [35, 37] confirmed the positive association between HSP90 and FKBP4 (Fig. 3m). Notable, further analysis of single-cell data from ACT-treated SKCM [38] and ICB-treated BCC [39] showed high expression of FKBP4 and HSP90 in CD8+ T cells of some non-responders but not in responders (Supplementary Figs S4 and S5).

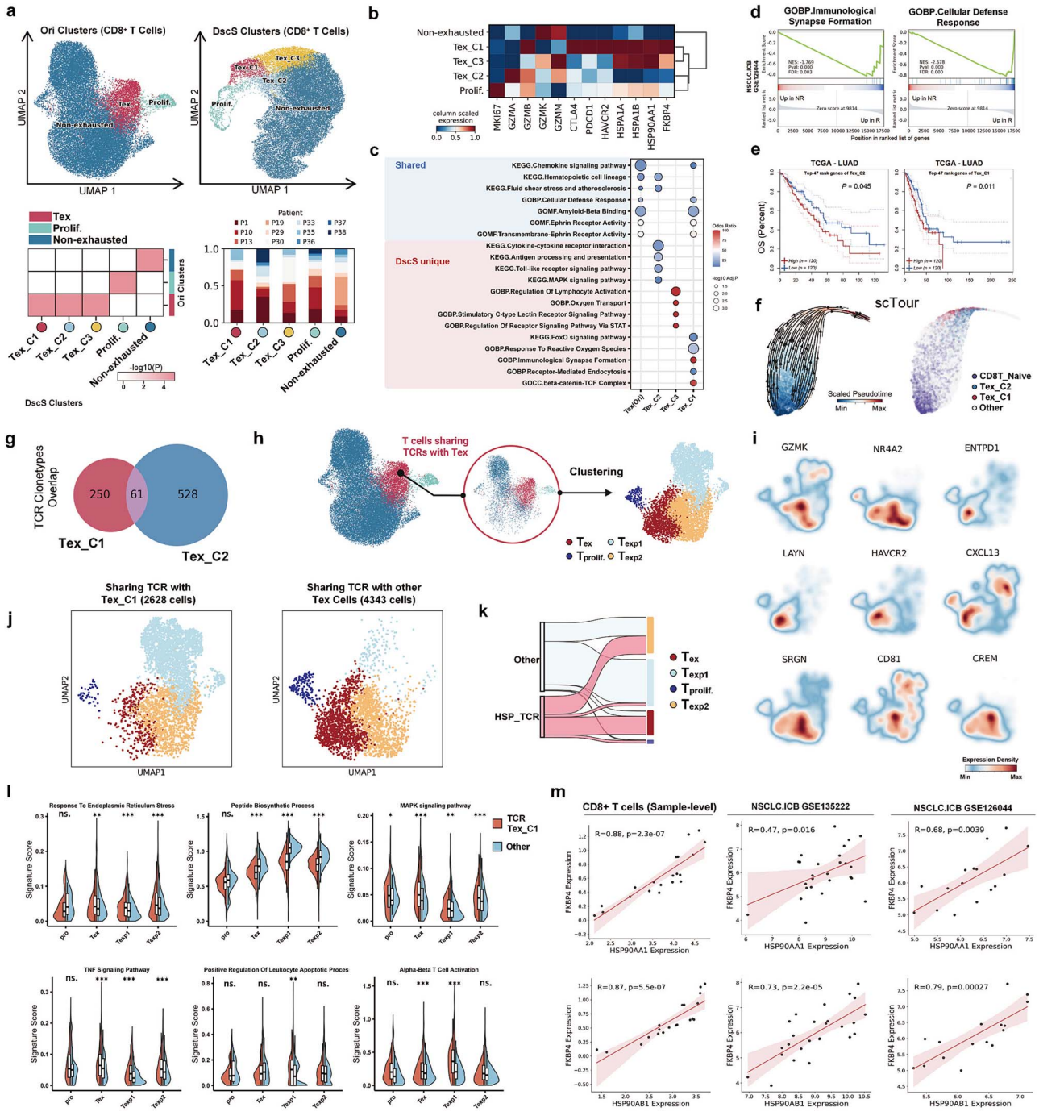

Figure 3. HSP-associated exhausted CD8+ T cell subpopulation related to immunotherapy (a) UMAP embeddings of 32 528 CD8+ T cells from lung tumors treated with anti-PD-1, before and after treatment, Leiden clusters based on raw data (left), Leiden clusters refined by DscSTAR (right). Heatmap shows the associations between clusters and cell types via hypergeometric testing. Stacked bar chart shows sample distribution across clusters. (b) Heatmap showing marker gene expression across DscSTAR-defined clusters. (c) Dot plot of enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) terms for Tex-related clusters, with dot size representing −log10 adjusted *P*-value and color indicating odds ratio. (d) GSEA results of immunological synapse formation and cellular Defense response in the NSCLC ICB dataset (GSE126044). (e) Kaplan–Meier survival curves assessing clinical relevance of marker genes in DscSTAR-defined clusters (Tex_C1 and Tex_C2), analyzed using the log-rank test. (f) Trajectory estimated by scTour, shows pseudotime trajectories (left) and DscSTAR-defined clusters. (g) Venn diagram illustrating clonotype overlap between Tex_C1 and Tex_C2. (h–j) Identification of Tex and Tex-irrelevant cells using TCR information, including unsupervised clustering of selected cells. (h–j) Identification of Tex-related and Tex-irrelevant cells using TCR information. (h) Tex-related cell extraction and re-clustering. (i) Density map of marker gene expression in Tex-related clusters in the UMAP embedding. (j) UMAP embeddings comparing TCR-sharing cells, showing cells sharing TCR with Tex_C1 (left) and other cells (right). (k) Sankey diagram illustrating the relationship between re-clustering subclusters and the TCR labels. (l) Violin and boxplots showing differential single-cell signature scores between HSP-related and unrelated clones, analyzed by Wilcoxon rank-sum test with Bonferroni correction. (m) Scatter plot of Pearson correlations between selected gene expressions, with significance assessed by the Pearson test.

Deep scSTAR uncovered tumor-immune interactions in RCC through enhancing spatial transcriptomics

In ST, clustering spots reveal distinct functional niches defined by specific cell types [40]. However, standard analysis using overall spatial and expression data, often struggles with background noise, complicating niche identification [41]. To enhance niche analysis in ST, DscSTAR can minimize irrelevant noise and enhance the detection of relevant biological signals. This ability was demonstrated using a primary RCC ST dataset with 12 samples [42], annotated with tertiary lymphoid structures (TLS), where tumor regions were mapped (Fig. 4a) using the "TESLA" Python package [43].

DscSTAR was tested to improve signal detection for B cells, CD8+ T cells, and monocytes. Using CD8+ T cells as an example, we divided spots into case (expressing CD8A and CD8B) and control groups and processed them with DscSTAR. Our evaluation followed three steps: establishing a baseline using Robust Cell Type Decomposition (RCTD) based on raw counts [44], which precisely identifies and quantifies cell types, adapting well to subtle variations across various contexts [45, 46]. Then, assessing cell abundance with the Microenvironment Cell Populations-counter (MCP-counter) [47] on DscSTAR processed and unprocessed normalized data, and finally comparing the correlation between MCP-counter scores and the RCTD baseline. Results showed that DscSTAR-enhanced data had a stronger correlation with the baseline than unprocessed data, confirming its effectiveness in signal enhancement (Figs 4b and c, Supplementary S6, and Supplementary Table S2).

To demonstrate DscSTAR's role in niche delineation and analysis, we improved CD8+ T cell signals. Both DscSTAR-treated and untreated samples revealed eight distinct niches (Fig. 4d and Supplementary S7). The treated data showed a marked difference in niche spatial distribution (Fig. 4d), underscoring DscSTAR's influence on data interpretation. In the processed samples, regions were annotated with specific cell type markers (Fig. 4e), whereas in the untreated data, two primary tumor regions were identified: Raw_cluster1 (R1) and Raw_cluster2 (R2) (Fig. 4a and d). R2, with strong mesenchymal stem cell (MSC) traits [48], was classified as an MSC-like tumor region, while R1 was characterized as non-MSC-like (Fig. 4e).

Further analysis using DscSTAR highlighted dynamic T cell features. Processed data showed a detailed subdivision including peripheral parts of cluster C1, C5, and inner cluster C3. These areas corresponded to untreated R1 and displayed a gradient decrease in progenitor exhausted T cell signals with an increase in Tex T cell signals inward (Fig. 4f). Additionally, changes in TAM signals were noted, suggesting the formation of a T-cell-rich area along the edges of the non-MSC-like tumor region, near TLS (Fig. 4a). This pattern was not detected in the unprocessed data (Fig. 4a and d).

DscSTAR uncovered interactions between MSC-like cancer cells and T/TAM cells in ST data, specifically identifying regions C2, C4, and C6 in the untreated R2 (MSC-like tumor region). Marker gene analysis showed T/TAM cell signals predominantly in C4, plasma cells in C2, and MSC-like cancer cells in C6 (Fig. 4g, Supplementary Table S3). CellChat [49] analysis revealed significant communication between C4 (T/TAM) and C6 (MSC-like cancer cell) primarily via FN1 and CD99 pathways (Fig. 4h and i, Supplementary Fig. S7, and Supplementary Table S3), highlighting complex cell interactions in RCC with prognostic implications—unseen with raw data (Supplementary Fig. S7).

Independent RCC scRNA-seq data [50] corroborated these findings, identifying two cancer cell types, TP1 and TP2, with TP2 showing MSC traits and high FN1 and CD99 expression (Supplementary Fig. S8, Fig. 4k). CellChat analysis showed TP2 cells had high FN1 and CD99 ligand activity, impacting progenitor exhausted CD8+ T cells and M2-type TAMs via CD44 and PILRA, respectively (Fig. 4j, Supplementary Fig. S9). Survival analysis from TCGA for various RCC types indicated poor outcomes associated with high FN1 and CD99 levels (Fig. 4l).

These results suggest that MSC-like cancer cells in RCC might enhance immune suppression through FN1 interactions with CD8+ T cells and influence TAM polarization via FN1 and CD99. FN1's link to poor prognosis and increased aggressiveness is established in other cancers such as thyroid [51] and breast [52] and correlates with M2-type macrophage infiltration in gastric cancer [53]. This DscSTAR-based insight underscores FN1 and CD99 as crucial biomarkers for immune infiltration and prognosis in RCC.

Unveiling neutrophil-CAF interactions linked to immunotherapy resistance in hepatocellular carcinoma

DscSTAR was used on scRNA-seq and ST data from an HCC multi-omics dataset [54] to analyze neutrophil characteristics and spatial distribution in HCC. We began by isolating neutrophils from the scRNA-seq data, categorizing tumors, and adjacent normal tissues for DscSTAR processing. DscSTAR identified subtypes such as S100A12+ (Neu_C1), VEGFA+ (Neu_C2), and CXCR2+ (Neu_C4) (Fig. 5a and b) linked to poor prognosis in HCC [55]. Hypergeometric tests correlated these DscSTAR clusters with original data (Fig. 5c), revealing high Neu_C1 marker expression associated with poorer survival in the The Cancer Genome Atlas - Liver Hepatocellular Carcinoma (TCGA-LIHC) cohort (Fig. 5d, Supplementary Table S4).

In an independent HCC ICB RNA-seq dataset [56], Gene Set Variation Analysis (GSVA) showed higher levels of Neu_C1 in non-responders (Fig. 5e), implicating these neutrophils in immunotherapy resistance. Furthermore, using the ESTIMATE R package [57], we analyzed the TCGA-LIHC cohort's StromalScores, finding higher Neu_C1 levels in samples with elevated stromal scores (Fig. 5f), indicating potential interactions between Neu_C1 and stromal cells within the TME.

Subsequently, DscSTAR was utilized to enhance Neu_C1 signals in HCC ST data, examining the spatial distribution of Neu_C1. Samples were divided based on the expression of neutrophil markers FCGR3B and CSF3R for DscSTAR enhancing neutrophil signals in spatial data (Supplementary Fig. S10). Enhanced data revealed pronounced Neu_C1 signals at tumor margins in non-responders, a pattern absent in responders (Fig. 5g and h). In non-responders, clusters with high levels of SPP1+ macrophages/CAFs showed elevated Neu_C1 and CAF scores, indicating active interactions (Fig. 5i). Neu_C1 and CAF features correlated significantly in non-responders ($R = 0.9$, $P < 2.2 \times 10^{-16}$) but not in responders ($P = .099$, Fig. 5j), underscoring their potential role in immunotherapy resistance. Clearly, analysis relying solely on raw data may not elucidate the significance of Neu_C1 and its interactions with CAFs (Supplementary Fig. S11).

Further analysis using NicheNet [58] explored Neu_C1 and CAF interactions, revealing high ligand activity for S100A4, B2M, and IL1B. These ligands bind to receptors on CAFs, driving upregulation of extracellular matrix (ECM)-related genes like COL1A1, COL4A1, and TIMP1 (Supplementary Fig. S12). Pathway

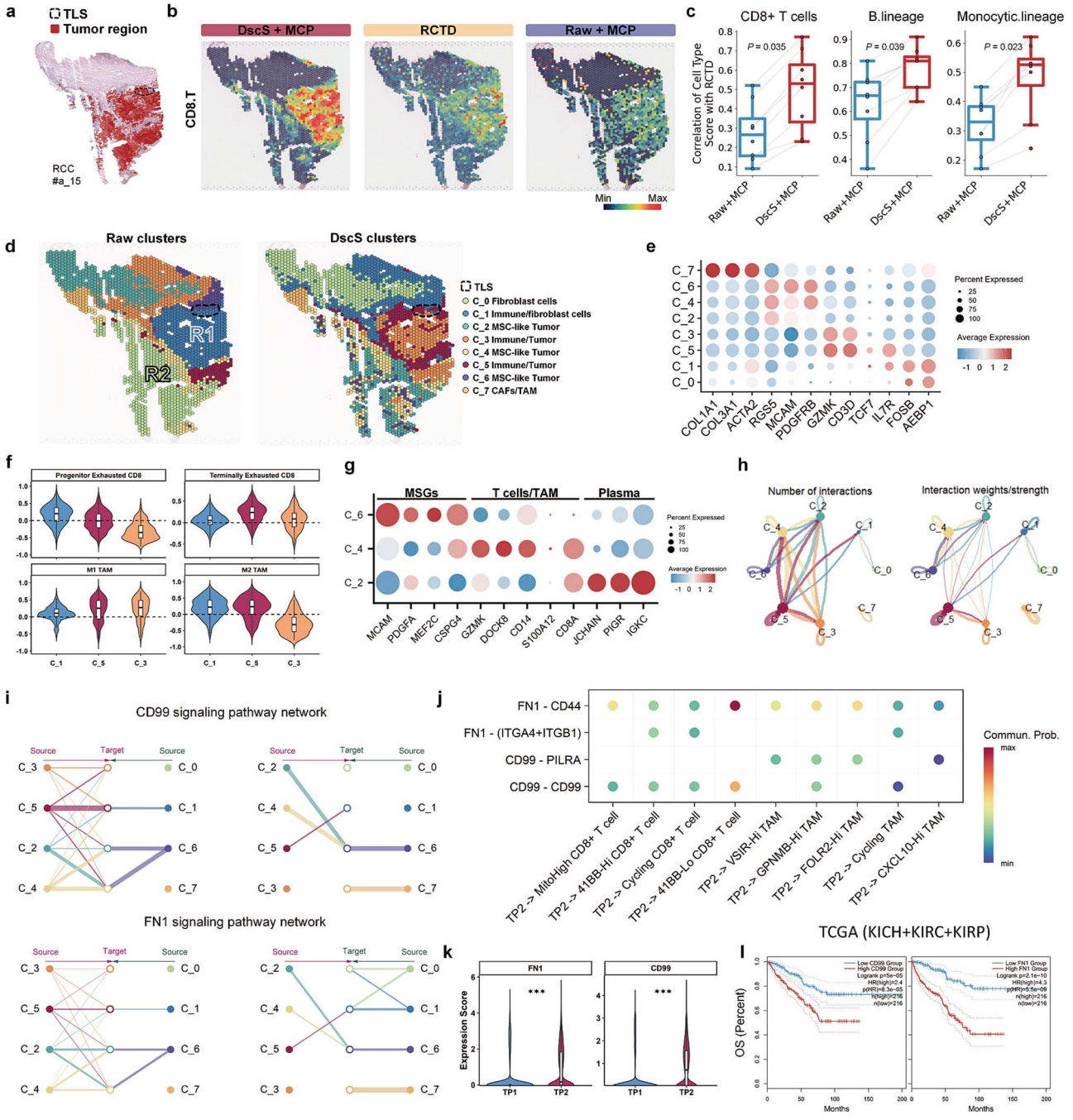

Figure 4. Unveiling tumor-immune interactions in RCC through enhanced ST analysis. (a) Hematoxylin and eosin (H&E) stained RCC tumor tissues, annotated using the "TESLA" Python package, annotations highlight tumor regions and TLS areas. (b) Feature plots showing improved visualization of MCP-counter scores of DscSTAR-treated samples compared to untreated samples. (c) Box plots displaying improved correlation between MCP-counter scores and RCTD baselines in DscSTAR-treated samples versus untreated samples; the tested cell types include CD8+ T cells, B cells, and monocytes, with significance assessed using the two-sided Wilcoxon rank-sum test. (d) Clustering of ST spots showing cell type definitions in DscSTAR-processed data compared to original R1/R2 regions. (e) Dot plots of marker gene expression in clusters, with dot size indicating cell proportion and color showing expression level. (f) Violin plots of CD8+ T cell and TAM signature scores across spatial sections, with median indicated by dashed line. (g) Dot plot showing marker gene expression across ST clusters in the untreated R2 region. Dot size represents the proportion of cells expressing the marker, and color indicates expression level. (h) Interaction counts and weights between spatial clusters. (i) Expression comparison of FN1 (top) and CD99 (bottom) pathways between spatial clusters. (j) Dot plot showing interactions between TP2 cancer cells and immune cells via FN1 and CD99 pathways, affecting progenitor exhausted CD8+ T cells and M2-type TAMs. (k) Expression comparison of FN1 and CD99 in TP1 and TP2 cancer cell types from RCC scRNA-seq data; significance was tested with a two-sided Wilcoxon rank-sum test. (l) Survival curves for various RCC types from TCGA datasets, analyzed using a two-tailed log-rank test.

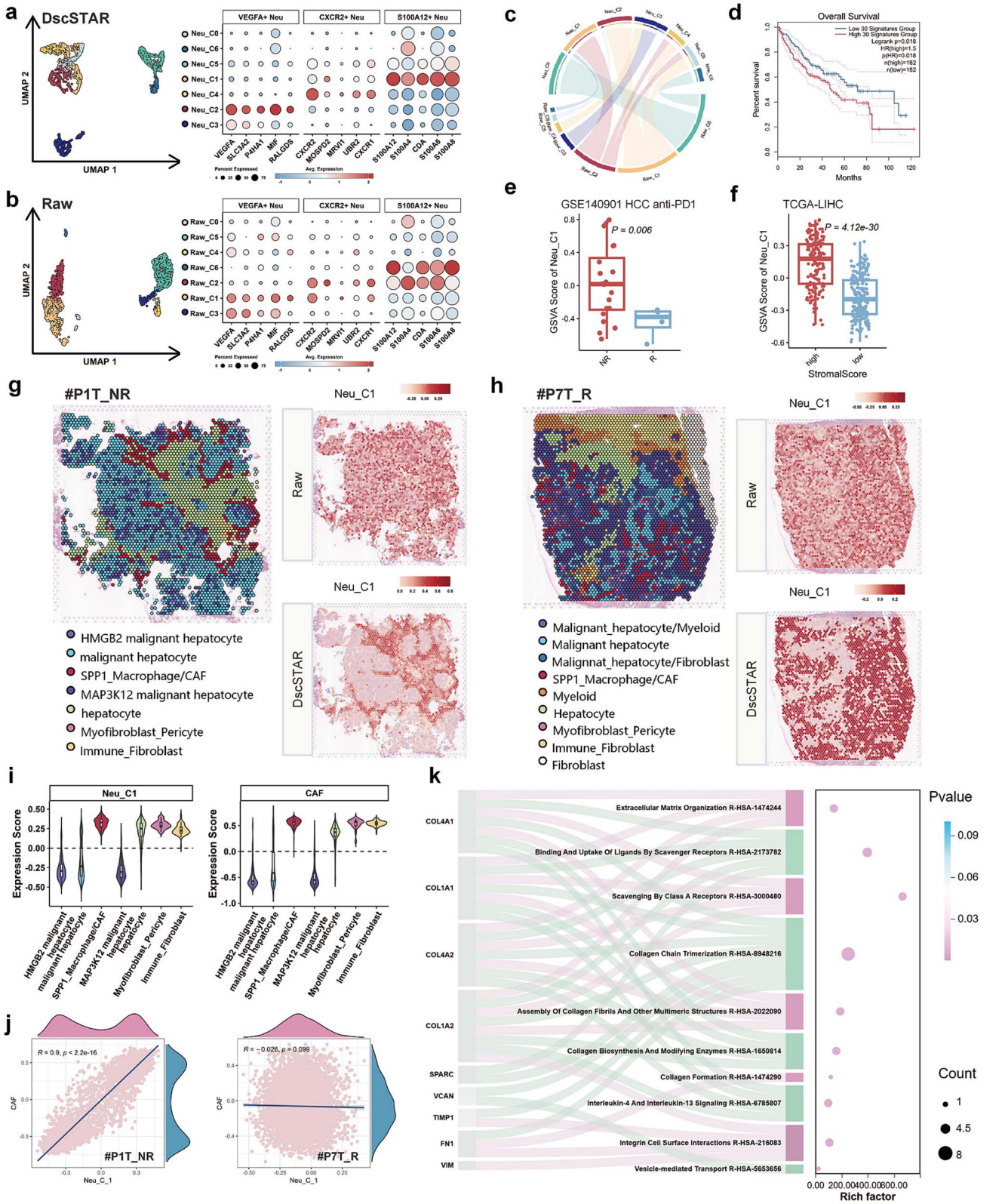

Figure 5. Neutrophil-CAF interactions in hepatocellular carcinoma and their impact on immunotherapy resistance (a) and (b) unbiased clustering of neutrophils in DscSTAR-processed (a) and original (b) data with dot plots. (c) Hypergeometric test associates DscSTAR-clusters with raw-clusters ($P < .01$). (d) Kaplan–Meier curves correlate high expression of Neu_C1 markers with poorer survival in TCGA-LIHC by two-tailed log-rank test. (e) Box plot comparing Neu_C1 levels in immunotherapy non-responders versus responders, significance tested with Welch's t-test. (f) Box plot showing differences in Neu_C1 levels between high and low StromalScore groups in TCGA-LIHC, tested with two-sided Wilcoxon rank-sum test. (g) and (h) Clustering of ST spots in ICB-treated non-responders' (g) and responders' (h) tissues shows spatial distribution of Neu_C1, enriched at tumor margins in non-responders. (i) Signature scores for Neu_C1 and CAF features across spatial clusters, median indicated by a dashed line. (j) Scatter plots of Neu_C1 and CAF feature correlations in non-responder samples, highlighting significant associations absent in responders. (k) Enrichr pathway analysis reveals REACTOME pathways in ECM organization activated by Neu_C1-CAF interactions.

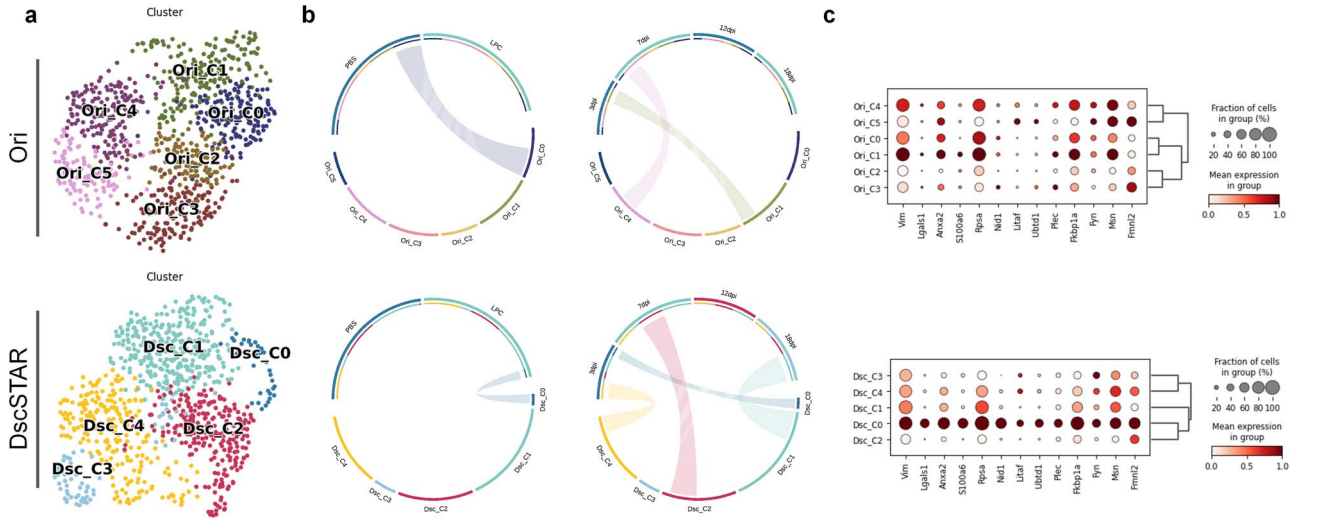

Figure 6. Identification of LPC-responsive endothelial cell subpopulations using DscSTAR (a) UMAP embeddings of endothelial cells from PBS (control) and LPC conditions. The top row represents clustering results from the standard workflow (Ori), while the bottom row represents DscSTAR-processed data. Colors indicate Leiden cluster assignments. (b) Associations between Leiden clusters and experimental conditions, as well as Leiden clusters and time points, evaluated using a hypergeometric test ($P < .01$). The top panel represents results from the standard workflow (Ori), while the bottom panel shows results from DscSTAR-processed data. (c) Dot plot showing the mean expression levels of marker genes associated with endothelial subpopulations affected in the early stages of demyelination. The top panel represents Leiden clusters from the standard workflow (Ori), while the bottom panel represents clusters identified by DscSTAR.

enrichment analysis with Enrichr [59] confirmed the involvement of these interactions in ECM organization (Fig. 5k). This suggests that Neu_C1 and CAFs within the tumor immune barrier (TIB) promote an immunosuppressive microenvironment, potentially impacting resistance to immunotherapy.

Identification of LPC-responsive endothelial cell subpopulations using DscSTAR

We then applied DscSTAR to single-nucleus RNA-seq profiles from a time-resolved dataset of a mouse model of demyelination to identify endothelial cell subpopulations associated with subtle perturbations. In this experiment, case animals received a corpus callosum injection of lysophosphatidylcholine (LPC), a compound toxic to oligodendrocytes, while control animals were injected with phosphate-Buffered Saline (PBS). Previous studies using in situ hybridization experiments identified the 3-day post-injection (3 dpi) time point in the LPC group as lesion-specific, characterized by the highly specific expression of Lgals1 and S100a6. These endothelial cells were identified as a key subpopulation responding to LPC-induced perturbation [16].

A standard dimensionality reduction workflow produced a uniform distribution of sample-level labels in the latent space, and clustering did not reveal any perturbation-enriched subpopulation (Fig. 6a and b). Subsequently, we applied DscSTAR to process both case and control groups, resulting in the identification of 5 groups (Fig. 6a). Using a hypergeometric test, we identified a subpopulation, Dsc_C0, that was strongly associated with the 3 dpi time point in the LPC group (Fig. 6b). Further analysis using previously reported marker genes for endothelial cells responsive to LPC perturbation confirmed that Dsc_C0 exhibited highly specific expression of these markers, whereas clusters identified by the standard workflow showed no specificity (Fig. 6c). This demonstrates the ability of DscSTAR to identify key phenotype-related subpopulations.

Ablation study on loss terms in deep scSTAR

The objective function of DscSTAR includes three loss terms: reconstruction loss, classification loss, and orthogonal loss (Materials and methods section). Ablation tests were performed to evaluate the impact of each term, using simulated datasets from 2C_1.2. Reconstruction loss, which is essential for the DAE model, was retained in all tests.

First, we evaluated the classification loss, which enhances phenotypic-associated factors in paired conditions. A comparison of F1 score, ARI, and ASW with and without the classification loss revealed significantly higher metrics in the default model with the classification loss (Supplementary Fig. S13).

Next, we assessed the orthogonal loss, which ensures the independence of the learned features and reduces distortions in the decoder output. The default model incorporating orthogonal loss exhibited higher ARI, with slight improvements in F1 and ASW. Further testing on RCC ST data [60] demonstrated that orthogonal loss preserved and reinforced biological signals, while its absence resulted in distortions (Supplementary Fig. S13).

Discussion

In this study, DscSTAR effectively dissected complex tumor microenvironment interactions, advancing our understanding of immune responses and ST. It identified key immune cell subsets and interactions linked to therapy resistance in cancers such as NSCLC, RCC, and HCC, highlighting its broad applicability and potential to uncover therapeutic targets.

Despite its advancements, DscSTAR has certain limitations. Its generalizability to complex or continuous phenotypes remains a challenge, as our simplification strategy of categorizing continuous phenotypes into binary classifications (e.g., “high” or “low”) may lead to the loss of nuanced information. However, this transformation improves interpretability and facilitates the identification of key subpopulations. Our applications in HSP-related T cells show that DscSTAR still effectively captures meaningful biological signals despite this simplification. Furthermore, future methodological developments may explore alternative classification approaches to better accommodate complex or continuous phenotype distributions.

Additionally, the method relies on high-quality phenotype labels and assumes the presence of phenotype-associated critical cell subpopulations, which may be subjective or incomplete in real-world datasets. Unbalanced phenotype samples further challenge the model's ability to detect rare subtypes or maintain performance under severe data imbalance. Lastly, the lack of orthogonal experimental validation, such as functional assays or proteomics, limits the biological confirmation of computational findings. Addressing these limitations through algorithmic refinements, validation experiments, and expansion to other omics datasets remains a priority for future work.

Looking ahead, we are exploring algorithmic enhancements to expand DscSTAR's capabilities for handling more intricate and continuous phenotypic analyzes. Furthermore, we aim to address its dependency on phenotype quality by developing approaches that can infer key features from less curated data. Finally, extending its applicability to other omics datasets, such as ATAC-seq and proteomics, remains a priority for future work.

Materials and methods

Concept of deep scSTAR

The original scSTAR method used a PLS model to extract differential features between two conditions, represented by matrices X and Y. The relationship between them was modeled as:

$$Y = \hat{X} + V \tag{1}$$

V and $\hat{X}$ are matrices with the same dimension as Y. $\hat{X}$ is the projection of Y in the control feature space. V represents the state differentiation matrix from $\hat{X}$ to Y. According to our previous derivation [17] and based on the assumption that V and X are unrelated, $\hat{X}$ and X have identical distributions, V can be obtained by $V = Y - P\left(P^{\mathrm{T}}P\right)^{-1}P^{\mathrm{T}}Y$, and P can be obtained by:

$$P = \underset{C^{\mathrm{T}}C=1,\ P^{\mathrm{T}}P=1}{\operatorname{argmax}} \operatorname{Cov}(XC, YP) \tag{2}$$

In the scSTAR method, the solution of Equation (2) was achieved by PLS, and C and P denote the PLS loading matrices of X and Y, respectively. We modify equation (2) as:

$$\underset{C^{\mathrm{T}}C=1,P^{\mathrm{T}}P=1}{\operatorname{argmax}} \operatorname{Cov}(XC, YP) = \underset{C^{\mathrm{T}}C=1,P^{\mathrm{T}}P=1}{\operatorname{argmax}} \rho(XC, YP)\,\sigma(XC)\sigma(YP) \tag{3}$$

where $\rho(XC, YP)$ is the Pearson correlation coefficient between the transformed cell states XC and YP. $\sigma(XC)$ and $\sigma(YP)$ are the standard deviations of XC and YP when we consider C and P represent the first eigenvectors of X and Y, respectively.

The optimization of equation (3) helps in emphasizing the features in X and Y that are most associated, thereby facilitating a insightful analysis of the cell state differentiation. However, PLS model has limited capability in dealing with complex and nonlinear relationships between cells. Therefore, the performance tends to compromise when processing over a few thousands of cells.

In this study, we aim to use deep learning model to improve the big data processing capability. New algorithm is named as DscSTAR. In DscSTAR, a combination of multiple deep learning models was designed to optimize the equation (3). First, to maximize standard deviation terms is equivalent to minimize the reconstruction error:

$$\min_{C^{\mathrm{T}}C=1} \left\| X - \hat{X} \right\|^2 = \min_{C^{\mathrm{T}}C=1} \left\| X - XCC^{\mathrm{T}} \right\|^2 \iff \max_{C^{\mathrm{T}}C=1} \sigma^2(XC) \tag{4}$$

$$\min_{P^{\mathrm{T}}P=1} \left\| Y - \hat{Y} \right\|^2 = \min_{P^{\mathrm{T}}P=1} \left\| Y - YPP^{\mathrm{T}} \right\|^2 \iff \max_{P^{\mathrm{T}}P=1} \sigma^2(YP) \tag{5}$$

In the realm of deep learning, the autoencoder (AE) stands out as notable tools. A DAE is specifically designed to learn a compressed representation of the input data through an encoder-decoder structure, with the added capability to reduce noise. It captures the main variation directions in the data by minimizing the mean squared error (MSE) between the input and output, which can be used to optimize Equations (4) and (5) taken X and Y as inputs and the reconstructed data $\hat{X}$ and $\hat{Y}$ as outputs. Here, $\hat{X}$ and $\hat{Y}$ denote the data reconstructed by the AE encoder and decoder networks.

Although Maximum Mean Discrepancy is a popular way to maximize the correlation coefficient in Equation (3), recognizing the necessity to capture phenotype-associated features more directly, we opted for a more targeted approach using a classification model, specifically MLP. This choice allows for a more precise differentiation of phenotype-associated components in the data, making the DscSTAR framework a robust tool for analyzing complex linear and nonlinear relationships between cell states across different conditions.

Deep learning model construction using DAE and MLP

DscSTAR uses a DAE for data encoding and a MLP for phenotype prediction. DAE is employed to learn a low-dimensional representation from the single cell expression matrix X_s. It consists of an encoder E_s that maps the input to a lower-dimensional space Z_s, and a decoder D_s that reconstructs the input from this reduced representation. In this study, the encoder has layers with 5120, 1024, and 512 neurons, while the decoder has layers with 512, 1024, and 5120 neurons, producing the input feature count. The ELU activation function is used, with dropout rates of 0.2 and 0.1 for the encoder, and 0.05 and 0.0 for the decoder. The model can be trained as:

$$Z_s = E^s\left(B\left(X_s, p^b\right)\right) \tag{6}$$

$$X_s{}' = D^s\left(Z_s\right) \tag{7}$$

$$\begin{aligned} \min \operatorname{loss}_{\mathrm{recon}}\left(E_s, D_s, X_s\right) &= \min \mathrm{MSE}\left(X_s, X_s{}'\right) \\ &= \min \frac{1}{N} \sum_{i=1}^{N} \left\| X_{s_i} - X_{s_i}{}' \right\|^2 \end{aligned} \tag{8}$$

where B is a noise adder that introduces random binomial noise to the single-cell expression matrix X_s, creating a noised expression matrix $B\left(X_s, p^b\right)$. The parameter p^b dictates the probability of having zeros in each row of the noise matrix. The input matrix X_s has dimensions corresponding to the number of genes by the number of cells. The encoder maps this noised input to a 512-dimensional feature vector Z_s, and the decoder reconstructs the expression matrix $X_s{}'$ from Z_s, maintaining the same dimensions as the original input matrix X_s. Simultaneously, a classifier (C_s) based on a MLP is applied to predict phenotype labels of each cell. The MLP takes the latent representation Z_s as input, which encoded by the DAE's encoder, and outputs the predicted phenotype labels Y_s'. The MLP consists of a single linear layer with an input dimension of 512 (corresponding to the output of the DAE) and an output dimension of 2 (representing the two phenotype classes). The phenotype labels have been normalized to fall within

the range of −1 to 1. Parameters inside the MLP are optimized using the MSE loss between the predictive phenotype score Y'_s and the true phenotype labels Y_s, accommodating the adjusted label range.

$$\begin{aligned}\text{minloss}_{\text{class}}(C_s, Z_s) &= \text{minMSE}\left(Y_s, Y'_s\right) \\ &= \min \frac{1}{N}\sum_{i=1}^{N} \| Y_{s_i} - Y_{s_i}{}' \|^2 \end{aligned} \tag{9}$$

$$Y'_s = C_s(Z_s) \tag{10}$$

This loss function ensures that the MLP model learns to approximate the true labels as closely as possible by minimizing the squared error between predicted and actual labels.

Joint training and loss optimization

The MTL can be trained by minimizing the equation (11):

$$\begin{aligned}\text{loss}_{\text{MTL}} &= \gamma_1 \cdot \text{loss}_{\text{recon}}(E_s, D_s, X_s, B) \\ &\quad + \gamma_2 \cdot \text{loss}_{\text{class}}(C_s, Z_s, Y_s) + \gamma_3 \cdot \text{loss}_{\text{orth}}(Z_s)\end{aligned} \tag{11}$$

$$\text{loss}_{\text{orth}}(Z_s) = \| G - I\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(G_{ij} - I_{ij}\right)^2 \tag{12}$$

$$G = Z_s Z_s{}^T \tag{13}$$

Here matrix G with elements g_{ij} representing the inner product between the i^{th} and j^{th} cells in the latent representation. I indicates an identity matrix. The orthogonality regularization loss $\text{loss}_{\text{orth}}$ is obtained by squaring the Frobenius norm of the difference between the Gram matrix and the identity matrix, denoted as $\| G - I\|_F^2$. Here $\|\cdot\|_F$ represents the Frobenius norm, and I_{ij} are the elements of the identity matrix. In this formulation, $G - I$ represents the difference between the Gram matrix and the identity matrix. The goal of the orthogonality regularization loss is to minimize the Frobenius norm of $G - I$. The model is trained for 400 epochs (Default) with a learning rate that follows a step decay schedule, beginning at 1e−3. Separate optimizers are used for the DAE and the classifier to accommodate their respective parameters.

Deep scSTAR workflow design

In realistic settings, variation V encompasses both relevant signals and noise [61]. V in each cell is a linear combination of distinct components, as described in [16]:

$$V = V_{\text{batch}} + V_{\text{noise}}^{r+b} + V_{\text{signal}} \tag{14}$$

where V_{batch} signifies the batch effect. V_{noise}^{r+b} is a composite of both random noise, which includes technical aspects r, and biological noise b, denoting the disturbances that do not pertain to the differentiation between the two conditions under comparison. V_{signal} denotes the variations in gene expression that are observed between the conditions being investigated, representing the phenotype-related features. The work-flow of DscSTAR includes the following three parts:

Step1. Identifying unchanged cells across phenotypes. We used scCURE [19] to identify unchanged cells, defined as cells with identical biological characteristics across conditions, differing only due to batch effects and noise. This step involves: (i) The utilization of GMM aids in the precise identification and categorization of different cell clusters, unveiling the heterogeneity within cells. (ii) Calculating the KL divergence between two GMM (each from a phenotype) to identify the unchanged cell [19].

Step2. Training a PLS-DA model on unchanged cells to mitigate batch effect and random noise. Training a PLS-DA model on unchanged cells to remove batch effect and noise. The PLS-DA model is used to extract batch effects and noise from unchanged cells. The model minimizes errors with:

$$V_{\text{batch}} + V_{\text{noise}}^{r} = X p_{\text{PLS}} p_{\text{PLS}}^{-1} \tag{15}$$

$$V' = X - V_{\text{batch}} + V_{\text{noise}}^{r} \tag{16}$$

where Y represents the observed data vector for a cell. The PLS loading matrix for X, denoted as p, comprises m PLS components. This procedure is the same as what is presented in the original scSTAR algorithm. The only difference lies in that the "anchor cells" in the original scSTAR are replaced by "unchanged cells" here.

Step3. Supervised MTL model reconstruct the noise reduced data to extract phenotype associated info matrix. A MTL model is formulated by incorporating cells from both groups. Given that the noise term, V_{noise}^{b} in Equation (13), does not contribute to distinguishing between the two conditions (or phenotypes) under comparison, MTL is specifically designed to isolate the V_{signal} variation. The signal component is discerned from V as follows:

$$\hat{V}_{\text{signal}} = D_{\text{MTL}}\left(E_{\text{MTL}}\left(V'\right)\right) \tag{17}$$

In this equation, $\hat{V}_{\text{signal}}$ represents the estimated cell dynamic features. E_{MTL} is the encoder of MTL and D_{MTL} is the Decoder of MTL. Consequently, all variation components not pertinent to the signal of interest are filtered out from $\hat{V}_{\text{signal}}$.

Evaluation metrics

Three metrics were used to assess DscSTAR on simulated datasets:

Adjusted rand index

Measures similarity between clustering assignments, ranging from −1 to 1. A score of 1 indicates perfect agreement, 0 represents random assignment, and negative values show disagreement. The ARI formula is:

$$\text{ARI} = \frac{\sum\binom{n_{ij}}{2} - \left[\sum\binom{a_i}{2}\sum\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum\binom{a_i}{2} + \sum\binom{b_j}{2}\right] - \left[\sum\binom{a_i}{2}\sum\binom{b_j}{2}\right]\Big/\binom{n}{2}} \tag{18}$$

Average silhouette widthAssesses clustering quality, measuring how similar an object is to its own cluster versus other clusters. ASW ranges from −1 to 1. The overall ASW for a dataset is the average of the silhouette width of all samples. A value close to 1 indicates good clustering, and values near −1 suggest misclassification:

$$\text{ASW} = \frac{1}{N}\sum_{i=1}^{N} s(i) \tag{19}$$

F1 score (for DE gene evaluation)

Evaluates the accuracy of identifying DE genes, balancing Precision and Recall. The F1 score ranges from 0 to 1, with 1 representing perfect identification. The formula is:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

where:

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN}\end{aligned} \quad . \tag{21}$$

scRNA-seq data integration, quality control

Cell clusters were identified using cell marker gene expression. Normalization of the data was performed via Seurat by adjusting UMI counts and log-transforming the normalized data. Integration of NSCLC datasets employed Scanpy's BBKNN method. For other datasets, batch integration was conducted using the Harmony R package.

Unsupervised cell clustering and subclustering analysis of scRNA-seq datasets

After HVG filtering (excluding mitochondrial, ribosomal, and T cell receptor genes), clustering and subclustering were performed using FindNeighbors and FindClusters to create a shared nearest neighbor graph, followed by UMAP for visualization. Optimal clusters were evaluated by UMAP and cluster markers. Gene signature scores for each cluster were calculated using Seurat's AddModuleScore and GSVA R package.

Single-cell trajectory inference

We used scTour to infer the differentiation state of $CD8^+$ T cells.

Mapping the spatial locations of tumor region

To map RCC tumor regions in histological images, we used the TESLA Python package (version 1.2.2). The cv2_detect_contour function outlined tissue boundaries, and gene expression was enhanced at the super-pixel level using imputation ($s = 1$, $k = 2$, num_nbs = 10). RCC markers (CA9, SLC17A3, NDUFA4L2, BUB1B, TOP2A, and ELF3) annotated tumor sites, which were visualized on histological images. This process yielded 12 annotated ST image slices, including TLS regions from the original study.

Single cell types mapping to spatial data

For the scRNA-seq dataset from ST, cell types with fewer than 25 cells were excluded for RCTD analysis. Using the spacexr R package (version 2.2.0), RCTD was performed on retained cell types and Visium ST data, with doublet mode activated. The 25-cell minimum ensures reliable deconvolution, and CD8+ T cell scores were normalized and aggregated to calculate the pan-CD8+ T cell score in the tumor microenvironment.

Spatial co-location analysis

Spatial co-location analysis was performed for each tumor by correlating gene or signature expression with Pearson's test. Expression levels were classified as "high" or "low" based on the median. Significant correlations and "high-high" spots indicated spatial co-location.

Receptor-ligand interaction and ligand-target inference

For the RCC dataset, receptor-ligand interactions were predicted using CellChat. In HCC, NicheNet was used to examine receptor-ligand impacts on gene expression with default parameters.

Statistics and reproducibility

Statistical analyzes for NSCLC datasets were performed in Python (version 3.10.12). Differential expression between cell groups was assessed using a two-sided Wilcoxon rank-sum test with Benjamini–Hochberg FDR correction. For other analyzes, R (version 4.2.2) was used. Differential expression between cell groups and comparisons of signature scores were performed with a two-sided Wilcoxon rank-sum test and Bonferroni FDR correction. Survival analysis for bulk RNA-Seq samples stratified by signature score used a log-rank test. Details of statistical tests are provided in Figure legends and Methods.

Key points

- Extracting meaningful phenotype-related features from single-cell RNA sequencing data is challenging due to the presence of noise and biologically irrelevant signals.
- Deep scSTAR utilizes deep learning to enhance key phenotype-associated cellular traits from single-cell data affected by noise and irrelevant biological signals.
- Detailed cell heterogeneities and phenotype-associated cell subtypes were revealed in various datasets, advancing the investigation of targeted biological questions.

Acknowledgements

The authors thank the technical support of National Engineering Center for Biochip at Shanghai, Shanghai Biochip Co., Ltd.

Authors' contributions

J.H., X.Z., L.G., and H.L. conceived and designed the algorithm. L.G. developed the software. J.H., X.Z., L.G., and H.L. provided data interpretation and biological explanation. J.H., X.Z., L.G., and Y.L. wrote the manuscript with feedback from all other authors. L.G., J.H., and X.Z. prepared the figs. Z.Z., Z.L, L.C., Y.J., H.Y.T., Y.L., J.Z., F.D., and XR.Z. provided scientific insights on the applications. J.H. and X.Z. supervised the study. All authors read and approved the final manuscript.

Supplementary data

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

Conflict of interest: None declared.

Funding

This work was supported in part by the National Natural Science Foundation of China [82170045 to JH]; the Special Fund for Scientific Research of Shanghai Landscaping & City Appearance Administrative Bureau [G222410 to JH and XZ]; the Translational Medicine Cross Research Fund of Shanghai Jiao Tong University [ZH2018QNB29 to JH]. The Innovative Research Team of High-level Local Universities in Shanghai [SHSMU-ZLCX20212301 to JH]; the Fundamental Research Funds for the Central Universities [21X010301839,23X010300645 to HL].

Data availability

The datasets supporting this study are available in public repositories. NSCLC scRNA-seq data are under GSE179994 (Liu et al., 2022), with validation data under GSE126044. SKCM HSP+ FKBP4+ CD8+ T cell data are under GSE222448, and BCC post-ICB data under GSE123814. RCC single-cell data are from SCP1288 (Bi et al., 2021), with spatial transcriptomics under GSE175540. HCC

single-cell and spatial data are on Mendeley Data (ID skrx2fz79n), with RNA-array validation post-ICB under GSE140901.

Code availability

The Deep scSTAR algorithm is available and can be accessed at the following GitHub repository: https://github.com/Hao-Zou-lab/Deep-scSTAR. Additionally, scripts for reproducing results in this article are also provided in the repository. The hyperparameters, datasets and required computational resources needed to reproduce results are also available in the GitHub repository.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Jiang X, Wang Y, Xiao Z. *et al.* A differentiation roadmap of murine placentation at single-cell resolution. *Cell Discov* 2023;**9**:30. https://doi.org/10.1038/s41421-022-00513-z.
2. Ren X, Wen W, Fan X. *et al.* Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021;**184**:1895–913. https://doi.org/10.1016/j.cell.2021.01.053.
3. Chu Y, Dai E, Li Y. *et al.* Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nat Med* 2023;**29**:1550. https://doi.org/10.1038/s41591-023-02371-y.
4. Oliveira G, Egloff AM, Afeyan AB. *et al.* Preexisting tumor-resident T cells with cytotoxic potential associate with response to neoadjuvant anti–PD-1 in head and neck cancer. *Sci Immunol* 2023;**8**:eadf4968. https://doi.org/10.1126/sciimmunol.adf4968.
5. Tian Y, Carpp LN, Miller HE. *et al.* Single-cell immunology of SARS-CoV-2 infection. *Nat Biotechnol* 2022;**40**:30–41. https://doi.org/10.1038/s41587-021-01131-y.
6. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**:35–45. https://doi.org/10.1038/nri.2017.76.
7. Zhang Z, Wang Z, Chen Y. *et al.* Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. *Genome Med* 2022;**14**:45. https://doi.org/10.1186/s13073-022-01050-w.
8. Trapnell C, Cacchiarelli D, Grimsby J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6. https://doi.org/10.1038/nbt.2859.
9. Korsunsky I, Millard N, Fan J. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96. https://doi.org/10.1038/s41592-019-0619-0.
10. Lin Y, Ghazanfar S, Wang KY. *et al.* scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci* 2019;**116**:9775–84. https://doi.org/10.1073/pnas.1820006116.
11. Lin Y, Cao Y, Willie E. *et al.* Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. *Nat Commun* 2023;**14**:4272. https://doi.org/10.1038/s41467-023-39923-2.
12. Haghverdi L, Lun AT, Morgan MD. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7. https://doi.org/10.1038/nbt.4091.
13. 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.
14. Welch JD, Kozareva V, Ferreira A. *et al.* Single-cell multiomic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**:1873–87. https://doi.org/10.1016/j.cell.2019.05.006.
15. Zhang L, Nie Q. Scmc learns biological variation through the alignment of multiple single-cell genomics datasets. *Genome Biol* 2021;**22**:10. https://doi.org/10.1186/s13059-020-02238-2.
16. Goeva A, Dolan M, Luu J. *et al.* Hidden: a machine learning method for detection of disease-relevant populations in case-control single-cell transcriptomics data. *Nat Commun* 2024;**15**:9468. https://doi.org/10.1038/s41467-024-53666-8.
17. Hao J, Zou J, Zhang J. *et al.* scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across conditions in single-cell RNA sequencing data. *Brief Bioinform* 2023;**24**:bbad62. https://doi.org/10.1093/bib/bbad062.
18. Sun Y, Zhang H, Zhang Y. *et al.* Li–Mg–Si bioceramics provide a dynamic immuno-modulatory and repair-supportive microenvironment for peripheral nerve regeneration. *Bioact Mater* 2023;**28**:227–42. https://doi.org/10.1016/j.bioactmat.2023.05.013.
19. Zeng F, Cao J, Hong Z. *et al.* Single-cell analyses reveal the dynamic functions of Itgb2^{+} microglia subclusters at different stages of cerebral ischemia-reperfusion injury in transient middle cerebral occlusion mice model. *Front Immunol* 2023;**14**:1114663. https://doi.org/10.3389/fimmu.2023.1114663.
20. Kha Q, Tran T, Nguyen T. *et al.* An interpretable deep learning model for classifying adaptor protein complexes from sequence information. *Methods* 2022;**207**:90–6. https://doi.org/10.1016/j.ymeth.2022.09.007.
21. Chen J, Wu Z, Qi R. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun* 2022;**13**:6494. https://doi.org/10.1038/s41467-022-34277-7.
22. Hao M, Gong J, Zeng X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91. https://doi.org/10.1038/s41592-024-02305-7.
23. Tran T, Vo TH, Le NQK. Omics-based deep learning approaches for lung cancer decision-making and therapeutics development. *Brief Funct Genomics* 2024;**23**:181–92. https://doi.org/10.1093/bfgp/elad031.
24. Fleming SJ, Chaffin MD, Arduini A. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using cellbender. *Nat Methods* 2023;**20**:1323. https://doi.org/10.1038/s41592-023-01943-7.
25. Zou X, Liu Y, Wang M. *et al.* scCURE identifies cell types responding to immunotherapy and enables outcome prediction. *Cell Rep Methods* 2023;**3**:100643. https://doi.org/10.1016/j.crmeth.2023.100643.
26. 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.
27. Santos JM, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification. In: *International Conference on Artificial Neural Networks*. Edinburgh, UK: Springer, 2009, pp. 175–84. https://doi.org/10.1007/978-3-642-04277-5_18.

28. Hu C, Yang J, Qi Z. *et al.* Heat shock proteins: biological functions, pathological roles, and therapeutic opportunities. *MedComm (2020)* 2022;**3**:e161. https://doi.org/10.1002/mco2.161.
29. Liu B, Zhang Y, Wang D. *et al.* Single-cell meta-analyses reveal responses of tumor-reactive CXCL13^{+} T cells to immune-checkpoint blockade. *Nat Cancer* 2022;**3**:1123. https://doi.org/10.1038/s43018-022-00433-7.
30. Liu B, Hu X, Feng K. *et al.* Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer. *Nat Cancer* 2022;**3**:108. https://doi.org/10.1038/s43018-021-00292-8.
31. Simoni Y, Becht E, Fehlings M. *et al.* Bystander CD8 $^{+}$ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature* 2018;**557**:575. https://doi.org/10.1038/s41586-018-0130-2.
32. Zong S, Jiao Y, Liu X. *et al.* FKBP4 integrates FKBP4 / Hsp90 / IKK with FKBP4 / Hsp70 / RelA complex to promote lung adenocarcinoma progression via IKK / NF-κB signaling. *Cell Death Dis* 2021;**12**:602. https://doi.org/10.1038/s41419-021-03857-8.
33. Zheng X, Hou Z, Qian Y. *et al.* Tumors evade immune cytotoxicity by altering the surface topology of NK cells. *Nat Immunol* 2023;**24**:802. https://doi.org/10.1038/s41590-023-01462-9.
34. Xiong W, Chen Y, Kang X. *et al.* Immunological synapse predicts effectiveness of chimeric antigen receptor cells. *Mol Ther* 2018;**26**:963–75. https://doi.org/10.1016/j.ymthe.2018.01.020.
35. Cho J, Hong MH, Ha S. *et al.* Genome-wide identification of differentially methylated promoters and enhancers associated with response to anti-PD-1 therapy in non-small cell lung cancer. *Exp Mol Med* 2020;**52**:1550–63. https://doi.org/10.1038/s12276-020-00493-8.
36. Li Q. Sctour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics. *Genome Biol* 2023;**24**:149. https://doi.org/10.1186/s13059-023-02988-9.
37. Jung H, Kim HS, Kim JY. *et al.* DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. *Nat Commun* 2019;**10**:4278. https://doi.org/10.1038/s41467-019-12159-9.
38. Barras D, Ghisoni E, Chiffelle J. *et al.* Response to tumor-infiltrating lymphocyte adoptive therapy is associated with preexisting CD8 $^{+}$ T-myeloid cell networks in melanoma. *Sci Immunol* 2024;**9**. https://doi.org/10.1126/sciimmunol.adg7995.
39. Yost KE, Satpathy AT, Wells DK. *et al.* Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat Med* 2019;**25**:1251. https://doi.org/10.1038/s41591-019-0522-3.
40. Williams CG, Lee HJ, Asatsuma T. *et al.* An introduction to spatial transcriptomics for biomedical research. *Genome Med* 2022;**14**:68. https://doi.org/10.1186/s13073-022-01075-1.
41. Jain S, Eadon MT. Spatial transcriptomics in health and disease. *Nat Rev Nephrol.* 2024;**20**:659–71. https://doi.org/10.1038/s41581-024-00841-1.
42. Meylan M, Petitprez F, Becht E. *et al.* Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing cells in renal cell cancer. *Immunity* 2022;**55**:527. https://doi.org/10.1016/j.immuni.2022.02.001.
43. Hu J, Coleman K, Zhang D. *et al.* Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst* 2023;**14**:404. https://doi.org/10.1016/j.cels.2023.03.008.
44. Cable DM, Murray E, Zou LS. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:517. https://doi.org/10.1038/s41587-021-00830-w.
45. Chen J, Liu W, Luo T. *et al.* A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbac245.
46. Li B, Zhang W, Guo C. *et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* 2022;**19**:662–70. https://doi.org/10.1038/s41592-022-01480-9.
47. Becht E, Giraldo NA, Lacroix L. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* 2016;**17**:218. https://doi.org/10.1186/s13059-016-1070-5.
48. Davidson G, Helleux A, Vano YA. *et al.* Mesenchymal-like tumor cells and myofibroblastic cancer-associated fibroblasts are associated with progression and immunotherapy response of clear cell renal cell carcinoma. *Cancer Res* 2023;**83**:2952–69. https://doi.org/10.1158/0008-5472.CAN-22-3034.
49. Jin S, Guerrero-Juarez CF, Zhang L. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021;**12**:1088. https://doi.org/10.1038/s41467-021-21246-9.
50. Bi K, He MX, Bakouny Z. *et al.* Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell* 2021;**39**:649. https://doi.org/10.1016/j.ccell.2021.02.015.
51. Geng Q, Huang T, Li L. *et al.* Over-expression and prognostic significance of FN1, correlating with immune infiltrates in thyroid cancer. *Front Med* 2022;**8**:812278. https://doi.org/10.3389/fmed.2021.812278.
52. Zhang X, Luo J, Wu L. FN1 overexpression is correlated with unfavorable prognosis and immune infiltrates in breast cancer. *Front Genet* 2022;**13**:913659. https://doi.org/10.3389/fgene.2022.913659.
53. Wang H, Zhang J, Li H. *et al.* FN1 is a prognostic biomarker and correlated with immune infiltrates in gastric cancers. *Front Oncol* 2022;**12**:918719. https://doi.org/10.3389/fonc.2022.918719.
54. Liu Y, Xun Z, Ma K. *et al.* Identification of a tumour immune barrier in the HCC microenvironment that determines the efficacy of immunotherapy. *J Hepatol* 2023;**78**:770–82. https://doi.org/10.1016/j.jhep.2023.01.011.
55. Wu Y, Ma J, Yang X. *et al.* Neutrophil profiling illuminates anti-tumor antigen-presenting potency. *Cell* 2024;**187**:1422–39. https://doi.org/10.1016/j.cell.2024.02.005.
56. Hsu C, Ou D, Bai L. *et al.* Exploring markers of exhausted CD8 T cells to predict response to immune checkpoint inhibitor therapy for hepatocellular carcinoma. *Liver Cancer* 2021;**10**:346–59. https://doi.org/10.1159/000515305.
57. Yoshihara K, Shahmoradgoli M, Martinez E. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;**4**:2612. https://doi.org/10.1038/ncomms3612.
58. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* 2020;**17**:159. https://doi.org/10.1038/s41592-019-0667-5.
59. Kuleshov MV, Jones MR, Rouillard AD. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7. https://doi.org/10.1093/nar/gkw377.
60. Brennecke P, Anders S, Kim JK. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013;**10**:1093–5. https://doi.org/10.1038/NMETH.2645.

Briefings in Bioinformatics, 2025, **26(3)**, bbaf160
https://doi.org/10.1093/bib/bbaf160
Problem Solving Protocol