Molecular Therapy
Nucleic Acids
Review

Computational elucidation of spatial gene expression variation from spatially resolved transcriptomics data

Ke Li,[1,2] Congcong Yan,[1,2] Chenghao Li,[1,2] Lu Chen,[1] Jingting Zhao,[1] Zicheng Zhang,[1] Siqi Bao,[1] Jie Sun, PhD,[1] and Meng Zhou, PhD[1]

[1]School of Biomedical Engineering, School of Ophthalmology & Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325027, P. R. China

Recent advances in spatially resolved transcriptomics (SRT) have revolutionized biological and medical research and enabled unprecedented insight into the functional organization and cell communication of tissues and organs *in situ*. Identifying and elucidating gene spatial expression variation (SE analysis) is fundamental to elucidate the SRT landscape. There is an urgent need for public repositories and computational techniques of SRT data in SE analysis alongside technological breakthroughs and large-scale data generation. Increasing efforts to use *in silico* techniques in SE analysis have been made. However, these attempts are widely scattered among a large number of studies that are not easily accessible or comprehensible by both medical and life scientists. This study provides a survey and a summary of public resources on SE analysis in SRT studies. An updated systematic overview of state-of-the-art computational approaches and tools currently available in SE analysis are presented herein, emphasizing recent advances. Finally, the present study explores the future perspectives and challenges of *in silico* techniques in SE analysis. This study guides medical and life scientists to look for dedicated resources and more competent tools for characterizing spatial patterns of gene expression.

INTRODUCTION

Recent advances in sequencing technologies, such as single-cell RNA sequencing (scRNA-seq), have allowed for the dissection of gene expression in tens of thousands of individual cells at a single-cell resolution as well as have enabled the analysis of cellular composition and heterogeneity of complex tissues using high-throughput methods.[1–4] However, scRNA-seq results in the loss of spatial context during the separation process even if it provides an estimate of the whole transcriptome, thus hindering the further understanding of the tissue structure and cell state.[5] Moreover, increasing evidence has suggested that the heterogeneity of gene expression patterns is closely associated with the characteristics of cell types and the microenvironment in which cells are located.[6–8] Although scRNA-seq allows for detecting highly variable genes that contribute strongly to cell-type differences, as a result of ignoring spatial separation, these cell-type-specific genes may not have spatially coherent expression patterns. The introduction of spatially resolved transcriptomics (SRT) technologies provides significant opportunities to link the gene expression profiling and location information of cells within the tissues.[9] A critical task in SRT studies is to identify spatially variable genes (SVGs) that have distinct spatial expression patterns across spatial locations. Identifying SVGs provides an opportunity to systematically analyze the state of cells in specific locations, infer the communication between cells, and determine vital phenotypes and functions in organisms. For example, Navarro et al. identified spatially different genes showing specific spatial expression patterns in the brains of Alzheimer model mice, and these could be used as novel molecular targets for the treatment of Alzheimer disease (AD).[10] Wang et al. explored the tumor microenvironment of prostate cancer with SRT data and revealed a series of new metabolic genes with spatial heterogeneity, which may drive vital functions of tumor cells.[11]

A traditional method to identify genes with spatial expression heterogeneity is to perform a differential expression analysis on different spatial regions directly.[12,13] However, this method can only reveal variations obtained by differences between discrete clusters or regions and cannot detect those spatial genes that show a gradient-like expression across the spatial region. Therefore, with the development of bioinformatics in SRT, a series of emerging computational methods have been developed and proposed to make full use of spatial gene expression data for elucidating spatial gene expression variation.

The present study aimed to provide an overview of public repositories and state-of-the-art computational algorithms, approaches, and tools that are currently available in SE analysis, with an emphasis on recent advances. This summary does not only comprise a list of dedicated resources and more competent tools for characterizing spatial

https://doi.org/10.1016/j.omtn.2021.12.009.

[2]These authors contributed equally

Correspondence: Jie Sun, School of Biomedical Engineering, School of Ophthalmology & Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325027, P. R. China.
E-mail: suncarajie@wmu.edu.cn
Correspondence: Meng Zhou, School of Biomedical Engineering, School of Ophthalmology & Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325027, P. R. China.
E-mail: zhoumeng@wmu.edu.cn

Table 1. Overview of resources and databases for spatially resolved transcriptomics

Database	Description	URL
SpatialDB	a database for spatially resolved transcriptomic datasets	https://www.spatialomics.org/SpatialDB/
Single Cell Portal	a comprehensive database for single-cell and SRT studies	https://singlecell.broadinstitute.org/single_cell

patterns of gene expression in the fields of medical and life sciences but also may be helpful in the fields of bioinformatics and computational science for the development of more powerful and efficient *in silico* techniques in SRT studies.

RESULTS

Data repositories and resources for SRT

With the rapid development of SRT technologies, SRT-related data are gradually accumulating. However, currently available data resources and databases are limited, and numerous data have not been systematically organized. Here, we list existing data repositories and resources for SRT, thereby offering convenience for researchers in this field (Table 1).

SpatialDB (https://www.spatialomics.org/SpatialDB/)[14] is currently the most comprehensive manually curated database collecting SRT datasets. It integrates 24 publicly available datasets of tissues from humans, mice, *Caenorhabditis elegans*, drosophila, and zebrafish that have been generated by eight SRT techniques. In addition, SpatialDB shows SVGs identified by SpatialDE and trendsceek, as well as data visualization, comparison, and Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analyses.

Single Cell Portal (https://singlecell.broadinstitute.org/single_cell) is a growing comprehensive single-cell database, which has collected and integrated 17,640,076 cells from 400 studies, including from SRT studies and datasets; most of them are from SRT technologies developed at the Broad Institute, such as high-definition spatial transcriptomics and slide-seqV2.[15,16]

Computational methods for identifying SVGs

Many computational efforts have been made during the past few years to help elucidate spatial gene expression variation. However, these attempts are scattered among a large number of studies. In this section, these state-of-the-art computational methods and tools will be systematically introduced and summarized. Based on the intrinsic principle, these existing methods could be classified into three categories, based on the methods they use: (1) statistical-modeling-based methods; (2) machine-learning-based methods; and (3) spatial-grid-based methods (Table 2).

Statistical-modeling-based methods

Based on known cell spatial coordinates and their gene expression levels, statistical-modeling-based methods provide statistical frameworks to elucidate spatial gene expression heterogeneity. A schematic workflow of statistical-modeling-based methods is illustrated in Figure 1. First, the gene expression profile and the location information of cells are input. According to the input information, statistical frameworks to clarify the dependence between the gene expression values and the spatial location of cells were constructed. Subsequently, significant SVGs are determined by different statistical methods.

trendsceek uses marked point processes to model the association between gene expression and cell coordinates.[17] trendsceek represents each point as a cell and the mark of the point as the gene expression value and calculates the distance between points. For a specific distance, evaluating whether the mark of the gene is dependent on the location of the point; in other words, whether marker separation occurred. Four types of dependency evaluation methods are used to test marker separation (V-mark, E-mark, the mark-variogram, and Stoyan's mark correlation). The score should be variable in different distances if the marks and the distribution of points are dependent.

SpatialDE is a Gaussian-process-regression-based method.[18] The Gaussian Process (GP) is a random process also known as normal distribution.[19] It allows for non-linear regression and quantification of the association between the measurement process and latent function, and GPs are useful for SRT data to model spatial gradient changes in gene expression. SpatialDE establishes a linear mixed model for gene expression profiles with Gaussian kernels and decomposes the variation of each gene as spatial or non-spatial variations.[18] The non-spatial variation is modeled using observation noise, and the spatial variation is denoted by the covariance matrix of gene expression values and spatial cell coordinates. For each Gaussian kernel, SpatialDE calculates an approximate p value using the likelihood test compared with a null model and identifies genes with significant spatial variability.

Compared to SpatialDE, SPARK makes some specific improvements.[20] SPARK recognizes SVGs based on a spatial generalized linear mixed model with multiple spatial kernels, including Gaussian kernels and periodic kernels, to directly model spatial count data. In order to adapt to different spatial patterns, SPARK uses ten spatial kernels by default, including the most common spatial expression patterns. SPARK relies on the mixture χ^2 distributions to accurately test the p value of each spatial kernel and then integrates all p values using the Cauchy combination rule to obtain a well-calibrated p value, which can effectively control the occurrence of type I errors. In addition, a Gaussian version of SPARK has been developed that can keep a stable model performance when facing SRT data with high counts.

Table 2. Overview of computational tools and methods for identification of spatially variable genes

Method	Description	Platform	URL	Reference
Statistical-modeling-based methods				
trendsceek	based on marked point processes	R	https://github.com/edsgard/trendsceek	Edsgard et al.[17]
SpatialDE	based on Gaussian process regression	Python	https://github.com/Teichlab/SpatialDE	Svensson et al.[18]
SPARK	based on spatial generalized linear mixed model with multiple spatial kernels	R	https://xzhoulab.github.io/SPARK/	Sun et al.[20]
SPARK-X	based on the non-parametric model	R	https://xzhoulab.github.io/SPARK/	Zhu et al.[21]
GPcounts	based on GP regression using negative binomial likelihood functions	Python	https://github.com/ManchesterBioinference/Gpcounts	BinTayyash et al.[24]
BayesSpace	based on the Bayesian statistical model	R	https://github.com/edward130603/BayesSpace	Zhao et al.[25]
Machine-learning-based methods				
RayleighSelection	based on the extension of the graph Laplacian method	R	https://github.com/CamaraLab/RayleighSelection	Govek et al.[29]
SOMDE	based on a self-organizing map neural network	Python	https://github.com/XuegongLab/somde	Hao et al.[35]
SPADE	based on convolutional neural network	Python/R	https://github.com/mexchy1000/spade	Bae et al.[36]
Spatial-grid-based methods				
singleCellHaystack	based on the grid and grid points and binary gene expression values	R	https://github.com/alexisvdb/singleCellHaystackhttps://cran.r-project.org/package=singleCellHaystack	Vandenbon et al.[37]
HMRF	based on spatial genes and neighborhood network to detect spatial domains	Python	https://bitbucket.org/qzhudfci/smfishhmrf-py/src/master/	Zhu et al.[40]
Meringue	based on Delaunay triangulation and spatial auto-correlation statistic	R	https://jef.works/MERINGUE/	Miller et al.[38]
BinSpect	based on Delaunay triangulation and statistical enrichment test	R	http://spatialgiotto.rc.fas.harvard.edu/	Dries et al.[39]

SPARK-X is an effective supplement to SPARK when dealing with large and sparse SRT data. Based on the non-parametric modeling, SPARK-X effectively reduces memory requirements and computational times while keeping a reliable model effectiveness.[21]

Negative binomial distribution refers to a discrete probability distribution, which is suitable for the characteristics of overdispersion and zero-inflated counts in single-cell data.[22,23] GPcounts takes advantage of the Gaussian process regression method, which implements negative binomial likelihood models (sometimes zero-inflated negative binomial [ZINB]) for modeling SRT data, achieving a better fit than the Gaussian likelihood function when dealing with count data.[24] The average of the negative binomial likelihood variation is modeled based on the logarithmic link function. GPcounts provides one- and two-sample tests to infer differentially expressed genes across space in spatial count data. In the one-sample test, the null hypothesis is a Gaussian model with no spatial variability in gene expression and no covariance between cells. There are two null hypotheses in a two-sample test: (1) the gene expression under two conditions shows no difference and (2) after constructing three GPs, each sample uses a GP, and the remaining GP is shared between the two samples. GPcounts implemented χ^2 distribution to assess the p value of each gene. Furthermore, GPcounts can use the ZINB model instead of the negative binomial model to cope with the data containing too many zeros.

A fully Bayesian statistical method called BayesSpace has recently been introduced to improve the resolution of SRT data based on the information from spatial neighborhoods and to perform spatially clustering analysis to infer clusters with similar gene expression patterns. BayesSpace overcomes the limitation in efficiently utilizing spatial information for gene expression data clustering and the limited resolution of the original data.[25]

Machine-learning-based methods

Spectral-based methods

Machine learning methods have been widely used in single-cell and SRT studies.[26,27] As a type of machine-learning-based method, spectral-based methods have emerged as a way to perform an unsupervised feature selection based on the degree of consistency between features and the underlying structure. A schematic workflow of spectral-based methods is illustrated in Figure 2. For each input feature (gene), the nearest neighbor graph is constructed firstly to connect each node (cell) associated with the same topic in space using the k-nearest neighbor (KNN) algorithm. The weight of the edge is measured using distance measurement methods, such as Euclidean distance. According to this nearest neighbor graph, an adjacency matrix (A) is constructed to represent the weight value between edges, where a node without an edge connection is 0. The larger the weight, the smaller the distance or difference between nodes.

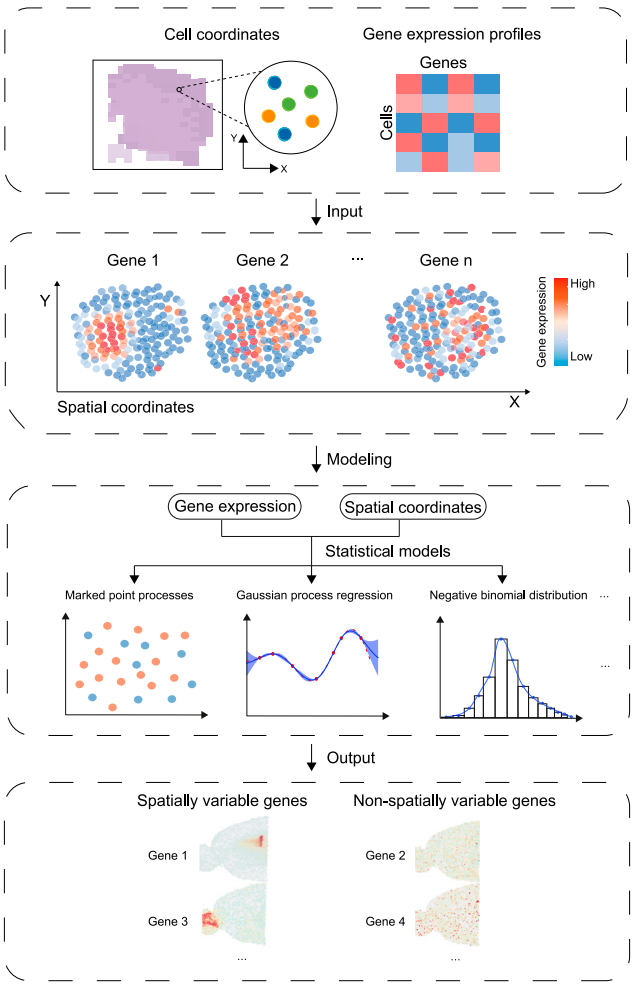

Figure 1. Schematic workflow of statistical-modeling-based strategies
Cell coordinates and gene expression profiles are input to represent the spatial distribution of gene expression. Then, spatial coordinates and gene expression values are modeled. Finally, significantly spatially variable genes are obtained by calculating statistical indicators.

$$A = e^{-\frac{\|X_i - X_j\|_2^2}{t}} \; X_i \text{ is not } X_j'\text{s nearest neighbor,}$$
and t is a suitable constant (Equation 1)

$$A = 0 \; X_i \text{ is } X_j'\text{s nearest neighbor} \quad \text{(Equation 2)}$$

Next, a degree matrix (D) is constructed to represent the number of edges at each node. And the Laplacian matrix (L) can be expressed as follows:

$$L = D - A \quad \text{(Equation 3)}$$

Finally, the Laplacian score is calculated to measure the correlation between the cell coordinates in the space and the gene expression value for each gene. In general, a minor Laplacian gene score indicates a higher value in nodes connected in a local structure and a strong correlation between the gene expression and the spatial coordinates.[28]

RayleighSelection expanded the graph-based Laplacian method, used a simplicial complex that significantly simplified the association among data, and performed a feature selection on features with a complex combinatorial structure.[29] The principle of RayleighSelection is the construction of a simplicial complex using the Vietoris-Rips complex (the simplicial complex is similar to the nearest neighbor graph in graph-based Laplacian methods). RayleighSelection introduced the combinatorial Laplacian score that extended ordinary Laplacian scores to high-dimensional relationships (such as triangles and tetrahedrons) in data. The spatial expression patterns of genes in SRT data are ranked according to their scores, and genes with low scores have highly variable spatial patterns. The statistical significance of the combinatorial Laplacian score is measured by random estimation.

Neural-network-based methods

Due to the feature-enriched and well-structured input data, neural networks, another important branch of machine learning, have been widely used to analyze scRNA-seq and SRT data.[30–34] As shown in Figure 2, neural networks are composed of three parts: an input layer for accepting input information, an output layer, and hidden layers composed of multiple neurons and links. Neurons are commonly divided into multiple hidden layers. Data with high complexity enter from the input layer and activate each hidden layer. Finally, the simplified data are output, effectively reducing the dimensionality of SRT data for the subsequent analysis.

SOMDE uses a self-organizing map (SOM), a competitive learning algorithm for dimensionality reduction according to the topological relationship between hidden layers.[35] A condensed map is constructed with fewer nodes based on the density and topology of the input data while maintaining the original spatial information, and then SVGs are detected by GP.[29] The process of SOMDE can be divided into two main steps. During the first step, SOMDE integrates cells that were close to each other into different nodes using the SOM and generates the weight vectors using the coordinates of nodes in space. Each node in the SOM contained a set of adjacent cells mapped to the node. In the second step, SOMDE models the spatial correlation in each node using specific statistical models, such as the GP. SOMDE scored the spatial variability of each gene with the maximum likelihood value and ranked SVGs.

SPADE uses imaging data and spatial transcriptomic data as inputs, extracting the morphological features around each spot by the convolutional neural network and combining them with gene expression data to identify critical genes associated with spatial and morphological heterogeneity.[36] In addition, a functional analysis can be performed based on these critical genes to further elucidate the biological processes responsible for distinct morphological features.

Spatial-grid-based methods

The schematic workflow of spatial-grid-based methods is illustrated in Figure 3. This class of methods aim to divide the space into multiple

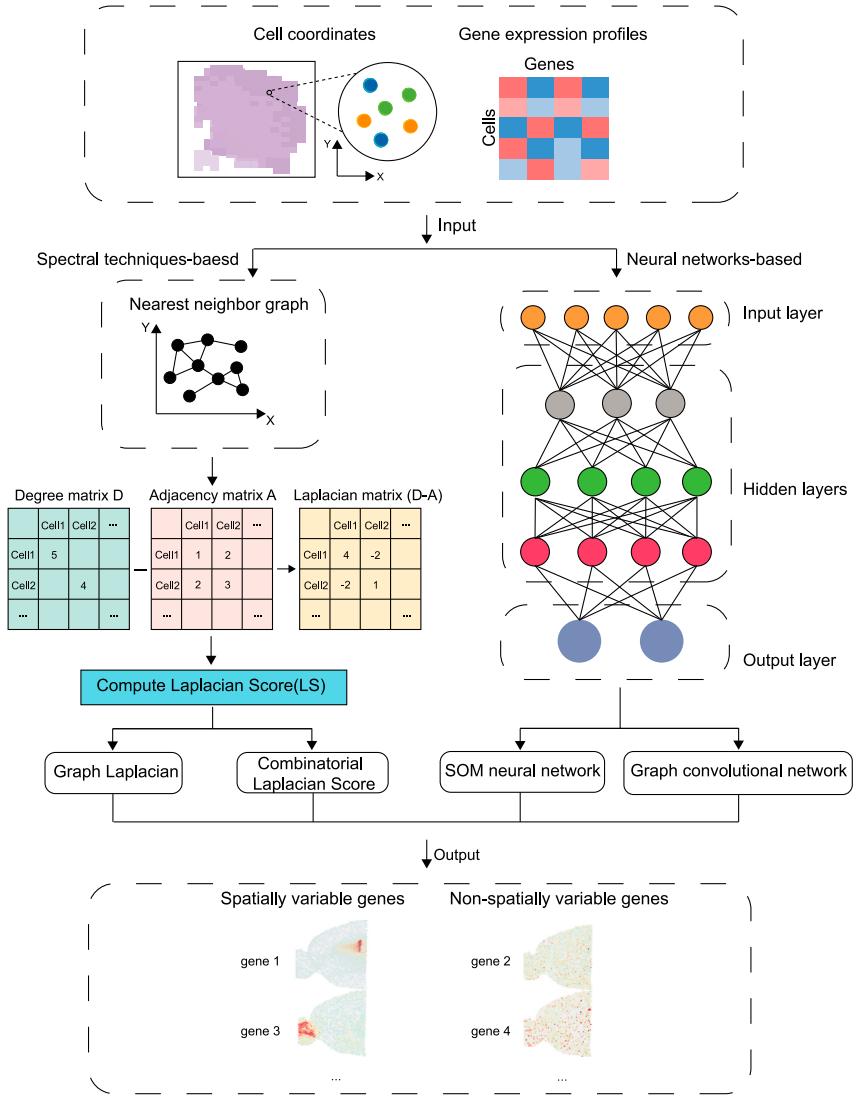

Figure 2. Schematic workflow of machine-learning-based strategies

Spectral-based methods first construct a nearest neighbor graph according to the input data and then calculate the Laplacian matrix and score. Graph and combinatorial Laplacian scores are used to identify the spatially variable genes. Neural-network-based methods process the input data through a graph convolutional neural network or SOM to identify spatially variable genes. SOM, self-organizing map.

grids and to encode spatial relationships among different cells or infer the distribution of cells then apply subsequent steps, such as binarizing the cells' spatial adjacent relationships or gene expression levels for the identification of SVGs.

SingleCellHaystack divides the space into grids and determines multiple grid points on this grid according to the density of cells.[37] For each gene, SingleCellHaystack clusters all cells into two categories by a hard threshold: cells with the gene detected, and cells with the gene not detected. Then, SingleCellHaystack calculates the distribution of these two categories of cells and compares them with a random distribution of cells in space. Kullback-Leibler divergence is used to calculate the D_{KL} score for each gene as the degree of variation and identify genes not uniformly expressed in a multidimensional space. Based on this score, the spatial variability of genes can be evaluated.

MERINGUE considers each cell in SRT data as a neighborhood through Delaunay triangulation then determines whether each cell pair is adjacent according to these neighborhoods and applies a binary adjacency weight matrix to represent this relationship. Depending on the constructed adjacency matrix and gene expression matrix, MERINGUE computes the spatial auto-correlation statistic, Moran's I, to obtain significant spatial genes.[38] In addition, MERINGUE classifies the identified spatial genes into multiple spatial expression patterns through a spatial cross-correlation index.

Giotto has been developed as a toolbox for analyzing and visualizing SRT data and incorporates four approaches to identify spatial genes, including trendsceek, SpatialDE, SPARK, and BinSpect. BinSpect first creates a spatial grid using Delaunay triangulation to represent the association between cells.[39] For each gene being inputted, BinSpect will binarize the gene expression value through K-means clustering or rank threshold and calculate a contingency table between neighboring cells according to these binarized expression values. Using a statistical enrichment test, if a gene is significantly highly expressed in neighboring cells, this gene will be regarded as a SVG.

As a graph-based model, the hidden Markov random fields (HMRFs) approach utilizes spatial genes and the spatial neighborhood network to summarize primary spatial domains.[40] First, the state of each cell will be inferred, which is determined by two factors (the gene expression pattern and the state of its neighbor surrounding cells), then each cell is assigned to a specific spatial domain according to their cell state.

DISCUSSION

In recent years, SRT technologies have continued to develop and become a novel paradigm of disease research. Accumulating evidence suggests that the association between spatial locations and gene expression levels of cells in tissue plays a critical role in diseases, particularly in the tumor mechanism and microenvironment.[41,42]

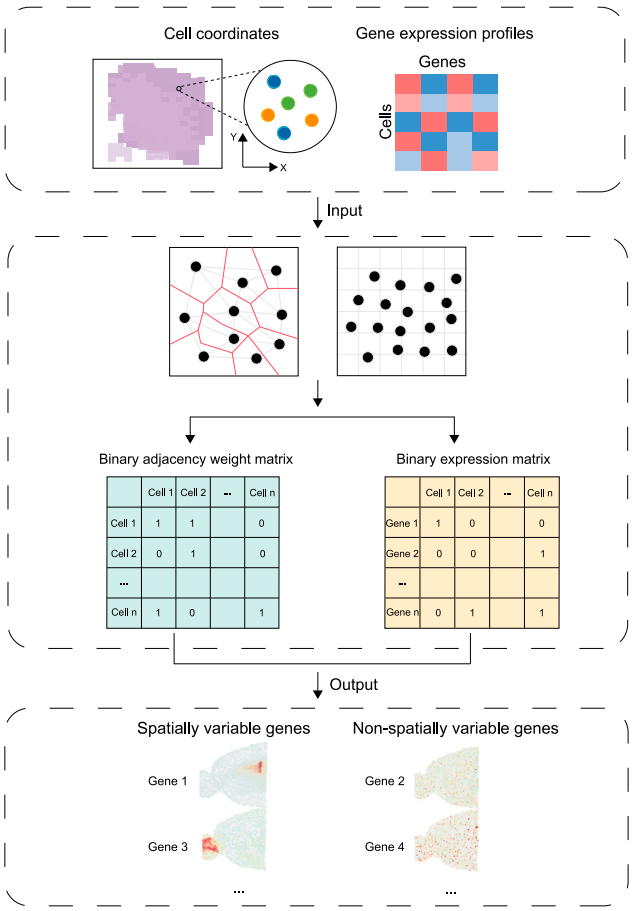

Figure 3. Schematic workflow of spatial-grid-based methods

Spatial-grid-based methods acquire cell coordinates and gene expression profiles as the input and divide the space into grids and encode spatial relationships or infer the distribution of cells, then they apply subsequent steps such as binarizing the cells' spatial adjacent relationships or gene expression levels to identify spatially variable genes.

Identifying genes with spatially variable expression patterns is the critical task of SRT data analysis, reflecting communication between adjacent cells, position-specific states, or cells that migrate to a specific tissue.[18] It has provided a broad range of applicability in identifying tumor markers associated with specific tissue areas, as well as informed targeted treatment, explored the spatial expression patterns related to the specific functions, and provided insights into the origins of tumor heterogeneity.

Currently available databases and computational resources relevant to SRT studies were systematically collected and reviewed in the present study. Next, a detailed overview of current computational and analytical strategies and tools for elucidating spatial gene expression variation from SRT data at single-cell or subcellular resolution was presented. Based on the intrinsic principle, these existing methods could be classified into three categories: (1) statistical-modeling-based methods, (2) machine-learning-based methods, and (3) spatial-grid-based methods.

Typically, the statistical-modeling-based methods involved a statistical analysis performed on each gene and detected significant SVGs through different statistical methods. However, trendsceek calculates test statistics using the permutation test, but this significantly increases the time-cost, and the downstream biological explanation of genes is lacking. The advantage of SpatialDE over trendsceek is that it uses the automatic expression histology method, which can detect related biological features. In addition, due to the exertion of efficient linear mixed models, SpatialDE guarantees a higher computational efficiency than trendsceek. However, SpatialDE approximates the data distribution to a typical model, which may lead to type I errors, and the p value generated by SpatialDE is relatively conservative. SPARK models the spatial counts data directly and shows an efficient performance in low counts data. Compared with SPARK, SPARK-X exhibits a higher calculation efficiency when used for larger-scale SRT data. GPcounts implements negative binomial (NB) or ZINB on GP; although ZINB is suitable for zero-inflated single-cell data, the NB distribution is usually used to model gene count data due to its simplicity.[43] Unlike trendsceek's and SpatialDE's modelings of normalized data, SPARK, SPARK-X, and GPcounts directly input the raw counts to start processing, which could more effectively account for the mean-variance relationship reflecting the discrete characteristics of data. However, for most methods, the defect of poor computational efficiency and high time-cost is still the most significant challenge preventing their application to large-scale SRT data.

On the contrary, machine-learning-based methods significantly reduce the calculation time due to their high computational efficiency and their feature selection for complex data. High-complexity data where the number of features is much greater than the number of observations usually leads to dimensional redundancy and processing difficulties;[44] in order to simplify the understanding of data and reduce redundancy while retaining most of the data information, performing a feature selection and a dimensionality reduction on complex data can significantly reduce the time-cost and improve the efficiency of analysis, which can be achieved using spectral techniques and neural networks.[45] SOMDE is a powerful method for applying large-scale datasets. The running time of SOMDE is associated with the number of genes but not the sample size. The SVGs recognized by the neural network have a more robust biological interpretation. Due to the image data input, SPADE can obtain the SVGs with morphological heterogeneity. However, the results of SPADE may be affected by the size of image patches and the density of spots.

Besides, SingleCellHaystack is clustering-independent and will not cause inaccurate recognition because of clustering deviation. SingleCellHaystack has universal applicability, not only for SRT data but also for scRNA-seq or bulk RNA sequencing (RNA-seq) data, to identify differentially expressed genes. A series of methods incorporate the distance information between cells into the calculation process, but the gene-expression-related spatial variation may be covered up due

to cell-distance differences. As a solution, Meringue put forward a cell neighborhood representation method, which is more stable than the traditional distance-neighbor relationship method, such as KNN, making it suitable for tissues with non-uniform cell density. For a part of spatial-grid-based methods (SingleCellHaystack and BinSpect), one common defect is that the gene expression values are binarized rather than being provided in continuous observations. There were only two indicators for gene expression values in cells (present or not). Even if their expression levels differ significantly, two genes may be detected as "present" or "high expression" in the same cell. Furthermore, choosing an appropriate threshold to distinguish the presence or absence of gene expression was still time-consuming and required multiple attempts.

Although these state-of-the-art computational algorithms, approaches, and tools have highlighted the power and necessity of *in silico* techniques in SRT data analysis, the efficient and accurate computer-aided identification and elucidation of spatial gene expression variations is in its infancy, with significant challenges remaining. For example, most of these existing methods face high memory requirements and the prevalence of zero values in data. In addition, there is a lack of ground truth to better systematically compare these methods. Usually, the running time, memory, and detected gene numbers are commonly used for improved systematic comparisons. Some studies also use statistical indicators, such as the Moran's I statistic, to assess the credibility of the SVGs identified by these methods. With the development of SRT technologies, tens of thousands of spatial sequencing sites can be measured in one sample, significantly increasing the scale of SRT data. Therefore, the need to adapt to the increasingly larger data size also increases the methods' scalability and computational complexity. Various sequencing technologies have been established to measure gene expression levels with spatial context. However, few approaches can measure the spatial information of a single cell at the scale of the whole transcriptome, resulting in the exclusion of some actual SVGs in data. As the sequencing depth of the genes measured by SRT increases, the methods of identifying SVGs will be further developed. Finally, these advanced computational tools were implemented in R or Python package. They required command-line operations, limiting their use by medical and life scientists who did not have a computational background.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (grant nos. 61973240 and 62072341). The funders had no roles in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

M.Z. and J.S. designed the study. K.L., C.Y., C.L., L.C., J.Z., Z.Z., and S.B. collected and reviewed literature. K.L., M.Z., and J.S. drafted the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zheng, Y., Chen, Z., Han, Y., Han, L., Zou, X., Zhou, B., Hu, R., Hao, J., Bai, S., Xiao, H., et al. (2020). Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment. Nat. Commun. *11*, 6268. https://doi.org/10.1038/s41467-020-20019-0.

2. Chen, Y.P., Yin, J.H., Li, W.F., Li, H.J., Chen, D.P., Zhang, C.J., Lv, J.W., Wang, Y.Q., Li, X.M., Li, J.Y., et al. (2020). Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. Cell Res. *30*, 1024–1042. https://doi.org/10.1038/s41422-020-0374-x.

3. Chen, Z., Zhou, L., Liu, L., Hou, Y., Xiong, M., Yang, Y., Hu, J., and Chen, K. (2020). Single-cell RNA sequencing highlights the role of inflammatory cancer-associated fibroblasts in bladder urothelial carcinoma. Nat. Commun. *11*, 5077. https://doi.org/10.1038/s41467-020-18916-5.

4. Bao, S., Li, K., Yan, C., Zhang, Z., Qu, J., and Zhou, M. (2021). Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. Brief. Bioinform. https://doi.org/10.1093/bib/bbab473.

5. Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat. Biotechnol. *38*, 333–342. https://doi.org/10.1038/s41587-019-0392-8.

6. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell *173*, 1307. https://doi.org/10.1016/j.cell.2018.05.012.

7. Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J.R., Bartonicek, N., et al. (2021). A single-cell and spatially resolved atlas of human breast cancers. Nat. Genet. *53*, 1334–1347. https://doi.org/10.1038/s41588-021-00911-1.

8. Asp, M., Giacomello, S., Larsson, L., Wu, C., Furth, D., Qian, X., Wardell, E., Custodio, J., Reimegard, J., Salmen, F., et al. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell *179*, 1647–1660.e19. https://doi.org/10.1016/j.cell.2019.11.025.

9. Hu, J., Schroeder, A., Coleman, K., Chen, C., Auerbach, B.J., and Li, M. (2021). Statistical and machine learning methods for spatially resolved transcriptomics with histology. Comput. Struct. Biotechnol. J. *19*, 3829–3841. https://doi.org/10.1016/j.csbj.2021.06.052.

10. Navarro, J.F., Croteau, D.L., Jurek, A., Andrusivova, Z., Yang, B., Wang, Y., Ogedegbe, B., Riaz, T., Stoen, M., Desler, C., et al. (2020). Spatial transcriptomics reveals genes associated with dysregulated mitochondrial functions and stress signaling in Alzheimer disease. iScience *23*, 101556. https://doi.org/10.1016/j.isci.2020.101556.

11. Wang, Y., Ma, S., and Ruzzo, W.L. (2020). Spatial modeling of prostate cancer metabolic gene expression reveals extensive heterogeneity and selective vulnerabilities. Sci. Rep. *10*, 3490. https://doi.org/10.1038/s41598-020-60384-w.

12. Stahl, P.L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., and Huss, M. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

13. Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse Hippocampus. Neuron *92*, 342–357. https://doi.org/10.1016/j.neuron.2016.10.001.

14. Fan, Z., Chen, R., and Chen, X. (2020). SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Res. *48*, D233–D237. https://doi.org/10.1093/nar/gkz934.

15. Vickovic, S., Eraslan, G., Salmen, F., Klughammer, J., Stenbeck, L., Schapiro, D., Aijo, T., Bonneau, R., Bergenstrahle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*, 987–990. https://doi.org/10.1038/s41592-019-0548-y.

16. Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. *39*, 313–319. https://doi.org/10.1038/s41587-020-0739-1.

17. Edsgard, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. Nat. Methods *15*, 339–342. https://doi.org/10.1038/nmeth.4634.

18. Svensson, V., Teichmann, S.A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. Nat. Methods *15*, 343–346. https://doi.org/10.1038/nmeth.4636.

19. Gibbs, M.N. (1997). Bayesian Gaussian Processes for Regression and Classification (Ph.D. thesis, Inferential Sciences Group of the Cavendish Laboratory, Cambridge University).

20. Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat. Methods *17*, 193–200. https://doi.org/10.1038/s41592-019-0701-7.

21. Zhu, J., Sun, S., and Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biol. *22*, 184. https://doi.org/10.1186/s13059-021-02404-0.

22. Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. *16*, 241. https://doi.org/10.1186/s13059-015-0805-z.

23. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. https://doi.org/10.1038/s41467-017-02554-5.

24. BinTayyash, N., Georgaka, S., John, S.T., Ahmed, S., Boukouvalas, A., Hensman, J., and Rattray, M. (2021). Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab486.

25. Zhao, E., Stone, M.R., Ren, X., Guenthoer, J., Smythe, K.S., Pulliam, T., Williams, S.R., Uytingco, C.R., Taylor, S.E.B., Nghiem, P., et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-00935-2.

26. Loher, P., and Karathanasis, N. (2020). Machine learning approaches identify genes containing spatial information from single-cell transcriptomics data. Front Genet. *11*, 612840. https://doi.org/10.3389/fgene.2020.612840.

27. Turki, T., and Taguchi, Y.H. (2020). SCGRNs: novel supervised inference of single-cell gene regulatory networks of complex diseases. Comput. Biol. Med. *118*, 103656. https://doi.org/10.1016/j.compbiomed.2020.103656.

28. Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. *15*, 1373–1396.

29. Govek, K.W., Yamajala, V.S., and Camara, P.G. (2019). Clustering-independent analysis of genomic data using spectral simplicial theory. PLoS Comput. Biol. *15*, e1007509. https://doi.org/10.1371/journal.pcbi.1007509.

30. Camp, J.G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., Kanton, S., Kageyama, J., Damm, G., Seehofer, D., et al. (2017). Multilineage communication regulates human liver bud development from pluripotency. Nature *546*, 533–538. https://doi.org/10.1038/nature22796.

31. Kimmel, J.C., and Kelley, D.R. (2020). scNym: semi-supervised adversarial neural networks for single cell classification. bioRxiv. https://doi.org/10.1101/2020.06.04.132324.

32. Yuan, Y., and Bar-Joseph, Z. (2020). GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. Genome Biol. *21*, 300. https://doi.org/10.1186/s13059-020-02214-w.

33. Ma, F., and Pellegrini, M. (2020). ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics *36*, 533–538. https://doi.org/10.1093/bioinformatics/btz592.

34. Kimmel, J.C., and Kelley, D.R. (2021). Semisupervised adversarial neural networks for single-cell classification. Genome Res. https://doi.org/10.1101/gr.268581.120.

35. Hao, M., Hua, K., and Zhang, X. (2021). SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab471.

36. Bae, S., Choi, H., and Lee, D.S. (2021). Discovery of molecular features underlying the morphological landscape by integrating spatial transcriptomic data with deep features of tissue images. Nucleic Acids Res. *49*, e55. https://doi.org/10.1093/nar/gkab095.

37. Vandenbon, A., and Diez, D. (2020). A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. Nat. Commun. *11*. https://doi.org/10.1038/s41467-020-17900-3.

38. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X., and Fan, J. (2021). Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities. Genome Res. https://doi.org/10.1101/gr.271288.120.

39. Dries, R., Zhu, Q., Dong, R., Eng, C.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. *22*, 78. https://doi.org/10.1186/s13059-021-02286-2.

40. Zhu, Q., Shah, S., Dries, R., Cai, L., and Yuan, G.C. (2018). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat. Biotechnol. https://doi.org/10.1038/nbt.4260.

41. Saviano, A., Henderson, N.C., and Baumert, T.F. (2020). Single-cell genomics and spatial transcriptomics: discovery of novel cell states and cellular interactions in liver physiology and disease biology. J. Hepatol. *73*, 1219–1230. https://doi.org/10.1016/j.jhep.2020.06.004.

42. Smith, E.A., and Hodges, H.C. (2019). The spatial and genomic hierarchy of tumor ecosystems revealed by single-cell technologies. Trends Cancer *5*, 411–425. https://doi.org/10.1016/j.trecan.2019.05.009.

43. Choi, K., Chen, Y., Skelly, D.A., and Churchill, G.A. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. Genome Biol. *21*, 183. https://doi.org/10.1101/2020.03.03.974808.

44. Zhang, J., Huang, H., and Wang, J. (2010). Manifold learning for visualizing and analyzing high-dimensional data. IEEE Intell. Syst. *25*, 54–61. https://doi.org/10.1109/MIS.2010.8.

45. Roweis, S., and Saul, L.J.S. (2000). Nonlinear dimensionality reduction by locally. Linear Embedding *290*, 2323–2326.