cells

MDPI

Article

SINFONIA: Scalable Identification of Spatially Variable Genes for Deciphering Spatial Domains

Rui Jiang [1], Zhen Li [1], Yuhang Jia [2], Siyu Li [2] and Shengquan Chen [3,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, BNRIST/Department of Automation, Tsinghua University, Beijing 100084, China
[2] School of Statistics and Data Science, Nankai University, Tianjin 300071, China
[3] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
* Correspondence: chenshengquan@nankai.edu.cn

Abstract: Recent advances in spatial transcriptomics have revolutionized the understanding of tissue organization. The identification of spatially variable genes (SVGs) is an essential step for downstream spatial domain characterization. Although several methods have been proposed for identifying SVGs, inadequate ability to decipher spatial domains, poor efficiency, and insufficient interoperability with existing standard analysis workflows still impede the applications of these methods. Here we propose SINFONIA, a scalable method for identifying spatially variable genes via ensemble strategies. Implemented in Python, SINFONIA can be seamlessly integrated into existing analysis workflows. Using 15 spatial transcriptomic datasets generated with different protocols and with different sizes, dimensions and qualities, we show the advantage of SINFONIA over three baseline methods and two variants via systematic evaluation of spatial clustering, domain resolution, latent representation, spatial visualization, and computational efficiency with 21 quantitative metrics. Additionally, SINFONIA is robust relative to the choice of the number of SVGs. We anticipate SINFONIA will facilitate the analysis of spatial transcriptomics.

Keywords: spatial transcriptomics; spatial autocorrelation; spatially variable genes; spatial domains

check for updates

Citation: Jiang, R.; Li, Z.; Jia, Y.; Li, S.; Chen, S. SINFONIA: Scalable Identification of Spatially Variable Genes for Deciphering Spatial Domains. *Cells* **2023**, *12*, 604. https://doi.org/10.3390/cells12040604

Academic Editor: Sang Woo Kim

Received: 4 January 2023
Revised: 29 January 2023
Accepted: 7 February 2023
Published: 13 February 2023

1. Introduction

Spatially resolved transcriptomics grants us a unique perspective on coherent spatial and gene expression patterns and hence allows for insights into the molecular organization of tissues. Advanced technologies for spatial transcriptomics (STs) have enabled genome-wide profiling of expression levels, demanding scalable methods that take advantage of spatial context to facilitate the identification of spatially variable genes (SVGs), which is regarded as the first critical step in ST data analysis [1]. We note that the task of SVG identification, an essential step before spatial domain characterization, is different from that of spatial pattern visualization and detected-gene evaluation, which are performed after spatial domain characterization, and that the former aims to select individual genes and visualize the expression of the gene in different spatial domains [2,3], while the latter aims to evaluate the differentially expressed genes between different spatial domains [4,5].

Li et al. have provided a survey of computational methods for SVG detection, including eight tools implemented in R and five tools implemented in Python [6]. In recent years, many researchers have preferred to analyze data in Python due to its attractive syntax and highly optimized scientific computing libraries for machine learning [7]. Additionally, in the single-cell community, a number of advanced frameworks have been implemented in Python, such as Scanpy [8], scvi-tools [9] and Squidy [3], suggesting the wide acceptance of Python in single-cell studies. In this study, we focus on Python-based methods and aim to provide an effective and scalable method for the Python community.

Among the tools implemented in Python, SPADE integrates high-level image features and spatial transcriptomes to identify genes that show rich patterns in morphological context [10]. SpatialDE decomposes the variation of each gene into spatial and nonspatial components and identifies SVGs through the comparison between the full model and a model without spatial random effect term [1]. GPcounts is built on a Gaussian process regression model and utilizes a negative binomial likelihood function to model the temporal and spatial variations in gene expression profiles [11]. SOMDE first clusters the neighboring cells into separated nodes based on the self-organizing map (SOM), and then fits the node-level gene expression variation with a Gaussian process to identify SVGs [12]. HMRF uses hidden Markov random fields to identify spatial domains based on cross-platform cell-type mapping [13].

However, these methods are still impeded by relatively low effectiveness and scalability. Additionally, considering that many newly developed methods for STs (e.g., STAGATE [14], stPlus [15], Squidpy [3]) are built on the AnnData format and can be seamlessly integrated into the SCANPY workflow [8], new tools should provide interoperability with AnnData and SCANPY. In addition, the performance of the identified SVGs for deciphering spatial domains still lacks systematic, and especially quantitative, evaluation using the ST datasets with a gold standard.

Spatial autocorrelation statistics, such as Moran's I and Geary's C, have been adopted to select individual genes for spatial pattern visualization (e.g., Seurat [2] and Squidpy [3]), and to quantify the existence of spatial patterns of detected genes (e.g., SpaGE [4] and SpaGCN [5]). However, spatial autocorrelation statistics have not been used to identify SVGs for the characterization of ST data (e.g., dimension reduction) and downstream analyses of ST data (e.g., spatial domain identification). Additionally, the low computational efficiency of existing implementations for calculating the statistics limits their scalability to high-throughput data.

To fill these gaps, we propose a method named SINFONIA to identify spatially variable genes for deciphering spatial domains. Specifically, we construct a spatial neighbor graph and, for the first time, apply spatial autocorrelation statistics to identify SVGs for deciphering spatial domains. We propose an ensemble strategy based on spatial autocorrelation statistics to better identify SVGs and achieve about a four- to two-hundred-and-sixty-two-times speedup compared with other spatial autocorrelation statistics implementations. Additionally, it is the first time that a systematic and quantitative evaluation of SVG identification performance has been conducted using a number of ST datasets.

2. Materials and Methods

2.1. Construction of Spatial Neighbor Graph

As the SINFONIA framework shown in Figure 1, SINFONIA identifies SVGs based on Moran's I and Geary's C statistics. To calculate the spatial autocorrelation statistics, we construct a spatial neighbor graph (SNG), considering that aggregating information from each spot's neighbors can improve the characterization of spatial patterns and, consequently, SVGs [16]. Specifically, for each spot we use spatial coordinates to identify k nearest neighbors in a Euclidean space. The weights $\mathbf{W}$ of SNG are then defined by (1).

$$w_{ij} = \begin{cases} 1 - \frac{D_{ij}}{\max(\mathbf{D}_{i.})} & if\ spots\ i\ and\ j\ are\ neighbors \\ 0 & otherwise \end{cases}, \tag{1}$$

where w_{ij} denotes the weight between spots i and j, D_{ij} denotes the distance between spots i and j, and $\max(\mathbf{D}_{i.})$ denotes the maximum distance between spot i and its nearest neighbors. We implemented the calculation via sparse matrix operations, which decrease the space complexity from $O(n^2)$ to $O(n)$ $(n \gg k)$, and thus enable memory-efficient and high-speed calculation for the large-scale spatial transcriptomic (ST) data.

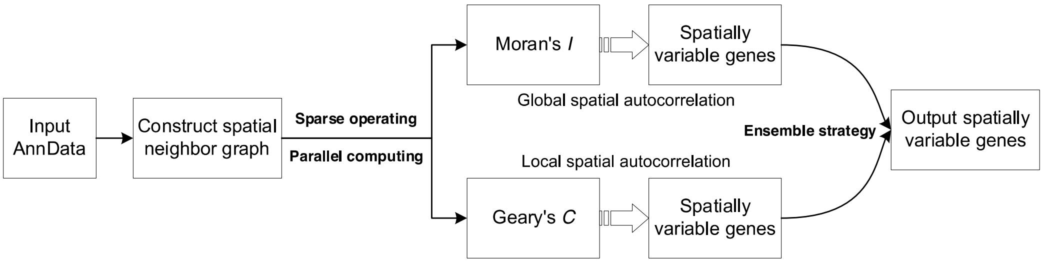

Figure 1. The framework of SINFONIA.

2.2. Calculation of Spatial Autocorrelation Statistics

We then compute Moran's *I* and Geary's *C* for each gene based on **W**, the former of which characterize spatial autocorrelation in the given gene among neighbors in the SNG. In ST data, the gene expression in spots from localized spatial neighborhoods tends to be closer in value than those at distant locations. For example, the expression levels of a gene at nearby locations may be closer in value than expression levels at distant locations. Therefore, genes with stronger spatial autocorrelation can exhibit more organized spatial expression patterns. Here, we calculate spatial autocorrelation statistics based on our constructed SNG to quantify the spatial autocorrelation degree of each gene.

Specifically, Moran's *I* measures the overall spatial autocorrelation of a dataset [17]. Given a gene and SNG of the spots, Moran's *I* evaluates whether the spatial expression pattern is clustered, dispersed, or random. The Moran's *I* score ranges from −1 to 1, where a score close to 1 indicates a clear spatial pattern, a score close to 0 indicates random spatial expression, and a score close to −1 indicates a chess board-like pattern. For each gene, the Moran's *I* score is computed as (2):

$$I = \frac{N}{W}\frac{\sum_i \sum_j \left[w_{ij}(x_i - \overline{x})(x_j - \overline{x})\right]}{\sum_i (x_i - \overline{x})^2}, \tag{2}$$

where x_i and x_j denote the gene expression levels of spots i and j, respectively, $\overline{x}$ is the mean expression level of the gene, N is the total number of spots, w_{ij} is the edge weight between spots i and j in SNG, and W is the sum of SNG weights.

Geary's *C*, another commonly used statistic [18], is inversely related to Moran's *I*, but is not identical to Moran's *I*. Compared with Moran's *I*, which measures global spatial autocorrelation, Geary's *C* is more sensitive to local spatial autocorrelation. For each gene, the Geary's *C* score is computed as (3):

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

The score of Geary's *C* ranges from 0 to 2. To make it on the same scale as that of Moran's *I*, we rescale it by (4):

$$C^* = 1 - C, \tag{4}$$

where a C^* close to 1 means positive spatial autocorrelation, a C^* close to 0 means poor spatial autocorrelation, and a C^* close to −1 means negative spatial autocorrelation.

2.3. Identification of Spatially Variable Genes

SINFONIA identifies SVGs for spatial domain detection based on an ensemble strategy for Moran's *I* and Geary's *C*. Given a specified number *n* of SVGs, SINFONIA_*I*, a variant of SINFONIA, selects the top *n* genes with the highest Moran's *I* scores as SVGs, while SINFONIA_*C*, another variant of SINFONIA, selects the top *n* genes with the highest rescaled Geary's *C* scores as SVGs. In this work, following the SCANPY workflow for ST data, we identified the top 2000 SVGs with SINFONIA_*I*, SINFONIA_*C* and other baseline methods by default on all datasets. SINFONIA adopts the union of SVGs identified by SINFONIA_*I* and SINFONIA_*C* as SVGs, based on an ensemble idea that integrates both global and local spatial autocorrelation. The ensemble strategy provides overall better performance than the two variants and is more robust than the variants, which fluctuate using different downstream clustering methods.

2.4. Implementation and Usage of SINFONIA

With the increase of data throughput and data scale, computational methods should meet the critical principles of efficiency and scalability [14]. To achieve superior efficiency and scalability, we implemented the calculation of Moran's *I* and Geary's *C* via sparse matrix operations and implemented parallel computing for genes. Specifically, SINFONIA performs sparse matrix operations using SciPy 2-D sparse array grammar for numeric data [19], which focuses on the calculation of only non-zero values, and thus enables memory-efficient and high-speed calculation for large-scale data. Additionally, SINFONIA adopts Numba (https://numba.pydata.org/ accessed on 17 July 2022) to translate functions to optimized fast machine code and achieve parallel computing. In addition, with tailored programming, SINFONIA computes Moran's *I* and Geary's *C* simultaneously with almost no extra time consumption compared to computing a single statistic.

The AnnData format offers a convenient way to store data matrices and annotations together. The use of AnnData format also facilitates interoperability with existing single-cell analysis tools such as SCANPY [8]. We designed the structure of the SINFONIA class to smoothly interface with the widely-used AnnData format. Therefore, SINFONIA can be seamlessly integrated into the SCANPY vignette by taking the AnnData object as input, and replace the default SVG identification method via just a single line of code. In addition, its modularity makes it efficient for data representation and storage, and makes it flexible to be interfaced with various external tools in the Python data science and machine-learning ecosystem, such as advanced deep-learning frameworks. Nevertheless, existing SVG identification methods in the Python community, such as SpatialDE [1] and SOMDE [12], have poor interoperability with the widely-used AnnData format and SCANPY workflow, and thus hamper the usage and application of these methods.

Additionally, SINFONIA provides rich documentation in the form of functional application programming interface documentation, tutorials and example workflows. It is easy to navigate and is accessible to both experienced developers and beginner analysts.

2.5. Data Collection

We systematically assessed the performance of SINFONIA using 15 ST datasets generated with different protocols, and with different sizes, dimensions and qualities. We note that limited sequencing coverage of single-cell technologies, especially for STs, results in large fractions of observed zeros, namely a high degree of data sparsity, providing data with low quality [20].

First, we collected a 10X Genomics Visium dataset containing spatial expressions of 12 human dorsolateral prefrontal cortex (DLPFC) sections from spatialLIBD [21]. The spots were manually annotated based on morphological features and gene markers [22].

Second, we downloaded 10X Genomics Visium ST data of a mouse brain coronal (MBC) section [23]. The annotation was collected from Squidpy [3], which annotated the spots using several resources, including the Allen Brain Atlas [24], the Mouse Brain gene expression atlas (http://mousebrain.org/ accessed on 17 July 2022), and this study [25].

Third, we collected a Slide-seqV2 dataset with 10 μm spatial resolution profiled from mouse hippocampus [26]. This dataset does not come with annotations that can serve as ground-truth since spatial domain detection was not performed in the original study. To evaluate the performance of spatial domain characterization visually, we also retrieved the annotation of hippocampus structures from the Allen Brain Atlas [24].

Fourth, we downloaded a mouse olfactory bulb ST dataset generated by Stereo-seq [27]. Analogously, because of a lack of ground-truth domain annotation for this dataset, we relied on the Allen Brain Atlas [24] to evaluate the consistency between the identified clusters and the annotated tissue structures.

A summary of the fifteen ST datasets used for performance evaluation is provided in Table 1.

Table 1. Summary of the fifteen benchmarking datasets.

Dataset	# of Spots	# of Genes	# of Domains	Sparsity	Protocol	Species	Reference
DLPFC_151507	4221	33,538	7	0.958	10X Visium	*Homo sapiens*	[22]
DLPFC_151508	4381	33,538	7	0.964	10X Visium	*Homo sapiens*	[22]
DLPFC_151509	4788	33,538	7	0.957	10X Visium	*Homo sapiens*	[22]
DLPFC_151510	4595	33,538	7	0.959	10X Visium	*Homo sapiens*	[22]
DLPFC_151669	3636	33,538	5	0.946	10X Visium	*Homo sapiens*	[22]
DLPFC_151670	3484	33,538	5	0.950	10X Visium	*Homo sapiens*	[22]
DLPFC_151671	4093	33,538	5	0.945	10X Visium	*Homo sapiens*	[22]
DLPFC_151672	3888	33,538	5	0.947	10X Visium	*Homo sapiens*	[22]
DLPFC_151673	3611	33,538	7	0.934	10X Visium	*Homo sapiens*	[22]
DLPFC_151674	3635	33,538	7	0.920	10X Visium	*Homo sapiens*	[22]
DLPFC_151675	3566	33,538	7	0.946	10X Visium	*Homo sapiens*	[22]
DLPFC_151676	3431	33,538	7	0.942	10X Visium	*Homo sapiens*	[22]
Brain coronal	2800	32,285	15	0.870	10X Visium	*Mus musculus*	[23]
Hippocampus	53,208	23,264	-	0.982	Slide-seqV2	*Mus musculus*	[26]
Olfactory bulb	19,527	27,106	-	0.987	Stereo-seq	*Mus musculus*	[27]

2.6. Performance Evaluation

Deciphering spatial domains (i.e., regions with similar spatial expression patterns) is one of the great challenges for STs [14]. We illustrate the benefits of SINFONIA for deciphering spatial domains from five perspectives: spatial clustering, domain resolution, latent representation, spot visualization, and computational efficiency. We identified SVGs by various methods and performed data analysis following the SCANPY vignette for ST data. All the experiments were performed on a standard desktop with an AMD Ryzen 7 1700X Eight-Core CPU with 32GB of RAM.

2.6.1. Spatial Clustering

We performed clustering on the latent representation of spots obtained from principal component analysis on expression levels of the identified SVGs following the SCANPY vignette for ST data [8]. Two widely-used community detection-based clustering methods, i.e., Louvain clustering and Leiden clustering, with default resolution in SCANPY, were adopted. In addition, for the situation where the number of clusters is specified, we implemented a binary search to tune the resolution parameter in clustering to make the number of clusters and the specified number as close as possible. The number was specified as the unique number of ground-truth labels in our benchmark experiments. For the binary search strategy, we searched the resolution in the range from 0.0 to 3.0. If the number of clusters did not match the specified number, the resolution value inducing the closest number of clusters to the specified number was used to perform clustering in the next iteration.

We considered the first scenario since the number of ground-truth spatial domains is unknown in advance in most studies and Louvain or Leiden clustering with default resolution is usually adopted [8,28]. The benchmark results in this scenario can indicate the performance of different SVG identification methods in general ST applications. We also considered the second scenario, one inspired by several studies for benchmarking single-cell analysis performance [15,29–32]. The benchmark results in this scenario can indicate the performance of different methods more fairly and objectively because the number of clusters is made to be as close as possible to the number of spatial domains.

Briefly, the former is more general and can be considered as an evaluation of real data, while the latter is more specific and can be considered as an evaluation of simulation data.

We evaluated the clustering results based on four widely-used metrics: adjusted mutual information (AMI) [33], adjusted Rand index (ARI) [34], homogeneity (Homo) [35], and normalized mutual information (NMI) [36]. Note that AMI is preferred for unbalanced clusters, while ARI is more suitable for ST datasets with large balanced clusters [15,32,37]. Since the sizes of spot populations in most ST data, including the datasets used in this study, are unbalanced, AMI is more appropriate than ARI in the benchmarking results.

To be more specific, let T denote the ground-truth labels of spots, P denote the clustering results, N denote the total number of spots, x_i denote the number of spots assigned to the i-th unique cluster of P, y_j denote the number of spots that belong to the j-th unique label of T, n_{ij} denote the number of overlapping spots between the i-th cluster and the j-th unique label, $MI(\cdot,\cdot)$ denote the mutual entropy, $E(\cdot)$ denote the expectation function, $H(\cdot)$ denote the entropy function, and $H(T|P)$ denote the uncertainty of ground-truth labels based on the knowledge of clustering results. Then the AMI score is calculated as (5):

$$\text{AMI} = \frac{MI(P,T) - E[MI(P,T)]}{avg[H(P),H(T)] - E[MI(P,T)]} \tag{5}$$

The ARI score is computed as (6):

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{x_i}{2}\sum_j \binom{y_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{x_i}{2} + \sum_j \binom{y_j}{2}\right] - \left[\sum_i \binom{x_i}{2}\sum_j \binom{y_j}{2}\right]/\binom{N}{2}} \tag{6}$$

The Homo score is calculated as (7):

$$\text{Homo} = 1 - \frac{H(T|P)}{H(T)} \tag{7}$$

The NMI score is computed as (8):

$$\text{NMI} = \frac{MI(P,T)}{\sqrt{H(P)H(T)}} \tag{8}$$

2.6.2. Domain Resolution

Intuitively, based on the ground-truth domain labels and the spot embeddings learned from the identified SVGs, discovery of more nearest neighbors with the same domain label indicates better domain resolution and, consequently, better performance for the characterization of spatial domains. As suggested by [38], we evaluated domain resolution by mean average precision (MAP). Supposing that the domain label of the i-th spot is $\mathrm{y}^{(i)}$, and the domain labels of its K ordered nearest neighbors are $\mathrm{y}_1^{(i)}, \mathrm{y}_2^{(i)}, \ldots, \mathrm{y}_K^{(i)}$, the MAP score is computed as (9) and (10).

$$\text{MAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}^{(i)} \tag{9}$$

$$\text{AP}^{(i)} = \begin{cases} \dfrac{\sum_{k=1}^{K} 1_{y^{(i)}=y_k} \cdot \frac{\sum_{j=1}^{k} 1_{y^{(i)}=y_j^{(i)}}}{k}}{\sum_{k=1}^{K} 1_{y^{(i)}=y_k^{(i)}}}, & if \ \sum\limits_{k=1}^{K} 1_{y^{(i)}=y_k^{(i)}} > 0 \\ 0, & otherwise \end{cases} \tag{10}$$

where $1_{y^{(i)}=y_k^{(i)}}$ is an indicator function that equals 1 if $y^{(i)} = y_k^{(i)}$ and 0 otherwise, and K is set to 30. For each spot, average precision (AP) measures the average label precision up to each domain-matched neighbor. MAP, ranging from 0 to 1, is the average AP across all spots, and a higher MAP indicates better domain resolution.

2.6.3. Latent Representation

We also investigated the representation performance of latent embeddings learned from the identified SVGs. First, we directly evaluated the information contained in the latent representation for predicting the true spatial domains in a supervised manner. Specifically, we treated the latent representation as the input and treated the true spatial domains as the output. As suggested by two benchmark studies on supervised cell type identification for single-cell RNA-seq data [39,40], we adopted the support vector machine (SVM) with radial basis function kernel as the classifier. We conducted five-fold cross-validation experiments by randomly splitting all spots into five folds and iteratively predicting spatial domains of the spots in each fold using the model trained with the remaining four folds. We computed the mean accuracy of the five folds (mean cross-validation accuracy, MCVA) to evaluate the predictive ability of the latent representation in predicting the spatial domains. A higher MCVA score suggests that the latent representation learned from the identified SVGs can better extract information in accurately predicting spatial domains.

We further adopted the local inverse Simpson's index (LISI), a commonly-used integration quality quantification metric in single-cell RNA-seq [41,42], to evaluate the characterized spatial domain patterns in latent representation for each spot. We built Gaussian kernel-based distributions of neighborhoods in the latent space, and calculated the LISI score using the *compute_lisi* function from the LISI R package with default parameters (perplexity = 30). A lower LISI score means more homogeneous neighborhood spatial domains of the spot [41,42]. To facilitate the benchmark, we used the transformed LISI scores, namely the tinverse of the median or mean LISI (iLISImd or iLISIm), to evaluate the performance. A higher iLISImd or iLISIm score suggests that the latent representation contains more effective information for deciphering spatial domains.

2.6.4. Spot Visualization

With the spatial coordinates and clustering results of spots, we assessed the performance of SVG identification methods via spot visualization. We collected annotations of related tissues from the Allen Reference Atlas [24], and used the annotation as a reference. We can evaluate the performance by visually investigating whether different domains were clearly segregated, whether the detected domains were spatially continuous and smooth, whether the revealed spatial domains were well consistent with the collected annotations, and whether the known tissue structures were successfully identified.

2.6.5. Computational Efficiency

Advanced high-throughput technologies for spatial transcriptomics have generated massive ST datasets, impeding exploratory data analysis on standard desktops. Therefore, we also benchmarked the computation time of various methods. We first compared the computation time of SINFONIA and its two variants with other SVG identification methods, including hvg [8], SOMDE [12], and SpatialDE [1]. Second, we compared the computation time of SINFONIA and its two variants with other Python implementations for the calculation of Moran's I and Geary's C, including Squidpy [3] and SpaGCN [5]. We note that the implementation in SpaGE [4] is the same as that in SpaGCN.

To sum up, we evaluated SVG identification performance by twenty-one metrics, including MAP, MCVA, iLISImd, iLISIm, computation time, and four clustering metrics for two clustering methods with default resolution or specified cluster number.

2.7. Baseline Methods

We compared the performance of SINFONIA with three baseline methods, including SpatialDE [1], SOMDE [12], and hvg [8], which is adopted in the widely-used SCANPY workflow for ST data. Specifically, SpatialDE and SOMDE were executed using the

Python packages SpatialDE (v1.1.3, https://github.com/Teichlab/SpatialDE accessed on 17 July 2022) and somde (v0.1.8, https://github.com/XuegongLab/somde accessed on 17 July 2022), respectively, while hvg was executed using the SCANPY (v1.9.1, https://github.com/scverse/scanpy accessed on 17 July 2022) implementation of the Seurat method for hvg filtering. We identified SVGs by various methods with the default parameters or settings provided in the accompanying examples. Following the SCANPY vignette for ST data, the top 2000 SVGs identified by various methods were used for benchmarking. We used the standard SCANPY workflow for ST data with default settings to evaluate the performance for characterizing spatial domains.

We also tried to benchmark the performance of other three Python-based methods, including SPADE [10], GPcounts [11], and HMRF [13]. However, SPADE requires additional high-resolution tissue images as input, which are not available on most of the ST datasets. Additionally, it would have led to an unfair comparison, even though the images are available. We then ran GPcounts following its tutorials. However, when we performed GPcounts on the smallest dataset, i.e., the MBC dataset, the status bar demonstrated that the computational time of the method would be more than 1668 h, making it impossible to benchmark the performance of GPcounts. We also tried to benchmark HMRF. Nevertheless, we failed to access the instructions of HMRF (http://spatial.rc.fas.harvard.edu/install.html accessed on 17 July 2022). Therefore, we finally adopted SpatialDE [1], SOMDE [12], and hvg [8] as baseline methods.

3. Results

3.1. SINFONIA Enables Accurate Spatial Clustering

It is intuitive that, based on the ground-truth domain labels and expression levels of the identified SVGs, a higher score of clustering metric indicates better performance for deciphering spatial domains. We first quantitatively evaluated the spatial clustering performance of SINFONIA. To mimic the scenario where the number of spatial domains is unknown, we used Louvain clustering with default resolution. This is a general scenario, since the number of ground-truth spatial domains is unknown in advance in most studies. We performed spatial clustering on each of the 12 DLPFC datasets. As shown in Figure 2A, SINFONIA achieved significantly higher AMI, ARI, Homo and NMI values than did the baseline methods. SpatialDE also achieved satisfactory performance, while with high variance across the 12 DLPFC datasets. Additionally, we considered the scenario where a specified number of clusters is given. The scenario is often adopted in benchmarking studies [15,29–32]. We used a binary search strategy to obtain clusters, and SINFONIA again outperformed the baseline methods (Figure 2B). In addition, we also evaluated the performance of two SINFONIA variants, namely SINFONIA_*I* and SINFONIA_*C*, to demonstrate the advantages of the ensemble strategy in SINFONIA. The evaluation of SINFONIA and its two variants is equivalent to ablation experiments. As shown in Figure 2A,B, SINFONIA achieved better performance than SINFONIA_*I*, and SINFONIA_*I* achieved better performance than SINFONIA_*C*, indicating that SINFONIA integrates the advantages of Moran's *I* and Geary's *C* and thus achieves the best performance. In addition to Louvain clustering, we also assessed the spatial clustering performance of SVGs identified by various methods via Leiden clustering, which is another widely-used clustering method in the single-cell community [2,8]. As shown in Figure 2C,D, Leiden clustering with default resolution or searched resolution also provided similar results, suggesting that the evaluation results are robust relative to clustering methods and SINFONIA consistently outperformed the baseline methods regardless of clustering methods and scenarios.

To demonstrate the superiority of spatial clustering results across multiple datasets, we further benchmarked SINFONIA on the MBC dataset. As shown in Figure 3A–D, SINFONIA again deciphered the spatial domains better than baseline methods. Additionally, SINFONIA provided overall better performance than its two variants and was more robust than the variants. We note that SINFONIA_*C* outperformed SINFONIA_*I* when clustering by Leiden with default resolution on this dataset, while SINFONIA_*I* outperformed SIN-

FONIA_*C* in other cases and on the 12 DLPFC datasets, suggesting that SVG identification based on an individual statistic may provide fluctuating performance while SINFONIA can provide the overall best performance based on its ensemble strategy.

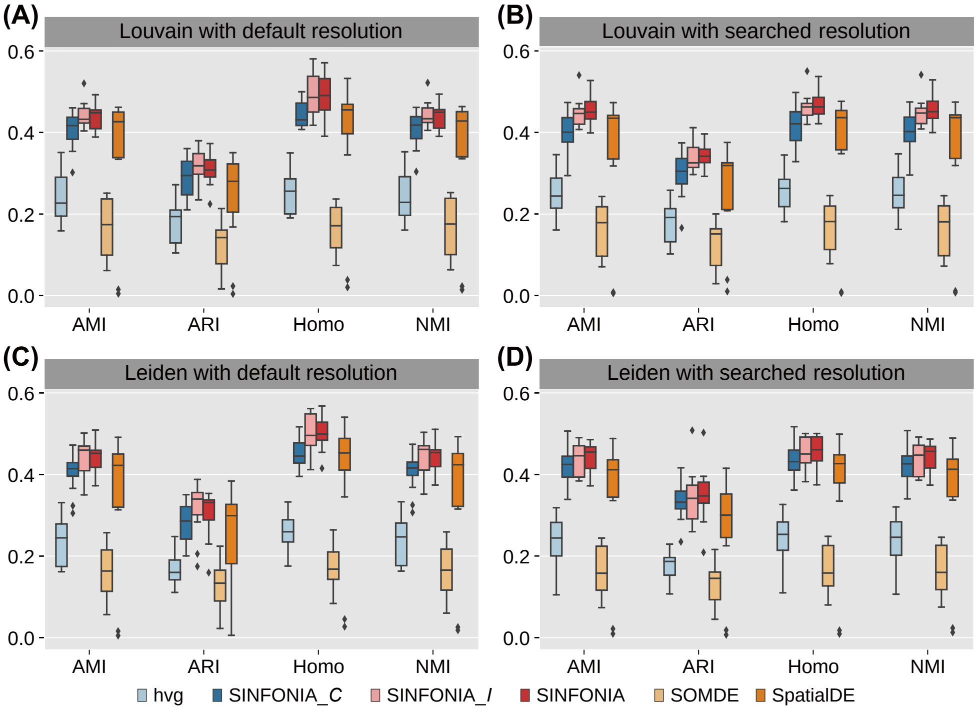

Figure 2. Louvain clustering results with (**A**) default resolution and (**B**) searched resolution (given a specified number of clusters) on 12 DLPFC datasets. Leiden clustering results with (**C**) default resolution and (**D**) searched resolution on 12 DLPFC datasets.

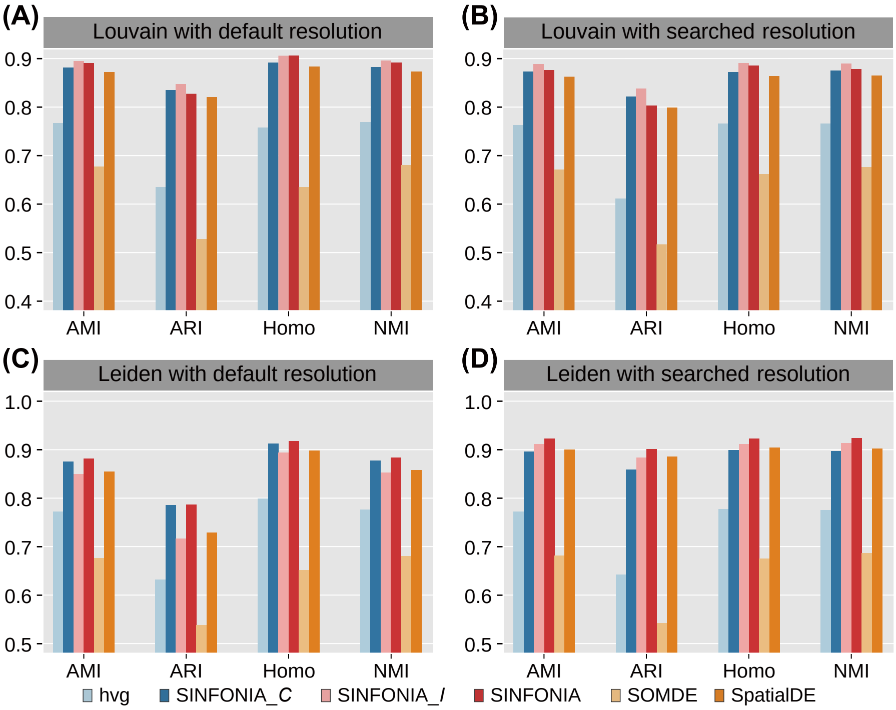

Figure 3. Louvain clustering results with (**A**) default resolution and (**B**) searched resolution on the MBC dataset. Leiden clustering results with (**C**) default resolution and (**D**) searched resolution on the MBC dataset.

3.2. SINFONIA Effectively Characterizes Spatial Patterns

To further investigate the effectiveness of spatial patterns that were characterized based on the SVGs identified by the different methods, we assessed the domain resolution by MAP, and assessed the prediction ability by MCVA, iLISImd and iLISIm. Again, SINFONIA achieved consistent superiority over baseline methods and, overall, performed better than its two variants on 12 DLPFC datasets (Figure 4A,B) and the MBC dataset (Figure 4C). Note that hvg, the SVG identification method used in the SCANPY vignette tailored for ST data analysis, provided relatively undesirable performance, which is consistent with the evaluation results in spatial clustering. The above results indicate that a simple replacement of the SVG identification step with SINFONIA can significantly improve the domain resolution and prediction ability of characterized spatial patterns.

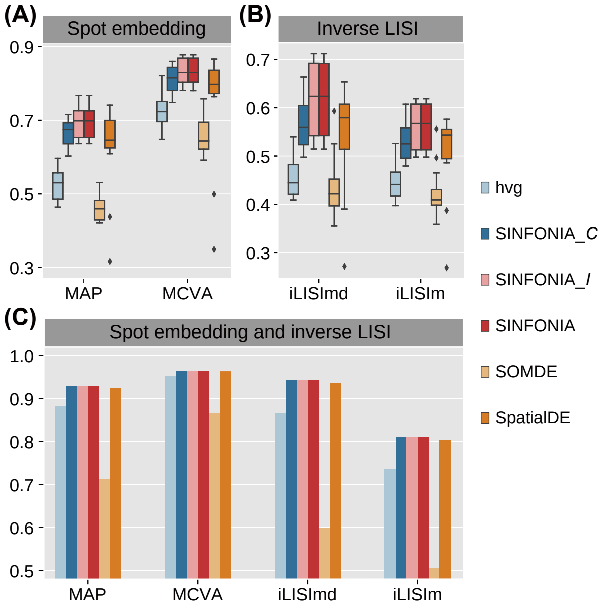

Figure 4. (**A**) Performance of MAP and MCVA on the 12 DLPFC datasets. (**B**) Performance of iLISImd and iLISIm on the 12 DLPFC datasets. (**C**) Performance of MAP, MCVA, iLISImd and iLISIm on the MBC dataset.

3.3. SINFONIA Facilitates Interpretable Spot Visualization

Interpretable visualization of spatial domains is important for researchers to better understand tissue structures and study biological functions. We next applied SINFONIA to a mouse hippocampus dataset sequenced by Slide-seqV2 [26] and a mouse olfactory bulb dataset sequenced by Stereo-seq [27]. Since these two datasets lack manually annotated spatial domain labels, we evaluated the interpretability by comparing the spatial domains obtained by clustering (Louvain with default resolution) with the annotated function structures in the Allen Brain Atlas [24].

We visualized the spots with the spatial coordinates and clustering results. As shown in Figure 5A,B, on both of these two datasets, the spatial domains revealed based on the SVGs identified by SINFONIA were well consistent with the annotation from the Allen Reference Atlas, while the clusters identified by other methods lacked clear spatial separation, indicating the advantage of SINFONIA for characterizing and visualizing spatial domains. For example, based on the SVGs identified by SINFONIA, we successfully detected the third ventricle (V3), dentate gyrus-granule cell layer (DG-sg), and CA field of pyramidal layer in the mouse hippocampus, and detected the subependymal zone (SEZ) in the mouse olfactory bulb. However, based on the SVGs identified by other methods, most of the tissue structures can hardly be characterized. Additionally, the baseline methods tended to claim significantly more spatial domains than did SINFONIA, which may be due to the limited spatial patterns and the even noise in the identified SVGs. In addition, we note that errors were encountered when performing SpatialDE and SOMDE on these two datasets, respectively, highlighting the importance of user-friendliness and software quality to the success of bioinformatics tools [43].

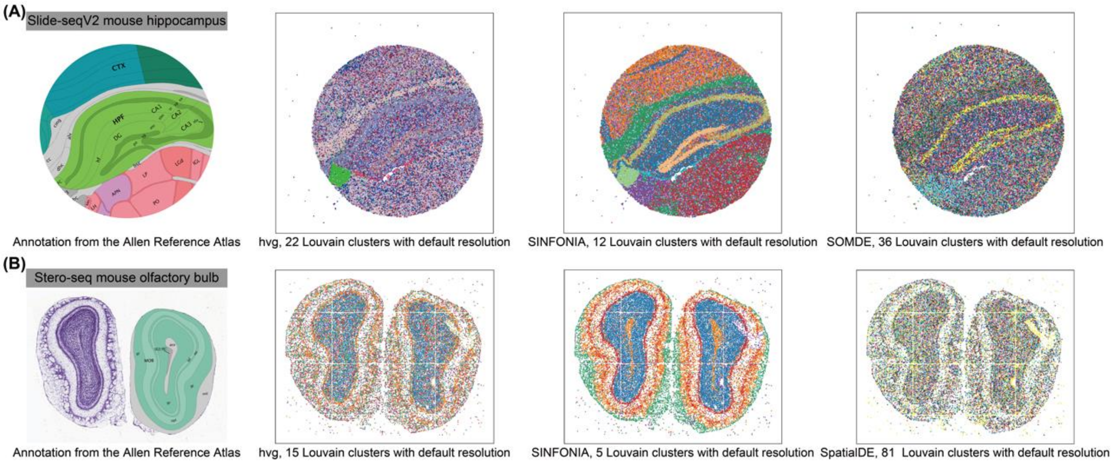
(A)
Slide-seqV2 mouse hippocampus
CTX
HPF
CA1
CA2
CA3
DG
Annotation from the Allen Reference Atlas
hvg, 22 Louvain clusters with default resolution
SINFONIA, 12 Louvain clusters with default resolution
SOMDE, 36 Louvain clusters with default resolution
(B)
Stero-seq mouse olfactory bulb
MOB
Annotation from the Allen Reference Atlas
hvg, 15 Louvain clusters with default resolution
SINFONIA, 5 Louvain clusters with default resolution
SpatialDE, 81 Louvain clusters with default resolution

Figure 5. Visualization of datasets of (**A**) a mouse hippocampus and (**B**) a mouse olfactory bulb with tissue regions annotated from the Allen Brain Atlas and spatial domains detected by different methods.

3.4. *SINFONIA Is Robust and Computationally Efficient*

The robustness of hyperparameters is an important aspect that affects the ease of use of a method. The major hyperparameter of SVG identification is the number of SVGs to be selected. In order to investigate the robustness of SINFONIA relative to the choice of the number of SVGs, we further performed SINFONIA on the 12 DLPFC datasets to select different numbers of SVGs. We adopted 20 metrics (i.e., four clustering metrics for two clustering methods with default resolution or specified cluster number: MAP, MCVA, iLISImd, and iLISIm) to demonstrate the robustness. For each choice of the number of SVGs, we computed the median of each metric on the 12 datasets, and thus obtained 20 median metrics. We then performed two-sided Wilcoxon signed-rank tests on different pairs of choice of the number of SVGs to test if the performance has a significant difference using different numbers of SVGs. As shown in Figure 6A, the performance of SINFONIA did not differ significantly using various numbers of SVGs (all the p-values were greater than 0.1), indicating that SINFONIA was robust to the choice of the number of SVGs.

To catch up with the growth in data throughput, computational methods should be designed with scalability in mind. Therefore, in addition to the comparison of spatial characterization performance of different SVG identification methods, we also compared the computational efficiency of different methods on the 12 DLPFC datasets. As shown in Figure 6B, SINFONIA and its two variants significantly outperformed the other methods except for the conventional hvg method. Additionally, we also systematically benchmarked the computational efficiency of different implementations for calculating Moran's I and Geary's C. In the Python community of ST data analysis, Squidpy [3] and SpaGCN [5] provide implementations for the calculation of Moran's I and Geary's C. The Moran's I implementation of SpaGCN is the same as that of SpaGE [4]. Hence, we compared the computational efficiency of SINFONIA and its two variants with Squidpy and SpaGCN. We used various implementations to calculate spatial autocorrelation statistics based on the same spatial neighbor graph to ensure the fairness of assessment. As shown in Figure 6C, SINFONIA offered about four- to two-hundred-and-sixty-two- times speedup compared with other implementations for calculating spatial autocorrelation statistics Moran's I and Geary's C. In addition, SINFONIA computed Moran's I and Geary's C simultaneously with almost no extra time consumption compared to computing a single statistic. The above results highlight that the contribution of SINFONIA to the ST data analysis community lies in not only the superior spatial characterization performance, but also the advanced implementation of spatial autocorrelation statistics.

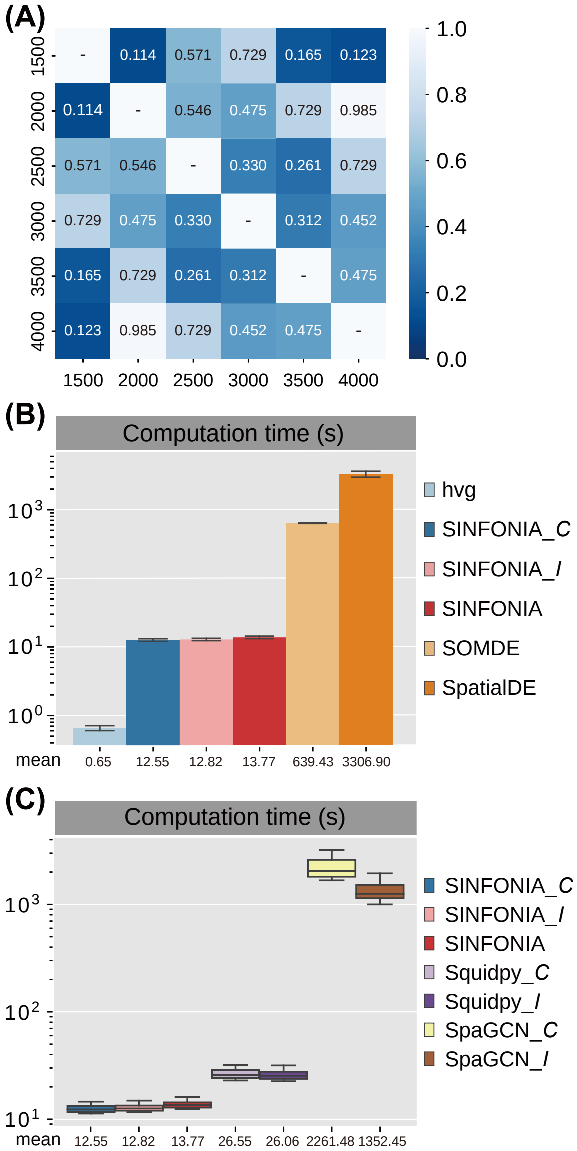

Figure 6. (**A**) *p*-values of two-sided Wilcoxon signed-rank tests on the performance of different numbers of SVGs identified by SINFONIA. (**B**) Computation time of different methods on 12 DLPFC datasets. (**C**) Computation time of different implementations for calculating spatial autocorrelation statistics on 12 DLPFC datasets.

3.5. SINFONIA Improves the Performance of Other Spatial Embedding Methods

In addition to the principal component analysis used in the SCANPY vignette tailored for ST data analysis, we further adopted STAGATE, the state-of-the-art method for ST data embedding [14], to demonstrate that using the SVGs identified by SINFONIA can also improve the performance of more advanced spatial embedding methods. We systematically evaluated the performance of STAGATE on the 12 DLPFC datasets and the MBC section dataset using the SVGs identified by hvg and SINFONIA, respectively. We assessed the performance with spatial clustering, domain resolution and prediction ability. As shown in Figure 7A,B, by simply replacing the SVG identification method from hvg used in SCANPY workflow to SINFONIA, STAGATE can achieve significant improvements. The results of the combination of SINFONIA and STAGATE indicate that our method can be used as a general step in ST data analysis to enhance the performance of existing ST data embedding methods.

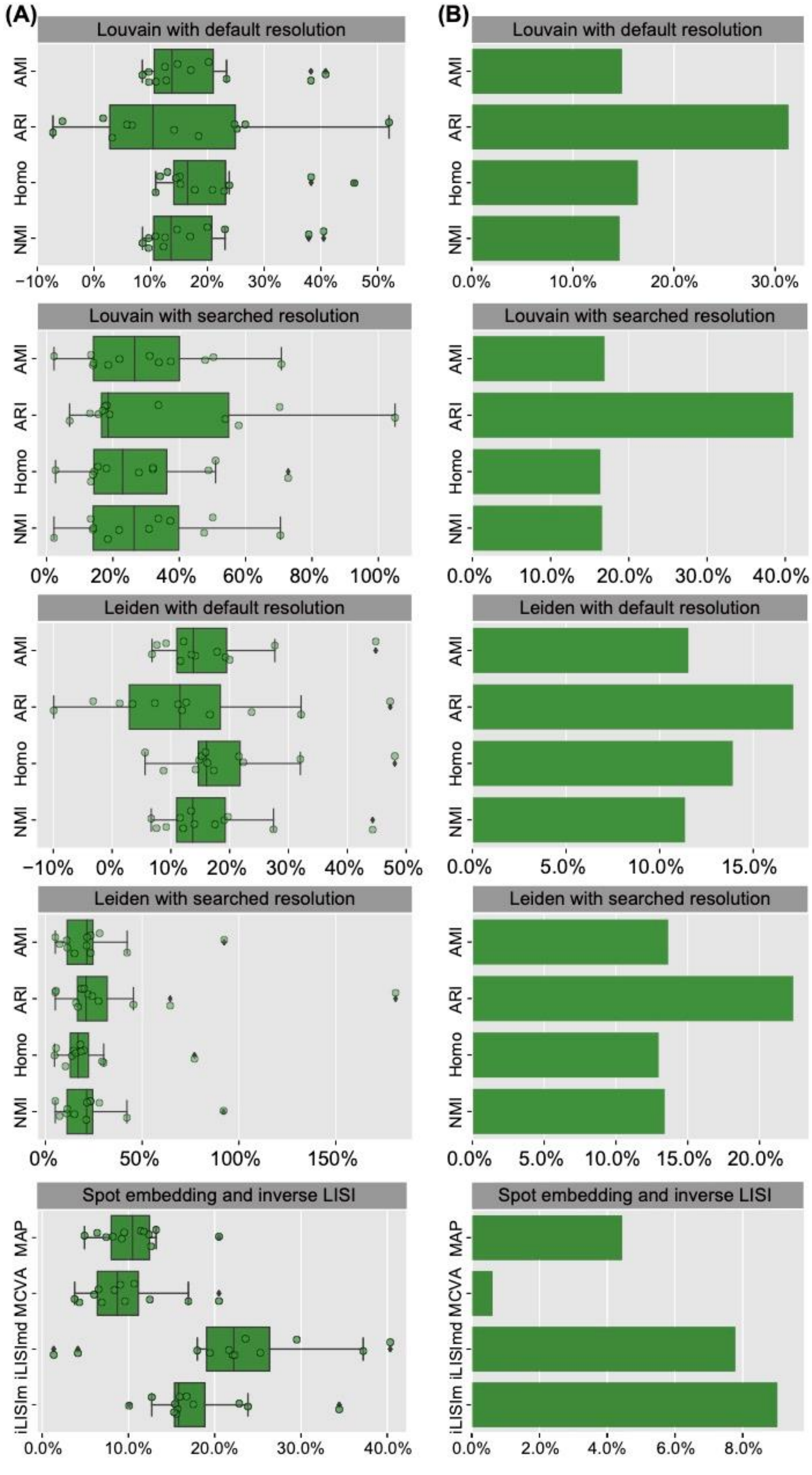

Figure 7. Performance improvement of STAGATE using SINFONIA on (**A**) the 12 DLPFC datasets and (**B**) the MBC dataset. The performance of spatial clustering, domain resolution and prediction ability were evaluated.

4. Discussion

Rapid advances in spatially resolved transcriptomics have revolutionized the interrogation of spatial heterogeneity and granted us a novel perspective on the cellular transcriptome. The identification of spatially variable genes (SVGs) is an essential step for downstream spatial domain characterization, and is regarded as the first critical step in spatial transcriptomic (ST) data analysis [1]. In this study, we propose SINFONIA, a scalable method for identifying spatially variable genes via ensemble strategies. We systematically benchmarked the performance of SINFONIA on 15 ST datasets generated with different protocols, and with different sizes, dimensions and qualities. We have demonstrated the applications of SINFONIA from five perspectives, including spatial clustering, spatial domain annotation, spot visualization, spatial embedding and applicability to various ST datasets. Furthermore, we have also shown that SINFONIA is robust relative to the number of SVGs and is computationally efficient. As an ensemble-based method, SINFONIA can provide the overall best performance based on both the global autocorrelation (Moran's *I*) and the local spatial autocorrelation (Geary's *C*), highlighting the contribution of the ensemble strategy in SINFONIA for SVG identification. Certainly, although SINFONIA successfully integrates the spatial coordinates and gene expression profiles in ST data to identify SVGs, its performance can be further improved in the future. For example, more advanced computational technologies and hardware such as the graphics processing unit (GPU) can be used to further accelerate the computation of spatial autocorrelation statistics, and thus facilitate the analysis of large-scale ST data.

5. Conclusions

This study is the first attempt to apply spatial autocorrelation statistics to the identification of SVGs for deciphering spatial domains. Additionally, it is the first time that a systematic and quantitative evaluation of SVG identification performance has been conducted using a number of ST datasets. SINFONIA, the ensemble-based method proposed in this study, provides an effective way to identify SVGs to facilitate ST data analysis, and achieves about four- to two-hundred-sixty-two-times speedup compared with other implementations. With its smooth interfaces to SCANPY and the Python data science ecosystem, we anticipate broad application of SINFONIA within ST data analysis.

Author Contributions: Conceptualization, S.C. and R.J.; methodology, R.J.; software, S.C.; validation, R.J., Z.L. and S.C.; formal analysis, S.C.; investigation, S.C.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, S.C.; writing—review and editing, R.J. and S.C.; visualization, Y.J. and S.L.; supervision, S.C.; project administration, S.C.; funding acquisition, R.J. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2021YFF1200902, and the National Natural Science Foundation of China, grant numbers 62203236, 62273194, 61873141 and 61721003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The 10X Genomics Visium dataset of 12 DLPFC sections [22] can be downloaded from spatialLIBD at http://spatial.libd.org/spatialLIBD/ (accessed on 17 July 2022) [21]. The 10X Genomics Visium dataset of a coronal mouse brain section is available at https://support.10xgenomics.com/spatial-gene-expression/datasets (accessed on 17 July 2022) [23]. The Slide-seqV2 dataset of mouse hippocampus is accessible at https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2#study-summary (accessed on 17 July 2022) [26]. The Stereo-seq dataset of mouse olfactory bulb tissues can be downloaded from https://db.cngb.org/search/project/CNP0001543/ (accessed on 17 July 2022) [27]. The annotated function structures in Allen Brain Atlas are available at https://atlas.brain-map.org/ (accessed on 17 July 2022) [24]. SINFONIA is freely accessible at https://sinfonia-svg.readthedocs.io/ (accessed on 17 July 2022) under the MIT License, along with detailed tutorials and examples. It can be seamlessly integrated into existing analysis workflows. The source code is available at https://github.com/BioX-NKU/SINFONIA (accessed on 17 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Svensson, V.; Teichmann, S.A.; Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **2018**, *15*, 343–346. [CrossRef] [PubMed]
2. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M., III; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587.e29. [CrossRef] [PubMed]
3. Palla, G.; Spitzer, H.; Klein, M.; Fischer, D.; Schaar, A.C.; Kuemmerle, L.B.; Rybakov, S.; Ibarra, I.L.; Holmberg, O.; Virshup, I.; et al. Squidpy: A scalable framework for spatial omics analysis. *Nat. Methods* **2022**, *19*, 171–178. [CrossRef] [PubMed]
4. Abdelaal, T.; Mourragui, S.; Mahfouz, A.; Reinders, M.J.T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **2020**, *48*, E107. [CrossRef] [PubMed]
5. Hu, J.; Li, X.; Coleman, K.; Schroeder, A.; Ma, N.; Irwin, D.J.; Lee, E.B.; Shinohara, R.T.; Li, M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **2021**, *18*, 1342–1351. [CrossRef]
6. Li, K.; Yan, C.; Li, C.; Chen, L.; Zhao, J.; Zhang, Z.; Bao, S.; Sun, J.; Zhou, M. Computational elucidation of spatial gene expression variation from spatially resolved transcriptomics data. *Mol. Ther. Nucleic Acids* **2022**, *27*, 404–411. [CrossRef]
7. Lu, L.; Welch, J.D. PyLiger: Scalable single-cell multi-omic data integration in Python. *Bioinformatics* **2022**, *38*, 2946–2948. [CrossRef]
8. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 15. [CrossRef]
9. Gayoso, A.; Lopez, R.; Xing, G.; Boyeau, P.; Valiollah Pour Amiri, V.; Hong, J.; Wu, K.; Jayasuriya, M.; Mehlman, E.; Langevin, M.; et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **2022**, *40*, 163–166. [CrossRef]
10. Bae, S.; Choi, H.; Lee, D.S. Discovery of molecular features underlying the morphological landscape by integrating spatial transcriptomic data with deep features of tissue images. *Nucleic Acids Res.* **2021**, *49*, e55. [CrossRef]
11. BinTayyash, N.; Georgaka, S.; John, S.T.; Ahmed, S.; Boukouvalas, A.; Hensman, J.; Rattray, M. Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. *Bioinformatics* **2021**, *37*, 3788–3795. [CrossRef] [PubMed]
12. Hao, M.; Hua, K.; Zhang, X. SOMDE: A scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics* **2021**, *37*, 4392–4398. [CrossRef] [PubMed]
13. Zhu, Q.; Shah, S.; Dries, R.; Cai, L.; Yuan, G.C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **2018**, *36*, 1183–1190. [CrossRef] [PubMed]
14. Dong, K.; Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **2022**, *13*, 1739. [CrossRef]
15. Chen, S.; Zhang, B.; Chen, X.; Zhang, X.; Jiang, R. stPlus: A reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **2021**, *37*, i299–i307. [CrossRef]
16. Zeng, Z.; Li, Y.; Li, Y.; Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol.* **2022**, *23*, 83. [CrossRef]
17. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef]
18. Geary, R.C. The Contiguity Ratio and Statistical Mapping. *Inc. Stat.* **1954**, *5*, 115–146. [CrossRef]
19. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
20. Lahnemann, D.; Koster, J.; Szczurek, E.; McCarthy, D.J.; Hicks, S.C.; Robinson, M.D.; Vallejos, C.A.; Campbell, K.R.; Beerenwinkel, N.; Mahfouz, A.; et al. Eleven grand challenges in single-cell data science. *Genome Biol* **2020**, *21*, 31. [CrossRef]
21. Pardo, B.; Spangler, A.; Weber, L.M.; Page, S.C.; Hicks, S.C.; Jaffe, A.E.; Martinowich, K.; Maynard, K.R.; Collado-Torres, L. spatialLIBD: An R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genom.* **2022**, *23*, 434. [CrossRef] [PubMed]
22. Maynard, K.R.; Collado-Torres, L.; Weber, L.M.; Uytingco, C.; Barry, B.K.; Williams, S.R.; Catallini, J.L., II; Tran, M.N.; Besich, Z.; Tippani, M.; et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **2021**, *24*, 425–436. [CrossRef]
23. 10XGenomics. Visium Spatial Gene Expression Reagent Kits User Guide. 2021. Available online: https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen/documentation/steps/library-construction/visium-spatial-gene-expression-reagent-kits-user-guide (accessed on 17 July 2022).
24. Sunkin, S.M.; Ng, L.; Lau, C.; Dolbeare, T.; Gilbert, T.L.; Thompson, C.L.; Hawrylycz, M.; Dang, C. Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **2013**, *41*, D996–D1008. [CrossRef] [PubMed]

25. Gracia Villacampa, E.; Larsson, L.; Mirzazadeh, R.; Kvastad, L.; Andersson, A.; Mollbrink, A.; Kokaraki, G.; Monteil, V.; Schultz, N.; Appelberg, K.S.; et al. Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genom.* **2021**, *1*, 100065. [CrossRef]
26. Stickels, R.R.; Murray, E.; Kumar, P.; Li, J.; Marshall, J.L.; Di Bella, D.J.; Arlotta, P.; Macosko, E.Z.; Chen, F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **2021**, *39*, 313–319. [CrossRef] [PubMed]
27. Chen, A.; Liao, S.; Cheng, M.; Ma, K.; Wu, L.; Lai, Y.; Qiu, X.; Yang, J.; Xu, J.; Hao, S.; et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **2022**, *185*, 1777–1792.e1721. [CrossRef]
28. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M., 3rd; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **2019**, *177*, 1888–1902.e21. [CrossRef]
29. Chen, H.; Lareau, C.; Andreani, T.; Vinyard, M.E.; Garcia, S.P.; Clement, K.; Andrade-Navarro, M.A.; Buenrostro, J.D.; Pinello, L. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **2019**, *20*, 241. [CrossRef]
30. Danese, A.; Richter, M.L.; Chaichoompu, K.; Fischer, D.S.; Theis, F.J.; Colome-Tatche, M. EpiScanpy: Integrated single-cell epigenomic analysis. *Nat. Commun.* **2021**, *12*, 5228. [CrossRef]
31. Chen, S.; Wang, R.; Long, W.; Jiang, R. ASTER: Accurately estimating the number of cell types in single-cell chromatin accessibility data. *Bioinformatics* **2023**, *39*, btac842. [CrossRef]
32. Chen, S.; Yan, G.; Zhang, W.; Li, J.; Jiang, R.; Lin, Z. RA3 is a reference-guided approach for epigenetic characterization of single cells. *Nat. Commun.* **2021**, *12*, 2177. [CrossRef] [PubMed]
33. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 1073–1080.
34. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
35. Rosenberg, A.; Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. 2007, pp. 410–420. Available online: https://aclanthology.org/D07-1043.pdf (accessed on 17 July 2022).
36. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2003**, *3*, 583–617. [CrossRef]
37. Romano, S.; Vinh, N.X.; Bailey, J.; Verspoor, K. Adjusting for chance clustering comparison measures. *J. Mach. Learn. Res.* **2016**, *17*, 1–32.
38. Cao, Z.-J.; Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **2022**, *40*, 1458–1466. [CrossRef] [PubMed]
39. Abdelaal, T.; Michielsen, L.; Cats, D.; Hoogduin, D.; Mei, H.; Reinders, M.J.T.; Mahfouz, A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **2019**, *20*, 194. [CrossRef] [PubMed]
40. Ma, W.; Su, K.; Wu, H. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: Classifier, feature selection, and reference construction. *Genome Biol.* **2021**, *22*, 264. [CrossRef] [PubMed]
41. Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.R.; Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **2019**, *16*, 1289–1296. [CrossRef]
42. Shang, L.; Zhou, X. Spatially Aware Dimension Reduction for Spatial Transcriptomics. *Nat. Commun.* **2022**, *13*, 7203. [CrossRef]
43. Mangul, S.; Martin, L.S.; Eskin, E.; Blekhman, R. Improving the usability and archival stability of bioinformatics software. *Genome Biol.* **2019**, *20*, 47. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.