Graph reinforcement and smoothing for improved spatial gene expression prediction

Eric D. Sun [1] **Rong Ma** [2] **James Zou** [1]

Abstract

Single-cell spatial transcriptomics technologies are often limited to a small number of measured genes. Several computational methods for the prediction of additional spatial gene expression profiles have been developed to address this limitation. However, relational information encoded by the co-expression patterns of genes and the spatial organization of cells are rarely utilized in these prediction methods. Here we introduce SPRITE (Spatial Propagation and Reinforcement of Imputed Transcript Expression) as a flexible post-processing algorithm to improve spatial gene expression predictions from any existing method by propagating errors along gene co-expression networks and smoothing predicted gene expression across spatial graphs of cells. SPRITE generally improves the quality of predicted gene expression profiles and downstream clustering and visualization of the data across several benchmark spatial transcriptomics datasets.

1. Introduction

Spatial transcriptomics technologies measure gene expression in relation to spatial organization, enabling the characterization of cell types and transcripts across various tissues and organisms (Moses & Pachter, 2022). However, most spatial transcriptomics methods with single-cell resolution can only measure a limited number of genes (Li et al., 2022). Due to the resource-intensive nature of acquiring single-cell spatial transcriptomics data, there is a need for computational methods to expand the number of genes profiled or predict the expression of additional genes of interest.

Several computational methods address the imputation or prediction of spatial gene expression using paired spatial transcriptomics and single-cell RNA-seq datasets. These approaches typically involve joint embedding of the spatial and RNA-seq data, followed by the prediction of expression for new spatial genes by aggregating neighboring cells in the RNA-seq data (Abdelaal et al., 2020; Allen et al., 2023; Shengquan et al., 2021; Welch et al., 2019), or by optimal transport (Biancalani et al., 2021). In most cases, relational information between genes (e.g. co-expression) and between cells (e.g. spatial proximity) are not utilized in the prediction method. Given that accurate prediction of spatial gene expression is necessary for downstream analysis, leveraging this relational information to improve the quality of predictions is desirable.

Here we introduce SPRITE, a simple wrapper method around existing spatial gene expression prediction methods that uses gene correlation networks and spatial neighborhood graphs to refine the baseline predictions. SPRITE is flexible and can be used with any method for predicting spatial gene expression. We show that post-processing of spatial gene expression predictions with SPRITE generally leads to more accurate predictions and that these improvements translate to common downstream analysis pipelines for spatial transcriptomics.

2. Methods

2.1. SPRITE workflow

Spatial gene expression prediction generally involves paired data from spatial transcriptomics and RNA-seq (paired in that they are approximately from the same tissue and organism, i.e. from the same distribution). We denote the spatial transcriptomic data as $X_{\text{spatial}} \in \mathbb{R}^{n \times p}$ and the RNA-seq data as $X_{\text{rna}} \in \mathbb{R}^{m \times q}$, where rows are cells and columns are genes. Generally, spatial gene prediction considers the case where $q >> p$ and the genes represented in X_{rna} are a superset of the genes in X_{spatial}. Spatial gene prediction proceeds for a gene j that is measured in X_{rna} but not in X_{spatial}. Using various procedures, a prediction method predicts the expression of gene j for each cell in X_{spatial} from the information in X_{spatial} and X_{rna}. We refer to the predicted gene expression matrix as $G \in \mathbb{R}^{n \times r}$, which contains r predicted genes.

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA [2]Department of Statistics, Stanford University, Stanford, CA. Correspondence to: Eric D. Sun <edsun@stanford.edu>, James Zou <jamesz@stanford.edu>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

The general workflow for SPRITE is depicted in Figure 1A, and can be summarized as follows. Given a black-box gene expression prediction method and paired data X_{spatial} and X_{rna}, we predict gene expression for a panel of target genes that are present in X_{rna} but not in X_{spatial} and also a set of "calibration" genes that are present in both data. The latter prediction is done in a leave-one-gene-out approach as to avoid overly optimistic predictions from models exposed to the real expression of that gene. After obtaining the set of predicted spatial gene expression values, G, we then reinforce those predictions using the prediction residuals observed in the calibration genes that were measured in the spatial transcriptomics data. The Reinforce step results in the propagation of residuals across a graph and correction of the raw predicted expression values. Afterwards, the reinforced gene expression is spatially smoothed in the Smooth step so that neighbors (cells in close spatial proximity) with similar gene expression profiles will have more similar expression of predicted transcripts. In all use cases, we perform a single pass of the Reinforce step followed by a single pass of the Smooth step.

In both the Reinforce and Smooth steps, we adapt an iterative smoothing procedure (Zhou et al., 2003; Huang et al., 2020), which is succinctly represented as:

$$X^{(t+1)} = (1-\alpha)X + \alpha A X^{(t)} \tag{1}$$

where X is the matrix to smooth over, A is a normalized adjacency matrix between rows of X, and α is the smoothing parameter. Updates continue until empirical convergence of $X^{(t)} \to X^{(t+1)}$. In Reinforce, we set $X = E$ to be the prediction residuals and $A = S_{\text{gene}}$ to be the adjacency matrix corresponding to a gene correlation network. In Smooth, we set $X = G$ to be the predicted gene expression and $A = S_{\text{spatial}}$ to be the adjacency matrix corresponding to a spatial cell graph.

2.2. Reinforce using gene correlation network

The Reinforce update rule is:

$$E^{(t+1)} = (1-\alpha_r)E + \alpha_r S_{\text{gene}} E^{(t)}. \tag{2}$$

We build a gene correlation network by (1) computing pairwise Spearman correlations between all genes in the combined target and calibration gene panels (r genes in total); (2) computing a cutoff value to keep an edge between genes such that all genes have at least one neighbor; (3) using the cutoff to construct a binary adjacency matrix $S_{\text{gene}} \in \mathbb{R}^{r \times r}$. The cutoff values are context-specific and selected to ensure that information can be propagated across all genes in the spatial transcriptomics data. Using fixed cutoff values across all experiments did not significantly change the results. We propagate residuals on the gene correlation network represented by S_{gene} using Eq. (2). We define the residuals as $E = (Y - G)^T$, where G is the predicted gene expressions and Y are the measured gene expressions in X_{spatial}. For genes that were not measured in the spatial data, we set $E = 0$. We select the optimal value of α_r using nested cross-validation where for each cross-validation fold, we mask out a subset of the calibration genes (i.e. set their residual to zero) and then evaluate the improvement in mean absolute error (MAE) with respect to the known true residual values after reinforcement. We repeat this across all cross-validation folds and then select the α_r that returns the lowest MAE.

2.3. Smooth using spatial neighbors graph

The Smooth update rule is:

$$G^{(t+1)} = (1-\alpha_s)G + \alpha_s S_{\text{spatial}} G^{(t)}. \tag{3}$$

The input to the Smooth step is constructed from the output of the Reinforce step as $G^{(0)} = G + E^{(t)}$, where $E^{(t)}$ is converged residual matrix. Our approach for building spatial connectivity graphs for the cells in X_{spatial} is to first draw edges between each cell and its k-nearest neighbors by Euclidean distance. We used $k = 50$ in all experiments, but the results are generally robust to the exact choice of k. After constructing the spatial graph, we prune outlier edges by removing all edges between cells with Euclidean distance greater than 1.5 times the interquartile range of neighbor Euclidean distances. The final normalized adjacency matrix is denoted S_{spatial}. Then, SPRITE propagates the predicted gene expression values across the spatial graph according to Eq. (3) until convergence. To select the optimal α_s, we employ a simple line search across discrete choices of α_s and select the value that minimizes the MAE of predictions with respect to the true expression values of calibration genes.

2.4. Evaluation of SPRITE

We evaluate the performance of the SPRITE across eight benchmark spatial transcriptomics and RNAseq dataset pairs (Li et al., 2022; Long et al., 2023; Sun et al., 2023b; Codeluppi et al., 2018) and three spatial gene expression prediction methods (SpaGE, Harmony, Tangram) (Abdelaal et al., 2020; Biancalani et al., 2021; Allen et al., 2023). We use the Spearman correlation coefficient (SCC) and mean absolute error (MAE) as metrics of spatial gene prediction quality. To evaluate the concordance of low-dimensional visualizations to the original high-dimensional data, we computed the Spearman correlation coefficient between the pairwise Euclidean distance vectors between all cells across both modalities using the concordance score method in DynamicViz (Sun et al., 2023a).

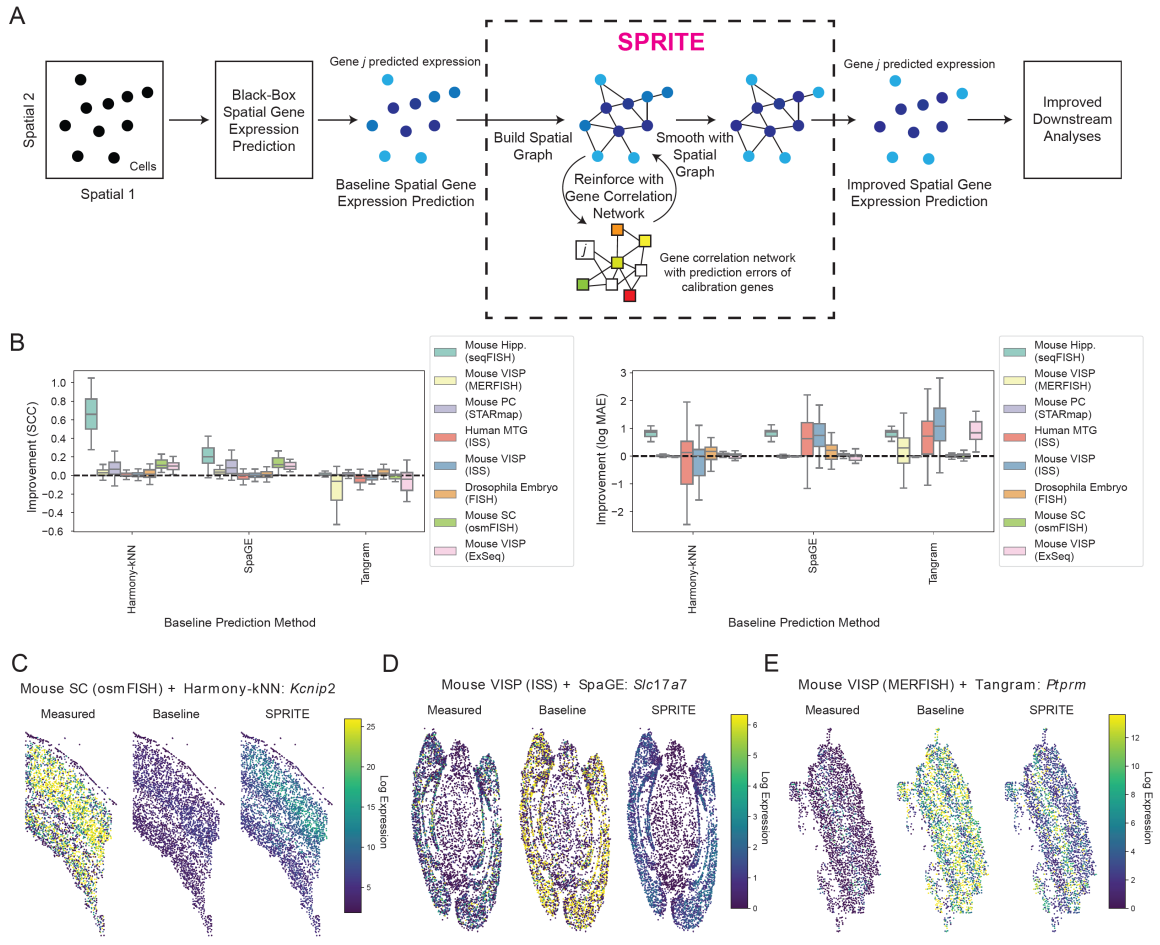

Figure 1. SPRITE improves prediction of spatial gene expression. (A) General workflow for the SPRITE algorithm including baseline spatial gene expression prediction followed by SPRITE post-processing with Reinforce and Smooth. (B-C) Improvement provided by SPRITE over baseline spatial gene expression prediction across eight spatial transcriptomics and RNAseq dataset pairs and three spatial baseline prediction methods as measured by: (B) the Spearman correlation coefficient between the predicted and measured spatial gene expression values computed for each gene across all cells; (C) the log mean absolute error between the predicted and measured spatial gene expression values computed for each gene across all cells. (C-E) Representative spatial plots of the measured gene expression, baseline predicted gene expression, and SPRITE-predicted gene expression profiles for (C) *Kcnip2* in osmFISH dataset of mouse somatosensory cortex with Harmony baseline prediction, (D) *Slc17a7* in ISS dataset of mouse visual cortex with SpaGE baseline prediction, (E) *Ptprm* in MERFISH dataset of mouse visual cortex with Tangram baseline prediction.

3. Results

3.1. SPRITE improves spatial gene expression prediction

To assess the performance of SPRITE, we applied SPRITE post-processing to spatial gene expression predictions from three methods (SpaGE, Tangram, Harmony) across eight spatial transcriptomics and RNAseq data pairs. Compared to the baseline predictions, SPRITE-predicted gene expression values were generally better correlated with the measured ground truth expression and had lower prediction errors (Figure 1B). Qualitatively, these improvements can result from either the recovery of spatial expression patterns that were missing in the original predictions (Figure 1C) or selective attenuation of overly optimistic gene expression predictions to better resemble the true spatial patterns of gene expression (Figure 1DE).

3.2. SPRITE improves downstream analyses of spatial transcriptomics

To determine whether the more accurate predicted gene expression values obtained through SPRITE would yield improvements in common downstream analyses, we considered clustering of cells, which is often used to identify

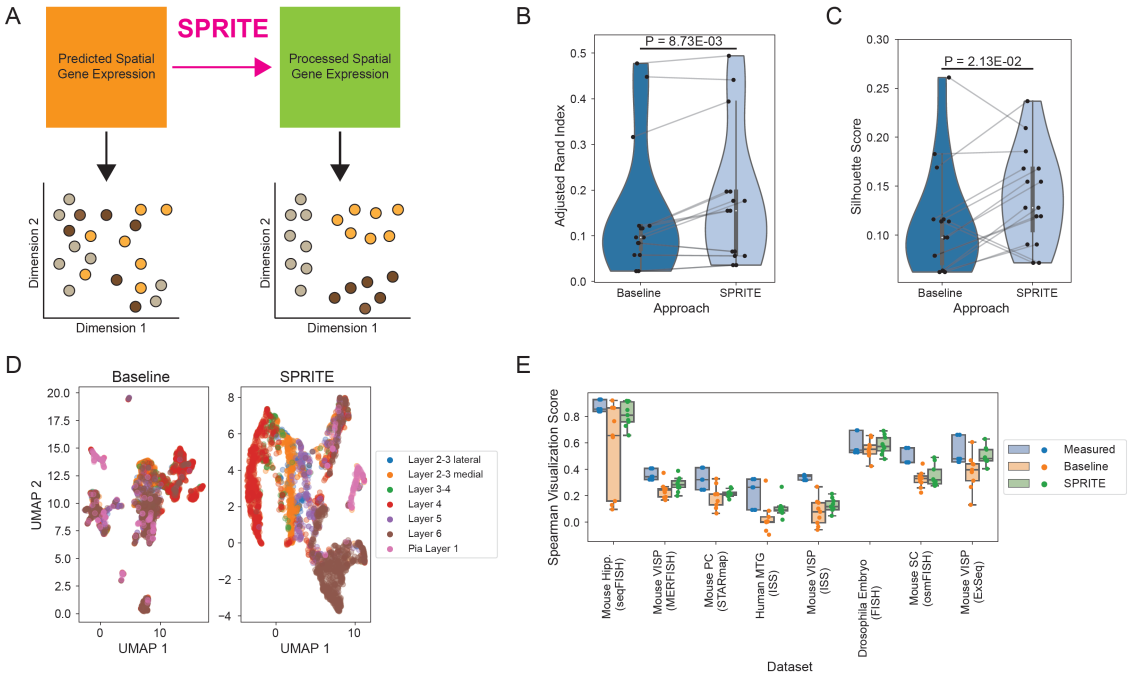

Figure 2. SPRITE improves clustering and visualization of predicted spatial gene expression. (A) Schematic diagram of clustering and visualization of cells using their baseline predicted spatial transcriptomes and SPRITE-predicted transcriptomes. (B-C) Leiden clustering quality of spatial single-cell transcriptomes across three spatial transcriptomics datasets and five metadata labels including cell type and anatomic region as measured by (B) adjusted rand index with respect to the Leiden clustering obtained on the measured spatial transcriptomes, and (C) silhouette score averaged across all cells. (D) UMAP visualization of all cells in the osmFISH mouse somatosensory cortex dataset with predicted expression of all genes. Colors correspond to anatomic region labels. (E) Spearman visualization scores for two-dimensional visualizations compared to the original high-dimensional data for all spatial transcriptomics datasets, spatial gene expression prediction methods, and three dimensionality reduction methods (UMAP, t-SNE, PCA) using either the measured gene expression, baseline predicted gene expression, or SPRITE-predicted gene expression profiles.

cell types or states, and data visualization, which is often used to intuitively understand high-dimensional spatial transcriptomics data, as two representative use cases. We compared Leiden clustering and low-dimensional visualization (PCA, UMAP, or t-SNE) of the baseline predicted spatial gene expression to those for the SPRITE-predicted spatial gene expression (Figure 2A). The clustering obtained on the SPRITE-predicted gene expression values had greater concordance with metadata labels such as cell type and anatomic region than clustering on the baseline predicted gene expression values (Figure 2B), and also resulted in higher-quality clustering of cells (Figure 2C).

The low-dimensional visualizations obtained from SPRITE-predicted gene expression were consistently better than visualizations obtained from baseline predicted gene expression at separating metadata labels and preserving the pairwise distance relations between cells with respect to the original high-dimensional data, and in some cases, were comparable in performance to visualizations obtained from the measured ground truth gene expression (Figure 2DE).

4. Discussion

SPRITE is a flexible wrapper method around any existing spatial gene expression prediction method that is highly scalable, with computational complexity on the same order of magnitude as the original prediction algorithm. We show that SPRITE generally leads to improved spatial gene expression predictions, rescues spatial gene expression patterns, and is extendable to improved clustering and visualization of the predicted gene expression data.

Further evaluation of SPRITE in the context of other downstream analyses, particularly in training machine learning models or analyzing differential gene expression patterns, would highlight the extent of expected improvements from SPRITE predictions. Experiments with additional passes or permutations of the Reinforce and Smooth steps may further optimize the reported improvements in SPRITE predictions. Combining SPRITE with tools for estimating prediction uncertainty (Sun et al., 2023b) may provide a richer context in which to interpret the reliability of downstream conclusions.

Software and Data

All associated code notebooks and scripts can be found at https://github.com/sunericd/sprite-figures-and-analyses. All data can be found from publicly available sources that are elaborated in (Li et al., 2022) and (Sun et al., 2023b).

Acknowledgements

Funding support was provided by Knight-Hennessy Scholars program (E.D.S.), Paul and Daisy Soros Fellowship for New Americans (E.D.S.), the National Science Foundation Graduate Research Fellowship Program (E.D.S.), Professor David Donoho at Stanford University (R.M.), NSF CAREER 1942926 (J.Z.), NIH P30AG059307 (J.Z.), 5RM1HG010023 (J.Z.) and grants from the Silicon Valley Foundation (J.Z.) and the Chan-Zuckerberg Initiative (J.Z.).

References

Abdelaal, T., Mourragui, S., Mahfouz, A., and Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Research*, 48(18):e107, October 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa740. URL https://doi.org/10.1093/nar/gkaa740.

Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C., and Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1):194–208.e18, January 2023. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2022.12.010. URL https://www.cell.com/cell/abstract/S0092-8674(22)01523-9. Publisher: Elsevier.

Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C. R., Segerstolpe, A., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N. B., Fanelli, D., Zhuang, X., Macosko, E. Z., and Regev, A. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, November 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01264-7. URL https://www.nature.com/articles/s41592-021-01264-7. Number: 11 Publisher: Nature Publishing Group.

Codeluppi, S., Borm, L. E., Zeisel, A., La Manno, G., van Lunteren, J. A., Svensson, C. I., and Linnarsson, S. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, November 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0175-z. URL https://www.nature.com/articles/s41592-018-0175-z. Number: 11 Publisher: Nature Publishing Group.

Huang, Q., He, H., Singh, A., Lim, S.-N., and Benson, A. R. Combining Label Propagation and Simple Models Out-performs Graph Neural Networks, November 2020. URL http://arxiv.org/abs/2010.13993. arXiv:2010.13993 [cs].

Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., Liu, K., Zhao, X., Lin, J., Cheng, L., Chen, F., Xue, T., and Qu, K. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6):662–670, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01480-9. URL https://www.nature.com/articles/s41592-022-01480-9. Number: 6 Publisher: Nature Publishing Group.

Long, B., Miller, J., and Consortium, T. S. SpaceTx: A Roadmap for Benchmarking Spatial Transcriptomics Exploration of the Brain, January 2023. URL http://arxiv.org/abs/2301.08436. arXiv:2301.08436 [q-bio].

Moses, L. and Pachter, L. Museum of spatial transcriptomics. *Nature Methods*, pp. 1–13, March 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01409-2. URL https://www.nature.com/articles/s41592-022-01409-2. Publisher: Nature Publishing Group.

Shengquan, C., Boheng, Z., Xiaoyang, C., Xuegong, Z., and Rui, J. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics*, 37(Supplement_1):i299–i307, July 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab298. URL https://doi.org/10.1093/bioinformatics/btab298.

Sun, E. D., Ma, R., and Zou, J. Dynamic visualization of high-dimensional data. *Nature Computational Science*, 3(1):86–100, January 2023a. ISSN 2662-8457. doi: 10.1038/s43588-022-00380-4. URL https://www.nature.com/articles/s43588-022-00380-4. Number: 1 Publisher: Nature Publishing Group.

Sun, E. D., Ma, R., and Zou, J. TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses, April 2023b. URL https://www.biorxiv.org/content/10.1101/2023.04.25.538326v1. Pages: 2023.04.25.538326 Section: New Results.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, June 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.05.006.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. Learning with Local and Global Consistency. In Thrun, S., Saul, L., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2003/file/87682805257e619d49b8e0dfdc14affa-Paper.pdf.