Supplemental Notes

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

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

Note 1. Validation of the rationality of the supervised evaluation strategy

To validate that our classification-based evaluation strategy is appropriate, we performed classification using the principal components (PCs) derived from the standard Seurat v4 workflow (Hao et al. 2021), which were originally used for clustering, as input features. The resulting models exhibited high classification performance (Supplemental Fig. S2), supporting the use of classification accuracy as a meaningful proxy for evaluating the effectiveness of feature extraction. It also suggests that more refined feature extraction, namely more precisely capturing biologically relevant signals potentially reflective of gene program activity, may contribute to improved cell identity identification.

On the other hand, we performed principal component analysis (PCA) on the gene expression profiles from the simulated gene programs across three simulation tests. We used the resulting principal components (PCs) as features to perform classification, similar to the evaluation process of DeCEP and existing methods. Based on the elbow plots, we determined that three was the optimal number of PCs for all three simulation tests (Supplemental Fig. S3A). Across all three tests, the first three PCs consistently demonstrated high performance in distinguishing cell identity (Supplemental Fig. S3B and S3C). These findings suggest that the simulated gene programs successfully introduced sufficient differential gene expression signals among cells within a specific context.

Note 2. Evaluation of the various components within the DeCEP framework on simulated data

Breaking down the various components of the DeCEP framework for performance evaluation helps in thoroughly assessing the contribution of each component to the final, context-specific gene program characterization at the cellular level. We conducted additional analyses to further strengthen this evaluation. In the three simulated datasets, we systematically remove components of the DeCEP framework to evaluate their respective contributions to the performance metrics. We focused on two key factors: (1) whether functional network construction was employed to derive gene weights based on network topology, and (2) whether an imputation step was incorporated to enable more precise microscopic dissection. In the condition without functional network construction, we treated all genes within a specific simulated gene program equally. This resulted in four groups: ‘w/ weight, w/ imputation,’ ‘w/o weight, w/ imputation,’ ‘w/ weight, w/o imputation,’ and ‘w/o weight, w/o imputation.’ The results showed that most of the evaluation metrics yielded the highest scores in the ‘w/ weight, w/ imputation’ group (Supplemental Fig. S6). This reflects the combined effects of imputation and gene weighting on the characterization of context-specific gene program activity at the cellular level. In particular, the weighted quantification of context-specific gene programs, using weights derived from functional network construction, generally performed better than unweighted quantification under both imputation and non-imputation conditions. This finding also indirectly suggests the potential accuracy and effectiveness of our function network construction and the associated gene weights.

Note 3. Initial characterization of the normal liver tissue scRNA-seq dataset used in this study

We focused on hepatocytes exhibiting high expression of the hepatocyte marker *Alb* and clustered them into four clusters, labeled clusters 0 to 3 (Supplemental Fig. S8A and S8B). Next, we performed differential expression analysis across these clusters, revealing significant variability in the expression of marker genes associated with zonation of the hepatic lobule. For example, *Oat* and *Cyp2e1,* as marker genes of the pericentral region, presented the highest expression levels in cluster 3 and the lowest expression levels in cluster 2. *Cyp2f2*, as a marker gene of the periportal region, exhibited expression levels following the order: cluster 2 > clusters 0 and 1 > cluster 3. Hence, we rearranged the four clusters along the hepatic lobule axis based on these expression patterns, where cluster 2 is in the periportal region, clusters 0 and 1 are in the mid-lobule region, and cluster 3 is in the pericentral region (Supplemental Fig. S8A and S8C).

Note 4. Evaluation of the various components within the DeCEP framework using the normal liver tissue scRNA-seq dataset

We further assessed the contribution of each component within the DeCEP framework to the final characterization of context-specific gene programs at the cellular level using the normal liver tissue dataset and the strategy described in Supplemental Note 2. Consistently, most evaluation metrics indicated that the ‘w/ weight, w/ imputation’ group exhibited improved performance relative to the others (Supplemental Fig. S11).

Note 5. Comparison between DeCEP and CoGAPS

We applied CoGAPS using the parameter settings outlined in the vignette (https://www.bioconductor.org/packages/release/bioc/vignettes/CoGAPS/inst/doc/CoGAPS.html) (Fertig et al. 2010). The number of patterns was set to 8 based on the empirical settings outlined in the single-cell analysis section of the vignette, meaning the expression matrix was factorized into 8 factors. As shown in Supplemental Fig. S13A, the activity levels of these 8 patterns at the cellular level are directly visualized in the UMAP plots. Since it was unclear which pattern specifically characterized liver detoxification, we used the *getPatternGeneSet* function to perform gene set enrichment analysis of the significant genes across all eight patterns. We focused on four gene sets, including drug metabolism - cytochrome P450, metabolism of xenobiotics by cytochrome, glutathione metabolism, and glutamate and glutamine metabolism, all of which were consistent with those used in this study. We found that three of the four gene sets were enriched in Pattern 7, suggesting that it was associated with liver detoxification (Supplemental Fig. S13B). However, in contrast to the DeCEP scores for liver detoxification-related gene programs, the activity levels of Pattern 7 did not exhibit a significant gradient distribution along the hepatic lobule axis (Supplemental Fig. S13A). Further annotation of Pattern 7, using the hallmark gene sets referenced in the vignette, revealed that the most significantly enriched gene set was HALLMARK_XENOBIOTIC_METABOLISM (Supplemental Fig. S13C). This finding supports our earlier conclusion that CoGAPS identified gene programs related to liver detoxification in Pattern 7. However, the pattern also exhibited significant enrichment for a variety of other functions. This complexity complicates the direct characterization of liver detoxification and likely explains the absence of a gradient distribution in Pattern 7 (Supplemental Fig. S13C).

Moreover, compared to traditional gene set scoring methods, matrix factorization-based methods like CoGAPS offer the advantage of characterizing gene programs at the gene level. We used the *patternMarkers* function to identify the marker genes for Pattern 7 and their corresponding PatternScores. As outlined in the CoGPAS's vignette, lower PatternScores indicate a stronger association of the marker genes with Pattern 7. Among the top 20 marker genes associated with Pattern 7 identified by CoGAPS, we did not find genes significantly linked to liver detoxification (Supplemental Fig. S13D). However, when we compared the context-dependent hub genes linked to the four detoxification-related gene sets identified by our method with the marker genes of Pattern 7, we found that the hub genes from three of the four programs exhibited varying degrees of overlap with the marker genes of Pattern 7 (Supplemental Fig. S13E). The exception was the glutamate and glutamine metabolism program, which aligned with its lack of enrichment in Pattern 7. In other words, the marker genes for Pattern 7 identified by CoGAPS include detoxification-related genes, but their significance is partially diminished by complex signals, causing them to appear in less prominent positions. To further investigate this, we examined the ranking of these genes among all marker genes for Pattern 7 and found that most of them were ranked well beyond the top 20 (Supplemental Fig. S13E), which supports the inference that detoxification signals are de-emphasized in this pattern. Additionally, compared to our method, CoGAPS failed to identify *Cyp1a2*, a key gene encoding a detoxification enzyme (Thorn et al. 2012). This suggests that the patterns identified by CoGAPS, which involve the entanglement of multiple functional gene programs, were less effective at pinpointing genes specifically related to individual gene programs.

Overall, these comparisons further highlight that our function-centric approach allows for the direct and independent characterization of distinct functional gene programs at both the gene and cellular levels, effectively disentangling them and enhancing interpretability.

Note 6. Initial characterization of the Alzheimer's disease snRNA-seq dataset used in this study

By applying DeCEP to a single-nucleus RNA sequencing (snRNA-seq) dataset from the hippocampus of AD and wild-type (WT) mice, we focused on astrocytes that exhibited high expression levels of the astrocytic marker *Slc1a3* (Habib et al. 2020) (Supplemental Fig. S16A). We distinguished between two conditions of astrocytes based on *Gfap* expression levels, which represented high and low levels of NI, termed NI-high and NI-low, respectively (Supplemental Fig. S16A and S16B). The NI-high condition mainly existed in AD mice, whereas the NI-low condition was present in both

AD and WT mice (Supplemental Fig. S16C). We observed that the gene expression levels of the cells under the two conditions were highly heterogeneous. For example, cells in the NI-high condition exhibited relatively high expression levels of *Vim*, *Gfap*, *Serpina3n*, *Apoe*, and *Ctsb* (Supplemental Fig. S16D), all of which are related to NI (Pekny et al. 2016; Parhizkar and Holtzman 2022; Wu et al. 2023; Han et al. 2024).

Note 7. DeCEP identifies the spatially associated neuroinflammatory phenotype in mouse brain tissue sections

Using the Allen brain atlas (Wang et al. 2020) as a reference, we assigned the spatial domains derived from spatial clustering in the two tissue sections to six regions: the cerebral cortex (CTX), cerebral nuclei (CNU), hippocampal formation (HPF), thalamus (TH), hypothalamus (HY), and fiber tracts (Supplemental Fig. S19A-S19C). We used chemokine signaling as a representative gene program to characterize the degree of NI in each spot, showing that, compared with the WT tissue section, the AD tissue section exhibited an increased number of spots with high DeCEP states and a decreased number of spots with low DeCEP states (Supplemental Fig. S20A). The distribution of the DeCEP states in these spots showed a significant spatial correlation, similar to the expression of *Gfap*, a marker of astrocyte reactivity (Lawrence et al. 2023) (Supplemental Fig. S20A and S20B). We quantified the proportions of different DeCEP states in the six regions, showing that in the AD tissue section, the highest proportion of spots exhibiting high DeCEP states was observed in the fiber tracts, followed by the HPF, TH, and CTX. In contrast, for the WT tissue section, the proportion of spots with high DeCEP states in these regions decreased compared with those in the AD section

(Supplemental Fig. S20C).

Based on our data, it appears that the HPF could be the initial brain area affected by inflammation during the onset of AD. Through an in-depth analysis of gene expression variances between spots with high and low DeCEP states within the HPF, we found that inflammatory-related genes, such as the typical marker gene *Gfap*, were prominently expressed in spots with high DeCEP states, whereas neural-related genes, such as the typical marker gene *Hpca*, were notably expressed in spots with low DeCEP states (Supplemental Fig. S20D-S20F). The spatial correlation of the expression pattern suggests that inflammatory glial cells may progressively move toward and damage neurons in the HPF during the onset of AD.

Additionally, we further employed DeCEP on a high-resolution Slide-seqV2 dataset of the AD mouse hippocampus (Cable et al. 2022). The clustering results showed that clusters 4, 5, and 8, enriched in granule and pyramidal cells, were surrounded by clusters 3, 6, and 2, respectively (Supplemental Fig. S21A). We counted the proportion of spots with a specific DeCEP state in different clusters, showing that the proportion of mixed DeCEP states in clusters 3, 6, and 2 was relatively high (Supplemental Fig. S21B), which suggests that AD, often triggered by chronic inflammation, is characterized by persistent and weak interactions between glial cells and neurons. To investigate whether certain factors may influence the mixed DeCEP states in space, we assigned the region containing these spots as the ROI. Using the "spatial" mode of DeCEP, we identified the ROI and its corresponding neighborhood as the specific spatial context associated with the chronic inflammation phenotype. The identification of a series of spatially dependent hub genes within this spatial context suggests potential regulation of the ROI by the neighborhood (Supplemental Fig. S21C). *Vav2*, in particular, has the highest gene weight in this spatial context (Supplemental Fig. S21C), implying a possible interaction between amyloid-β (Aβ) and glial cells under the regulation of *Vav2*. Actually, a previous study showed that Vav2 can interact with amyloid precursor protein (APP) and positively regulate APP's protein level (Zhang et al. 2022), thus confirming the potential of this inference.

Note 8. Initial characterization of the cSCC scRNA-seq dataset used in this study

These tumor keratinocytes exhibited high expression of *KRT5*, a marker associated with basal tumors (Supplemental Fig. S22A). They were further clustered into 5 clusters, labeled clusters 0 to 4 (Supplemental Fig. S22B). The high expression of *KRT1*, a marker of terminal differentiation, in cluster 0 indicates that the cells in this cluster are mainly differentiated. In contrast, the loss of *KRT1* expression in cluster 1 suggests that these cells are undergoing dedifferentiation. The presence of *MMP10* expression in cluster 4 indicates that the cells here are undergoing epithelial-mesenchymal transition (EMT) (Garg 2022). Clusters 2 and 3 represent proliferative cell populations associated with dedifferentiation and tumor progression, respectively, as indicated by their high expression levels of *MKI67* and *TOP2A* (Supplemental Fig. S22C).

Note 9. Comparison of clusters 2 and 5 as well as clusters 3, 8, and 9 in the ST data of the human cSCC tissue section

We investigated the ten TME-related genes in Supplemental Fig. S26A, and as shown in Supplemental Table S1, the overall differential expression significance of these ten genes was significantly higher in clusters 2 and 5 compared to clusters 3, 8, and 9. Further differential expression analysis comparing clusters 2 and 5 with clusters 3, 8, and 9 confirmed that nine of the ten genes were expressed at notably higher levels in clusters 2 and 5 (as detailed in the table below). These results indicate that clusters 2 and 5 exhibited a more active TME compared to clusters 3, 8, and 9, which aligns with the finding that the spots with high DeCEP states of TGFB and Wnt signaling were mainly enriched in clusters 2 and 5. While clusters 3, 8, and 9 displayed some degree of TME activity, they had not yet reached a highly active state, resulting in fewer spots with high DeCEP states within these clusters.

Clusters 2 and 5 vs. Clusters 3, 8, and 9

gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
ACTB	4.67E-14	0.496642953	1	1	1.56E-09
TNC	5.60E-19	1.094225125	0.985	0.711	1.87E-14
HLA-B	3.26E-08	0.346578144	1	1	0.001090115
HLA-A	1.25E-08	0.323408529	1	0.993	0.000418573
TMSB4X	3.70E-15	0.449690944	1	0.987	1.24E-10
TMSB10	1.09E-13	0.501921664	1	0.993	3.65E-09
MMP1	2.73E-13	1.143078161	0.956	0.698	9.11E-09
HLA-A	1.25E-08	0.323408529	1	0.993	0.000418573
SAT1	1.14E-08	0.460718699	0.985	0.899	0.00038152

Note 10. DeCEP enhances the characterization of spatially dependent gene programs

DeCEP provides deeper biological insights than existing methods in ST data analysis. Specifically, when applying DeCEP to adult mouse liver ST data, DeCEP successfully reconstructed the nonuniform distribution of detoxification-related gene program activity states along the liver lobule axis, which existing methods were unable to achieve. This finding further validates the accuracy of anchoring DeCEP states derived from scRNA-seq reference data to ST data in the DeCEP framework. Additionally, in our application to the human cSCC tissue section, DeCEP was employed to identify the ROIs associated with tumor invasion and their neighborhoods, highlighting its capability to uncover biological insights within spatial contexts. We also quantified the activity of cancer-related gene programs in these ROIs and their neighborhoods, demonstrating the variability of DeCEP scores across spatial gradients. These insights are difficult to obtain with existing methods.

Note 11. Further discussion on the limitations and prospects of this study

Characterizing gene programs from imaging-based ST data presents a significant challenge, as current technologies can measure only a few hundred genes (Park et al. 2023). In future studies, we plan to explore the reference-free identification of spatially dependent gene programs in sequencing-based ST data and establish a basis for characterizing gene programs of imaging-based ST data through imputation and gridding approaches. The current imputation methods can predict the expression of undetected genes in imaging-based ST data (Wan et al. 2023; Li et al. 2024; Qiao and Huang 2024), suggesting that incorporating such methods to characterize gene programs in this type of data is potentially feasible. These efforts will enhance the applicability and scalability of the DeCEP framework, enabling new biological insights into the characterization of gene programs in spatial organization.

On the other hand, although DeCEP provides a direct and independent framework for characterizing context-specific gene programs, its requirement for a priori functional gene lists introduces an inherent constraint, as the analysis is restricted to the genes within the selected gene lists. This approach implicitly assumes that all genes associated with a given gene program are already included within the corresponding gene list. While curated resources such as GO (Consortium 2019), KEGG (Kanehisa et al. 2023), and MSigDB (Liberzon et al. 2015) offer comprehensive and high-quality prior knowledge, it remains possible that currently unannotated genes play critical roles in specific biological processes. Consequently, DeCEP may limit the discovery of novel gene associations beyond the scope of predefined annotations. This limitation reflects a broader, ongoing challenge in the field of gene program characterization. In the future, we plan to use unsupervised strategies that take known genes as seed genes. By applying clustering and propagation approaches, we would uncover previously unrecognized genes potentially associated with specific gene programs, enabling further investigation.

Note 12. Hierarchical design of parent-child node relationships within spatial contexts of the DeCEP framework

In the “spatial” mode of the DeCEP framework, we designed constrained parent-child node relationships tailored to spatial contexts, where the neighborhoods are treated as parent nodes and the regions of interest (ROIs) as child nodes. Here, we mainly consider that the ROIs are the regions of true interest to users, and this hierarchical design is intended to capture the context-specific gene programs at the gene level that affect the ROIs. In other words, our design aims to uncover the molecular factors that are potentially associated with the functional phenotypes of the ROIs. If the parent-child node relationships were reversed, it would suggest that the context-specific gene programs captured at the gene level are related to the functional phenotypes of the neighborhoods. While this approach is also feasible, it deviates from our original intention in defining the ROIs.

Note 13. The supervised evaluation strategy and fairness principles

In the simulated data section, simulation tests 1-3 were each conducted to represent a single cell type. DeCEP treated each simulated cell type as a specific cellular context. Accordingly, DeCEP processed the expression matrix from each simulation as a unified dataset, aligning with the analysis approach adopted by the five existing methods. The classification labels were employed solely as ground truth references to assess the ability of each method to identify similarities and differences among cells within a specific context.

References

Cable DM, Murray E, Shanmugam V, Zhang S, Zou LS, Diao M, Chen H, Macosko EZ, Irizarry RA, Chen F. 2022. Cell type-specific inference of differential expression in spatial transcriptomics. *Nat Methods* **19**: 1076-1087.

Consortium GO. 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research* **47**: D330-D338.

Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF. 2010. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* **26**: 2792-2793.

Garg M. 2022. Emerging roles of epithelial-mesenchymal plasticity in invasion-metastasis cascade and therapy resistance. *Cancer Metastasis Rev* **41**: 131-145.

Habib N, McCabe C, Medina S, Varshavsky M, Kitsberg D, Dvir-Szternfeld R, Green G, Dionne D, Nguyen L, Marshall JL et al. 2020. Disease-associated astrocytes in Alzheimer's disease and aging. *Nat Neurosci* **23**: 701-706.

Han X, Lei Q, Liu H, Zhang T, Gou X. 2024. SerpinA3N Regulates the Secretory Phenotype of Mouse Senescent Astrocytes Contributing to Neurodegeneration. *J Gerontol A Biol Sci Med Sci* **79**.

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.

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.

Lawrence JM, Schardien K, Wigdahl B, Nonnemacher MR. 2023. Roles of neuropathology-associated reactive astrocytes: a systematic review. *Acta Neuropathol Commun* **11**: 42.

Li K, Li J, Tao Y, Wang F. 2024. stDiff: a diffusion model for imputing spatial transcriptomics through single-cell transcriptomics. *Brief Bioinform* **25**.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417-425.

Parhizkar S, Holtzman DM. 2022. APOE mediated neuroinflammation and neurodegeneration in Alzheimer's disease. *Semin Immunol* **59**: 101594.

Park HE, Jo SH, Lee RH, Macks CP, Ku T, Park J, Lee CW, Hur JK, Sohn CH. 2023. Spatial transcriptomics: technical aspects of recent developments and their applications in neuroscience and cancer research. *Advanced Science* **10**: 2206939.

Pekny M, Pekna M, Messing A, Steinhauser C, Lee JM, Parpura V, Hol EM, Sofroniew MV, Verkhratsky A. 2016. Astrocytes: a central element in neurological diseases. *Acta Neuropathol* **131**: 323-345.

Qiao C, Huang Y. 2024. Reliable imputation of spatial transcriptomes with uncertainty estimation and spatial regularization. *Patterns (N Y)* **5**: 101021.

Thorn CF, Aklillu E, Klein TE, Altman RB. 2012. PharmGKB summary: very important pharmacogene information for CYP1A2. *Pharmacogenet Genomics* **22**: 73-77.

Wan X, Xiao J, Tam SST, Cai M, Sugimura R, Wang Y, Wan X, Lin Z, Wu AR, Yang C. 2023. Integrating spatial and single-cell transcriptomics data using deep generative models with SpatialScope. *Nat Commun* **14**: 7848.

Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, Graddis N, Naeemi M, Facer B, Ho A et al. 2020. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181**: 936-953 e920.

Wu Y, Mumford P, Noy S, Cleverley K, Mrzyglod A, Luo D, van Dalen F, Verdoes M, Fisher EMC, Wiseman

FK. 2023. Cathepsin B abundance, activity and microglial localisation in Alzheimer's disease-Down syndrome and early onset Alzheimer's disease; the role of elevated cystatin B. *Acta Neuropathol Commun* **11**: 132.

Zhang Y, Yang X, Liu Y, Ge L, Wang J, Sun X, Wu B, Wang J. 2022. Vav2 is a novel APP-interacting protein that regulates APP protein level. *Scientific Reports* **12**: 12752.