

A Comprehensive Simulator for Spatially Resolved Transcriptomics

By

Jing Fu

B.S., The Pennsylvania State University, 2022

Thesis

Submitted in partial fulfillment of the requirements for the Degree of
Master of Science in the Department of Biostatistics at Brown University

PROVIDENCE, RHODE ISLAND
MAY 2025

This thesis by Jing Fu is accepted in its present form
by the Brown University Department of Biostatistics as satisfying the
thesis requirements for the degree of Master of Science.

Date _____
Ying Ma, Ph.D., Advisor

Date _____
Christopher Schmid, Ph.D., Reader

Date _____
Stavroula Chrysanthopoulou, Ph.D.
Director, Master's Program in Biostatistics

Approved by the Graduate Council

Date _____
Thomas A. Lewis, Ph.D., Dean of the Graduate School

AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other institutions or individuals for the purpose of scholarly research.

Date _____
Jing Fu, Author

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Date _____
Jing Fu, Author

Acknowledgements

I would like to express my sincere gratitude to Dr. Ying Ma and Dr. Christopher Schmid for their insightful guidance, constant support, and invaluable feedback throughout my research. Their mentorship has played a crucial role in shaping both the direction and quality of my work, and I am deeply thankful for their dedication and encouragement.

Abstract of A Comprehensive Simulator for Spatially Resolved Transcriptomics, by Jing Fu, Degree Sc.M., Brown University, May 2025.

Spatially resolved transcriptomics (SRT) has enabled transcriptome-wide gene expression profiling with spatial localization, leading to significant advances in understanding tissue architecture and cellular interactions. As the number of computational methods for spatial clustering, spatially variable gene detection, and gene network analysis continues to grow, there is a critical need for high-fidelity synthetic SRT data with known ground truth to support systematic benchmarking and experimental design. However, existing simulators either lack spatial modeling or fail to preserve gene–gene co-expression, limiting their utility for comprehensive method evaluation. To address this gap, we developed SPASI, a biologically informed and flexible simulator for generating synthetic SRT data. SPASI supports both reference-based and reference-free simulations and uniquely preserves two essential features: spatial expression patterns across tissue regions and gene–gene correlation structures. We achieved this by integrating gene-wise count modeling with a NORTA framework, allowing for realistic simulation of count distributions and network dependencies. We evaluated SPASI using datasets from multiple SRT platforms and demonstrated that it outperformed existing simulators, in preserving both gene-level and spatial-level statistical properties. SPASI maintained the lowest root-mean-square error and highest Pearson correlation to real data, and was the only method to preserve spatial patterns of key marker genes and the gene co-expression network. In reference-free mode, SPASI remained robust in retaining both spatial and gene fidelity. Furthermore, we applied SPASI-simulated data to benchmark downstream methods, including five spatially variable gene detection tools and two co-expression network inference algorithms. Our results demonstrated SPASI’s capacity to distinguish method performance under realistic spatial complexity. SPASI provides a general-purpose, extensible simulation platform for spatial transcriptomics, supporting reproducible benchmarking and advancing computational method development in spatial omics.

Contents

1	Background	1
2	Results	3
2.1	Workflow of SPASI	3
2.2	SPASI-generated synthetic data closely resemble real SRT data in reference-based simulations	4
2.3	SPASI is able to generate realistic SRT data in reference-free sim- ulations	8
2.4	Applications of SPASI Synthetic Data for Benchmarking	13
2.4.1	Application 1: Benchmarking SVG detection methods using synthetic SRT data	13
2.4.2	Application 2: Benchmarking gene co-expression methods using synthetic SRT data	14
3	Discussion	17
4	Conclusions	19
5	Methods	20
5.1	Overview of SPASI framework	20
5.2	Existing Simulation Models	20
5.3	SRT reference-based simulation	21
5.4	SRT reference-free simulation	25
5.5	Compared Simulators	26
5.6	Simulated Data Scenarios and Evaluations	26
5.6.1	Reference-based Simulation	26
5.6.2	Reference-Free Simulation and Comparison with SRTsim	27
5.7	Downstream Benchmarking Analysis	28
5.7.1	Evaluation of Spatially Variable Gene Detection Methods Based on Simulated Data	28

5.7.2	Evaluation of Gene Co-expression Network Analysis Based on Simulated Data	29
	References	34
	Appendices	35
	Appendix A. Data Pre-processing	35
	Appendix B. Availability of Data and Materials	36
	Appendix C.The Distributions of RMSE and Correlation for Genes . . .	37
	Appendix D.The Distributions of F1 Scores Across 4 Simulations	39

List of Figures

1	Benchmarking SPASI against ten existing simulation methods under the reference-based simulation setting using Xenium breast cancer data.(A) Comparison of six evaluation metrics, including four gene-wise metrics and two location-wise metrics, between simulated and real data across different simulators. Higher values indicate better preservation of the reference data’s statistical properties.(B) Comparison of root-mean-square error (RMSE) and Pearson correlation between simulated and real expression counts. Lower RMSE and higher Pearson correlation indicate higher fidelity to the original dataset. *Note: scDesign3 (Gaussian copula) results are missing in panel B due to failure to complete simulation on the high-resolution Xenium dataset within resource constraints (20 CPUs and 200 GB memory over 2 days). And SymSim and all three Splatter variants were excluded from panel B due to their inability to preserve either gene-wise or location-wise statistics.	6
---	--	---

2	Evaluation of SPASI and other simulators under the reference-based simulation setting using the Xenium breast cancer dataset.(A) Comparison of spatial expression patterns of three biologically significant marker genes (HPX, KLF5, and VOPP1) between simulated and real data. SPASI uniquely preserves the spatial expression signatures across all three genes.(B) Quantitative evaluation of marker gene preservation based on RMSE and Pearson correlation between simulated and real expression profiles.(C) Heatmap visualization of the gene-gene correlation structure among the top 20 highly expressed genes in the simulated data.(D) Root-mean-square error (RMSE) between the simulated and real gene-gene correlation matrices.(E) Pearson correlation between the simulated and real gene-gene correlation matrices, assessing the preservation of co-expression structure. *Note: SymSim and all three Splatter variants were excluded from panel A-E due to their inability to preserve either gene-wise or location-wise statistics.	8
3	Benchmarking SPASI against SRTsim under the reference-free simulation setting using the DLPFC 151674 sample.(A) Comparison of six evaluation metrics, including four gene-wise metrics and two location-wise metrics, between simulated and real data. Only SPASI successfully preserved both gene-wise and location-wise properties.(B) Comparison of root-mean-square error (RMSE) and Pearson correlation between simulated and real expression counts. Metrics are only computed for SPASI, as SRTsim does not retain gene-level information in the reference-free mode.	10

4	Evaluation of SPASI and SRTsim under the reference-free simulation setting using the DLPFC 151674 sample.(A) Spatial expression patterns of eight biologically important genes (PCP4, HOPX, CLDN11, HPCAL1, HS3ST2, MBP, NRGN, and OLFM1). SPASI accurately preserves spatial signatures, whereas SRTsim fails to replicate these patterns.(B) Quantitative evaluation of marker gene preservation based on RMSE and Pearson correlation between simulated and real expression profiles.(C) Heatmap visualization of gene–gene co-expression structure among the top 20 highly expressed genes in the simulated data.(D) Root-mean-square error (RMSE) and Pearson correlation between the simulated and real gene–gene correlation matrices, evaluating the preservation of co-expression relationships.	12
5	Benchmarking five spatially variable gene (SVG) detection methods using SPASI-simulated spatial transcriptomic data under the reference-free setting.(A) Visualization of four domain-specific spatial expression patterns (cortical layers L1, L3, L5, and white matter) used in the simulation design.(B) F1 score performance of SPARK, SPARK-G, SPARK-X, nnSVG, and MERINGUE across the four simulated spatial patterns. Higher F1 scores indicate better balance between sensitivity and precision in detecting SVGs.	14
6	Benchmarking two gene co-expression network inference methods using SPASI-simulated spatial transcriptomic data under the reference-free setting.(A) Root-mean-square error (RMSE) between the inferred and ground-truth gene–gene co-expression networks for CSCORE and Giotto.(B) Pearson correlation between the inferred and ground-truth co-expression networks. Lower RMSE and higher correlation indicate more accurate network reconstruction.	16

7	Model Selection Workflow in SPASI.	23
8	Distribution of RMSE.	37
9	Distribution of Correlation.	38
10	Distribution of F1 scores for five spatially variable gene (SVG) de- tection methods across four domain-specific simulation settings (L1, L3, L5, and WM). Each F1 score represents the method-level per- formance in detecting 100 true SVGs out of 1,000 simulated genes per setting.	39

1 Background

Spatially resolved transcriptomics (SRT) is a transformative technology that combines spatial localization with transcriptomic profiling, enabling the characterization of gene expression patterns within intact tissues and complex organs. A growing array of SRT platforms—including imaging-based technologies such as MERFISH[1] and seqFISH+[2], and capture-based techniques such as 10x Visium[3, 4] and Slide-seq[5, 6]—now enable transcriptome-wide measurement at cellular or near-cellular resolution across diverse tissue types. These technologies have uncovered critical spatial patterns of gene expression in development, disease progression, and tissue organization, and have driven the development of thousands of SRT-specific computational tools for spatially variable gene detection[7, 8, 9, 10] and co-expression network analysis[11, 12].

To systematically evaluate such tools and guide experimental design, synthetic SRT data with known ground truths have become essential. Synthetic datasets allow for objective benchmarking of computational methods, characterization of performance across varying noise levels or spatial complexity, and design of *in silico* experiments. However, existing simulation frameworks face substantial limitations: most are adapted from single-cell RNA-seq simulators and fail to incorporate spatial coordinates, or preserve only partial aspects of the data structure such as marginal distributions or dropout rates[13]. More recent simulators such as SRTsim[14] model spatial patterns realistically, but ignore gene–gene dependencies, while others like scDesign2[15] and scDesign3[16] preserve gene correlation structures using copulas, but lack spatial modeling.

Preserving both spatial patterns and gene–gene co-expression networks is critical for high-fidelity simulation of SRT data, especially for benchmarking methods that rely on spatial correlations or regulatory network inference. Yet to date, no existing simulator simultaneously captures both dimensions.

Here, we present a novel simulator for spatially resolved transcriptomics, which we call SPASI (Spatial Simulator for Spatially resolved transcriptomics), designed

to generate synthetic SRT data that preserve both spatial and gene characteristics of real tissues. SPASI enables the simulation of spatial transcriptomic count data in a flexible, reproducible, and biologically informed manner, thereby supporting method development, benchmarking, and experimental design in spatial transcriptomics. A key advantage of SPASI is its ability to not only maintain gene- and location-specific expression properties, but also to simultaneously preserve spatial expression patterns and gene–gene co-expression networks, a feature not available in existing simulators. This makes SPASI uniquely suited for evaluating computational methods that rely on both spatial structure and gene network information. We comprehensively characterize the properties of SPASI-generated data and demonstrate its utility in benchmarking co-expression inference methods, and spatially variable gene detection approaches.

2 Results

2.1 Workflow of SPASI

SPASI is a flexible and integrated computational framework developed to simulate spatially resolved transcriptomics (SRT) data with high biological fidelity (Methods and Materials). It offers two complementary modes of simulation: a reference-based mode, which utilizes existing SRT datasets as input, and a reference-free mode, which functions independently of any real data. In both scenarios, SPASI is capable of producing synthetic datasets that closely resemble real SRT data, making it particularly valuable for benchmarking computational methods.

In reference-based simulations, SPASI models gene expression counts by fitting an appropriate distribution to each gene based on a reference dataset. Spatial coordinates can either be taken directly from the reference or newly generated based on user specifications. To realistically reproduce gene-gene dependencies, SPASI integrates a co-expression network—either user-defined or inferred from the input data—into the simulation process. It then applies the NORTA (Normal To Anything) algorithm[17] to generate multivariate count data that reflects the intended marginal distributions while maintaining the prescribed correlation structure. The simulated gene expression values are subsequently mapped to spatial locations in a way that preserves the spatial expression patterns present in the original data (Methods and Materials).

In reference-free simulations, SPASI allows users to flexibly define the spatial structure, including resolution and geometry, as well as gene-level statistical properties such as mean and variance. It employs the same NORTA-based approach to simulate expression counts under user-defined or data-derived parameters. Depending on whether spatial domain labels are provided, SPASI supports two reference-free strategies: domain-based simulation, in which spatial domains are separately modeled and counts are allocated within each domain, and tissue-

based simulation, where all spatial locations are treated as a single unit and simulated counts are distributed randomly across them. In both cases, the framework ensures that spatial expression structure and gene correlation patterns are coherently reflected in the synthetic data. To evaluate the performance of SPASI in comparison with existing simulators, we conducted benchmarking analyses using two datasets—one designated for reference-based simulation and the other for reference-free simulation—sourced from two distinct platforms: 10x Visium[3, 4] and Xenium[18].

2.2 SPASI-generated synthetic data closely resemble real SRT data in reference-based simulations

In the reference-based setting, SPASI was compared against a total of ten other simulation tools, including eight originally developed for single-cell RNA sequencing and two tailored for spatial transcriptomics. These include ZINB-WaVE[19], SPARSim[13], SymSim[20], two variants of scDesign2[15] (with and without copula), and three variants of Splatter[21] (Splat, Splat Simple, and Kersplat) for single-cell RNA simulations; and SRTsim[14] and scDesign3[16] (with Gaussian copula) for SRT simulations. To quantitatively assess the realism of the synthetic data generated by each simulator, we computed a suite of four gene-wise metrics and two location-wise metrics (detailed in Materials and Methods), evaluating the extent to which each simulator could reproduce the statistical properties of the reference SRT data. For the reference-based benchmark, we selected the Xenium Breast Cancer dataset as the primary example for performance comparison across all methods.

Within the reference-based simulations—focusing on the tissue-based setting—we assessed the similarity between synthetic and real data by comparing the distribution of six evaluation metrics across simulators. One exception was scDesign3, which was unable to complete simulations for the high-resolution 10x Xenium dataset due to excessive computational demands: even after two days using 20

CPU cores and 200 GB of memory, no results were returned. Among the simulators evaluated, only SPASI and SRTsim were able to effectively preserve both gene-wise and location-wise statistical characteristics of the real data (Figure 1A). Specifically, ZINB-WaVE failed to capture gene-level variance; both scDesign2 variants (with and without copula modeling) successfully maintained gene-wise metrics but did not preserve location-wise properties. In contrast, SPARSim was able to retain location-wise features, but did not capture gene-level characteristics. SymSim and all three versions of Splat were unable to preserve either gene-wise or location-wise features, leading us to exclude them from subsequent comparisons due to their inability to maintain basic gene-level structure. Further analysis based on the Xenium reference data indicated that SPASI achieved the lowest root-mean-square error (RMSE) and the highest Pearson correlation between simulated and real expression values (Figure 1B), highlighting its superior fidelity in recapitulating the original data.

A defining characteristic of spatial transcriptomics is that gene expression is recorded at precise spatial coordinates, allowing spatial patterns to be retained and examined. To assess whether such patterns were preserved in the simulated data, we focused on three biologically significant marker genes from the 10x Xenium breast cancer (BC) dataset. These include HPX[22], identified as the gene most strongly associated with proteomic profiles; KLF5[23], a transcription factor known to regulate numerous downstream genes and play roles in key cellular processes such as stemness, proliferation, apoptosis, autophagy, and migration; and VOPP1[24], whose loss induces apoptosis, while ectopic overexpression promotes survival in cancer cells, indicating its function as a potential anti-apoptotic factor. Among all evaluated simulators, SPASI was the only one capable of preserving the spatial expression patterns of all three marker genes in the synthetic datasets (Figure 2A). In contrast, SRTsim failed to retain the spatial pattern for HPX and VOPP1, while both versions of scDesign2 were unable to preserve any of the three genes' spatial expression. To quantitatively confirm these observations, we

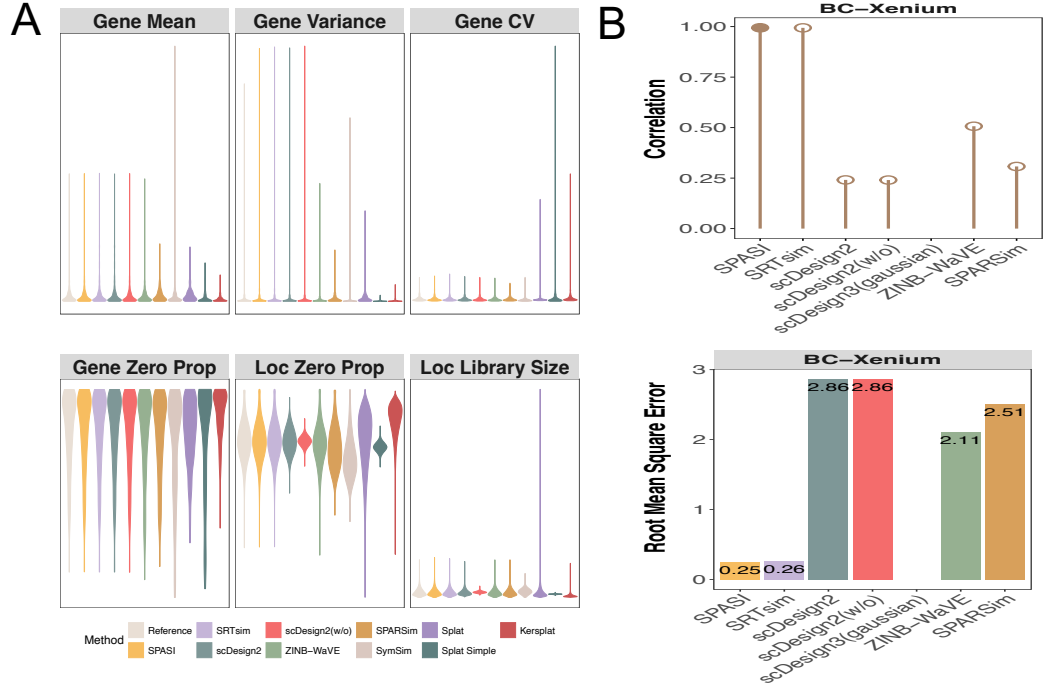


Figure 1: Benchmarking SPASI against ten existing simulation methods under the reference-based simulation setting using Xenium breast cancer data. (A) Comparison of six evaluation metrics, including four gene-wise metrics and two location-wise metrics, between simulated and real data across different simulators. Higher values indicate better preservation of the reference data’s statistical properties. (B) Comparison of root-mean-square error (RMSE) and Pearson correlation between simulated and real expression counts. Lower RMSE and higher Pearson correlation indicate higher fidelity to the original dataset. *Note: scDesign3 (Gaussian copula) results are missing in panel B due to failure to complete simulation on the high-resolution Xenium dataset within resource constraints (20 CPUs and 200 GB memory over 2 days). And SymSim and all three Splatter variants were excluded from panel B due to their inability to preserve either gene-wise or location-wise statistics.

computed RMSE and Pearson correlation coefficients between the simulated and reference gene expression profiles. SPASI achieved the highest correlation and lowest RMSE across all three genes (Figure 2B), emphasizing its strength in replicating biologically meaningful spatial expression signatures. Another distinctive aspect of SPASI is its integration of gene–gene co-expression networks into the simulation process. As visualized through the correlation heatmap of the top 20 highly expressed genes (Figure 2C), SPASI accurately retained the co-expression structure. Further quantitative analysis comparing the gene–gene correlation matrices between simulated and real data confirmed that SPASI generated the closest match, achieving the highest correlation and lowest RMSE among all simulators (Figure 2D and E).

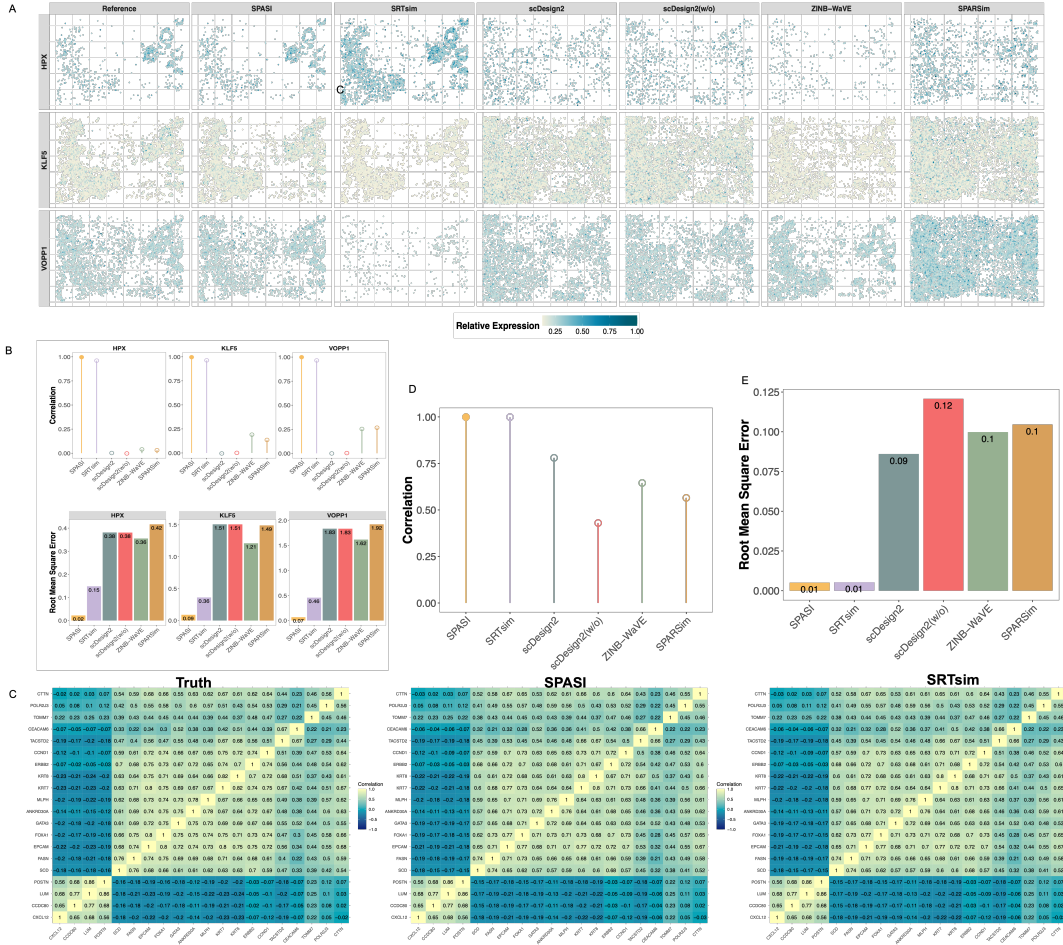


Figure 2: Evaluation of SPASI and other simulators under the reference-based simulation setting using the Xenium breast cancer dataset. (A) Comparison of spatial expression patterns of three biologically significant marker genes (HPX, KLF5, and VOPP1) between simulated and real data. SPASI uniquely preserves the spatial expression signatures across all three genes. (B) Quantitative evaluation of marker gene preservation based on RMSE and Pearson correlation between simulated and real expression profiles. (C) Heatmap visualization of the gene-gene correlation structure among the top 20 highly expressed genes in the simulated data. (D) Root-mean-square error (RMSE) between the simulated and real gene-gene correlation matrices. (E) Pearson correlation between the simulated and real gene-gene correlation matrices, assessing the preservation of co-expression structure. *Note: SymSim and all three Splatter variants were excluded from panel A-E due to their inability to preserve either gene-wise or location-wise statistics.

2.3 SPASI is able to generate realistic SRT data in reference-free simulations

We next evaluated SPASI's performance in reference-free simulations by conducting domain-based synthetic data generation using model parameters inferred

from real datasets. Given that SRTsim is currently the only other simulator capable of operating in a reference-free mode, it served as our sole comparator in this setting. For this analysis, we utilized the human dorsolateral prefrontal cortex (DLPFC) sample 151674 from the 10x Visium platform as the representative dataset for assessing reference-free simulation performance.

Unlike reference-based simulations that require a full SRT expression matrix as input, reference-free simulations operate solely on parameter estimates. Although SPASI allows users to freely define these parameters, for the purpose of objective benchmarking with known ground truth, we chose to extract parameters directly from the real DLPFC data. This strategy ensured that the resulting synthetic datasets bore strong resemblance to the reference, enabling meaningful comparisons with SRTsim. To quantify performance, we applied six established evaluation metrics, along with root-mean-square error (RMSE) and Pearson correlation, to assess how well the simulated data replicated the characteristics of the real data. In line with our findings from the reference-based setting, only SPASI successfully retained both gene-wise and location-wise features, while SRTsim failed to preserve either (Figure 3A). Accordingly, we did not compute RMSE or Pearson correlation between simulated and real gene expression counts, nor did we evaluate gene-gene correlation or similarity between correlation matrices for SRTsim in Figures 3B, 4B, 4C, and 4D, as SRTsim does not retain gene identity or preserve gene-level information in the reference-free setting, making such comparisons inappropriate. This drop in performance can be attributed to several limitations of SRTsim in the reference-free mode: (1) it uses a single set of parameters for each domain rather than gene-specific parameters, leading to less accurate modeling of gene expression; (2) gene identities are not retained, resulting in loss of gene-level information; and (3) unlike in the reference-based setting where simulated values are reordered to match reference ranks, this adjustment is not possible without access to ranks or gene identities in the reference-free mode. These factors collectively contribute to the poor performance of SRTsim in this setting. SPASI’s synthetic

data, generated using parameters learned from the real dataset, maintained high Pearson correlation and achieved low RMSE when compared to the original expression matrix (Figure 3B). Although there is no strict universal threshold for what constitutes a “low” RMSE, it can be interpreted as the standard deviation of the residual error between the simulated and real expression values. In our dataset, expression counts typically ranged from 0 to 25 or higher, and the observed RMSE values, 1.85 indicate small deviations relative to this scale. This suggests that SPASI captures the essential structure of the reference data with high fidelity.

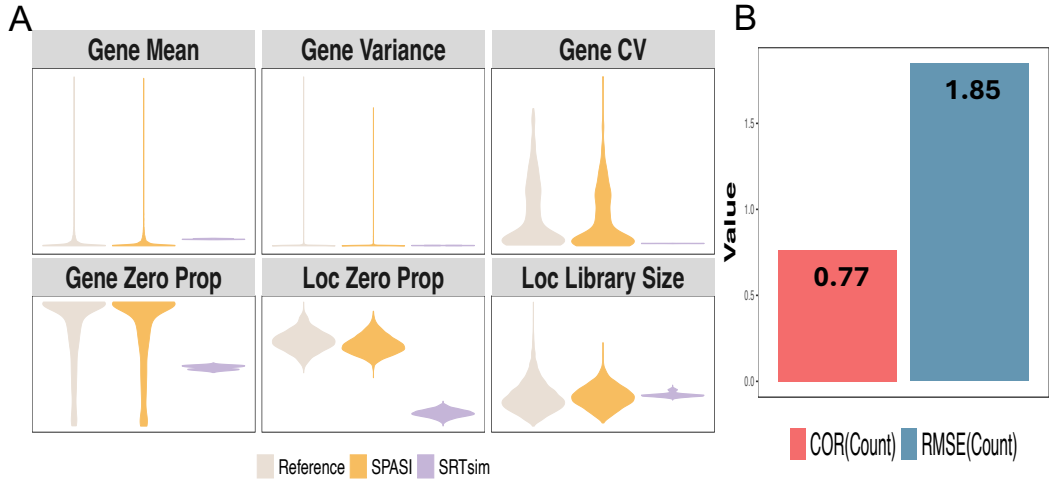


Figure 3: Benchmarking SPASI against SRTsim under the reference-free simulation setting using the DLPFC 151674 sample. (A) Comparison of six evaluation metrics, including four gene-wise metrics and two location-wise metrics, between simulated and real data. Only SPASI successfully preserved both gene-wise and location-wise properties. (B) Comparison of root-mean-square error (RMSE) and Pearson correlation between simulated and real expression counts. Metrics are only computed for SPASI, as SRTsim does not retain gene-level information in the reference-free mode.

We further evaluated the preservation of spatial expression patterns by examining eight biologically important genes. For example, PCP4[25] is known to regulate Purkinje cell excitability and cardiac rhythm, and HOPX[26, 27] encodes a key homeodomain protein essential for cardiac development; CLDN11, associated with oligodendrocyte tight junctions and brain development[28]; HPCAL1, involved in calcium signaling and neuronal activity[29]; HS3ST2, a sulfotrans-

ferase implicated in neuronal development and neurodegenerative conditions[30]; MBP, a major structural protein of the myelin sheath[31]; NRG1, a postsynaptic calmodulin-binding protein involved in synaptic plasticity[32]; and OLFML1, which plays a role in neurogenesis and neural crest development[33]. SPASI accurately reproduced the spatial expression patterns of all eight genes, whereas SRTsim failed to replicate these patterns (Figure 4A). Correlation and RMSE metrics computed between SPASI-simulated gene counts and reference data confirmed high fidelity for all eight genes (Figure 4B). In addition, we visualized gene–gene co-expression structures using correlation plots of the top 20 highly expressed genes. These revealed that SPASI was the only simulator to maintain co-expression relationships in a reference-free context (Figure 4C). Quantitative analysis further validated this result, showing that SPASI produced the highest correlation and lowest RMSE between the simulated and reference correlation matrices, affirming that SPASI’s synthetic data retained the complex data structure present in the real dataset (Figure 4D). While these metrics were computed globally across all genes and spatial locations, we acknowledge that local variations may exist. In supplementary analyses, we examined RMSE and correlation distributions across individual genes (see Figures 8 and 9 in Appendix C), calculated between the simulated expression vector for each gene and the corresponding reference vector across all spatial locations. Interestingly, while SPASI maintained consistently low RMSE across most genes, gene-wise correlation coefficients were often low. This reflects a limitation of the reference-free simulation mode: since SPASI focuses on preserving global gene–gene correlation structures rather than fitting the spatial expression of each individual gene, the synthetic data may not accurately recover the precise correlation network of every gene. As a result, the overall gene–gene co-expression matrix is well preserved (Figure 4D), but per-gene correlation with real expression can be low. This trade-off is an inherent feature of reference-free simulation, and future work may explore hybrid strategies that balance global structure with individual gene fidelity.

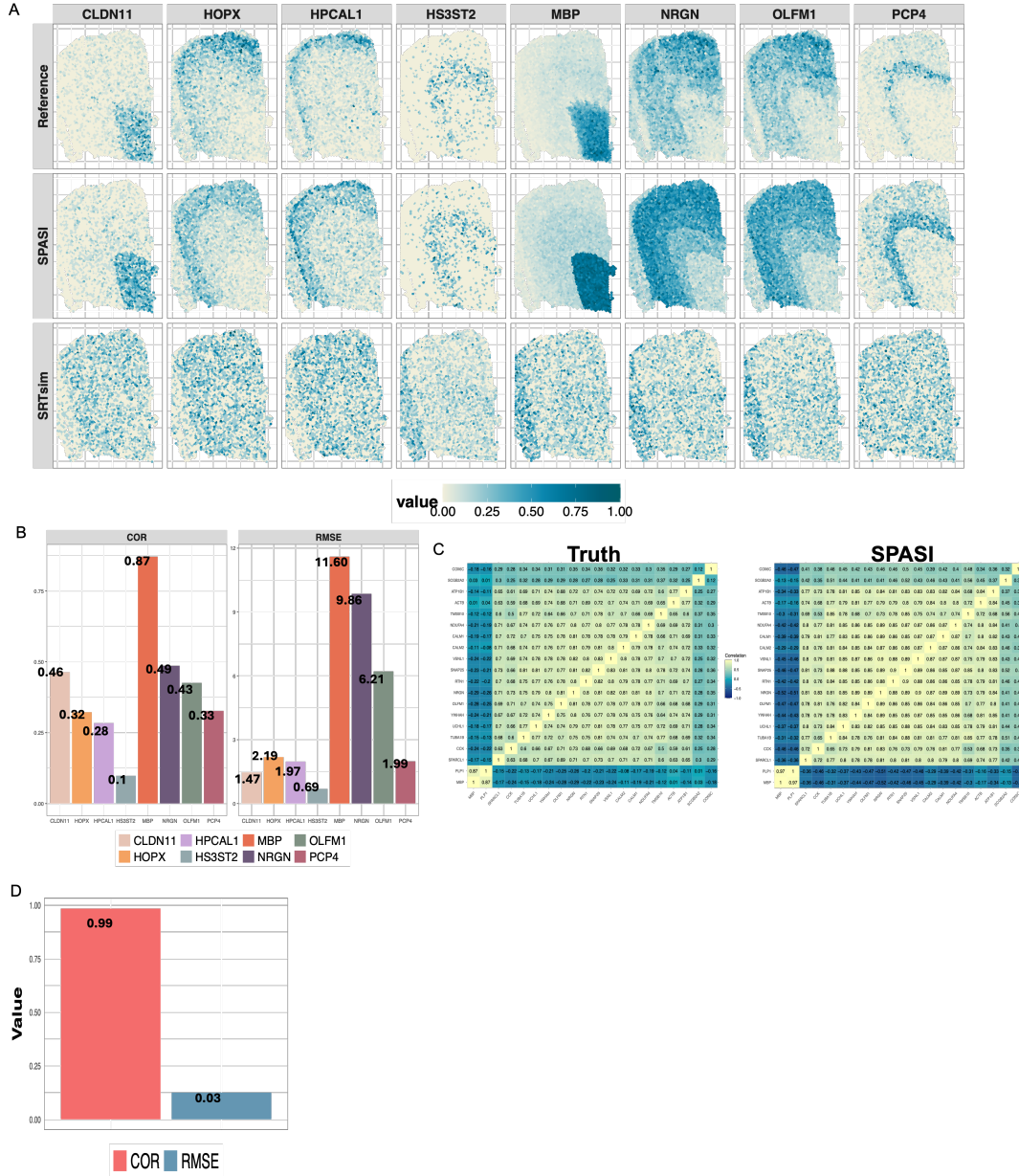


Figure 4: Evaluation of SPASI and SRTsim under the reference-free simulation setting using the DLPFC 151674 sample. (A) Spatial expression patterns of eight biologically important genes (PCP4, HOPX, CLDN11, HPCAL1, HS3ST2, MBP, NRGN, and OLFM1). SPASI accurately preserves spatial signatures, whereas SRTsim fails to replicate these patterns. (B) Quantitative evaluation of marker gene preservation based on RMSE and Pearson correlation between simulated and real expression profiles. (C) Heatmap visualization of gene-gene co-expression structure among the top 20 highly expressed genes in the simulated data. (D) Root-mean-square error (RMSE) and Pearson correlation between the simulated and real gene-gene correlation matrices, evaluating the preservation of co-expression relationships.

2.4 Applications of SPASI Synthetic Data for Benchmarking

2.4.1 Application 1: Benchmarking SVG detection methods using synthetic SRT data

We demonstrated the applicability of SPASI-generated synthetic data for benchmarking computational methods that identify spatially variable genes (SVGs). To this end, we performed reference-free simulations using parameters derived from the 10x Visium DLPFC dataset, and assessed the performance of five widely used SVG detection methods: SPARK[7], SPARK-G[8], SPARK-X[8], nnSVG[9], and MERINGUE[10]. To emulate realistic spatial expression, we designed four synthetic spatial patterns reflective of those observed in the DLPFC reference data. Each simulation comprised 1,000 genes, among which 100 were designated as SVGs following one of the four spatial patterns. These SVGs were simulated with strong spatial signals, where the expression in their active domain was set to be fivefold higher than in all other domains, mimicking domain-specific enrichment. We evaluated each method’s performance by computing the F1 score, which balances sensitivity and precision. Here, $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$, and $\text{sensitivity} = \text{TP} / (\text{TP} + \text{FN})$. For each of the four spatial settings (L1, L3, L5, and WM), we performed one representative simulation, generating 1,000 genes including 100 true SVGs with strong domain-specific enrichment. The F1 score for each method was calculated as the harmonic mean of precision and sensitivity, based on the 2×2 confusion matrix comparing predicted and true SVGs. The F1 score reported in Figure 5B corresponds to the average F1 across all 1,000 genes in each simulation. To visualize the distribution of F1 scores across the four simulations, we provide the corresponding boxplots in Appendix D. The four spatial patterns—corresponding to cortical layers L1, L3, L5, and white matter (WM)—are shown in Figure 5A. Across all scenarios, both SPARK-X and SPARK-G consistently achieved the highest F1 scores and demonstrated robust

performance across spatial contexts (Figure 5B). While nnSVG and MERINGUE outperformed SPARK in most cases, nnSVG showed reduced sensitivity in detecting genes enriched within layer L3, likely due to the lower expression levels of some SVGs in that region. These results highlight SPASI’s utility in generating synthetic spatial transcriptomic datasets with known spatial structure, enabling fine-grained evaluation of SVG detection methods under realistic and controlled conditions.

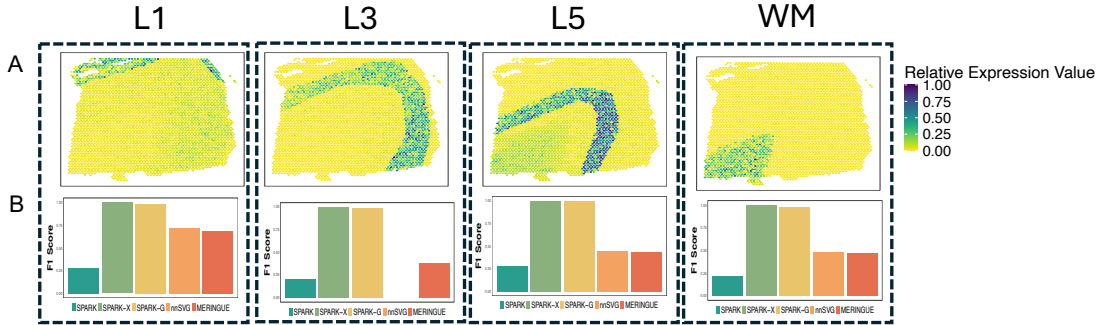


Figure 5: Benchmarking five spatially variable gene (SVG) detection methods using SPASI-simulated spatial transcriptomic data under the reference-free setting. (A) Visualization of four domain-specific spatial expression patterns (cortical layers L1, L3, L5, and white matter) used in the simulation design. (B) F1 score performance of SPARK, SPARK-G, SPARK-X, nnSVG, and MERINGUE across the four simulated spatial patterns. Higher F1 scores indicate better balance between sensitivity and precision in detecting SVGs.

2.4.2 Application 2: Benchmarking gene co-expression methods using synthetic SRT data

We further demonstrated the utility of SPASI-generated data for benchmarking gene co-expression network inference methods. For this evaluation, we employed reference-free simulations using model parameters derived from renal cancer data generated by the Decoder-seq platform[34]. Specifically, we first estimated a gene–gene co-expression network from the real dataset and then used it as ground truth for simulating synthetic data (see Materials and Methods).

To assess method performance, we compared two representative network inference tools: CS-CORE[11], originally developed for single-cell RNA-seq analysis,

and Giotto[12], designed specifically for spatial transcriptomics. Both methods were applied to the SPASI-simulated dataset, and their inferred co-expression networks—defined as gene–gene Pearson correlation matrices—were compared to the ground-truth network derived from the real expression matrix. Figure 6A shows heatmaps of the gene–gene correlation matrices estimated by each method, as well as the ground-truth matrix used in the simulation. To ensure readability, we visualized only the top 200 most highly expressed genes, although the full correlation matrices used for evaluation were computed across all genes. Due to the large number of genes, axis labels are omitted in the heatmaps to highlight overall correlation structure patterns rather than individual gene names. To quantitatively assess network recovery, we computed the root-mean-square error (RMSE) and Pearson correlation between the full inferred and ground-truth correlation matrices (Figure 6B). Since this experiment involved a single simulation, no variability or confidence intervals are reported. The corresponding RMSE and correlation values are shown directly on the bar plots.

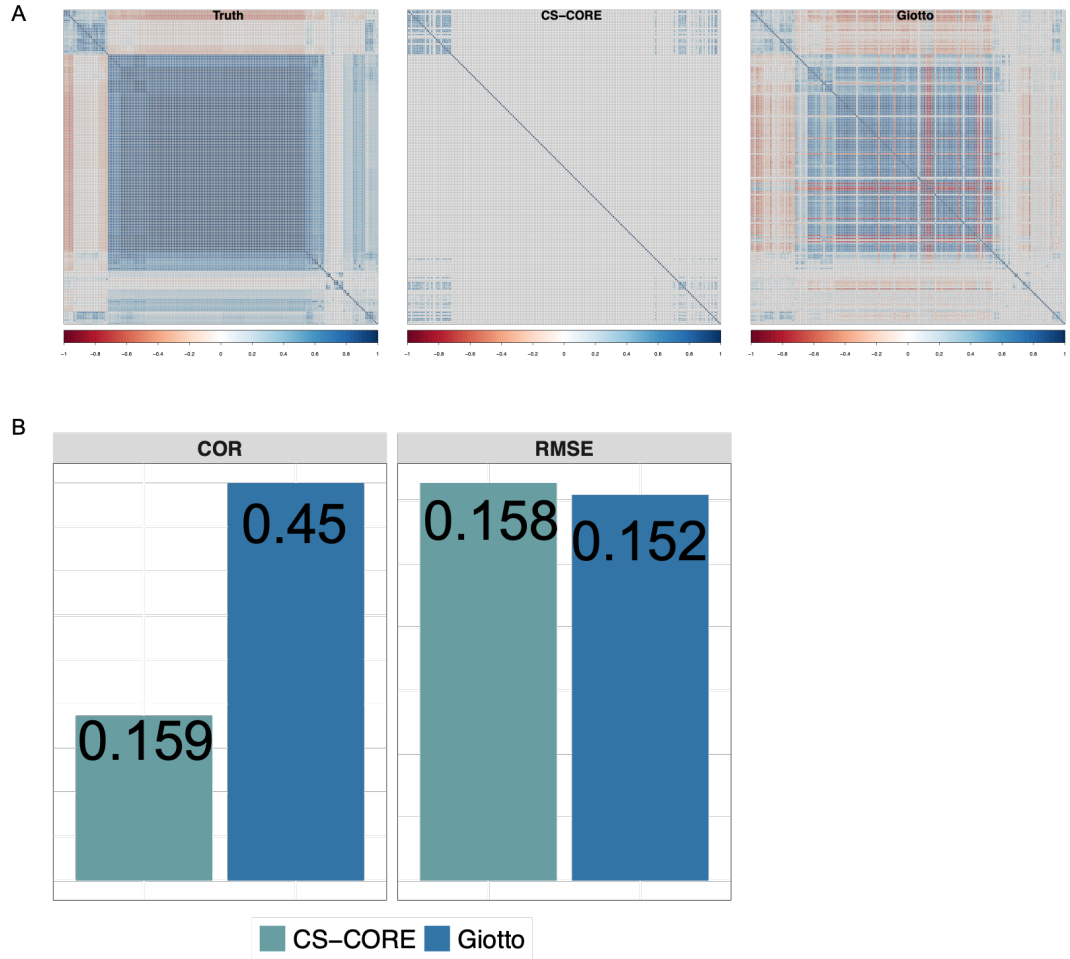


Figure 6: Benchmarking two gene co-expression network inference methods using SPASI-simulated spatial transcriptomic data under the reference-free setting. (A) Root-mean-square error (RMSE) between the inferred and ground-truth gene–gene co-expression networks for CS-CORE and Giotto. (B) Pearson correlation between the inferred and ground-truth co-expression networks. Lower RMSE and higher correlation indicate more accurate network reconstruction.

3 Discussion

We have developed SPASI, a flexible and biologically informed simulator for spatially resolved transcriptomics (SRT) that preserves both spatial gene expression patterns and gene–gene co-expression structures. By addressing limitations of existing simulators—such as the inability to jointly model spatial architecture and gene correlation—SPASI enables high-fidelity simulations for benchmarking a wide range of spatial transcriptomic methods.

Our evaluations demonstrate that SPASI consistently produces synthetic datasets that closely resemble real SRT data across both reference-based and reference-free settings. However, as our benchmarking was performed on one dataset per setting, additional evaluation across a broader range of datasets will be valuable to further assess the generalizability of SPASI’s performance. In reference-based simulations, SPASI outperformed existing methods by achieving the lowest root-mean-square error (RMSE) and highest Pearson correlation between simulated and reference expression values, particularly for key marker genes with known spatial patterns. Unlike SRTsim and scDesign2, which failed to preserve spatial features in certain genes, SPASI accurately maintained biologically meaningful spatial expression patterns across all evaluated genes. A distinguishing feature of SPASI is its incorporation of gene–gene correlation structures using NORTA algorithm. Visualization and quantitative analyses of correlation matrices confirmed that SPASI was the only simulator capable of reproducing the gene co-expression network observed in real datasets. This dual preservation is crucial for benchmarking methods such as spatial gene network inference or multi-gene spatial domain detection.

SPASI also supports reference-free simulations, allowing users to generate synthetic spatial transcriptomic data for hypothetical or synthetic tissue layouts. Despite operating without reference data, SPASI retained strong fidelity to real data in terms of gene-wise and location-wise metrics, outperforming SRTsim in both spatial pattern preservation and co-expression recovery. We further demonstrated

the practical utility of SPASI-generated data for method benchmarking. In testing five spatially variable gene detection methods, SPASI revealed differences in detection power across cortical layers and methods, illustrating SPASI’s capacity to disentangle performance under realistic spatial complexity. Additionally, SPASI-enabled synthetic datasets supported benchmarking of co-expression network inference, highlighting its flexibility for both spatial and regulatory modeling tasks.

Despite these strengths, SPASI currently does not model within-spot cellular heterogeneity and assumes a fixed cell-per-spot structure. Future development will focus on integrating spatially varying cell type mixtures and extending support to multi-modal data such as spatial proteomics. Furthermore, adapting SPASI to model dynamic or temporal spatial processes would expand its utility in developmental and disease progression studies.

4 Conclusions

In conclusion, SPASI offers a robust and extensible platform for simulating spatial transcriptomic data with high biological realism. Its ability to preserve both spatial topology and gene interaction networks makes it uniquely positioned to support reproducible benchmarking, method development, and hypothesis-driven exploration in the rapidly evolving field of spatial omics.

5 Methods

5.1 Overview of SPASI framework

SPASI is a versatile, flexible, and comprehensive simulator developed for generating gene expression count data in spatially resolved transcriptomics (SRT) studies (Figure 1). There are two key components of SPASI’s framework: (1) SRT reference-based (2) SRT reference-free. In the SRT reference-based component, SPASI simulates the gene expression data that resembles the provided reference SRT data. In the SRT reference-free framework, SPASI simulates the data accurately and robustly without the need for user-provided SRT data and spatial information. In comparison to existing simulation tools, SPASI captures a broader spectrum of biological complexity. It accounts for varying count distributions, spatial expression heterogeneity, cell type heterogeneity, and gene–gene co-expression relationships—features that are often inadequately modeled or entirely omitted by currently available simulators[13, 14, 15, 16, 19, 20, 21].

5.2 Existing Simulation Models

Several simulation frameworks have been developed to generate synthetic transcriptomic data, each with distinct modeling strategies. Simulators such as Splatter and SPARSim focus on modeling gene expression counts using hierarchical models like gamma–Poisson or Gamma–Multivariate Hypergeometric distributions, capturing basic properties such as sparsity, variability, and count intensity. ZINB-WaVE introduces a zero-inflated negative binomial model with latent factors to account for overdispersion and dropout. scDesign2 and scDesign3 extend beyond marginal modeling by incorporating gene–gene correlations via copula-based approaches, with scDesign3 additionally supporting spatial and multimodal simulation under a unified probabilistic framework. SRTsim, developed specifically for spatial transcriptomics, preserves spatial expression patterns through reference-based or user-defined spatial coordinates but does not model gene–gene

co-expression. While these simulators address different aspects of transcriptomic complexity, most are limited to either spatial structure or gene correlation, but not both.

5.3 SRT reference-based simulation

In the reference-based simulator, SPASI utilizes existing SRT datasets that provide both gene expression measurements and corresponding spatial coordinates. SPASI utilizes a gene co-expression network from either publicly available databases (e.g., STRING[35, 36]) or calculated from the user’s reference data using `cor` function from the `stats` package. SPASI can either use spatial coordinates directly from the reference SRT data or generate new spatial coordinates based on user requirements. For each gene, SPASI fits an appropriate statistical model to capture its count distribution, selecting the best-fitting distribution based on the reference expression data. Additionally, SPASI integrates the gene co-expression network into the model to account for the gene-gene correlations. Users can choose between tissue-based simulations, which utilize the entire tissue’s reference expression data, or domain-based simulations, where simulations are performed within each domain and results are combined across domains. Specifically, SPASI carries out the following four steps to generate reference-based simulations:

Step 1: Obtain spatial location information. SPASI uses spatial coordinates directly from the SRT reference data or creates new location coordinates if required by the user. Specifically, to create new spatial locations, we apply a concave hull algorithm [37] to obtain the peripheral coordinates of the tissue in the shape of a polygon. Subsequently, users can opt to create grids of location coordinates or generate a specified number of random coordinates within the original polygon.

Step 2: Fit the model. We fit an expression count model for each gene based on the reference SRT data. SPASI considers four candidate count models: Poisson,

zero-inflated Poisson (ZIP), negative binomial (NB), and zero-inflated negative binomial (ZINB). SPASI automatically selects the most appropriate model using dispersion information, Maximum Likelihood Estimation (MLE), and Akaike Information Criterion (AIC).

Specifically, for gene g , we first compare the empirical mean and variance of its count data. If the mean is greater than or equal to the variance, a Poisson distribution is fit. Otherwise, SPASI initially fits a negative binomial (NB) distribution. Then we assess whether zero inflation is present by comparing the empirical proportion of zeros to the expected proportion under the fitted model. For the Poisson distribution with mean μ , the expected zero proportion is calculated as $\exp(-\mu)$. For the NB distribution with mean μ and dispersion parameter θ , the expected zero proportion is calculated as:

$$\left(\frac{\theta}{\theta + \mu} \right)^\theta \quad (1)$$

where θ is the dispersion parameter and μ is the mean.

If the empirical proportion of zeros exceeds the expected value, SPASI attempts to fit a zero-inflated negative binomial (ZINB) model. A likelihood ratio test is conducted to assess whether ZINB provides a significantly better fit than NB. If ZINB significantly improves the fit, it is selected.

If the ZINB model does not yield a significantly better fit, SPASI fits a ZIP model as a fallback option. Although ZIP is based on the Poisson distribution, it may still better accommodate zero inflation in certain scenarios. The final model is selected based on AIC: if ZIP's AIC is lower than NB's by more than 2 points, ZIP is chosen; otherwise, NB is retained.

For Poisson-distributed genes, the expected proportion of zeros is calculated as $\exp(-\mu)$. However, SPASI does not consider zero-inflated models when the mean is greater than or equal to the variance, under the assumption that neither overdispersion nor zero inflation is severe enough to warrant a more complex model.

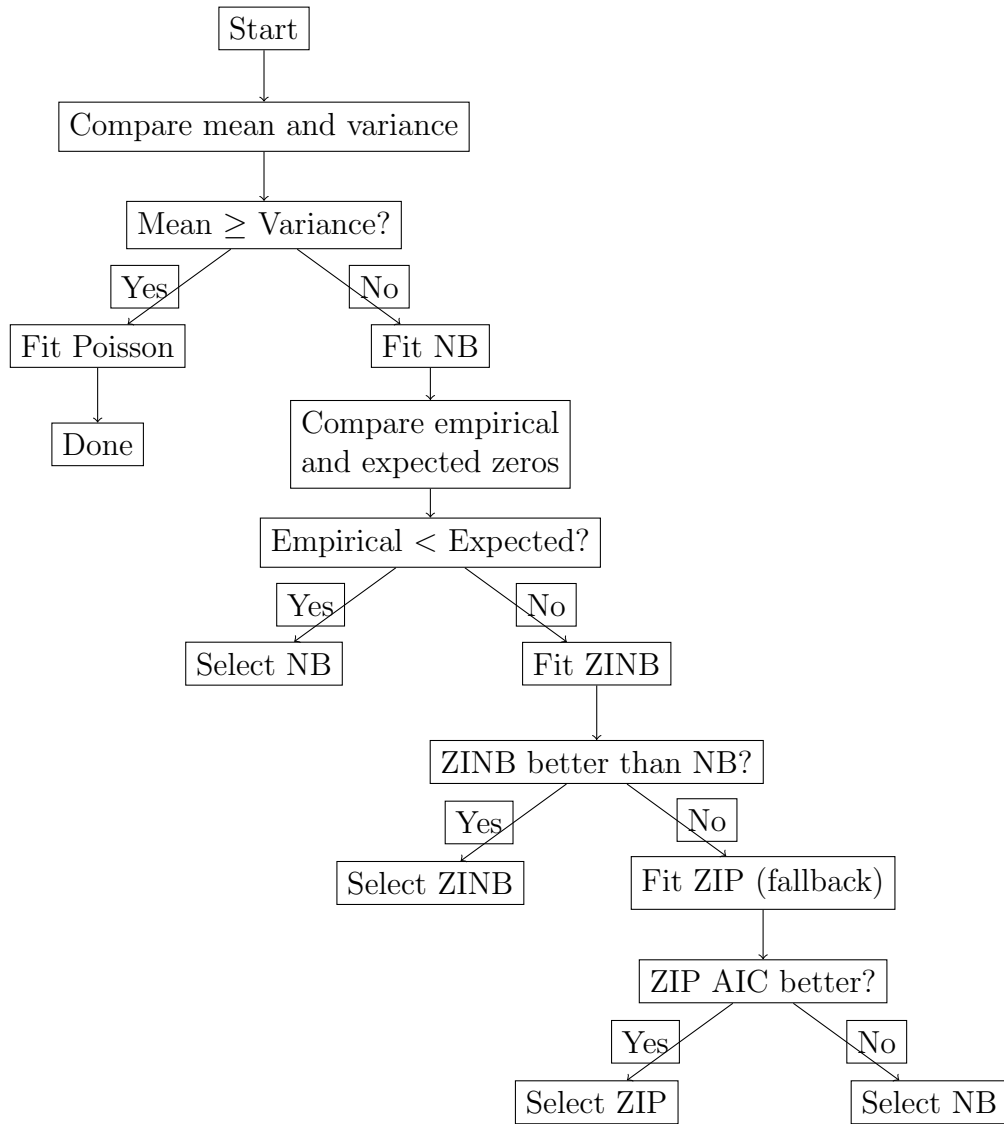


Figure 7: Model Selection Workflow in SPASI.

Step 3: Incorporate gene co-expression network. We generate simulated data by applying the Normal to Anything (NORTA) algorithm [17]. Specifically, given an SRT expression matrix $X \in \mathbb{R}^{G \times N}$ with G genes and N spatial locations, we generate the synthetic expression matrix

$$X' = \begin{bmatrix} X'_1 \\ \vdots \\ X'_G \end{bmatrix}$$

where each row vector X'_g represents the simulated expression data for gene g , fitted to marginal distribution functions $F_1, \dots, F_G$ learned in Step 2. A multivariate standard-normal matrix Z is generated such that $\text{Corr}(Z) = \Sigma$, where Σ represents the gene-gene co-expression network. The spatial locations are used collectively to estimate the gene-gene co-expression structure. Specifically, Σ is constructed as the empirical Pearson correlation matrix of the raw expression matrix across the N spatial locations, capturing the co-expression relationships between genes. Each vector Z_g of Z is then transformed using the inverse-transformation method into X'_g , corresponding to the given marginal distributions:

$$X'_g = F_g^{-1}(\Phi(Z_g))$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution, and F_g^{-1} is the inverse function for gene g . Therefore, the simulated data reflect the gene-gene co-expression structure defined by Σ .

Step 4: Spatial arrangement. The synthetic count expression data are allocated to corresponding locations according to their spatial patterns. SPASI assigns expression counts of every gene to the locations based on the rank for this gene across locations observed in the reference data.

In both tissue-based and domain-based simulations, SPASI generates gene expression data by following the steps outlined above. For tissue-based simulations,

SPASI fits a gene-specific distribution using all spatial locations from the reference dataset and applies the gene–gene co-expression network across the entire tissue. In contrast, domain-based simulations involve fitting separate distributions for each gene within individual spatial domains and integrating co-expression networks that are specific to each domain. When new spatial coordinates are generated, SPASI identifies the K nearest neighbors for each new location from the reference spatial locations. It then randomly samples expression values from these neighboring points to serve as reference expressions for the simulation. Based on these sampled values, simulated expression counts are assigned to the new locations. For tissue-based simulations, SPASI performs a unified allocation of counts across the entire tissue. In domain-based simulations, however, the allocation process is carried out independently within each domain.

5.4 SRT reference-free simulation

In addition to reference-based simulations, SPASI supports a reference-free simulation mode, which allows users to define spatial information in one of two ways: (1) domain-based: users input a spatial information file with coordinates and domain labels in a $N \times 3$ data frame, where N is the number of spatial locations and each row contains the x -coordinate, y -coordinate, and domain label for a spot; or (2) tissue-based: SPASI allow users to generate random or regular coordinates based on the convex hull algorithm[37] on various predefined shapes. If using domain-based simulations, users need to assign domain labels for each spot. Subsequently, the simulator requires users to provide gene models and a gene–gene co-expression network. Specifically, users must supply the desired model parameters for each gene (e.g., the mean parameter μ for Poisson models), a gene-by-gene correlation matrix for each domain representing the co-expression structure, and the corresponding spatial information file. Using the provided models and co-expression networks, SPASI applies the NORTA algorithm to simulate synthetic data. In tissue-based simulations, the resulting expression counts are randomly

assigned across all spatial coordinates. In contrast, under the domain-based approach, expression values are distributed randomly within each domain, ensuring that spatial domains are preserved in the simulated dataset.

5.5 Compared Simulators

We compared our results with eight single-cell simulators and two SRT simulator. The single-cell simulators included ZINB-WaVE[19] (v.1.24.0), SPARSim[13] (v.0.9.5), SymSim[20] (v.0.0.0.9000), two variants of scDesign2[15] (with/without copula) (v.0.1.0), and three Splat[21] variants from the Splatter package (v.1.26.0) (Splat, Simple Splat, and Kersplat). The SRT simulator we compared against was SRTsim[14] (v.0.99.6), and scDesign3[16] with gaussian copula (v.1.1.2). Specifically, we mainly follow their corresponding Github tutorials to run the above simulators for synthesizing SRT gene expression data.

5.6 Simulated Data Scenarios and Evaluations

We compared SPASI with 10 simulators aforementioned in reference-based simulations when taking a SRT data as reference. Because only SRTsim can simulate SRT data without a reference, we compared SPASI with SRTsim in reference-free simulations.

5.6.1 Reference-based Simulation

We utilized Xenium breast cancer data, which includes a total of 118,752 spatial spots, to perform reference-based simulations. To evaluate how well the simulated data reproduced the original dataset, we assessed six key metrics: four at the gene level and two at the location level. The gene-wise metrics captured expression characteristics for each gene across all locations, including mean expression, variance, coefficient of variation, and zero proportion. The location-wise metrics, on the other hand, evaluated properties across genes within each spatial location, specifically focusing on zero proportion and library size.

To further quantify the overall similarity between the simulated and reference datasets, we computed the root-mean-square error (RMSE) and Pearson correlation coefficient between the two full expression matrices (genes $\times$ locations). These metrics reflect global agreement across all genes and all spatial spots. Additionally, we selected three biologically relevant marker genes to visually assess spatial expression patterns. For each marker gene, we computed RMSE and Pearson correlation between its simulated and real expression vectors (i.e., gene-specific expression profiles across all locations), allowing for a more targeted evaluation of spatial preservation. Finally, to evaluate how well the gene–gene correlation structure was preserved, we first computed pairwise Pearson correlations among all genes in both the real and simulated datasets. We then compared the two resulting correlation matrices by calculating RMSE and Pearson correlation between them, thereby quantifying the degree to which the co-expression structure was maintained. We also visualized pairwise Pearson correlations among the top 20 most highly expressed genes.

5.6.2 Reference-Free Simulation and Comparison with SRTsim

We obtained the human dorsolateral prefrontal cortex (DLPFC) sample 151674 from the 10x Visium platform, which includes 3,673 spatial spots, and used it to perform reference-free simulations. To benchmark the performance of SPASI in this context, we compared its output against that of SRTsim (version 0.99.6), currently the only other simulator capable of conducting reference-free SRT simulations.

As in the reference-based setting, we evaluated synthetic data using four gene-level metrics and two location-level metrics, along with root-mean-square error (RMSE) and Pearson correlation coefficients to quantify the similarity to real data. Unlike reference-based simulations, which require both an expression matrix and spatial coordinates, reference-free simulations rely solely on user-specified model parameters. In our evaluation, we used parameters derived from the DLPFC

dataset to generate SPASI’s synthetic data, and then computed the aforementioned metrics using the original dataset as the reference to enable meaningful comparisons.

To assess the preservation of spatial expression patterns, we selected eight biologically informative genes for visualization. For each gene, we compared its simulated and real expression vectors across all spatial locations by computing root-mean-square error (RMSE) and Pearson correlation, providing gene-specific evaluations of spatial accuracy. To examine whether gene–gene co-expression structures were retained in the synthetic data, we computed pairwise Pearson correlations among all genes in both the real and simulated datasets, and visualized the pairwise Pearson correlations among top 20 highly expressed gene as correlation heatmaps. We then quantitatively assessed the similarity between the two correlation matrices by calculating RMSE and Pearson correlation between the matrix entries. Note that SRTsim does not preserve gene identity in the reference-free mode; as a result, we could not compute gene-wise expression correlations or evaluate gene–gene co-expression preservation for this method.

5.7 Downstream Benchmarking Analysis

5.7.1 Evaluation of Spatially Variable Gene Detection Methods Based on Simulated Data

To evaluate the effectiveness of SPASI for benchmarking spatially variable gene (SVG) detection methods, we conducted tissue-based, reference-free simulations using model parameters inferred from the 10x Visium dorsolateral prefrontal cortex (DLPFC) sample 151673 dataset. This dataset includes 3,639 spatial locations, each annotated with one of seven spatial domains, comprising six cortical layers and white matter.

For the simulation, we randomly selected 1,000 genes from the dataset. Among these, 100 genes were designated as SVGs, and the remaining 900 genes were modeled as non-SVGs. To introduce spatial variation for the SVGs, we adjusted the

model parameters such that, for each SVG, the mean parameter in its designated domain was set to be three times higher than the mean parameter in other domains. In contrast, for non-SVGs, the model parameters were kept consistent across all spatial locations to ensure random spatial expression without spatially specific patterns.

We evaluated three SVG detection methods, including SPARK[7], SPARK-G[8], and SPARK-X[8] from the SPARK package (v.1.1.1), nnSVG[9] (v.1.6.4), and MERINGUE[10], following their GitHub tutorials. To quantify the performance of SVG detection methods, we assessed the power of detecting SVGs based on the False Discovery Rate (FDR).

5.7.2 Evaluation of Gene Co-expression Network Analysis Based on Simulated Data

To assess the benefits of using SPASI-generated synthetic data for benchmarking gene co-expression network analysis methods, we carried out reference-free simulations. Model parameters were extracted from a renal cancer dataset produced by the Decoder-seq platform. This dataset comprised 54,368 genes across 2,500 spatial locations. For downstream analysis, we used Seurat to identify the top 1,000 most variable genes. A gene–gene co-expression network was then constructed using Pearson correlation, and this network served as the ground truth for simulating synthetic SRT data with SPASI. Specifically, we computed the Pearson correlation coefficients between each pair of genes across spatial locations, and used the resulting correlation matrix directly as the gene–gene co-expression network without additional thresholding or binarization. We evaluated the performance of two gene co-expression network analysis methods, including CS-CORE[11] (v.0.0.0.9000) and Giotto[12] (v.1.1.2). Specifically, we calculated the RMSE and Pearson’s correlation between the estimated gene-gene network and the true gene-gene network we used to simulate the SRT data.

References

- [1] K. H. Chen et al. “Spatially resolved, highly multiplexed RNA profiling in single cells”. In: *Science* 348.6233 (2015), aaa6090. DOI: 10.1126/science.aaa6090.
- [2] C.-H. L. Eng et al. “Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+”. In: *Nature* 568.7751 (2019), pp. 235–239. DOI: 10.1038/s41586-019-1049-y.
- [3] P. L. Ståhl et al. “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics”. In: *Science* 353.6294 (2016), pp. 78–82. DOI: 10.1126/science.aaf2403.
- [4] N. Rao, S. Clark, and O. Habern. “Bridging genomics and tissue pathology: 10x genomics explores new frontiers with the visium spatial gene expression solution”. In: *Genetic Engineering and Biotechnology News* 40.2 (2020), pp. 50–51.
- [5] S. G. Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. In: *Science* 363.6434 (2019), pp. 1463–1467. DOI: 10.1126/science.aaw1219.
- [6] R. R. Stickels et al. “Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2”. In: *Nature Biotechnology* 39.3 (2021), pp. 313–319. DOI: 10.1038/s41587-020-0739-1.
- [7] S. Sun, J. Zhu, and X. Zhou. “Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies”. In: *Nature Methods* (2020). DOI: 10.1038/s41592-019-0701-7.
- [8] J. Zhu, S. Sun, and X. Zhou. “SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies”. In: *Genome Biology* (2021). DOI: 10.1186/s13059-021-02404-0.

- [9] L. M. Weber et al. “nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes”. In: *Nature Communications* (2023). DOI: 10.1038/s41467-023-39748-z.
- [10] B. F. Miller et al. “Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities”. In: *Genome Research* 31.10 (2021), pp. 1843–1855. DOI: 10.1101/gr.271288.120.
- [11] C. Su et al. “Cell-type-specific co-expression inference from single cell RNA-sequencing data”. In: *Nature Communications* (2023). DOI: 10.1038/s41467-023-40503-7.
- [12] R. Dries et al. “Giotto: a toolbox for integrative analysis and visualization of spatial expression data”. In: *Genome Biology* (2021). DOI: 10.1186/s13059-021-02286-2.
- [13] G. Baruzzo, I. Patuzzi, and B. Di Camillo. “SPARSim single cell: a count data simulator for scRNA-seq data”. In: *Bioinformatics* 36.5 (2020), pp. 1468–1475. DOI: 10.1093/bioinformatics/btz752.
- [14] J. Zhu, L. Shang, and X. Zhou. “SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics”. In: *Genome Biology* (2023). DOI: 10.1186/s13059-023-02879-z.
- [15] T. Sun et al. “scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured”. In: *Genome Biology* (2021). DOI: 10.1186/s13059-021-02394-z.
- [16] D. Song et al. “scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics”. In: *Nature Biotechnology* (2024). DOI: 10.1038/s41587-023-01772-1.
- [17] M. C. Cario and B. L. Nelson. “Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix”. In: *Technical Report* (1997).

- [18] A. Janesick and et al. “High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis”. In: *Nature Communications* 14 (2023), p. 8353. DOI: 10.1038/s41467-023-40864-z.
- [19] D. Risso et al. “A general and flexible method for signal extraction from single-cell RNA-seq data”. In: *Nature Communications* (2018). DOI: 10.1038/s41467-017-02554-5.
- [20] X. Zhang, C. Xu, and N. Yosef. “Simulating multiple faceted variability in single cell RNA sequencing”. In: *Nature Communications* (2019). DOI: 10.1038/s41467-019-10500-w.
- [21] L. Zappia, B. Phipson, and A. Oshlack. “Splatter: simulation of single-cell RNA sequencing data”. In: *Genome Biology* (2017). DOI: 10.1186/s13059-017-1305-0.
- [22] N. Cine et al. “Identification of ApoA1, HPX and POTEE genes by omic analysis in breast cancer”. In: *Oncology Reports* (2014). DOI: 10.3892/or.2014.3277.
- [23] Y. Luo and C. Chen. “The roles and regulation of the KLF5 transcription factor in cancers”. In: *Cancer Science* (2021). DOI: 10.1111/cas.14910.
- [24] F. Bonin and et al. “VOPP1 promotes breast tumorigenesis by interacting with the tumor suppressor WWOX”. In: *BMC Biology* (2018). DOI: 10.1186/s12915-018-0576-6.
- [25] E. E. Kim and et al. “PCP4 regulates Purkinje cell excitability and cardiac rhythmicity”. In: *Journal of Clinical Investigation* (2014). DOI: 10.1172/JCI77495.
- [26] F. Chen and et al. “Hop is an unusual homeobox gene that modulates cardiac development”. In: *Cell* (2002). DOI: 10.1016/s0092-8674(02)00932-7.
- [27] C. H. Shin and et al. “Modulation of cardiac growth and development by HOP, an unusual homeodomain protein”. In: *Cell* (2002). DOI: 10.1016/s0092-8674(02)00933-9.

- [28] K. Morita et al. “Claudin-11/OSP-based tight junctions of myelin sheaths in brain and Sertoli cells in testis”. In: *Journal of Cell Biology* 145.3 (1999), pp. 579–588. DOI: 10.1083/jcb.145.3.579.
- [29] Xixi. Zhou, Xi. Sun, and et al. “Arsenite Interacts Selectively with Zinc Finger Proteins Containing C3H1 or C4 Motifs*”. In: *Journal of Biological Chemistry* 286.26 (2011), pp. 22855–22863. DOI: 10.1074/jbc.M111.232926.
- [30] JE. Sepulveda-Diaz, SM. Alavi Naini, and et al. “HS3ST2 expression is critical for the abnormal phosphorylation of tau in Alzheimer’s disease-related tau pathology”. In: *Brain* 138 (2015), pp. 1339–54. DOI: 10.1093/brain/awv056.
- [31] J. M. Boggs. “Myelin basic protein: a multifunctional protein”. In: *Cellular and Molecular Life Sciences* 63 (2006), pp. 1945–1961. DOI: 10.1007/s00018-006-6094-7.
- [32] S. L. Eastwood and P. J. Harrison. “Synaptic pathology in the anterior cingulate cortex in schizophrenia and mood disorders. A review and a Western blot study of synaptophysin, GAP-43 and the complexins”. In: *Brain Res Bull* 301.1 (2001), pp. 569–78. DOI: 10.1016/s0361-9230(01)00530-5.
- [33] M. Barembaum and M. Bronner-Fraser. “Olfactomedin 1 (OLFM1) promotes neural crest formation and migration”. In: *Developmental Biology* 224.2 (2000), pp. 255–266. DOI: 10.1006/dbio.2000.9809.
- [34] J. Cao et al. “Decoder-seq enhances mRNA capture efficiency in spatial RNA sequencing”. In: *Nature Biotechnology* 42 (2024), pp. 1735–1746. DOI: 10.1038/s41587-023-02086-y.
- [35] D. Szklarczyk and et al. “The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D605–D612.

- [36] D. Szklarczyk and et al. “The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D638–D646.
- [37] A. Moreira and M. Y. Santos. “Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points”. In: *Proceedings of the Second International Conference on Computer Graphics Theory and Applications* (2007). DOI: 10.5220/0002080800610068.

Appendices

Appendix A. Data Pre-processing

To maintain consistency with prior work, we followed the same gene selection strategy as described in the scDesign3 study to identify 1,000 informative genes for simulation purposes.

For both the 10x Visium and Decoder-seq datasets, we applied the Seurat R package (version 5.0.2) to extract the top 1,000 spatially variable genes. Specifically, we first constructed a Seurat object using the `CreateSeuratObject` function, setting `min.cells` and `min.features` to 0 to avoid any initial filtering of cells or genes. Next, we applied the `FindVariableFeatures` function using the `vst` method to rank genes by variability, retaining the top 1,000 most variable genes for downstream simulation. The expression matrices filtered by this selection formed the basis for reference in synthetic data generation.

For the 10x Xenium dataset, we directly used the raw gene count matrix without additional filtering, as the genes included were pre-selected during experimental design and considered representative for simulation.

Appendix B. Availability of data and materials

The DLPFC data generated with 10x Visium are available at <http://spatial.libd.org/spatialLIBD/>. The Xenium data are available at <https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast>. The Decoder-seq data are available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7585149>.

Appendix C.The Distributions of RMSE and Correlation for Genes

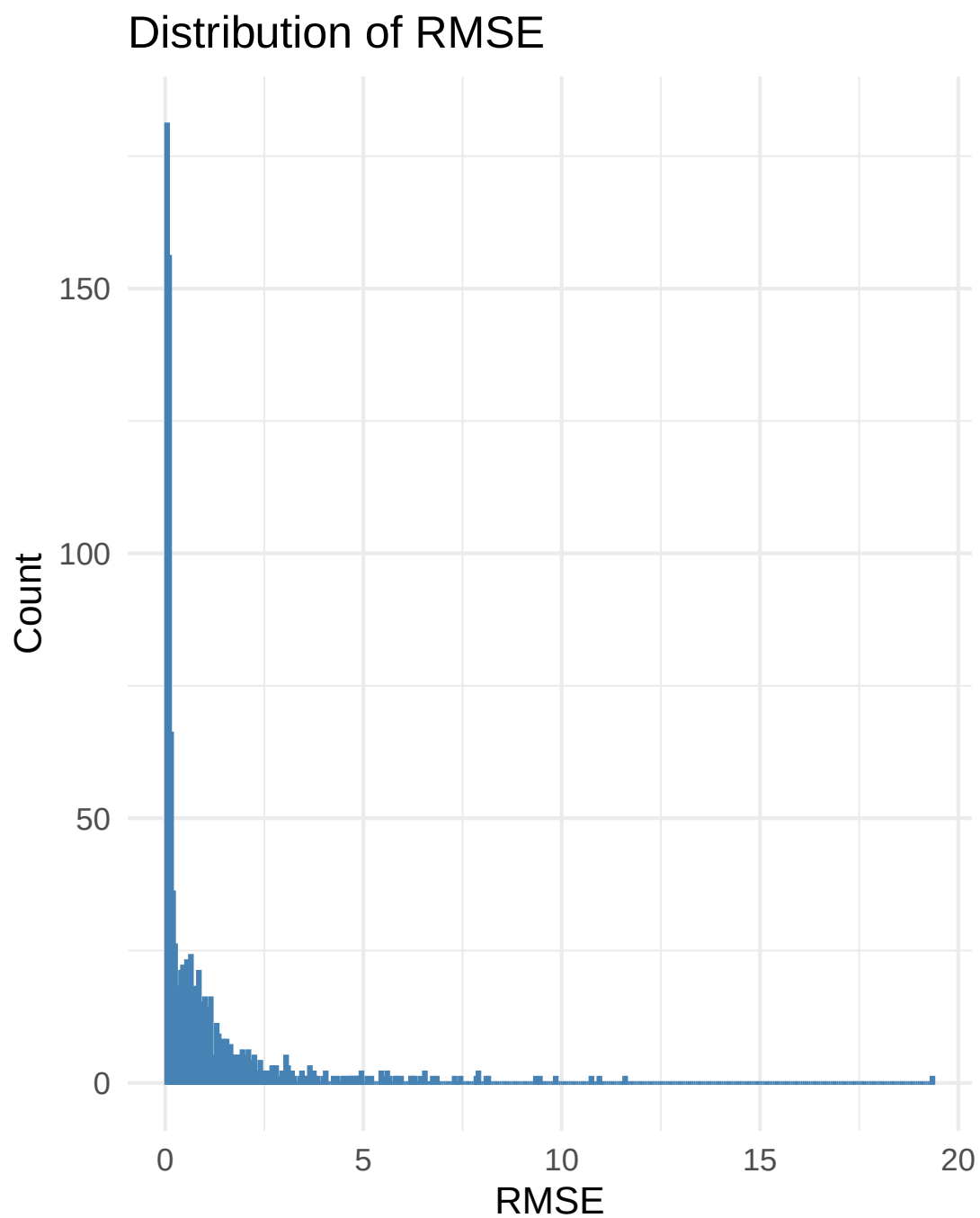


Figure 8: Distribution of RMSE.

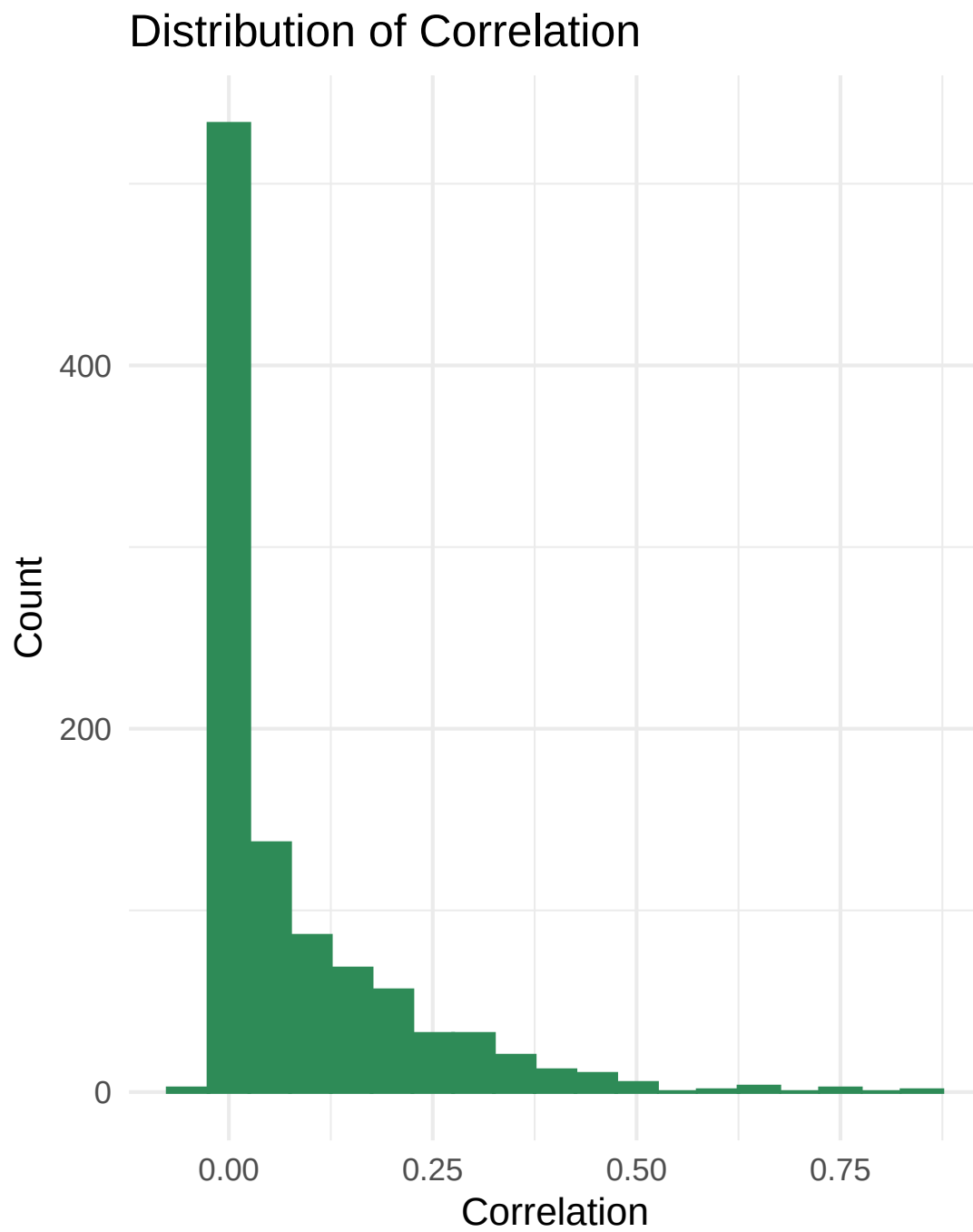


Figure 9: Distribution of Correlation.

Appendix D. The Distributions of F1 Scores Across 4 Simulations

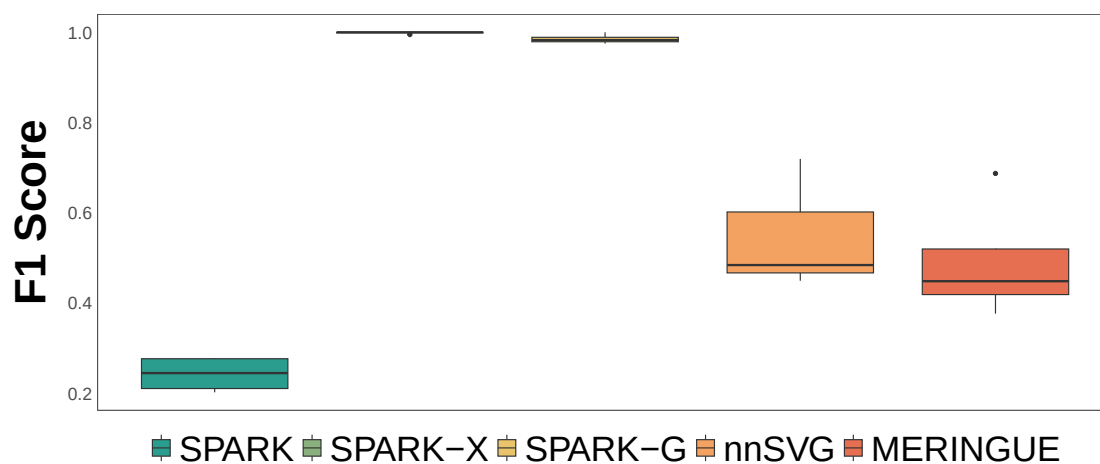


Figure 10: Distribution of F1 scores for five spatially variable gene (SVG) detection methods across four domain-specific simulation settings (L1, L3, L5, and WM). Each F1 score represents the method-level performance in detecting 100 true SVGs out of 1,000 simulated genes per setting.