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

Guoxin Cai[1], Yichang Chen[1], Shuqing Chen[1], Xun Gu[2], Zhan Zhou[1,3,4,5*]

1 Institute of Drug Metabolism and Pharmaceutical Analysis 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 Zhejiang University Innovation Institute for Artificial Intelligence in Medicine, Hangzhou 310018, China.

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

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

* Corresponding author: Zhan Zhou.

Email: zhanzhou@zju.edu.cn

Author Contributions: G.X.C., X.G., and Z.Z. designed the study; G.X.C. implemented the method; G.X.C., Y.C.C. and S.Q.C. analyzed the data; G.X.C. and Z.Z. wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Keywords: spatial transcriptomics, single-cell sequencing, statistical model, spatially variable genes, clustering, data imputation, large-scale data

Abstract

The depiction of in situ gene expression through spatial transcriptomics facilitates the inference of cell function mechanisms. To build spatial maps of transcriptomes, the first and crucial step is to identify spatially variable (SV) genes. However, current methods fall short in dealing with large-scale spatial transcriptomics data and may result in a high false positive rate due to the modeling of gene expression into parametric distributions. This paper introduces Spanve (https://github.com/zjupgx/Spanve), a non-parametric statistical approach based on modeling space dependence as a distance of two distributions for detecting SV genes. The high computing efficiency and accuracy of Spanve is demonstrated through comprehensive benchmarking. Additionally, Spanve can detect clustering-friendly SV genes and spatially variable co-expression, facilitating the identification of spatial tissue domains by an imputation.

Introduction

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 (1, 2). 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. 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 (1) and are thus informative for multiple downstream analyses, such as reconstructing tissue structure (3). SV genes are also helpful for providing biological insight, as different spatial expression patterns reflect different gene functions.

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 (4)), it is possible to locate cellular or subcellular

transcriptomes (5). However, the progress of new technology raises new challenges in that the volume and complexity of data increases, which have not been well solved by current SV gene detection methods. Covariance-matrix-based (6) and clustering-based (7) methods increase computational cost dramatically with the increased number of cells or spots, limiting 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, especially genes with high spatial heterogeneity (Fig. 1D).

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 parametric 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). Spanve detects more precise and spatially heterogeneous SV genes by avoiding the assumption of gene expression distributions. It also considers the relationship between cells or spots, which facilitates downstream clustering. Additionally, Spanve exhibits exceptional computational performance due to its implementation of parallel processing and a simple yet effective algorithm. 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.

Results

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 (8), to benchmark the statistical sensitivity of SV identification methods. 46 datasets with only HVGs sorted by Li et al. (9) are used. The results show that Spanve achieves the best score in both metrics (Fig. 2A) 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. 2B). 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 SV genes reflect clustering labels, 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 clustering-friendly 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. 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, 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 (10, 11). The results in DLPFC data (12) (Fig. 2E) indicate the large improvement of clustering after Spanve imputation, and clearer boundary of tissue layers can be obtained (Fig. 2D). A similar strategy can be used into spatially variable co-expression (SVC) by converting correlation into individual inter correlation. By doing so, Spanve is able to identify spatial co-expression (Fig. 2F), which might be helpful for identifying tissue structure. In the osmFISH data, ARI increased from 0.28 (SV genes imputation) to 0.35 (combined SV and SVC imputation).

Discussion

In conclusion, we developed a spatially variable gene identification method called Spanve, which provides clustering-friendly genes scalably. We anticipated that scalability would become more crucial as spatial transcriptomics data expand in volume. 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.

Materials and 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 $x \in \mathbb{N}^{N}$:

$$\tilde{x} = \left[F(x) \times \frac{\mathrm{Med}(x)}{\mathrm{Med}\big(F(x)\big)}\right] \in \mathbb{N}_0^+ \tag{1}$$

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:

$$\hat{X}_{ij} = \mid X_i - X_j \mid \tag{2}$$

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:

$$P_{\exp}(\hat{x} = k \mid x) = \sum_{i=k}^{\max(x)-k} P\,(i \mid x)P(i \pm k \mid x) \tag{3}$$

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 $P_{obs}\big(\hat{X} \mid X; G_{spa}\big)$ 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.

$$D_{KL(P_{obs}\|P_{\exp})} = \sum_{i=0} P_{obs}\,(i)\log\frac{P_{obs}(i)}{P_{\exp}(i)} \tag{4}$$

$$G = 2ND_{KL(P_{obs}\|P_{\exp})} \sim \chi^2_{\max(x)} \tag{5}$$

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.

$$G'_{ij} = \frac{\mid SF_{\exp}(\hat{X}_{ij}) \leq \alpha; g \in SV \mid}{\mid SV \mid} \tag{6}$$

$$X' = X(G' + \mathbf{I})^2 \tag{7}$$

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 (6), SPARK-X (13)), clustering-based (SOMDE (14), Gitto-rank and Gitto-silhouette-rank (7)) and others (sepal (15), 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 with two metrics Moran's I and Geary's C. The benchmarking is running in 44 datasets with only HVGs sorted by Li et al. (9). 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.

Spatial co-expression

For any two gene expressions x and y, let

$$C_i^{xy} = \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \tag{8}$$

where σ is the standard variance of a gene, $\bar{x}$ and $\bar{y}$ are the mean of the expression. Then the mean of C is equal to the Pearson's correlation coefficient ρ.

$$\bar{C}^{xy} = \frac{1/N \times \sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})}{\sigma_x \sigma_y} \equiv \rho_{x,y} \tag{9}$$

Thus, C can be seen as the sample resolution co-expression strength of the two genes. In this way, a similar strategy can be used to explore the gene pair with spatially variable co-expression.

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. (12).

Availability of data and materials

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 downloaded by a script based on BayesSpace. osmFISH mouse secondary somatosensory cortex dataset comes from http://linnarssonlab.org/osmFISH/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/.

Acknowledgments

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.

References

1. L. Moses, L. Pachter, Museum of spatial transcriptomics. *Nat Methods* **19**, 534–546 (2022).
2. S. K. Longo, M. G. Guo, A. L. Ji, P. A. Khavari, Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* **22**, 627–644 (2021).
3. A. Rao, D. Barkley, G. S. França, I. Yanai, Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
4. A. Chen, *et al.*, Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
5. J. Liao, X. Lu, X. Shao, L. Zhu, X. Fan, Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. *Trends in Biotechnology* **39**, 43–58 (2021).
6. V. Svensson, S. A. Teichmann, O. Stegle, SpatialDE: identification of spatially variable genes. *Nat Methods* **15**, 343–346 (2018).
7. R. Dries, *et al.*, Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* **22**, 78 (2021).
8. P. A. P. MORAN, NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika* **37**, 17–23 (1950).
9. B. Li, *et al.*, Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* **19**, 662–670 (2022).
10. E. Zhao, *et al.*, Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375–1384 (2021).

231 11. J. Hu, *et al.*, SpaGCN: Integrating gene expression, spatial location and histology to
232 identify spatial domains and spatially variable genes by graph convolutional network.
233 *Nat Methods* **18**, 1342–1351 (2021).
234 12. K. R. Maynard, *et al.*, Transcriptome-scale spatial gene expression in the human
235 dorsolateral prefrontal cortex. *Nat Neurosci* **24**, 425–436 (2021).
236 13. J. Zhu, S. Sun, X. Zhou, SPARK-X: non-parametric modeling enables scalable and
237 robust detection of spatial expression patterns for large spatial transcriptomic studies.
238 *Genome Biol* **22**, 1–25 (2021).
239 14. M. Hao, K. Hua, X. Zhang, SOMDE: a scalable method for identifying spatially
240 variable genes with self-organizing map. *Bioinformatics* **37**, 4392–4398 (2021).
241 15. A. Andersson, J. Lundeberg, sepal: identifying transcript profiles with spatial patterns
242 by diffusion-based modeling. *Bioinformatics* **37**, 2644–2650 (2021).
243

Figures

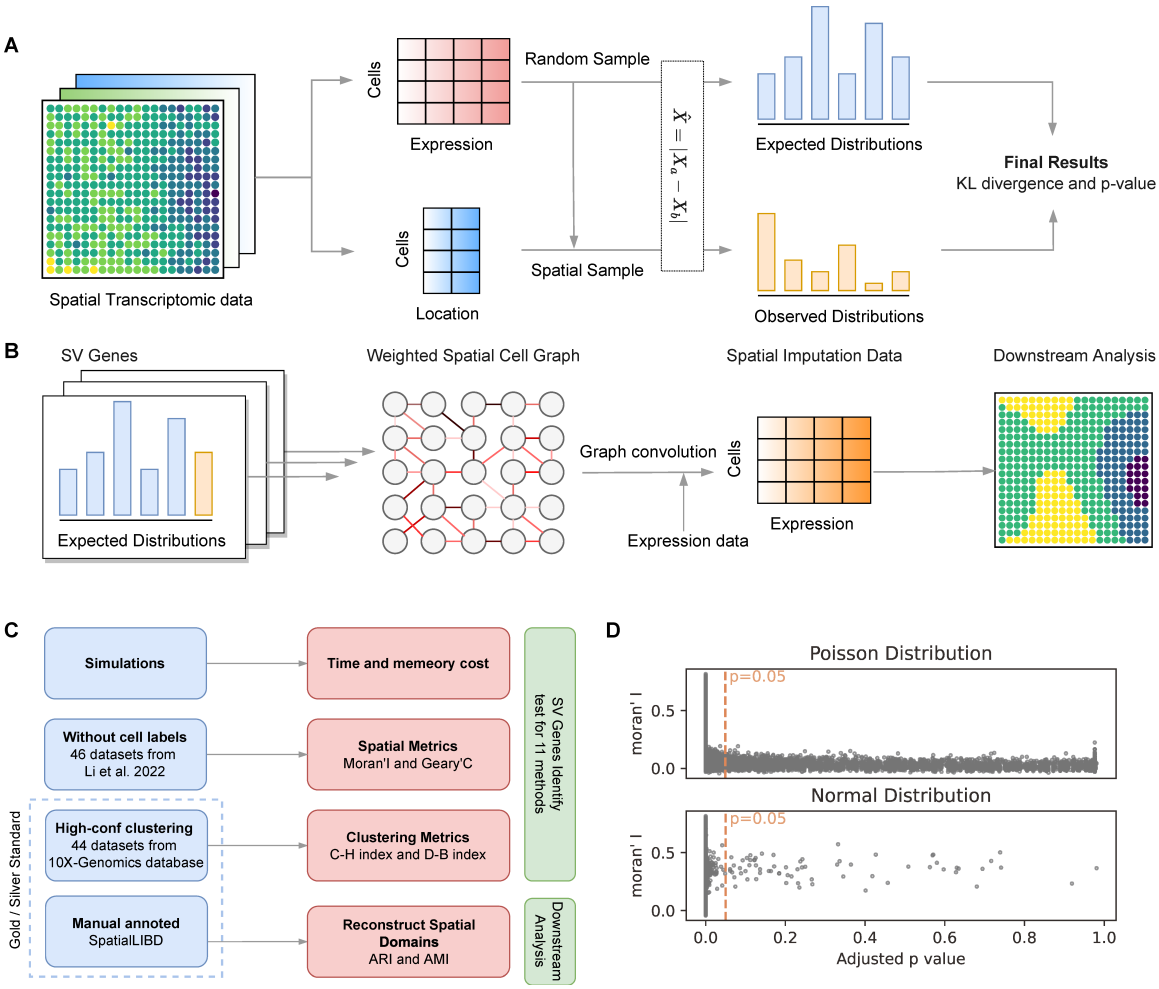

Figure 1. The framework of Spanve and experiments design in this study. (a) In spatial transcriptomic data, Spanve considers two types of sampling of cell or spot pairs: random sampling as null hypothesis and location based sampling as alternative hypothesis. Spatial dependency of genes is defined as 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-friendly are all considered while detecting SV genes. Reconstruction of spatial domains is an additional examination. (d) Most of the genes with high spatial variance do not conform to Poisson or Gaussian distributions. x-axis represents the Benjamini-Hochberg adjusted p-value and y-axis represents Moran's I. The dashed lines highlight when p-value is 0.05. The result is based on dataset Visium Mouse Olfactory Bulb from 10X Genomics.

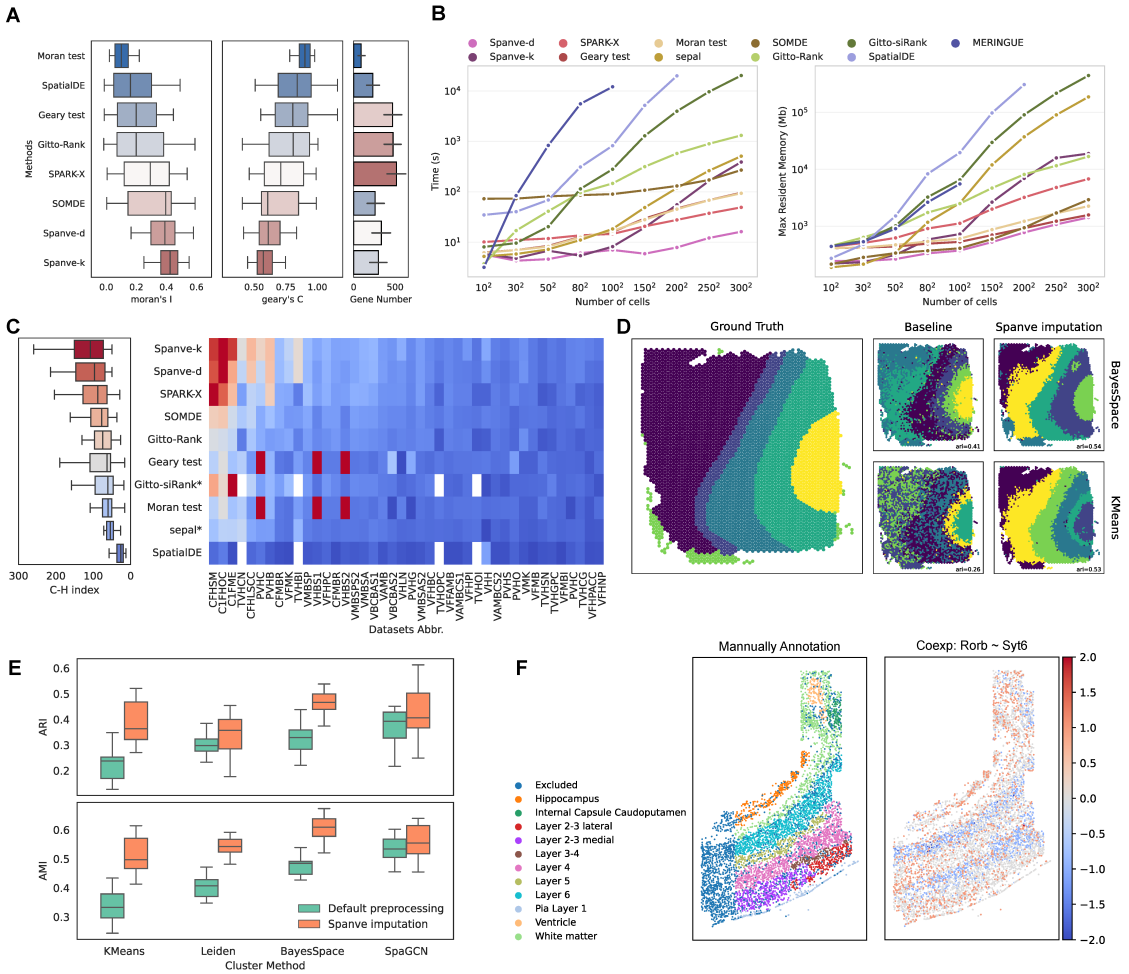

Figure 2. Benchmark results of Spanve. (a) Spatial heterogeneity of SV genes detected by methods. Elements in box plots in this study are defined as followed: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; color, performance. Error bar in bar plot represents 95% confidence interval. (b) The cost of time (left) and memory (right) in 1000 genes as the number of cells increases. (c) Clustering metric (C-H index) of SV genes 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 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). (e) Clustering results of the DLPFC dataset (n=12) with (orange) or without (green) Spanve imputation. (f) Mannually annotation (left) and a example of spatial co-expression (right) of osmFISH mouse secondary somatosensory cortex.

Supporting Information for

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

Guoxin Cai[1], Yichang Chen[1], Shuqing Chen[1], Xun Gu[2], Zhan Zhou[1,3,4,5*]

* Corresponding author: Zhan Zhou.

Email: zhanzhou@zju.edu.cn

Supplementary Methods

Goodness-of-fit test

To test whether gene expression can be modeled by Poisson distribution, the Poisson dispersion test (1) is performed. For the expression of a gene x,

$$D = \sum_{i=1}^{N} \frac{(x - \bar{x})^2}{\bar{x}} \sim \chi^2_{N-1} \quad [1]$$

where χ^2_{N-1} is a chi-squared distribution with degrees of freedom $N - 1$ and $\bar{x}$ is the mean of x. The omnibus normality test is performed on preprocessed data to test the null hypothesis that gene expression comes from a Gaussian distribution. The test is based on D'Agostino and Pearson's test (2, 3) and is implemented by the scipy python package.

Spatial autocorrelation Metrics

Moran's I and Geary's C, are defined as follows:

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

$$C = \frac{(N-1) \sum_i \sum_j w_{ij} \left(x_i - x_j\right)^2}{2W \sum_i (x_i - \bar{x})^2} \quad [3]$$

where w_{ij} is the spatial weight and W is the sum of w_{ij}. Two spatial autocorrelation metrics and their corresponding tests are implemented by squidpy (4).

Clustering metrics

The CH index is defined as:

$$CH = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{N - k}{k - 1} \quad [4]$$

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. The CH index is implemented by scikit-learn (5).

Two metrics, adjusted mutual information (AMI) and adjusted Rand index (ARI), are used to evaluate the similarity between ground truth and predicted labels.

$$AMI = \frac{MI - E[MI]}{\text{mean}\big(H(U), H(V)\big) - E[MI]} \quad [5]$$

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

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad [6]$$

where RI is the Rand index.

Implementation. PCA, K-means, AMI and ARI are implemented by scikit-learn (5). Louvain is implemented by the scanpy package (6). SpaGCN was implemented by Hu et al. (7).

References

1. D. Karlis, E. Xekalaki, A Simulation Comparison of Several Procedures for Testing the Poisson Assumption. *J Royal Statistical Soc D* **49**, 355–382 (2000).
2. R. B. D'Agostino, An omnibus test of normality for moderate and large size samples. *Biometrika* **58**, 341–348 (1971).
3. R. D'Agostino, E. S. Pearson, Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b}^1$. *Biometrika* **60**, 613–622 (1973).
4. G. Palla, *et al.*, Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* **19**, 171–178 (2022).
5. F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011) (April 7, 2022).
6. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
7. J. Hu, *et al.*, SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* **18**, 1342–1351 (2021).