Supplemental Methods

Deciphering context-specific gene programs from single-cell and spatial transcriptomics data with DeCEP

Lin Li[1], Xianbin Su[1], Ze-Guang Han[1,*]

Details of simulation

To benchmark the “discrete” mode of DeCEP, we simulated a single-cell gene expression matrix containing 1,000 cells and 20,000 genes at a time and set the *de.facLoc* parameter with gradients. To benchmark the “continuous” mode of DeCEP, we lowered the number of cells per simulation to 500, set the *path.nSteps* parameter to 50, and adjusted the *de.facLoc* parameter with gradients accordingly. Additionally, for both simulation modes, we set the *de.prob* parameter to 0.2. Here, the *de.facLoc* parameter represents the differential expression factor, which determines the extent of gene expression differences among simulated cells. The *path.nSteps* parameter establishes continuous cell steps and governs the smoothness of cell dynamic changes in the simulation. The *de.prob* parameter controls the probability that a gene is selected for differential expression, thus influencing the diversity of gene expression profiles across simulated cells.

We performed consensus clustering (Wilkerson and Hayes 2010) on the 1000 highly variable genes in the simulated gene expression matrix, setting the output to 10 clusters. We assumed that each cluster corresponds to a unique gene program, where genes are clustered together due to potential transcriptional co-regulation. The

collection of highly variable genes within each cluster thus represents a functional gene list. We evaluated the performance of different methods in characterizing 10 gene programs in each dataset.

Details of existing methods employed for benchmarking

Seurat. The R package Seurat (v4.3.0) was employed to calculate gene program activity scores. We executed Seurat using the *AddModuleScore* function, with all the parameters set to their default values.

AUCell. The R package AUCell (v1.20.2) was used to calculate gene program activity scores. We first applied the *AUCell_buildRankings* function, setting the parameter *splitByBlocks* to TRUE while keeping the other parameters at their default values. We subsequently executed the *AUCell_calcAUC* function using the default settings.

VISION. The R package VISION (v3.0.1) was used to calculate gene program activity scores. We sequentially executed the *Vision* function and the *analyze* function, using the default parameter settings for both.

VAM. The R package VAM (v1.0.0) was employed to calculate gene program activity scores. We used the *vamForSeurat* function, maintaining its default parameters.

UCell. The R package UCell (v2.0.1) was used to calculate gene program activity scores. We applied the *AddModuleScore_UCell* function with its default values.

Details of evaluation metrics

To evaluate the performance of the “discrete” mode of DeCEP, we first performed unsupervised clustering on the simulated datasets using the standard Seurat v4

workflow. Next, we used simulated gene programs as features and assigned cell clusters identified by the workflow with a resolution parameter of 0.6 as classification labels. We applied the *k*-nearest neighbors (kNN) algorithm with $k = 7$ to build kNN classifiers, using 10-fold cross-validation to predict the classification labels. The kNN classifier calculates the Euclidean distance between training and test samples and classifies test samples based on the majority label of their nearest neighbors. In this way, we assessed the effectiveness of activity scores for gene programs calculated by different methods in distinguishing cell clusters. We calculated the accuracy, sensitivity, specificity, F1 score, and AUC as our measures of the effectiveness of different methods to characterize gene programs. To eliminate the influence of predefined classification labels in the supervised evaluation, we performed an unsupervised assessment for each method. We used simulated gene programs as features for *k*-means clustering. The number of clusters for *k*-means clustering was fixed at two (i.e., $k = 2$). This binary clustering strategy provides a simplified and reasonable scheme for evaluating the effectiveness of gene program activity scores to cluster and separate cells. By fixing $k = 2$, we ensured a controlled evaluation strategy that allowed a fair comparison between DeCEP and existing methods in terms of their separation-and-clustering capability, thereby eliminating potential confounders introduced by the estimation of the number of clusters. For each simulated dataset, we calculated the silhouette coefficient (SC) for each cell and reported the mean SC value.

To evaluate the performance of DeCEP in its “continuous” mode, we used generalized additive models (GAMs) to fit the simulated cell steps on the gene program

activity scores and reported the goodness of fit. Specifically, we used the simulated cell steps as the response variable and the gene program activity scores as the explanatory variables to construct the nonparametric regression model. Here, the goodness of fit refers to the degree of closeness between the model's predicted values and the actual response variable. We employed the adjusted R-squared as our measure of the goodness of fit. We performed the fitting using the R package mgcv (v1.8-42), selected splines as the basis function with the parameter *bs = tp*, and set the smoothing parameter *method = REML*.

Collection of datasets and functional gene lists

The adult mouse liver scRNA-seq and ST data were downloaded from the Gene Expression Omnibus (GEO) with accession numbers GSE171993 and GSE192742. The scRNA-seq data for the hippocampus of AD and WT mice was downloaded from GSE143758, and the corresponding ST data from 10x Visium and Slide-seqV2 platforms were downloaded from GSE174321 and https://singlecell.broadinstitute.org/single_cell/study/SCP1663, respectively. The human cSCC scRNA-seq and ST data were downloaded from GSE144240. The functional gene lists used in this study were downloaded from the KEGG (Kanehisa et al. 2023) (https://www.genome.jp/kegg/pathway.html), WikiPathways (Agrawal et al. 2024) (https://www.wikipathways.org/), and Reactome (Milacic et al. 2024) (https://reactome.org/) databases.

Clustering

We performed unsupervised clustering of scRNA-seq data using the standard Seurat v4 (Hao et al. 2021) workflow, which employs a graph-based clustering approach. The identification of spatial domains from ST data was performed with STAGATE (Dong and Zhang 2022).

Trajectory inference

We reconstructed single-cell trajectories to generate pseudotime or pseudo-space for each cell using Monocle 2 (Qiu et al. 2017).

Cell-cell communication inference

We employed CellChat (Jin et al. 2021) to perform cell-cell communication inference, analysis, and visualization.

Differential expression and functional enrichment analysis

We used the Wilcoxon Rank Sum test to identify differentially expressed genes among cell populations or spatial domains implemented by the Seurat package (Hao et al. 2021) with a false discovery rate (FDR) less than 0.05. The functional enrichment of these genes was performed by clusterProfiler (Wu et al. 2021).

References

Agrawal A, Balcı H, Hanspers K, Coort SL, Martens M, Slenter DN, Ehrhart F, Digles D, Waagmeester A, Wassink I et al. 2024. WikiPathways 2024: next generation pathway database. *Nucleic Acids Res* **52**: D679-d689.

Dong K, Zhang S. 2022. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* **13**: 1739.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-3587 e3529.

Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. 2021. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**: 1088.

Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. 2023. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**: D587-d592.

Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, Haw R, Jassal B, Matthews L, May B et al. 2024. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* **52**: D672-d678.

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. 2017. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**: 979-982.

Wilkerson MD, Hayes DN. 2010. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**: 1572-1573.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**: 100141.