Spanve: an Effective Statistical Method to Detect Spatially Variable Genes in Large-scale Spatial Transcriptomics Data

Guoxin Cai[1], Yichang Chen[1], Xun Gu[2], Zhan Zhou[1,3,4*]

1 Innovation Institute for Artificial Intelligence in Medicine and Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China.

2 Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

3 The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu 322000, China.

4 Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 310058, China.

*Corresponding author. E-mail: zhanzhou@zju.edu.cn.

Abstract

Spatial transcriptomics enables the depiction of *in situ* gene expression and could further be applied to infer the mechanism of cell functions. In this study, we present Spanve (Spatial Neighborhood Variably Expressed Genes), a statistic based approach to detect space-dependent expressed genes from spatial transcriptomics data, which has a high computation efficiency and accuracy by modeling the dependence of space and expression as a distance of two distributions. We demonstrate that Spanve is capable of imputating the spatial transcriptome and improving the identification of spatial tissue regions.

Main

Spatially resolved transcriptomics provides high-throughput sequencing of the transcriptome while retaining the spatial information of the tissue, which has facilitated biological insights in many areas in the past three years (Moses and Pachter 2022; Longo et al. 2021). Spatial transcriptomics enables the analysis of cells under the background of tissue structure, providing a new perspective of cell status. In theory, analysis approaches designed for single-cell RNA sequencing (scRNA-seq) data can be used in spatial transcriptomics, but new methods are still urgently needed to make full use of location information. The first step of processing scRNA-seq data is dealing with high dimensions, considering the efficiency and accuracy of calculation. In general, pipelines filtering genes with variance of expression is used to reduce data dimensionality, i.e. identifying highly variable genes (HVGs). In spatial transcriptomics, detecting genes with spatial expression patterns or spatially variable (SV) genes is the first and important step in describing complex spatial transcriptomic landscapes. SV genes can be caused by cell type localization or space-dependent cell interactions (Moses and Pachter 2022) and are thus informative for multiple downstream analyses, such as reconstructing tissue structure (Rao et al. 2021). SV genes are also helpful for providing biological insight, as different spatial expression patterns reflect different gene functions.

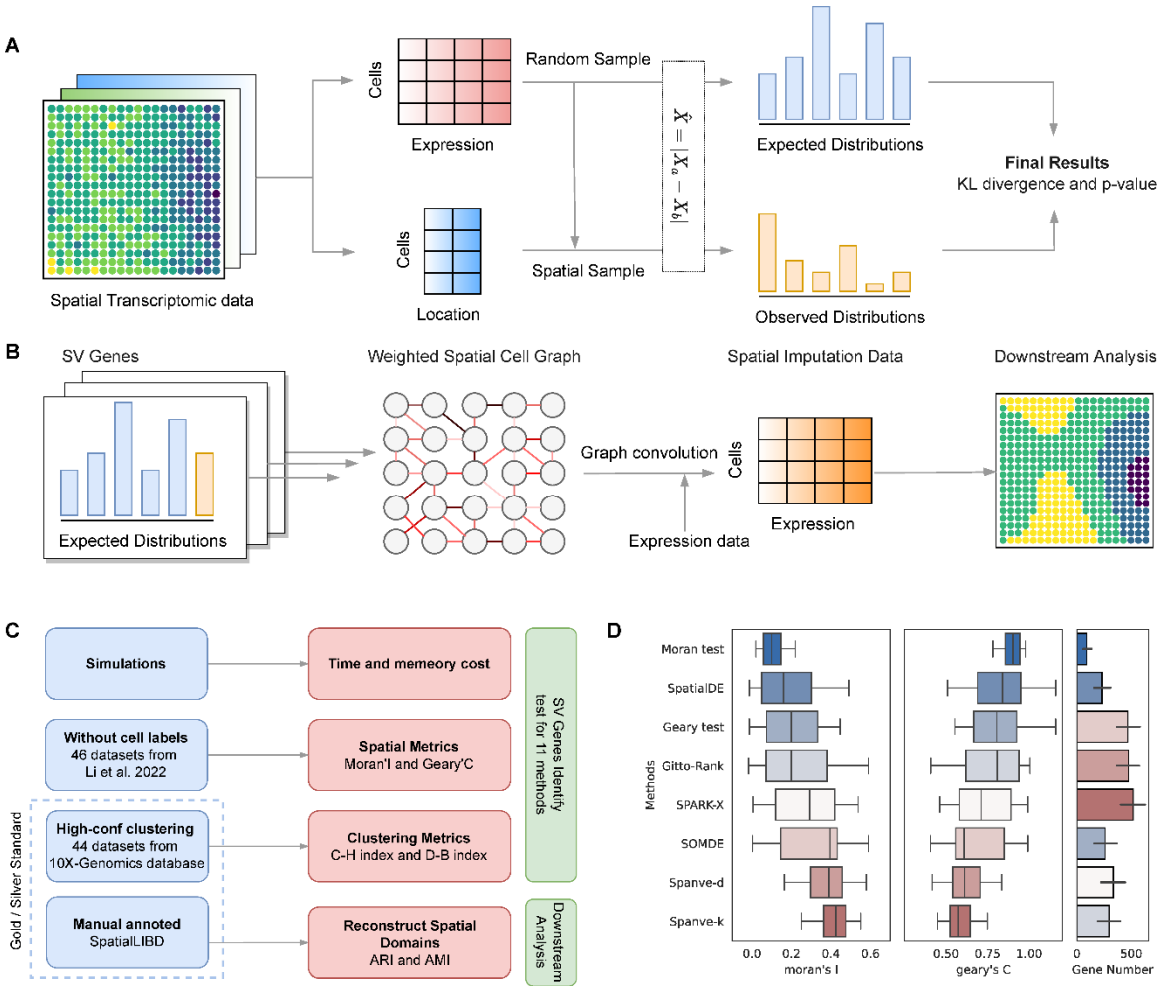


Figure 1. The framework of Spanve and experiments design in this study. (a) In spatial transcriptomic data, Spanve considers two types of sampling of cell or spot pairs: random sampling as null hypothesis and location based sampling as alternative hypothesis. Spatial dependency of genes is defined as the difference of two sampling distribution. (b) Spatial imputation employs graph convolution on a weighted spatial cell graph based on cell pairs that deviate from the expected distribution. (c) Evaluations of Spanve and other 10 methods divides into four parts. Computational cost, geographical heterogeneity, and clustering awareness are all considered while detecting SV genes. Reconstruction of spatial domains is an additional examination. (d) Spatial heterogeneity of SV genes detected by methods. In box plots, the score is represented by the color of the box; the closer the box is to

red, the better the performance. The center of box represents the median value, while the hinges indicate the quartiles.

At present, commonly used spatial transcriptome sequencing technology (for example, Visium) contains tens or more cells in a spot. With the development of sequencing technology (for example, stereo-seq (Chen et al. 2022)), it is possible to locate cellular or subcellular transcriptomes (Liao et al. 2021). However, the progress of new technology raises new challenges in that the volume and complexity of data increases. Unfortunately, these challenges have not been well solved by current SV gene detection methods. Covariance-matrix-based (Svensson, Teichmann, and Stegle 2018) and clustering-based (Dries et al. 2021) methods increase computational cost dramatically with the increased number of cells or spots, which limits the usage of these methods in big data. Furthermore, gene expression modeling techniques frequently rely on Gaussian or Poisson distributions, which can result in a high rate of false positive results because not all gene expressions conform to these assumptions.

To this end, we developed Spanve (Spatial Neighborhood Variably Expressed Genes), a method for analyzing the dependence of gene expression on spatial location based on a simple but effective statistical test (Fig. 1A). Spanve does not directly consider modeling gene expression in a cell to classical distributions but rather the difference between a cell and its spatial neighbors, which is viewed as a kind of sampling. If a gene has a spatial expression pattern, there would be a great difference between space-based sampling and random sampling, which is quantified by Kullback-Leibler divergence (KL divergence). As avoiding the assumption of gene expression distributions, more precise and spatial heterogeneous SV genes can be detected. Furthermore, by considering the relationship of cells or spots, SV genes is able to be awareness of spatial domains, which facilitates the downstream clustering task. Besides, through the implementation of parallel processing and simple but effective algorithm, Spanve has been demonstrated to exhibit exceptional

computational performance. Here, we implemented two types of spatial sampling, including k-nearest-neighbor-based (Spanve-k) and Delaunay-network-based (Spanve-d). A spatial imputation method (Fig. 1B) is also implemented for clustering and a comprehensive benchmarking of the state-of-the-art (Fig. 1C) is used for evaluation.

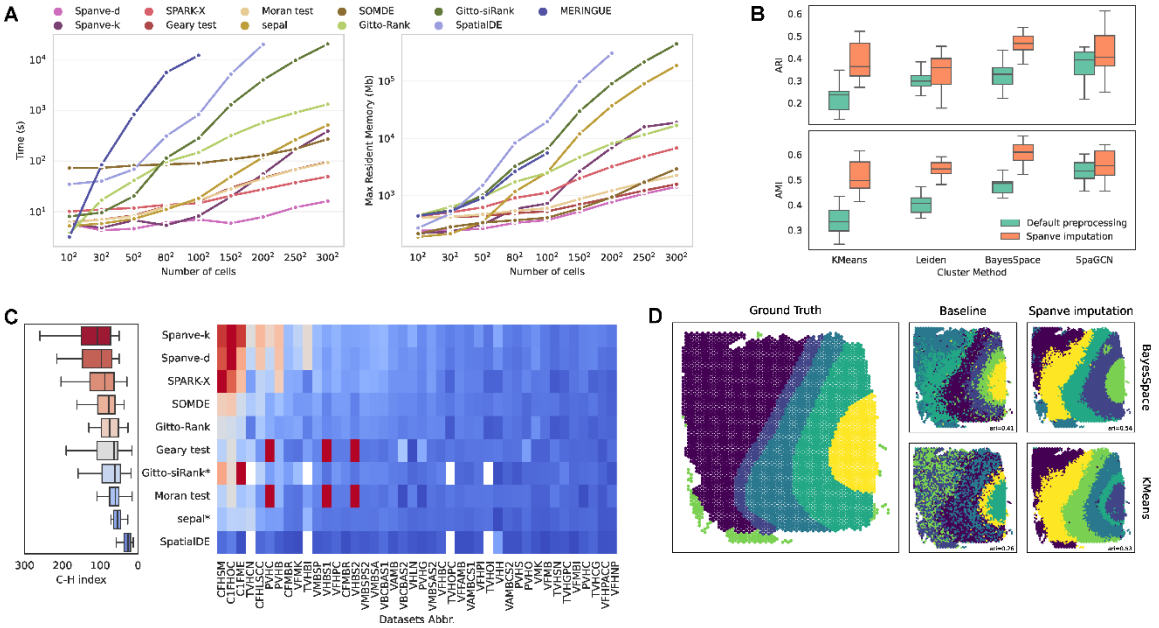


Figure 2. Benchmark results of Spanve. (a) The cost of time (left) and memory (right) in 1000 genes as the number of cells increases. (b) Clustering results of the DLPFC dataset with (orange) or without (green) Spanve imputation. The ARI and AMI rise in all four clustering techniques after Spanve imputation, particularly in statistically based approaches. The box's center and hinges symbolize the median and quartiles, respectively. (c) Performance of clustering in 46 datasets. Methods were ranked by the median of the C-H index across all datasets. The color of the box and heatmap represents the score; the closer it is to red (blue), the better (worse) the performance. The color white in the heatmap shows that the technique discovers no SV genes in the dataset. (d) An example of how Spanve imputation may help with tissue structure identification (NO.151672 in the DLPFC dataset). The ARI rose from 0.41 to 0.54 (BayesSpace) and from 0.26 to 0.53 (KMeans). Following Spanve imputation, a more distinct margin of spatial domains is observed.

As location effects may exist, including cell type distribution, local environment and cell interactions, simple distributions may not describe the gene expression, while complex distributions largely increase the number of parameters and fitting time. Without the hypothesis of gene expression distribution, Spanve is able to accurately identify genes with spatial expression patterns. As there is no gold standard for SV genes, we used two spatial autocorrelation metrics, Moran's I and Geary's C (MORAN 1950), to benchmark the statistical sensitivity of SV identification methods. Forty-six datasets with only HVGs sorted by Li et al. (Li et al. 2022) are used. The results show that Spanve achieves the best score in both metrics (Fig. 1D) and still identified a moderate number of genes. Moreover, by avoiding modeling gene distribution, Spanve reduces the complexity of the calculation and avoids subjective parameter selection. The K nearest neighbor (KNN) network contains more edges than the Delaunay network and is thus more sensitive to spatial variability but requires more computation. In our simulation with 1000 genes and an increased number of cells, Spanve-d has the highest calculation speed and lowest memory costs. Spanve-k is also low-cost when the number of cells is less than 10000, but not in data containing larger cells as edges in the KNN network grow at a polynomial rate. Spanve is able to deal with spatial transcriptome data with one hundred thousand cells in just more than ten seconds (Fig. 2A).

Compared with the existing methods, Spanve considers the complicated interaction between cells in the detection of spatial patterns, which is helpful for the next crucial downstream job of recognizing the tissue hierarchy and structure. For benchmarking the relationship of clusters and SV genes, 44 datasets from the 10X-Genomics database are used, which contain clustering labels performed by Space Ranger. The Calinski–Harabasz Index (CH index) is used as an indicator to show how well selected genes are related to cell clustering results by dispersions of data between clusters and within clusters. The higher the CH index, the more likely it is that data with SV genes as features will attain the clustering label based on raw data, implying that SV genes are more representative of raw data. As shown in Fig. 2C, we

compared the CH index of genes identified by different methods and showed the advantage of Spanve in identifying cluster-aware genes.

We further implemented a spatial imputation algorithm for clustering on the basis of detected SV genes (Fig. 1B). As SV genes reflect the location effects of the expression, it may be possible to capture the relationships between cells by scanning the expression of cell pairs. As previously described, these effects are driven by a variety of complicated factors. Consequently, imputation combining spatial locations and gene expression can better capture cellular characteristics, which may be beneficial for downstream clustering. Based on the established expectation distribution of SV gene expression, we can recognize cell pairs with an unexpected expression variation to construct a network. A strategy of graph convolution with a self-loop is performed to impute the data, which is regarded as an enhancement of location effects. To validate whether imputation is beneficial for clustering, we aggregated spatial imputation data with several typical tissue structure detection methods (Zhao et al. 2021; Hu et al. 2021). The results in DLPFC data (Maynard et al. 2021) (Fig. 2B) indicate the large improvement of clustering after Spanve imputation, and clearer boundary of tissue layers can be obtained (Fig. 2D).

In conclusion, we developed a spatially variable gene identification method called Spanve, which provides cluster-aware genes scalably. We anticipated that scalability would become more crucial as spatial transcriptomics data expand in size. Spanve requires the least amount of time and computational cost in large-scale spatial transcriptomics data when compared to current methods. Furthermore, the genes identified by Spanve are more spatially heterogeneous and favorable to further clustering. We also created a spatial imputation method that allows for popular clustering approaches identifying more accurate tissue structures. A Limitation is that it is not able to deal with data from similar tissue slices simultaneously for detecting robust SV genes and making use of prior data. As the volume of spatial transcriptomics data continues to increase, overcoming the

current limitations of our approach, the Spanve, has the potential to become even more powerful in the future.

Methods

Spanve algorithm

Spanve takes spatial transcriptomics data as inputs, which consists of two parts: the expression of genes and cell locations. As Spanve does not fit the expression of genes to any distribution, such as a Poisson distribution, expression data need to be positive integers for follow-up calculation. While preprocessing may play an important role in denoising spatial transcriptomes, a special strategy is used, where the median of data remains consistent for each gene. For N cells and M genes spatial transcriptome raw count data $X \in \mathbb{N}^{N \times M}$, and any of a gene expression $x \in \mathbb{N}^{N}$:

$$\tilde{x} = \left[F(x) \times \frac{\mathrm{Med}(x)}{\mathrm{Med}\big(F(x)\big)} \right] \in \mathbb{N}_0^+$$

where $\tilde{x}$ is preprocessed data, F is any of a preprocess function, and Med calculates the median of data. Spanve attempts to identify genes whose expression is dependent on cell location by focusing on the expression difference of cells. Here, we define the difference of any of two cells $\hat{X}$ as follows:

$$\hat{X}_{ij} = \mid X_i - X_j \mid$$

Two cell pair samplings are used in this study: random sampling and spatial sampling. It can be expected that the two sampling results differ when considering genes with space-dependent expression. For any gene g, as the expression of all cells is observed, it is easy to obtain the expected random sampling result of the distribution (expectation or null hypothesis) of cell pair expression differences by listing all combinations:

$$P_{\exp}(\hat{x} = k \mid x) = \sum_{i=k}^{\max(x)-k} P\,(i \mid x) P(i \pm k \mid x)$$

For spatial sampling, we first construct a network G_{spa} based on cell locations, for example, the K-neighbor network or Delaunay network, where nodes are cells and edges are unweighted. By calculating all cell pairs in this network, the distribution of spatial sampling $P_{obs}\left(\hat{X} \mid X; G_{spa}\right)$ can also be obtained. Here, we use KL divergence to define the distance of two distributions and obtain the p value by the G-test.

$$D_{KL(P_{obs} \| P_{\exp})} = \sum_{i=0} P_{obs}\,(i) \log \frac{P_{obs}(i)}{P_{\exp}(i)}$$

$$G = 2ND_{KL(P_{obs} \| P_{\exp})} \sim \chi^2_{\max(x)}$$

Spatial Imputation based on space-dependent genes

We further implement a spatial imputation method based on previous results similar to graph imputation. The expression of each gene of each cell pair in G_s is inspected to see if this cell pair breaks the expectation distribution under a confidence level. In addition, construct a graph G' where edge is the proportion of genes that break expectation.

$$G'_{ij} = \frac{\mid \{SF_{\exp}\left(\hat{X}_{ij}\right) \leq \alpha; g \in SV\} \mid}{\mid SV \mid}$$

$$X' = X(G' + \mathbf{I})^2$$

where SV is the spatially expressed gene set, $SF_{\exp}$ is the survival function of expectation, $\mathbf{I}$ is an identity matrix and α controls confidence.

Benchmarking of spatially variable genes

Eleven SV identification methods are used here, which can be roughly categorized into spatial autocorrelation-based (Moran test, Geary test, MERINGUE), Gaussian process regression-based (SpatialDE (Svensson, Teichmann, and Stegle 2018),

SPARK-X (Zhu, Sun, and Zhou 2021)), cluster-based (SOMDE (Hao, Hua, and Zhang 2021), Gitto-rank and Gitto-silhouette-rank (Dries et al. 2021)) and others (sepal (Andersson and Lundeberg 2021), Spanve-k and Spanve-d). For nine of them, the adjusted p value is the output, while for Gitto-silhouette-rank and speal, only the score is given. Thus, for score-only methods, the top 10% of score genes are treated as SV genes; for others, the threshold of the adjusted p value is set to 0.05. For MERINGUE, the computation cost is too high to benchmark it. First, we want to show the gene spatial heterogeneity of each method, and two metrics, Moran's I and Geary's C, are defined as follows:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \, (x_i - x)\left(x_j - x\right)}{\sum_{i=1}^{N} (x_i - x)^2}$$

$$C = \frac{(N-1) \sum_i \sum_j w_{ij} \left(x_i - x_j\right)^2}{2W \sum_i (x_i - x)^2}$$

where w_{ij} is the spatial weight and W is the sum of w_{ij}. The benchmarking is running in 44 datasets with only HVGs sorted by Li et al. (Li et al. 2022). Furthermore, genes' relationships with clustering results are scanned in silver standard datasets, which are the result of the Louvain algorithm in the decomposition of whole data. The CH index is used as the indicator for showing how easy it is for genes to obtain the cluster label. With a higher CH index, genes used as features are more related to cell types. The CH index is defined as:

$$CH = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{N-k}{k-1}$$

where tr calculates the trace of the matrixes, B_k and W_k are between-cluster and within-cluster dispersion, respectively, and k is the number of clusters. Two spatial autocorrelation metrics and their corresponding tests are implemented by squidpy (Palla et al. 2022), and the CH index is implemented by scikit-learn (Pedregosa et al. 2011).

Identification of tissue structure

In this part, four clustering methods are used to identify the layer structure of DLPFC, including K-means (distance-based), Louvain (graph-based), BayesSpace (Bayesian-based) and SpaGCN (deep-learning-based). Default parameters are used for benchmarking, and principal component analysis (PCA) embeddings used by four methods are replaced with PCA results of Spanve imputation. As K-Means clustering requires the parameter of cluster number, the elbow method based on inertia determines it. The inertia of K-means is defined as the sum of squared distances of samples to their closest cluster center. For BayesSpace, there is no such method; thus, the ground truth number of clusters is used. For SpaGCN, tissue images are not included in the analysis. The gold standard of the tissue layers is manual annotation performed by Maynard et al. (Maynard et al. 2021). Two metrics, adjusted mutual information (AMI) and adjusted Rand index (ARI), are used to evaluate the similarity between ground truth and predicted labels:

$$AMI = \frac{MI - E[MI]}{\text{mean}\big(H(U), H(V)\big) - E[MI]}$$

where MI is mutual information and H calculates the entropy of two label assignments U and V.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

where RI is the Rand index. PCA, K-means, AMI and ARI are implemented by scikit-learn (Pedregosa et al. 2011). Louvain is implemented by the scanpy package (Wolf, Angerer, and Theis 2018). SpaGCN was implemented by Hu et al. (Hu et al. 2021).

Declarations

Data availability

Datasets from 10X genomics datasets can be downloaded from https://www.10xgenomics.com/resources/datasets. Sorted datasets are available from https://github.com/QuKunLab/SpatialBenchmarking. DLPFC datasets are extracted from https://data.humancellatlas.org/explore/projects/7b393e4d-65bc-4c03-b402-aae769299329/.

Code availability

All source codes in this study are available from https://github.com/zjupgx/Spanve. Spanve is also available as a Python package from https://pypi.org/project/Spanve/.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 31971371), and the Zhejiang Provincial Natural Science Foundation of China (Grant No. LDT23H19011H19). We thank the Information Technology Center and State Key Lab of CAD&CG, the Innovation Institute for Artificial Intelligence in Medicine, Zhejiang University, and Alibaba Cloud for the support of computing resources.

Contributions

G.X.C., X.G., and Z.Z. designed the study; G.X.C. implemented the method; G.X.C. and Y.C.C. analyzed the data; G.X.C. and Z.Z. wrote the paper.

References

Andersson, Alma, and Joakim Lundeberg. 2021. "Sepal: Identifying Transcript Profiles with Spatial Patterns by Diffusion-Based Modeling." *Bioinformatics* 37 (17): 2644–50. https://doi.org/10.1093/bioinformatics/btab164.

Chen, Ao, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, et al. 2022. "Spatiotemporal Transcriptomic Atlas of Mouse Organogenesis Using DNA Nanoball-Patterned Arrays." *Cell* 185 (10): 1777-1792.e21. https://doi.org/10.1016/j.cell.2022.04.003.

Dries, Ruben, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, et al. 2021. "Giotto: A Toolbox for Integrative Analysis and Visualization of Spatial Expression Data." *Genome Biology* 22 (1): 78. https://doi.org/10.1186/s13059-021-02286-2.

Hao, Minsheng, Kui Hua, and Xuegong Zhang. 2021. "SOMDE: A Scalable Method for Identifying Spatially Variable Genes with Self-Organizing Map." *Bioinformatics* 37 (23): 4392–98. https://doi.org/10.1093/bioinformatics/btab471.

Hu, Jian, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. 2021. "SpaGCN: Integrating Gene Expression, Spatial Location and Histology to Identify Spatial Domains and Spatially Variable Genes by Graph Convolutional Network." *Nature Methods* 18 (11): 1342–51. https://doi.org/10.1038/s41592-021-01255-8.

Li, Bin, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, et al. 2022. "Benchmarking Spatial and Single-Cell Transcriptomics Integration Methods for Transcript Distribution Prediction and Cell Type Deconvolution." *Nature Methods* 19 (6): 662–70. https://doi.org/10.1038/s41592-022-01480-9.

Liao, Jie, Xiaoyan Lu, Xin Shao, Ling Zhu, and Xiaohui Fan. 2021. "Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics." *Trends in Biotechnology* 39 (1): 43–58. https://doi.org/10.1016/j.tibtech.2020.05.006.

Longo, Sophia K., Margaret G. Guo, Andrew L. Ji, and Paul A. Khavari. 2021. "Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics." *Nature Reviews Genetics* 22 (10): 627–44. https://doi.org/10.1038/s41576-021-00370-8.

Maynard, Kristen R., Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, et al. 2021. "Transcriptome-Scale Spatial Gene Expression in the Human Dorsolateral Prefrontal Cortex." *Nature Neuroscience* 24 (3): 425–36. https://doi.org/10.1038/s41593-020-00787-0.

MORAN, P. A. P. 1950. "NOTES ON CONTINUOUS STOCHASTIC PHENOMENA." *Biometrika* 37 (1–2): 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

Moses, Lambda, and Lior Pachter. 2022. "Museum of Spatial Transcriptomics." *Nature Methods* 19 (5): 534–46. https://doi.org/10.1038/s41592-022-01409-2.

Palla, Giovanni, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, et al. 2022. "Squidpy: A Scalable Framework for Spatial Omics Analysis." *Nature Methods* 19 (2): 171–78. https://doi.org/10.1038/s41592-021-01358-2.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, October. https://hal.inria.fr/hal-00650905.

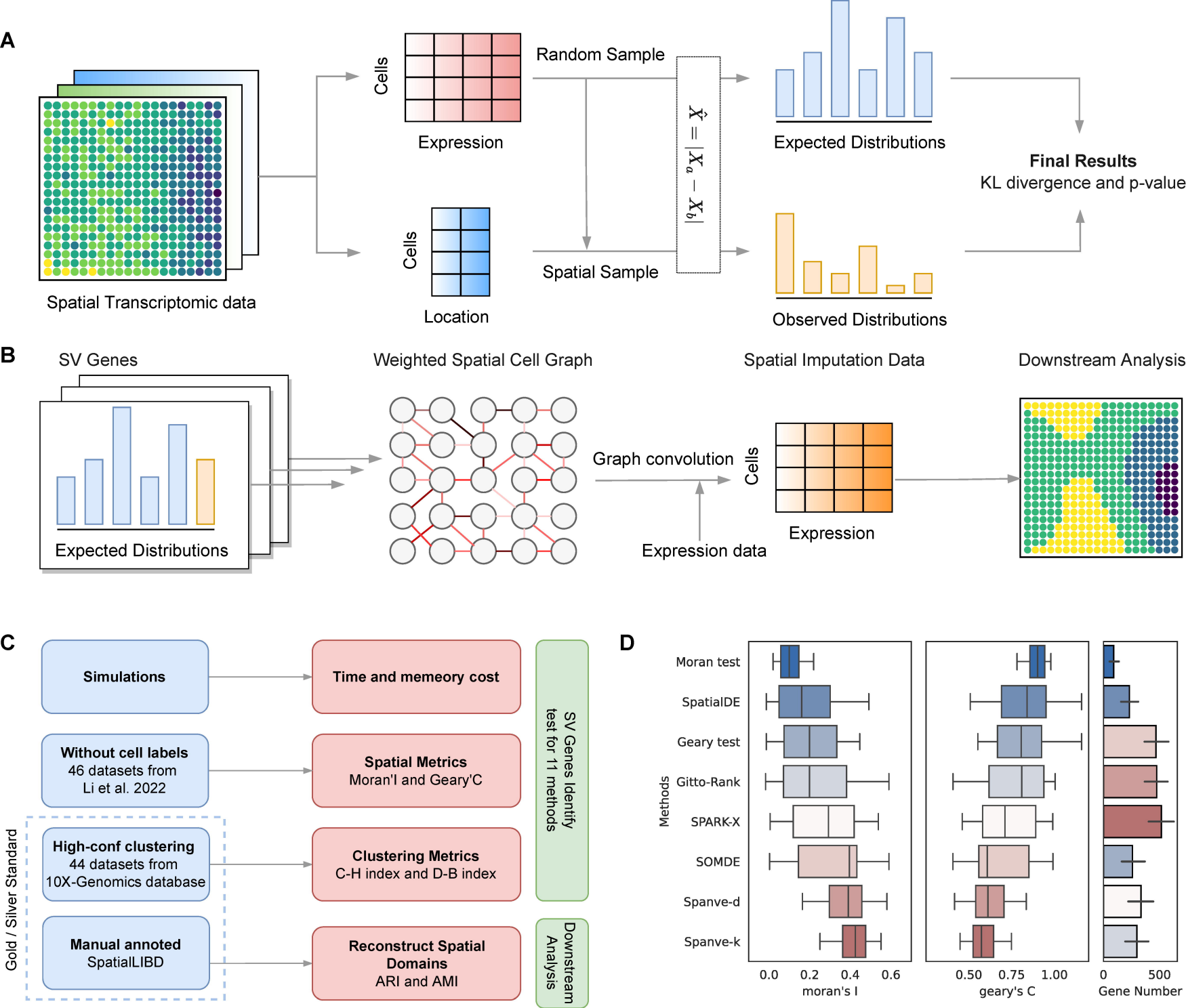
Rao, Anjali, Dalia Barkley, Gustavo S. França, and Itai Yanai. 2021. "Exploring Tissue Architecture Using Spatial Transcriptomics." *Nature* 596 (7871): 211–20. https://doi.org/10.1038/s41586-021-03634-9.

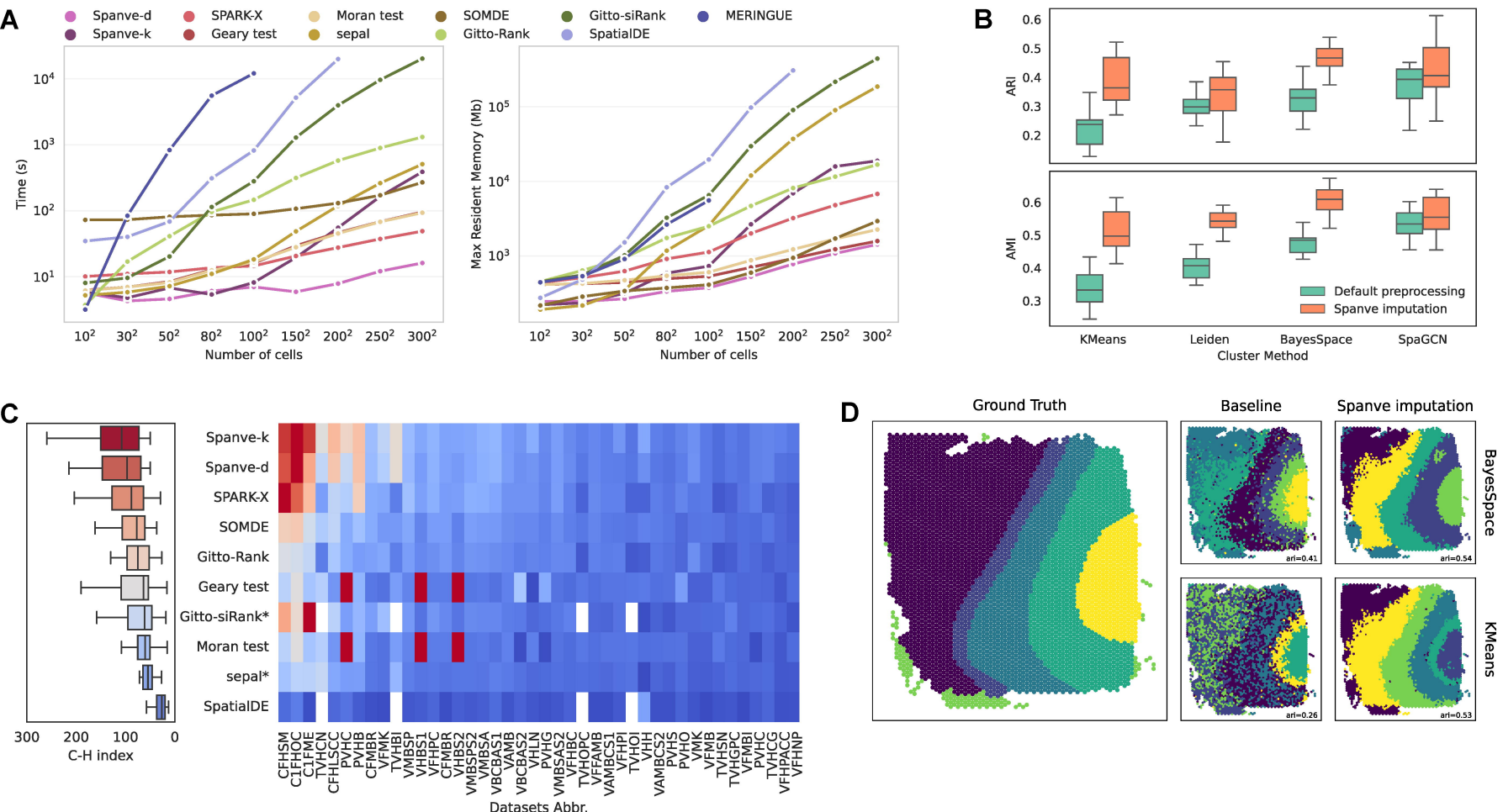
Svensson, Valentine, Sarah A. Teichmann, and Oliver Stegle. 2018. "SpatialDE: Identification of Spatially Variable Genes." *Nature Methods* 15 (5): 343–46. https://doi.org/10.1038/nmeth.4636.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1): 15. https://doi.org/10.1186/s13059-017-1382-0.

Zhao, Edward, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, et al. 2021. "Spatial Transcriptomics at Subspot Resolution with BayesSpace." *Nature Biotechnology* 39 (11): 1375–84. https://doi.org/10.1038/s41587-021-00935-2.

Zhu, Jiaqiang, Shiquan Sun, and Xiang Zhou. 2021. "SPARK-X: Non-Parametric Modeling Enables Scalable and Robust Detection of Spatial Expression Patterns for Large Spatial Transcriptomic Studies." *Genome Biology* 22 (1): 1–25. https://doi.org/10.1186/s13059-021-02404-0.


A
Spatial Transcriptomic data
Cells
Expression
Cells
Location
Random Sample
Spatial Sample
$|\hat{X}_a - X_b| = \hat{X}$
Expected Distributions
Observed Distributions
Final Results
KL divergence and p-value
B
SV Genes
Expected Distributions
Weighted Spatial Cell Graph
Graph convolution
Expression data
Spatial Imputation Data
Cells
Expression
Downstream Analysis
C
Simulations
Time and memeory cost
Without cell labels
46 datasets from
Li et al. 2022
Spatial Metrics
Moran'I and Geary'C
High-conf clustering
44 datasets from
10X-Genomics database
Clustering Metrics
C-H index and D-B index
Manual annoted
SpatialLIBD
Reconstruct Spatial Domains
ARI and AMI
Gold / Silver Standard
SV Genes Identify test for 11 methods
Downstream Analysis
D
Methods
Moran test
SpatialDE
Geary test
Gitto-Rank
SPARK-X
SOMDE
Spanve-d
Spanve-k
0.0
0.2
0.4
0.6
moran's I
0.50
0.75
1.00
geary's C
0
500
Gene Number


A
Spanve-d
Spanve-k
SPARK-X
Geary test
Moran test
sepal
SOMDE
Gitto-Rank
Gitto-siRank
SpatialDE
MERINGUE
Time (s)
Number of cells
Max Resident Memory (Mb)
Number of cells
B
ARI
AMI
Default preprocessing
Spanve imputation
KMeans
Leiden
BayesSpace
SpaGCN
Cluster Method
C
Spanve-k
Spanve-d
SPARK-X
SOMDE
Gitto-Rank
Geary test
Gitto-siRank*
Moran test
sepal*
SpatialDE
C-H index
Datasets Abbr.
D
Ground Truth
Baseline
Spanve imputation
BayesSpace
KMeans
ari=0.41
ari=0.54
ari=0.26
ari=0.53