MIRAGE: Manifold-Informed Gene-module Extraction for Disentangling Simultaneous Dynamics in Single-Cell RNA-seq

Zhaoheng Li
Department of Biostatistics
University of Washington
zli1@uw.edu

Yifan Lin
Department of Biostatistics
University of Washington
yifanlin@uw.edu

Kevin Z. Lin
Department of Biostatistics
University of Washington
kzlin@uw.edu

Single-cell RNA-seq (scRNA-seq) has transformed our ability to dissect cellular dynamics, yet deciphering the intertwined biology of complex tissues – such as diseased brain or inflamed stroma – remains challenging where multiple biological processes such as cell cycle, immune activation, and intercellular signaling often operate simultaneously. For example, in brains with Alzheimer's disease, microglia, and astrocytes span homeostatic to neurotoxic states while co-pathologies like vascular dysfunction and protein aggregation further modulate such transitions (3; 8). High dimensionality, sparsity, and noise compound the problem. Without robust computational methods to disentangle these concurrent processes, attempts to construct development, regeneration, or immune dynamics may risk yielding incomplete or misleading insights, slowing progress to translate scRNA-seq discoveries for clinical therapies.

Recent methods, such as Palantir (14), DELVE (12), and GeneTrajectory (11), identify gene modules based on correlating gene expression with a cell-cell graph structure. However, all three rely on a single graph built from highly variable genes (HVGs). Mixing all the HVGs together can mask secondary signals or erroneously merge distinct trajectories in complex or dysregulated systems (13). Correlation-based tools such as Canonical Correlation Analysis (CCA) and WGCNA (6) cluster genes into modules, but correlations alone overlook cell-type heterogeneity. Consequently, existing approaches can partition genes, yet none fully disentangles modules arising from simultaneously operating distinct dynamic processes.

To address the above challenges, we propose **MIRAGE** (Manifold-Informed Gene-module Extraction), a statistical test that determines if two gene sets share similar dynamical processes based on the cellular manifold they induce (Fig. 1A). MIRAGE has three stages: (1) forming the initial gene sets, (2) constructing the manifolds defined by each gene set, and (3) testing for pairwise relations between any two manifolds. Since neither the true number of biological processes nor their gene composition is known beforehand, we perform spectral clustering on the gene-correlation matrix estimated by SAVER (5) on the HVGs to obtain K fine-resolution gene modules $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(K)}$. We over-cluster with an initialization of $K = 20$ to ensure that functionally distinct programs are initially separated. For any gene set $\mathcal{G}^{(k)}$, taking inspiration from previous manifold-based work to model single-cell data (4; 9; 14), we can construct a cellular manifold (i.e., k-nearest-neighbor (KNN) graph $A^{(k)}$) using only the genes in $\mathcal{G}^{(k)}$. If two gene modules are functionally coordinated, then their associated cellular manifolds (i.e., the cell-cell nearest-neighbor graphs by those genes) should be structurally similar. Intuitively, for any gene that does not belong to either of these two gene sets, the extent of smoothness in variation it demonstrates on the two induced manifolds should be similar. Hence, to compare two gene programs $\mathcal{G}^{(k_1)}$ and $\mathcal{G}^{(k_2)}$, we define the so-called *bystander genes* to be the set $\mathcal{B}^{(k_1,k_2)} = \cup_{k' \in [K] \setminus \{k_1,k_2\}} \mathcal{G}^{(k')}$ – that is, these are all the genes not in manifolds k_1 and k_2 themselves. The usage of bystander genes takes inspiration from works that test for shared graph structure (18; 2). We compute the Laplacian score (12) of these bystander genes to quantify how smoothly its expression varies over a manifold $A^{(k)}$. For each $A^{(k)}$, we form the unnormalized Laplacian $L^{(k)} = D^{(k)} - A^{(k)}$, where D^k is the diagonal degree matrix, and compute the Laplacian score as

$$L_g^{(k)} = \frac{\widetilde{f}_g^\top L^{(k)} \widetilde{f}_g}{\widetilde{f}_g^\top D^{(k)} \widetilde{f}_g} \in [0,1] \quad \text{where} \quad \widetilde{f}_g = f_g - \left(\frac{f_g^\top D^{(k)} \mathbf{1}}{\mathbf{1}^T D^{(k)} \mathbf{1}}\right) \cdot \mathbf{1} \in \mathbb{R}^n.$$

Here, f_g is the centered expression vector of gene g across n cells. The smaller $L_g^{\mathcal{G}}$ is, the smoother the bystander gene g varies on the manifold induced by $\mathcal{G}$. To compare two gene programs $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$, we apply the Kolmogorov–Smirnov test to assess differences between the distributions of $L^{(k_1)}_{\mathcal{B}^{(k_1,k_2)}}$ and $L^{(k_2)}_{\mathcal{B}^{(k_1,k_2)}}$, and merge $\mathcal{G}^{(k_1)}$ and $\mathcal{G}^{(k_2)}$ if the null hypothesis fails to be rejected. Altogether, MIRAGE uncovers and disentangles multiple biological processes by bypassing the need to rely solely on gene-wise correlations or a single lower-dimensional manifold defined by the agglomeration of these genes. Currently, we restrict MIRAGE to operate only on HVGs to prevent noisy genes

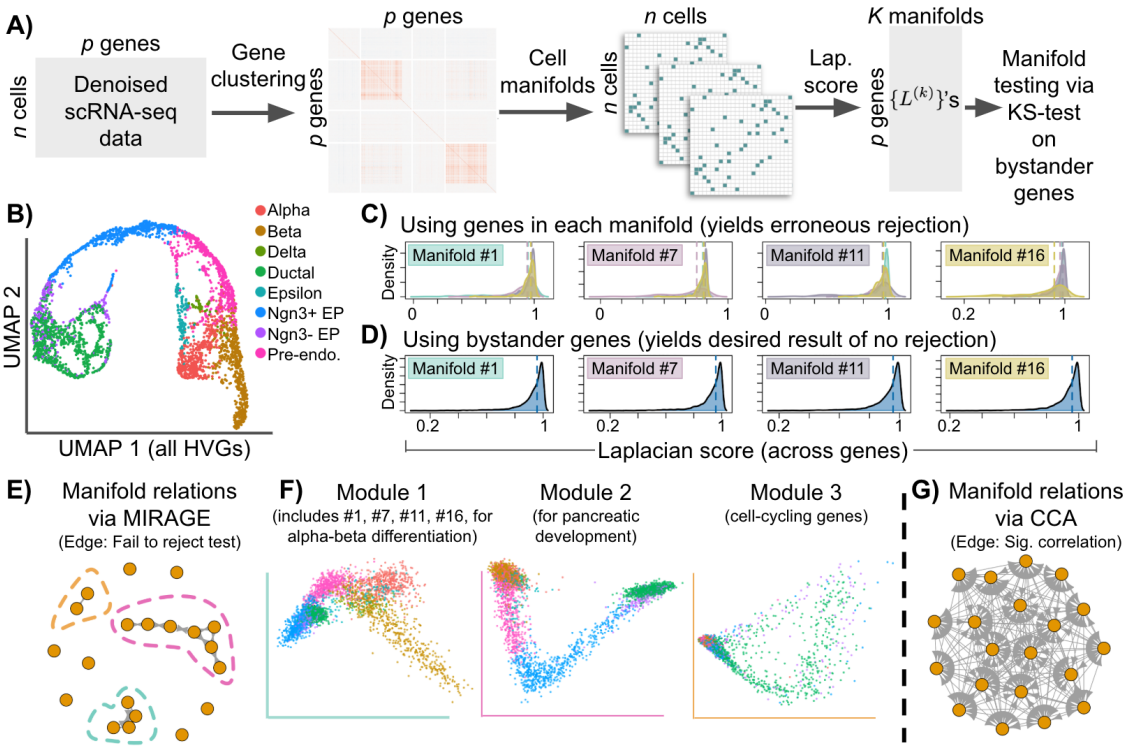

Figure 1: **(A)** Workflow of MIRAGE. **(B)** UMAP of pancreatic cells from (1). **(C)** Distribution of Laplacian scores of gene clusters that form Module 1 in B on their corresponding manifolds. The dotted lines denote the average Laplacian score of each gene set. **(D)** Distribution of bystander genes on the manifolds in C. **(E)** Relations estimated manifold relations among all 20 manifolds, where each node is a manifold (i.e., an initial cluster of genes). The KS test on bystander genes identified 3 merged gene modules. The green module includes the 4 manifolds shown in C and D. **(F)** UMAPs created using the 3 main merged modules. **(G)** Comparison, using testing the significance of canonical correlation to determine if gene clusters share similar dynamics. Plotted in a similar style as in E.

from negatively affecting the workflow, With our results, we apply gene ontology analysis to identify the functional annotations of the aggregated set of genes within each set of merged manifolds.

We applied MIRAGE to a pancreas scRNA-seq dataset (1) to illustrate its effectiveness in isolating distinct dynamical processes. From the top 3000 HVGs we formed 20 spectral clusters that merged into three main gene modules (Fig. 1E). We note that a gene set almost always tends to demonstrate the smoothest variation across the very manifold it defines (Fig. 1B) – this yields an incorrect rejection among these four manifolds (i.e., incorrectly assessing these gene sets are unrelated). In contrast, using the bystander gene framework provides the intended Type-I error control (Fig. 1C). Module 1 highlights the distinction between β and α cells, while Module 2 highlights the development of pre-endocrine cells to terminal pancreatic cells (Fig. 1F). In contrast, Module 3 illustrates the ongoing cell cycle among ductal cells, similar to what was found in previous work (20). To compare our graph-based hypothesis testing approach to the correlation-based approach, we tested the significance of canonical correlations among the 20 initial gene clusters. That is, two gene clusters will be merged if they have a significant canonical correlation. However, this approach produced a single connected component that merges all genes into a single module (Fig. 1G) and failed to separate the cell cycle from cell differentiation. This comparison between MIRAGE and CCA demonstrates that our approach could disentangle simultaneous cellular dynamics that correlation-based approaches may overlook.

MIRAGE demonstrates the potential to disentangle multiple coexisting biological programs through a hypothesis testing approach using the cellular manifolds without relying on a pre-computed pseudotime that might be affected by the selection of HVGs. Next, we will apply the method on a metastatic pancreatic cancer dataset with lineage barcoding (16) to explore the synergy between lineage identity and metastasis states, and a microglia dataset to understand the dynamics during the progression of Alzheimer's disease (10). As downstream analysis, we will use each finalized gene module to construct a cellular trajectory using TFVelo (7), a method that infers RNA velocity without requiring access to spliced and unspliced counts, where we can use methods such as TradeSeq (19) to identify the biological processes being activated along each manifold.

References

[1] BASTIDAS-PONCE, A., TRITSCHLER, S., DONY, L., SCHEIBNER, K., TARQUIS-MEDINA, M., SALINNO, C., SCHIRGE, S., BURTSCHER, I., BÖTTCHER, A., THEIS, F. J., LICKERT, H., AND BAKHTI, M. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development 146*, 12 (June 2019).

[2] BUNEA, F., GIRAUD, C., LUO, X., ROYER, M., AND VERZELEN, N. Model assisted variable clustering: Minimax-optimal recovery and algorithms. *Annals of Statistics 48*, 1 (2020), 111.

[3] GUVENEK, A., PARIKSHAK, N., ZAMOLODCHIKOV, D., GELFMAN, S., MOSCATI, A., DOBBYN, L., STAHL, E., SHULDINER, A., AND COPPOLA, G. Transcriptional profiling in microglia across physiological and pathological states identifies a transcriptional module associated with neurodegeneration. *Communications Biology 7*, 1 (Sept. 2024).

[4] HAGHVERDI, L., BUETTNER, F., AND THEIS, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics 31*, 18 (2015), 2989–2998.

[5] HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M., AND ZHANG, N. R. SAVER: Gene expression recovery for single-cell RNA sequencing. *Nature Methods 15*, 7 (2018), 539–542.

[6] LANGFELDER, P., AND HORVATH, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics 9*, 1 (Dec. 2008).

[7] LI, J., PAN, X., YUAN, Y., AND SHEN, H.-B. TFvelo: Gene regulation inspired RNA velocity estimation. *Nature Communications 15*, 1 (Feb. 2024).

[8] MALLACH, A., ZIELONKA, M., VAN LIESHOUT, V., AN, Y., KHOO, J. H., VANHEUSDEN, M., CHEN, W.-T., MOECHARS, D., ARANCIBIA-CARCAMO, I. L., FIERS, M., AND DE STROOPER, B. Microglia-astrocyte crosstalk in the amyloid plaque niche of an Alzheimer's disease mouse model, as revealed by spatial transcriptomics. *Cell Reports 43*, 6 (June 2024), 114216.

[9] MOON, K. R., VAN DIJK, D., WANG, Z., GIGANTE, S., BURKHARDT, D. B., CHEN, W. S., YIM, K., ELZEN, A. V. D., HIRN, M. J., COIFMAN, R. R., ET AL. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology 37*, 12 (2019), 1482–1492.

[10] PRATER, K. E., GREEN, K. J., MAMDE, S., SUN, W., COCHOIT, A., SMITH, C. L., CHIOU, K. L., HEATH, L., ROSE, S. E., WILEY, J., ET AL. Human microglia show unique transcriptional changes in Alzheimer's disease. *Nature Aging 3*, 7 (2023), 894–907.

[11] QU, R., CHENG, X., SEFIK, E., STANLEY III, J. S., LANDA, B., STRINO, F., PLATT, S., GARRITANO, J., ODELL, I. D., COIFMAN, R., FLAVELL, R. A., MYUNG, P., AND KLUGER, Y. Gene trajectory inference for single-cell data by optimal transport metrics. *Nature Biotechnology 43*, 2 (Apr. 2024), 258–268.

[12] RANEK, J. S., STALLAERT, W., MILNER, J. J., REDICK, M., WOLFF, S. C., BELTRAN, A. S., STANLEY, N., AND PURVIS, J. E. DELVE: Feature selection for preserving biological trajectories in single-cell data. *Nature Communications 15*, 1 (Mar. 2024).

[13] SAUNDERS, L. M., SRIVATSAN, S. R., DURAN, M., DORRITY, M. W., EWING, B., LINBO, T. H., SHENDURE, J., RAIBLE, D. W., MOENS, C. B., KIMELMAN, D., AND TRAPNELL, C. Embryo-scale reverse genetics at single-cell resolution. *Nature 623*, 7988 (Nov. 2023), 782–791.

[14] SETTY, M., KISELIOVAS, V., LEVINE, J., GAYOSO, A., MAZUTIS, L., AND PE'ER, D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology 37*, 4 (Mar. 2019), 451–460.

[15] SETTY, M., KISELIOVAS, V., LEVINE, J., GAYOSO, A., MAZUTIS, L., AND PE'ER, D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology 37*, 4 (2019), 451–460.

[16] SIMEONOV, K. P., BYRNS, C. N., CLARK, M. L., NORGARD, R. J., MARTIN, B., STANGER, B. Z., SHENDURE, J., MCKENNA, A., AND LENGNER, C. J. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell 39*, 8 (Aug. 2021), 1150–1162.e9.

[17] SU, C., XU, Z., SHAN, X., CAI, B., ZHAO, H., AND ZHANG, J. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nature Communications 14*, 1 (2023), 4846.

[18] SU, Y., WONG, R. K., AND LEE, T. C. Network estimation via graphon with node features. *IEEE Transactions on Network Science and Engineering 7*, 3 (2020), 2078–2089.

[19] VAN DEN BERGE, K., ROUX DE BÉZIEUX, H., STREET, K., SAELENS, W., CANNOODT, R., SAEYS, Y., DUDOIT, S., AND CLEMENT, L. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications 11*, 1 (2020), 1201.

[20] ZHENG, S. C., STEIN-O'BRIEN, G., AUGUSTIN, J. J., SLOSBERG, J., CAROSSO, G. A., WINER, B., SHIN, G., BJORNSSON, H. T., GOFF, L. A., AND HANSEN, K. D. Universal prediction of cell-cycle position using transfer learning. *Genome Biology 23*, 1 (Jan. 2022).